DATA 621 Final Project

Farhana Zahir, Vijaya Cherukuri, Scott Reed, Shovon Biswas, Habib Khan, Alain Kuiete Tchoupou

11/18/2020

## OVERVIEW

The self-management of asthma help improve patient health. Asthma self-management provide to the patient and caregivers the 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 In this project, we study the characteristics that influence asthma self-management.

## [1] 13922 899

The data set come from CDC with ulr = “<https://www.cdc.gov/brfss/acbs/2016_documentation.html>”. It is a survey study. The download file is “2016 ACBS Adult Data SAS [ZIP – 3.10 MB]” The unzip file has 899 variables and 13,922 cases. We have selected the variables to use on our studies.

## EXPLORATORY DATA ANALYSIS

Meaning of variables used in the dataset

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”

HOSPPLAN = “HOSPITAL FOLLOW-UP”

HOSPTIME = “NUMBER OF HOSPITAL VISITS”

HOSP\_VST = “HOSPITAL VISIT”

QSTLANG\_F = “LANGUAGE IDENTIFIER”

SCR\_MED3 = “HAVE ALL THE MEDICATIONS”

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”

LASTSYMP = “LAST HAD ANY SYMPTOMS OF ASTHMA”

LAST\_MD = “LAST TALKED TO A DOCTOR”

LAST\_MED = “LAST TOOK ASTHMA MEDICATION”

COMPASTH = “TYPICAL ATTACK”

### Constructing the Data Frame by Selecting variables

We select all possible variable that we can use in our dataset. We also start to clean the dataset

## TCH.SIGN TCH.RESP TCH.MON MGT.PLAN MGT.CLAS INHALERW MOD.ENV SEX   
## 1:8639 1:9936 1:5694 1:3814 1: 1190 1:9382 1:4401 1:4786   
## 2:4972 2:3669 2:8014 2:9759 2:12675 2:2925 2:9407 2:9136   
## 7: 301 7: 299 7: 195 7: 336 7: 53 5: 386 7: 110   
## 9: 10 9: 18 9: 19 9: 13 9: 4 6: 762 9: 4   
## 7: 466   
## 9: 1   
##   
## AGEG.F7 X\_RACEGR3 EDUCAL X\_INCOMG X\_RFBMI5 SMOKE100 COPD   
## 1: 763 1:10806 1: 24 1:1910 1:3640 1:6546 1 : 2686   
## 2:1186 2: 766 2: 350 2:2264 2:9647 2:7312 2 :11027   
## 3:1419 3: 513 3: 722 3:1272 9: 635 7: 59 7 : 149   
## 4:2149 4: 459 4:3307 4:1659 9: 5 9 : 16   
## 5:3476 5: 1209 5:4177 5:5265 NA's: 44   
## 6:3199 9: 169 6:5321 9:1552   
## 7:1730 9: 21   
## EMPHY DEPRESS BRONCH DUR.30D INCINDT LAST.MD   
## 1 : 1133 1 :5185 1 : 3670 12 :4735 1: 316 4 :7977   
## 2 :12638 2 :8628 2 :10027 6 :4351 2: 1002 5 :1820   
## 7 : 93 7 : 36 7 : 166 10 :1618 3:12549 6 : 819   
## 9 : 14 9 : 29 9 : 15 1 :1288 7: 44 7 :3025   
## NA's: 44 NA's: 44 NA's: 44 2 : 899 9: 11 77: 131   
## 11 : 749 88: 135   
## (Other): 282 99: 15   
## LAST.MED LAST.SYMP EPIS.12M COMPASTH INS1 INS2   
## 1 :4924 1 :3567 1:5210 11 :4361 1:13121 1: 683   
## 7 :2978 7 :2616 2:4237 6 :4351 2: 767 2:12415   
## 3 :1555 3 :2515 6:4351 3 :3205 7: 26 5: 767   
## 4 :1238 4 :1618 7: 117 1 :1169 9: 8 7: 46   
## 5 :1072 2 :1613 9: 7 2 : 781 9: 11   
## 2 :1059 5 :1113 7 : 41   
## (Other):1096 (Other): 880 (Other): 14   
## ER.VISIT HOSP.VST ASMDCOST ASRXCOST ASSPCOST WORKTALK   
## 1:1347 1: 380 1 : 770 1 :1559 1 : 506 1 : 2592   
## 2:6700 2:6702 2 :10370 2 :9592 2 :10647 2 :10851   
## 5:2737 4: 979 5 : 2736 5 :2736 5 : 2736 6 : 281   
## 6:3105 5:2737 7 : 31 7 : 18 7 : 15 7 : 144   
## 7: 32 6:3105 9 : 10 9 : 12 9 : 13 8 : 31   
## 9: 1 7: 19 NA's: 5 NA's: 5 NA's: 5 9 : 11   
## NA's: 12

Here we collapse certain variables with to many classes, and factors with few cases.

asthma.mgt.adult2 <- asthma.mgt.adult2 %>%  
 mutate(EDUCAL = fct\_collapse(EDUCAL,  
 "1" = "1",  
 "2" = "2",  
 "3" = "3",  
 "4" = "4",  
 "5" = "5",  
 "6" = c("6", "9")  
 ),  
 LAST.MD = fct\_collapse(LAST.MD,  
 "4" = "4",  
 "5" = "5",  
 "6" = "6",  
 "7" = "7",  
 "9" = c("77", "88", "99"),  
 ),  
 LAST.MED = fct\_collapse(LAST.MED,  
 "4" = "1",  
 "5" = "2",  
 "6" = "3",  
 "7" = "4",  
 "9" = c("77", "88", "99"),  
 ),  
 INCINDT = fct\_collapse(INCINDT,  
 "1" = "1",  
 "2" = "2",  
 "3" = "3",  
 "7" = c("7", "9")  
 ),  
 LAST.SYMP = fct\_collapse(LAST.SYMP,  
 "1" = "1",  
 "2" = "2",  
 "3" = "3",  
 "4" = "4",  
 "5" = "5",  
 "7" = "7",  
 "9" = c("77", "88", "99")  
 ),  
 DUR.30D = fct\_collapse(DUR.30D,  
 "1" = "1",  
 "2" = "2",  
 "6" = "6",  
 "7" = c("7", "9","77" ,"99"),  
 "10" = "10",  
 "11" = "11",  
 "12" = "12"  
 ),  
 EPIS.12M = fct\_collapse(EPIS.12M,  
 "1" = "1",  
 "2" = "2",  
 "6" = "6",  
 "7" = c("7", "9")  
 ),  
 ER.VISIT = fct\_collapse(ER.VISIT,  
 "1" = "1",  
 "2" = "2",  
 "5" = c("5", "7", "9"),  
 "7" = c("7", "9")  
 ),  
 COMPASTH = fct\_collapse(COMPASTH,  
 "1" = "1",  
 "2" = "2",  
 "3" = c("3","4"),  
 "6" = "6",  
 "7" = c("7", "9"),  
 "11" = "11"  
 ),  
 )  
summary(asthma.mgt.adult2)

## TCH.SIGN TCH.RESP TCH.MON MGT.PLAN MGT.CLAS INHALERW MOD.ENV SEX   
## 1:8639 1:9936 1:5694 1:3814 1: 1190 1:9382 1:4401 1:4786   
## 2:4972 2:3669 2:8014 2:9759 2:12675 2:2925 2:9407 2:9136   
## 7: 301 7: 299 7: 195 7: 336 7: 53 5: 386 7: 110   
## 9: 10 9: 18 9: 19 9: 13 9: 4 6: 762 9: 4   
## 7: 466   
## 9: 1   
##   
## AGEG.F7 X\_RACEGR3 EDUCAL X\_INCOMG X\_RFBMI5 SMOKE100 COPD   
## 1: 763 1:10806 1: 24 1:1910 1:3640 1:6546 1 : 2686   
## 2:1186 2: 766 2: 350 2:2264 2:9647 2:7312 2 :11027   
## 3:1419 3: 513 3: 722 3:1272 9: 635 7: 59 7 : 149   
## 4:2149 4: 459 4:3307 4:1659 9: 5 9 : 16   
## 5:3476 5: 1209 5:4177 5:5265 NA's: 44   
## 6:3199 9: 169 6:5342 9:1552   
## 7:1730   
## EMPHY DEPRESS BRONCH DUR.30D INCINDT LAST.MD LAST.MED  
## 1 : 1133 1 :5185 1 : 3670 1 :1288 1: 316 4:7977 4:4924   
## 2 :12638 2 :8628 2 :10027 10:1618 2: 1002 5:1820 5:2131   
## 7 : 93 7 : 36 7 : 166 11: 749 3:12549 6: 819 6:2136   
## 9 : 14 9 : 29 9 : 15 12:4735 7: 55 7:3025 7:4216   
## NA's: 44 NA's: 44 NA's: 44 2 : 899 9: 281 9: 515   
## 6 :4351   
## 7 : 282   
## LAST.SYMP EPIS.12M COMPASTH INS1 INS2 ER.VISIT HOSP.VST  
## 1 :3567 1:5210 1 :1169 1:13121 1: 683 1:1347 1: 380   
## 7 :2616 2:4237 11:4361 2: 767 2:12415 2:6700 2:6702   
## 3 :2515 6:4351 2 : 781 7: 26 5: 767 5:2737 4: 979   
## 4 :1618 7: 124 3 :3217 9: 8 7: 46 6:3105 5:2737   
## 2 :1613 6 :4351 9: 11 7: 33 6:3105   
## 5 :1113 7 : 43 7: 19   
## (Other): 880   
## ASMDCOST ASRXCOST ASSPCOST WORKTALK   
## 1 : 770 1 :1559 1 : 506 1 : 2592   
## 2 :10370 2 :9592 2 :10647 2 :10851   
## 5 : 2736 5 :2736 5 : 2736 6 : 281   
## 7 : 31 7 : 18 7 : 15 7 : 144   
## 9 : 10 9 : 12 9 : 13 8 : 31   
## NA's: 5 NA's: 5 NA's: 5 9 : 11   
## NA's: 12

