brute_force_SEM_parallel.R

au653261

2021-02-06

```
# BRUTE FORCE VARIABLE SELECTION ##############################################################################
rm(list=ls())
library(lavaan)
```

```
## This is lavaan 0.6-7
```

```
## lavaan is BETA software! Please report any bugs.
```

```
library(stringr)
library(ggplot2)
```

```
## Warning: replacing previous import 'vctrs::data_frame' by 'tibble::data_frame'
## when loading 'dplyr'
```

```
library(parallel)
library(foreach)
library(doParallel)
```

```
## Loading required package: iterators
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v tibble  3.0.4     v purrr   0.3.3
## v tidyr   1.0.2     v dplyr   1.0.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x purrr::accumulate() masks foreach::accumulate()
## x dplyr::filter()     masks stats::filter()
## x dplyr::lag()        masks stats::lag()
## x purrr::when()       masks foreach::when()
```

```
dat_no.na <- readRDS("../data/sem_input_data.rds")

# set max number of variables (to add to the fixed ones) used and register cores
var.max <- 8
registerDoParallel(4)
loop_active <- 1 # set to length(main_reg) to run the entire loop for actual analysis

# build regression character strings ##########################################################################
sr_variables <- c("sub_trop_dbf",  "pre_sd", "mont_gs" , "pre_m", "sea_sd", "tra_m"
