Chapter 8: Selection of variables

Contents

1 Problem formulation

In practice, there are sometimes a **lot of possible explanatory variables** (e.g., m candidate explanatory variables: $X_1, X_2, ..., X_m$) to explain the response variable Y.

- How to reduce this number of explanatory variables?
- Examine whether all of the potential explanatory variables are needed or whether a subset of them is adequate.
- Search the optimal subset of explanatory variables.

Look for the optimal subset of p X-variables.
What could be a criterion for finding the optimal subset?

2 All possible regressions

2.1 Purpose

The **all-possible-regressions** procedure considers all possible subsets of the pool of potential explanatory variables X_i (with $i = 1, 2, ..., m$). It then identifies a small group of regression models which are "good" according to a specified criterion. A detailed examination of these models can lead to the selection of the final model.

All possible subsets?

If there are m candidate explanatory variables: → 2^m regressions for all possible subsets
(e.g. if $m = 10$, then there are 1024 possible regression models)
→ A lot of computation
→ Only possible for models with not too many candidates ($m < 25$)

Possible criteria?

- The smallest *Error SS* or highest R^2 (Remember: $R^2 = 1 - \frac{Error\ SS}{TotalSS}$)
- The highest *adjusted* R^2

2.2 Example

Example *Insurance*

Import the data set *insurance.txt* as `insurance`.

Variable	Description
$LOSS$	Amount of damage caused by the driver (in 1000$)
AGE	Age of the driver
$EXPER$	Number of years the driver has a driver license
$TOWORK$	Distance from home to work (expressed in miles)
$MILES$	Total number of miles in one year (in 1000 miles)

We want to explain the $LOSS$ of the driver based on following possible variables: AGE, $EXPER$, $TOWORK$, $MILES$. (thus $m = 4$)

Part of the data:

```
insurance[1:10, c(2, 4, 5, 8, 9)]
```

```
##     LOSS AGE EXPER TOWORK MILES
## 1  0.432  59    10     14     1
## 2  0.488  70    42     10     6
## 3  0.491  54    26      0     6
## 4  1.030  35    17     16     4
## 5  0.853  33    12      6     5
## 6  0.128  72    30      8     7
## 7  0.297  63    33      7     3
## 8  0.542  60    35      8     4
## 9  0.874  42     8     14     2
## 10 0.558  51    17      5    20
```

```
loss <- insurance$LOSS
age <- insurance$AGE
exper <- insurance$EXPER
miles <- insurance$MILES
towork <- insurance$TOWORK
```

Possible regression models with these 4 explanatory variables:

- Regression with only intercept: 1 regression
- All regressions with 1 variable: 4 regressions
- All regressions with 2 variables: 6 regressions
- All regressions with 3 variables: 4 regressions
- All regressions with 4 variables: 1 regression
  → In total: 16 regressions

In R

The function `leaps()` (from package `leaps`) performs an exhaustive search for the best subsets of the explanatory variables for predicting the response variable in linear regression.

```
install.packages("leaps")
library(leaps)
?leaps
```

Usage:
`leaps(x=, y=, method=c("Cp", "adjr2", "r2"), ...)`

- `x`: A matrix of predictors
- `y`: A response vector
- `method`: Calculate Cp, adjusted R-squared or R-squared

```
leap <- leaps(x = cbind(age, exper, miles, towork), y = loss, method = c("r2"))
combine <- cbind(leap$which,leap$size, leap$r2)
dimnames(combine) <- list(1:15,c("age","exper","miles","towork","size","r2"))
round(combine, digits=3)
```

```
##    age exper miles towork size    r2
## 1    1     0     0      0    2 0.873
## 2    0     1     0      0    2 0.623
## 3    0     0     0      1    2 0.036
## 4    0     0     1      0    2 0.000
## 5    1     0     0      1    3 0.909
## 6    1     1     0      0    3 0.873
## 7    1     0     1      0    3 0.873
## 8    0     1     0      1    3 0.658
## 9    0     1     1      0    3 0.623
## 10   0     0     1      1    3 0.036
## 11   1     0     1      1    4 0.909
## 12   1     1     0      1    4 0.909
## 13   1     1     1      0    4 0.873
## 14   0     1     1      1    4 0.658
## 15   1     1     1      1    5 0.909
```

Interpretation:

1. `size` is the number of parameters in the model. In this example, this corresponds to the number of explanatory variables +1.
2. The best model with 2 parameters is the first one: The regression model with the explanatory variable AGE. ($R^2 = 0.87$)
3. The best model with 3 parameters is the one with the variables AGE and $TOWORK$. ($R^2 = 0.90$)
4. The best model with 4 parameters is the one with the variables AGE, $TOWORK$, and $MILES$. ($R^2 = 0.90$)
5. There is only one model with 5 parameters.

