Exercise & HW Assignment: Dealing with High Dimensional Data (ML\_HW8)

Judy Fordjuoh

March 20, 2022

Q1 in part 2

# Exercise 1: Feature selection using regularization methods

This exercise is *loosely* based on the following paper: Integration of an interpretable machine learning algorithm to identify early life risk factors of childhood obesity among preterm infants: a prospective birth cohort <https://doi.org/10.1186/s12916-020-01642-6>. The data used in this exercise are an altered version of data available from the HHEAR Data Center, with dois <https://doi.org/10.36043/2017-1740_EPI_58> and <https://doi.org/10.36043/2017-1740_MAIN_84>

In this exercise, you will utilize the caret package to optimize a regularization algorithm for feature selection. You will compare results when you include variables that could induce confounding as features entered into the algorithm. You will also consider how study design and source of data can impact the conclusions drawn by a machine learning analysis.

### Description of the Theoretical Study and Data

The goal of this study is to identify prenatal social and environmental risk factors for childhood overweight/obesity among preterm infants.This study is a prospective birth cohort involving mother-child pairs. Women were enrolled during the first or second trimester of pregnancy and were followed up via visiting clinics until the birth of their children. Women and children were then followed up periodically during infancy and childhood. A total of 1447 singleton children were born preterm, prior to 37 weeks gestation and had complete data on maternal demographics and pregnancy, birth characteristics, lifestyle factors, biospecimen analyzed for exposure to metals and to define childhood obesity at age 5. You have recently been hired as a research data analyst, and tasked with performing the analysis for this study.You are provided with a dataset containing a number of features, in addition to a binary outcome indicating childhood overweight or obesity vs normal weight.

Features in the dataset have informative names. The following categorical features use codes to indicate the different labels:

Child.Human.Biological.Sex: sex assigned at birth of child; 110:Female 111:Male

HHIncome: Household income during pregnancy; 159: <$5,000 204:$5,000-$10,000, 205: $10,000-$20,000, 206: $20,000-$40,000, 207: $40,000-$70,000, 208: >$70,000

Race\_ethnicity: Race or ethnicity of child, as reported by parent; 47:“Hispanic or Latino Ethnicity” 54:“Multiracial” 210:“Black Non-Hispanic” 212:“American Indian Non-Hispanic” 214:“White Non-Hispanic” 217:“Asian Non-Hispanic” 855824:“Other race/ethnicity than white/black/hispanic/asian/american indian/multiracial”

Mother\_Education: Highest Educational Attainment at time of Pregnancy 4:“Advanced Graduate Degree” 12:“College Graduate” 32:“Graduated From High School” 203:“Some College or Technical School” 215:“Less than High School”

Smoking\_Preg: maternal smoking during pregnancy 1: No smoking during pregnancy 2: Active smoker during pregnancy 81: Quit smoking before pregnancy

ow\_obesity: overweight or obesity during childhood 1: Overweight or Obese >= 85th percentile 0: Typical weight <85th percentile (no underweight children in sample)

## BEFORE DATA ANALYSIS

### Question 1: What additional information, if any, would you want from the principal study investigators in regards to the above features?

#### I would love to know why they chose the above features and how they decided to come up with the way they measured or the classifications they created for each feature. I would also like to know if theyh considered measuring other factors such as gestational weight gain, maternal overweight/obesity, access to healthy food during pregnancy, and access to prenatal care, as these are all characteristics that could be related to the outcome of childhood obesity.

### Question 2: Look at the features in the dataset before you start your analysis. Are there any you want to exclude from your analysis completely? Why or why not? Are there any you want to recode or transform? Why or why not?

#### I would recode the household income into quintiles (from lowest income to highest income). I would do this so that approximately 20% of the sample/population is in each group. This makes it easier to compare between groups in this study. Equally important, based on the amount of participants in each quintile, it will provide a better understanding of the socioeconomic status’ represented in the study. Immediately exclude the features that measure specific RBC concentration of elements in the participants pregnancy sample. Some elements like lead may be interesting to analyze because if there are high levels, it can give us some insight on how the participant was living (i.e. if they are living in a poorer neighborhood with buildings with old lead paint etc.). However, overall I do not think including the element features will provide much value in our study of the development of childhood obesity.

