chapter 10

Regression analysis III

Instructor: Li, Han

Contents

- Wellfitting, overfitting and underfitting
- Model selection appropaches
- ANOVA
- AIC/BIC
- Stepwise regression
- Mallows’ Cp
- Cross Validation (CV)

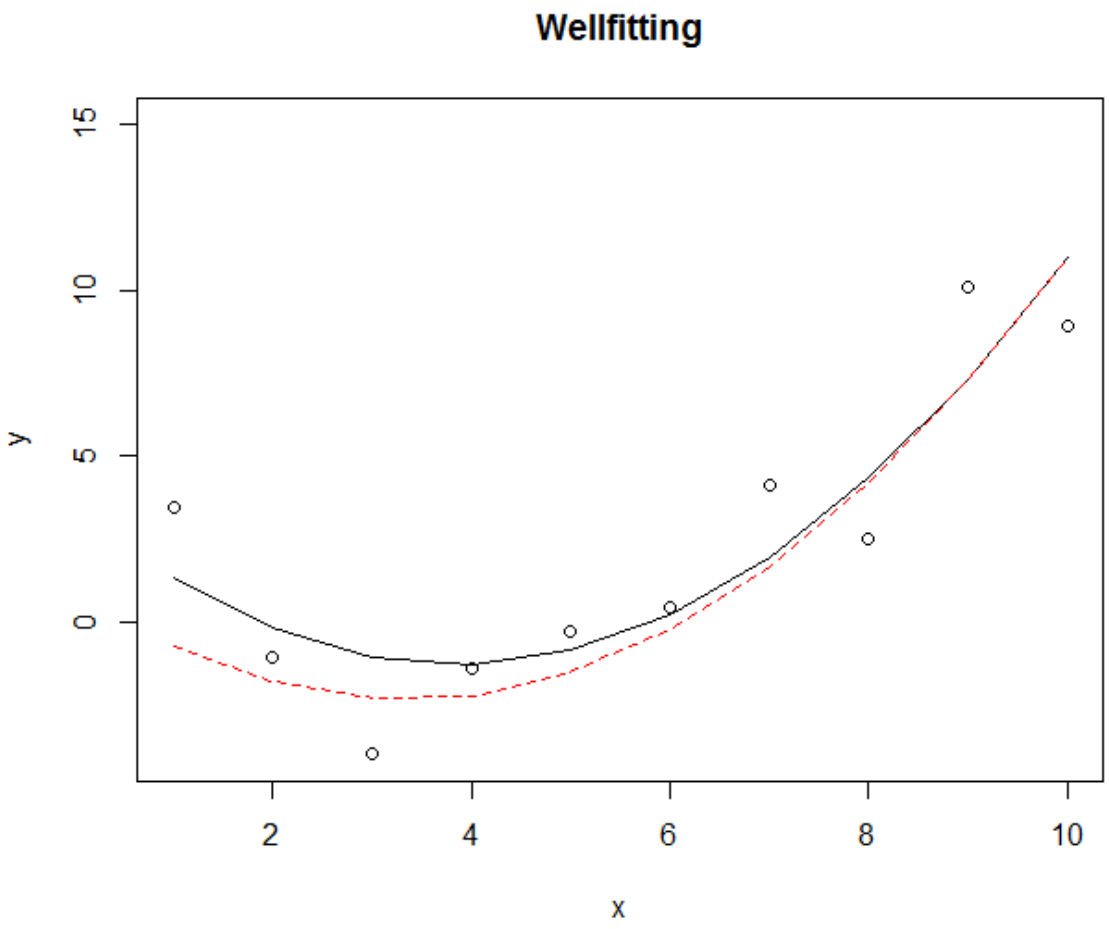
$$Y_i = \beta_o + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i, \;\; i = 1, \ldots, n.$$

When we have multiple explanatory variables and we are not sure which variables are truely associated with the response variable, for the sake of data interpretation and prediction, we face the problem of variable selection, which is also called model selection.