2.3 Mallows' C_p

Mallows' C_p is a criterion which is concerned with the total mean squared error of the fitted values for each subset regression model. This statistic is invented by Mallows in 1973.

Mallows' C_p can be used to assure

- that all important explanatory variables are in the model
- that there aren't too many or not enough variables in the model.

Mallows' C_p is calculated as
$C_p = \frac{SSE(RM)}{MSE(FM)} - (n - 2(p+1))$
where

- FM stands for *Full Model.* It refers to the regression model with all m candidate variables.
- RM stands for *Reduced Model* which is the regression model with p regressors.
- $MSE(FM)$ is the *Mean Squared Error* of the *Full Model.* It is thus an estimate of the variance σ^2 of Y for the model with m variables.
  $MSE(FM) = \frac{SSE(FM)}{n-m-1}$
- $SSE(RM)$ (=*Error SS* for reduced model) is the *error sum of squares* of the model with p regressors.
- n is the total number of observations
- p is the number of explanatory variables ($p+1$ is the number of parameters)

2.3.1 Theory of Mallows' C_p

> When there is no bias in the regression model with p X variables, the expected value of C_p is approximately $p+1$ (number of parameters).

Hence we make a plot of C_p values against p+1 and detect:

- Models with little bias will tend to fall near the line $C_p = p+1$.
- Models with substantial bias will tend to fall considerably above this line.
- Models with C_p values below this line are interpreted as showing no bias, being below this line due to sampling error.

In using the C_p criterion, we seek to identify subsets of X variables for which

1. the C_p value is small and
2. the C_p value is near p+1

- If $C_p > p+1$, the model is under specified. The error is because there are too less variables in the model.
- If $C_p < p+1$, the model is over specified. There are too many variables in the model.

> There where C_p crosses the line $(p+1)$, gives you an indication about the number of variables in the model.

Example *Insurance*

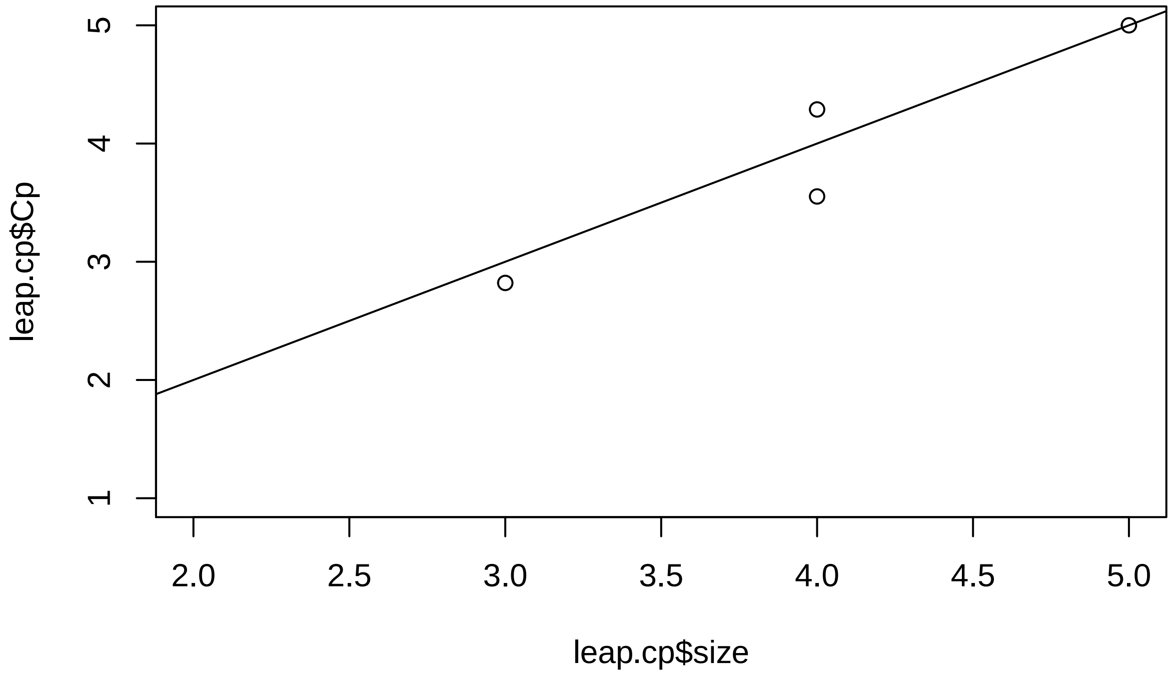
```
leap.cp <- leaps(x = cbind(age, exper, miles, towork), y=loss, method="Cp")
# Note that method is set to "Cp" in this leaps command
combine.cp <- cbind(leap.cp$which,leap.cp$size, leap.cp$Cp)
dimnames(combine.cp) <- list(1:15,c("age","exper","miles","towork","size","cp"))
round(combine.cp, digits=3)
```

```
##    age exper miles towork size        cp
## 1    1     0     0      0    2  1944.741
## 2    0     1     0      0    2 15623.407
## 3    0     0     0      1    2 47658.806
## 4    0     0     1      0    2 49639.573
## 5    1     0     0      1    3     2.821
## 6    1     1     0      0    3  1945.972
## 7    1     0     1      0    3  1946.434
## 8    0     1     0      1    3 13698.865
## 9    0     1     1      0    3 15620.675
## 10   0     0     1      1    3 47653.857
## 11   1     0     1      1    4     3.552
## 12   1     1     0      1    4     4.288
## 13   1     1     1      0    4  1947.654
## 14   0     1     1      1    4 13693.331
## 15   1     1     1      1    5     5.000
```

