Exposome (ML\_HW9)

Judy Fordjuoh

March 27, 2022

We aimed to evaluate the association between a broad range of prenatal and postnatal lifestyle and environmental exposures and lung function in children.

#### Question 1

### Examine the different features within the data frames and provide some descriptive measures (for continuous measures: means and ranges, for categorical/binary: frequency counts). Feel free to look at correlations between features, examine missingness, and other exploratory data analyses. You don’t need to do this for all features within the data frames (although you could) but you should document summaries for at least 5 features from exposome and 1 feature from phenotype. These don’t need to be formatted into tables or data frames. Just document that you’ve done some data exploration.

#analyzing the covariates data with only the features I decided to analyze  
covariates = covariates %>%  
 mutate(h\_edumc\_None = recode(h\_edumc\_None,  
 "1" = "Primary School",  
 "2" = "Secondary School",  
 "3" = "University Degree or Higher")) %>%  
 mutate(h\_native\_None = recode(h\_native\_None,  
 "0" = "No Native Parent",  
 "1" = "Only One Native Parent",  
 "2" = "Both Parents Native")) %>%  
 select(ID,h\_edumc\_None, h\_native\_None, h\_mbmi\_None, hs\_wgtgain\_None) %>%  
 drop\_na()  
  
#categorical data  
covar = covariates %>%  
 select(h\_edumc\_None, h\_native\_None) %>%  
 group\_by(h\_edumc\_None, h\_native\_None) %>%  
 summarize(   
 count = n()  
 )

## `summarise()` has grouped output by 'h\_edumc\_None'. You can override using the `.groups` argument.

#continuous variables  
covar2 = covariates %>%  
 select(h\_mbmi\_None, hs\_wgtgain\_None) %>%  
 summary(covar2)

As for the level attained in the mothers education, 178 have only completed primary school level, 449 has only completed up to secondary school, and 674 has completed a univiersity degree or higher. In this dataset, 146 of the children had no native parents, 67 had only one native parent, and 1088 had both parents being natives. The average maternal pre-pregnancy BMI was 25.03 with a maximum of 51.42 and a minimum of 15.88. The average pregnancy weight gain was 13.5 with a maximum of 55 and a minimum of 0.0.

#analyzing the exposome data with only the features I decided to analyze  
exposome = exposome %>%  
 mutate(e3\_alcpreg\_yn\_None = recode(e3\_alcpreg\_yn\_None,  
 "0" = "None",  
 "1" = "<1m for KANC")) %>%  
 select(ID, e3\_alcpreg\_yn\_None, h\_folic\_t1\_None, h\_pamod\_t3\_None, h\_greenyn300\_preg\_None, h\_pm10\_ratio\_preg\_None, h\_accesslines300\_preg\_dic0, h\_accesspoints300\_preg\_Log,   
h\_builtdens300\_preg\_Sqrt, h\_connind300\_preg\_Sqrt, h\_walkability\_mean\_preg\_None) %>%  
 drop\_na()  
   
#categorical data  
ex1 = exposome %>%  
 select(e3\_alcpreg\_yn\_None, h\_folic\_t1\_None, h\_pamod\_t3\_None, h\_greenyn300\_preg\_None, h\_accesslines300\_preg\_dic0) %>%  
 group\_by(e3\_alcpreg\_yn\_None, h\_folic\_t1\_None) %>%  
 summarize(   
 count = n()  
 )

## `summarise()` has grouped output by 'e3\_alcpreg\_yn\_None'. You can override using the `.groups` argument.

#continuous variables  
ex2 = exposome %>%  
 select(h\_pm10\_ratio\_preg\_None, h\_accesspoints300\_preg\_Log, h\_builtdens300\_preg\_Sqrt, h\_connind300\_preg\_Sqrt, h\_walkability\_mean\_preg\_None) %>%  
 summary(ex2)

I selected a few variable I found interesting in the exposome dataset in order to reduce the amount of variables. In the ex dataset, there is no missing data. There were 405 cases of mothers who had <1m for KANC for the variable discussing alcohol consumption while pregnant. Although 606 mothers did not take folic acid while pregnant, 695 mothers in this data set did take folic acid while pregnant. 980 mothers reported that they did have access to greenspace at a distance of at most 300m of their home while they were pregnant. While 42 mothers reported doing no walking/cycling activity while pregnant, 474 reported doing this often, 191 reported doing this sometimes, and 594 reported walking/cycling activity very often during their pregnancy. The average walkability index was 0.2674, with a maximum walkability index of 0.6250 and minimum of 0.1000. The average connectivity density was 12.737, with a maximum connectivity density of 27.276 and minimum of 1.887. The average building density was 417.06, with a maximum building density of 807.57 and minimum of 11.02. The average building density was 417.06, with a maximum building density of 807.57 and minimum of 11.02. The average number of bus public transport mode stops inside each 300m buffer was 2.670, with a maximum of 4.58 and minimum of 1.270. The average PM10 value was 23.504, with a maximum PM10 value of 47.698 and minimum of 8.066.

#analyzing the phenotype data with only the features I decided to analyze  
phenotype = phenotype %>%  
 mutate(hs\_asthma = recode(hs\_asthma,  
 "0" = "No Asthma",  
 "1" = "Asthma")) %>%  
 select(ID, e3\_bw, hs\_asthma, hs\_zbmi\_who) %>%  
 drop\_na()  
   
#categorical data  
p1 = phenotype %>%  
 select(hs\_asthma) %>%  
 group\_by(hs\_asthma) %>%  
 summarize(   
 count = n()  
 )   
  
#continuous variables  
p2 = phenotype %>%  
 select(e3\_bw,hs\_zbmi\_who) %>%  
 summary(p2)

In the phenotype data set there was no missing data. There were 142 cases of Asthma. The mean child birth weight in grams was 3389g, with a maximum of 5260g and a minimum of 1100g (range = 4160). The average body mass index z-score of the child at 6-11 years old was 0.4032, with a maximum of 4.72 and minimum of -3.5800 (range = 8.3).

#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  
  
#Partition data for use in demonstration  
set.seed(100)  
train.indices <- createDataPartition(y = studydata$e3\_bw,p = 0.7,list = FALSE)  
train.data <- studydata[train.indices, ]  
test.data <- studydata[-train.indices, ]

#### Question 2

### Develop a single research question. It can be a prediction or hypothesis generating research question. It doesn’t need to involve all features but it must involve at least one outcome from phenotype and 5 features from exposome (and covariates if relevant). This can be as simple or complex as a question as you want it to be. Clearly state the research question in a complete sentence outside of a code chunk, not as a comment within one.

#### Question 3

### Implement one algorithm from class that we’ve used previously to address your question. For this assignment, you should partition your data, develop the model in training (including hyperparameter turning and cross-validation) and then get final evaluation metrics in testing data. You can use package defaults for tuning or create your own vectors/grids. This is to keep getting practice with implementing algorithms so do as much as you would like and find useful.

#I need something where the postiive is having asthma

#### Question 4

### Optional: Adapt the sample code for elastic net within the provided .RMD file to examine two-way interactions using the features and outcome you selected.