## [1] 11494 33

## Structure of the data

## 'data.frame': 11494 obs. of 33 variables:  
## $ TCH.SIGN : num 1 2 1 2 2 1 1 2 1 2 ...  
## ..- attr(\*, "label")= chr "EVER TAUGHT RECOGNIZE EARLY SIGN OR SYMPTOMS"  
## ..- attr(\*, "format.sas")= chr "TCH\_SIGN"  
## $ TCH.RESP : num 1 1 1 2 1 1 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 2 2 1 1 2 2 2 ...  
## ..- attr(\*, "label")= chr "EVER TAUGHT HOW TO USE A PEAK FLOW"  
## ..- attr(\*, "format.sas")= chr "TCH\_MON"  
## $ MGT.PLAN : num 2 2 2 2 2 2 1 2 2 2 ...  
## ..- 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 2 ...  
## ..- attr(\*, "label")= chr "EVER TAKEN A COURSE TO MANAGE ASTHMA"  
## ..- attr(\*, "format.sas")= chr "MGT\_CLAS"  
## $ INHALERW : num 2 2 1 1 1 1 1 1 1 1 ...  
## ..- attr(\*, "label")= chr "INHALER USE WATCHED"  
## ..- attr(\*, "format.sas")= chr "INHALERW"  
## $ MOD.ENV : num 2 2 2 2 1 2 2 2 1 2 ...  
## ..- attr(\*, "label")= chr "EVER ADVISED CHANGE THINGS IN YOUR HOME"  
## ..- attr(\*, "format.sas")= chr "MOD\_ENV"  
## $ SEX : num 1 2 2 2 2 2 1 2 2 2 ...  
## ..- attr(\*, "label")= chr "RESPONDENTS SEX"  
## ..- attr(\*, "format.sas")= chr "SEX"  
## $ AGEG.F7 : num 4 5 5 3 6 5 4 6 6 7 ...  
## ..- attr(\*, "label")= chr "AGE COLLAPSED TO 7 GROUPS FOR ASTHMA CALL-BACK"  
## ..- attr(\*, "format.sas")= chr "AGEG\_F7Z"  
## $ X\_RACEGR3: num 3 1 1 5 1 5 1 1 1 1 ...  
## ..- attr(\*, "label")= chr "COMPUTED FIVE LEVEL RACE/ETHNICITY CATEGORY."  
## ..- attr(\*, "format.sas")= chr "\_3RACEGR"  
## $ EDUCAL : num 6 4 4 5 6 6 6 6 6 5 ...  
## ..- attr(\*, "label")= chr "EDUCATION LEVEL"  
## ..- attr(\*, "format.sas")= chr "EDUCA"  
## $ X\_INCOMG : num 5 1 1 5 5 5 5 5 3 9 ...  
## ..- attr(\*, "label")= chr "COMPUTED INCOME CATEGORIES"  
## ..- attr(\*, "format.sas")= chr "\_INCOMG"  
## $ X\_RFBMI5 : num 2 2 2 2 2 2 1 2 2 1 ...  
## ..- attr(\*, "label")= chr "OVERWEIGHT OR OBESE CALCULATED VARIABLE"  
## ..- attr(\*, "format.sas")= chr "\_5RFBMI"  
## $ SMOKE100 : num 2 1 1 2 1 2 1 1 2 2 ...  
## ..- attr(\*, "label")= chr "SMOKED AT LEAST 100 CIGARETTES"  
## ..- attr(\*, "format.sas")= chr "SMOK100\_"  
## $ COPD : num 2 1 2 2 2 2 2 2 2 1 ...  
## ..- 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 2 2 ...  
## ..- attr(\*, "label")= chr "EVER TOLD HAVE EMPHYSEMA"  
## ..- attr(\*, "format.sas")= chr "EMPHY"  
## $ DEPRESS : num 2 1 2 2 2 2 2 2 1 1 ...  
## ..- attr(\*, "label")= chr "EVER TOLD DEPRESSED"  
## ..- attr(\*, "format.sas")= chr "DEPRESS"  
## $ BRONCH : num 2 1 2 2 1 2 2 2 1 2 ...  
## ..- attr(\*, "label")= chr "EVER TOLD HAVE CHRONIC BRONCHITIS"  
## ..- attr(\*, "format.sas")= chr "BRONCH"  
## $ DUR.30D : num 10 2 12 6 12 10 12 6 1 6 ...  
## ..- attr(\*, "label")= chr "CONSTANT SYMPTOMS"  
## ..- attr(\*, "format.sas")= chr "DUR\_30D"  
## $ INCINDT : num 3 2 3 3 3 2 3 3 3 3 ...  
## ..- attr(\*, "label")= chr "TIME SINCE DIAGNOSIS"  
## ..- attr(\*, "format.sas")= chr "INCIDNT"  
## $ LAST.MD : num 5 4 4 7 4 4 4 5 4 5 ...  
## ..- attr(\*, "label")= chr "LAST TALKED TO A DOCTOR"  
## ..- attr(\*, "format.sas")= chr "LAST\_MD"  
## $ LAST.MED : num 4 1 3 7 3 1 1 6 1 5 ...  
## ..- attr(\*, "label")= chr "LAST TOOK ASTHMA MEDICATION"  
## ..- attr(\*, "format.sas")= chr "LAST\_MED"  
## $ LAST.SYMP: num 4 1 3 7 3 4 3 5 1 5 ...  
## ..- attr(\*, "label")= chr "LAST HAD ANY SYMPTOMS OF ASTHMA"  
## ..- attr(\*, "format.sas")= chr "LASTSYMP"  
## $ EPIS.12M : num 1 1 1 6 1 2 1 6 2 6 ...  
## ..- attr(\*, "label")= chr "ASTHMA EPISODE OR ATTACK"  
## ..- attr(\*, "format.sas")= chr "EPIS\_12M"  
## $ COMPASTH : num 1 3 1 6 3 11 3 6 11 6 ...  
## ..- attr(\*, "label")= chr "TYPICAL ATTACK"  
## ..- attr(\*, "format.sas")= chr "COMPASTH"  
## $ INS1 : num 1 1 1 2 1 1 1 1 1 1 ...  
## ..- attr(\*, "label")= chr "INSURANCE"  
## ..- attr(\*, "format.sas")= chr "INS1Z"  
## $ INS2 : num 2 2 2 5 2 2 2 2 2 2 ...  
## ..- attr(\*, "label")= chr "INSURANCE OR COVERAGE GAP"  
## ..- attr(\*, "format.sas")= chr "INS2Z"  
## $ ER.VISIT : num 6 2 2 5 2 2 2 5 2 6 ...  
## ..- attr(\*, "label")= chr "EMERGENCY ROOM VISIT"  
## ..- attr(\*, "format.sas")= chr "ER\_VISIT"  
## $ HOSP.VST : num 6 2 2 5 2 2 2 5 2 6 ...  
## ..- attr(\*, "label")= chr "HOSPITAL VISIT"  
## ..- attr(\*, "format.sas")= chr "HOSP\_VST"  
## $ ASMDCOST : num 2 2 2 5 2 2 2 5 2 2 ...  
## ..- attr(\*, "label")= chr "COST BARRIER: PRIMARY CARE DOCTOR"  
## ..- attr(\*, "format.sas")= chr "ASMDCOST"  
## $ ASRXCOST : num 2 2 2 5 2 2 2 5 1 2 ...  
## ..- attr(\*, "label")= chr "COST BARRIER: MEDICATION"  
## ..- attr(\*, "format.sas")= chr "ASRXCOST"  
## $ ASSPCOST : num 2 2 2 5 2 2 2 5 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"

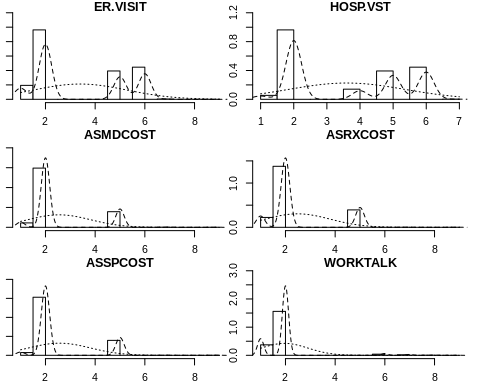
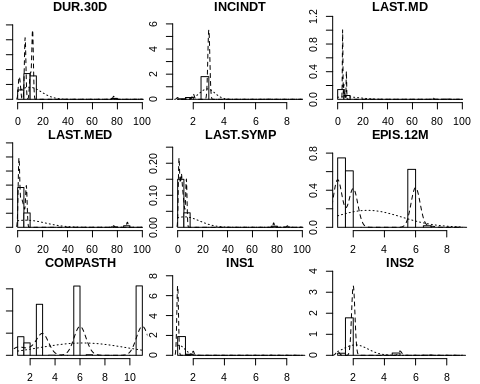
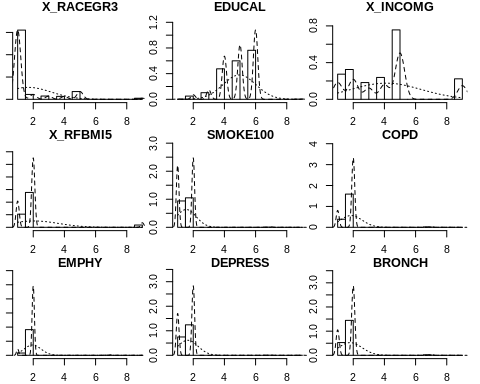
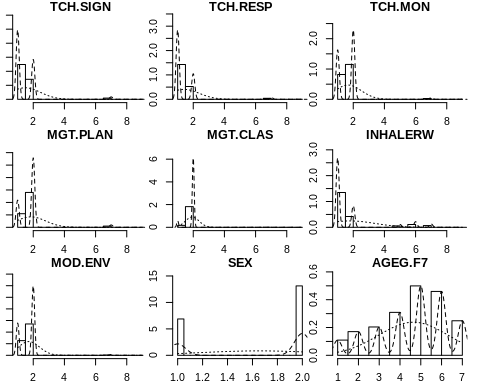
### Summary of the Data