Interpretation:

How to obtain the optimal number of variables in the model?
We make a plot of C_p versus size $(= p + 1)$.

```
plot(leap.cp$size, leap.cp$Cp, ylim=c(1,5))
abline(a=0, b=1)
```



A model with 3 parameters (and hence 2 variables) is the best model. This is the model with variables *AGE* and *TOWORK*.

Remark:
Use the option `nbest=3` in the `leaps` function to report only the 3 best subsets of each size.

3 Model selection criteria

When the set of candidate variables is too large, we have to make use of automatic stepwise selection techniques (see further). What is the **criterion used to determine whether a certain model is better or worse than another model?** The *partial F-test* or the *Akaike Information criterion* can be used.

3.1 Partial F-test

Consider two **nested models** for Y with $p > k$:
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + ... + \beta_p x_p$ (with $p + 1$ parameters)

and
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$ (with $p + 1$

We want to **test**
$H_0 : \beta_{k+1} = ... = \beta_p = 0$
versus
H_1 : **not all** $\beta_i = 0$ **for** $i = k + 1, k + 2, ..., p$.

Construction of a test statistic:

Since $p > k \Rightarrow SSE(p) < SSE(k)$

Extra sum of squares $= SSE(k) - SSE(p) \quad (> 0)$

If the new parameters are not really important (if H_0 is true), then there should be little difference between the sums of squares when computed with or without the new parameters. If they are important (if H_0 is not true), then there should be a big difference. To measure big or small, we divide the extra sum of squares by the residual sum of squares for the largest model. This ratio, $\frac{SSE(k)-SSE(p)}{SSE(p)}$ should measure the influence of the extra parameters.
If we divide the numerator and the denominator by their respective degrees of freedom, we obtain the **partial F-test statistic**:

$$F = \frac{\frac{Extra\ sum\ of\ squares}{p-k}}{\frac{SSE(p)}{n-p-1}} = \frac{\frac{SSE(k)-SSE(p)}{p-k}}{\frac{SSE(p)}{n-p-1}}$$

which has an F-distribution with $p - k$ and $n - p - 1$ degrees of freedom.

This partial F can be used to test hypothesis (for nested models only!):
$H_0 : \beta_{k+1} = ... = \beta_p = 0$
versus
H_1 : **not all** $\beta_i = 0$ **for** $i = k + 1, k + 2, ..., p$.

Based on the corresponding $p - value$, we come to a conclusion .

In R

Consider the following two nested models:

1. A model with 2 explanatory variables: AGE and $TOWORK$
2. A model with 4 explanatory variables: $AGE, EXPER, MILES$, and $TOWORK$

How to compare these nested models?

```r
reg.lm1 <- lm(loss ~ age + towork)
reg.lm2 <- lm(loss ~ age + towork + miles + exper)
anova(reg.lm1, reg.lm2)
```

```
## Analysis of Variance Table
##
## Model 1: loss ~ age + towork
## Model 2: loss ~ age + towork + miles + exper
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   4997 84.045
## 2   4995 84.015  2  0.030626 0.9104 0.4024
```

Interpretation
Model 1: $LOSS = \beta_0 + \beta_1 \cdot AGE + \beta_2 \cdot TOWORK$
Model 2: $LOSS = \beta_0 + \beta_1 \cdot AGE + \beta_2 \cdot TOWORK + \beta_3 \cdot MILES + \beta_4 \cdot EXPER$

$H_0 : \beta_3 = \beta_4 = 0$
versus
H_1 : **not both** β_3 **and** β_4 **are** 0 .

