Final Project

G5

11/18/2020

## OVERVIEW

Patient self-management, besides expert care, is necessary to improve asthma signs and symptoms and reduce related morbidity and mortality. Asthma self-management education is one of the four key components of asthma control and management. It provides people with asthma and their caregivers with the knowledge and skills to understand the disease and its treatment. It teaches them to take medications appropriately, recognize early signs and symptoms of asthma episodes, seek medical care as appropriate, and identify and avoid environmental asthma allergens and irritants

## EXPLORATORY DATA ANALYSIS

From the questionnaires ask to each participant, we collect the response variables below that explain how each participant manages it asthma.

ASTHNOW Have you ever been told by a doctor or other health professional that you have asthma?

TCH\_SIGN Has a doctor or other health professional ever taught you… a. How to recognize early signs or symptoms of an asthma episode?

TCH\_RESP Has a doctor or other health professional ever taught you… b. What to do during an asthma episode or attack?

TCH\_MON A peak flow meter is a hand held device that measures how quickly you can blow air out of your lungs. Has a doctor or other health professional ever taught you… c. How to use a peak flow meter to adjust your daily medications?

MGT\_PLAN An asthma action plan, or asthma management plan, is a form with instructions about when to change the amount or type of medicine, when to call the doctorfor advice, and when to go to the emergency room. Has a doctor or other health professional EVER given you an asthma action plan?

MOD\_ENV (7.13) INTERVIEWER READ: Now, back to questions specifically about you. Has a health professional ever advised you to change things in your home, school, or work to improve your asthma

MGT\_CLAS Have you ever taken a course or class on how to manage your asthma?

INHALERH (8.3) Did a doctor or other health professional show you how to use the inhaler?

INHALERW (8.4) Did a doctor or other health professional watch you use the inhaler?

Responses types (1) YES (2) NO (7) DON’T KNOW (9) REFUSED

Predictors

MISS\_DAY = “NUMBER OF MISSED DAYS” MOD\_

ENV = “EVER ADVISED CHANGE THINGS IN YOUR HOME”

AGEDX = “AGE AT ASTHMA DIAGNOSIS”

AGEG\_F6\_M = “MODIFIED SIX AGE GROUPS USED IN ASTHMA ADULT POST-STRATIFICATION”

AIRCLEANER = “AIR CLEANER USED”

ASMDCOST = “COST BARRIER: PRIMARY CARE DOCTOR”

ASRXCOST = “COST BARRIER: MEDICATION”

ASSPCOST = “COST BARRIER: SPECIALIST”

CATTMPTS\_F = “DISPOSITION CODES FOR CALL ATTEMPTS 1 THROUGH 20 …”

EMP\_STAT = “CURRENT EMPLOYMENT STATUS”

EPIS\_12M = “ASTHMA EPISODE OR ATTACK”

EPIS\_TP = “NUMBER OF EPISODES / ATTACKS”

ER\_TIMES = “NUMBER OF EMERGENCY ROOM VISITS”

ER\_VISIT = “EMERGENCY ROOM VISIT”

EVER\_ASTH = “EVER HAVE ASTHMA INCONSISTENT WITH BRFSS”

AGEDX, INCIDNT INCONSISTENCY

HOSPPLAN = “HOSPITAL FOLLOW-UP”

HOSPTIME = “NUMBER OF HOSPITAL VISITS”

HOSP\_VST = “HOSPITAL VISIT”

UNEMP\_R = “REASON NOT NOW EMPLOYED”

URG\_TIME = “NUMBER OF URGENT VISITS”

WORKENV5 = “ASTHMA AGGRAVATED BY CURRENT JOB”

WORKENV6 = “ASTHMA CAUSED BY CURRENT JOB”

WORKENV7 = “ASTHMA AGGRAVATED BY PREVIOUS JOB”

WORKENV8 = “ASTHMA CAUSED BY PREVIOUS JOB”

WORKQUIT1 = “EVER CHANGE OR QUIT A JOB”

WORKSEN3 = “DOCTOR DIAGNOSED WORK ASTHMA”

WORKSEN4 = “SELF-IDENTIFIED WORK ASTHMA”

WORKTALK = “DOCTOR DISCUSSED WORK ASTHMA”

INS1 = “INSURANCE”

INS2 = “INSURANCE OR COVERAGE GAP”

LANDWT\_F = “FINAL ADULT ASTHMA CALL-BACK WEIGHT”

LASTSYMP = “LAST HAD ANY SYMPTOMS OF ASTHMA”

LAST\_MD = “LAST TALKED TO A DOCTOR” LAST\_MED = “

LAST TOOK ASTHMA MEDICATION”

COMPASTH =”TYPICAL ATTACK" REGION FOR ASTHMA CALL-BACK. COMPASTH =”TYPICAL ATTACK"

### Constructing the data frame by selecting predictors

## [1] 11494 45

The initial data has 11,494 with 45 varibles.

We are going to clean and restructure this dataset.

Structure of the data frame.

## 'data.frame': 5880 obs. of 45 variables:  
## $ TCH.SIGN : num 2 1 2 1 1 2 1 1 1 1 ...  
## ..- attr(\*, "label")= chr "EVER TAUGHT RECOGNIZE EARLY SIGN OR SYMPTOMS"  
## ..- attr(\*, "format.sas")= chr "TCH\_SIGN"  
## $ TCH.RESP : num 1 1 1 1 1 2 1 1 1 1 ...  
## ..- attr(\*, "label")= chr "EVER TAUGHT WHAT TO DO DURING ASTHMA EPISODE OR ATTACK"  
## ..- attr(\*, "format.sas")= chr "TCH\_RESP"  
## $ TCH.MON : num 2 2 2 1 2 1 1 1 1 2 ...  
## ..- attr(\*, "label")= chr "EVER TAUGHT HOW TO USE A PEAK FLOW"  
## ..- attr(\*, "format.sas")= chr "TCH\_MON"  
## $ MGT.PLAN : num 2 2 2 1 2 2 1 2 2 1 ...  
## ..- attr(\*, "label")= chr "EVER GIVEN AN ASTHMA ACTION PLAN"  
## ..- attr(\*, "format.sas")= chr "MGT\_PLAN"  
## $ MGT.CLAS : num 2 2 2 2 2 2 2 2 2 1 ...  
## ..- attr(\*, "label")= chr "EVER TAKEN A COURSE TO MANAGE ASTHMA"  
## ..- attr(\*, "format.sas")= chr "MGT\_CLAS"  
## $ INHALERW : num 2 1 1 1 1 2 1 1 1 1 ...  
## ..- attr(\*, "label")= chr "INHALER USE WATCHED"  
## ..- attr(\*, "format.sas")= chr "INHALERW"  
## $ MOD.ENV : num 2 2 1 2 1 2 2 2 2 1 ...  
## ..- attr(\*, "label")= chr "EVER ADVISED CHANGE THINGS IN YOUR HOME"  
## ..- attr(\*, "format.sas")= chr "MOD\_ENV"  
## $ SEX : num 2 2 2 1 2 2 2 1 2 1 ...  
## ..- attr(\*, "label")= chr "RESPONDENTS SEX"  
## ..- attr(\*, "format.sas")= chr "SEX"  
## $ AGEG.F7 : num 5 5 6 4 6 5 6 6 4 5 ...  
## ..- attr(\*, "label")= chr "AGE COLLAPSED TO 7 GROUPS FOR ASTHMA CALL-BACK"  
## ..- attr(\*, "format.sas")= chr "AGEG\_F7Z"  
## $ X\_RACEGR3: num 1 1 1 1 1 1 3 1 1 3 ...  
## ..- attr(\*, "label")= chr "COMPUTED FIVE LEVEL RACE/ETHNICITY CATEGORY."  
## ..- attr(\*, "format.sas")= chr "\_3RACEGR"  
## $ X\_EDUCAG : num 2 2 4 4 4 3 1 3 2 3 ...  
## ..- attr(\*, "label")= chr "COMPUTED LEVEL OF EDUCATION COMPLETED CATEGORIES"  
## ..- attr(\*, "format.sas")= chr "\_EDUCAG"  
## $ EDUCAL : num 4 4 6 6 6 5 3 5 4 5 ...  
## ..- attr(\*, "label")= chr "EDUCATION LEVEL"  
## ..- attr(\*, "format.sas")= chr "EDUCA"  
## $ INCOMEL : num 2 2 8 8 5 99 1 77 3 2 ...  
## ..- attr(\*, "label")= chr "INCOME LEVEL"  
## ..- attr(\*, "format.sas")= chr "IN2COME"  
## $ X\_INCOMG : num 1 1 5 5 3 9 1 9 2 1 ...  
## ..- attr(\*, "label")= chr "COMPUTED INCOME CATEGORIES"  
## ..- attr(\*, "format.sas")= chr "\_INCOMG"  
## $ X\_BMI5CAT: num 4 3 4 2 4 NA NA 2 4 3 ...  
## ..- attr(\*, "label")= chr "COMPUTED BODY MASS INDEX CATEGORIES"  
## ..- attr(\*, "format.sas")= chr "\_BMI5CAT"  
## $ X\_RFBMI5 : num 2 2 2 1 2 9 9 1 2 2 ...  
## ..- attr(\*, "label")= chr "OVERWEIGHT OR OBESE CALCULATED VARIABLE"  
## ..- attr(\*, "format.sas")= chr "\_5RFBMI"  
## $ SMOKE100 : num 1 1 1 1 2 2 2 2 1 1 ...  
## ..- attr(\*, "label")= chr "SMOKED AT LEAST 100 CIGARETTES"  
## ..- attr(\*, "format.sas")= chr "SMOK100\_"  
## $ COPD : num 1 2 2 2 2 2 2 2 1 2 ...  
## ..- attr(\*, "label")= chr "EVER TOLD HAVE CHRONIC OBSTRUCTIVE PULMONARY DISEASE"  
## ..- attr(\*, "format.sas")= chr "COPD"  
## $ EMPHY : num 2 2 2 2 2 2 2 2 1 2 ...  
## ..- attr(\*, "label")= chr "EVER TOLD HAVE EMPHYSEMA"  
## ..- attr(\*, "format.sas")= chr "EMPHY"  
## $ DEPRESS : num 1 2 2 2 1 1 1 2 1 1 ...  
## ..- attr(\*, "label")= chr "EVER TOLD DEPRESSED"  
## ..- attr(\*, "format.sas")= chr "DEPRESS"  
## $ BRONCH : num 1 2 1 2 1 2 2 2 1 2 ...  
## ..- attr(\*, "label")= chr "EVER TOLD HAVE CHRONIC BRONCHITIS"  
## ..- attr(\*, "format.sas")= chr "BRONCH"  
## $ SYMP.30D : num 30 7 17 10 30 5 30 7 14 30 ...  
## ..- attr(\*, "label")= chr "SYMPTOM DAYS"  
## ..- attr(\*, "format.sas")= chr "SYMP\_30D"  
## $ DUR.30D : num 2 12 12 12 1 12 1 12 12 1 ...  
## ..- attr(\*, "label")= chr "CONSTANT SYMPTOMS"  
## ..- attr(\*, "format.sas")= chr "DUR\_30D"  
## $ ASLEEP30 : num 15 6 88 88 20 88 4 88 7 30 ...  
## ..- attr(\*, "label")= chr "NIGHT SYMPTOMS"  
## ..- attr(\*, "format.sas")= chr "ASLEEP30Z"  
## $ SYMPFREE : num 88 6 11 7 88 5 3 7 88 88 ...  
## ..- attr(\*, "label")= chr "SYMPTOM-FREE DAYS"  
## ..- attr(\*, "format.sas")= chr "SYMPFREE"  
## $ INCINDT : num 2 3 3 3 3 3 3 3 3 2 ...  
## ..- attr(\*, "label")= chr "TIME SINCE DIAGNOSIS"  
## ..- attr(\*, "format.sas")= chr "INCIDNT"  
## $ LAST.MD : num 4 4 4 4 4 7 6 5 7 4 ...  
## ..- attr(\*, "label")= chr "LAST TALKED TO A DOCTOR"  
## ..- attr(\*, "format.sas")= chr "LAST\_MD"  
## $ LAST.MED : num 1 3 3 1 1 5 3 1 7 1 ...  
## ..- attr(\*, "label")= chr "LAST TOOK ASTHMA MEDICATION"  
## ..- attr(\*, "format.sas")= chr "LAST\_MED"  
## $ LAST.SYMP: num 1 3 3 3 1 3 3 3 3 1 ...  
## ..- attr(\*, "label")= chr "LAST HAD ANY SYMPTOMS OF ASTHMA"  
## ..- attr(\*, "format.sas")= chr "LASTSYMP"  
## $ EPIS.12M : num 1 1 1 1 2 2 2 2 2 1 ...  
## ..- attr(\*, "label")= chr "ASTHMA EPISODE OR ATTACK"  
## ..- attr(\*, "format.sas")= chr "EPIS\_12M"  
## $ EPIS.TP : num 2 2 38 2 1111 ...  
## ..- attr(\*, "label")= chr "NUMBER OF EPISODES / ATTACKS"  
## ..- attr(\*, "format.sas")= chr "EPIS\_TP"  
## $ DUR.ASTH : num 204 120 115 201 1111 ...  
## ..- attr(\*, "label")= chr "DURATION OF ATTACK"  
## ..- attr(\*, "format.sas")= chr "DUR\_ASTH"  
## $ COMPASTH : num 3 1 3 3 11 11 11 11 11 3 ...  
## ..- attr(\*, "label")= chr "TYPICAL ATTACK"  
## ..- attr(\*, "format.sas")= chr "COMPASTH"  
## $ INS1 : num 1 1 1 1 1 1 1 2 1 1 ...  
## ..- attr(\*, "label")= chr "INSURANCE"  
## ..- attr(\*, "format.sas")= chr "INS1Z"  
## $ INS2 : num 2 2 2 2 2 2 2 5 2 2 ...  
## ..- attr(\*, "label")= chr "INSURANCE OR COVERAGE GAP"  
## ..- attr(\*, "format.sas")= chr "INS2Z"  
## $ ER.TIMES : num 677 677 677 677 677 666 666 666 666 677 ...  
## ..- attr(\*, "label")= chr "NUMBER OF EMERGENCY ROOM VISITS"  
## ..- attr(\*, "format.sas")= chr "ER\_TIMES"  
## $ ER.VISIT : num 2 2 2 2 2 6 6 6 6 2 ...  
## ..- attr(\*, "label")= chr "EMERGENCY ROOM VISIT"  
## ..- attr(\*, "format.sas")= chr "ER\_VISIT"  
## $ URG.TIMES: num 888 3 888 888 888 666 666 666 666 888 ...  
## ..- attr(\*, "label")= chr "NUMBER OF URGENT VISITS"  
## ..- attr(\*, "format.sas")= chr "URG\_TIME"  
## $ HOSP.VST : num 2 2 2 2 2 6 6 6 6 2 ...  
## ..- attr(\*, "label")= chr "HOSPITAL VISIT"  
## ..- attr(\*, "format.sas")= chr "HOSP\_VST"  
## $ HOSPTIME : num 455 455 455 455 455 666 666 666 666 455 ...  
## ..- attr(\*, "label")= chr "NUMBER OF HOSPITAL VISITS"  
## ..- attr(\*, "format.sas")= chr "HOSPTIME"  
## $ HOSPPLAN : num 45 45 45 45 45 6 6 6 6 45 ...  
## ..- attr(\*, "label")= chr "HOSPITAL FOLLOW-UP"  
## ..- attr(\*, "format.sas")= chr "HOSPPLAN"  
## $ ASMDCOST : num 2 2 2 2 2 2 2 2 2 1 ...  
## ..- attr(\*, "label")= chr "COST BARRIER: PRIMARY CARE DOCTOR"  
## ..- attr(\*, "format.sas")= chr "ASMDCOST"  
## $ ASRXCOST : num 2 2 2 2 1 1 2 1 2 2 ...  
## ..- attr(\*, "label")= chr "COST BARRIER: MEDICATION"  
## ..- attr(\*, "format.sas")= chr "ASRXCOST"  
## $ ASSPCOST : num 2 2 2 2 2 2 2 2 2 2 ...  
## ..- attr(\*, "label")= chr "COST BARRIER: SPECIALIST"  
## ..- attr(\*, "format.sas")= chr "ASSPCOST"  
## $ WORKTALK : num 2 2 2 2 2 2 2 2 2 2 ...  
## ..- attr(\*, "label")= chr "DOCTOR DISCUSSED WORK ASTHMA"  
## ..- attr(\*, "format.sas")= chr "WORKTALK"