### Question 3: Are any of the features not of interest as modifiable contributors to childhood overweight/obesity themselves, but in an explanatory model, you would typically include them? Will you include them in your analysis?

#### Features such as race/ethnicity of child is not an interest as a modifiable contributor to childhood overweight/obesity itself. Race is a social construct and based on my limited knowledge, certain races are not genetically more or less predisposed to childhood obesity in comparison to others. However, I would still include it in the model since we know that certain races and communities of color have been systematically oppressed and the effects of this can influence the rate of childhood obesity.

## STEP ONE: Load Packages and Prepare Data

You will start by loading the needed packages. Some are already listed, but you can choose to use different ones. You will need to clean the data, check that values are plausible, ensure that all variables are the correct type for the algorithm and packages you want to use, etc.

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.5 ✓ purrr 0.3.4  
## ✓ tibble 3.1.5 ✓ dplyr 1.0.7  
## ✓ tidyr 1.1.3 ✓ stringr 1.4.0  
## ✓ readr 2.0.1 ✓ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library(glmnet)

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack

## Loaded glmnet 4.1-3

library(Amelia)

## Loading required package: Rcpp

## ##   
## ## Amelia II: Multiple Imputation  
## ## (Version 1.8.0, built: 2021-05-26)  
## ## Copyright (C) 2005-2022 James Honaker, Gary King and Matthew Blackwell  
## ## Refer to http://gking.harvard.edu/amelia/ for more information  
## ##

library(dplyr)  
library(readxl)  
library(knitr)  
library(gbm)

## Loaded gbm 2.1.8

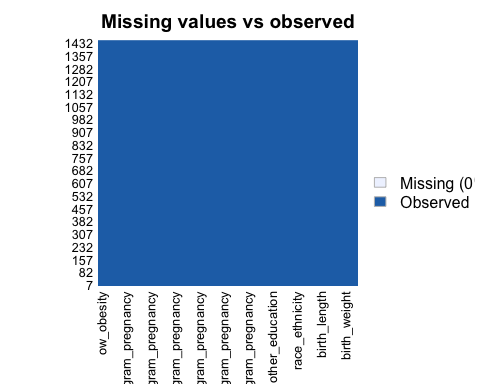
library(e1071)  
library(rpart)  
library(pROC)