The $p-value(=0.402)$ is large (>0.05), hence the H_0 is not rejected. We can use *model 1* with only 2 explanatory variables AGE and $TOWORK$.

3.2 Akaike information criterion

We have to realize that there are no 'true' models. The proposed models only approximate reality!

The more parameters that there are in the model, the better the fit. You can obtain a perfect fit when you have a separate parameter for every data point, but this model will not have any explanatory power.

There is always a trade-off between the goodness of fit of the model and the number of parameters required. **AIC** (**Akaike's Information Criterion**, 1973) is useful because it explicitly penalizes any superfluous parameters in the model. In a regression context, $AIC(k)$ for a model with $(k+1)$ parameters (hence k explanatory variables) and sample size n can be written as:

$AIC(k) = n \cdot \ln(SSE(k)) - n \cdot \ln(n) + 2 \cdot (k+1)$

Remark:

1. When k increases, then $SSE(k)$ decreases
2. $n \cdot \ln(n)$ is independent of the choice of k
3. $2 \cdot (k+1)$ increases when k increases
4. **When comparing two models, the smaller the AIC, the better the fit.**
5. An advantage of the AIC is that it can be used to compare models that are not nested. In contrast, nested models was a requirement for the partial F-test.

In R
The function `AIC()` will compute the AIC value for a given model.

```
reg.lm1 <- lm(loss ~ age + towork)
reg.lm2 <- lm(loss ~ age + towork + miles + exper)

list(AIC_model1 = AIC(reg.lm1), AIC_model2 = AIC(reg.lm2))
```

```
## $AIC_model1
## [1] -6231.8
##
## $AIC_model2
## [1] -6229.622
```

Interpretation:
The AIC value for the model with the variables AGE and $TOWORK$ is smaller. This model is to be preferred.

4 Selection methods

4.1 Introduction

In those cases when the pool of potential explanatory variables X contains 40 to 60 (or even more) variables, an automatic procedure that develops the "best" subset of explanatory variables X sequentially may be helpful. Therefore we have the **selection methods**. These methods develop a sequence of regression models. At each step, they add or delete an certain explanatory variable X. The criterion used in R for comparing the sequential models is the AIC criterion.

Weakness of the procedure:
The search methods end with a 'single' regression model, while the all-possible-regression procedure can identify several regression models as good for final consideration.

4.2 Stepwise selection methods

There are 3 **selection methods**:

- **Forward selection**
  This method start with a model with no variables.
  → *Step 1:* Add a variable to the model. This variable is selected according a certain selection criterion (e.g. AIC of the model that includes the considered variable). We obtain a regression model with one variable.
  → *Step 2:* Add another variable to the model. This variable is selected according a certain selection criterion, given the fact that the first variable is already in the model. We then obtain a model with two variables.
  → *Step 3:* . . .
  Variables are added one by one to the model until no variable satisfy or improves the selection criterion (e.g. when the AIC value of the models obtained by adding one of the resting variables is higher than the AIC value for the last created model).

- **Backward selection**
  This method start with a model with all possible variables in it. Try to delete one variable (step by step) in terms of a certain criterion.

- **Bidirectional selection**
  This method start with a model with no variables.
  Try to add variables one by one (as in the Forward procedure). Once there are variables in the model, try to delete variables (as in the Backward procedure). Continue by adding variables, deleting, . . .

The criterion which is used in R is the AIC criterion.

4.3 In R

Forward selection in R

```
slm.forward <- step(lm(LOSS ~ 1, data = insurance), scope = ~ AGE + EXPER + MILES + TOWORK,
                    direction = "forward", data = insurance)
```

```
## Start:  AIC=-8467.38
## LOSS ~ 1
## 
##          Df Sum of Sq    RSS      AIC
## + AGE     1    802.29 116.74 -18782.2
## + EXPER   1    572.21 346.81 -13338.0
## + TOWORK  1     33.39 885.64  -8650.4
## <none>                919.03  -8467.4
## + MILES   1      0.07 918.96  -8465.8
## 
## Step:  AIC=-18782.15
## LOSS ~ AGE
## 
##          Df Sum of Sq     RSS    AIC
## + TOWORK  1    32.696  84.045 -20423
## <none>                116.742 -18782
## + EXPER   1     0.013 116.729 -18781
## + MILES   1     0.005 116.736 -18780
## 
## Step:  AIC=-20423.19
## LOSS ~ AGE + TOWORK
## 
```

```
##         Df Sum of Sq    RSS    AIC
## <none>               84.045 -20423
## + MILES  1 0.0213362 84.024 -20423
## + EXPER  1 0.0089564 84.036 -20422
```

