Assignment 9: Using real-world data for hypothesis generation

Fang Liu

03/27/22

Table of Contents

Demonstrate Interaction using Regression Models and Tree-based Methods using Exposome Data from HELIX

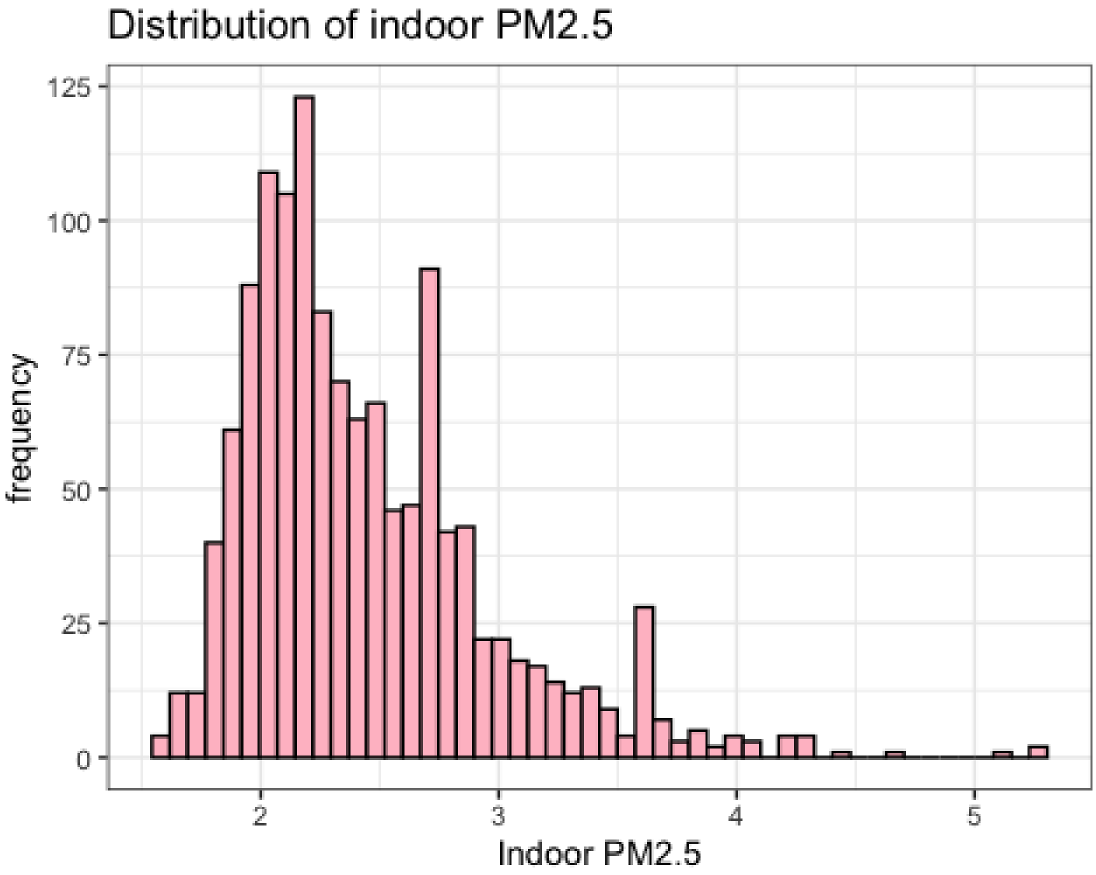
Load .Rdata file and merge into single data frame

Reminder: Merging into a single data frame is optional. Depends upon how you program. This example will assume you've merged everything into a single data frame.

```
library(tidyverse)
library(caret)
library(rpart.plot)
library(pROC)

#Load data using path of where file is stored
load("./exposome.RData")

#Merge all data frames into a single data frame. FYI, this is just a shortcut
by combining baseR with piping from tidyverse. There are other ways of
merging across three data frames that are likely more elegant.
studydata <- merge(exposome,phenotype,by="ID") %>% merge(covariates, by="ID")

#Strip off ID Variable
studydata$ID<-NULL

#factor the outcome variable 'hs_asthma'
studydata$hs_asthma <- factor(studydata$hs_asthma)
str(studydata$hs_asthma)

##  Factor w/ 2 levels "0","1": 1 1 2 1 1 2 1 2 1 2 ...
```

Step 1: Data Exploration of Training Data

```
#exposure 1 - indoor PM2.5 (postnatal;continuous)
summary(studydata$h_PM_Log)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.549   2.069   2.304   2.443   2.699   5.236

studydata %>%
  ggplot() +
  geom_histogram(aes(h_PM_Log), bins = 50, color = "black", fill = "pink") +
  labs(title = "Distribution of indoor PM2.5", x = "Indoor PM2.5", y =
"frequency") +
  theme_bw()
```



```
#exposure 2 - pm10 during pregnancy (continuous)
summary(studydata$h_pm10_ratio_preg_None)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.066  17.535  23.018  23.504  27.677  47.698

#exposure 3 - humidity average during pregnancy (continuous)
summary(studydata$h_humidity_preg_None)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   55.83   70.63   77.10   76.56   86.54   90.67

#exposure 4 - tobacco smoke status of parents  (factor with 3 levels)
str(studydata$hs_smk_parents_None)

##  Factor w/ 3 levels "both","neither",..: 1 2 3 3 2 2 2 2 2 3 ...

summary(studydata$hs_smk_parents_None) #note: total of 1301 mother-child 
pairs

##    both neither     one 
##     142     814     345

#exposure 5 - traffic density on nearest road at home (postnatal; continuous)
summary(studydata$hs_trafnear_h_pow1over3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   8.434  14.841  15.977  22.104  49.348

#correlations
cor(studydata$h_PM_Log,studydata$h_pm10_ratio_preg_None)

## [1] 0.2825068

cor(studydata$h_pm10_ratio_preg_None, studydata$hs_trafnear_h_pow1over3)

## [1] 0.1708348

#outcome of interest: asthma (outcome at 6-11 years old); factor with 2 
levels
str(studydata$hs_asthma)

##  Factor w/ 2 levels "0","1": 1 1 2 1 1 2 1 2 1 2 ...

summary(studydata$hs_asthma) #142 asthma; 1159 without asthma

##    0    1 
## 1159  142
```

