CLSA cross omics code - example code for creating 1- 2- and 3-way CRP risk scores

AYM

2025-05-17

R Markdown

```
library(glmnet)
library (Biobase)
library (utils)
library(devtools)
library(tibble)
library(ggplot2)
library (tidyverse)
library(data.table)
library(readr)
library (methods)
library (dplyr)
library(readxl)
library(stats)
```

```
#1-way CRP ERS
#Establish 1-way risk score (PRS, MRS, ERS) for each omics as predictor
#(genetics, metabolomics, epigenetics) using e.g., penalizedregression or
#other continuous shrinkage methods to predict CRP:

bVals <-read_csv("bVals_quantile_preprocessing_N1476.csv")
bVals <- as.data.frame (bVals)
colnames (bVals)
bVals = subset(bVals, select = -c(...1) )

target <- read.csv("CLSA_metadata_FINAL_Apr2021-deidentified.csv")
target$Sentrix_ID <- sub("\\_.*", "", target$IDAT_FILE_NAME_Grn_Cy3)
target$Basename1 <- paste0(paste(target$Sentrix_ID, target$Sentrix_Position,sep = "_"))
ADM_EPIGEN2_COM_to_basename <- dplyr::select (target, Basename1, ADM_EPIGEN2_COM)
#write.csv (ADM_EPIGEN2_COM_to_basename, "ADM_EPIGEN2_COM_to_basename.csv", row.names = FALSE)

metadata <- read.csv("baseline_1478.csv")
data <- dplyr::select (metadata, ADM_EPIGEN2_COM, BLD_HSCRP_COM)

target <- merge (target, data, by="ADM_EPIGEN2_COM")

#log transforme CRP
CRP_data <- dplyr::select (target,Basename1, BLD_HSCRP_COM)
CRP_data<- na.omit (CRP_data)
#remove values below 0
```

```
CRP_data <- subset (CRP_data, BLD_HSCRP_COM > 0)
CRP_data$BLD_HSCRP_COM <- log(CRP_data$BLD_HSCRP_COM, base=10)

bVals <- merge (CRP_data, bVals, by="Basename1")

colnames (bVals)

df = bVals

nrFolds <- 10
folds <- sample(rep_len(1:nrFolds, nrow(df)),replace=F)

Predict = rep(NULL, nrow(df))

coeffi <- list ()
lambdas <- list ()

for(k in 1:10){
  fold <- which(folds == k)
  data.train <- df[-fold,]
  data.test <- df[fold,]
  x.train <- as.matrix(data.train[,3:865861])
  y.train <- as.matrix(data.train$BLD_HSCRP_COM)
  x.test <- as.matrix(data.test[,3:865861])

  cv <- cv.glmnet(x.train, y.train, alpha = 0.5, nfold=10)
  Predict[fold] = predict(cv, x.test ,type="response", s="lambda.min")
  lambdas[k] = cv[["lambda.min"]]

  myCoefs <- coef(cv, s="lambda.min")
  myCoefs[which(myCoefs != 0 ) ]
  myCoefs@Dimnames[[1]][which(myCoefs != 0 ) ]

  ## Asseble into a data.frame
  myResults <- data.frame(
    features = myCoefs@Dimnames[[1]][ which(myCoefs != 0 ) ], #intercept included
    coefs    = myCoefs              [ which(myCoefs != 0 ) ]  #intercept included
  )
  coeffi[k]=myResults
}

RMSE <- RMSE(Predict, df$BLD_HSCRP_COM, na.rm=T)
Rsquare <- R2(Predict, df$BLD_HSCRP_COM)

RMSE
Rsquare

Predict<- as.data.frame (Predict)

data_1= select (df, Basename1, BLD_HSCRP_COM)
data_1 <- cbind (data_1, Predict)

cor.test (data_1$Predict, data_1$BLD_HSCRP_COM, method="pearson")
```

```r
plot (data_1$Predict, data_1$BLD_HSCRP_COM)

#write.csv (myResults, file= "myResults_CRPLog10_10fold_test_QuantilePP", row.names = TRUE)
#write.csv (data_1, file= "Predictions_CRPLog10_10fold_test_QuantilePP", row.names = TRUE)
#saveRDS(lambdas, file="lambdas_CRPLog10_10fold_test_QuantilePP.RData")
#saveRDS(coeffi, file="Coefficients_CRPLog10_10fold_test_QuantilePP.RData")
#save(cv, file="cv_results_CRPLog10_10fold_test_QuantilePP.RData")