Summary of the dataset. The summary show the minimun, maximun, mean, median, and quantiles of each variable. But the variables are categorical because it represent the tpye of response to a question. It is better to factor each variable so that we can have the count of each type of response or class.

## TCH.SIGN TCH.RESP TCH.MON MGT.PLAN   
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:1.000   
## Median :1.000 Median :1.000 Median :2.000 Median :2.000   
## Mean :1.329 Mean :1.229 Mean :1.524 Mean :1.669   
## 3rd Qu.:2.000 3rd Qu.:1.000 3rd Qu.:2.000 3rd Qu.:2.000   
## Max. :2.000 Max. :2.000 Max. :2.000 Max. :2.000   
##   
## MGT.CLAS INHALERW MOD.ENV SEX   
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:2.000 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:1.000   
## Median :2.000 Median :1.000 Median :2.000 Median :2.000   
## Mean :1.891 Mean :1.233 Mean :1.622 Mean :1.703   
## 3rd Qu.:2.000 3rd Qu.:1.000 3rd Qu.:2.000 3rd Qu.:2.000   
## Max. :2.000 Max. :2.000 Max. :2.000 Max. :2.000   
##   
## AGEG.F7 X\_RACEGR3 X\_EDUCAG EDUCAL   
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:4.000 1st Qu.:1.000 1st Qu.:2.000 1st Qu.:4.000   
## Median :5.000 Median :1.000 Median :3.000 Median :5.000   
## Mean :4.712 Mean :1.585 Mean :2.957 Mean :4.928   
## 3rd Qu.:6.000 3rd Qu.:1.000 3rd Qu.:4.000 3rd Qu.:6.000   
## Max. :7.000 Max. :9.000 Max. :9.000 Max. :9.000   
##   
## INCOMEL X\_INCOMG X\_BMI5CAT X\_RFBMI5   
## Min. : 1.00 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.: 4.00 1st Qu.:2.000 1st Qu.:3.000 1st Qu.:2.000   
## Median : 6.00 Median :4.000 Median :3.000 Median :2.000   
## Mean :12.97 Mean :3.889 Mean :3.203 Mean :2.073   
## 3rd Qu.: 8.00 3rd Qu.:5.000 3rd Qu.:4.000 3rd Qu.:2.000   
## Max. :99.00 Max. :9.000 Max. :4.000 Max. :9.000   
## NA's :259   
## SMOKE100 COPD EMPHY DEPRESS   
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:2.000 1st Qu.:1.000   
## Median :1.000 Median :2.000 Median :2.000 Median :2.000   
## Mean :1.512 Mean :1.788 Mean :1.924 Mean :1.582   
## 3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:2.000   
## Max. :9.000 Max. :9.000 Max. :9.000 Max. :9.000   
## NA's :14 NA's :14 NA's :14   
## BRONCH SYMP.30D DUR.30D ASLEEP30   
## Min. :1.000 Min. : 1.00 Min. : 1.000 Min. : 1.00   
## 1st Qu.:1.000 1st Qu.: 4.00 1st Qu.: 2.000 1st Qu.:10.00   
## Median :2.000 Median :14.00 Median :12.000 Median :88.00   
## Mean :1.707 Mean :15.46 Mean : 8.683 Mean :53.83   
## 3rd Qu.:2.000 3rd Qu.:30.00 3rd Qu.:12.000 3rd Qu.:88.00   
## Max. :9.000 Max. :30.00 Max. :12.000 Max. :88.00   
## NA's :14   
## SYMPFREE INCINDT LAST.MD LAST.MED   
## Min. : 1.00 Min. :1.000 Min. : 4.000 Min. :1.000   
## 1st Qu.: 7.00 1st Qu.:3.000 1st Qu.: 4.000 1st Qu.:1.000   
## Median :13.00 Median :3.000 Median : 4.000 Median :1.000   
## Mean :38.92 Mean :2.855 Mean : 4.637 Mean :2.032   
## 3rd Qu.:88.00 3rd Qu.:3.000 3rd Qu.: 4.000 3rd Qu.:3.000   
## Max. :99.00 Max. :3.000 Max. :88.000 Max. :7.000   
##   
## LAST.SYMP EPIS.12M EPIS.TP DUR.ASTH   
## Min. :1.00 Min. :1.000 Min. : 1.0 Min. : 101.0   
## 1st Qu.:1.00 1st Qu.:1.000 1st Qu.: 3.0 1st Qu.: 130.0   
## Median :2.00 Median :1.000 Median : 90.0 Median : 303.0   
## Mean :1.78 Mean :1.393 Mean : 527.4 Mean : 567.4   
## 3rd Qu.:3.00 3rd Qu.:2.000 3rd Qu.:1111.0 3rd Qu.:1111.0   
## Max. :3.00 Max. :2.000 Max. :1111.0 Max. :1111.0   
##   
## COMPASTH INS1 INS2 ER.TIMES   
## Min. : 1.000 Min. :1.000 Min. :1.000 Min. : 1.0   
## 1st Qu.: 3.000 1st Qu.:1.000 1st Qu.:2.000 1st Qu.:666.0   
## Median : 3.000 Median :1.000 Median :2.000 Median :677.0   
## Mean : 5.794 Mean :1.048 Mean :2.101 Mean :568.7   
## 3rd Qu.:11.000 3rd Qu.:1.000 3rd Qu.:2.000 3rd Qu.:677.0   
## Max. :11.000 Max. :2.000 Max. :7.000 Max. :677.0   
##   
## ER.VISIT URG.TIMES HOSP.VST HOSPTIME   
## Min. :1.000 Min. : 1.0 Min. :1.000 Min. : 1.0   
## 1st Qu.:2.000 1st Qu.: 5.0 1st Qu.:2.000 1st Qu.:455.0   
## Median :2.000 Median :888.0 Median :2.000 Median :455.0   
## Mean :2.696 Mean :590.4 Mean :2.808 Mean :479.7   
## 3rd Qu.:2.000 3rd Qu.:888.0 3rd Qu.:2.000 3rd Qu.:455.0   
## Max. :6.000 Max. :888.0 Max. :6.000 Max. :666.0   
##   
## HOSPPLAN ASMDCOST ASRXCOST ASSPCOST WORKTALK   
## Min. : 1.00 Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.00   
## 1st Qu.: 6.00 1st Qu.:2.000 1st Qu.:2.000 1st Qu.:2.000 1st Qu.:2.00   
## Median :45.00 Median :2.000 Median :2.000 Median :2.000 Median :2.00   
## Mean :34.72 Mean :1.928 Mean :1.836 Mean :1.955 Mean :1.88   
## 3rd Qu.:45.00 3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:2.00   
## Max. :45.00 Max. :9.000 Max. :9.000 Max. :9.000 Max. :9.00   
## NA's :3 NA's :3 NA's :3 NA's :4