## 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 :1.337 Mean :1.238 Mean :1.555 Mean :1.694   
## 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 AGEG.F7   
## Min. :1.000 Min. :1.00 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:2.000 1st Qu.:1.00 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:4.000   
## Median :2.000 Median :1.00 Median :2.000 Median :2.000 Median :5.000   
## Mean :1.906 Mean :1.24 Mean :1.662 Mean :1.678 Mean :4.594   
## 3rd Qu.:2.000 3rd Qu.:1.00 3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:6.000   
## Max. :2.000 Max. :2.00 Max. :2.000 Max. :2.000 Max. :7.000   
##   
## X\_RACEGR3 EDUCAL X\_INCOMG X\_RFBMI5 SMOKE100   
## Min. :1.000 Min. :1.00 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:1.000 1st Qu.:4.00 1st Qu.:2.000 1st Qu.:1.000 1st Qu.:1.000   
## Median :1.000 Median :5.00 Median :4.000 Median :2.000 Median :2.000   
## Mean :1.653 Mean :4.98 Mean :4.057 Mean :2.062 Mean :1.549   
## 3rd Qu.:1.000 3rd Qu.:6.00 3rd Qu.:5.000 3rd Qu.:2.000 3rd Qu.:2.000   
## Max. :9.000 Max. :9.00 Max. :9.000 Max. :9.000 Max. :9.000   
##   
## COPD EMPHY DEPRESS BRONCH   
## Min. :1.00 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:2.00 1st Qu.:2.000 1st Qu.:1.000 1st Qu.:1.000   
## Median :2.00 Median :2.000 Median :2.000 Median :2.000   
## Mean :1.85 Mean :1.945 Mean :1.633 Mean :1.769   
## 3rd Qu.:2.00 3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:2.000   
## Max. :9.00 Max. :9.000 Max. :9.000 Max. :9.000   
## NA's :30 NA's :30 NA's :30 NA's :30   
## DUR.30D INCINDT LAST.MD LAST.MED   
## Min. : 1.000 Min. :1.000 Min. : 4.000 Min. : 1.000   
## 1st Qu.: 6.000 1st Qu.:3.000 1st Qu.: 4.000 1st Qu.: 1.000   
## Median :10.000 Median :3.000 Median : 4.000 Median : 3.000   
## Mean : 9.201 Mean :2.892 Mean : 5.729 Mean : 3.733   
## 3rd Qu.:12.000 3rd Qu.:3.000 3rd Qu.: 5.000 3rd Qu.: 5.000   
## Max. :99.000 Max. :9.000 Max. :99.000 Max. :99.000   
##   
## LAST.SYMP EPIS.12M COMPASTH INS1   
## Min. : 1.000 Min. :1.000 Min. : 1.000 Min. :1.000   
## 1st Qu.: 1.000 1st Qu.:1.000 1st Qu.: 3.000 1st Qu.:1.000   
## Median : 3.000 Median :2.000 Median : 6.000 Median :1.000   
## Mean : 4.742 Mean :2.703 Mean : 6.126 Mean :1.065   
## 3rd Qu.: 5.000 3rd Qu.:6.000 3rd Qu.:11.000 3rd Qu.:1.000   
## Max. :99.000 Max. :9.000 Max. :11.000 Max. :9.000   
##   
## INS2 ER.VISIT HOSP.VST ASMDCOST   
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:2.000 1st Qu.:2.000 1st Qu.:2.000 1st Qu.:2.000   
## Median :2.000 Median :2.000 Median :2.000 Median :2.000   
## Mean :2.127 Mean :3.256 Mean :3.468 Mean :2.416   
## 3rd Qu.:2.000 3rd Qu.:5.000 3rd Qu.:5.000 3rd Qu.:2.000   
## Max. :9.000 Max. :9.000 Max. :7.000 Max. :9.000   
## NA's :4   
## ASRXCOST ASSPCOST WORKTALK   
## Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:2.000 1st Qu.:2.000 1st Qu.:2.000   
## Median :2.000 Median :2.000 Median :2.000   
## Mean :2.353 Mean :2.433 Mean :1.928   
## 3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:2.000   
## Max. :9.000 Max. :9.000 Max. :9.000   
## NA's :4 NA's :4 NA's :8

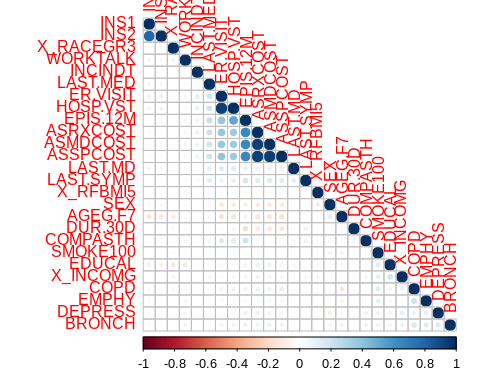
### Distribution of the Variables in the Data

#### Histograms

Histograms tell us how the data is distributed in the dataset (numeric fields).



### The correlations betweeen predictors



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] "DUR.30D" "INCINDT" "LAST.MD" "LAST.MED" "LAST.SYMP" "EPIS.12M"   
## [25] "COMPASTH" "INS1" "INS2" "ER.VISIT" "HOSP.VST" "ASRXCOST"   
## [31] "WORKTALK"

### CONSTRUCT THE RESPONSE VARIABLE

We first extract variables related to education, #### Selection of variables

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

#### Exploration of the clustering

## TCH.SIGN TCH.RESP TCH.MON MGT.PLAN MGT.CLAS INHALERW MOD.ENV   
## 1:7621 1:8760 1:5118 1:3522 1: 1078 1:8738 1:3889   
## 2:3873 2:2734 2:6376 2:7972 2:10416 2:2756 2:7605

#### Elbow method to find the number of clusters

We run kmeans with different clusters from 1 to 16 and we produce a scree plot to determine the number of cluster at the elbow. 

The number of cluster is 3

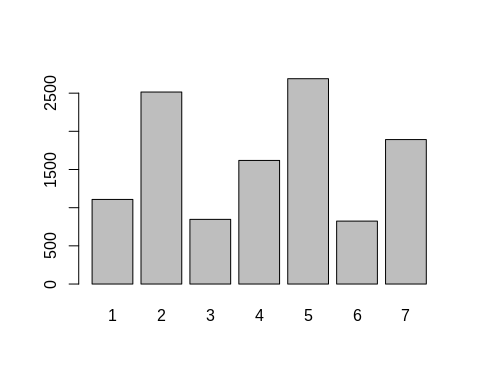
#### Now we do the clustering and extract the centers of resulting model

## TCH.SIGN TCH.RESP TCH.MON MGT.PLAN MGT.CLAS INHALERW MOD.ENV  
## 1 1.228339 1.057762 2.000000 1.935018 1.944043 1.251805 1.000000  
## 2 1.042545 1.022664 1.229423 1.000000 2.000000 1.073161 1.522465  
## 3 1.035419 1.021251 1.129870 1.216057 1.000000 1.066116 1.493506  
## 4 1.051266 1.013589 1.000000 2.000000 2.000000 1.140828 1.559605  
## 5 1.310152 1.040536 2.000000 1.948680 1.962068 1.289327 2.000000  
## 6 1.904126 1.694175 1.000000 1.916262 1.967233 1.348301 1.845874  
## 7 1.962474 2.000000 2.000000 1.966173 1.978858 1.498943 1.835624

#### We add the point classification to the original data

## TCH.SIGN TCH.RESP TCH.MON MGT.PLAN MGT.CLAS INHALERW MOD.ENV target  
## 1 1 1 2 2 2 2 2 5  
## 2 2 1 2 2 2 2 2 5  
## 3 1 1 2 2 2 1 2 5  
## 4 2 2 2 2 2 1 2 7  
## 5 2 1 2 2 2 1 1 1  
## 6 1 1 1 2 2 1 2 4

## TCH.SIGN TCH.RESP TCH.MON MGT.PLAN MGT.CLAS INHALERW MOD.ENV target   
## 1:7621 1:8760 1:5118 1:3522 1: 1078 1:8738 1:3889 1:1108   
## 2:3873 2:2734 2:6376 2:7972 2:10416 2:2756 2:7605 2:2515   
## 3: 847   
## 4:1619   
## 5:2689   
## 6: 824   
## 7:1892



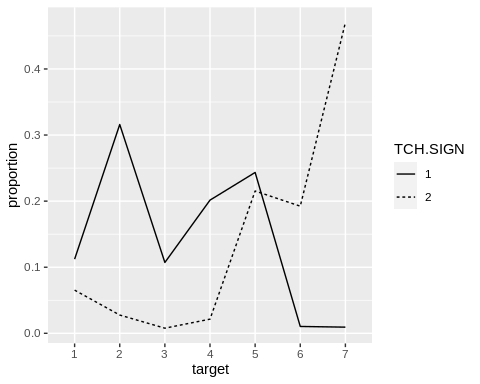
View of the clustering result

### Interpretation of the Selft-Management Response clustering

#### TCH.SIGN

## # A tibble: 14 x 5  
## TCH.SIGN target count etotal proportion  
## <fct> <fct> <int> <int> <dbl>  
## 1 1 1 855 7621 0.112   
## 2 1 2 2408 7621 0.316   
## 3 1 3 817 7621 0.107   
## 4 1 4 1536 7621 0.202   
## 5 1 5 1855 7621 0.243   
## 6 1 6 79 7621 0.0104   
## 7 1 7 71 7621 0.00932  
## 8 2 1 253 3873 0.0653   
## 9 2 2 107 3873 0.0276   
## 10 2 3 30 3873 0.00775  
## 11 2 4 83 3873 0.0214   
## 12 2 5 834 3873 0.215   
## 13 2 6 745 3873 0.192   
## 14 2 7 1821 3873 0.470