                  , "mat_m" ,"tra_sd", "mat_sd" , "sea_m" , "medit_fws" , "temp_gss" , "tri" , "pet_sd"
mrd_variables <- c("tra_m", "sea_sd", "soil" , "tri", "tra_sd" , "pre_sd" , "pet_m", "mat_m" , "area"
                   , "mat_sd" , "temp_bmf" , "sub_trop_mbf" , "pet_sd" , "medit_fws" , "sub_trop_cf" ,
fixed_sr <- c("sr_trans ~ soil + sub_trop_mbf + area + mrd +")
```

```r
fixed_mrd <- c("mrd ~ pre_m + sea_m +")

sr_var_list <- do.call("c", lapply(seq_along(sr_variables[1:var.max]), function(i) combn(sr_variables[1
sr_combs <- c()
Sys.time() # TAKES 20 secs
```

```
## [1] "2021-02-06 12:01:14 CET"
```

```r
for(i in 1:length(sr_var_list)){ #
  temp <- unlist(sr_var_list[i])
  sr_combs <- c(sr_combs, str_c(temp, collapse = "+"))
}
Sys.time()
```

```
## [1] "2021-02-06 12:01:14 CET"
```

```r
mrd_var_list <- do.call("c", lapply(seq_along(mrd_variables[1:var.max]), function(i) combn(mrd_variables
mrd_combs <- c()
Sys.time() # TAKES 30secs
```

```
## [1] "2021-02-06 12:01:14 CET"
```

```r
for(i in 1:length(mrd_var_list)){ #
  temp <- unlist(mrd_var_list[i])
  mrd_combs <- c(mrd_combs, str_c(temp, collapse = "+"))
}
Sys.time()
```

```
## [1] "2021-02-06 12:01:14 CET"
```

```r
# build all options combining the two regression combos (use)
temp <- expand.grid(sr_combs, mrd_combs)