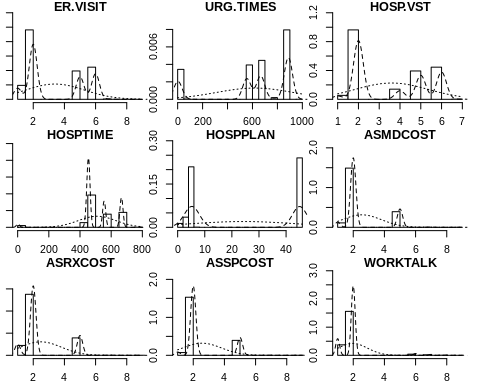
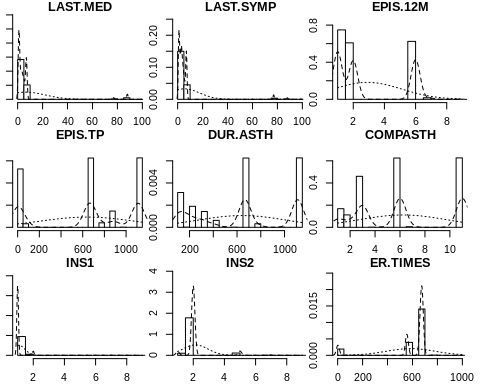
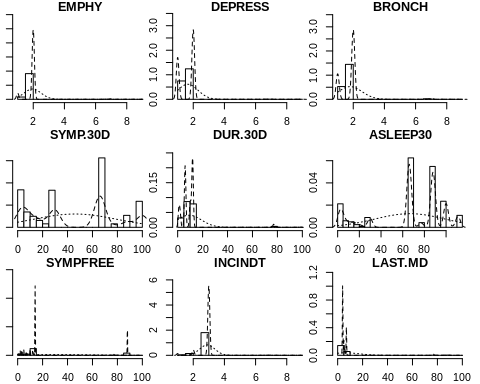
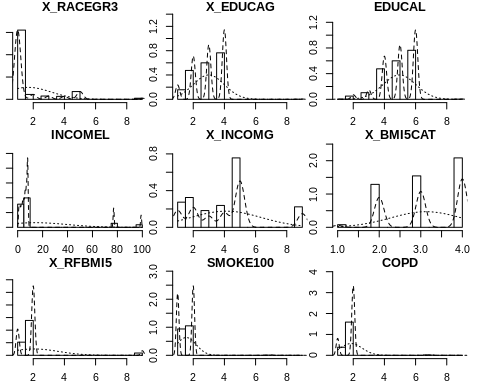
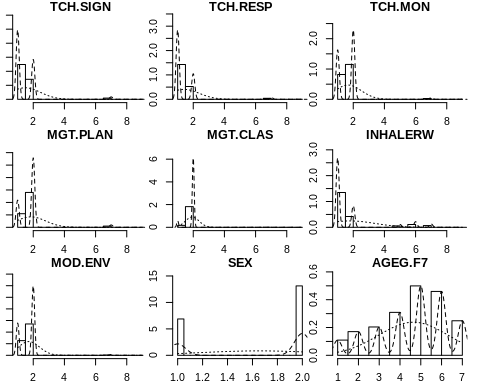
### Univariate

### EXPLORE THE DATA

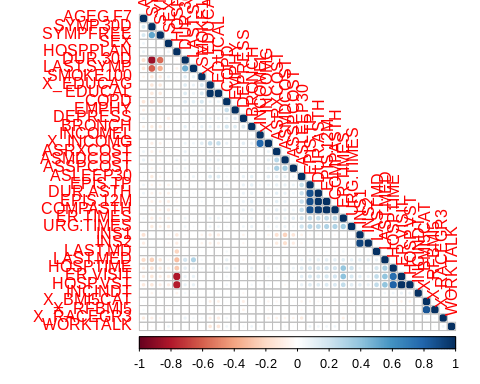
Histogram of the target variable

#### Histograms

Histograms tell us how the data is distributed in the dataset (numeric fields). There are no normal distribution in the histograms below. All the distributions are either skewed or



#### The correlations betweeen predictors



There are multiple correlation between predictors.

We are removing the predictor X\_BMISCAT because it has to much missing values There are highly correlated predictors. We are going to remove some of them.

## [1] "TCH.SIGN" "TCH.RESP" "TCH.MON" "MGT.PLAN" "MGT.CLAS" "INHALERW"   
## [7] "MOD.ENV" "SEX" "AGEG.F7" "X\_RACEGR3" "EDUCAL" "X\_INCOMG"   
## [13] "X\_RFBMI5" "SMOKE100" "COPD" "EMPHY" "DEPRESS" "BRONCH"   
## [19] "SYMP.30D" "DUR.30D" "ASLEEP30" "SYMPFREE" "INCINDT" "LAST.MD"   
## [25] "LAST.MED" "LAST.SYMP" "EPIS.12M" "DUR.ASTH" "COMPASTH" "INS1"   
## [31] "ER.TIMES" "ER.VISIT" "URG.TIMES" "HOSP.VST" "HOSPTIME" "ASRXCOST"   
## [37] "WORKTALK"

### CONSTRUCT THE RESPONSE VARIABLE

Our intention here is to squeeze the seven response variables to one so that it can fit in a regression model. To do that, we are going to cluster the seven variables and extract a binary response.

We first extract variables related to education,

Selection of variables

## TCH.SIGN TCH.RESP TCH.MON MGT.PLAN MGT.CLAS INHALERW MOD.ENV  
## 1 2 1 2 2 2 2 2  
## 2 1 1 2 2 2 1 2  
## 3 2 1 2 2 2 1 1  
## 4 1 1 1 1 2 1 2  
## 5 1 1 2 2 2 1 1  
## 6 2 2 1 2 2 2 2

#### Exploration



The elbow show that the best number of cluster is 3.

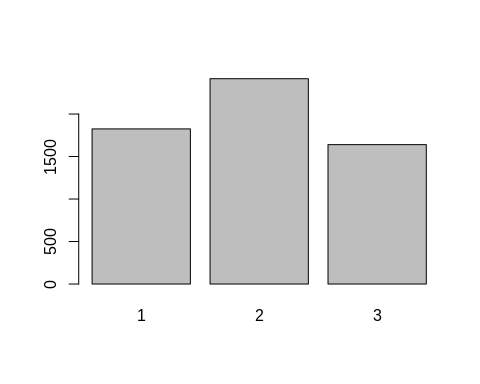
## TCH.SIGN TCH.RESP TCH.MON MGT.PLAN MGT.CLAS INHALERW MOD.ENV   
## 1:3947 1:4533 1:2801 1:1945 1: 639 1:4510 1:2222   
## 2:1933 2:1347 2:3079 2:3935 2:5241 2:1370 2:3658

## TCH.SIGN TCH.RESP TCH.MON MGT.PLAN MGT.CLAS INHALERW MOD.ENV  
## 1 1.979178 1.642192 1.788493 1.952877 1.979178 1.429589 1.797260  
## 2 1.053416 1.029814 1.000000 1.407039 1.811594 1.104348 1.487371  
## 3 1.010366 1.062805 2.000000 1.739634 1.910976 1.203659 1.625610

##### We cluster the dataset and add the point classification to the original data as target column.

## TCH.SIGN TCH.RESP TCH.MON MGT.PLAN MGT.CLAS INHALERW MOD.ENV target  
## 1 2 1 2 2 2 2 2 1  
## 2 1 1 2 2 2 1 2 3  
## 3 2 1 2 2 2 1 1 1  
## 4 1 1 1 1 2 1 2 2  
## 5 1 1 2 2 2 1 1 3  
## 6 2 2 1 2 2 2 2 1

## TCH.SIGN TCH.RESP TCH.MON MGT.PLAN MGT.CLAS INHALERW MOD.ENV target   
## 1:3947 1:4533 1:2801 1:1945 1: 639 1:4510 1:2222 1:1825   
## 2:1933 2:1347 2:3079 2:3935 2:5241 2:1370 2:3658 2:2415   
## 3:1640



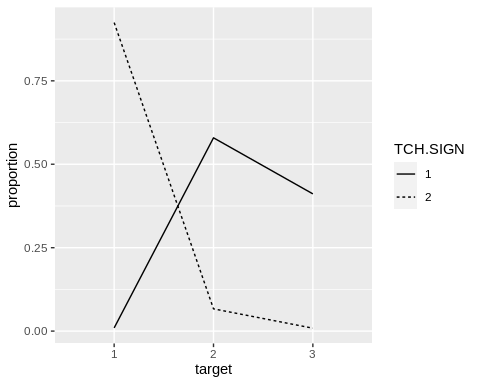
### Fig. Histogram of the clustering

### Interpretation of the Selft-Management Response clustering

#### TCH.SIGN

## # A tibble: 6 x 5  
## TCH.SIGN target count etotal proportion  
## <fct> <fct> <int> <int> <dbl>  
## 1 1 1 38 3947 0.00963  
## 2 1 2 2286 3947 0.579   
## 3 1 3 1623 3947 0.411   
## 4 2 1 1787 1933 0.924   
## 5 2 2 129 1933 0.0667   
## 6 2 3 17 1933 0.00879

## # A tibble: 2 x 6  
## # Groups: target [2]  
## TCH.SIGN target count etotal proportion group.max  
## <fct> <fct> <int> <int> <dbl> <int>  
## 1 1 2 2286 3947 0.579 2286  
## 2 2 1 1787 1933 0.924 1787

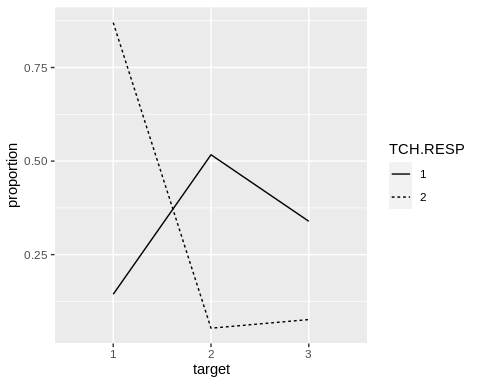


In the target response, 8 is the positive answer, 3 is the negative answer, 5 is don’t know and 6 is refused for the question: TCH\_SIGN Has a doctor or other health professional ever taught you… a. How to recognize early signs or symptoms of an asthma episode?

#### TCH.RESP

## # A tibble: 6 x 5  
## TCH.RESP target count etotal proportion  
## <fct> <fct> <int> <int> <dbl>  
## 1 1 1 653 4533 0.144   
## 2 1 2 2343 4533 0.517   
## 3 1 3 1537 4533 0.339   
## 4 2 1 1172 1347 0.870   
## 5 2 2 72 1347 0.0535  
## 6 2 3 103 1347 0.0765

## # A tibble: 2 x 6  
## # Groups: target [2]  
## TCH.RESP target count etotal proportion group.max  
## <fct> <fct> <int> <int> <dbl> <int>  
## 1 1 2 2343 4533 0.517 2343  
## 2 2 1 1172 1347 0.870 1172

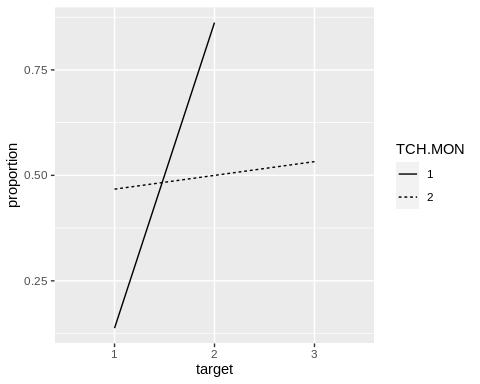


In the target response, 8 is the positive answer, 3 is the negative answer, 1 is don’t know and 1 is refused for the question: TCH\_RESP Has a doctor or other health professional ever taught you… b. What to do during an asthma episode or attack?

#### TCH.MON

## # A tibble: 4 x 5  
## TCH.MON target count etotal proportion  
## <fct> <fct> <int> <int> <dbl>  
## 1 1 1 386 2801 0.138  
## 2 1 2 2415 2801 0.862  
## 3 2 1 1439 3079 0.467  
## 4 2 3 1640 3079 0.533

## # A tibble: 2 x 6  
## # Groups: target [2]  
## TCH.MON target count etotal proportion group.max  
## <fct> <fct> <int> <int> <dbl> <int>  
## 1 1 2 2415 2801 0.862 2415  
## 2 2 3 1640 3079 0.533 1640

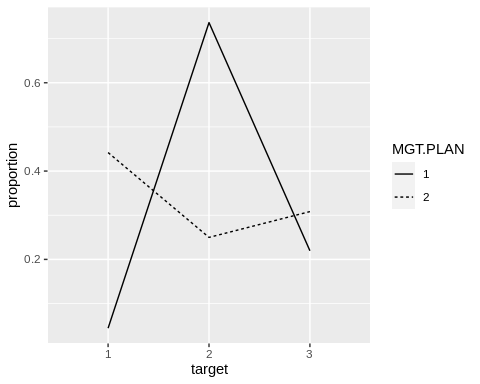


In the target response, 8 is the positive answer, 7 are the negative answers, 2 is don’t know and 2 is refused for the question: TCH\_MON A peak flow meter is a hand held device that measures how quickly you can blow air out of your lungs. Has a doctor or other health professional ever taught you… c. How to use a peak flow meter to adjust your daily medications?

