Model Selection

DS 301

Iowa State University

This week's agenda

- Midterm 1 details.
- HW 4 due this Wednesday. HW 5 will be posted next week.
- Introduction to Model Selection.

Motivation for model selection

Up to this point, we know how to fit a model for a set of p predictors $X_1, \ldots, X_p$.

- How do we know which subset of predictors are important?
- Can we streamline our model? $X_1, \ldots, X_p$
- Data acquisition can be expensive.

$M_1: Y \sim X_1 + X_1^2 + X_1^3 + X_2$

$M_2: Y \sim X_1 + X_2 + X_3$

Model selection

The process of removing “irrelevant” or “less important” predictors is called:

- Model selection
- Feature selection/ feature screening. (CS)
- Variable selection (statistics)

We know that we cannot just look at a model with all p predictors and remove non-significant predictors (why?). We need some more systematic techniques.

Possible models

X_1, X_2, X_3

Suppose we are considering 3 potential predictors. How many possible models are there? $2^3 = 8$ possible models

$Y \sim X_1 + X_2 + X_3$

$X_1 + X_2$

$X_1 + X_3$

$X_2 + X_3$

X_1

X_2

X_3

intercept only

p predictors $\Rightarrow$ 2^p possible models

$2^{10} = 1024$ possible models

$2^{20} = (1024)^2 > 1$ million possible models.

Model selection Strategy

(1) p is small ($p < 30$)

↳ # of predictors

↳ exhaustive search

$\Rightarrow$ 'best' subset selection

(2) p is large ($p \geq 30$)

↳ greedy algorithms

- forward selection
- backward selection
- hybrid (stepwise) selection

Best subset selection (p<30) p is 'small'

algorithm:

(1) For all $k = 1, \ldots, p$

(a) Fit all $\binom{p}{k}$ models that contain exactly k predictors (size k).

(b) pick the best among these $\binom{p}{k}$ models. call this M_k.

'best' is the model that has the smallest RSS.

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

(2) $M_1, M_2, M_3, \ldots\ldots, M_p$.

select best model from among these candidate models.

'best' is according to some criteria.

Best subset selection

This is an exhaustive search.
It will always produce an optimal model based on some criteria.

But not computationally efficient (np hard problem).

⇒ What criteria should we use to pick our final model from $M_1, M_2, \ldots, M_p$?

What criteria should we use to pick final model?

Why can we not define 'best' here as the model with the smallest RSS?

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

RSS will decrease (or stay the same) as we add predictors to our model.

$$R^2 = 1 - \frac{RSS}{TSS}$$

Instead:

AIC

BIC

adjusted R^2

Mallow's C_p

Intuition

adding irrelevant predictors will lead to only a small decrease in RSS.

↳ each predictor pays a price to be in the model (penalty).

↳ if the decrease in RSS is not enough to offset this price

→ this will lead to a relatively large BIC/AIC

(we want model w/ smallest AIC/BIC).

Criteria for model selection

- $AIC = n \cdot \log\left(\frac{RSS}{n}\right) + \underbrace{2p}_{\text{penalty}}$.
- $BIC = n \cdot \log\left(\frac{RSS}{n}\right) + \underbrace{p \cdot \log(n)}_{\text{penalty}}$

smaller is better.

- Mallow's $Cp = \frac{RSS}{\hat{\sigma}^2} - n + \underbrace{2p}_{\text{penalty}}$ (smaller is better)

  $\hat{\sigma}^2$ (estimate)

  $\hat{\sigma}^2 = \frac{\sum^n e_i^2}{(n-(p+1))}$.

- adjusted $R^2 = 1 - \frac{RSS/(n-(p+1))}{\sum_{i=1}^n (y_i - \bar{y})^2/(n-1)}$

(bigger is better).

AIC vs. BIC

$M_1, M_2, \ldots, M_p.$

$M_4: \; Y \sim X_1 + X_2 + X_5 + X_7.$

- AIC used more frequently
  - Assumes true model not in candidate pool $M_1, \ldots, M_p$
  - Tries to mimic the true model.

$M_2: \; Y \sim X_4 + X_6.$

- BIC has a heavier penalty term.
  - Will usually pick a simpler/more parsimonious model.
  - BIC Is consistent: if true model is among candidate pool it will eventually lead you to the true model (if n is large enough).

For you to think about:

$M_1, M_2, \ldots, M_p.$

$M_1: Y \sim X_3$

$M_2: Y \sim X_3 + X_1$

$RSS_2 \leq RSS_1$

- Will these 4 criteria (AIC, BIC, Mallow's C_p, adjusted R^2) lead you to pick the same final model ?

$M_1: Y \sim X_3$

$M_2: Y \sim X_1 + X_2$

temp

# of clouds

- Suppose I have two models:

$$M_3 : Y \sim X_1 + X_2 + X_4$$
$$M_4 : Y \sim X_4 + X_5 + X_6 + X_7$$

Is it true that $\text{RSS}_3 \geq \text{RSS}_4$?

$RSS_4 \leq RSS_3$?

For you to think about:

- Suppose I have 3 models to pick from:

$$
\begin{aligned}
M_A &: Y \sim X_1 + X_2 + X_3 + X_4 + X_5 \\
M_B &: Y \sim X_6 + X_7 + X_8 + X_9 + X_{10} \\
M_C &: Y \sim X_1 + X_2 + X_7 + X_9 + X_{10}
\end{aligned}
$$

Will using AIC, BIC, Mallow's C_p, adjusted R^2 lead you to pick the same final model ?

Implementation

See R script: `subsetselection.R`

Test MSE

An alternative to the approaches we discussed is to directly estimate the test error by splitting the dataset into a training set and test set.

Drawbacks to this approach