Backward selection in R

```r
slm.backward <- step(lm(LOSS ~ AGE + EXPER + MILES + TOWORK, data = insurance),
                     direction = "backward")
```

```
## Start:  AIC=-20421.01
## LOSS ~ AGE + EXPER + MILES + TOWORK
## 
##          Df Sum of Sq     RSS    AIC
## - EXPER   1     0.009  84.024 -20423
## - MILES   1     0.022  84.036 -20422
## <none>                 84.015 -20421
## - TOWORK  1    32.709 116.723 -18779
## - AGE     1   230.268 314.283 -13826
## 
## Step:  AIC=-20422.45
## LOSS ~ AGE + MILES + TOWORK
## 
##          Df Sum of Sq    RSS      AIC
## - MILES   1      0.02  84.05 -20423.2
## <none>                 84.02 -20422.5
## - TOWORK  1     32.71 116.74 -18780.4
## - AGE     1    801.50 885.52  -8649.1
## 
## Step:  AIC=-20423.19
## LOSS ~ AGE + TOWORK
## 
##          Df Sum of Sq    RSS      AIC
## <none>                 84.05 -20423.2
## - TOWORK  1      32.7 116.74 -18782.2
## - AGE     1     801.6 885.64  -8650.4
```

Bidirectional selection in R

```r
slm.both <- step(lm(LOSS ~ AGE + EXPER + MILES + TOWORK, data = insurance), direction = "both")
```

```
## Start:  AIC=-20421.01
## LOSS ~ AGE + EXPER + MILES + TOWORK
## 
##          Df Sum of Sq     RSS    AIC
## - EXPER   1     0.009  84.024 -20423
## - MILES   1     0.022  84.036 -20422
## <none>                 84.015 -20421
## - TOWORK  1    32.709 116.723 -18779
## - AGE     1   230.268 314.283 -13826
## 
## Step:  AIC=-20422.45
## LOSS ~ AGE + MILES + TOWORK
## 
##          Df Sum of Sq    RSS      AIC
## - MILES   1      0.02  84.05 -20423.2
## <none>                 84.02 -20422.5
## + EXPER   1      0.01  84.01 -20421.0
## - TOWORK  1     32.71 116.74 -18780.4
## - AGE     1    801.50 885.52  -8649.1
##
## Step:  AIC=-20423.19
## LOSS ~ AGE + TOWORK
##
##          Df Sum of Sq    RSS      AIC
## <none>                 84.05 -20423.2
## + MILES   1      0.02  84.02 -20422.5
## + EXPER   1      0.01  84.04 -20421.7
## - TOWORK  1     32.70 116.74 -18782.2
## - AGE     1    801.60 885.64  -8650.4
```

5 Exercise

In this exercise, the data set *cars1.txt* is used.

93 cars are investigated in order to compare their prices. There is also information on the specifications of the car. In this way, a customer can decide which car to take.

```
##   Manufacturer    Type MinimumPrice MidrangePrice MaximumPrice CityMPG
## 1      Mercury  Sporty         13.3          14.1         15.0      23
## 2        Honda  Sporty         17.0          19.8         22.7      24
## 3     Plymouth  Sporty         11.4          14.4         17.4      23
## 4      Hyundai  Sporty          9.1          10.0         11.0      26
## 5        Lexus Midsize         34.7          35.2         35.6      18
## 6          Geo   Small          6.7           8.4         10.0      46
##   HighwayMPG AirBags Drivetrain cylinders Enginesize Horsepower  RPM
## 1         26       1          1         4        1.6        100 5750
## 2         31       2          1         4        2.3        160 5800
## 3         30       0          2         4        1.8         92 5000
## 4         34       0          1         4        1.5         92 5550
## 5         23       2          0         6        3.0        225 6000
## 6         50       0          1         3        1.0         55 5700
##   revolutions transmission FuelTank Passengers Length Wheelbase Width Uturn
## 1        2475            1     11.1          4    166        95    65    36
## 2        2855            1     15.9          4    175       100    70    39
## 3        2360            1     15.9          4    173        97    67    39
## 4        2540            1     11.9          4    166        94    64    34
## 5        2510            1     20.6          4    191       106    71    39
## 6        3755            1     10.6          4    151        93    63    34
##   RearSeat Luggage Weight Domestic
## 1     19.0       6   2450        1
## 2     23.5       8   2865        0
## 3     24.5       8   2640        1
## 4     23.5       9   2285        0
## 5     25.0       9   3515        0
## 6     27.5      10   1695        0
```

Variable	Description
`Manufacturer`	Name of manufacturer
`Type`	Small, sporty, compact, midsize, large