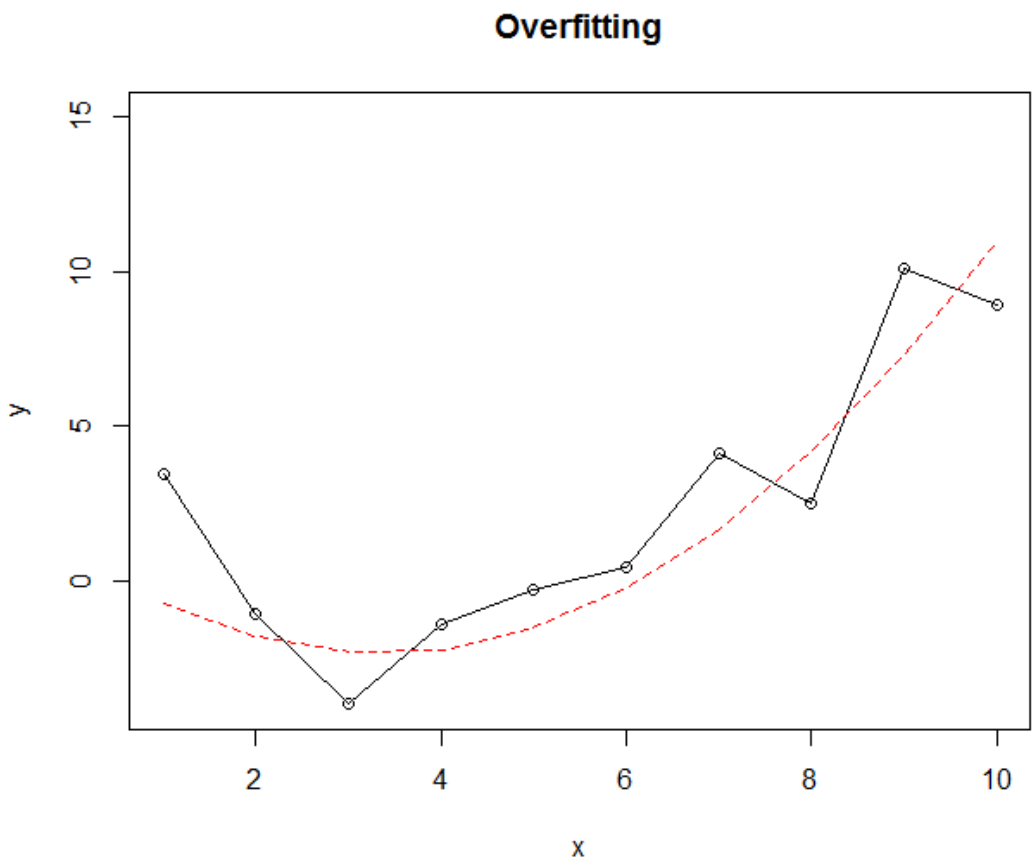
Three cases for fitting the model:

- **Wellfitting** (close to most of the data and the population regression line)
- **Overfitting** (too close to the data and thus deviate from the population regression line)
- **Underfitting** (far from most of the data and the population regression line)