## Type 'citation("pROC")' for a citation.

##   
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

## child\_human\_biological\_sex birth\_weight infant\_growth\_zscore  
## Min. :110.0 Min. : 481 Min. :-3.4300   
## 1st Qu.:110.0 1st Qu.:3165 1st Qu.:-0.5800   
## Median :111.0 Median :3515 Median : 0.3400   
## Mean :110.5 Mean :3486 Mean : 0.3375   
## 3rd Qu.:111.0 3rd Qu.:3855 3rd Qu.: 1.1700   
## Max. :111.0 Max. :5528 Max. : 4.1700   
## gestational\_length birth\_length child\_height hh\_income   
## Min. :23.29 Min. :43.10 Min. :55.60 Min. :159.0   
## 1st Qu.:34.57 1st Qu.:48.60 1st Qu.:64.90 1st Qu.:207.0   
## Median :35.43 Median :50.10 Median :66.90 Median :208.0   
## Mean :35.26 Mean :49.87 Mean :66.78 Mean :207.2   
## 3rd Qu.:36.29 3rd Qu.:51.40 3rd Qu.:68.50 3rd Qu.:208.0   
## Max. :37.00 Max. :56.10 Max. :76.00 Max. :208.0   
## race\_ethnicity mother\_age mother\_bmi\_pre\_preg mother\_education  
## Min. : 47 Min. :16.28 Min. :16.19 Min. : 4.0   
## 1st Qu.: 210 1st Qu.:29.61 1st Qu.:21.17 1st Qu.: 4.0   
## Median : 214 Median :32.43 Median :23.54 Median : 12.0   
## Mean : 1374 Mean :32.30 Mean :24.96 Mean : 56.4   
## 3rd Qu.: 214 3rd Qu.:35.49 3rd Qu.:27.16 3rd Qu.: 32.0   
## Max. :855824 Max. :44.90 Max. :50.14 Max. :215.0   
## smoking\_preg   
## Min. : 1.00   
## 1st Qu.: 1.00   
## Median : 1.00   
## Mean :16.82   
## 3rd Qu.: 2.00   
## Max. :81.00   
## pregnancy\_sample\_red\_blood\_cell\_concentration\_aluminum\_nanogram\_per\_gram\_pregnancy  
## Min. :-51.10   
## 1st Qu.: 8.18   
## Median : 17.70   
## Mean : 20.15   
## 3rd Qu.: 27.25   
## Max. :648.00   
## pregnancy\_sample\_red\_blood\_cell\_concentration\_antimony\_nanogram\_per\_gram\_pregnancy  
## Min. :-1.1400   
## 1st Qu.: 0.0404   
## Median : 0.0747   
## Mean : 0.1466   
## 3rd Qu.: 0.1190   
## Max. :56.5000   
## pregnancy\_sample\_red\_blood\_cell\_concentration\_arsenic\_nanogram\_per\_gram\_pregnancy  
## Min. :-0.250   
## 1st Qu.: 0.449   
## Median : 0.849   
## Mean : 1.296   
## 3rd Qu.: 1.550   
## Max. :34.100   
## pregnancy\_sample\_red\_blood\_cell\_concentration\_barium\_nanogram\_per\_gram\_pregnancy  
## Min. :-0.249   
## 1st Qu.: 2.010   
## Median : 3.190   
## Mean : 5.389   
## 3rd Qu.: 5.955   
## Max. :59.200   
## pregnancy\_sample\_red\_blood\_cell\_concentration\_cadmium\_nanogram\_per\_gram\_pregnancy  
## Min. :0.0830   
## 1st Qu.:0.2790   
## Median :0.3950   
## Mean :0.4987   
## 3rd Qu.:0.5645   
## Max. :6.9400   
## pregnancy\_sample\_red\_blood\_cell\_concentration\_cesium\_nanogram\_per\_gram\_pregnancy  
## Min. : 0.455   
## 1st Qu.: 2.010   
## Median : 2.530   
## Mean : 2.672   
## 3rd Qu.: 3.180   
## Max. :10.000   
## pregnancy\_sample\_red\_blood\_cell\_concentration\_chromium\_nanogram\_per\_gram\_pregnancy  
## Min. : 0.0529   
## 1st Qu.: 0.9285   
## Median : 1.3900   
## Mean : 1.8521   
## 3rd Qu.: 2.0500   
## Max. :25.1000   
## pregnancy\_sample\_red\_blood\_cell\_concentration\_cobalt\_nanogram\_per\_gram\_pregnancy  
## Min. :-0.115000   
## 1st Qu.: 0.002135   
## Median : 0.030600   
## Mean : 0.037027   
## 3rd Qu.: 0.060550   
## Max. : 1.130000   
## pregnancy\_sample\_red\_blood\_cell\_concentration\_copper\_nanogram\_per\_gram\_pregnancy  
## Min. : 86.3   
## 1st Qu.: 516.0   
## Median : 564.0   
## Mean : 573.5   
## 3rd Qu.: 617.0   
## Max. :3970.0   
## pregnancy\_sample\_red\_blood\_cell\_concentration\_lead\_nanogram\_per\_gram\_pregnancy  
## Min. : 4.11   
## 1st Qu.:13.50   
## Median :17.70   
## Mean :19.74   
## 3rd Qu.:23.70   
## Max. :90.80   
## pregnancy\_sample\_red\_blood\_cell\_concentration\_magnesium\_nanogram\_per\_gram\_pregnancy  
## Min. : 7070   
## 1st Qu.:37000   
## Median :41100   
## Mean :41860   
## 3rd Qu.:46100   
## Max. :77500   
## pregnancy\_sample\_red\_blood\_cell\_concentration\_manganese\_nanogram\_per\_gram\_pregnancy  
## Min. : 5.36   
## 1st Qu.:13.10   
## Median :16.20   
## Mean :17.30   
## 3rd Qu.:20.30   
## Max. :45.30   
## pregnancy\_sample\_red\_blood\_cell\_concentration\_molybdenum\_nanogram\_per\_gram\_pregnancy  
## Min. :-0.44400   
## 1st Qu.: 0.08475   
## Median : 0.26000   
## Mean : 0.28416   
## 3rd Qu.: 0.42050   
## Max. : 4.74000   
## pregnancy\_sample\_red\_blood\_cell\_concentration\_nickel\_nanogram\_per\_gram\_pregnancy  
## Min. : -0.3180   
## 1st Qu.: 0.4375   
## Median : 0.6630   
## Mean : 0.8084   
## 3rd Qu.: 0.9000   
## Max. :121.0000   
## pregnancy\_sample\_red\_blood\_cell\_concentration\_selenium\_nanogram\_per\_gram\_pregnancy  
## Min. : 39.0   
## 1st Qu.: 220.0   
## Median : 248.0   
## Mean : 255.5   
## 3rd Qu.: 282.0   
## Max. :1090.0   
## pregnancy\_sample\_red\_blood\_cell\_concentration\_thallium\_nanogram\_per\_gram\_pregnancy  
## Min. :-0.213000   
## 1st Qu.:-0.029600   
## Median :-0.014200   
## Mean :-0.027899   
## 3rd Qu.:-0.004895   
## Max. : 0.128000   
## pregnancy\_sample\_red\_blood\_cell\_concentration\_tin\_nanogram\_per\_gram\_pregnancy  
## Min. :-0.176   
## 1st Qu.: 1.120   
## Median : 1.470   
## Mean : 1.529   
## 3rd Qu.: 1.870   
## Max. : 7.490   
## pregnancy\_sample\_red\_blood\_cell\_concentration\_vanadium\_nanogram\_per\_gram\_pregnancy  
## Min. :-0.00762   
## 1st Qu.: 0.06060   
## Median : 0.08540   
## Mean : 0.10274   
## 3rd Qu.: 0.11800   
## Max. : 6.27000   
## pregnancy\_sample\_red\_blood\_cell\_concentration\_zinc\_nanogram\_per\_gram\_pregnancy  
## Min. : 2310   
## 1st Qu.: 9260   
## Median :10400   
## Mean :10473   
## 3rd Qu.:11600   
## Max. :26000   
## ow\_obesity   
## Typical Weight :1105   
## Overweight/Obese: 342   
##   
##   
##   
##



