Homework 4

Zachary Foster, Elizabeth Orning, Sarah Nalven, Spencer Ledoux

10/28/2014

Perpare R envriomentt

```
library(leaps)
library(knitr)
library(reshape2)
library(ggplot2)
library(grid)
library(plyr)
library(glmulti)
```

```
## Loading required package: rJava
```

```
opts_chunk$set(fig.width = 8, message = FALSE, warning = FALSE)
```

Data preparation

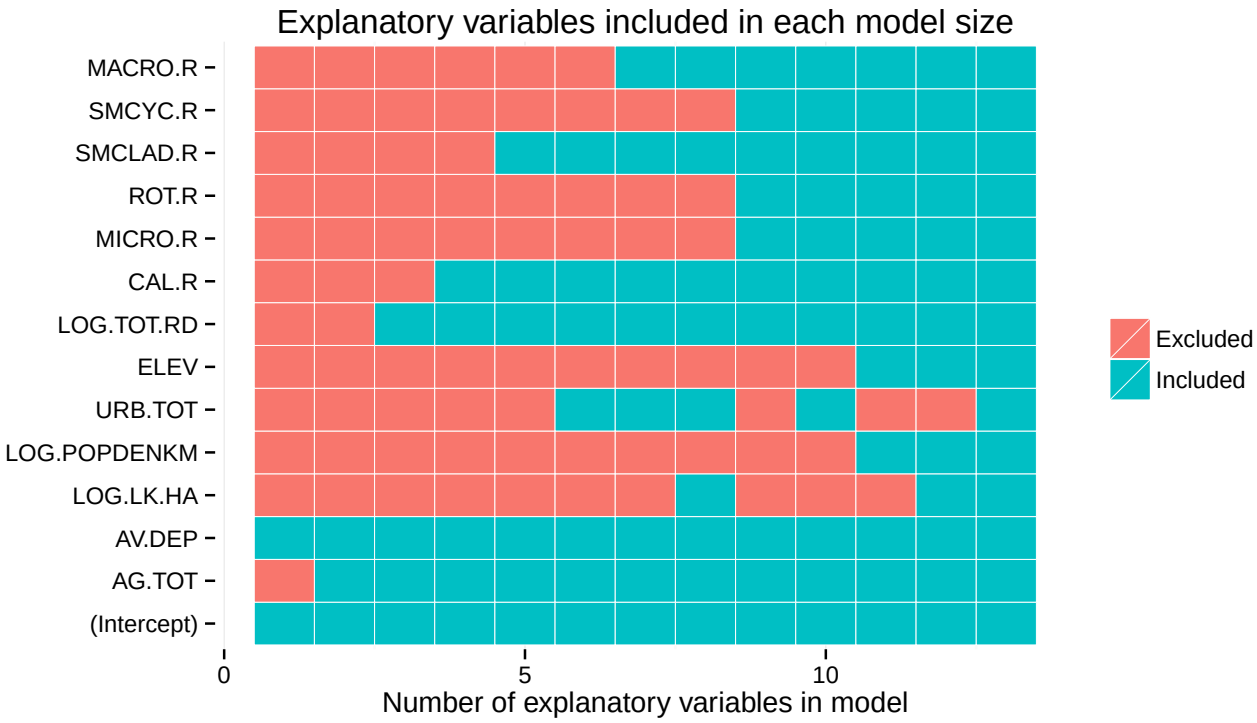
```
# Read input data ------------------------------------------------------------
data <- read.table("emap.build08.txt", header = TRUE)
data <- data[ , !(names(data) %in% c("LAT", "LON"))]
# Log varaibles --------------------------------------------------------------
vars_to_log <- c("LK.HA", "POPDENKM", "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[ , names(data) != "SECMEAN"]
response <-  data[ , "SECMEAN", drop = FALSE]
```