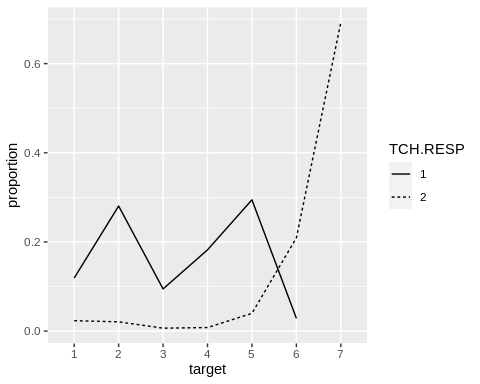
## # A tibble: 2 x 6  
## # Groups: target [2]  
## TCH.SIGN target count etotal proportion group.max  
## <fct> <fct> <int> <int> <dbl> <int>  
## 1 1 2 2408 7621 0.316 2408  
## 2 2 7 1821 3873 0.470 1821

 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: 13 x 5  
## TCH.RESP target count etotal proportion  
## <fct> <fct> <int> <int> <dbl>  
## 1 1 1 1044 8760 0.119   
## 2 1 2 2458 8760 0.281   
## 3 1 3 829 8760 0.0946   
## 4 1 4 1597 8760 0.182   
## 5 1 5 2580 8760 0.295   
## 6 1 6 252 8760 0.0288   
## 7 2 1 64 2734 0.0234   
## 8 2 2 57 2734 0.0208   
## 9 2 3 18 2734 0.00658  
## 10 2 4 22 2734 0.00805  
## 11 2 5 109 2734 0.0399   
## 12 2 6 572 2734 0.209   
## 13 2 7 1892 2734 0.692

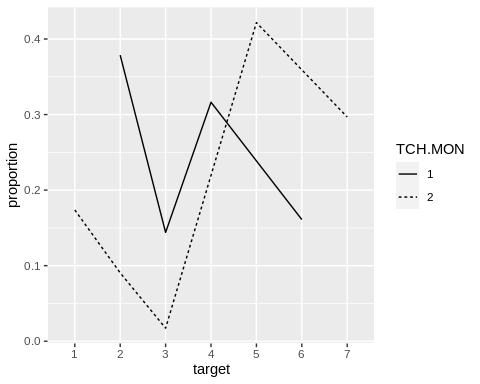
## # A tibble: 2 x 6  
## # Groups: target [2]  
## TCH.RESP target count etotal proportion group.max  
## <fct> <fct> <int> <int> <dbl> <int>  
## 1 1 5 2580 8760 0.295 2580  
## 2 2 7 1892 2734 0.692 1892

 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: 9 x 5  
## TCH.MON target count etotal proportion  
## <fct> <fct> <int> <int> <dbl>  
## 1 1 2 1938 5118 0.379   
## 2 1 3 737 5118 0.144   
## 3 1 4 1619 5118 0.316   
## 4 1 6 824 5118 0.161   
## 5 2 1 1108 6376 0.174   
## 6 2 2 577 6376 0.0905  
## 7 2 3 110 6376 0.0173  
## 8 2 5 2689 6376 0.422   
## 9 2 7 1892 6376 0.297

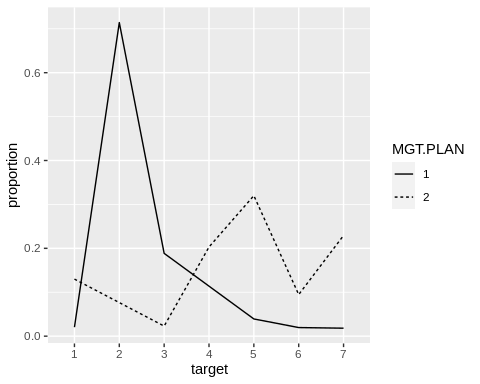
## # A tibble: 2 x 6  
## # Groups: target [2]  
## TCH.MON target count etotal proportion group.max  
## <fct> <fct> <int> <int> <dbl> <int>  
## 1 1 2 1938 5118 0.379 1938  
## 2 2 5 2689 6376 0.422 2689

 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: 12 x 5  
## MGT.PLAN target count etotal proportion  
## <fct> <fct> <int> <int> <dbl>  
## 1 1 1 72 3522 0.0204  
## 2 1 2 2515 3522 0.714   
## 3 1 3 664 3522 0.189   
## 4 1 5 138 3522 0.0392  
## 5 1 6 69 3522 0.0196  
## 6 1 7 64 3522 0.0182  
## 7 2 1 1036 7972 0.130   
## 8 2 3 183 7972 0.0230  
## 9 2 4 1619 7972 0.203   
## 10 2 5 2551 7972 0.320   
## 11 2 6 755 7972 0.0947  
## 12 2 7 1828 7972 0.229

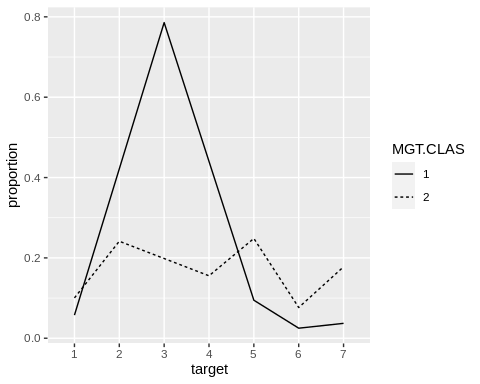
## # A tibble: 2 x 6  
## # Groups: target [2]  
## MGT.PLAN target count etotal proportion group.max  
## <fct> <fct> <int> <int> <dbl> <int>  
## 1 1 2 2515 3522 0.714 2515  
## 2 2 5 2551 7972 0.320 2551

 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: 11 x 5  
## MGT.CLAS target count etotal proportion  
## <fct> <fct> <int> <int> <dbl>  
## 1 1 1 62 1078 0.0575  
## 2 1 3 847 1078 0.786   
## 3 1 5 102 1078 0.0946  
## 4 1 6 27 1078 0.0250  
## 5 1 7 40 1078 0.0371  
## 6 2 1 1046 10416 0.100   
## 7 2 2 2515 10416 0.241   
## 8 2 4 1619 10416 0.155   
## 9 2 5 2587 10416 0.248   
## 10 2 6 797 10416 0.0765  
## 11 2 7 1852 10416 0.178

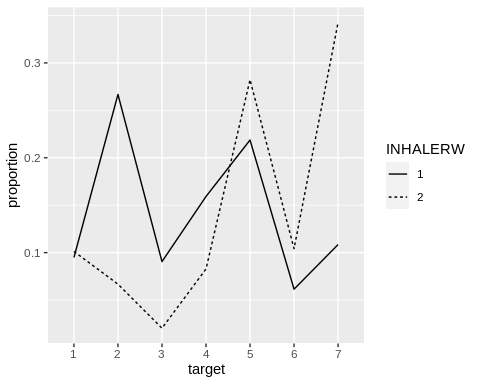
## # A tibble: 2 x 6  
## # Groups: target [2]  
## MGT.CLAS target count etotal proportion group.max  
## <fct> <fct> <int> <int> <dbl> <int>  
## 1 1 3 847 1078 0.786 847  
## 2 2 5 2587 10416 0.248 2587

 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: 14 x 5  
## INHALERW target count etotal proportion  
## <fct> <fct> <int> <int> <dbl>  
## 1 1 1 829 8738 0.0949  
## 2 1 2 2331 8738 0.267   
## 3 1 3 791 8738 0.0905  
## 4 1 4 1391 8738 0.159   
## 5 1 5 1911 8738 0.219   
## 6 1 6 537 8738 0.0615  
## 7 1 7 948 8738 0.108   
## 8 2 1 279 2756 0.101   
## 9 2 2 184 2756 0.0668  
## 10 2 3 56 2756 0.0203  
## 11 2 4 228 2756 0.0827  
## 12 2 5 778 2756 0.282   
## 13 2 6 287 2756 0.104   
## 14 2 7 944 2756 0.343

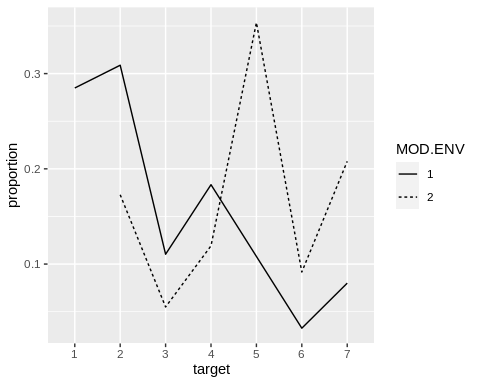
## # A tibble: 2 x 6  
## # Groups: target [2]  
## INHALERW target count etotal proportion group.max  
## <fct> <fct> <int> <int> <dbl> <int>  
## 1 1 2 2331 8738 0.267 2331  
## 2 2 7 944 2756 0.343 944

 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: 12 x 5  
## MOD.ENV target count etotal proportion  
## <fct> <fct> <int> <int> <dbl>  
## 1 1 1 1108 3889 0.285   
## 2 1 2 1201 3889 0.309   
## 3 1 3 429 3889 0.110   
## 4 1 4 713 3889 0.183   
## 5 1 6 127 3889 0.0327  
## 6 1 7 311 3889 0.0800  
## 7 2 2 1314 7605 0.173   
## 8 2 3 418 7605 0.0550  
## 9 2 4 906 7605 0.119   
## 10 2 5 2689 7605 0.354   
## 11 2 6 697 7605 0.0917  
## 12 2 7 1581 7605 0.208

## # 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 1201 3889 0.309 1201  
## 2 2 5 2689 7605 0.354 2689



#### 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 5 2 2 3 2 2   
## 2 2=NO 7 7 5 5 5 7 5

##### For the response variable TARGET, an excellent management skill has number 2 but a poor management skill has number 7 and 5.

