Generating new prediction models for 1960 cohort

Eric Chow
July 23, 2019

Introduction: The LCRFsim R package contains stripped-down prediction models from PLCO data for 4 risk ractors, (BMI, FHL, PH, and COPD), stratified by male/female, and 1950/1960 (4 strata of 4 models). We cannot distribute PLCO data (or the residuals stored in the prediction models) with the LCRFsim however. The models are also extremely large due to the residuals (~40MB) and bloat the package to a non-distributable size.

Objective of walkthrough: At the end of this walkthrough, you will obtain 16X prediction models saved in a single `sysdata.rda` file (without stored residuals they are <0.2MB each, but keeping splines in the model environment for the ns() [ie: natural spline] terms to work). The `sysdata.rda` file is distributed with the LCRFsim package in the LCRFsim/R folder.

---

1. Running the PLCO regressions

The models are ran by the `RUN_18_0509_Extension_1960.cohort_PLCO.ANYCOHORT_ERIC.r` file (Summer Han), which calls these 4 functions (defined in `source.NLST.functions_EC.R`) that do the actual fitting: `myBMI.cond3()`, `myFHL.cond3()`, `myPH.cond2()`, and `myCOPD.cond3().`

Around line 577 of `RUN_18_0509_Extension_1960[…].r, you will find the 4 output files of the regressions:`

```
OUT.bmi  = myBMI.cond3()
OUT.fh   = myFHL.cond3()
OUT.ph   = myPH.cond2()
Out.copd = myCOPD.cond3()
```

For each of 4 strata (1950/1960 & M/F), you will need to obtain the 4 OUT.xyz files above. So run the code 4 times, setting:

```
cohort = 1950; Gender = "F"
cohort = 1950; Gender = "M"
cohort = 1960; Gender = "F"
cohort = 1960; Gender = "M"
```