main_reg = paste(paste(fixed_sr,temp$Var1), paste(fixed_mrd, temp$Var2), sep="\n")
cat(main_reg[54])
```

```
## sr_trans ~ soil + sub_trop_mbf + area + mrd + sub_trop_dbf+sea_sd+tra_sd
## mrd ~ pre_m + sea_m + tra_m
```

```r
# set indirect path dependencies ##############################################################

dir_path_soil_area <- "soil ~ area"                    # always present
dir_path_sea_sd_area <- "sea_sd ~ area"                # grepl("sea_sd", mod)
dir_path_tra_sd_area <- "tra_sd ~ area"                # grepl("tra_sd", mod)
subtrop_path <- "sub_trop_mbf ~ pre_m + tra_m + pet_m"  # always present
desert_path <- "deserts_x_shrub ~ pre_m + sea_m + pet_m"  # if habitat included

# loop through model options  _ TAKES TIME! ###################################################

#### %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
#### ~ 5 hours for 2^9 * 2^9 combinations for serial analysis
#### %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

Sys.time()
```

```
## [1] "2021-02-06 12:01:14 CET"
```

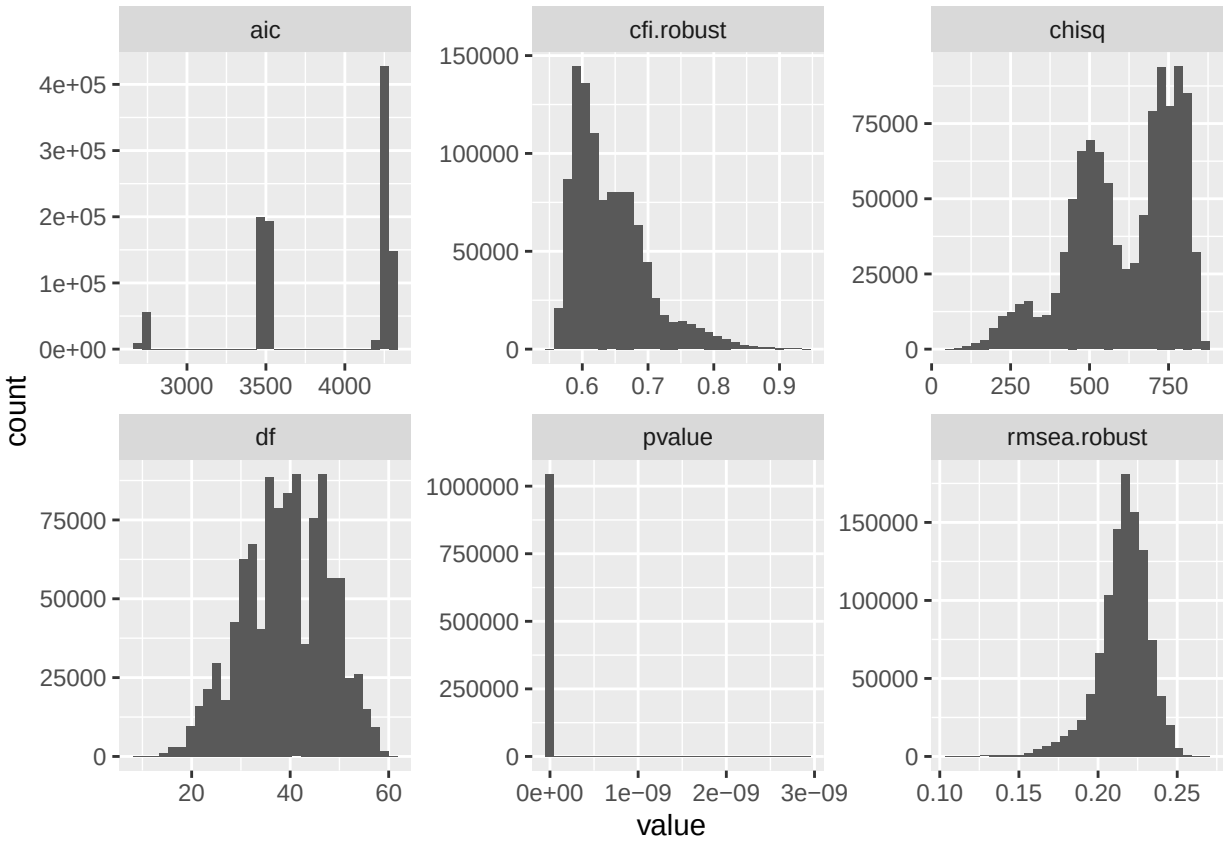
```r
system.time({
# r2list <- list()
```

```
# mod.list <- list()
temp <- foreach(i = 1:loop_active) %dopar% {
    # build regression models
    mod_main <- paste(c(main_reg[i], dir_path_soil_area, subtrop_path), collapse = "\n")