Naive all subsets selction without interactions

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

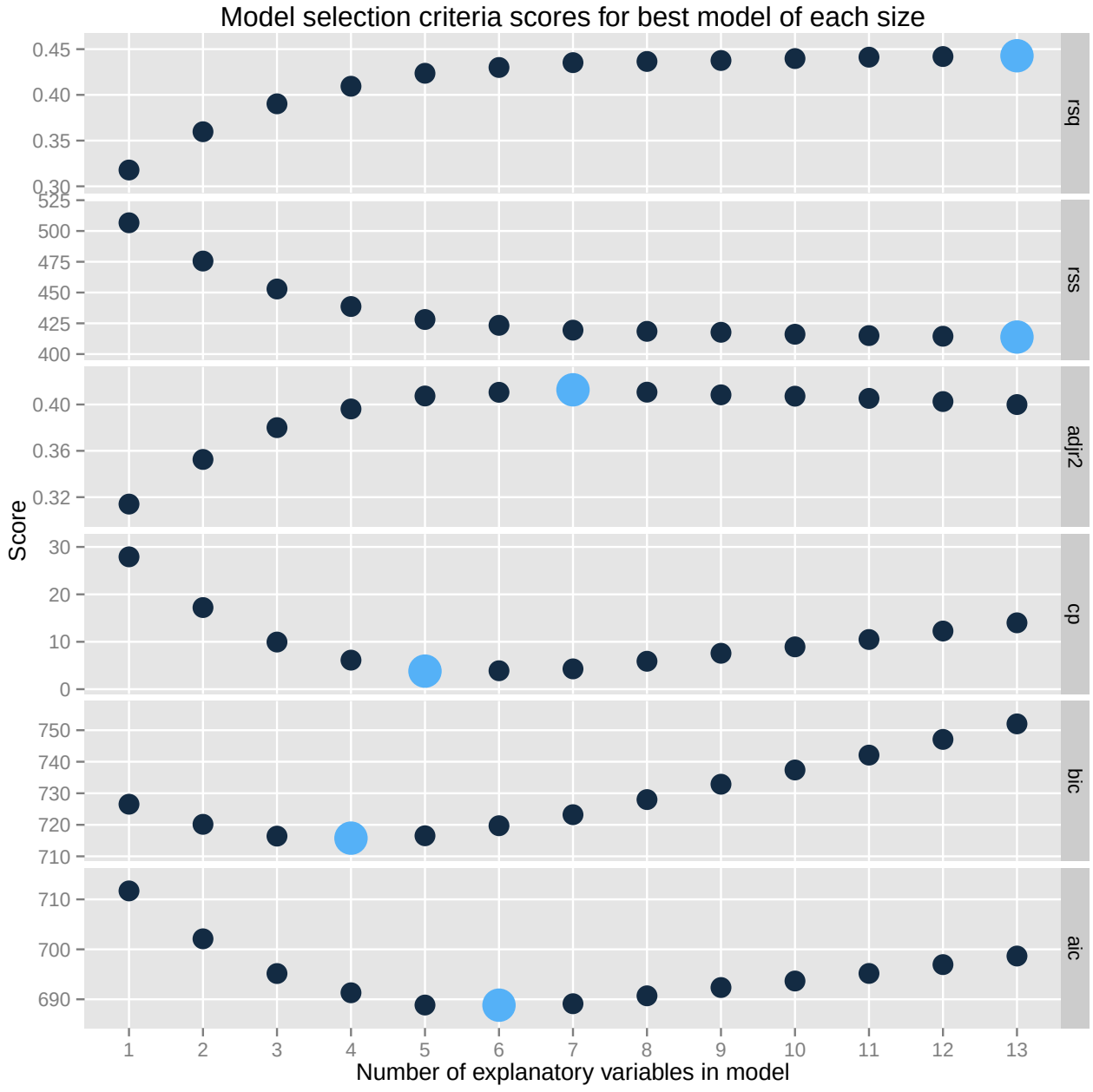
```
# Conduct all subsets model evaluation ---------------------------------------
naive_raw <- regsubsets(explanatory, response[[1]], nvmax = length(explanatory))
naive_results <- summary(naive_raw)
naive_results$bic <- naive_results$bic + bic(lm(data$SECMEAN~1))

# 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$size <- 1:nrow(scores)
calculate_aic <- function(included) {
  vars <- explanatory[included]
  model <- as.formula(paste("response[[1]] ~", paste("explanatory", names(vars),
                                                     collapse = " + ", sep = "$")))
  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())
```




```
naive_best_model <- lm(SECMEAN ~ 1 + AV.DEP + AG.TOT + LOG.TOT.RD + CAL.R, data = data)
naive_best_model
```

```
##
## Call:
## lm(formula = SECMEAN ~ 1 + AV.DEP + AG.TOT + LOG.TOT.RD + CAL.R,
```

```
##     data = data)
##
## Coefficients:
## (Intercept)       AV.DEP       AG.TOT   LOG.TOT.RD        CAL.R
##      2.6237       0.2563      -0.0179      -0.1087       0.3220
```