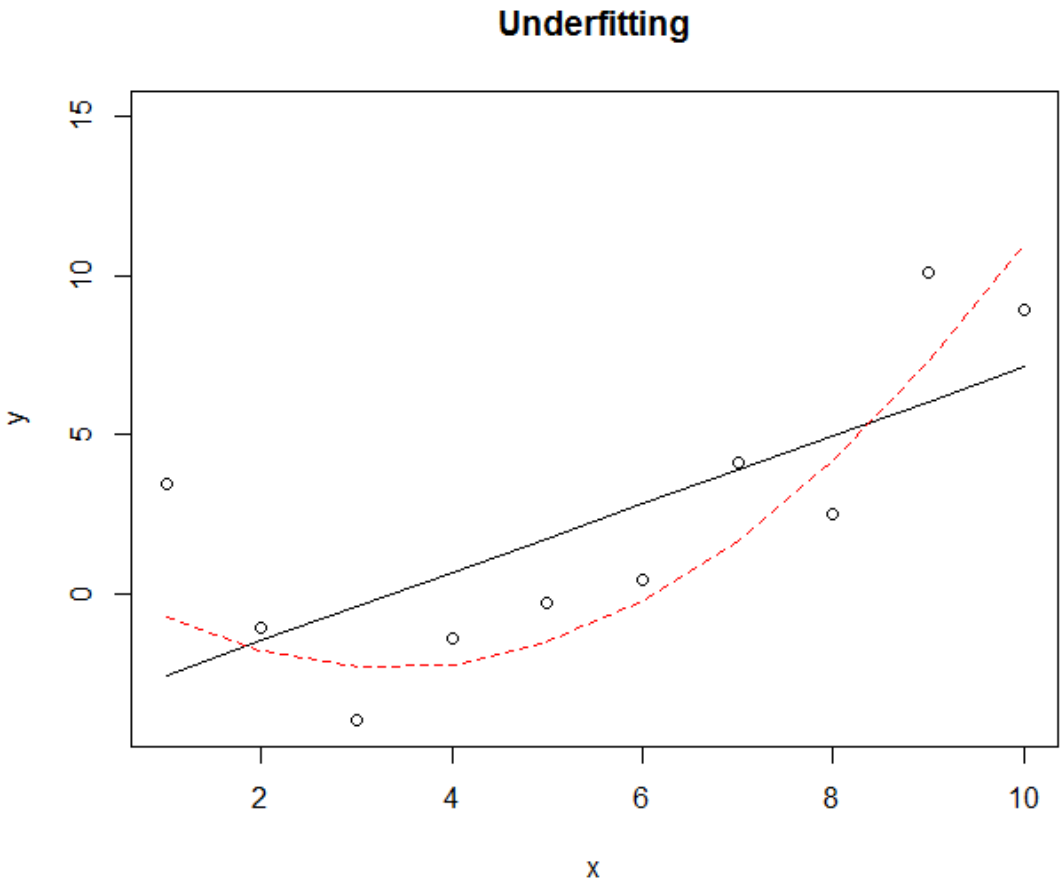
Wellfitting



Overfitting



Underfitting



Selecting the "best" model

The selection of a final regression model always involves a compromise between

- fitting accuracy (a model that fits the data as well as possible) , and
- model complexity (the number of explanatory variables).

If we have two models with approximately equal fitting accuracy, we favor the simpler one.

Model selection approaches

Generally, we have three approaches for model selection.

- Hypothesis testing (for instance, t test or F test)
- Optimize an objective function, which balances two terms: fitting accuracy and model complexity, for instance, AIC or BIC, Mallows' Cp.
- Cross validation, use the data iteself to validate the model

ANOVA

For nested models, we could use anova() function. A nested model is one whose terms are completely included in the other model.

Example 1 (anova test)

```
states <- as.data.frame(state.x77[,c("Murder", "Population",
"Illiteracy", "Income", "Frost")])
fit1 <- lm(Murder ~ Population + Illiteracy, data=states)
fit2 <- lm(Murder ~ Population + Illiteracy + Income + Frost,
data=states)
anova(fit1, fit2)       # hypothesis testing beta3=beta4=0
```

```
> fit1 <- lm(Murder ~ Population + Illiteracy, data=states)
> fit2 <- lm(Murder ~ Population + Illiteracy + Income + Frost, $
> anova(fit1, fit2)
Analysis of Variance Table

Model 1: Murder ~ Population + Illiteracy
Model 2: Murder ~ Population + Illiteracy + Income + Frost
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     47 289.25
2     45 289.17  2  0.078505 0.0061 0.9939
>
```

H_0: beta3=beta4=0. For p-value =0.9939 >>0.05, we retain H_0, and thus prefer the simpler model, that's, Model 1.