## Typical Weight Overweight/Obese   
## 1105 342

## STEP TWO: Decide on a pipeline

### Question 4: In previous exercises, we often partition our sample into training and testing. We optimize hyperparameters using cross-validation. Is this pipeline still necessary if our goal is feature selection and not building a prediction model to apply to new data? What do you think?

#### It is still necessary to perform cross validation even if our goal is feature selection and not building a prediction model. Cross-validation is a means of estimating the performance of a method for fitting a model, rather than of the model itself, so all steps in fitting the model (including feature selection and optimizing the hyper-parameters) need to be performed independently in each fold of the cross-validation procedure. If you don’t do this, then you will end up with an optimistically biased performance estimate.

Regardless of your answer above, partition the data into a 70/30 split just to get the practice with the programming code to partition.

#tidyverse way to create data partition  
train\_indices <- createDataPartition(y = bc$ow\_obesity,p = 0.7,list = FALSE)  
train\_data <- bc[train\_indices, ]  
test\_data <- bc[-train\_indices, ]

## STEP THREE: Construct a model using a regularization algorithm (lasso, ridge, elastic net) and the features of interest in the training data

### Question 5: Which regularization algorithm seems most appropriate for this research question? Justify your choice.

#### Q5 Ans: I will use LASSO. If we use LASSO, which is L1 regularization, some of the features can shrink down to a zero. This basically works as feature selection. I would not recommend use a ridge regression because with ridge, the features can shrink down but it’ll never be able to shrink it to zero, so I would retain all the features in the dataset.

### Question 6: Which metrics will you use to evaluate your model? Consider your research question and the outcome of interest.

#### Q6 Ans: Since our outcome is binary we will use the confusion matrix/accuracy to evaluate our model. If we were using a continuous outcome/linear regression model, we would use postResample.

Assess how the metric(s) change(s) based on values of the hyperparameters. Construct a grid to explore various values (do not just use the default parameters). Once you have a final model, determine the features that are considered “important” based on the model output.