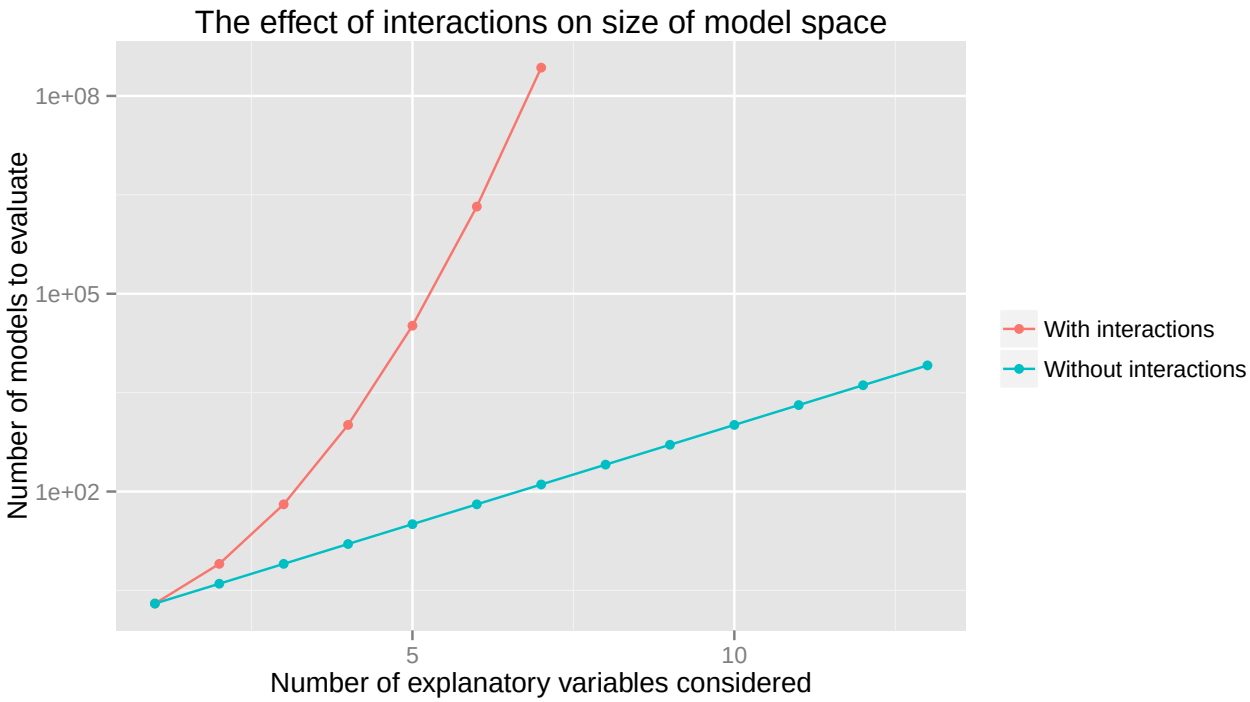
$$
\begin{aligned}
X_D &= \text{Estimated mean depth (m)} \\
X_W &= \text{Percent watershed agricultural} \\
X_R &= \text{Meters of road in watershed} \\
X_C &= \text{Number of calanoid species} \\
Y &= \text{Secchi depth (m)}
\end{aligned}
$$