data_1 <- read.csv ("Predictions_CRPLog10_10fold_test_QuantilePP")
myResults <- read.csv ("myResults_CRPLog10_10fold_test_QuantilePP")

ggplot(data_1, aes(x=Predict, y=BLD_HSCRP_COM)) + xlab ("Predicted Log10 CRP") + ylab ("Blood CRP (log-
  geom_point() + geom_smooth(method=lm,  linetype="dashed",  color="darkred", fill="blue")
```

```r
#2-way CRP signature - metabolomics + epigenetics (MRS-ERS)

#Step 1: Save the residuals from a linear regression predicting CRP by
#metabolomics from the 1-way risk score:

baseline_data <- read.csv ("2109001_Harvard_LLiang_CoPv7_Baseline_fixed.csv")
data_1 <- read.csv ("Predictions_CRPLog10_10fold__metabolomics_standardized.csv")

data_1 <- data_1 %>%
  dplyr::rename(Log_CRP= BLD_HSCRP_COM)
data_1 <- data_1 %>%
  dplyr::rename(Predicted_by_metabolomics= Predict)

lm <- lm (Log_CRP ~ Predicted_by_metabolomics, data=data_1)
data_1$Res1 <- residuals(lm)
summary (lm)
plot(data_1$Predicted_by_metabolomics*1.050574, data_1$Res1) +abline(0, 0)

#Step 2i: Use elastic net regression to build a model with the metabolomics
#as a predictor for Res1:

load("RGset_CLSA_1476.RData")
mSetSq <- preprocessQuantile(RGset1476)
bVals <- getBeta(mSetSq)
colnames (bVals)
rownames (bVals)
bVals <- t(bVals)
bVals <- as.data.frame(bVals)
bVals <- tibble::rownames_to_column(bVals, "Basename1")

target <- read.csv("CLSA_metadata_FINAL_Apr2021-deidentified.csv")
target$Sentrix_ID <- sub("\\_.*", "", target$IDAT_FILE_NAME_Grn_Cy3)
target$Basename1 <- paste0(paste(target$Sentrix_ID, target$Sentrix_Position,sep = "_"))

ID1478 <- dplyr::select (target, Basename1,ADM_EPIGEN2_COM)
bVals <- merge (ID1478, bVals, by="Basename1")
colnames (bVals)
bVals = subset(bVals, select = -c(Basename1) )
colnames (bVals)
```

```
df = bVals
epi_data <- dplyr::select (baseline_data,ADM_EPIGEN2_COM, entity_id)
data_1 <- merge (data_1, epi_data, by="entity_id")
res1_data <- dplyr::select (data_1,ADM_EPIGEN2_COM, Res1)
res1_data <- na.omit (res1_data)
df <- merge (res1_data, df, by="ADM_EPIGEN2_COM")
colnames (df)

nrFolds <- 10
folds <- sample(rep_len(1:nrFolds, nrow(df)),replace=F)

Predict = rep(NULL, nrow(df))

coeffi <- list ()
lambdas <- list ()

for(k in 1:10){
  fold <- which(folds == k)
  data.train <- df[-fold,]
  data.test <- df[fold,]
  x.train <- as.matrix(data.train[,3:865861])
  y.train <- as.matrix(data.train$Res1)
  x.test <- as.matrix(data.test[,3:865861])

  cv <- cv.glmnet(x.train, y.train, alpha = 0.5, nfold=10)
  Predict[fold] = predict(cv, x.test ,type="response", s="lambda.min")
  lambdas[k] = cv[["lambda.min"]]

  myCoefs <- coef(cv, s="lambda.min")
  myCoefs[which(myCoefs != 0 ) ]
  myCoefs@Dimnames[[1]][which(myCoefs != 0 ) ]

  ## Asseble into a data.frame
  myResults <- data.frame(
    features = myCoefs@Dimnames[[1]][ which(myCoefs != 0 ) ], #intercept included
    coefs    = myCoefs              [ which(myCoefs != 0 ) ]  #intercept included
  )
  coeffi[k]=myResults
}

RMSE <- RMSE(Predict, df$Res1, na.rm=T)
Rsquare <- R2(Predict, df$Res1)

RMSE
Rsquare

save_this= Predict

Predict<- as.data.frame (Predict)

data_2= dplyr::select (df, ADM_EPIGEN2_COM, Res1)
data_2 <- cbind (data_2, Predict)
```

```