    mod <- mod_main
    #cat(main_reg)

    if(grepl("sea_sd", mod_main)){
      mod <- paste(c(mod_main, dir_path_sea_sd_area), collapse = "\n")
    }
    if(grepl("tra_sd", mod_main)){
      mod <- paste(c(mod_main, dir_path_tra_sd_area), collapse = "\n")
    }
    if(grepl("sea_sd", mod_main) & grepl("tra_sd", mod_main)){
      mod <- paste(c(mod_main, dir_path_sea_sd_area, dir_path_tra_sd_area), collapse = "\n")
    }
    if(grepl("deserts_x_shrub", mod_main)){
      mod <- paste(c(mod_main, dir_path_sea_sd_area, dir_path_tra_sd_area, desert_path), collapse = "\n
    }

    modfit <- sem(mod, data = dat_no.na, estimator="MLM")
    #if(i==1){
      res <- fitmeasures(modfit, c("chisq", "df", "pvalue", "cfi.robust", "rmsea.robust", "aic"))
    #}else{
    #  res <- rbind(res, fitmeasures(modfit, c("chisq", "df", "pvalue", "cfi.robust", "rmsea.robust", "
    #}
    #r2list[[i]] <- lavInspect(modfit, "r2")
    #mod.list[[i]] <- mod
    list(res, mod, lavInspect(modfit, "r2")[1:2])
  }
})
```

```
##    user  system elapsed
##   0.109   0.012   0.122
```

```
#save(temp, file="brute_force_mod_selection_results_parallel.RData")

# Analyse output ##############################################################################################
load("../data/brute_force_mod_selection_results_parallel.RData")