##### We can build a logistics regression on the dataset.

### Here we remove the varibles used to calculate the target variable and reformat the data frame.

## 'data.frame': 11494 obs. of 25 variables:  
## $ TARGET : num 0 0 0 0 0 0 1 0 0 0 ...  
## $ SEX : num 1 2 2 2 2 2 1 2 2 2 ...  
## ..- attr(\*, "label")= chr "RESPONDENTS SEX"  
## ..- attr(\*, "format.sas")= chr "SEX"  
## $ AGEG.F7 : num 4 5 5 3 6 5 4 6 6 7 ...  
## ..- attr(\*, "label")= chr "AGE COLLAPSED TO 7 GROUPS FOR ASTHMA CALL-BACK"  
## ..- attr(\*, "format.sas")= chr "AGEG\_F7Z"  
## $ X\_RACEGR3: num 3 1 1 5 1 5 1 1 1 1 ...  
## ..- attr(\*, "label")= chr "COMPUTED FIVE LEVEL RACE/ETHNICITY CATEGORY."  
## ..- attr(\*, "format.sas")= chr "\_3RACEGR"  
## $ EDUCAL : num 6 4 4 5 6 6 6 6 6 5 ...  
## ..- attr(\*, "label")= chr "EDUCATION LEVEL"  
## ..- attr(\*, "format.sas")= chr "EDUCA"  
## $ X\_INCOMG : num 5 1 1 5 5 5 5 5 3 9 ...  
## ..- attr(\*, "label")= chr "COMPUTED INCOME CATEGORIES"  
## ..- attr(\*, "format.sas")= chr "\_INCOMG"  
## $ X\_RFBMI5 : num 2 2 2 2 2 2 1 2 2 1 ...  
## ..- attr(\*, "label")= chr "OVERWEIGHT OR OBESE CALCULATED VARIABLE"  
## ..- attr(\*, "format.sas")= chr "\_5RFBMI"  
## $ SMOKE100 : num 2 1 1 2 1 2 1 1 2 2 ...  
## ..- attr(\*, "label")= chr "SMOKED AT LEAST 100 CIGARETTES"  
## ..- attr(\*, "format.sas")= chr "SMOK100\_"  
## $ COPD : num 2 1 2 2 2 2 2 2 2 1 ...  
## ..- 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 2 2 ...  
## ..- attr(\*, "label")= chr "EVER TOLD HAVE EMPHYSEMA"  
## ..- attr(\*, "format.sas")= chr "EMPHY"  
## $ DEPRESS : num 2 1 2 2 2 2 2 2 1 1 ...  
## ..- attr(\*, "label")= chr "EVER TOLD DEPRESSED"  
## ..- attr(\*, "format.sas")= chr "DEPRESS"  
## $ BRONCH : num 2 1 2 2 1 2 2 2 1 2 ...  
## ..- attr(\*, "label")= chr "EVER TOLD HAVE CHRONIC BRONCHITIS"  
## ..- attr(\*, "format.sas")= chr "BRONCH"  
## $ DUR.30D : num 10 2 12 6 12 10 12 6 1 6 ...  
## ..- attr(\*, "label")= chr "CONSTANT SYMPTOMS"  
## ..- attr(\*, "format.sas")= chr "DUR\_30D"  
## $ INCINDT : num 3 2 3 3 3 2 3 3 3 3 ...  
## ..- attr(\*, "label")= chr "TIME SINCE DIAGNOSIS"  
## ..- attr(\*, "format.sas")= chr "INCIDNT"  
## $ LAST.MD : num 5 4 4 7 4 4 4 5 4 5 ...  
## ..- attr(\*, "label")= chr "LAST TALKED TO A DOCTOR"  
## ..- attr(\*, "format.sas")= chr "LAST\_MD"  
## $ LAST.MED : num 4 1 3 7 3 1 1 6 1 5 ...  
## ..- attr(\*, "label")= chr "LAST TOOK ASTHMA MEDICATION"  
## ..- attr(\*, "format.sas")= chr "LAST\_MED"  
## $ LAST.SYMP: num 4 1 3 7 3 4 3 5 1 5 ...  
## ..- attr(\*, "label")= chr "LAST HAD ANY SYMPTOMS OF ASTHMA"  
## ..- attr(\*, "format.sas")= chr "LASTSYMP"  
## $ EPIS.12M : num 1 1 1 6 1 2 1 6 2 6 ...  
## ..- attr(\*, "label")= chr "ASTHMA EPISODE OR ATTACK"  
## ..- attr(\*, "format.sas")= chr "EPIS\_12M"  
## $ COMPASTH : num 1 3 1 6 3 11 3 6 11 6 ...  
## ..- attr(\*, "label")= chr "TYPICAL ATTACK"  
## ..- attr(\*, "format.sas")= chr "COMPASTH"  
## $ INS1 : num 1 1 1 2 1 1 1 1 1 1 ...  
## ..- attr(\*, "label")= chr "INSURANCE"  
## ..- attr(\*, "format.sas")= chr "INS1Z"  
## $ INS2 : num 2 2 2 5 2 2 2 2 2 2 ...  
## ..- attr(\*, "label")= chr "INSURANCE OR COVERAGE GAP"  
## ..- attr(\*, "format.sas")= chr "INS2Z"  
## $ ER.VISIT : num 6 2 2 5 2 2 2 5 2 6 ...  
## ..- attr(\*, "label")= chr "EMERGENCY ROOM VISIT"  
## ..- attr(\*, "format.sas")= chr "ER\_VISIT"  
## $ HOSP.VST : num 6 2 2 5 2 2 2 5 2 6 ...  
## ..- attr(\*, "label")= chr "HOSPITAL VISIT"  
## ..- attr(\*, "format.sas")= chr "HOSP\_VST"  
## $ ASRXCOST : num 2 2 2 5 2 2 2 5 1 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"

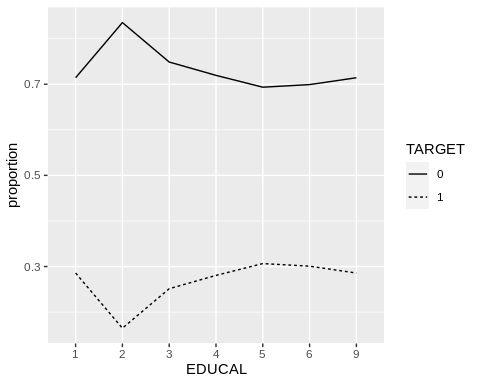
### PREPARE THE DATA FOR MODELISATION

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

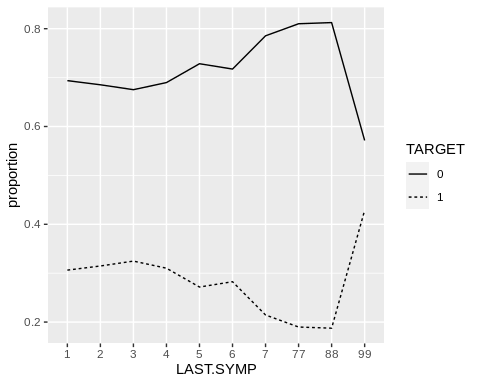
Here were are going to drop missing data because they are only 12 over 13,922 rows. We also transform all predictors to categorical.

## TARGET SEX AGEG.F7 X\_RACEGR3 EDUCAL X\_INCOMG X\_RFBMI5 SMOKE100  
## 0:8114 1:3686 1: 622 1:8935 1: 14 1:1583 1:2917 1:5401   
## 1:3350 2:7778 2: 996 2: 636 2: 255 2:1872 2:8027 2:6017   
## 3:1217 3: 420 3: 573 3:1042 9: 520 7: 43   
## 4:1869 4: 384 4:2702 4:1394 9: 3   
## 5:2924 5: 964 5:3514 5:4412   
## 6:2555 9: 125 6:4392 9:1161   
## 7:1281 9: 14   
## COPD EMPHY DEPRESS BRONCH DUR.30D INCINDT LAST.MD   
## 1:2311 1: 973 1:4428 1:3198 12 :4181 1: 259 4 :7077   
## 2:9038 2:10425 2:7000 2:8159 6 :3094 2: 864 5 :1563   
## 7: 106 7: 61 7: 18 7: 100 10 :1425 3:10307 6 : 673   
## 9: 9 9: 5 9: 18 9: 7 1 :1149 7: 28 7 :2014   
## 2 : 804 9: 6 77: 76   
## 11 : 613 88: 55   
## (Other): 198 99: 6   
## LAST.MED LAST.SYMP EPIS.12M COMPASTH INS1 INS2   
## 1 :4479 1 :3161 1:4724 11 :3646 1:10833 1: 561   
## 7 :2027 3 :2205 2:3565 6 :3094 2: 611 2:10257   
## 3 :1388 7 :1669 6:3094 3 :2915 7: 14 5: 611   
## 4 :1111 4 :1425 7: 77 1 :1060 9: 6 7: 27   
## 2 : 950 2 :1414 9: 4 2 : 703 9: 8   
## 5 : 931 5 : 924 7 : 33   
## (Other): 578 (Other): 666 (Other): 13   
## ER.VISIT HOSP.VST ASRXCOST WORKTALK  
## 1:1229 1: 345 1:1393 1:2305   
## 2:5896 2:5986 2:8279 2:8841   
## 5:1772 4: 808 5:1772 6: 219   
## 6:2546 5:1772 7: 13 7: 76   
## 7: 20 6:2546 9: 7 8: 15   
## 9: 1 7: 7 9: 8   
##