AIC

The Akaike Information Criterion (AIC) index takes into account a model's statistical fit and the number of parameters needed to achieve this fit. Models with smaller AIC values—indicating adequate fit with fewer parameters — are preferred.

$$\mathbf{AIC=-2lnL+2k}$$

fitting model complexity

Example 2 (AIC)

AIC(fit1, fit2)

RGui (32-bit) - [R Console]

File Edit View Misc Packages Windows Help

```
> AIC(fit1, fit2)
     df      AIC
fit1  4 237.6565
fit2  6 241.6429
>
```

BIC

The Bayesian Information Criterion (BIC, also called Schwarz Information Criterion) is similar to AIC, but adds a heavier model complexity penalty, by replacing 2 with log(n), where n is the sample size.

$$\mathbf{BIC=-2lnL+log(n)k}$$

fitting model complexity

Example 3 (BIC)

AIC(fit1, fit2, k=log(n))

RGui (32-bit) - [R Console]

File Edit View Misc Packages Windows Help

```
> n=nrow(states)
> AIC(fit1, fit2, k=log(n))
     df      AIC
fit1  4 245.3046
fit2  6 253.1151
>
```

Both the AIC and BIC values suggest that the model without Income and Frost is the better model. Note that although the ANOVA approach requires nested models, the AIC/BIC approach doesn't.

AIC or BIC

Choosing AIC or BIC depends on the problem's context.

- For small or moderate samples, BIC often chooses models that are too simple, because of its heavy penalty on complexity.
- For large sample size, AIC tends to choose more complex models than BIC.

Generally, BIC is more accurate.

Mallows Cp

Mallows Cp addresses the issue of overfitting, with a small value of C_p meaning that the model is relatively precise.

$$\mathbf{Cp=(SSE+2k \times se)/n}$$

fitting model complexity

Where SSE is the sum of square errors, k is the number of predictors, n is the sample size and se is the standard error of the residuals.

Stepwise regression

Given the model selection criterion, variables are added to or deleted from a model one at a time, until the stopping criterion (AIC/BIC/Mallows Cp) is reached. There are three ways to update the model.

- forward
- backward
- stepwise

- **Step forward -** add one predictor variable that improves the model most at a time, stopping when the addition of variables would no longer improve the model.

- **Step backward-** start with a model that includes all predictor variables, and then delete one predictor at a time such that the resulting model has better performance, until removing variables would degrade the quality of the model.

Stepwise

- **Stepwise** - combine the step forward and step backward approaches. Variables are entered one at a time, but at each step, the variables in the model are reevaluated, and those that don’t contribute to the model are deleted.

The above approaches do NOT guarantee to reach the optimal model, due to their discrete variable selection process.

The implementation of stepwise regression methods vary by the criteria used to enter or remove variables. The stepAIC() function in the MASS package performs stepwise model selection (forward, backward, stepwise) using an exact AIC criterion.

Example 4 (backward AIC)

```
library(MASS)
fit1 <- lm(Murder ~ Population + Illiteracy + Income + Frost, data=states)
stepAIC(fit1, direction="backward")
```

RGui (32-bit) - [R Console]

File Edit View Misc Packages Windows Help

```
> library(MASS)
> fit1 <- lm(Murder ~ Population + Illiteracy + Income + Fros$
> stepAIC(fit1, direction="backward")
Start:  AIC=97.75
Murder ~ Population + Illiteracy + Income + Frost

             Df Sum of Sq    RSS     AIC
- Frost       1     0.021 289.19  95.753
- Income      1     0.057 289.22  95.759
<none>                    289.17  97.749
- Population  1    39.238 328.41 102.111
- Illiteracy  1   144.264 433.43 115.986

Step:  AIC=95.75
Murder ~ Population + Illiteracy + Income

             Df Sum of Sq    RSS     AIC
- Income      1     0.057 289.25  93.763
<none>                    289.19  95.753
- Population  1    43.658 332.85 100.783
- Illiteracy  1   236.196 525.38 123.605
```

RGui (32-bit) - [R Console]