#NTS: first create a grid to search lambda  
lambda <- 10^seq(-5,5, length = 100)  
  
set.seed(200)  
  
#NTS: replace tuneLength with tuneGrid and alpha is 1 because we are doing lasso. If we were doing rigid it would be 0.   
las <- train(  
 ow\_obesity ~., data = train\_data, method = "glmnet", trControl = trainControl("cv", number = 10, sampling = "up"), preProc = c("center", "scale"), tuneGrid = expand.grid(alpha = 1, lambda = lambda)  
)  
  
#Print the values of alpha and lambda that gave best prediction  
las$bestTune %>% knitr::kable() # 1(alpha)|0.0005(lambda)|0.9734 (Accuracy)

|  |  |  |
| --- | --- | --- |
|  | alpha | lambda |
| 6 | 1 | 3.2e-05 |

#Print all of the options examined  
las$results %>% knitr::kable()  
  
# Model coefficients  
coef(las$finalModel, las$bestTune$lambda)

#Confusion Matrix  
confusionMatrix(las)

## Cross-Validated (10 fold) Confusion Matrix   
##   
## (entries are percentual average cell counts across resamples)  
##   
## Reference  
## Prediction Typical Weight Overweight/Obese  
## Typical Weight 75.0 0.9  
## Overweight/Obese 1.3 22.8  
##   
## Accuracy (average) : 0.9783

#### Based on the training data model, the two features that are important are birth weight and the infant growth Z-score with a beta of 21.03 and 21.32, respectively. Features like child height and mother age went to zero. Some of the pregnancy RBC concentration of elements also went to 0. The accuracy in this model is 0.9734.

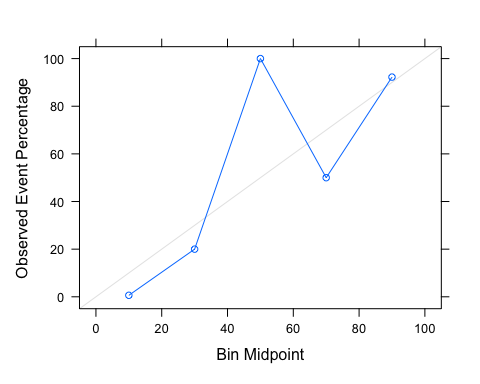
### STEP FOUR: Test your final model in the testing dataset

Use the implementation of your model in the testing set to obtain final performance metrics and perform the inference needed to address your research question.

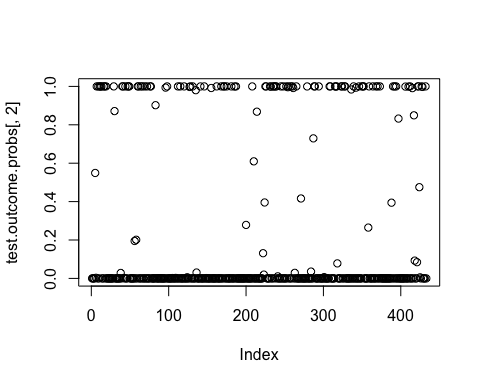
#Using the test data to make predictions  
  
las2 <- las %>% predict(test\_data)  
confusionMatrix(las2,test\_data$ow\_obesity, positive = "Overweight/Obese")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Typical Weight Overweight/Obese  
## Typical Weight 322 5  
## Overweight/Obese 9 97  
##   
## Accuracy : 0.9677   
## 95% CI : (0.9463, 0.9822)  
## No Information Rate : 0.7644   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.9114   
##   
## Mcnemar's Test P-Value : 0.4227   
##   
## Sensitivity : 0.9510   
## Specificity : 0.9728   
## Pos Pred Value : 0.9151   
## Neg Pred Value : 0.9847   
## Prevalence : 0.2356   
## Detection Rate : 0.2240   
## Detection Prevalence : 0.2448   
## Balanced Accuracy : 0.9619   
##   
## 'Positive' Class : Overweight/Obese  
##

#Obtain predicted probabilities  
test.outcome.probs <- predict(las, test\_data, type = "prob")  
  
testProbs.rmodel <- data.frame(obs = test\_data$ow\_obesity,  
 pred.las = test.outcome.probs[,2])  
  
#Create calibration plot  
obesity\_PlotData.rmodel <- calibration(obs ~ pred.las, data = testProbs.rmodel, class = "Overweight/Obese", cuts = 5)  
  
xyplot(obesity\_PlotData.rmodel, auto.key = list(columns = 2))



plot(test.outcome.probs[,2])