#### Visualization of some combine variables



Proportion of Good Skill Management in terme of Education Level



Proportion of Good skill management in terme of Duration of Asthma Attack

#### 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 AGEG.F75   
## 0.815274 0.376929 0.046073 0.110514 -0.009079 -0.224369   
## AGEG.F76 AGEG.F77 X\_RACEGR32 X\_RACEGR33 X\_RACEGR34 X\_RACEGR35   
## -0.355010 -0.545024 0.446352 -0.215877 0.210427 0.251571   
## X\_RACEGR39 EDUCAL2 EDUCAL3 EDUCAL4 EDUCAL5 EDUCAL6   
## 0.307141 -1.394910 -0.877344 -0.670978 -0.538958 -0.543019   
## EDUCAL9 X\_INCOMG2 X\_INCOMG3 X\_INCOMG4 X\_INCOMG5 X\_INCOMG9   
## -0.073281 0.024733 0.122684 -0.120826 -0.006362 -0.092898   
## X\_RFBMI52 X\_RFBMI59 SMOKE1002 SMOKE1007 SMOKE1009 COPD2   
## -0.077778 -0.057911 0.025179 0.463083 1.509425 -0.223620   
## COPD7 COPD9 EMPHY2 EMPHY7 EMPHY9 DEPRESS2   
## -0.432349 0.231118 -0.047004 -0.496181 -1.281916 0.131436   
## DEPRESS7 DEPRESS9 BRONCH2 BRONCH7 BRONCH9 DUR.30D10   
## -0.436594 0.399594 -0.096194 -0.245964 -1.653375 0.361595   
## DUR.30D11 DUR.30D12 DUR.30D2 DUR.30D6 DUR.30D7 DUR.30D77   
## -0.044773 -0.028588 -0.248560 -1.261388 -1.089232 -0.277066   
## DUR.30D9 DUR.30D99 INCINDT2 INCINDT3 INCINDT7 INCINDT9   
## -12.485576 -0.143906 0.070017 0.661739 -0.138097 0.811971   
## LAST.MD5 LAST.MD6 LAST.MD7 LAST.MD77 LAST.MD88 LAST.MD99   
## 0.288263 0.015621 -0.306763 -0.885122 0.129965 -0.405754   
## LAST.MED2 LAST.MED3 LAST.MED4 LAST.MED5 LAST.MED6 LAST.MED7   
## -0.175013 -0.370797 -0.531523 -0.449710 -0.425216 -0.267810   
## LAST.MED77 LAST.MED99 LAST.SYMP2 LAST.SYMP3 LAST.SYMP4 LAST.SYMP5   
## -0.689832 -14.116510 0.179669 0.237474 NA 1.234484   
## LAST.SYMP6 LAST.SYMP7 LAST.SYMP77 LAST.SYMP88 LAST.SYMP99 EPIS.12M2   
## 1.249424 1.085248 -0.175410 NA 1.712970 -0.399376   
## EPIS.12M6 EPIS.12M7 EPIS.12M9 COMPASTH11 COMPASTH2 COMPASTH3   
## NA -0.864829 0.500115 NA -0.302154 -0.228848   
## COMPASTH4 COMPASTH6 COMPASTH7 COMPASTH9 INS12 INS17   
## -1.194822 NA -1.058498 13.639362 -0.079325 0.697682   
## INS19 INS22 INS25 INS27 INS29 ER.VISIT2   
## 0.439947 -0.041448 NA -0.884905 1.593803 -0.146146   
## ER.VISIT5 ER.VISIT6 ER.VISIT7 ER.VISIT9 HOSP.VST2 HOSP.VST4   
## -0.811271 -1.117798 -0.541108 -12.841299 -0.370558 -0.330771   
## HOSP.VST5 HOSP.VST6 HOSP.VST7 ASRXCOST2 ASRXCOST5 ASRXCOST7   
## NA NA -0.246388 0.001970 NA 0.099281   
## ASRXCOST9 WORKTALK2 WORKTALK6 WORKTALK7 WORKTALK8 WORKTALK9   
## 0.022983 -0.645312 -0.875433 -0.736890 0.493321 -1.026428   
##   
## Degrees of Freedom: 9171 Total (i.e. Null); 9067 Residual  
## Null Deviance: 11080   
## Residual Deviance: 10290 AIC: 10500

#### 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 +   
## DEPRESS + INCINDT + LAST.MD + LAST.MED + LAST.SYMP + COMPASTH +   
## HOSP.VST + WORKTALK, family = binomial, data = training1)  
##   
## Coefficients:  
## (Intercept) SEX2 AGEG.F72 AGEG.F73 AGEG.F74 AGEG.F75   
## 0.67503 0.38978 0.03151 0.09901 -0.02593 -0.24730   
## AGEG.F76 AGEG.F77 X\_RACEGR32 X\_RACEGR33 X\_RACEGR34 X\_RACEGR35   
## -0.37922 -0.57735 0.44897 -0.20453 0.21221 0.26320   
## X\_RACEGR39 EDUCAL2 EDUCAL3 EDUCAL4 EDUCAL5 EDUCAL6   
## 0.28926 -1.48356 -0.96518 -0.76571 -0.64278 -0.64865   
## EDUCAL9 COPD2 COPD7 COPD9 DEPRESS2 DEPRESS7   
## -0.15141 -0.26254 -0.48888 -0.65500 0.12697 -0.48893   
## DEPRESS9 INCINDT2 INCINDT3 INCINDT7 INCINDT9 LAST.MD5   
## 0.10701 0.05560 0.65730 -0.17102 0.81770 0.23577   
## LAST.MD6 LAST.MD7 LAST.MD77 LAST.MD88 LAST.MD99 LAST.MED2   
## -0.05437 -0.35912 -0.94164 0.07989 -0.44613 -0.17772   
## LAST.MED3 LAST.MED4 LAST.MED5 LAST.MED6 LAST.MED7 LAST.MED77   
## -0.36298 -0.53144 -0.44380 -0.42466 -0.26556 -0.74123   
## LAST.MED99 LAST.SYMP2 LAST.SYMP3 LAST.SYMP4 LAST.SYMP5 LAST.SYMP6   
## -11.31099 0.19962 0.25733 0.43099 0.01780 0.04559   
## LAST.SYMP7 LAST.SYMP77 LAST.SYMP88 LAST.SYMP99 COMPASTH11 COMPASTH2   
## -0.13301 -0.17353 -1.18106 1.84112 -0.43808 -0.27741   
## COMPASTH3 COMPASTH4 COMPASTH6 COMPASTH7 COMPASTH9 HOSP.VST2   
## -0.22703 -1.21507 NA -1.04322 12.62674 -0.45987   
## HOSP.VST4 HOSP.VST5 HOSP.VST6 HOSP.VST7 WORKTALK2 WORKTALK6   
## -0.43338 -0.71792 -1.02477 -0.25267 -0.65732 -0.90508   
## WORKTALK7 WORKTALK8 WORKTALK9   
## -0.75831 0.25672 -1.02508   
##   
## Degrees of Freedom: 9171 Total (i.e. Null); 9104 Residual  
## Null Deviance: 11080   
## Residual Deviance: 10330 AIC: 10470

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 1541 584  
## 1 82 85  
##   
## Accuracy : 0.7094   
## 95% CI : (0.6904, 0.728)  
## No Information Rate : 0.7081   
## P-Value [Acc > NIR] : 0.4555   
##   
## Kappa : 0.0982   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.12706   
## Specificity : 0.94948   
## Pos Pred Value : 0.50898   
## Neg Pred Value : 0.72518   
## Prevalence : 0.29188   
## Detection Rate : 0.03709   
## Detection Prevalence : 0.07286   
## Balanced Accuracy : 0.53827   
##   
## '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   
## 0.080651 0.361343 0.021414 0.084859 -0.037847 -0.240055   
## AGEG.F76 AGEG.F77 X\_RACEGR32 X\_RACEGR33 X\_RACEGR34 X\_RACEGR35   
## -0.357383 -0.514889 0.447556 -0.220765 0.210588 0.255396   
## X\_RACEGR39 EDUCAL2 EDUCAL3 EDUCAL4 EDUCAL5 EDUCAL6   
## 0.314758 -1.186373 -0.649726 -0.457224 -0.331784 -0.336729   
## EDUCAL9 X\_INCOMG2 X\_INCOMG3 X\_INCOMG4 X\_INCOMG5 X\_INCOMG9   
## 0.082770 0.029843 0.125526 -0.132588 -0.007516 -0.085219   
## BRONCH2 BRONCH7 BRONCH9 DUR.30D10 DUR.30D11 DUR.30D12   
## -0.146313 -0.267072 -1.345563 0.298953 -0.104421 -0.091564   
## DUR.30D2 DUR.30D6 DUR.30D7 DUR.30D77 DUR.30D9 DUR.30D99   
## -0.301763 -1.263734 -1.130751 -0.339126 -11.569494 -0.065062   
## INCINDT2 INCINDT3 INCINDT7 INCINDT9 LAST.MD5 LAST.MD6   
## 0.036932 0.622928 -0.184242 0.791634 -0.320354 -0.561904   
## LAST.MD7 LAST.MD77 LAST.MD88 LAST.MD99 LAST.MED2 LAST.MED3   
## -0.837523 -1.019408 -0.498965 -0.396666 -0.217678 -0.422067   
## LAST.MED4 LAST.MED5 LAST.MED6 LAST.MED7 LAST.MED77 LAST.MED99   
## -0.590182 -0.481048 -0.449286 -0.261555 -0.766662 -11.440581   
## LAST.SYMP2 LAST.SYMP3 LAST.SYMP4 LAST.SYMP5 LAST.SYMP6 LAST.SYMP7   
## 0.175750 0.233560 NA 1.214932 1.265219 1.136694   
## LAST.SYMP77 LAST.SYMP88 LAST.SYMP99 COMPASTH11 COMPASTH2 COMPASTH3   
## -0.133388 NA 1.899011 -0.440892 -0.259834 -0.218830   
## COMPASTH4 COMPASTH6 COMPASTH7 COMPASTH9 WORKTALK2 WORKTALK6   
## -1.195201 NA -0.989726 12.669093 -0.642399 -0.866024   
## WORKTALK7 WORKTALK8 WORKTALK9   
## -0.747603 0.252681 -1.013315   
##   
## Degrees of Freedom: 9171 Total (i.e. Null); 9100 Residual  
## Null Deviance: 11080   
## Residual Deviance: 10350 AIC: 10490

#### Confusion Matrix with the testingset

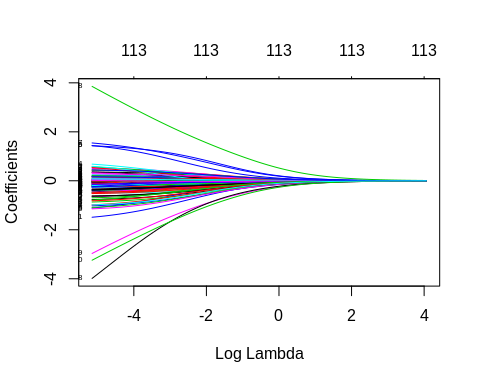
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 1545 581  
## 1 78 88  
##   
## Accuracy : 0.7125   
## 95% CI : (0.6935, 0.7309)  
## No Information Rate : 0.7081   
## P-Value [Acc > NIR] : 0.3322   
##   
## Kappa : 0.1072   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.13154   
## Specificity : 0.95194   
## Pos Pred Value : 0.53012   
## Neg Pred Value : 0.72672   
## Prevalence : 0.29188   
## Detection Rate : 0.03839   
## Detection Prevalence : 0.07243   
## Balanced Accuracy : 0.54174   
##   
## 'Positive' Class : 1   
##