Variable	Description
MinimumPrice	Price basic version (in \$1000 \$ \$)
MidrangePrice	Price in between minimum and maximum price
MaximumPrice	Maximum price (in \$1000 \$ \$)
CityMPG	'Miles per Gallon' in city
HighwayMPG	'Miles per Gallon' on highway
AirBags	0: No airbags 1: Driver only 2: Driver and passenger
Drivetrain	0: rear wheel drive 1: front wheel drive 2: all wheel drive
cylinders	Number of cylinders
Enginesize	Engine size (in liters)
Horsepower	Horsepower (max)
RPM	Revolutions per minute at maximum horsepower
revolutions	Engine revolutions per mile
transmission	Manual transmission available? 0: no 1: yes
FuelTank	Fuel tank capacity (in gallons)
Passengers	Passenger capacity (number of passengers)
Length	Length in inches
Wheelbase	Wheel base in inches
Width	Width in inches
Uturn	U-turn space expressed in feet
RearSeat	Rear seat room expressed in inches
Luggage	Luggage capacity
Weight	Weight in pond
Domestic	0: Non-US 1: US

Formulate a good regression model forMidrangePrice. You may use following variables: Horsepower, Length, Luggage, Uturn, Wheelbase, and Width.

a. Use all-possible-subsets selection
b. Use one of the automatic-selection techniques

6 Solution

Import *cars1.txt* as cars

```
#import dataset cars1.txt
cars <- read.table(file=file.choose(), header=TRUE)
names(cars)
```

6.1 Question (a): Use of all-possible-subset selection

Method in leaps function: R^2

```
attach(cars)
leap <- leaps(x = cbind(Horsepower, Length, Luggage, Uturn, Wheelbase, Width),
              y=MidrangePrice, method=c("r2"), nbest=3)

combine <- cbind(leap$which, leap$size, leap$r2)
n <- length(leap$size)
dimnames(combine) <- list(1:n, c("horsep", "length", "Lug", "Uturn", "WB", "Width",
                                "size", "r2"))
round(combine, digits=3)
```

```
##    horsep length Lug Uturn WB Width size    r2
## 1       1      0   0     0  0     0    2 0.620
## 2       0      0   0     0  1     0    2 0.395
## 3       0      1   0     0  0     0    2 0.305
## 4       1      0   0     0  1     0    3 0.640
## 5       1      0   1     0  0     0    3 0.628
## 6       1      0   0     0  0     1    3 0.624
## 7       1      0   0     0  1     1    4 0.699
## 8       1      0   0     1  1     0    4 0.670
## 9       1      1   0     0  0     1    4 0.658
## 10      1      0   0     1  1     1    5 0.703
## 11      1      0   1     0  1     1    5 0.701
## 12      1      1   0     0  1     1    5 0.699
## 13      1      0   1     1  1     1    6 0.705
## 14      1      1   0     1  1     1    6 0.703
## 15      1      1   1     0  1     1    6 0.701
## 16      1      1   1     1  1     1    7 0.705
```

```
detach(cars)
```