File Edit View Misc Packages Windows Help

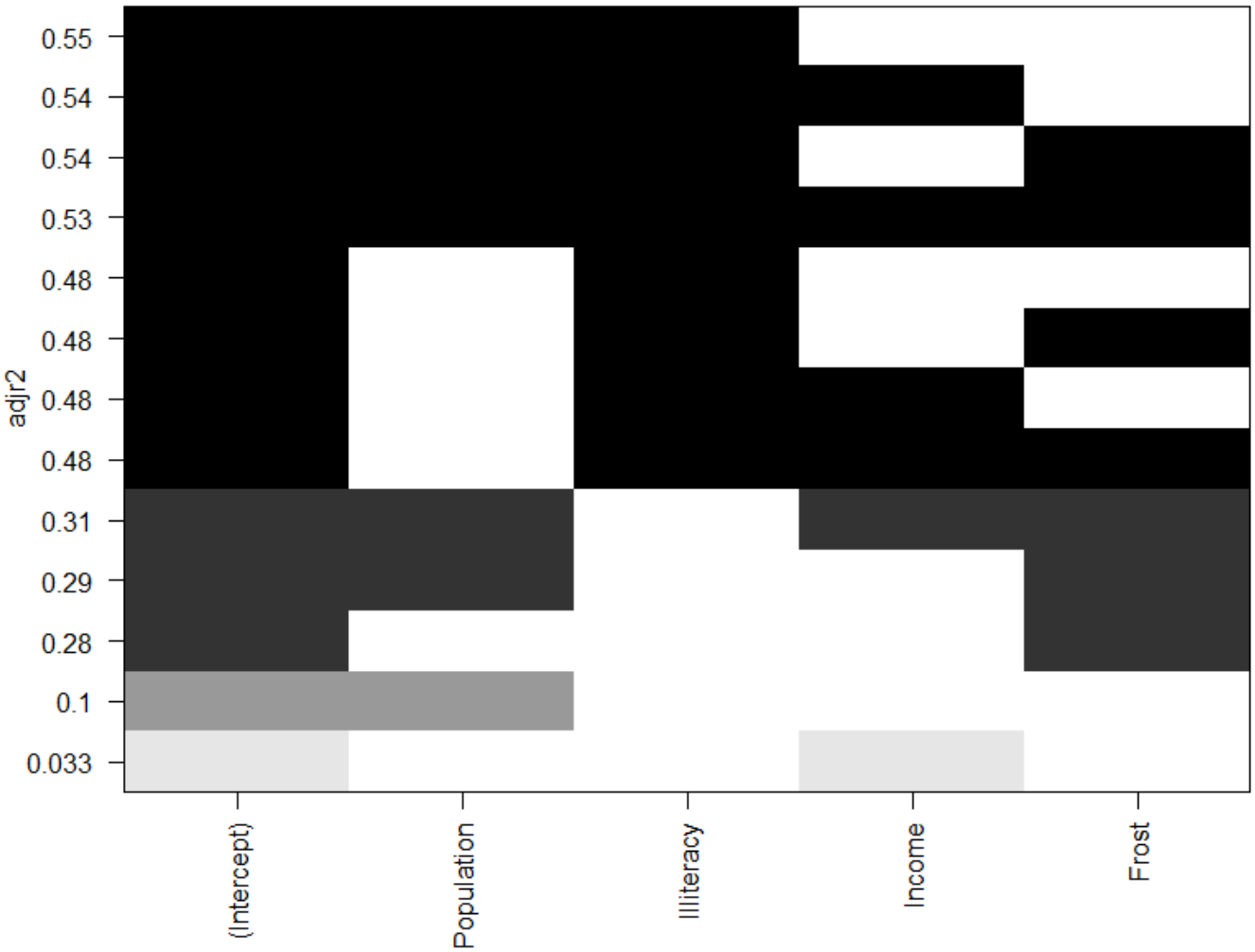
```
- Income       1     0.057 289.25  93.763
<none>                    289.19  95.753
- Population   1    43.658 332.85 100.783
- Illiteracy   1   236.196 525.38 123.605

Step:  AIC=93.76
Murder ~ Population + Illiteracy

             Df Sum of Sq    RSS     AIC
<none>                    289.25  93.763
- Population   1    48.517 337.76  99.516
- Illiteracy   1   299.646 588.89 127.311

Call:
lm(formula = Murder ~ Population + Illiteracy, data = states)

Coefficients:
(Intercept)   Population   Illiteracy
  1.6515497    0.0002242    4.0807366

>
```