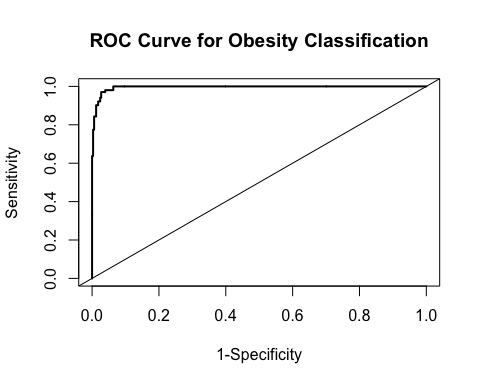


#Create ROC Curve for Analysis  
pred.prob <- predict(las, test\_data, type = "prob")  
  
analysis <- roc(response = test\_data$ow\_obesity, predictor = pred.prob[,2])

## Setting levels: control = Typical Weight, case = Overweight/Obese

## Setting direction: controls < cases

plot(1 - analysis$specificities,analysis$sensitivities,type = "l",  
ylab = "Sensitivity",xlab = "1-Specificity",col = "black",lwd = 2,  
main = "ROC Curve for Obesity Classification")  
abline(a = 0,b = 1)



#Our ROC curve is greater than 0.5 which indicates that we’re classifying the two classes – overweight/obese and typical weight – better than random. And the closer we get to the shape on the right, which we would correspond to an increase in the area under the curve suggests that we are classifying or predicting better, so often you will see these types of curves as part of people’s evaluations for their predictive algorithms.

#### Our ROC curve is greater than 0.5 which indicates that we’re classifying the two classes – overwieght/obese and typical weight – better than random.

### Question 7: Summarize your final conclusion in 2-3 sentences

#### Q7 Ans:The accuracy of the model on the test data was 97.00 % which was similar to the accuracy of the model on the training data. Along witht the 97% accuracy, there was a sensitivity of 0.9902 and a specificity of 0.9637.

### STEP FIVE: Construct another model, making a different choice of variable inclusion.

Redo the above, but now make the opposite choice about variable inclusion. That is, if you did not include the features that themselves might not be modifiable contributors to childhood overweight/obesity, but you would typically include in an explanatory model, include them now. Conversely, if you include those variables previously, exclude them now

Question 7: Do the “important” features change when you make a differnt choice about the other features? Do the hyperparameters that optimize the model change when the additional variables are included? What about model performance?

lambda <- 10^seq(-5,5, length = 100)  
  
set.seed(150)  
  
#Note replacing tuneLength with tuneGrid  
ridge <- train(  
 ow\_obesity ~., data = train\_data, method = "glmnet", trControl = trainControl("cv", number = 10, sampling = "up"), preProc = c("center", "scale"), tuneGrid = expand.grid(alpha = 0, lambda = lambda)  
)  
  
#Print the values of alpha and lambda that gave best prediction  
ridge$bestTune %>% knitr::kable() # 0(alpha)|0.022(lambda)|0.8846(Accuracy)

|  |  |  |
| --- | --- | --- |
|  | alpha | lambda |
| 33 | 0 | 0.0170735 |

#Print all of the options examined  
ridge$results %>% knitr::kable()  
  
# Model coefficients  
coef(ridge$finalModel, ridge$bestTune$lambda)

#Confusion Matrix  
confusionMatrix(ridge)

## Cross-Validated (10 fold) Confusion Matrix   
##   
## (entries are percentual average cell counts across resamples)  
##   
## Reference  
## Prediction Typical Weight Overweight/Obese  
## Typical Weight 67.5 2.1  
## Overweight/Obese 8.9 21.6  
##   
## Accuracy (average) : 0.8905

#### Since this is a ridge model, all of the features have remained in the model. The two main features (birth weight and the infant growth Z-score) remained the same but the betas decreased significantly to 1.58 and 1.92, respectively. The accuracy of the ridge model also reduced to 0.8846.

# Exercise 2: Creating more refined phenotypes for an explanatory analysis.

This exercise is *loosely* based on the following paper: Deploying unsupervised clustering analysis to derive clinical phenotypes and risk factors associated with mortality risk in 2022 critically ill patients with COVID-19 in Spain <doi:10.1186/s13054-021-03487-8>. Data were simulated and are not true COVID data.