This code is available in the attached `strip_model.R` Which has all the supplemental code required (besides Summer's code) to run Summer's code 16X and create a sysdata.rda file to include in the package.

```
# ------------------------------------------------------------------------------
# Trimming the size of saved models in R
#
# Eric Chow, 2017 (updated 2019)
#
# The code below uses RDS files, R environments, and the strip library to reduce
# the size of saved models from R.
#
# Models in R are saved with a lot of information so that they can be fully
# functional (for updating, predicting, etc). For complex analyses, the R
# environment may be bloated with the actual data that was used for constructing
# splines or other "functional" terms.  This environment (the namespace of R
# at the time the model was ran) gets tacked onto the saved my_model under:
#
# attr(my_model$terms, ".Environment")
#
# You can set it to NULL, but that also removes the necessary package references
# So an alternative is to start a blank R session, load the libraries that your
# model needs, and insert your current global environment, .GlobalEnv into the
# saved model's .Environment.
#
#
# ------------------------------------------------------------------------------

# ------------------------------------------------------------------------------
# what is in the sysdata.rda file that ships with LCRFsim?
rm(list=ls()); gc()
load(file = "~/QSU/LCsim/LCRFsim/R/sysdata.rda") # doesn't put into object, because it's a package of
objects already
ls()  # hrmmm I see.
# ------------------------------------------------------------------------------
```

```
# ------------------------------------------------------------------------------
# RUNNING THE PREDICTION MODELS TO GET 16X rds files:
# copy the follow code snippets into:
#
#   19_0709_18_0509_1960.cohort.extension_risk.generator___________________ ERIC.R
#
# at line 4:

                # ERIC ADDED 7/15/2019 ---------------------------------
                cohort = 1950; Gender = "F"
                # cohort = 1950; Gender = "M"
                # cohort = 1960; Gender = "F"
                # cohort = 1960; Gender = "M"
                # ------------------------------------------------------

# at line 30 (after RUN__18_0509_Extension_1960.cohort_PLCO.ANYCOHORT__ERIC.r has been sourced)
# and the OUT.xyz objects exist.
```

```
                # ERIC ADDED 7/15/2019 --------------------------------------------------
                library(stringr)
                # CREATE /rds directory!
                saveRDS(OUT.bmi,  str_c("rds/OUT.bmi.", Gender , ".", cohort, ".rds")  )
                saveRDS(OUT.copd, str_c("rds/OUT.copd.", Gender , ".", cohort, ".rds") )
                saveRDS(OUT.fh,   str_c("rds/OUT.fh.", Gender , ".", cohort, ".rds")   )
                saveRDS(OUT.ph,   str_c("rds/OUT.ph.", Gender , ".", cohort, ".rds")   )
                saveRDS(PROB.edu.race, str_c("rds/PROB.edu.race", Gender , ".", cohort, ".rds")   )
                # -----------------------------------------------------------------------

# now you should have 4 files in the /rds directory, do it again x4 for all 16!
# ------------------------------------------------------------------------------
```

```
# ------------------------------------------------------------------------------
# which files didn't work? I needed the imputation files from summer
# 18_0509_PLCO.fit.models_male_1960_imputation.file.csv
# 16_0303_PLCO.fit.models_male_imputation.file.csv

library(stringr)
setwd("~/QSU/LCsim/fit_plco_1960")

rdss <- c(      "OUT.bmi.F.1950.rds", "OUT.bmi.F.1960.rds", "OUT.bmi.M.1950.rds", "OUT.bmi.M.1960.rds",
                                                "OUT.fh.F.1950.rds", "OUT.fh.F.1960.rds",
"OUT.fh.M.1950.rds", "OUT.fh.M.1960.rds",
                                                "OUT.copd.F.1950.rds", "OUT.copd.F.1960.rds",
"OUT.copd.M.1950.rds", "OUT.copd.M.1960.rds",
                                                "OUT.ph.F.1950.rds", "OUT.ph.F.1960.rds",
"OUT.ph.M.1950.rds", "OUT.ph.M.1960.rds",
                                                "PROB.edu.raceF.1950.rds", "PROB.edu.raceF.1960.rds",
"PROB.edu.raceM.1950.rds", "PROB.edu.raceM.1960.rds")

for (rds in rdss) {
        fit  <- readRDS(str_c("rds/",rds))
        cat(rds, "\t",names(fit),"\n\n")
}
# ------------------------------------------------------------------------------
```

```
# ------------------------------------------------------------------------------
# Reduce the object size of each OUT.xyz in this codes. Unfortunately, I can't
# figure out a way to loop it b/c I have to clear the environment each time,
# so you'll just have to run it 16X for each of these rds:
```

```
                                 # "OUT.bmi.F.1950.rds"                   X
                                 # "OUT.bmi.F.1960.rds"                   X
                                 # "OUT.bmi.M.1950.rds"                   X
                                 # "OUT.bmi.M.1960.rds"                   X