fit <- sapply(temp, "[", 1)
#rm(temp)
fit2 <- data.frame(matrix(unlist(fit), nrow=length(fit), byrow=T))
rm(fit)
names(fit2) <-  c("chisq", "df","pvalue","cfi.robust","rmsea.robust","aic")
#saveRDS(fit2, "goodness_of_fit_10.rds")

ggplot(pivot_longer(fit2, cols = names(fit2)), aes(value))+
  geom_histogram()+
  facet_wrap(~name, scales = "free")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
load("../data/brute_force_mod_selection_results_parallel.RData")
fit2 <- readRDS("goodness_of_fit_10.rds")
model.name <- sapply(temp, "[", 2)
fit2$model.name <- unlist(model.name)
rm(model.name)

head(fit2)
```

```
##      chisq df       pvalue cfi.robust rmsea.robust      aic
## 1 132.5533 18 0.000000e+00  0.8763446    0.1401042 2759.345
## 2 175.2246 18 0.000000e+00  0.8384671    0.1649410 2748.201
## 3 122.5489 18 0.000000e+00  0.8821956    0.1359243 2760.815
## 4 105.5278 14 4.440892e-16  0.8965835    0.1432904 2758.495
## 5 288.4482 22 0.000000e+00  0.7699103    0.1968836 3529.806
## 6 106.9192 14 2.220446e-16  0.8965711    0.1432990 2759.886
##
## 1          sr_trans ~ soil + sub_trop_mbf + area + mrd + sub_trop_dbf\nmrd ~ pre_m + sea_m + tra_m\ns
## 2               sr_trans ~ soil + sub_trop_mbf + area + mrd + pre_sd\nmrd ~ pre_m + sea_m + tra_m\ns
## 3              sr_trans ~ soil + sub_trop_mbf + area + mrd + mont_gs\nmrd ~ pre_m + sea_m + tra_m\ns
## 4                sr_trans ~ soil + sub_trop_mbf + area + mrd + pre_m\nmrd ~ pre_m + sea_m + tra_m\ns
## 5 sr_trans ~ soil + sub_trop_mbf + area + mrd + sea_sd\nmrd ~ pre_m + sea_m + tra_m\nsoil ~ area\nsul
## 6                sr_trans ~ soil + sub_trop_mbf + area + mrd + tra_m\nmrd ~ pre_m + sea_m + tra_m\ns
```

```
which.max(fit2$cfi.robust) # max CFI
```

```
## [1] 224086
```

```
which.min(fit2$rmsea.robust) # min RMSEA
```

```
## [1] 125832
```

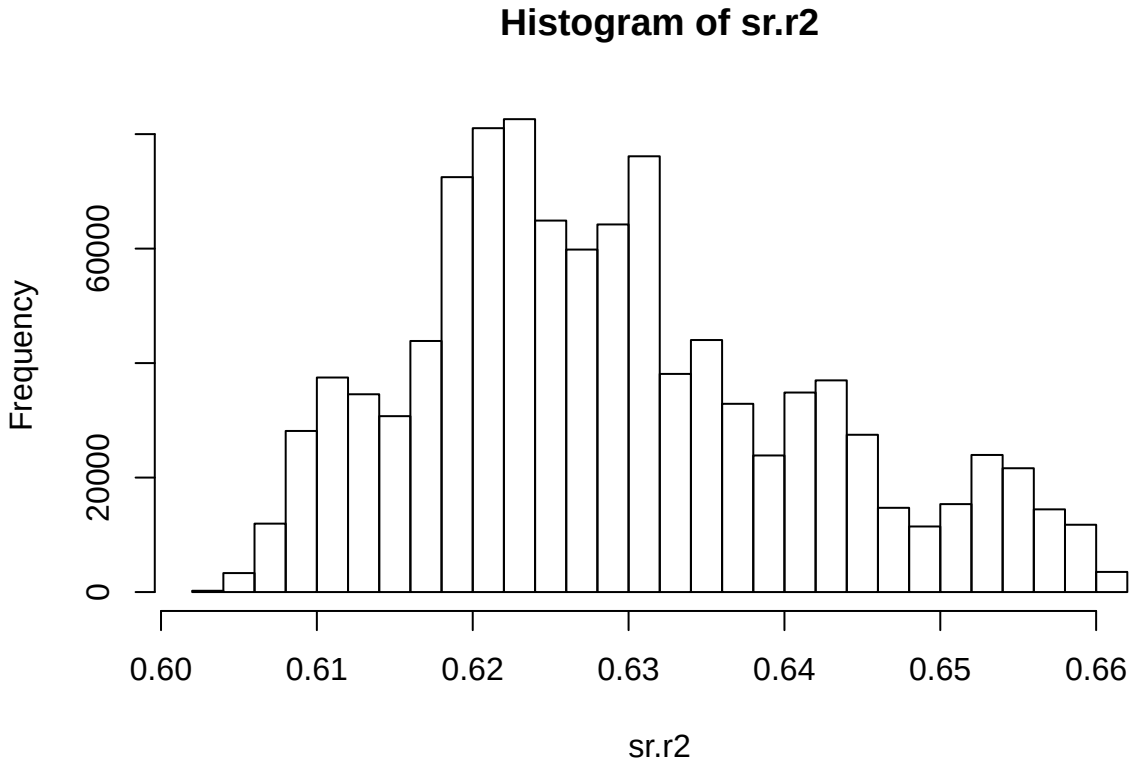
```
# model fit
length(which(fit2$cfi.robust>0.9))/nrow(fit2) # 753 models have an acceptable robust CFI (with 9 variab
```

```
## [1] 0.0007758982
```

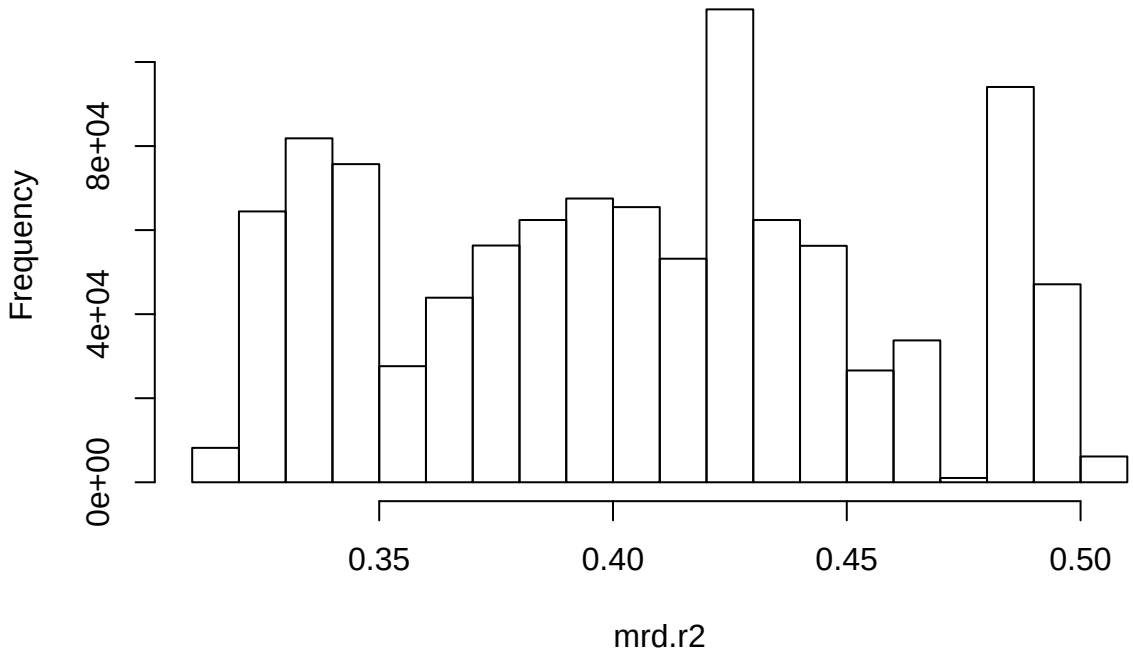
```
length(which(fit2$rmsea.robust<0.11))/nrow(fit2) # 14 with 10 variables
```

```
## [1] 1.337756e-05
```

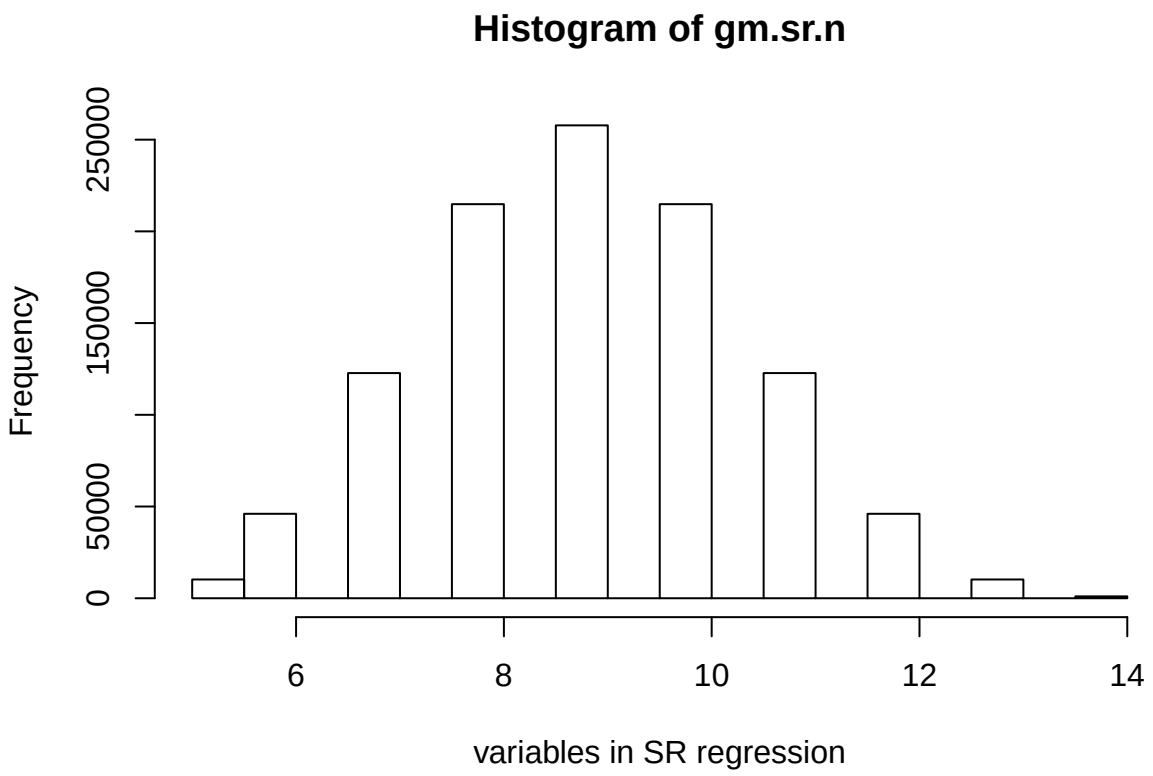
```
# explanatory power
## extract first two
pow <- sapply(temp, "[", 3)
rm(temp)
sr.r2 <- sapply(pow, "[[", 1)
mrd.r2 <- sapply(pow, "[[", 2)
rm(pow)
hist(sr.r2)
```



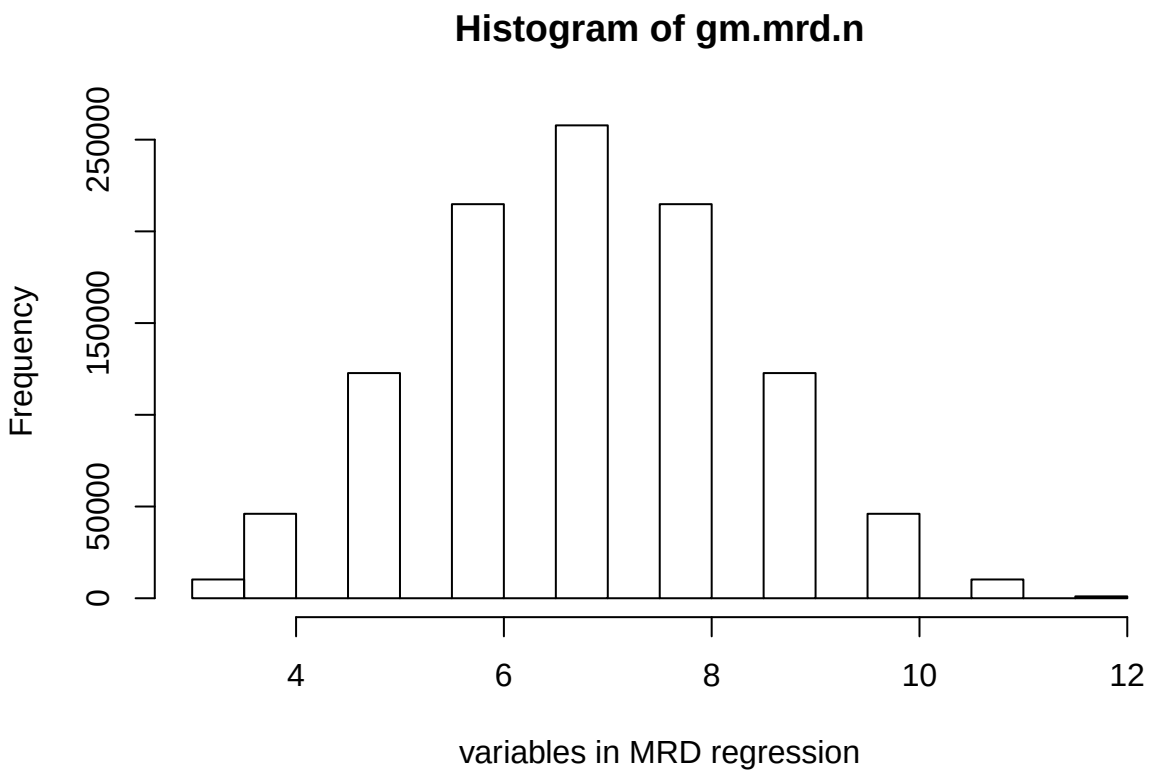

```
hist(mrd.r2)
```



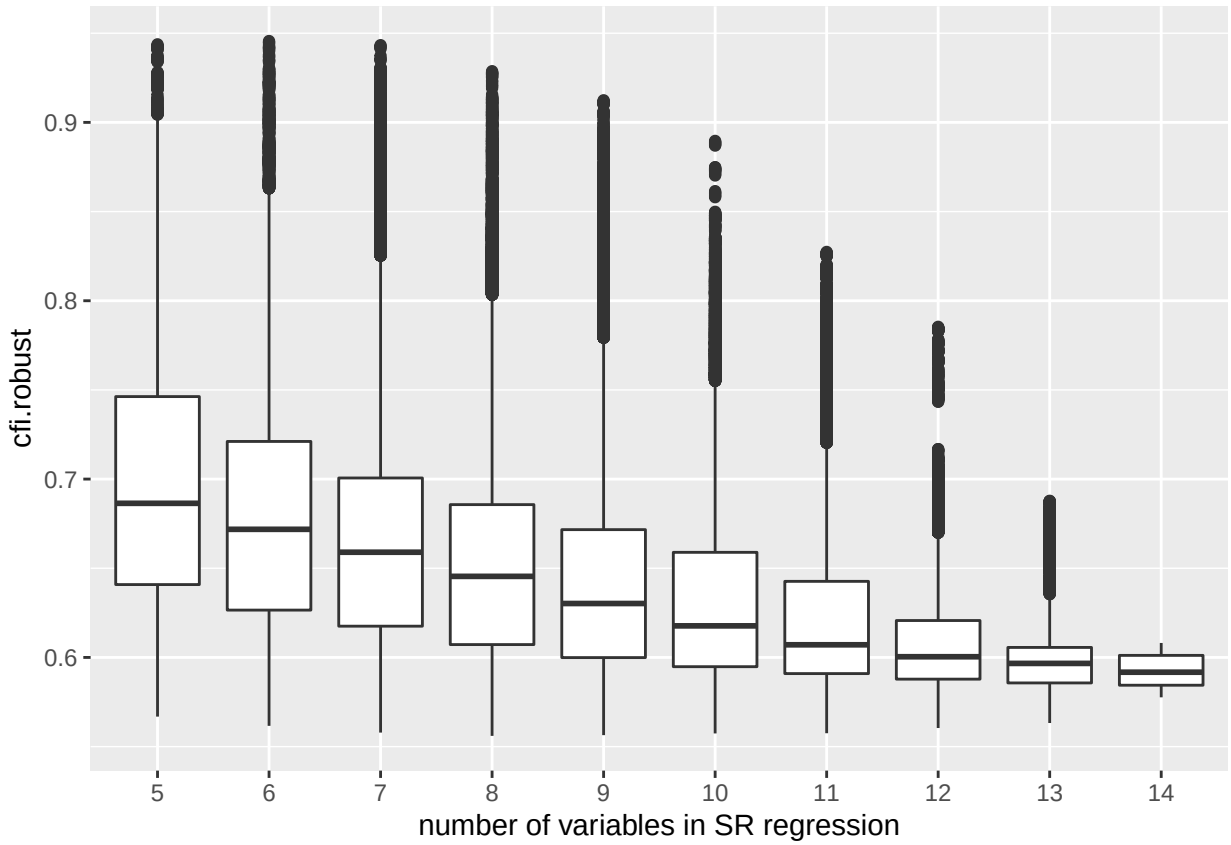

```
# get models props
## number of variables connected to CFI
# fit2$model.name
sr1 <- sub("\n.*", "", fit2$model.name)