Researchers are interested in understanding the factors associated with ICU mortality among COVID-19 patients. They hypothesize there are different clinical phenotypes that could be at different risks for mortality and require different medical interventions. The goal of this research is to determine if patient features including demographics and clinical data at ICU admission could be used to separate COVID-19 patients into distinct phenotypic clusters. The secondary aim was to determine if identified phenotypic clusters had different risk of mortality.

You are provided with the following dataset for a subset of 178 COVID-19 patients, with the instructions to conduct an unsupervised analysis to identify phenotypic clusters within the patient population, describe the clusters in terms of the input features and then determine if there are differences in mortality rate across the clusters. All feature data in the dataset have been centered and scaled. The outcome, mortality, is a binary indicator.

Variables in the dataset (covid.csv) are:

Ageyr: the age of the patient at admission to the ICU APACHE:APACHE II Score, Acute Physiology and Chronic Health Evaluation II Score SOFA: SOFA Score Sequential Organ Failure Assessment DDimer: D dimer, a fibrin degradation product, can indicate blood clot formation and breakdown SerumLactate: measures level of lactic acid in the blood, can be indicator of hypoxia Ferritin: measure of iron sufficiency in blood CRP: C-reactive protein, measure of inflammation Creatinine: product of protein metabolism, high levels can indicate impaired kidney function WBC: concentration of white blood cells, marker of infection DBP: diastolic blood pressure Procalcitonin: marker of bacterial infection IGA: immunoglobulin A, measure of antibodies found in mucous membranes Oxmetric.1: measure of oxygen saturation mortality: 1=Died in ICU, 0=Survived

### STEP 1: Load needed libraries and implement the unsupervised analysis

library(stats)  
library(cluster)  
library(factoextra)

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

covid = read.csv("/Users/judyfordjuoh/Desktop/Machine Learning/covid.csv") %>%  
 janitor::clean\_names()

### Question 1: Name one potential risk/concern of obtaining data that has already been centered and scaled.

#### Q1 Answer: The issue is what represents a good measure of distance between cases.

Centering and scaling the data is a process by which you transform each feature such that its mean becomes 0, and variance becomes 1. If you center and scale your data before splitting it into training and test sets, then you have used information from your training set to make calculations on your test set. The issue with having data that is already centered and scaled is the measure of distance between cases. If each feature is different/has different units (ex. ft and lb), these features are not comparable. So if we clustered individuals their height in feet and clustered individuals by their weight in lbs, is a 1ft difference as significant as a 1lb difference in weight?

### Question 2: What unsupervised analysis do you think is appropriate for this research question? Justify your answer.

#### Q2 Answer: Hieracrhical clustering is an appropriate unsupervised analysis approach because we are trying to identify clusters based on similarity across features. We would not do a principal components analysis because it is a form of dimensionality reduction, which is based on reducing the number of variables or features under consideration. Dimensionality reduction can either be feature selection or feature extraction, not for identifiying clusters. K-means is an inappropriate approach because our outcome is binary. Since K-means needs to compute means, if we have binary values, discrete attributes or categorial attributes, we tend to stay away from k-means since the mean value is not meaningful on this kind of data.

Implement the unsupervised analysis you chose in Question 2, implementing appropriate analyses to determine the number of phenotypic clusters to keep within the analysis.

#Determining if scaling is necessary. Since murder is really low, I'll scale the data  
colMeans(covid, na.rm = TRUE) #the other variables are e-16 while mortality is e-01 so I think I don't have to scale #apply(covid, 2, sd, na.rm=TRUE)

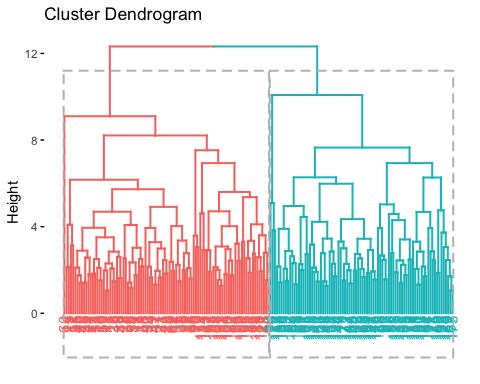
## apache sofa d\_dimer serum\_lactate ferritin   
## -2.810252e-16 3.442939e-16 -2.836175e-16 6.928953e-16 2.583130e-16   
## crp creatinine wbc dbp procalcitonin   
## 1.949712e-16 1.304941e-17 -5.551895e-16 -3.983432e-16 -3.229314e-16   
## iga ageyr oxmetric\_1 mortality   
## -6.786862e-17 -7.906831e-16 2.583130e-16 1.741573e-01

