Homework 4

Zachary Foster

10/28/2014

Perpare R envrioment

```
library(leaps)
library(knitr)
library(reshape2)
library(ggplot2)
library(grid)
library(plyr)
opts_chunk$set(fig.width = 8, message = FALSE, warning = FALSE)
```

Data preparation

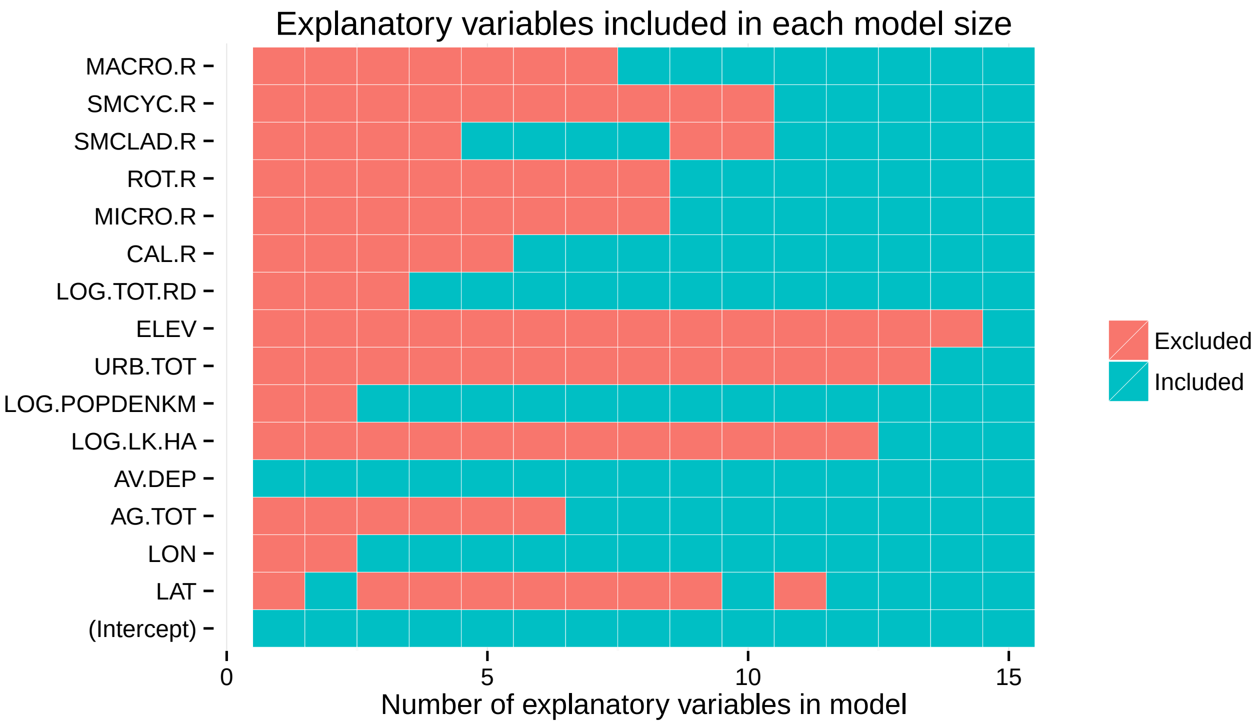
```
data <- read.table("emap.build08.txt", header = TRUE)
vars_to_log <- c("LK.HA", "POPDENKM", "TOT.RD")
data[vars_to_log] <- lapply(data[vars_to_log], function(x) log(x + 1))
names(data) <- ifelse(names(data) %in% vars_to_log,
                      paste("LOG", names(data), sep = "."),
                      names(data))
explanatory <- data[, -3]
response <-  data[, 3]
```