gm.sr <- sub(".*~ ", "", sr1)
gm.sr.sp <- strsplit(gm.sr, "+", fixed = TRUE)
gm.sr.n <- unlist(lapply(gm.sr.sp, length))
hist(gm.sr.n, xlab="variables in SR regression") # the number of variables in the good models
```



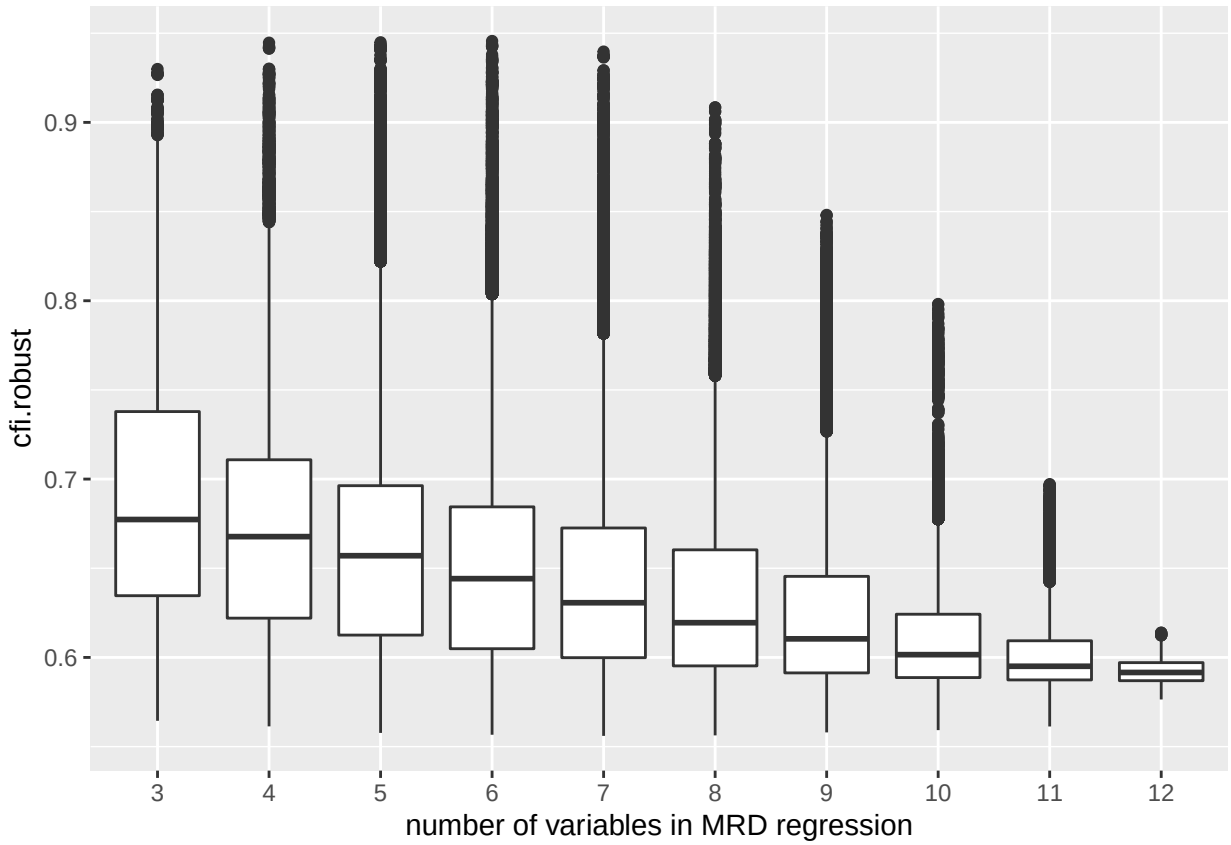

```
fit2$gm.sr.n <- gm.sr.n
rm(gm.sr.sp)

mrd1 <- sub(".*\nmrd ~ ", "", fit2$model.name)
gm.mrd <- sub("\nsoil.*", "", mrd1)
gm.mrd.sp <- strsplit(gm.mrd, "+", fixed = TRUE)
gm.mrd.n <- unlist(lapply(gm.mrd.sp, length))
hist(gm.mrd.n, xlab="variables in MRD regression") # the number of variables in the good models
```



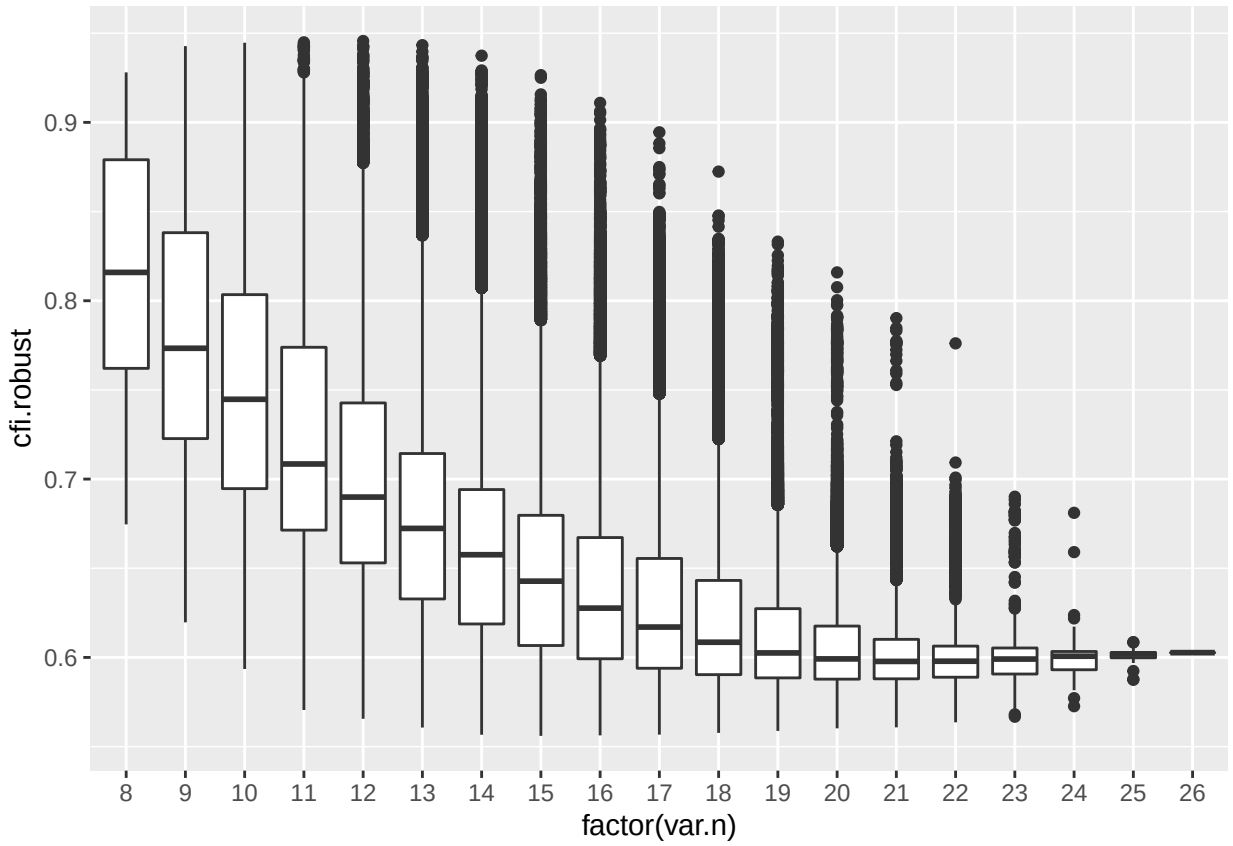

```
fit2$gm.mrd.n <- gm.mrd.n
rm(gm.mrd.sp)

rm(gm.mrd, gm.mrd.n, gm.sr, gm.sr.n)

ggplot(fit2, aes(factor(gm.sr.n), cfi.robust)) +
  geom_boxplot()+
  xlab("number of variables in SR regression")
```



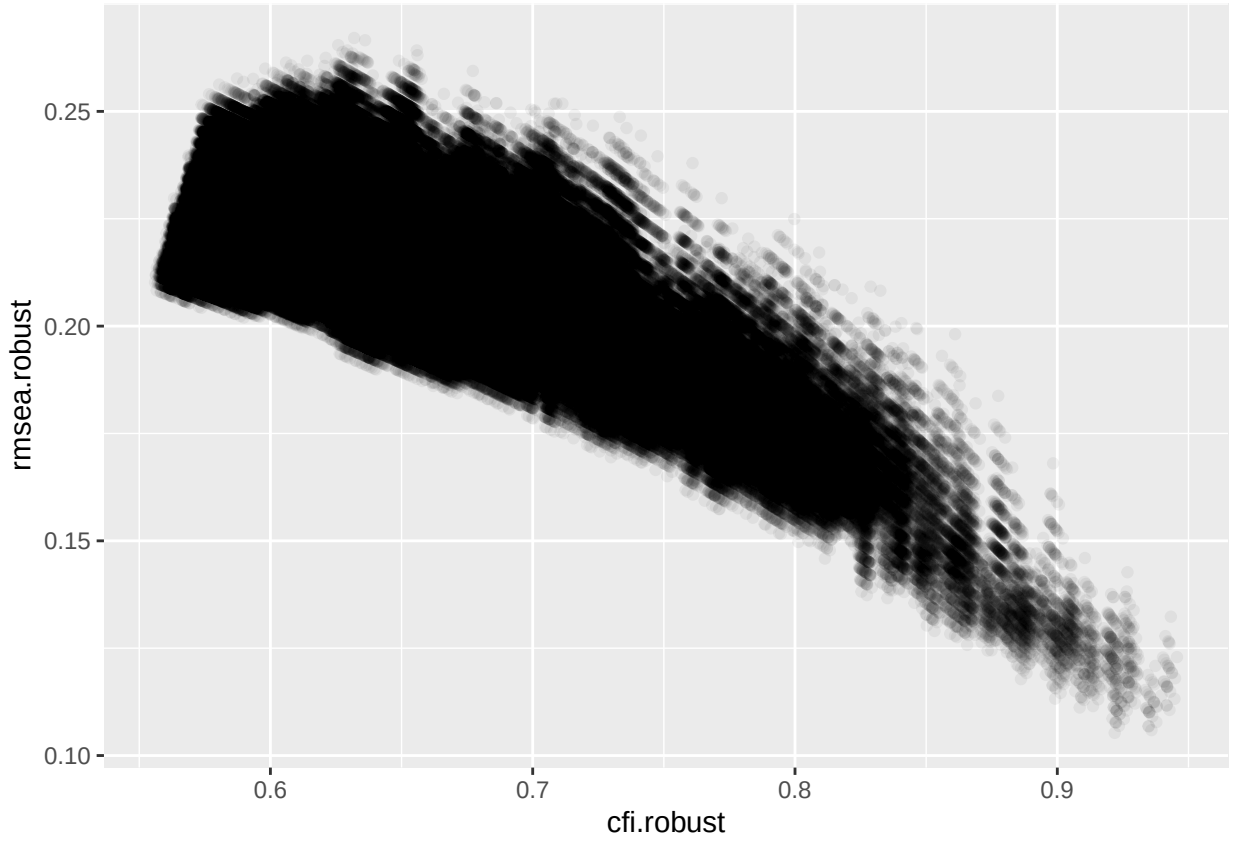

```
ggplot(fit2, aes(factor(gm.mrd.n), cfi.robust)) +
  geom_boxplot()+
  xlab("number of variables in MRD regression")
```




```
fit2$var.n <- fit2$gm.sr.n+fit2$gm.mrd.n
ggplot(fit2, aes(factor(var.n), cfi.robust)) +
  geom_boxplot()
```




```
# ggplot(fit2, aes(gm.mrd.n, gm.sr.n, col=cfi.robust)) +
#   geom_jitter()

ggplot(fit2, aes(cfi.robust, rmsea.robust)) +
  geom_point(alpha=0.05)
```




```
# find model that has most variables with best fit
a <- max(fit2$var.n[which(fit2$cfi.robust>0.9)])