#### MGT.PLAN

## # A tibble: 6 x 5  
## MGT.PLAN target count etotal proportion  
## <fct> <fct> <int> <int> <dbl>  
## 1 1 1 86 1945 0.0442  
## 2 1 2 1432 1945 0.736   
## 3 1 3 427 1945 0.220   
## 4 2 1 1739 3935 0.442   
## 5 2 2 983 3935 0.250   
## 6 2 3 1213 3935 0.308

## # A tibble: 2 x 6  
## # Groups: target [2]  
## MGT.PLAN target count etotal proportion group.max  
## <fct> <fct> <int> <int> <dbl> <int>  
## 1 1 2 1432 1945 0.736 1432  
## 2 2 1 1739 3935 0.442 1739

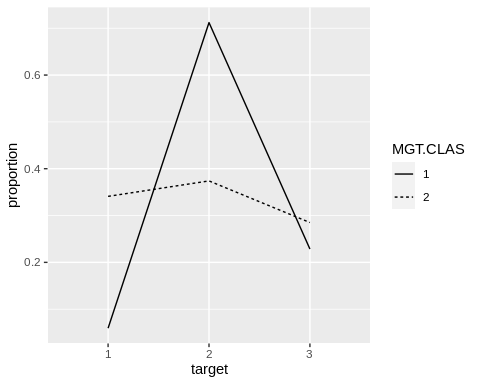


In the target response, 8 is the positive answer, 3 is the negative answer, 9 is don’t know and 9 is refused for the question: MGT\_PLAN An asthma action plan, or asthma management plan, is a form with instructions about when to change the amount or type of medicine, when to call the doctor for advice, and when to go to the emergency room. Has a doctor or other health professional EVER given you an asthma action plan?

#### MGT.CLAS

## # A tibble: 6 x 5  
## MGT.CLAS target count etotal proportion  
## <fct> <fct> <int> <int> <dbl>  
## 1 1 1 38 639 0.0595  
## 2 1 2 455 639 0.712   
## 3 1 3 146 639 0.228   
## 4 2 1 1787 5241 0.341   
## 5 2 2 1960 5241 0.374   
## 6 2 3 1494 5241 0.285

## # A tibble: 2 x 6  
## # Groups: target [1]  
## MGT.CLAS target count etotal proportion group.max  
## <fct> <fct> <int> <int> <dbl> <int>  
## 1 1 2 455 639 0.712 455  
## 2 2 2 1960 5241 0.374 1960

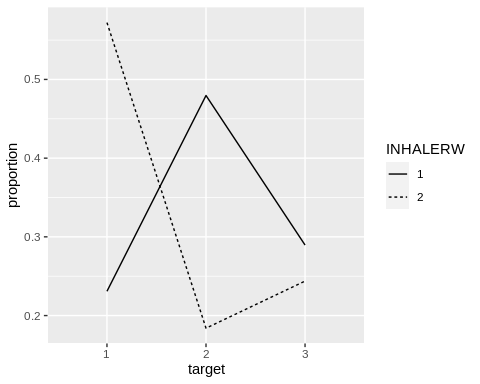


In the target response, 8 is the positive answer, 8 or(3,7) is the negative answer, 8 is don’t know and 6 is refused for the question: MGT\_CLAS Have you ever taken a course or class on how to manage your asthma?

#### INHALERW

## # A tibble: 6 x 5  
## INHALERW target count etotal proportion  
## <fct> <fct> <int> <int> <dbl>  
## 1 1 1 1041 4510 0.231  
## 2 1 2 2163 4510 0.480  
## 3 1 3 1306 4510 0.290  
## 4 2 1 784 1370 0.572  
## 5 2 2 252 1370 0.184  
## 6 2 3 334 1370 0.244

## # A tibble: 2 x 6  
## # Groups: target [2]  
## INHALERW target count etotal proportion group.max  
## <fct> <fct> <int> <int> <dbl> <int>  
## 1 1 2 2163 4510 0.480 2163  
## 2 2 1 784 1370 0.572 784

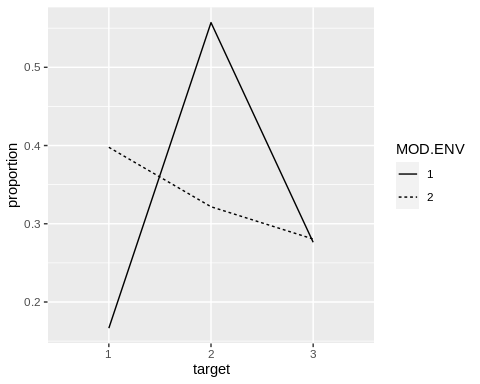


In the target response, 8 is the positive answer, 3 is the negative answer, 4 is don’t know and 1 is refused for the question: INHALERW (8.4) Did a doctor or other health professional watch you use the inhaler?

#### MOD.ENV

## # A tibble: 6 x 5  
## MOD.ENV target count etotal proportion  
## <fct> <fct> <int> <int> <dbl>  
## 1 1 1 370 2222 0.167  
## 2 1 2 1238 2222 0.557  
## 3 1 3 614 2222 0.276  
## 4 2 1 1455 3658 0.398  
## 5 2 2 1177 3658 0.322  
## 6 2 3 1026 3658 0.280

## # A tibble: 2 x 6  
## # Groups: target [2]  
## MOD.ENV target count etotal proportion group.max  
## <fct> <fct> <int> <int> <dbl> <int>  
## 1 1 2 1238 2222 0.557 1238  
## 2 2 1 1455 3658 0.398 1455



#### Summary of the response variables

## # A tibble: 2 x 8  
## RESPONSE TCH.SIGN TCH.RES TCH.MON MGT.PLAN MGT.CLAS INHALERW MOD.ENV  
## <chr> <fct> <fct> <fct> <fct> <fct> <fct> <fct>   
## 1 1=YES 2 2 2 2 2 2 2   
## 2 2=NO 1 1 3 1 2 1 1

##### For the response variable TARGET, an excellent management skill has number 2 but a poor management skill has number 1 and 3.

##### We can biuld two logistics regression on the dataset.

### Here we remove the varibles used to calculate the target variable

## 'data.frame': 5880 obs. of 31 variables:  
## $ TARGET : num 0 0 0 1 0 0 1 1 1 0 ...  
## $ SEX : num 2 2 2 1 2 2 2 1 2 1 ...  
## ..- attr(\*, "label")= chr "RESPONDENTS SEX"  
## ..- attr(\*, "format.sas")= chr "SEX"  
## $ AGEG.F7 : num 5 5 6 4 6 5 6 6 4 5 ...  
## ..- attr(\*, "label")= chr "AGE COLLAPSED TO 7 GROUPS FOR ASTHMA CALL-BACK"  
## ..- attr(\*, "format.sas")= chr "AGEG\_F7Z"  
## $ X\_RACEGR3: num 1 1 1 1 1 1 3 1 1 3 ...  
## ..- attr(\*, "label")= chr "COMPUTED FIVE LEVEL RACE/ETHNICITY CATEGORY."  
## ..- attr(\*, "format.sas")= chr "\_3RACEGR"  
## $ EDUCAL : num 4 4 6 6 6 5 3 5 4 5 ...  
## ..- attr(\*, "label")= chr "EDUCATION LEVEL"  
## ..- attr(\*, "format.sas")= chr "EDUCA"  
## $ X\_INCOMG : num 1 1 5 5 3 9 1 9 2 1 ...  
## ..- attr(\*, "label")= chr "COMPUTED INCOME CATEGORIES"  
## ..- attr(\*, "format.sas")= chr "\_INCOMG"  
## $ X\_RFBMI5 : num 2 2 2 1 2 9 9 1 2 2 ...  
## ..- attr(\*, "label")= chr "OVERWEIGHT OR OBESE CALCULATED VARIABLE"  
## ..- attr(\*, "format.sas")= chr "\_5RFBMI"  
## $ SMOKE100 : num 1 1 1 1 2 2 2 2 1 1 ...  
## ..- attr(\*, "label")= chr "SMOKED AT LEAST 100 CIGARETTES"  
## ..- attr(\*, "format.sas")= chr "SMOK100\_"  
## $ COPD : num 1 2 2 2 2 2 2 2 1 2 ...  
## ..- attr(\*, "label")= chr "EVER TOLD HAVE CHRONIC OBSTRUCTIVE PULMONARY DISEASE"  
## ..- attr(\*, "format.sas")= chr "COPD"  
## $ EMPHY : num 2 2 2 2 2 2 2 2 1 2 ...  
## ..- attr(\*, "label")= chr "EVER TOLD HAVE EMPHYSEMA"  
## ..- attr(\*, "format.sas")= chr "EMPHY"  
## $ DEPRESS : num 1 2 2 2 1 1 1 2 1 1 ...  
## ..- attr(\*, "label")= chr "EVER TOLD DEPRESSED"  
## ..- attr(\*, "format.sas")= chr "DEPRESS"  
## $ BRONCH : num 1 2 1 2 1 2 2 2 1 2 ...  
## ..- attr(\*, "label")= chr "EVER TOLD HAVE CHRONIC BRONCHITIS"  
## ..- attr(\*, "format.sas")= chr "BRONCH"  
## $ SYMP.30D : num 30 7 17 10 30 5 30 7 14 30 ...  
## ..- attr(\*, "label")= chr "SYMPTOM DAYS"  
## ..- attr(\*, "format.sas")= chr "SYMP\_30D"  
## $ DUR.30D : num 2 12 12 12 1 12 1 12 12 1 ...  
## ..- attr(\*, "label")= chr "CONSTANT SYMPTOMS"  
## ..- attr(\*, "format.sas")= chr "DUR\_30D"  
## $ ASLEEP30 : num 15 6 88 88 20 88 4 88 7 30 ...  
## ..- attr(\*, "label")= chr "NIGHT SYMPTOMS"  
## ..- attr(\*, "format.sas")= chr "ASLEEP30Z"  
## $ SYMPFREE : num 88 6 11 7 88 5 3 7 88 88 ...  
## ..- attr(\*, "label")= chr "SYMPTOM-FREE DAYS"  
## ..- attr(\*, "format.sas")= chr "SYMPFREE"  
## $ INCINDT : num 2 3 3 3 3 3 3 3 3 2 ...  
## ..- attr(\*, "label")= chr "TIME SINCE DIAGNOSIS"  
## ..- attr(\*, "format.sas")= chr "INCIDNT"  
## $ LAST.MD : num 4 4 4 4 4 7 6 5 7 4 ...  
## ..- attr(\*, "label")= chr "LAST TALKED TO A DOCTOR"  
## ..- attr(\*, "format.sas")= chr "LAST\_MD"  
## $ LAST.MED : num 1 3 3 1 1 5 3 1 7 1 ...  
## ..- attr(\*, "label")= chr "LAST TOOK ASTHMA MEDICATION"  
## ..- attr(\*, "format.sas")= chr "LAST\_MED"  
## $ LAST.SYMP: num 1 3 3 3 1 3 3 3 3 1 ...  
## ..- attr(\*, "label")= chr "LAST HAD ANY SYMPTOMS OF ASTHMA"  
## ..- attr(\*, "format.sas")= chr "LASTSYMP"  
## $ EPIS.12M : num 1 1 1 1 2 2 2 2 2 1 ...  
## ..- attr(\*, "label")= chr "ASTHMA EPISODE OR ATTACK"  
## ..- attr(\*, "format.sas")= chr "EPIS\_12M"  
## $ DUR.ASTH : num 204 120 115 201 1111 ...  
## ..- attr(\*, "label")= chr "DURATION OF ATTACK"  
## ..- attr(\*, "format.sas")= chr "DUR\_ASTH"  
## $ COMPASTH : num 3 1 3 3 11 11 11 11 11 3 ...  
## ..- attr(\*, "label")= chr "TYPICAL ATTACK"  
## ..- attr(\*, "format.sas")= chr "COMPASTH"  
## $ INS1 : num 1 1 1 1 1 1 1 2 1 1 ...  
## ..- attr(\*, "label")= chr "INSURANCE"  
## ..- attr(\*, "format.sas")= chr "INS1Z"  
## $ ER.TIMES : num 677 677 677 677 677 666 666 666 666 677 ...  
## ..- attr(\*, "label")= chr "NUMBER OF EMERGENCY ROOM VISITS"  
## ..- attr(\*, "format.sas")= chr "ER\_TIMES"  
## $ ER.VISIT : num 2 2 2 2 2 6 6 6 6 2 ...  
## ..- attr(\*, "label")= chr "EMERGENCY ROOM VISIT"  
## ..- attr(\*, "format.sas")= chr "ER\_VISIT"  
## $ URG.TIMES: num 888 3 888 888 888 666 666 666 666 888 ...  
## ..- attr(\*, "label")= chr "NUMBER OF URGENT VISITS"  
## ..- attr(\*, "format.sas")= chr "URG\_TIME"  
## $ HOSP.VST : num 2 2 2 2 2 6 6 6 6 2 ...  
## ..- attr(\*, "label")= chr "HOSPITAL VISIT"  
## ..- attr(\*, "format.sas")= chr "HOSP\_VST"  
## $ HOSPTIME : num 455 455 455 455 455 666 666 666 666 455 ...  
## ..- attr(\*, "label")= chr "NUMBER OF HOSPITAL VISITS"  
## ..- attr(\*, "format.sas")= chr "HOSPTIME"  
## $ ASRXCOST : num 2 2 2 2 1 1 2 1 2 2 ...  
## ..- attr(\*, "label")= chr "COST BARRIER: MEDICATION"  
## ..- attr(\*, "format.sas")= chr "ASRXCOST"  
## $ WORKTALK : num 2 2 2 2 2 2 2 2 2 2 ...  
## ..- attr(\*, "label")= chr "DOCTOR DISCUSSED WORK ASTHMA"  
## ..- attr(\*, "format.sas")= chr "WORKTALK"