cor.test (data_2$Predict, data_2$Res1, method="pearson")

plot (data_2$Predict, data_2$Res1)

#write.csv (myResults, file= "CRP_Log10_2way_metabolomic_epig_10fold.csv", row.names = TRUE)
#write.csv (data_2, file= "Predictions_CRPLog10_10fold_2way_metabolomics_epig.csv", row.names = TRUE)
#saveRDS(lambdas, lambdas_CRPLog10_10fold_2way_metabolomics_epig_new_data.RData")
#saveRDS(coeffi, file="Coefficients_CRPLog10_10fold_2way_metabolomics_epig.RData")
#save(cv, file="cv_results_CRPLog10_10fold_2way_metabolomics_epig_new_data.RData")

#Step 2ii: The 2 omics risk score (2-way MRS-ERS) is the sum of: beta1
#(from step 1 LR) MRS + EN model by ERS (from step 2.i EN):

data_2 <- read.csv ("Predictions_CRPLog10_10fold_2way_metabolomics_epig.csv")
link <- dplyr::select (baseline_data, entity_id, ADM_EPIGEN2_COM)
data_2 <- merge (data_2, link, by="ADM_EPIGEN2_COM")
data <- merge (data_1, data_2, by="entity_id")
data$metab_beta1 <- data$Predicted_by_metabolomics*1.050574
data$metab_beta1_epigen <- data$metab_beta1 + data$Predict
coeffi <- readRDS ("Coefficients_CRPLog10_10fold_2way_metabolomics_epig.RData")

cor.test (data$Log_CRP, data$metab_beta1_epigen)

ggplot(data, aes(x=metab_beta1_epigen, y=Log_CRP)) + xlab ("Predicted logCRP 2-way MRS ERS") + ylab ("Bl
  geom_point() + geom_smooth(method="lm", linetype="dashed",  color="darkred", fill="blue")
RMSE <- RMSE(data$metab_beta1_epigen, data$Log_CRP, na.rm=T)
```

```
#Step 1: save the residuals from log10CRP ~ PRS. This is Res1.
CRP_PRS <- read.delim("CLSA_NHS_inflammatory_PRS/CLSA__CRP.txt")
CRP_PRS <- CRP_PRS %>%
  dplyr::rename(SCORESUM_CRP= SCORESUM)
baseline_data <- read.csv ("/2109001_Harvard_LLiang_CoPv7_Baseline_fixed.csv")
CRP_PRS <- CRP_PRS %>%
  dplyr::rename(ADM_GWAS3_COM=IID)

baseline_data <- merge (baseline_data, CRP_PRS, by="ADM_GWAS3_COM")
#clean CRP data
baseline_data <- subset (baseline_data, BLD_HSCRP_COM > 0 )
min (baseline_data$BLD_HSCRP_COM) #
max (baseline_data$BLD_HSCRP_COM) #
baseline_data$Log_CRP <- log(baseline_data$BLD_HSCRP_COM, base=10)

lm <- lm (Log_CRP ~ SCORESUM_CRP, data=baseline_data)
baseline_data$Res1 <- residuals(lm)
plot(baseline_data$SCORESUM_CRP, baseline_data$Res1) +abline(0, 0)
summary (lm)
plot(baseline_data$SCORESUM_CRP*0.748406, baseline_data$Res1) +abline(0, 0)

#Step 2: Res 1 ~ 10 fold CV for metabolomics
data_1 <- read.csv ("Predictions_CRPLog10_10fold_2way_PRS_metabolomics_standardized.csv")

#Step 3: Save the residuals from the linear regression (LR) predicting CRP by genetics and #metabolomic
step3_data <- dplyr::select (data_1, entity_id, Predict)
more_data <- dplyr::select (baseline_data, entity_id,ADM_METABOLON_COM, SCORESUM_CRP,Log_CRP, ADM_EPIGE
```

```
step3_data <- merge (more_data, step3_data, by="entity_id")
lm2 <- lm (Log_CRP ~ SCORESUM_CRP + Predict, data=step3_data)
step3_data$Res2 <- residuals(lm2)

#Step 4i: Use elastic net (EN) with epigenetics as the predictor for Res 2
colnames(bVals)
df = bVals
res2_data <- dplyr::select (step3_data,ADM_EPIGEN2_COM, entity_id, Res2)
df <- merge (res2_data, df, by="ADM_EPIGEN2_COM")
colnames (df)

nrFolds <- 10
folds <- sample(rep_len(1:nrFolds, nrow(df)),replace=F)

Predict = rep(NULL, nrow(df))

coeffi <- list ()
lambdas <- list ()

for(k in 1:10){