clusters\_hcut <- hcut(covid, k = 2, hc\_func = "hclust", hc\_method = "complete", hc\_metric = "euclidian")  
  
clusters\_hcut$size

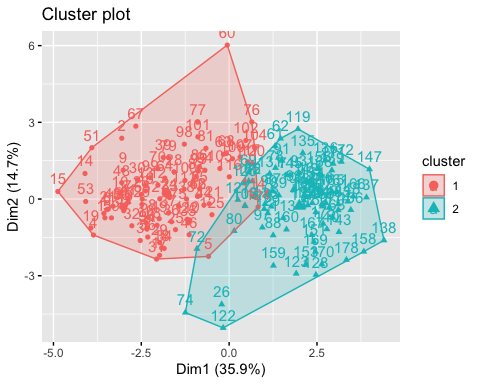
## [1] 94 84

fviz\_dend(clusters\_hcut, rect = TRUE)

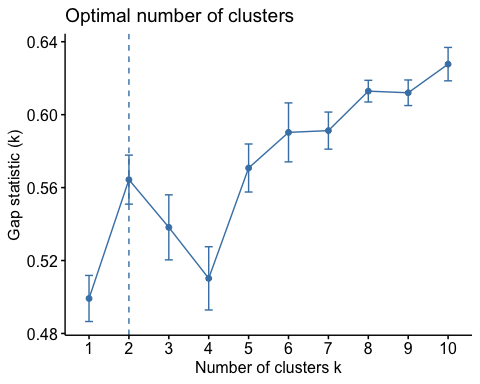
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =  
## "none")` instead.



fviz\_cluster(clusters\_hcut)



gap\_stat <- clusGap(covid, FUN = hcut, hc\_method = "complete", K.max = 10, B = 5)  
fviz\_gap\_stat(gap\_stat)



input.feature.vals <- cbind(covid,cluster = clusters\_hcut$cluster)  
  
input.feature.vals %>%  
 group\_by(cluster) %>%  
 summarise\_all(mean) %>% knitr::kable()

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| cluster | apache | sofa | d\_dimer | serum\_lactate | ferritin | crp | creatinine | wbc | dbp | procalcitonin | iga | ageyr | oxmetric\_1 | mortality |
| 1 | 0.5154620 | -0.3715702 | 0.4715748 | -0.2273347 | -0.5983223 | 0.1350353 | 0.5007861 | 0.618920 | 0.6341271 | -0.5739372 | 0.6624438 | 0.1968494 | -0.5983223 | 0.2127660 |
| 2 | -0.5768265 | 0.4158047 | -0.5277147 | 0.2543984 | 0.6695511 | -0.1511110 | -0.5604035 | -0.692601 | -0.7096184 | 0.6422630 | -0.7413062 | -0.2202838 | 0.6695511 | 0.1309524 |

### STEP 2: Interpret the clusters

Describe the clusters in terms of both their input features and the incidence of mortality within the cluster.

#### I used Euclidean as my distance metric and complete linkage so that I can use the distance between all of the different data points. After visualizing the gapstat, the optimal number of clusters was 2.

#### In the first cluster there were 94 participants and the second cluster had 84 participants. In cluster one, the average mortality was 0.2127660 while in cluster two it was 0.130954. Overall in cluster two, the average of all the features were all larger than the averages of the same features in cluster one. For instance, the measure of oxygen saturation was 0.70 in cluster two but in cluster one it was 0.6. We see that all the features averages are larger in cluster 2 except for mortality, which was larger in cluster one.

### Question 3: A researcher at a different medical institution has heard about your analysis and is interested in using your results to determine risk of mortality within their ICU. What are some limitations or concerns of using the results from your unsupervised analysis in a different setting?

#### The clusters formed in the data in my analysis may not align with the clusters that would form in a different setting. Therefore, the assumptions or inferences the researcher would make about risk of mortality within their ICU based on my results, may potentially be biased because the individuals in their ICU may have different, since their feature data may differ.