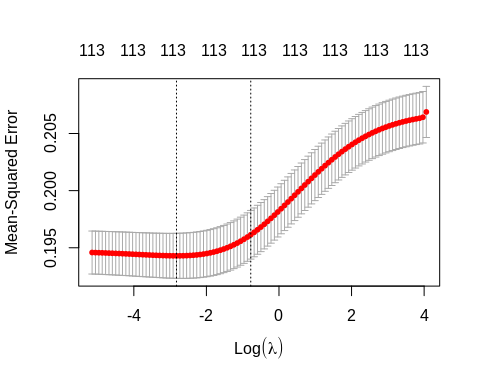
#### Lasso and Ridge model

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



Lambda that Minimises MSE

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] 0.05934789

#### Confusion matrix with lambda min

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 1620 668  
## 1 3 1  
##   
## Accuracy : 0.7072   
## 95% CI : (0.6881, 0.7258)  
## No Information Rate : 0.7081   
## P-Value [Acc > NIR] : 0.547   
##   
## Kappa : -5e-04   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.0014948   
## Specificity : 0.9981516   
## Pos Pred Value : 0.2500000   
## Neg Pred Value : 0.7080420   
## Prevalence : 0.2918848   
## Detection Rate : 0.0004363   
## Detection Prevalence : 0.0017452   
## Balanced Accuracy : 0.4998232   
##   
## 'Positive' Class : 1   
##

We observe overfitting with this ridge model

#### Confusion matrix with best lambda

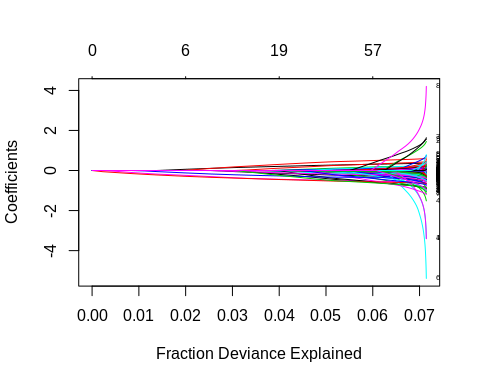
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 1623 669  
## 1 0 0  
##   
## Accuracy : 0.7081   
## 95% CI : (0.689, 0.7267)  
## No Information Rate : 0.7081   
## P-Value [Acc > NIR] : 0.5104   
##   
## Kappa : 0   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.0000   
## Specificity : 1.0000   
## Pos Pred Value : NaN   
## Neg Pred Value : 0.7081   
## Prevalence : 0.2919   
## Detection Rate : 0.0000   
## Detection Prevalence : 0.0000   
## Balanced Accuracy : 0.5000   
##   
## 'Positive' Class : 1   
##

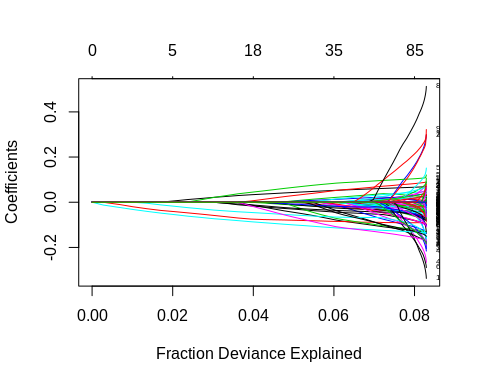
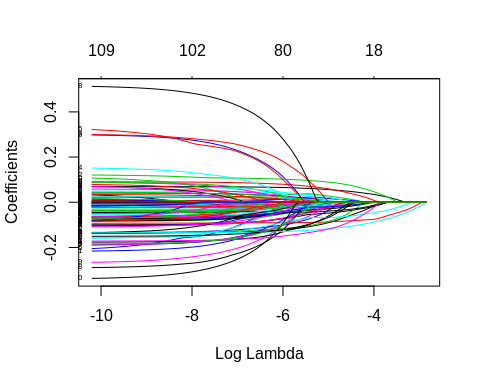
We observe overfitting with this second ridge model

#### Getting the coefficients

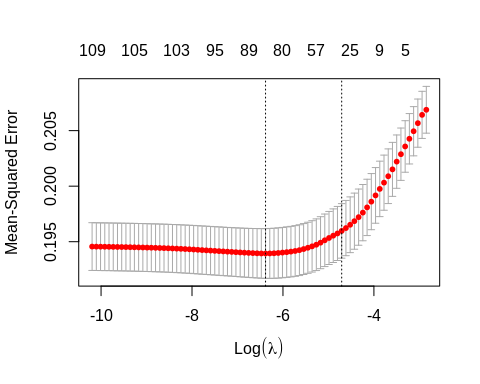
## 115 x 1 sparse Matrix of class "dgCMatrix"  
## s0  
## (Intercept) -0.311165635  
## (Intercept) .   
## SEX2 0.268905875  
## AGEG.F72 0.121885314  
## AGEG.F73 0.191736371  
## AGEG.F74 0.123772029  
## AGEG.F75 -0.032188058  
## AGEG.F76 -0.130739675  
## AGEG.F77 -0.276264679  
## X\_RACEGR32 0.345370100  
## X\_RACEGR33 -0.168141719  
## X\_RACEGR34 0.163555652  
## X\_RACEGR35 0.188481891  
## X\_RACEGR39 0.195951087  
## EDUCAL2 -0.563246481  
## EDUCAL3 -0.182972819  
## EDUCAL4 -0.042867391  
## EDUCAL5 0.062680154  
## EDUCAL6 0.055564892  
## EDUCAL9 0.371413396  
## X\_INCOMG2 0.025243211  
## X\_INCOMG3 0.096292186  
## X\_INCOMG4 -0.092595652  
## X\_INCOMG5 0.001257866  
## X\_INCOMG9 -0.077444585  
## X\_RFBMI52 -0.061302201  
## X\_RFBMI59 -0.016489589  
## SMOKE1002 0.028885927  
## SMOKE1007 0.291630495  
## SMOKE1009 1.088366129  
## COPD2 -0.160502579  
## COPD7 -0.317161991  
## COPD9 -0.008265946  
## EMPHY2 -0.044676385  
## EMPHY7 -0.357739837  
## EMPHY9 -0.571614550  
## DEPRESS2 0.077386564  
## DEPRESS7 -0.282525745  
## DEPRESS9 0.177231137  
## BRONCH2 -0.091278530  
## BRONCH7 -0.185188286  
## BRONCH9 -0.892552189  
## DUR.30D10 0.103059663  
## DUR.30D11 -0.007410897  
## DUR.30D12 0.018440319  
## DUR.30D2 -0.189528989  
## DUR.30D6 -0.043051137  
## DUR.30D7 -0.714063545  
## DUR.30D77 -0.199413085  
## DUR.30D9 -1.369560353  
## DUR.30D99 -0.149140530  
## INCINDT2 -0.177554947  
## INCINDT3 0.333653991  
## INCINDT7 -0.305724039  
## INCINDT9 0.327707463  
## LAST.MD5 0.046352354  
## LAST.MD6 -0.140558615  
## LAST.MD7 -0.324589440  
## LAST.MD77 -0.622666977  
## LAST.MD88 -0.092433891  
## LAST.MD99 -0.416059561  
## LAST.MED2 -0.069855890  
## LAST.MED3 -0.190342125  
## LAST.MED4 -0.282619266  
## LAST.MED5 -0.213971081  
## LAST.MED6 -0.204577487  
## LAST.MED7 -0.142499572  
## LAST.MED77 -0.462091222  
## LAST.MED99 -1.519208340  
## LAST.SYMP2 0.074695782  
## LAST.SYMP3 0.104344631  
## LAST.SYMP4 0.103427204  
## LAST.SYMP5 0.047617877  
## LAST.SYMP6 0.047414364  
## LAST.SYMP7 -0.092154999  
## LAST.SYMP77 -0.239520085  
## LAST.SYMP88 -0.731665053  
## LAST.SYMP99 1.029382576  
## EPIS.12M2 -0.131132983  
## EPIS.12M6 -0.045041216  
## EPIS.12M7 -0.455741521  
## EPIS.12M9 0.241892279  
## COMPASTH11 -0.143376396  
## COMPASTH2 -0.072476113  
## COMPASTH3 -0.060151767  
## COMPASTH4 -0.769612572  
## COMPASTH6 -0.044309830  
## COMPASTH7 -0.628683252  
## COMPASTH9 2.088878498  
## INS12 -0.041714894  
## INS17 0.165315864  
## INS19 0.226422818  
## INS22 -0.036741753  
## INS25 -0.041725122  
## INS27 -0.425696420  
## INS29 0.827497153  
## ER.VISIT2 -0.067959603  
## ER.VISIT5 -0.077432802  
## ER.VISIT6 -0.233869539  
## ER.VISIT7 -0.376234693  
## ER.VISIT9 -1.554564161  
## HOSP.VST2 -0.022813098  
## HOSP.VST4 0.066196704  
## HOSP.VST5 -0.076585501  
## HOSP.VST6 -0.233271485  
## HOSP.VST7 0.070767181  
## ASRXCOST2 -0.006830214  
## ASRXCOST5 -0.076131222  
## ASRXCOST7 0.043116310  
## ASRXCOST9 -0.029149814  
## WORKTALK2 -0.476959994  
## WORKTALK6 -0.543345939  
## WORKTALK7 -0.464366092  
## WORKTALK8 0.321695786  
## WORKTALK9 -0.669199573

##### 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 1555 593  
## 1 68 76  
##   
## Accuracy : 0.7116   
## 95% CI : (0.6926, 0.7301)  
## No Information Rate : 0.7081   
## P-Value [Acc > NIR] : 0.3663   
##   
## Kappa : 0.0932   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.11360   
## Specificity : 0.95810   
## Pos Pred Value : 0.52778   
## Neg Pred Value : 0.72393   
## Prevalence : 0.29188   
## Detection Rate : 0.03316   
## Detection Prevalence : 0.06283   
## Balanced Accuracy : 0.53585   
##   
## 'Positive' Class : 1   
##