$$Y = 2.623 + 0.322X_C + 0.256X_D - 0.018X_W - 0.109ln(X_R + 1)$$

All subsets with interactions

The R package `glmulti` allows for all subsets model selection with testing of all two way interactions. When interactions are included the number of potential models quickly increases with number oof explanatory variables used.

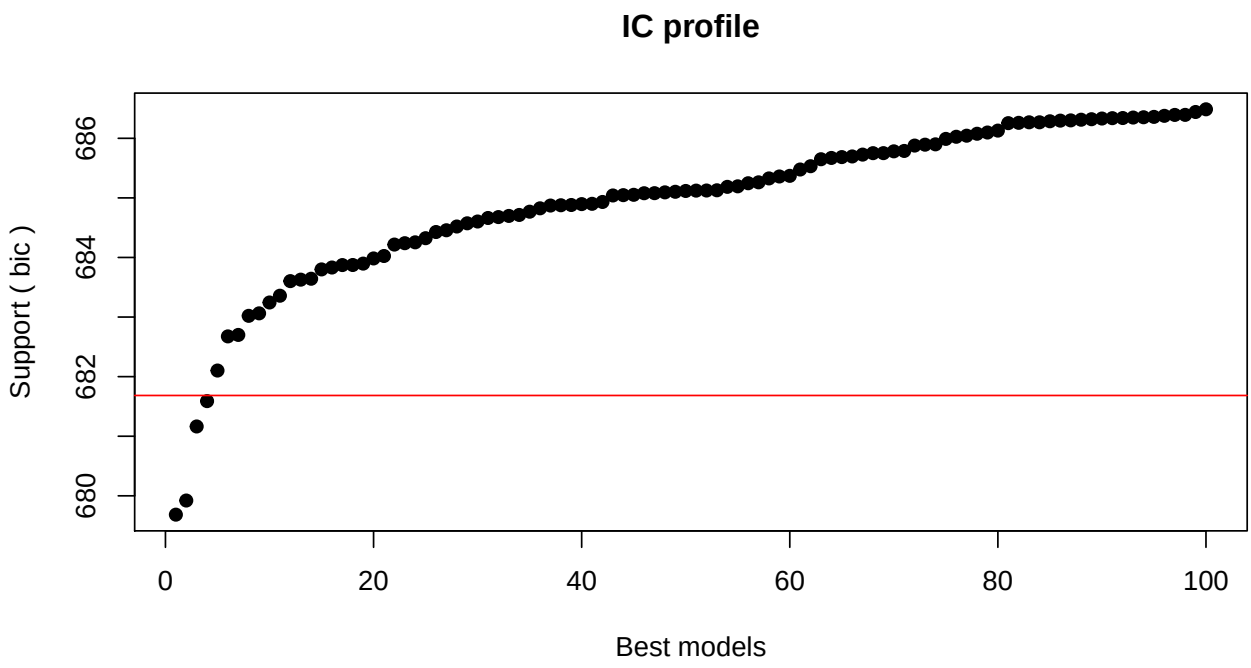
```
# Calculate the number of potential models to evaluate -------------------------
dont_print <- capture.output(
  count_with_iter <- sapply(1:7, function(x)
  glmulti(y = names(response), xr = names(explanatory[1:x]), data = data,
          level = 2, crit = "bic", method = "d")))
dont_print <- capture.output(
  count_without_iter <- sapply(1:length(explanatory), function(x)
  glmulti(y = names(response), xr = names(explanatory[1:x]), data = data,
          level = 1, crit = "bic", method = "d")))
# Graph results ----------------------------------------------------------------
counts <- data.frame(size = 1:length(explanatory),
                     with = c(count_with_iter, rep(NA, length(explanatory) - 7)),
                     without = count_without_iter)
counts <- melt(counts, id.vars = "size")
levels(counts$variable) <- c('With interactions', 'Without interactions')
ggplot(counts, aes(x = size, y = value, group = variable, color = variable)) +
  geom_line() +
  geom_point() +
  scale_y_log10() +
  labs(x = "Number of explanatory variables considered",
       y = "Number of models to evaluate",
       title = "The effect of interactions on size of model space",
       color = "")
```