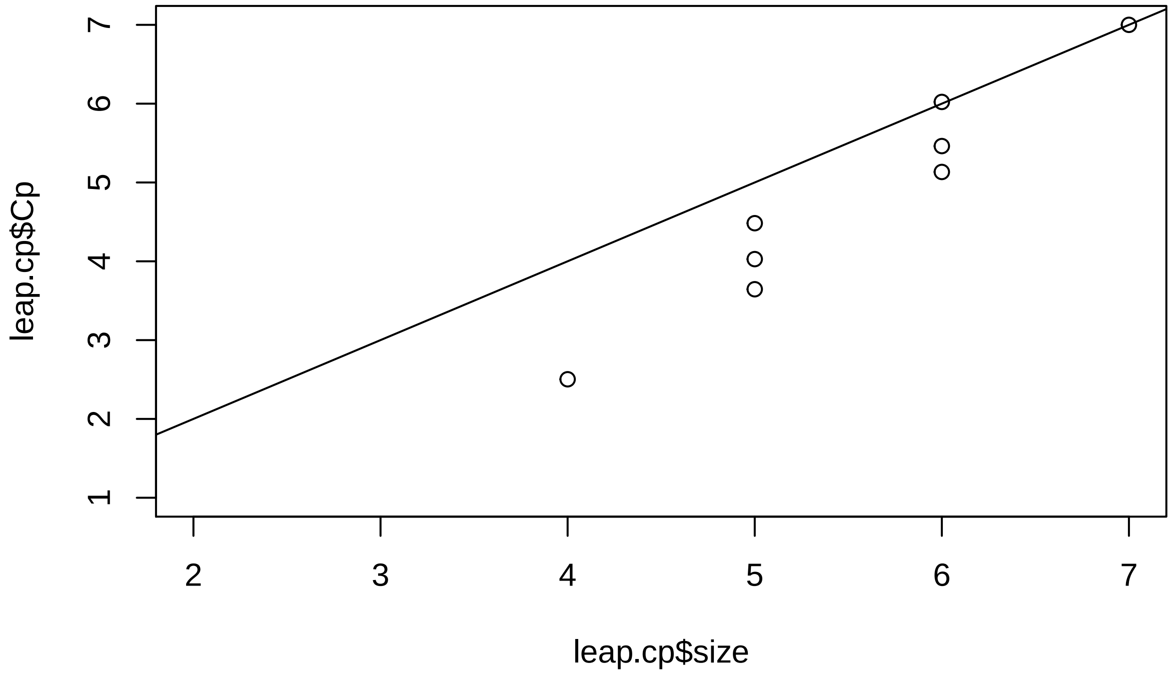
Method in `leaps` function: Mallows' C_p

```
attach(cars)
leap.cp <- leaps(x = cbind(Horsepower, Length, Luggage, Uturn, Wheelbase, Width),
                  y = MidrangePrice, nbest = 3)

combine.cp <- cbind(leap.cp$which, leap.cp$size, leap.cp$Cp)
n <- length(leap.cp$size)
dimnames(combine.cp) <- list(1:n, c("horsep", "length", "Lug", "Uturn", "WB", "Width",
                                    "size", "Cp"))
round(combine.cp, digits=3)
```

```
##    horsep length Lug Uturn WB Width size     Cp
## 1       1      0   0     0  0     0    2 18.747
## 2       0      0   0     0  1     0    2 75.975
## 3       0      1   0     0  0     0    2 98.886
## 4       1      0   0     0  1     0    3 15.605
## 5       1      0   1     0  0     0    3 18.696
## 6       1      0   0     0  0     1    3 19.677
## 7       1      0   0     0  1     1    4  2.503
## 8       1      0   0     1  1     0    4 10.032
## 9       1      1   0     0  0     1    4 13.106
## 10      1      0   0     1  1     1    5  3.646
## 11      1      0   1     0  1     1    5  4.028
## 12      1      1   0     0  1     1    5  4.483
## 13      1      0   1     1  1     1    6  5.133
## 14      1      1   0     1  1     1    6  5.462
## 15      1      1   1     0  1     1    6  6.022
## 16      1      1   1     1  1     1    7  7.000
```

```
plot(leap.cp$size, leap.cp$Cp, ylim=c(1,7))
abline(a=0, b=1)
```



```
detach(cars)
```

6.2 Question (b): Use one of the automatic-selection techniques

Use of Forward selection

```
slm.foward <- step(lm(MidrangePrice ~ 1, data=cars), scope = ~ Horsepower + Length +
                     Luggage + Uturn + Wheelbase + Width, direction = "forward", data = cars)
```

```
## Start:  AIC=377.95
## MidrangePrice ~ 1
##
##              Df Sum of Sq    RSS    AIC
## + Horsepower  1    4979.3 3054.9 300.66
## + Wheelbase   1    3172.3 4862.0 338.76
## + Length      1    2448.8 5585.4 350.14
## + Width       1    1969.2 6065.0 356.89
## + Uturn       1    1450.2 6584.0 363.63
## + Luggage     1    1079.6 6954.7 368.12
## <none>                    8034.2 377.95
##
## Step:  AIC=300.66
## MidrangePrice ~ Horsepower
##
##             Df Sum of Sq    RSS    AIC
## + Wheelbase  1   162.358 2892.5 298.18
## <none>                   3054.9 300.66
## + Luggage    1    64.739 2990.2 300.90
## + Width      1    33.758 3021.1 301.75
## + Length     1    28.638 3026.3 301.89
## + Uturn      1    26.582 3028.3 301.94
##
## Step:  AIC=298.18
## MidrangePrice ~ Horsepower + Wheelbase
##
##           Df Sum of Sq    RSS    AIC
## + Width    1    476.86 2415.7 285.41
## + Uturn    1    239.12 2653.4 293.11
## + Length   1    103.05 2789.5 297.20
## <none>                 2892.5 298.18
## + Luggage  1      2.01 2890.5 300.12
##
## Step:  AIC=285.41
## MidrangePrice ~ Horsepower + Wheelbase + Width
##
##           Df Sum of Sq    RSS    AIC
## <none>                 2415.7 285.41
## + Uturn    1   27.0742 2388.6 286.48
## + Luggage  1   15.0100 2400.7 286.90
## + Length   1    0.6207 2415.1 287.39
```