which(fit2$cfi.robust>0.9 & fit2$var.n==a)
```

```
## [1] 461823 483345 483450 739992
```

```
fin <- fit2$model.name[which(fit2$cfi.robust>0.9 & fit2$var.n==a)]
cat(fin[1])
```

```
## sr_trans ~ soil + sub_trop_mbf + area + mrd + sub_trop_dbf+mont_gs+pre_m+tra_m+sea_m
## mrd ~ pre_m + sea_m + tra_m+soil+tri+pet_m+area
## soil ~ area
## sub_trop_mbf ~ pre_m + tra_m + pet_m
```

```
cat(fin[2])
```

```
## sr_trans ~ soil + sub_trop_mbf + area + mrd + sub_trop_dbf+pre_m+tra_m+mat_m+sea_m
## mrd ~ pre_m + sea_m + tra_m+soil+pet_m+mat_m+area
## soil ~ area
## sub_trop_mbf ~ pre_m + tra_m + pet_m
```

```
cat(fin[3])
```

```
## sr_trans ~ soil + sub_trop_mbf + area + mrd + mont_gs+pre_m+tra_m+mat_m+sea_m
## mrd ~ pre_m + sea_m + tra_m+soil+pet_m+mat_m+area
## soil ~ area
## sub_trop_mbf ~ pre_m + tra_m + pet_m
```

```
cat(fin[4])
```

```
## sr_trans ~ soil + sub_trop_mbf + area + mrd + pre_m+tra_m+mat_m+sea_m
```

```
## mrd ~ pre_m + sea_m + tra_m+soil+tri+pet_m+mat_m+area
## soil ~ area
## sub_trop_mbf ~ pre_m + tra_m + pet_m
```

```
# models with mode indirect paths
b <- gregexpr("\n", fit2$model.name)
d <- lapply(b, length)
rm(b)
fit2$reg.n <- unlist(d)+1

ggplot(fit2, aes(factor(reg.n), cfi.robust)) +
  geom_boxplot()+
  xlab("Number of regressions")
```