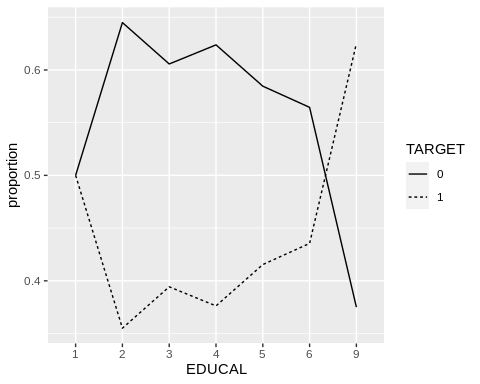
### PREPARE THE DATA FOR MODELISATION

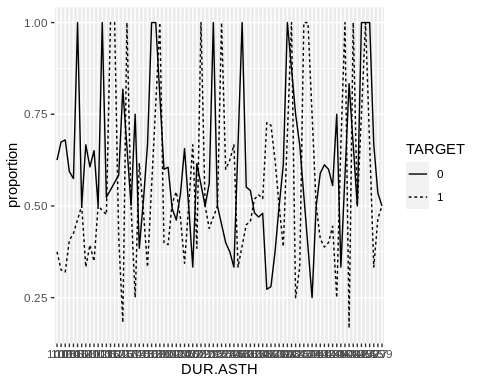
#### We remove the rows with missing values.

Here were are going to drop missing data because they are only 12 over 13,922 rows. We also transform all predictors to categorical.

## TARGET SEX AGEG.F7 X\_RACEGR3 EDUCAL X\_INCOMG X\_RFBMI5 SMOKE100  
## 0:3457 1:1737 1: 241 1:4675 1: 8 1: 931 1:1377 1:2985   
## 1:2409 2:4129 2: 416 2: 312 2: 138 2:1078 2:4230 2:2857   
## 3: 540 3: 206 3: 317 3: 529 9: 259 7: 22   
## 4:1049 4: 212 4:1454 4: 726 9: 2   
## 5:1605 5: 404 5:1856 5:2041   
## 6:1384 9: 57 6:2085 9: 561   
## 7: 631 9: 8   
## COPD EMPHY DEPRESS BRONCH SYMP.30D DUR.30D ASLEEP30   
## 1:1596 1: 664 1:2550 1:2001 30 :1857 1 :1066 88 :3110   
## 2:4202 2:5160 2:3300 2:3810 2 : 455 12:4009 30 : 475   
## 7: 61 7: 39 7: 7 7: 51 3 : 438 2 : 760 2 : 317   
## 9: 7 9: 3 9: 9 9: 4 5 : 431 7 : 31 3 : 272   
## 10 : 415 5 : 261   
## 15 : 412 1 : 259   
## (Other):1858 (Other):1172   
## SYMPFREE INCINDT LAST.MD LAST.MED LAST.SYMP EPIS.12M DUR.ASTH   
## 88 :2176 1: 174 4 :4620 1:3356 1:2935 1:3564 1111 :2302   
## 14 : 512 2: 501 5 : 702 2: 775 2:1293 2:2302 105 : 301   
## 10 : 491 3:5191 6 : 223 3: 954 3:1638 130 : 295   
## 7 : 383 7 : 301 4: 278 201 : 289   
## 12 : 366 88: 20 5: 196 202 : 266   
## 5 : 354 6: 99 110 : 249   
## (Other):1584 7: 208 (Other):2164   
## COMPASTH INS1 ER.TIMES ER.VISIT URG.TIMES HOSP.VST  
## 1 : 762 1:5584 677 :3698 1: 922 888 :2957 1: 264   
## 11:2302 2: 282 666 :1246 2:3698 666 :1246 2:4356   
## 2 : 552 1 : 456 6:1246 1 : 688 6:1246   
## 3 :2242 2 : 227 2 : 454   
## 4 : 8 3 : 106 3 : 202   
## 4 : 63 4 : 113   
## (Other): 70 (Other): 206   
## HOSPTIME ASRXCOST WORKTALK  
## 455 :4356 1:1043 1:1385   
## 666 :1246 2:4810 2:4331   
## 1 : 143 7: 10 6: 109   
## 2 : 49 9: 3 7: 32   
## 3 : 29 8: 5   
## 4 : 21 9: 4   
## (Other): 22