  fold <- which(folds == k)
  data.train <- df[-fold,]
  data.test <- df[fold,]
  x.train <- as.matrix(data.train[,4:865862])
  y.train <- as.matrix(data.train$Res2)
  x.test <- as.matrix(data.test[,4:865862])

  cv <- cv.glmnet(x.train, y.train, alpha = 0.5, nfold=10)
  Predict[fold] = predict(cv, x.test ,type="response", s="lambda.min")
  lambdas[k] = cv[["lambda.min"]]

  myCoefs <- coef(cv, s="lambda.min")
  myCoefs[which(myCoefs != 0 ) ]
  myCoefs@Dimnames[[1]][which(myCoefs != 0 ) ]

  ## Asseble into a data.frame
  myResults <- data.frame(
    features = myCoefs@Dimnames[[1]][ which(myCoefs != 0 ) ], #intercept included
    coefs    = myCoefs              [ which(myCoefs != 0 ) ]  #intercept included
  )
  coeffi[k]=myResults
}

RMSE <- RMSE(Predict, df$Res2, na.rm=T)
Rsquare <- R2(Predict, df$Res2)

RMSE
Rsquare

save_this= Predict

Predict<- as.data.frame (Predict)
```

```
data_1= dplyr::select (df, ADM_EPIGEN2_COM, Res2)
data_1 <- cbind (data_1, Predict)

cor.test (data_1$Predict, data_1$Res2, method="pearson")

plot (data_1$Predict, data_1$Res2)

#write.csv (myResults, file= "CRP_Log10_3way_PRS_metabolomic_10fold.csv", row.names = TRUE)
#write.csv (data_1, file= "Predictions_CRPLog10_10fold_3way_PRS_metabolomics.csv", row.names = TRUE)
#saveRDS(lambdas, file="lambdas_CRPLog10_10fold_3way.RData")
#saveRDS(coeffi, file="Coefficients_CRPLog10_10fold_3way.RData")
#save(cv, cv_results_CRPLog10_10fold_3way.RData")

#Step 4ii: Calculate 3 omics risk score (3-way PRS-MRS-ERS)
#beta1 (from step 3 LR) PRS + beta2 (from step 3 LR) EN model by MRS
#(from step 2.i EN***) + EN model by ERS (from step 4.i EN)
#*** step 2.i EN model by MRS beta is from PRS-MRS

myResults <- read.csv ("CRP_Log10_3way_PRS_metabolomic_10fold.csv")
data_1 <- read.csv ("Predictions_CRPLog10_10fold_3way_PRS_metabolomics.csv")

log10CRP <- dplyr::select (baseline_data, Log_CRP,ADM_EPIGEN2_COM)
data_1 <- merge (data_1, log10CRP, by="ADM_EPIGEN2_COM")
cor.test (data_1$Log_CRP, data_1$Predict)

data_metab <- read.csv ("Predictions_CRPLog10_10fold_2way_PRS_metabolomics_standardized.csv")

link <- dplyr::select (baseline_data, ADM_EPIGEN2_COM, ADM_GWAS3_COM, entity_id)
CRP_PRS <- merge (CRP_PRS, link, by="ADM_GWAS3_COM")
data_metab <- merge (data_metab, link, by="entity_id")
data_1 <- merge (data_1, link, by="ADM_EPIGEN2_COM")

all <- merge (CRP_PRS, data_metab, by="ADM_EPIGEN2_COM")
all <- merge (all, data_1, by="ADM_EPIGEN2_COM")
all$all_predictors <- all$SCORESUM_CRP + all$Predict.x + all$Predict.y
CRP_to_add <- dplyr::select (baseline_data, Log_CRP, ADM_EPIGEN2_COM)
all <- merge (all, CRP_to_add, by="ADM_EPIGEN2_COM")
cor.test (all$Log_CRP.x, all$all_predictors)
plot (all$all_predictors,all$Log_CRP.x)

all$PRS_beta1 <- all$SCORESUM_CRP*0.537409
all$Metab_beta2 <- all$Predict.x*1.042079
all$all_predictors2 <- all$PRS_beta1 + all$Metab_beta2 + all$Predict.y
cor.test (all$Log_CRP.x, all$all_predictors2)
plot (all$all_predictors2,all$Log_CRP.x)

RMSE <- RMSE(all$all_predictors2, all$Log_CRP.x, na.rm=T)

ggplot(all, aes(x=all_predictors2, y=Log_CRP.x)) + xlab ("Predicted logCRP 3-way PRS-MRS-ERS") + ylab (
  geom_point() + geom_smooth(method="lm", linetype="dashed",  color="darkred", fill="blue")
```