Naive all subsets selction without interactions

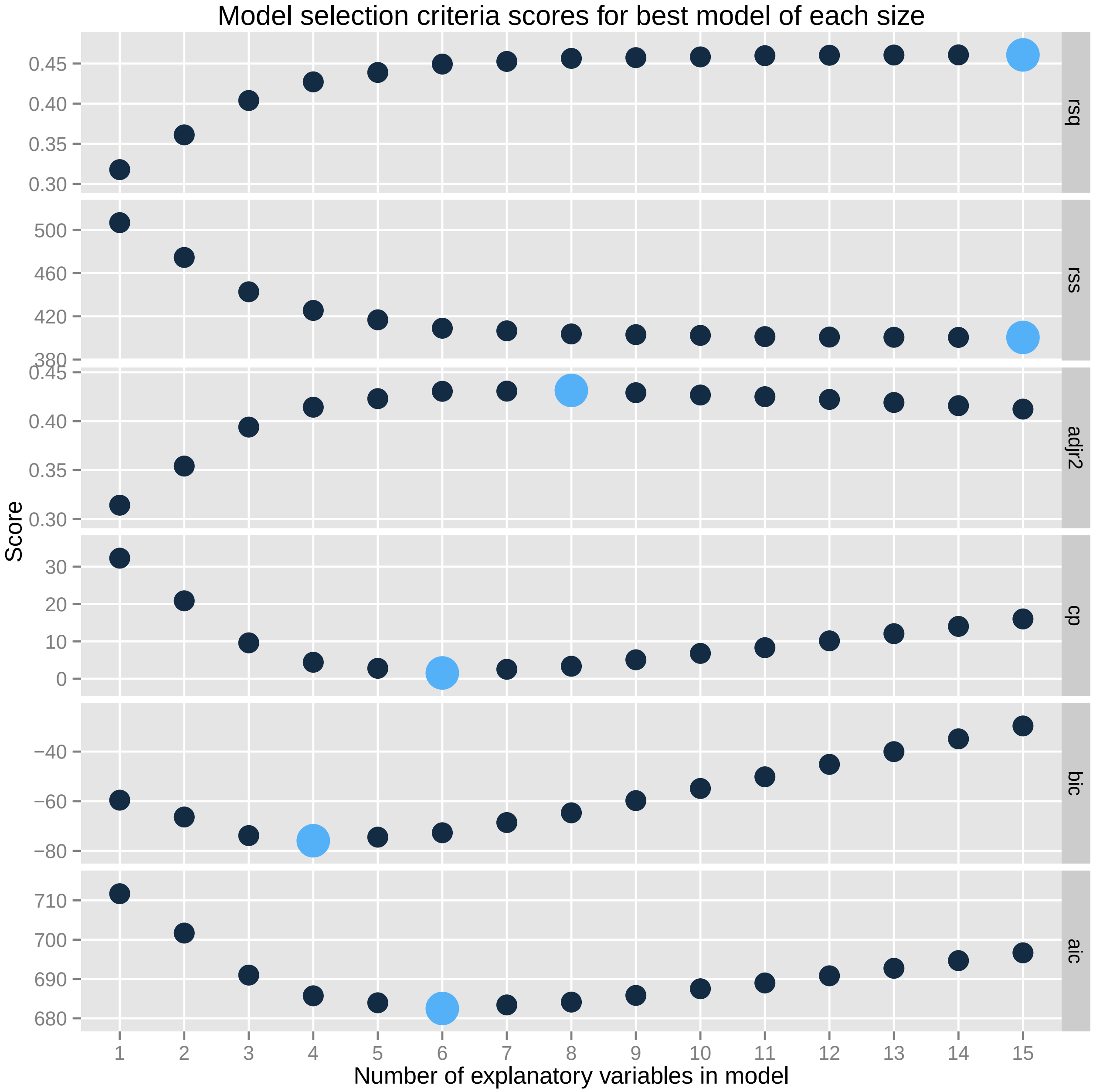
The `regsubsets` function from the `leaps` package provides a simple way of doing all subsets selection.

```
# Conduct all subsets model evaluation ------------------------------------------
naive_raw <- regsubsets(explanatory, response, nvmax = length(explanatory))
naive_results <- summary(naive_raw)

# Graph all subsets results -----------------------------------------------------
models <- melt(naive_results$which,
               varnames = c("size", "variables"),
               value.name = "included")
models$included = ifelse(models$included, "Included", "Excluded")
ggplot(models, aes(y = variables, x = size)) +
  geom_tile(aes(fill = included), color = "white") +
  labs(y = "",
       x = "Number of explanatory variables in model",
       fill = "",
       title = "Explanatory variables included in each model size") +
  theme_minimal() + theme(panel.grid.major.y = element_blank())
```



```
# Graph information cirteria scores -------------------------------------------
scores <- data.frame(naive_results[c("rsq", "rss", "adjr2", "cp", "bic")])
scores$size <- 1:nrow(scores)
calculate_aic <- function(included) {
  vars <- explanatory[included]
  model <- as.formula(paste("response ~", paste("explanatory", names(vars),
                                                collapse = " + ", sep = "$")))
  AIC(lm(model))
}
scores$aic <- apply(naive_results$which[, -1], 1, calculate_aic)
scores <- melt(scores, variable.name = "method", id.vars = "size")
is_best <- function(x) {
  if (x$method %in% c("rsq", "adjr2"))
    x$value == max(x$value)
  else x$value == min(x$value)
}
scores$best <- as.numeric(unlist(dlply(scores, "method", is_best)))
scores$size <- ordered(scores$size)
ggplot(scores, aes(x = size, y = value, color = best, size = best)) +
  geom_point(stat = "identity") +
  facet_grid(method ~ ., scales = "free_y") +
  scale_size(range = c(5, 8)) +
  scale_y_continuous(expand = c(.2,0)) +
  labs(y = "Score",
       x = "Number of explanatory variables in model",
       title = "Model selection criteria scores for best model of each size") +
  theme(legend.position = "none",
        panel.grid.minor = element_blank())
```


Model selection criteria scores for best model of each size
Score
Number of explanatory variables in model
rsq
rss
adjr2
cp
bic
aic
0.45
0.40
0.35
0.30
500
460
420
380
0.45
0.40
0.35
0.30
30
20
10
0
−40
−60
−80
710
700
690
680
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15