### Visualization





#### Splitting the data into train and test sets

### BUILDS MODELS

#### Model using full predictors with glm

##   
## Call: glm(formula = TARGET ~ ., family = binomial, data = training1)  
##   
## Coefficients:  
## (Intercept) SEX2 AGEG.F72 AGEG.F73 AGEG.F74   
## 3.148e-01 1.293e-01 1.131e-01 1.504e-01 9.236e-02   
## AGEG.F75 AGEG.F76 AGEG.F77 X\_RACEGR32 X\_RACEGR33   
## -1.835e-01 -4.512e-01 -5.215e-01 4.031e-01 -4.809e-02   
## X\_RACEGR34 X\_RACEGR35 X\_RACEGR39 EDUCAL2 EDUCAL3   
## 4.576e-01 6.589e-02 2.576e-02 -8.339e-01 -8.178e-01   
## EDUCAL4 EDUCAL5 EDUCAL6 EDUCAL9 X\_INCOMG2   
## -8.853e-01 -6.926e-01 -6.861e-01 2.897e-01 2.582e-02   
## X\_INCOMG3 X\_INCOMG4 X\_INCOMG5 X\_INCOMG9 X\_RFBMI52   
## 1.535e-01 1.394e-02 1.730e-01 -4.262e-02 -2.091e-02   
## X\_RFBMI59 SMOKE1002 SMOKE1007 SMOKE1009 COPD2   
## -1.833e-01 1.098e-01 5.766e-01 1.668e+01 -1.340e-01   
## COPD7 COPD9 EMPHY2 EMPHY7 EMPHY9   
## -2.604e-01 -1.588e+01 -9.427e-02 -6.301e-01 -5.219e-01   
## DEPRESS2 DEPRESS7 DEPRESS9 BRONCH2 BRONCH7   
## -8.341e-03 9.152e-01 -5.560e-01 -6.622e-02 -8.376e-01   
## BRONCH9 SYMP.30D10 SYMP.30D11 SYMP.30D12 SYMP.30D13   
## 1.703e+01 -9.024e-02 -1.696e+01 4.681e-02 -5.399e-01   
## SYMP.30D14 SYMP.30D15 SYMP.30D16 SYMP.30D17 SYMP.30D18   
## -2.700e-01 -7.077e-02 3.487e-01 5.211e-01 -2.048e-01   
## SYMP.30D19 SYMP.30D2 SYMP.30D20 SYMP.30D21 SYMP.30D22   
## 1.679e+01 1.342e-01 -1.241e-01 -5.655e-01 -1.607e+01   
## SYMP.30D23 SYMP.30D24 SYMP.30D25 SYMP.30D26 SYMP.30D27   
## -1.352e+00 -1.209e+00 2.490e-01 -1.702e+01 7.241e-01   
## SYMP.30D28 SYMP.30D29 SYMP.30D3 SYMP.30D30 SYMP.30D4   
## -4.543e-02 7.101e-01 1.874e-01 1.483e-01 -6.604e-02   
## SYMP.30D5 SYMP.30D6 SYMP.30D7 SYMP.30D8 SYMP.30D9   
## 1.316e-01 -2.286e-01 -2.617e-01 -2.095e-01 5.006e-01   
## DUR.30D12 DUR.30D2 DUR.30D7 ASLEEP3010 ASLEEP3011   
## NA -2.974e-01 -5.443e-01 3.808e-01 -1.575e+01   
## ASLEEP3012 ASLEEP3013 ASLEEP3014 ASLEEP3015 ASLEEP3016   
## 6.213e-01 -1.585e+01 1.460e-02 2.668e-01 2.043e+00   
## ASLEEP3017 ASLEEP3018 ASLEEP3019 ASLEEP302 ASLEEP3020   
## -1.592e+00 -5.210e-01 -1.556e+01 2.596e-01 3.092e-03   
## ASLEEP3021 ASLEEP3022 ASLEEP3024 ASLEEP3025 ASLEEP3026   
## 1.536e+00 -1.713e+01 -1.549e+01 2.309e-01 1.652e+01   
## ASLEEP3027 ASLEEP3028 ASLEEP3029 ASLEEP303 ASLEEP3030   
## 1.538e+01 -6.753e-01 3.006e-02 -5.431e-02 9.782e-02   
## ASLEEP304 ASLEEP305 ASLEEP306 ASLEEP307 ASLEEP3077   
## 9.922e-02 -2.422e-01 -8.901e-01 4.986e-01 1.468e-01   
## ASLEEP308 ASLEEP3088 ASLEEP309 SYMPFREE10 SYMPFREE11   
## 1.638e+00 6.594e-03 1.821e+00 3.441e-01 1.086e-01   
## SYMPFREE12 SYMPFREE13 SYMPFREE14 SYMPFREE2 SYMPFREE3   
## 3.239e-01 1.763e-01 2.656e-01 2.289e-01 2.224e-01   
## SYMPFREE4 SYMPFREE5 SYMPFREE6 SYMPFREE7 SYMPFREE77   
## -2.764e-01 2.955e-01 3.644e-01 4.241e-01 4.134e-01   
## SYMPFREE8 SYMPFREE88 SYMPFREE9 SYMPFREE99 INCINDT2   
## 6.427e-02 -3.411e-02 2.999e-01 -1.721e-01 3.213e-01   
## INCINDT3 LAST.MD5 LAST.MD6 LAST.MD7 LAST.MD88   
## 1.105e+00 -4.755e-01 -9.314e-01 -9.252e-01 -6.223e-01   
## LAST.MED2 LAST.MED3 LAST.MED4 LAST.MED5 LAST.MED6   
## -3.517e-01 -5.208e-01 -5.526e-01 -5.898e-01 -4.604e-01   
## LAST.MED7 LAST.SYMP2 LAST.SYMP3 EPIS.12M2 DUR.ASTH102   
## -8.205e-01 2.458e-01 2.176e-01 -7.685e-02 -2.437e-01   
## DUR.ASTH103 DUR.ASTH104 DUR.ASTH105 DUR.ASTH106 DUR.ASTH107   
## -1.478e-01 3.440e-01 7.052e-02 -1.623e+01 5.414e-01   
## DUR.ASTH108 DUR.ASTH110 DUR.ASTH1111 DUR.ASTH112 DUR.ASTH113   
## -8.825e-01 -3.022e-02 NA 3.253e-01 -1.571e+01   
## DUR.ASTH115 DUR.ASTH116 DUR.ASTH117 DUR.ASTH120 DUR.ASTH125   
## 2.165e-01 1.616e+01 1.767e+01 7.653e-02 -2.389e+00   
## DUR.ASTH127 DUR.ASTH130 DUR.ASTH135 DUR.ASTH140 DUR.ASTH145   
## 3.167e+01 3.672e-01 3.499e-01 4.244e-01 2.538e-01   
## DUR.ASTH150 DUR.ASTH152 DUR.ASTH160 DUR.ASTH180 DUR.ASTH190   
## 7.228e-01 -1.625e+01 -1.683e+01 1.943e+01 -1.712e-01   
## DUR.ASTH201 DUR.ASTH202 DUR.ASTH203 DUR.ASTH204 DUR.ASTH205   
## 1.119e-02 3.736e-01 4.943e-01 3.095e-01 1.072e-02   
## DUR.ASTH206 DUR.ASTH207 DUR.ASTH208 DUR.ASTH210 DUR.ASTH212   
## -2.081e-02 1.094e+00 -1.269e-01 7.457e-01 6.886e-01   
## DUR.ASTH213 DUR.ASTH216 DUR.ASTH220 DUR.ASTH224 DUR.ASTH230   
## -1.726e+01 3.983e-01 1.752e+01 9.331e-01 1.874e+00   
## DUR.ASTH236 DUR.ASTH248 DUR.ASTH259 DUR.ASTH301 DUR.ASTH302   
## 3.326e-01 -4.223e-01 -1.655e+01 2.308e-01 1.646e-01   
## DUR.ASTH303 DUR.ASTH304 DUR.ASTH305 DUR.ASTH306 DUR.ASTH307   
## 6.122e-01 4.797e-01 8.527e-01 9.096e-01 1.356e+00   
## DUR.ASTH308 DUR.ASTH309 DUR.ASTH310 DUR.ASTH311 DUR.ASTH312   
## 4.453e-01 -4.595e-02 -3.048e-01 9.586e-01 1.621e+01   
## DUR.ASTH314 DUR.ASTH315 DUR.ASTH317 DUR.ASTH320 DUR.ASTH321   
## -1.367e-02 -1.639e+01 1.851e+01 1.796e+01 1.677e+01   
## DUR.ASTH330 DUR.ASTH401 DUR.ASTH402 DUR.ASTH403 DUR.ASTH404   
## 1.478e+00 1.781e-01 -6.890e-01 2.291e-01 6.058e-01   
## DUR.ASTH405 DUR.ASTH406 DUR.ASTH407 DUR.ASTH408 DUR.ASTH410   
## 9.910e-02 2.114e+00 1.678e+01 -1.595e+01 1.621e+01   
## DUR.ASTH412 DUR.ASTH420 DUR.ASTH424 DUR.ASTH452 DUR.ASTH555   
## 5.372e-01 -1.550e+01 1.432e+01 -1.636e+01 3.397e-01   
## DUR.ASTH777 DUR.ASTH999 COMPASTH11 COMPASTH2 COMPASTH3   
## 4.177e-01 -2.681e-01 NA -3.773e-01 -1.572e-01   
## COMPASTH4 INS12 ER.TIMES10 ER.TIMES12 ER.TIMES14   
## -1.395e+00 -4.535e-02 -1.638e+01 4.603e+01 -1.681e+01   
## ER.TIMES2 ER.TIMES20 ER.TIMES25 ER.TIMES3 ER.TIMES30   
## 3.835e-01 1.602e+01 NA -5.714e-02 4.790e+01   
## ER.TIMES32 ER.TIMES4 ER.TIMES5 ER.TIMES6 ER.TIMES666   
## 1.805e+01 2.001e-02 1.068e+00 -1.913e+00 NA   
## ER.TIMES677 ER.TIMES7 ER.TIMES8 ER.TIMES9 ER.VISIT2   
## 6.679e-02 1.958e+00 1.055e+00 1.598e+01 NA   
## ER.VISIT6 URG.TIMES10 URG.TIMES12 URG.TIMES15 URG.TIMES16   
## NA -1.570e+00 3.802e-01 5.609e-01 -1.682e+01   
## URG.TIMES18 URG.TIMES2 URG.TIMES20 URG.TIMES24 URG.TIMES3   
## NA -2.872e-01 1.526e+01 -1.379e+01 1.018e-01   
## URG.TIMES30 URG.TIMES4 URG.TIMES45 URG.TIMES5 URG.TIMES52   
## -3.159e+01 1.385e-01 1.648e+01 -5.072e-02 1.615e+01   
## URG.TIMES6 URG.TIMES60 URG.TIMES666 URG.TIMES7 URG.TIMES77   
## 4.643e-01 1.735e+01 NA -1.315e+00 -1.505e+01   
## URG.TIMES8 URG.TIMES88 URG.TIMES888 URG.TIMES9 HOSP.VST2   
## 4.949e-01 -4.400e-01 -1.682e-01 -9.831e-01 -2.567e-01   
## HOSP.VST6 HOSPTIME10 HOSPTIME12 HOSPTIME13 HOSPTIME17   
## NA 1.324e+01 -6.296e+01 1.749e+01 NA   
## HOSPTIME2 HOSPTIME3 HOSPTIME4 HOSPTIME455 HOSPTIME5   
## 2.502e-01 6.003e-01 9.197e-01 NA -7.699e-01   
## HOSPTIME6 HOSPTIME666 HOSPTIME7 HOSPTIME8 HOSPTIME9   
## -4.379e-01 NA 1.621e+01 -4.374e+00 1.768e+01   
## ASRXCOST2 ASRXCOST7 ASRXCOST9 WORKTALK2 WORKTALK6   
## -7.717e-04 -1.617e+01 -4.035e-01 -6.936e-01 -1.122e+00   
## WORKTALK7 WORKTALK8 WORKTALK9   
## -9.093e-01 1.542e-01 1.181e-01   
##   
## Degrees of Freedom: 4693 Total (i.e. Null); 4424 Residual  
## Null Deviance: 6357   
## Residual Deviance: 5610 AIC: 6150

#### Confusion Matrix with the testingset

#### First glm model using backward elimination of step function

##   
## Call: glm(formula = TARGET ~ SEX + AGEG.F7 + X\_RACEGR3 + EDUCAL + COPD +   
## BRONCH + DUR.30D + INCINDT + LAST.MD + LAST.MED + LAST.SYMP +   
## COMPASTH + ASRXCOST + WORKTALK, family = binomial, data = training1)  
##   
## Coefficients:  
## (Intercept) SEX2 AGEG.F72 AGEG.F73 AGEG.F74 AGEG.F75   
## 0.572975 0.134428 0.019596 0.125438 0.059608 -0.203982   
## AGEG.F76 AGEG.F77 X\_RACEGR32 X\_RACEGR33 X\_RACEGR34 X\_RACEGR35   
## -0.454594 -0.525984 0.438177 -0.065017 0.385895 0.144983   
## X\_RACEGR39 EDUCAL2 EDUCAL3 EDUCAL4 EDUCAL5 EDUCAL6   
## 0.028981 -1.000982 -0.805075 -0.912798 -0.716403 -0.640010   
## EDUCAL9 COPD2 COPD7 COPD9 BRONCH2 BRONCH7   
## 0.326210 -0.116369 -0.201415 -14.506809 -0.122576 -0.825107   
## BRONCH9 DUR.30D12 DUR.30D2 DUR.30D7 INCINDT2 INCINDT3   
## 14.366382 -0.015806 -0.322705 -0.414395 0.222194 1.017538   
## LAST.MD5 LAST.MD6 LAST.MD7 LAST.MD88 LAST.MED2 LAST.MED3   
## -0.134753 -0.549793 -0.622004 -0.435851 -0.359625 -0.559291   
## LAST.MED4 LAST.MED5 LAST.MED6 LAST.MED7 LAST.SYMP2 LAST.SYMP3   
## -0.589799 -0.650823 -0.516207 -0.876367 0.247679 0.312643   
## COMPASTH11 COMPASTH2 COMPASTH3 COMPASTH4 ASRXCOST2 ASRXCOST7   
## -0.283554 -0.255527 -0.124567 -1.325465 -0.007984 -14.125953   
## ASRXCOST9 WORKTALK2 WORKTALK6 WORKTALK7 WORKTALK8 WORKTALK9   
## -0.132507 -0.650284 -1.117860 -1.004688 0.349332 0.080784   
##   
## Degrees of Freedom: 4693 Total (i.e. Null); 4640 Residual  
## Null Deviance: 6357   
## Residual Deviance: 5919 AIC: 6027

Call: glm(formula = TARGET ~ SEX + AGEG.F7 + X\_RACEGR3 + EDUCAL + BRONCH + DUR.30D + INCINDT + LAST.MD + LAST.MED + LAST.SYMP + COMPASTH + HOSPTIME + ASRXCOST + WORKTALK, family = binomial, data = training1)

#### Confusion Matrix with the testingset

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 530 302  
## 1 161 179  
##   
## Accuracy : 0.6049   
## 95% CI : (0.5763, 0.6331)  
## No Information Rate : 0.5896   
## P-Value [Acc > NIR] : 0.1493   
##   
## Kappa : 0.1456   
##   
## Mcnemar's Test P-Value : 7.7e-11   
##   
## Sensitivity : 0.3721   
## Specificity : 0.7670   
## Pos Pred Value : 0.5265   
## Neg Pred Value : 0.6370   
## Prevalence : 0.4104   
## Detection Rate : 0.1527   
## Detection Prevalence : 0.2901   
## Balanced Accuracy : 0.5696   
##   
## 'Positive' Class : 1   
##

#### Second glm model

##   
## Call: glm(formula = TARGET ~ SEX + AGEG.F7 + X\_RACEGR3 + EDUCAL + X\_INCOMG +   
## BRONCH + DUR.30D + INCINDT + LAST.MD + LAST.MED + LAST.SYMP +   
## COMPASTH + WORKTALK, family = binomial, data = training1)  
##   
## Coefficients:  
## (Intercept) SEX2 AGEG.F72 AGEG.F73 AGEG.F74 AGEG.F75   
## 0.51709 0.14366 0.01621 0.10818 0.05680 -0.20145   
## AGEG.F76 AGEG.F77 X\_RACEGR32 X\_RACEGR33 X\_RACEGR34 X\_RACEGR35   
## -0.43018 -0.49934 0.45018 -0.04577 0.39576 0.15849   
## X\_RACEGR39 EDUCAL2 EDUCAL3 EDUCAL4 EDUCAL5 EDUCAL6   
## 0.07587 -1.00516 -0.79980 -0.92160 -0.75202 -0.71345   
## EDUCAL9 X\_INCOMG2 X\_INCOMG3 X\_INCOMG4 X\_INCOMG5 X\_INCOMG9   
## 0.32313 0.02614 0.08805 -0.02705 0.13954 -0.07447   
## BRONCH2 BRONCH7 BRONCH9 DUR.30D12 DUR.30D2 DUR.30D7   
## -0.16285 -0.80582 1.41137 -0.04488 -0.35773 -0.43165   
## INCINDT2 INCINDT3 LAST.MD5 LAST.MD6 LAST.MD7 LAST.MD88   
## 0.22100 1.00806 -0.13356 -0.56097 -0.60689 -0.41855   
## LAST.MED2 LAST.MED3 LAST.MED4 LAST.MED5 LAST.MED6 LAST.MED7   
## -0.36450 -0.56518 -0.60133 -0.66289 -0.51592 -0.87298   
## LAST.SYMP2 LAST.SYMP3 COMPASTH11 COMPASTH2 COMPASTH3 COMPASTH4   
## 0.24389 0.30375 -0.29369 -0.26218 -0.12404 -1.33403   
## WORKTALK2 WORKTALK6 WORKTALK7 WORKTALK8 WORKTALK9   
## -0.65023 -1.07686 -0.99734 0.27756 -0.64155   
##   
## Degrees of Freedom: 4693 Total (i.e. Null); 4641 Residual  
## Null Deviance: 6357   
## Residual Deviance: 5928 AIC: 6034

