STAT 420: Methods of Applied Statistics

Model Selection

Ruoqing Zhu, Ph.D. <rqzhu@illinois.edu>

Course website: https://sites.google.com/site/teazrq/teaching/STAT420

Department of Statistics
University of Illinois at Urbana-Champaign
April 6, 2017

Model selection and variable selection

- When the number of variables is large, we may want to restrict the number of predictors to a reasonable size.
  - When $p < n$, we can still fit the linear model, however, it will be unstable when p is large.
  - When $p \geq n$, we need some other technique.
- We focus on the first situation.

Comparing different models

- Suppose we have a total of p predictors. For each predictor, we can choose to include or not to include it. Then the combination number suggests that there are a total of 2^p combinations.
- When $p = 20$, we have over 1 million models. How to compare them?
  - If we are interested in a nested model comparison of a reduced model and full model we could use the F-test.
  - What if they are not nested? We will introduce the adjusted R^2, Mallow's C_p, AIC, BIC.
  - We will also introduce algorithms that can help us select the best model

R^2 and adjusted R^2

- The most common measure of model fit is R^2

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

- However, R^2 does not take into account the number of parameters used in the model, or the sample size. Hence, we can easily overfit the data.
- We can adjust R^2 for sample size and the number of predictors used to obtain a better estimate:

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-p}$$

- Note that $R^2_{\text{adj}} < R^2$, and is a shrunken estimate.
- However, there is no guarantee that R^2_{adj} is > 0.

Adjusted R^2

```
> library(faraway)
> data(gala)
>
> fit = lm(Species~ Area + Elevation + Nearest + Scruz +
    Adjacent, data = gala)
>
> summary(fit)$r.squared
[1] 0.7658469
>
> summary(fit)$adj.r.squared
[1] 0.7170651
```

Mallow's C_p Criterion

- Many model selection criterion follows this logic:

$$\text{Goodness-of-fit} + \text{Complexity-Penalty}$$

- The first term will decrease as the model gets more complicated, which prefers "larger" model
- The second term increases with the number of predictor variables, which prefers "smaller" model
- The best model is the one that balances the two, i.e., the smallest overall score.
- A popular choice for selecting the best model is Mallow's C_p (1973)

$$\text{SSE} + 2\widehat{\sigma}^2_{\text{full}} \cdot p$$

where σ^2_{full} is the MSE from the full model.

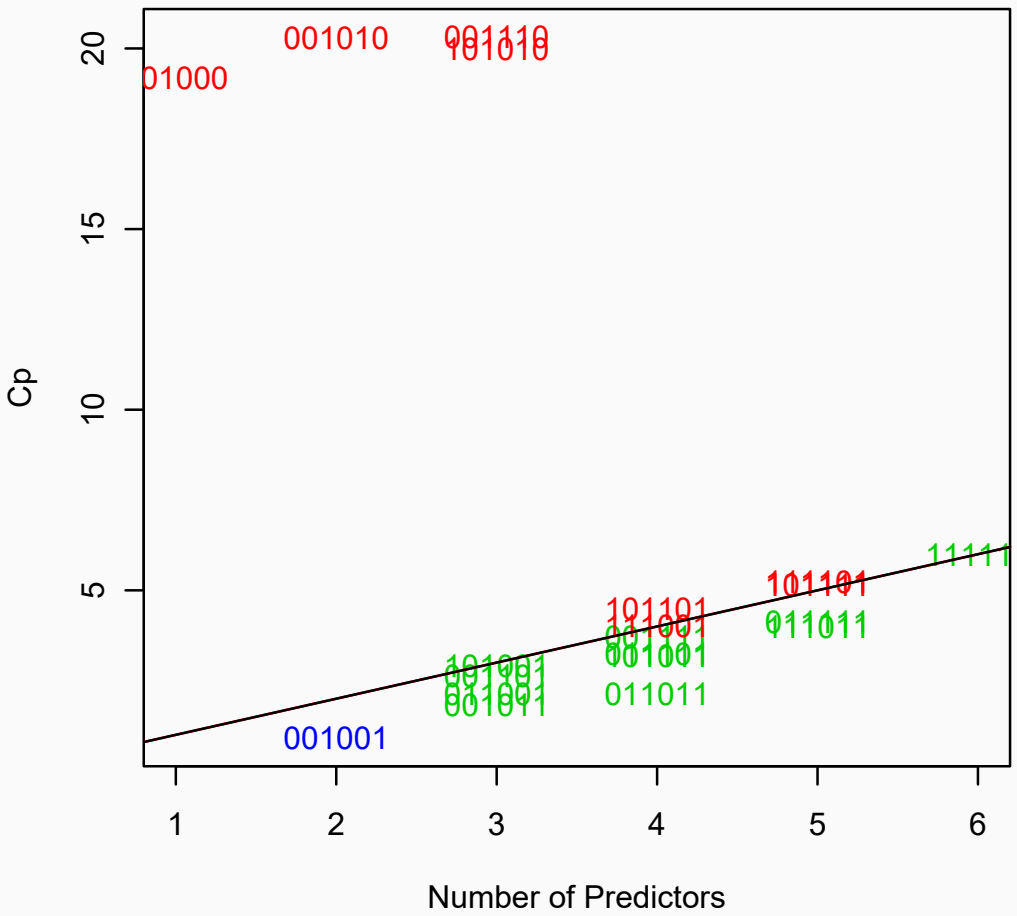
Adjusted R^2

```
> library(wle)
> mle.cp(fit)

Mallows Cp:
      (Intercept) Area Elevation Nearest Scruz Adjacent     cp
 [1,]           0    0         1       0     0        1 0.9062
 [2,]           0    0         1       0     1        1 1.8220
 [3,]           0    1         1       0     1        1 2.1370
 [4,]           0    1         1       0     0        1 2.1440
 [5,]           0    0         1       1     0        1 2.6270

> plot(mle.cp(fit))
```

Mallow's C_p


Cp
Number of Predictors
5
10
15
20
1
2
3
4
5
6
001010
001110
01000
001001
011011
101101
11111

Example

- Compare the following two models using Mallow's C_p. Which is better?
  A). Intercept + Area + Scruz
  B). Intercept + Elevation + Nearest + Adjacent
- What σ^2_{full} should you use?