Exploratory Analysis: My five choosen exposure of interest is h_PM_Log, h_pm10_ratio_preg_None, h_humidity_preg_None, hs_smk_parents_None, hs_trafnear_h_pow1over3 and my phenotype outcome of interest is hs_asthma. I choose these variables because I want to see if factors like indoor particulate matter and parent smoking status would impact a child's risk of asthma. From my exploratory analysis, I found the following: the mean indoor PM2.5 is 2.443 (range: 1.549 - 5.236), the mean outdoor pm10 value during pregnancy is 23.018 (range: 8.066 - 27.698), and that the average humidity is 77.10. I also found that of the 1301 mother-child pairs, 142 parents both smoke, 345 only one parent smoke, and the rest does not smoke at all. There is a weak and positive relationship between these variables. As for my outcome of interest hs_asthma, only 142 out of the 1301 was diagnosed with asthma at 6-11 years old.

Step 2: Research Question

Put your Research Question in this section. It can be a prediction question OR it can be a hypothesis-generating question about either combinations of features or interactions between features.

Prediction RQ: What is the probability of having a diagnosis of asthma for a child with certain characteristics (i.e., the 5 selected variables from step 1)?

Step 3: Implement pipeline to address research question

You only need to implement a single algorithm to address your research question.Tune hyperparameters to obtain optimal model in training then evaluate in test set.

```
#Data Partition
set.seed(100)
train_indices<-createDataPartition(y=studydata$hs_asthma,p=0.7,list=FALSE)
train_data<-studydata[train_indices, ] #912
test_data<-studydata[-train_indices, ] #389

summary(studydata$hs_asthma)

##    0    1 
## 1159  142

#highly unbalanced: no asthma = 1159, asthma = 142 --> upsampling needed!!
```

Elastic Net

```
set.seed(100)

en_asthma <- train(hs_asthma ~ h_PM_Log + h_pm10_ratio_preg_None + 
h_humidity_preg_None + hs_smk_parents_None + hs_trafnear_h_pow1over3, 
data=train_data, method="glmnet",family="binomial", trControl = 
trainControl("cv", number = 10, sampling= "up"), tuneLength=10)

en_asthma$bestTune

##    alpha     lambda
## 36   0.6 0.02330043

en_asthma$results[36,] #accuracy of 0.573495

##    alpha     lambda Accuracy      Kappa AccuracySD    KappaSD
## 36   0.6 0.02330043 0.573495 0.04131494 0.05953056 0.07138054
```

For fun: Ensemble method (bagging)

```
set.seed(100)

#Note: in bagging, ALL predictor features are eligible for selection at each
```

```
node
mtry_val1 <- expand.grid(.mtry = 5)

bag_asthma<-train(hs_asthma ~ h_PM_Log + h_pm10_ratio_preg_None +
h_humidity_preg_None + hs_smk_parents_None + hs_trafnear_h_pow1over3,
data=train_data, method="rf", metric="Accuracy", trControl =
trainControl("cv", number = 10, sampling= "up"), tuneGrid=mtry_val1,
ntree=100)

bag_asthma$results #accuracy = 0.8563545

##   mtry  Accuracy       Kappa AccuracySD    KappaSD
## 1    5 0.8563545 -0.00471495  0.0150786 0.07413176
```

Model Evaluation for Elastic Net

```
asthma_pred = predict(en_asthma, test_data)
asthma_pred_prob = predict(en_asthma, test_data, type = "prob")

#Confusion Matrix
en_eval = confusionMatrix(asthma_pred, test_data$hs_asthma, positive = "1")
en_eval #accuracy: 0.5398

## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   0   1
##          0 185  17
##          1 162  25
## 
##                Accuracy : 0.5398
##                  95% CI : (0.4889, 0.5902)
##     No Information Rate : 0.892
##     P-Value [Acc > NIR] : 1
## 
##                   Kappa : 0.051
## 
##  Mcnemar's Test P-Value : <2e-16
## 
##             Sensitivity : 0.59524
##             Specificity : 0.53314
##          Pos Pred Value : 0.13369
##          Neg Pred Value : 0.91584
##              Prevalence : 0.10797
##          Detection Rate : 0.06427
##    Detection Prevalence : 0.48072
##       Balanced Accuracy : 0.56419
## 
##        'Positive' Class : 1
## 
```

```
#AUC
auc = roc(response=test_data$hs_asthma, predictor=asthma_pred_prob[,2])

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

auc$auc #0.6118

## Area under the curve: 0.6118

#Variable importance
varImp(en_asthma)

## glmnet variable importance
##
##                              Overall
## hs_smk_parents_Noneone       100.000
## hs_smk_parents_Noneneither    25.058
## h_pm10_ratio_preg_None         5.933
## hs_trafnear_h_pow1over3        3.390
## h_humidity_preg_None           2.223
## h_PM_Log                       0.000
```

I chosen the **elastic net** algorithm to answer my research question. The model accuracy is only 0.573495 for the training data, with the hyperparameter alpha = 0.6. When I evaluate my model in the testing set, the model accuracy is 0.5498. The sensitivity and specificity is also fairly low; the area under the curve is **0.6118**. As for the variable importance, we see that the smoking status of the parent plays the biggest role.