#### Confusion Matrix with the testingset

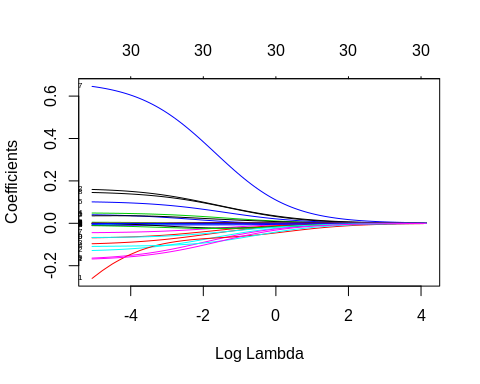
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 537 297  
## 1 154 184  
##   
## Accuracy : 0.6152   
## 95% CI : (0.5866, 0.6432)  
## No Information Rate : 0.5896   
## P-Value [Acc > NIR] : 0.03957   
##   
## Kappa : 0.1672   
##   
## Mcnemar's Test P-Value : 2.285e-11   
##   
## Sensitivity : 0.3825   
## Specificity : 0.7771   
## Pos Pred Value : 0.5444   
## Neg Pred Value : 0.6439   
## Prevalence : 0.4104   
## Detection Rate : 0.1570   
## Detection Prevalence : 0.2884   
## Balanced Accuracy : 0.5798   
##   
## 'Positive' Class : 1   
##

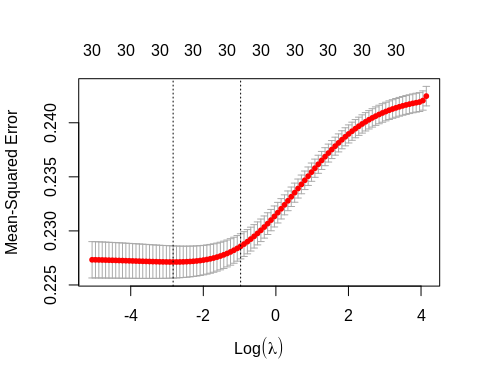
#### Lasso and Ridge model

Since our dataset has multiple variable, we can use penalized logistic regression to find an optimal performing model. Ridge Regression and Lasso Regression have two different approaches. Ridge Regression incorporates all variables in the model and gives the coefficients of variables with minor contribution close to zero Lasso Regression keeps only the most significant variables and gives zero to the coefficient of the rest of variables.

#### Split the data into trainset and testingset, Dumy code categorical predictors

#### Ridge Regression

We fit and obsrve the coefficients of rigde regression against the log of lambda.  The coefficients are significative for negative log lambda and start stabilize around -4



The plot shows that the log of the optimal value of lambda (i.e. the one that minimises the root mean square error) is approximately -3. The exact value can be viewed by examining the variable lambda\_min in the code below. In general though, the objective of regularisation is to balance accuracy and simplicity. In the present context, this means a model with the smallest number of coefficients that also gives a good accuracy. To this end, the cv.glmnet function finds the value of lambda that gives the simplest model but also lies within one standard error of the optimal value of lambda.

## [1] 0.05889484

#### Confusion matrix with lambda min

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 690 443  
## 1 14 26  
##   
## Accuracy : 0.6104   
## 95% CI : (0.5818, 0.6384)  
## No Information Rate : 0.6002   
## P-Value [Acc > NIR] : 0.2469   
##   
## Kappa : 0.042   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.05544   
## Specificity : 0.98011   
## Pos Pred Value : 0.65000   
## Neg Pred Value : 0.60900   
## Prevalence : 0.39983   
## Detection Rate : 0.02217   
## Detection Prevalence : 0.03410   
## Balanced Accuracy : 0.51778   
##   
## 'Positive' Class : 1   
##

We observe overfitting with this ridge model

#### Confusion matrix with best lambda

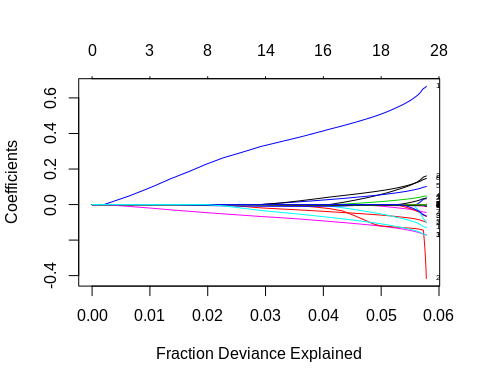
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 704 469  
## 1 0 0  
##   
## Accuracy : 0.6002   
## 95% CI : (0.5715, 0.6283)  
## No Information Rate : 0.6002   
## P-Value [Acc > NIR] : 0.5127   
##   
## Kappa : 0   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.0000   
## Specificity : 1.0000   
## Pos Pred Value : NaN   
## Neg Pred Value : 0.6002   
## Prevalence : 0.3998   
## Detection Rate : 0.0000   
## Detection Prevalence : 0.0000   
## Balanced Accuracy : 0.5000   
##   
## 'Positive' Class : 1   
##

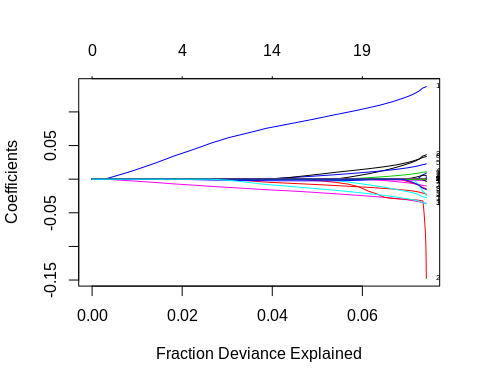
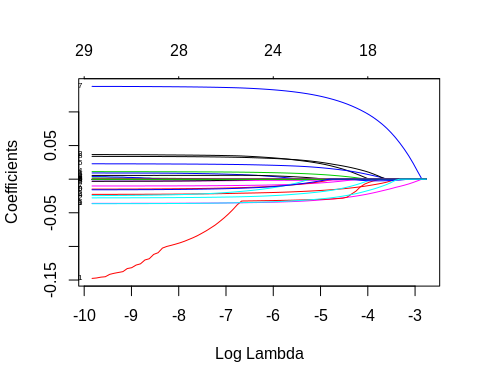
We observe overfitting with this second ridge model

#### Getting the coefficients

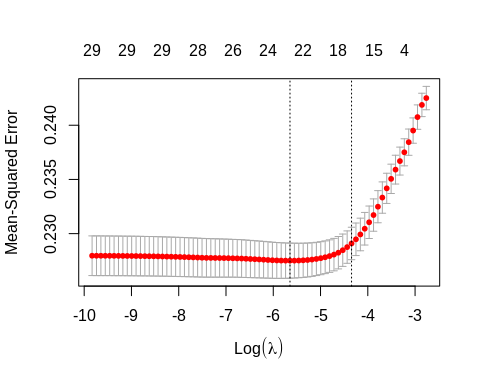
## 32 x 1 sparse Matrix of class "dgCMatrix"  
## s0  
## (Intercept) -2.163978e-01  
## (Intercept) .   
## SEX 1.265288e-01  
## AGEG.F7 -7.217141e-02  
## X\_RACEGR3 3.963990e-02  
## EDUCAL 8.261740e-02  
## X\_INCOMG -1.785201e-03  
## X\_RFBMI5 -3.518837e-02  
## SMOKE100 1.196469e-01  
## COPD -5.320546e-02  
## EMPHY 2.256841e-03  
## DEPRESS 2.532519e-02  
## BRONCH -9.988764e-02  
## SYMP.30D -5.963146e-03  
## DUR.30D 2.678729e-04  
## ASLEEP30 -1.807218e-03  
## SYMPFREE -1.728817e-03  
## INCINDT 5.013512e-01  
## LAST.MD -4.436136e-03  
## LAST.MED -1.209514e-01  
## LAST.SYMP 2.998216e-02  
## EPIS.12M -8.959034e-02  
## DUR.ASTH -1.092595e-07  
## COMPASTH -7.884146e-03  
## INS1 -1.026336e-01  
## ER.TIMES -2.007619e-04  
## ER.VISIT -1.507517e-02  
## URG.TIMES -1.763067e-04  
## HOSP.VST -2.085963e-02  
## HOSPTIME -8.309835e-04  
## ASRXCOST -5.868057e-02  
## WORKTALK -1.321715e-01

##### Lasso Regression





#### Find the best lambda using cross validation

 The plot shows that the log of the optimal value of lambda (i.e. the one that minimises the root mean square error) is approximately -10. The exact value can be viewed by examining the variable lambda\_min in the code below. In general though, the objective of regularisation is to balance accuracy and simplicity. In the present context, this means a model with the smallest number of coefficients that also gives a good accuracy. To this end, the cv.glmnet function finds the value of lambda that gives the simplest model but also lies within one standard error of the optimal value of lambda.

#### Confusion Matrix with lambda min

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 583 324  
## 1 121 145  
##   
## Accuracy : 0.6206   
## 95% CI : (0.5922, 0.6485)  
## No Information Rate : 0.6002   
## P-Value [Acc > NIR] : 0.08035   
##   
## Kappa : 0.148   
##   
## Mcnemar's Test P-Value : < 2e-16   
##   
## Sensitivity : 0.3092   
## Specificity : 0.8281   
## Pos Pred Value : 0.5451   
## Neg Pred Value : 0.6428   
## Prevalence : 0.3998   
## Detection Rate : 0.1236   
## Detection Prevalence : 0.2268   
## Balanced Accuracy : 0.5686   
##   
## 'Positive' Class : 1   
##

#### Getting the coefficients

## 32 x 1 sparse Matrix of class "dgCMatrix"  
## s0  
## (Intercept) -0.4016480184  
## (Intercept) .   
## SEX 0.1309947917  
## AGEG.F7 -0.0845307099  
## X\_RACEGR3 0.0397922864  
## EDUCAL 0.0884053764  
## X\_INCOMG .   
## X\_RFBMI5 -0.0347581104  
## SMOKE100 0.1271695364  
## COPD -0.0412685488  
## EMPHY .   
## DEPRESS 0.0067054721  
## BRONCH -0.1066137715  
## SYMP.30D -0.0073182969  
## DUR.30D .   
## ASLEEP30 -0.0018297498  
## SYMPFREE -0.0016882384  
## INCINDT 0.6193944347  
## LAST.MD -0.0015410918  
## LAST.MED -0.1564942956  
## LAST.SYMP 0.0230722354  
## EPIS.12M -0.1378563702  
## DUR.ASTH .   
## COMPASTH .   
## INS1 -0.0328559622  
## ER.TIMES -0.0001644335  
## ER.VISIT -0.0044621380  
## URG.TIMES -0.0001831538  
## HOSP.VST .   
## HOSPTIME -0.0012376466  
## ASRXCOST -0.0361901490  
## WORKTALK -0.1525769397

#### Confusion Matrix with best lambda

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 632 366  
## 1 72 103  
##   
## Accuracy : 0.6266   
## 95% CI : (0.5982, 0.6544)  
## No Information Rate : 0.6002   
## P-Value [Acc > NIR] : 0.03416   
##   
## Kappa : 0.1311   
##   
## Mcnemar's Test P-Value : < 2e-16   
##   
## Sensitivity : 0.21962   
## Specificity : 0.89773   
## Pos Pred Value : 0.58857   
## Neg Pred Value : 0.63327   
## Prevalence : 0.39983   
## Detection Rate : 0.08781   
## Detection Prevalence : 0.14919   
## Balanced Accuracy : 0.55867   
##   
## 'Positive' Class : 1   
##