                                 # "OUT.fh.F.1950.rds"                      X
                                 # "OUT.fh.F.1960.rds"       X
                                 # "OUT.fh.M.1950.rds"       X
                                 # "OUT.fh.M.1960.rds"       X
                                 # "OUT.copd.F.1950.rds"     X
                                 # "OUT.copd.F.1960.rds"     X
                                 # "OUT.copd.M.1950.rds"     X
                                 # "OUT.copd.M.1960.rds"     X
                                 # "OUT.ph.F.1950.rds"       X
                                 # "OUT.ph.F.1960.rds"       X
                                 # "OUT.ph.M.1950.rds"       X
                                 # "OUT.ph.M.1960.rds"       X
                                 # "PROB.edu.raceF.1950.rds"     # Don't need to strip these ones
                                 # "PROB.edu.raceF.1960.rds"
                                 # "PROB.edu.raceM.1950.rds"
                                 # "PROB.edu.raceM.1960.rds"

        # clear the memory entirely, needed to have a clean, empty Global Environment
        rm(list=ls()); gc();

        # load the packages that my saved model will need. For example, if my models
        # includes a spline term using the ns() function, it will need splines library
        library(splines)
        library(strip)

        # I open the saved model (which is bloated) - saved as an RDS file
        setwd("~/QSU/LCsim/fit_plco_1960")
        RDS <- "PROB.edu.raceF.1950.rds"
        OUT.xyz  <- readRDS(stringr::str_c("rds/", RDS ))

        # find the largest thing in the OUT.xys object, likely will be the model
        max_size = 0; for (item in names(OUT.xyz)) {
                this_size <- object.size(OUT.xyz[item])
                if (this_size > max_size) {
                        max_size = this_size
                        fit <- item
                }
        }; fit

  # how big is the fit?
        object.size(OUT.xyz[[fit]]) # it is very large

        # I use the strip function from library(strip) to trim it a little
        # but keep predict functionality
        less_bloated_model <- strip(OUT.xyz[[fit]], keep="predict")

        object.size(less_bloated_model)  # it is slightly smaller, but still very large

        # I now replace the model's Environment term with the current global environments
        # which has the splines and strip library loaded. the model's Environment
        # contains all the data and namespace that was present when the model was
        # originally ran.  You can also make it NULL, however, your model's namespace
        # will no longer include the packages it needs to function.  You can set it
        # to NULL if your model is simple and doesn't include any function calls like ns()
        attr(less_bloated_model$terms,".Environment") <- .GlobalEnv # reload splines into the model's
environment
```

```
        object.size(less_bloated_model) # it is much smaller now!

        # put it back into OUT objects
        OUT.xyz[[fit]] <- less_bloated_model; str(OUT.xyz)
        # now it is way smaller
        object.size(OUT.xyz)

        # save the shrunken model back out to a new RDS file to later be saved to an R package
        saveRDS(OUT.xyz, stringr::str_c("rds/stripped_", RDS ))
# ------------------------------------------------------------------------------
```

```
# ------------------------------------------------------------------------------
# open up all the RDS objects that are small and package them into on rda files

rm(list=ls()); gc();
library(stringr)
setwd("~/QSU/LCsim/fit_plco_1960")
ls()

stripped_rdss <- c(      "stripped_OUT.bmi.F.1950.rds", "stripped_OUT.bmi.F.1960.rds",
"stripped_OUT.bmi.M.1950.rds", "stripped_OUT.bmi.M.1960.rds",
                                                 "stripped_OUT.fh.F.1950.rds",
"stripped_OUT.fh.F.1960.rds", "stripped_OUT.fh.M.1950.rds", "stripped_OUT.fh.M.1960.rds",
                                                 "stripped_OUT.copd.F.1950.rds",
"stripped_OUT.copd.F.1960.rds", "stripped_OUT.copd.M.1950.rds", "stripped_OUT.copd.M.1960.rds",
                                                 "stripped_OUT.ph.F.1950.rds",
"stripped_OUT.ph.F.1960.rds", "stripped_OUT.ph.M.1950.rds", "stripped_OUT.ph.M.1960.rds",
                                                 "PROB.edu.raceF.1950.rds", "PROB.edu.raceF.1960.rds",
"PROB.edu.raceM.1950.rds", "PROB.edu.raceM.1960.rds")

# read in each file into the name
for (rds in stripped_rdss) {
        rds_name <- str_replace(rds, "stripped_", "") # remove stripped
        rds_name <- str_replace(rds_name, ".rds", "") # remove .rds
        rds_cmd <- str_c(rds_name, " <- readRDS(str_c('rds','/',rds))")
        eval(parse(text=rds_cmd))
        cat(rds_name, "\n")
}

# remove unnecessary objects and save stripped objects together into rda
rm(rds); rm(rds_cmd); rm(rds_name); rm(stripped_rdss)
# save the selected objects to an rda file
save(list=ls(), file = "rds/sysdata.rda")


# test open it
rm(list=ls()); gc();
load("rds/sysdata.rda")
ls()
```

# ~ fin ~