For this reason, it is not pheasable to test all 15 aviable explanatory varaibles with interactions. Perliminary testing indicated that only 6 explanatory varaibles can be evaluated. We decided to exclude `LAT`, `LON`, and `ELEV` since their effects, if any, are likly due to climatic effects on other explanatory variables. `LK_HA` was excluded because it is probably corelated with `AV.DEP`, which is a much stronger predictor, as determined by the naive all-subsets analysis. Using similar logic, `URB.TOT` was excluded, since `POPDENKM` likly measures the same effect and was a strong predictor in the first analysis. Of the planktonic variables, all but `MACRO.R` and `MICRO.R` were exculded. The other planktonic variables measure the abundance of differnet taxa. Since none of the taxa are photosynthetic and likly lack pigment, the size of the organism is probably more important than taxonomic group. This leaves `AG.TOT`, `AV.DEP`, `LOG.POPDENKM`, `LOG.TOT.RD`, `MACRO.R`, and `MICRO.R` as the predictors that will be evaluated when considering interactions.

```
relevant_exp <- c("AG.TOT", "AV.DEP", "LOG.POPDENKM", "LOG.TOT.RD", "MACRO.R", "MICRO.R")
glmulti_model <- glmulti(y = names(response), xr = relevant_exp, data = data, level = 2, crit = "bic")
save.image()
```

```
load(file = ".RData")
plot(glmulti_model)
```




```
glmulti_model@objects[[1]]
```

```
##
## Call:  fitfunc(formula = as.formula(x), data = data)
##
## Coefficients:
##        (Intercept)              AV.DEP          LOG.TOT.RD
##            1.41736             0.93104             -0.14314
##  AV.DEP:LOG.TOT.RD      MACRO.R:AG.TOT      AV.DEP:MACRO.R
##           -0.03342            -0.00337             -0.04335
## LOG.TOT.RD:MACRO.R
##            0.03139
##
## Degrees of Freedom: 182 Total (i.e. Null);  176 Residual
## Null Deviance:       743
## Residual Deviance: 350   AIC: 654
```

```
glmulti_model@objects[[1]]$formula
```

```
## SECMEAN ~ 1 + AV.DEP + LOG.TOT.RD + LOG.TOT.RD:AV.DEP + MACRO.R:AG.TOT +
##     MACRO.R:AV.DEP + MACRO.R:LOG.TOT.RD
## <environment: 0x722a840>
```

```
bic(glmulti_model@objects[[1]])
```

```
## [1] 679.7
```

$$
\begin{aligned}
X_D &= \text{Estimated mean depth (m)} \\
X_W &= \text{Percent watershed agricultural} \\
X_R &= \text{Meters of road in watershed} \\
X_M &= \text{Number of macrozooplankton species} \\
Y &= \text{Secchi depth}(m)
\end{aligned}
$$

$$
Y = 1.417 + 0.931 X_D - 0.143 ln(X_R + 1) - 0.033 X_D ln(X_R + 1) - 0.003 X_M X_W - 0.043 X_D X_M + 0.031 ln(X_R + 1) X_M
$$