##### Calculating the AICc of Ridge and Lasso Models

it <- glmnet(x, y, family = “multinomial”)

tLL <- fitdf n <- fit$nobs AICc <- -tLL+2*k+2*k\*(k+1)/(n-k-1) AICc

## [1] -288.8758

## [1] -310.9217

#### Partial Least Squared

#### Confusion Matrix with best lambda

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 526 298  
## 1 165 183  
##   
## Accuracy : 0.6049   
## 95% CI : (0.5763, 0.6331)  
## No Information Rate : 0.5896   
## P-Value [Acc > NIR] : 0.1493   
##   
## Kappa : 0.1479   
##   
## Mcnemar's Test P-Value : 8.54e-10   
##   
## Sensitivity : 0.3805   
## Specificity : 0.7612   
## Pos Pred Value : 0.5259   
## Neg Pred Value : 0.6383   
## Prevalence : 0.4104   
## Detection Rate : 0.1561   
## Detection Prevalence : 0.2969   
## Balanced Accuracy : 0.5708   
##   
## 'Positive' Class : 1   
##

#### Here we train the model with partial least square using tune parameter.

## Partial Least Squares   
##   
## 4694 samples  
## 30 predictor  
## 2 classes: 'F', 'T'   
##   
## Pre-processing: centered (286), scaled (286)   
## Resampling: Cross-Validated (10 fold, repeated 3 times)   
## Summary of sample sizes: 4224, 4224, 4226, 4225, 4224, 4225, ...   
## Resampling results across tuning parameters:  
##   
## ncomp ROC Sens Spec   
## 1 0.6234314 0.8042955 0.3552866  
## 2 0.6378337 0.7938053 0.3703287  
## 3 0.6384305 0.7866914 0.3777634  
## 4 0.6357423 0.7880191 0.3713740  
## 5 0.6359938 0.7810312 0.3789742  
## 6 0.6383510 0.7823541 0.3800122  
## 7 0.6404820 0.7834380 0.3841672  
## 8 0.6407412 0.7837994 0.3876268  
## 9 0.6401770 0.7839198 0.3853807  
## 10 0.6397273 0.7827160 0.3850299  
## 11 0.6392595 0.7831990 0.3843435  
## 12 0.6391967 0.7824757 0.3838254  
## 13 0.6391284 0.7836800 0.3864152  
## 14 0.6390115 0.7840431 0.3858970  
## 15 0.6389724 0.7839224 0.3852035  
##   
## ROC was used to select the optimal model using the largest value.  
## The final value used for the model was ncomp = 8.

#### Confusion Matrix with best lambda

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 526 298  
## 1 165 183  
##   
## Accuracy : 0.6049   
## 95% CI : (0.5763, 0.6331)  
## No Information Rate : 0.5896   
## P-Value [Acc > NIR] : 0.1493   
##   
## Kappa : 0.1479   
##   
## Mcnemar's Test P-Value : 8.54e-10   
##   
## Sensitivity : 0.3805   
## Specificity : 0.7612   
## Pos Pred Value : 0.5259   
## Neg Pred Value : 0.6383   
## Prevalence : 0.4104   
## Detection Rate : 0.1561   
## Detection Prevalence : 0.2969   
## Balanced Accuracy : 0.5708   
##   
## 'Positive' Class : 1   
##

### SELECT MODELS

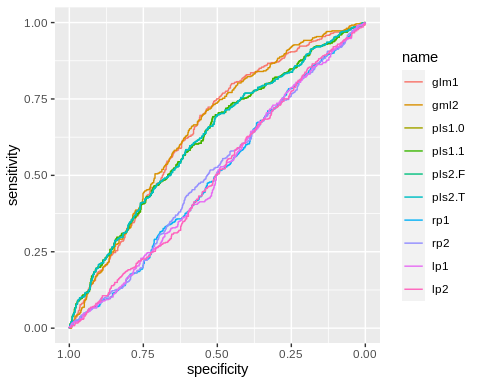
#### We compare the models with the accuray, precision, sensitivity, specificity, and F1 score from the confusion matrix

## glm.mod11 glm.mod12 ridge.mod1 ridge.mod2 lasso.mod1 lasso.mod2  
## Accuracy 0.6049488 0.6151877 0.6104007 0.6001705 0.6206309 0.6265985  
## Precision 0.5264706 0.5443787 0.6500000 NA 0.5451128 0.5885714  
## Sensitivity 0.3721414 0.3825364 0.0554371 0.0000000 0.3091684 0.2196162  
## Specificity 0.7670043 0.7771346 0.9801136 1.0000000 0.8281250 0.8977273  
## F1 0.4360536 0.4493284 0.1021611 NA 0.3945578 0.3198758  
## pls.mod1 pls.mod2  
## Accuracy 0.6049488 0.6049488  
## Precision 0.5258621 0.5258621  
## Sensitivity 0.3804574 0.3804574  
## Specificity 0.7612156 0.7612156  
## F1 0.4414958 0.4414958

With precision and specificity equal to 1, the ridge.mod2 model is overfitting. But glm.mod12 has the best accuracy, precision, sensivity, and specificity.

### Using pROC package.

We can plot the ROC curve and extract the AUC value.

 The glm model has the best Area Under the Curve.

### We run the glm model with the entire dataset

#### Second glm model

##   
## Call:  
## glm(formula = TARGET ~ SEX + AGEG.F7 + X\_RACEGR3 + EDUCAL + X\_INCOMG +   
## BRONCH + DUR.30D + INCINDT + LAST.MD + LAST.MED + LAST.SYMP +   
## COMPASTH + WORKTALK, family = binomial, data = asth.mgt.ad.min35)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.7951 -1.0128 -0.7147 1.1757 2.5382   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.136969 0.780112 0.176 0.860627   
## SEX2 0.189189 0.063119 2.997 0.002723 \*\*   
## AGEG.F72 0.021781 0.173823 0.125 0.900281   
## AGEG.F73 0.034508 0.167562 0.206 0.836835   
## AGEG.F74 -0.039075 0.156449 -0.250 0.802769   
## AGEG.F75 -0.277709 0.152909 -1.816 0.069344 .   
## AGEG.F76 -0.484091 0.155418 -3.115 0.001841 \*\*   
## AGEG.F77 -0.509155 0.169461 -3.005 0.002660 \*\*   
## X\_RACEGR32 0.439750 0.124452 3.533 0.000410 \*\*\*  
## X\_RACEGR33 0.014363 0.153340 0.094 0.925375   
## X\_RACEGR34 0.258029 0.148371 1.739 0.082020 .   
## X\_RACEGR35 0.322545 0.113626 2.839 0.004530 \*\*   
## X\_RACEGR39 0.143035 0.281651 0.508 0.611563   
## EDUCAL2 -0.559113 0.754273 -0.741 0.458535   
## EDUCAL3 -0.518114 0.743301 -0.697 0.485775   
## EDUCAL4 -0.525467 0.736426 -0.714 0.475514   
## EDUCAL5 -0.395211 0.736645 -0.537 0.591612   
## EDUCAL6 -0.358741 0.737092 -0.487 0.626472   
## EDUCAL9 0.918241 1.075043 0.854 0.393026   
## X\_INCOMG2 0.023649 0.096737 0.244 0.806872   
## X\_INCOMG3 0.018788 0.118975 0.158 0.874522   
## X\_INCOMG4 -0.002100 0.109835 -0.019 0.984747   
## X\_INCOMG5 0.149022 0.093661 1.591 0.111590   
## X\_INCOMG9 -0.099911 0.118132 -0.846 0.397687   
## BRONCH2 -0.155023 0.063188 -2.453 0.014154 \*   
## BRONCH7 -0.765089 0.339698 -2.252 0.024306 \*   
## BRONCH9 1.179099 1.219037 0.967 0.333425   
## DUR.30D12 -0.019133 0.084480 -0.226 0.820829   
## DUR.30D2 -0.276521 0.103822 -2.663 0.007735 \*\*   
## DUR.30D7 -0.382081 0.435759 -0.877 0.380585   
## INCINDT2 0.242276 0.211146 1.147 0.251203   
## INCINDT3 0.993496 0.185423 5.358 8.42e-08 \*\*\*  
## LAST.MD5 -0.094399 0.090328 -1.045 0.295991   
## LAST.MD6 -0.628298 0.165091 -3.806 0.000141 \*\*\*  
## LAST.MD7 -0.673877 0.163405 -4.124 3.72e-05 \*\*\*  
## LAST.MD88 -0.691748 0.578855 -1.195 0.232077   
## LAST.MED2 -0.395747 0.091363 -4.332 1.48e-05 \*\*\*  
## LAST.MED3 -0.568700 0.087858 -6.473 9.61e-11 \*\*\*  
## LAST.MED4 -0.606226 0.140052 -4.329 1.50e-05 \*\*\*  
## LAST.MED5 -0.778220 0.176073 -4.420 9.88e-06 \*\*\*  
## LAST.MED6 -0.539057 0.240229 -2.244 0.024837 \*   
## LAST.MED7 -0.922157 0.200288 -4.604 4.14e-06 \*\*\*  
## LAST.SYMP2 0.171761 0.081552 2.106 0.035190 \*   
## LAST.SYMP3 0.266576 0.080462 3.313 0.000923 \*\*\*  
## COMPASTH11 -0.311425 0.089904 -3.464 0.000532 \*\*\*  
## COMPASTH2 -0.249139 0.117608 -2.118 0.034142 \*   
## COMPASTH3 -0.080845 0.088299 -0.916 0.359888   
## COMPASTH4 -1.334457 1.129875 -1.181 0.237577   
## WORKTALK2 -0.609217 0.066038 -9.225 < 2e-16 \*\*\*  
## WORKTALK6 -0.864391 0.224236 -3.855 0.000116 \*\*\*  
## WORKTALK7 -0.818453 0.391397 -2.091 0.036518 \*   
## WORKTALK8 0.684934 0.940840 0.728 0.466612   
## WORKTALK9 -0.003068 1.205505 -0.003 0.997969   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 7943.8 on 5865 degrees of freedom  
## Residual deviance: 7428.4 on 5813 degrees of freedom  
## AIC: 7534.4  
##   
## Number of Fisher Scoring iterations: 4

## (Intercept) SEX2 AGEG.F72 AGEG.F73 AGEG.F74 AGEG.F75   
## 1.1467930 1.2082698 1.0220199 1.0351106 0.9616782 0.7575174   
## AGEG.F76 AGEG.F77 X\_RACEGR32 X\_RACEGR33 X\_RACEGR34 X\_RACEGR35   
## 0.6162574 0.6010030 1.5523191 1.0144663 1.2943762 1.3806364   
## X\_RACEGR39 EDUCAL2 EDUCAL3 EDUCAL4 EDUCAL5 EDUCAL6   
## 1.1537700 0.5717158 0.5956428 0.5912790 0.6735380 0.6985553   
## EDUCAL9 X\_INCOMG2 X\_INCOMG3 X\_INCOMG4 X\_INCOMG5 X\_INCOMG9   
## 2.5048801 1.0239306 1.0189658 0.9979024 1.1606990 0.9049179   
## BRONCH2 BRONCH7 BRONCH9 DUR.30D12 DUR.30D2 DUR.30D7   
## 0.8563958 0.4652923 3.2514425 0.9810490 0.7584175 0.6824395   
## INCINDT2 INCINDT3 LAST.MD5 LAST.MD6 LAST.MD7 LAST.MD88   
## 1.2741456 2.7006599 0.9099194 0.5334992 0.5097286 0.5007002   
## LAST.MED2 LAST.MED3 LAST.MED4 LAST.MED5 LAST.MED6 LAST.MED7   
## 0.6731772 0.5662612 0.5454054 0.4592228 0.5832983 0.3976603   
## LAST.SYMP2 LAST.SYMP3 COMPASTH11 COMPASTH2 COMPASTH3 COMPASTH4   
## 1.1873943 1.3054868 0.7324023 0.7794713 0.9223370 0.2633012   
## WORKTALK2 WORKTALK6 WORKTALK7 WORKTALK8 WORKTALK9   
## 0.5437762 0.4213080 0.4411134 1.9836400 0.9969364