All subsets regression

In all subsets regression, every possible model is inspected. All subsets regression is performed using the regsubsets() function from the leaps package. You can choose Adjusted R-squared, or Mallows Cp statistic as your criterion for reporting "best" models.

Example 5 (All subset selection)

```
library(leaps)
leaps <- regsubsets(Murder ~ Population + Illiteracy + Income + Frost,
data=states, nbest=4)
plot(leaps, scale="adjr2")
```


adjr2
0.55
0.54
0.54
0.53
0.48
0.48
0.48
0.48
0.31
0.29
0.28
0.1
0.033
(Intercept)
Population
Illiteracy
Income
Frost

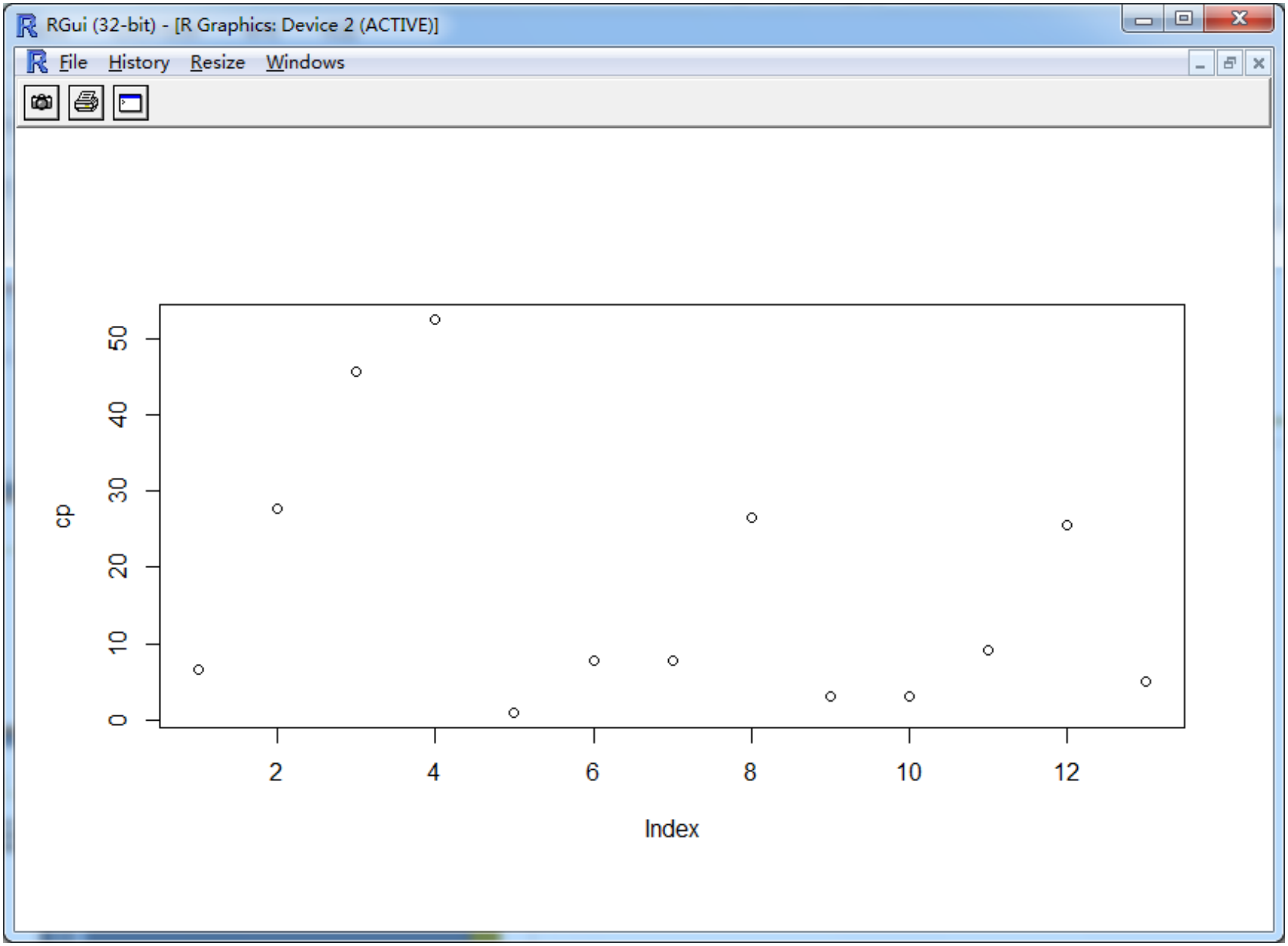
Example 6 (Mallows cp)

```
leaps <- regsubsets(Murder ~ Population + Illiteracy + Income +
    Frost, data=states, nbest=4)
cp=summary(leaps)$cp
index=which.min(cp)
summary(leaps)$which[index,]
summary(leaps)
```