Use of backward selection

```
reg.lm1 <- lm(MidrangePrice ~ Horsepower + Length + Luggage + Uturn + Wheelbase + Width,
              data = cars)
slm.backward <- step(reg.lm1, direction = "backward")
```

```
## Start:  AIC=289.78
## MidrangePrice ~ Horsepower + Length + Luggage + Uturn + Wheelbase +
##     Width
##
##              Df Sum of Sq    RSS    AIC
## - Length      1      4.21 2372.4 287.93
## - Luggage     1     14.59 2382.8 288.28
## - Uturn       1     32.27 2400.5 288.89
## <none>                    2368.2 289.78
## - Wheelbase   1    241.98 2610.2 295.76
## - Width       1    272.60 2640.8 296.71
## - Horsepower  1   2305.15 4673.4 343.52
##
## Step:  AIC=287.93
## MidrangePrice ~ Horsepower + Luggage + Uturn + Wheelbase + Width
##
##              Df Sum of Sq    RSS    AIC
## - Luggage     1     16.17 2388.6 286.48
## - Uturn       1     28.23 2400.7 286.90
## <none>                    2372.4 287.93
## - Width       1    280.35 2652.8 295.09
## - Wheelbase   1    413.05 2785.5 299.09
## - Horsepower  1   2301.51 4673.9 341.53
##
## Step:  AIC=286.48
## MidrangePrice ~ Horsepower + Uturn + Wheelbase + Width
##
##              Df Sum of Sq    RSS    AIC
## - Uturn       1     27.07 2415.7 285.41
## <none>                    2388.6 286.48
## - Width       1    264.81 2653.4 293.10
## - Wheelbase   1    630.27 3018.9 303.69
## - Horsepower  1   2421.50 4810.1 341.88
##
## Step:  AIC=285.41
## MidrangePrice ~ Horsepower + Wheelbase + Width
##
##              Df Sum of Sq    RSS    AIC
## <none>                    2415.7 285.41
## - Width       1    476.86 2892.5 298.18
## - Wheelbase   1    605.46 3021.1 301.75
## - Horsepower  1   2402.21 4817.9 340.02
```

Use of bidirectional selection

```
reg.lm1 <- lm(MidrangePrice ~ Horsepower + Length + Luggage + Uturn + Wheelbase + Width,
              data = cars)
slm.both <- step(reg.lm1, direction="both")
```

```
## Start:  AIC=289.78
## MidrangePrice ~ Horsepower + Length + Luggage + Uturn + Wheelbase +
##     Width
##
##              Df Sum of Sq    RSS    AIC
## - Length      1      4.21 2372.4 287.93
## - Luggage     1     14.59 2382.8 288.28
## - Uturn       1     32.27 2400.5 288.89
## <none>                    2368.2 289.78
## - Wheelbase   1    241.98 2610.2 295.76
## - Width       1    272.60 2640.8 296.71
## - Horsepower  1   2305.15 4673.4 343.52
##
## Step:  AIC=287.93
## MidrangePrice ~ Horsepower + Luggage + Uturn + Wheelbase + Width
##
##              Df Sum of Sq    RSS    AIC
## - Luggage     1     16.17 2388.6 286.48
## - Uturn       1     28.23 2400.7 286.90
## <none>                    2372.4 287.93
## + Length      1      4.21 2368.2 289.78
## - Width       1    280.35 2652.8 295.09
## - Wheelbase   1    413.05 2785.5 299.09
## - Horsepower  1   2301.51 4673.9 341.53
##
## Step:  AIC=286.48
## MidrangePrice ~ Horsepower + Uturn + Wheelbase + Width
##
##              Df Sum of Sq    RSS    AIC
## - Uturn       1     27.07 2415.7 285.41
## <none>                    2388.6 286.48
## + Luggage     1     16.17 2372.4 287.93
## + Length      1      5.79 2382.8 288.28
## - Width       1    264.81 2653.4 293.10
## - Wheelbase   1    630.27 3018.9 303.69
## - Horsepower  1   2421.50 4810.1 341.88
##
## Step:  AIC=285.41
## MidrangePrice ~ Horsepower + Wheelbase + Width
##
##              Df Sum of Sq    RSS    AIC
## <none>                    2415.7 285.41
## + Uturn       1     27.07 2388.6 286.48
## + Luggage     1     15.01 2400.7 286.90
## + Length      1      0.62 2415.1 287.39
## - Width       1    476.86 2892.5 298.18
## - Wheelbase   1    605.46 3021.1 301.75
## - Horsepower  1   2402.21 4817.9 340.02
```