Training vs. Testing error

- Training data $\mathcal{D}_n = \{x_i, y_i\}_{i=1}^n$
- Suppose $\{x_i, y_i^*\}_{i=1}^n$ is an independent (imaginary) testing dataset collected at the same location x_i's (aka, in-sample prediction)
- Assume that the data are indeed from a linear model

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{e} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$
$$\mathbf{y}^* = \boldsymbol{\mu} + \mathbf{e}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}^*$$

where both $\mathbf{y}$ and $\mathbf{y}^*$ are $n \times 1$ response vectors, $\mathbf{e}$ and $\mathbf{e}^*$ are i.i.d. error terms with mean 0 and variance σ^2.

Training vs. Testing error

$$
\begin{aligned}
\mathsf{E}[\text{Train Err}] &= \mathsf{E}\|\mathbf{y} - \widehat{\mathbf{y}}\|^2 = \mathsf{E}\|(\mathbf{I} - \mathbf{H})\mathbf{y}\|^2 \\
&= \mathsf{E}\|(\mathbf{I} - \mathbf{H})\mathbf{e}\|^2 \\
&= \mathsf{Trace}\big((\mathbf{I} - \mathbf{H})^{\mathsf{T}}(\mathbf{I} - \mathbf{H})\mathsf{Cov}(\mathbf{e})\big) \\
&= (n - p)\sigma^2
\end{aligned}
$$

$$
\begin{aligned}
\mathsf{E}[\text{Test Err}] &= \mathsf{E}\|\mathbf{y}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 \\
&= \mathsf{E}\|(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}) + (\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta})\|^2 \\
&= \mathsf{E}\|\mathbf{y}^* - \boldsymbol{\mu}\|^2 + \mathsf{E}\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2 \\
&= \mathsf{E}\|\mathbf{e}^*\|^2 + \mathsf{Trace}\big(\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathsf{Cov}(\widehat{\boldsymbol{\beta}})\big) \\
&= n\sigma^2 + p\sigma^2
\end{aligned}
$$

So the testing error increase with p and training error decreases with p. When p gets large, this is a big trouble...

Mallow's C_p Criterion

- The SSE is just the training error — thats why we estimated σ^2 by SSE$/(n-p)$.
- The difference between the training and testing errors is $2p\sigma^2$
- Hence, the Mallow's C_p is add the training error SSE by $2p\widehat{\sigma}^2$, so that it mimics the testing error.
- Then the model is generalizable to future data (prediction).

AIC and BIC

- Other popular choices of scores:
  - AIC (Akaike 1970): -2 Log-likelihood $+ 2 \cdot p$
  - BIC (Schwarz, 1978): -2 Log-likelihood $+ \log n \cdot p$
- Both AIC and BIC penalize the number of predictors.
- When n is large, adding an additional predictor costs a lot more in BIC than AIC (or C_p). So AIC tends to pick a larger model than BIC.
- C_p performs similarly to AIC.

AIC and BIC

- Akaike's information Criterion (AIC) is

$$\begin{aligned}\text{AIC} &= -2\,\text{Log-likelihood} + 2p \\ &= n + n\log(2\pi) + n\log(\text{SSE}/n) + 2p\end{aligned}$$

- The Bayes Information Criterion (BIC) is

$$\begin{aligned}\text{BIC} &= -2\,\text{Log-likelihood} + \log(n)p \\ &= n + n\log(2\pi) + n\log(\text{SSE}/n) + \log(n)p\end{aligned}$$

- Note: In R, some default functions will remove part of the constants: $n + n\log(2\pi)$.

Model Selection Algorithms

- Sometimes we may want to use an algorithm to systematically choose the best model.
- We will discuss backward elimination, forward selection, and stepwise regression.
    - These algorithms use a specified rejection level α to decide which variables to include or omit from a model sequentially.
- Another type of algorithm is the best subset selection, which will try all possible combinations.

Backward, forward, and stepwise regression

- For the backward regression:
  - start with all predictors in the model
  - Remove the predictor with the largest p-value above α
  - Continue refitting the model and omitting predictors with p-values $> \alpha$.
- Faraway notes that α could be 0.15 to 0.20 if the goal of the model fitting is prediction.
- See the R code for this approach. Use function update to manually selecting the best model.
- Forward and stepwise approach can be done similarly.

Stepwise regression with AIC or BIC

- The similar algorithm can be applied to AIC or BIC
- The idea is to compare the AIC or BIC values for adding or removing a variable. Process the change if it improves the measurement the best.
- Stop when there is nothing to improve anymore.
- See the R code for this approach. Use function step to directly select the best model.

Best subset selection

- The best subset selection is to exhaustively search for the best model.
- However, this approach can be computationally very expensive when p is large.
- Usually only feasible for $p < 50$
- Algorithm:
  1. For each $k = 1, \ldots, p$, check 2^k possible combinations, and find the model with smallest RSS
     - The penalty term is the same for models with the same size
  2. To choose the best k, use model selection criteria
- See the R code for this approach. Use function regsubsets in the leaps package to directly select the best model.