RGui (32-bit) - [R Console]

File Edit View Misc Packages Windows Help

```
> leaps <- regsubsets(Murder ~ Population + Illiteracy + In
> cp=summary(leaps)$cp
> index=which.min(cp)
> summary(leaps)$which[index,]
(Intercept)  Population  Illiteracy      Income       Frost
       TRUE        TRUE        TRUE       FALSE       FALSE
> plot(cp)
>
```


RGui (32-bit) - [R Graphics: Device 2 (ACTIVE)]
File History Resize Windows
cp
0
10
20
30
40
50
2
4
6
8
10
12
Index

RGui (32-bit) - [R Console]

File Edit View Misc Packages Windows Help

```
Frost           FALSE      FALSE
4 subsets of each size up to 4
Selection Algorithm: exhaustive
         Population Illiteracy Income Frost
1  ( 1 ) " "        "*"        " "    " "
1  ( 2 ) " "        " "        " "    "*"
1  ( 3 ) "*"        " "        " "    " "
1  ( 4 ) " "        " "        "*"    " "
2  ( 1 ) "*"        "*"        " "    " "
2  ( 2 ) " "        "*"        " "    "*"
2  ( 3 ) " "        "*"        "*"    " "
2  ( 4 ) "*"        " "        " "    "*"
3  ( 1 ) "*"        "*"        "*"    " "
3  ( 2 ) "*"        "*"        " "    "*"
3  ( 3 ) " "        "*"        "*"    "*"
3  ( 4 ) "*"        " "        "*"    "*"
4  ( 1 ) "*"        "*"        "*"    "*"
>
```

All subset or stepwise method

In most instances, all subsets regression is preferable to stepwise regression, because more models are considered. However, when the number of predictors is large, it takes much more computing time.

Cross validation

When data description is your primary goal, the selection and interpretation of a regression model is the end of data analysis. But when your goal is prediction, you can justifiably ask, "How well will this model perform in the real world?"

The simplest way to use the data iteself to test the model.

k-fold cross validation

In k-fold cross-validation, we split the data into roughly equal-size parts. For the k-th part (test samples), fit the model to the other k-1 parts (training samples) and calculate the prediction error of the fitted model when applying it to predict the k-th part of the data.

The optimal model is the one that has the smallest average prediction error.

Summary

Each model selection approach has its own advantages and disadvantages. No approach is optimal in any case.

In general, automated variable selection methods should be seen as an aid rather than a directing force in model selection. A wellfitting model that doesn't make sense doesn't help you. Ultimately, it's your knowledge of the subject matter that should guide you.