#### Getting the coefficients

## 115 x 1 sparse Matrix of class "dgCMatrix"  
## s0  
## (Intercept) -0.2579854899  
## (Intercept) .   
## SEX2 0.3542166859  
## AGEG.F72 0.0632343529  
## AGEG.F73 0.1399556825  
## AGEG.F74 0.0390792790  
## AGEG.F75 -0.1367723728  
## AGEG.F76 -0.2586333383  
## AGEG.F77 -0.4402801733  
## X\_RACEGR32 0.4020122135  
## X\_RACEGR33 -0.1657005319  
## X\_RACEGR34 0.1596831570  
## X\_RACEGR35 0.2233125366  
## X\_RACEGR39 0.1951977048  
## EDUCAL2 -0.7606948688  
## EDUCAL3 -0.2686823287  
## EDUCAL4 -0.1016518549  
## EDUCAL5 .   
## EDUCAL6 .   
## EDUCAL9 0.1981910680  
## X\_INCOMG2 0.0039926583  
## X\_INCOMG3 0.0911052087  
## X\_INCOMG4 -0.0979138979  
## X\_INCOMG5 .   
## X\_INCOMG9 -0.0712901224  
## X\_RFBMI52 -0.0532072941  
## X\_RFBMI59 .   
## SMOKE1002 0.0068398773  
## SMOKE1007 0.2528453161  
## SMOKE1009 0.9267670939  
## COPD2 -0.2077955878  
## COPD7 -0.3397625834  
## COPD9 .   
## EMPHY2 -0.0115236452  
## EMPHY7 -0.3225277669  
## EMPHY9 -0.3000741589  
## DEPRESS2 0.1033939415  
## DEPRESS7 -0.1313032009  
## DEPRESS9 .   
## BRONCH2 -0.0823806414  
## BRONCH7 -0.1481377298  
## BRONCH9 -0.7407051201  
## DUR.30D10 0.1671942061  
## DUR.30D11 .   
## DUR.30D12 .   
## DUR.30D2 -0.2071573169  
## DUR.30D6 .   
## DUR.30D7 -0.7771113181  
## DUR.30D77 -0.1641557803  
## DUR.30D9 -0.9766940404  
## DUR.30D99 .   
## INCINDT2 .   
## INCINDT3 0.5674152990  
## INCINDT7 .   
## INCINDT9 0.3871056134  
## LAST.MD5 .   
## LAST.MD6 -0.2343082060  
## LAST.MD7 -0.5630790101  
## LAST.MD77 -0.7754395287  
## LAST.MD88 -0.0395389804  
## LAST.MD99 -0.1722368343  
## LAST.MED2 -0.0967159868  
## LAST.MED3 -0.2750431383  
## LAST.MED4 -0.4149006454  
## LAST.MED5 -0.3401234455  
## LAST.MED6 -0.3126209215  
## LAST.MED7 -0.1919073171  
## LAST.MED77 -0.5040969504  
## LAST.MED99 -1.6110364164  
## LAST.SYMP2 0.1017268810  
## LAST.SYMP3 0.1563643114  
## LAST.SYMP4 0.1166262675  
## LAST.SYMP5 .   
## LAST.SYMP6 .   
## LAST.SYMP7 -0.1257986113  
## LAST.SYMP77 -0.1792473910  
## LAST.SYMP88 -0.9576660890  
## LAST.SYMP99 1.1412666330  
## EPIS.12M2 .   
## EPIS.12M6 .   
## EPIS.12M7 -0.3437286481  
## EPIS.12M9 .   
## COMPASTH11 -0.2837501453  
## COMPASTH2 -0.1067554496  
## COMPASTH3 -0.0820766622  
## COMPASTH4 -0.6426440521  
## COMPASTH6 .   
## COMPASTH7 -0.6743718403  
## COMPASTH9 1.7028612685  
## INS12 -0.0007142587  
## INS17 .   
## INS19 .   
## INS22 .   
## INS25 .   
## INS27 -0.2845170762  
## INS29 1.0038782839  
## ER.VISIT2 -0.1189195122  
## ER.VISIT5 -0.0030299001  
## ER.VISIT6 -0.4258432089  
## ER.VISIT7 -0.2794223153  
## ER.VISIT9 -0.9050853282  
## HOSP.VST2 -0.0716798701  
## HOSP.VST4 .   
## HOSP.VST5 -0.1627026609  
## HOSP.VST6 -0.0835855569  
## HOSP.VST7 .   
## ASRXCOST2 .   
## ASRXCOST5 -0.0233760057  
## ASRXCOST7 .   
## ASRXCOST9 .   
## WORKTALK2 -0.6171783972  
## WORKTALK6 -0.7728890333  
## WORKTALK7 -0.6132713063  
## WORKTALK8 0.0712529778  
## WORKTALK9 -0.6265422376

#### Confusion Matrix with best lambda

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 1614 658  
## 1 9 11  
##   
## Accuracy : 0.709   
## 95% CI : (0.6899, 0.7275)  
## No Information Rate : 0.7081   
## P-Value [Acc > NIR] : 0.4738   
##   
## Kappa : 0.0152   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.016442   
## Specificity : 0.994455   
## Pos Pred Value : 0.550000   
## Neg Pred Value : 0.710387   
## Prevalence : 0.291885   
## Detection Rate : 0.004799   
## Detection Prevalence : 0.008726   
## Balanced Accuracy : 0.505449   
##   
## '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] -495.6112

## [1] -593.2907

#### Partial Least Squared

#### Confusion Matrix with best lambda

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction F T  
## F 1556 593  
## T 67 76  
##   
## Accuracy : 0.712   
## 95% CI : (0.693, 0.7305)  
## No Information Rate : 0.7081   
## P-Value [Acc > NIR] : 0.3491   
##   
## Kappa : 0.0941   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.11360   
## Specificity : 0.95872   
## Pos Pred Value : 0.53147   
## Neg Pred Value : 0.72406   
## Prevalence : 0.29188   
## Detection Rate : 0.03316   
## Detection Prevalence : 0.06239   
## Balanced Accuracy : 0.53616   
##   
## 'Positive' Class : T   
##

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

## Partial Least Squares   
##   
## 9172 samples  
## 24 predictor  
## 2 classes: 'F', 'T'   
##   
## Pre-processing: centered (113), scaled (113)   
## Resampling: Cross-Validated (10 fold, repeated 3 times)   
## Summary of sample sizes: 8254, 8255, 8255, 8255, 8255, 8255, ...   
## Resampling results across tuning parameters:  
##   
## ncomp ROC Sens Spec   
## 1 0.6219778 1.0000000 0.00000000  
## 2 0.6429671 0.9875208 0.03605809  
## 3 0.6518623 0.9685715 0.08355250  
## 4 0.6540849 0.9612278 0.10630722  
## 5 0.6553881 0.9588149 0.11177801  
## 6 0.6560877 0.9580958 0.11588202  
## 7 0.6561267 0.9576333 0.11625238  
## 8 0.6559958 0.9562980 0.11799460  
## 9 0.6560079 0.9573250 0.11899147  
## 10 0.6557761 0.9572736 0.12011088  
## 11 0.6558837 0.9568113 0.11849350  
## 12 0.6559341 0.9569138 0.11787207  
## 13 0.6561282 0.9567598 0.11724926  
## 14 0.6561443 0.9564517 0.11637907  
## 15 0.6561146 0.9567085 0.11724972  
##   
## ROC was used to select the optimal model using the largest value.  
## The final value used for the model was ncomp = 14.

#### Confusion Matrix with best lambda

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction F T  
## F 1548 598  
## T 75 71  
##   
## Accuracy : 0.7064   
## 95% CI : (0.6873, 0.725)  
## No Information Rate : 0.7081   
## P-Value [Acc > NIR] : 0.5831   
##   
## Kappa : 0.0778   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.10613   
## Specificity : 0.95379   
## Pos Pred Value : 0.48630   
## Neg Pred Value : 0.72134   
## Prevalence : 0.29188   
## Detection Rate : 0.03098   
## Detection Prevalence : 0.06370   
## Balanced Accuracy : 0.52996   
##   
## 'Positive' Class : T   
##

### SELECT MODELS

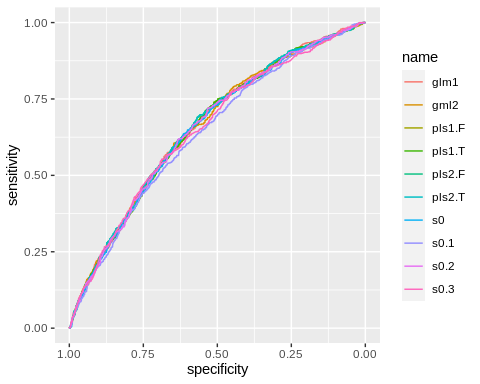
#### We compare the models with the accuray, precision, sensitivity, specificity, and F1 score from the confusion matrix

## glm.mod11 glm.mod12 ridge.mod1 ridge.mod2 lasso.mod1 lasso.mod2  
## Accuracy 0.7094241 0.7124782 0.707242583 0.7081152 0.7116056 0.70898778  
## Precision 0.5089820 0.5301205 0.250000000 NA 0.5277778 0.55000000  
## Sensitivity 0.1270553 0.1315396 0.001494768 0.0000000 0.1136024 0.01644245  
## Specificity 0.9494763 0.9519409 0.998151571 1.0000000 0.9581023 0.99445471  
## F1 0.2033493 0.2107784 0.002971768 NA 0.1869619 0.03193033  
## pls.mod1 pls.mod2  
## Accuracy 0.7120419 0.7063700  
## Precision 0.5314685 0.4863014  
## Sensitivity 0.1136024 0.1061286  
## Specificity 0.9587184 0.9537893  
## F1 0.1871921 0.1742331

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.



All models have approximally the same AUC Any of these model can be used to modelization

### 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.min33)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.7376 -0.8468 -0.6538 1.1955 2.3412   
##   
## Coefficients: (3 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.47880 0.64646 -0.741 0.458911   
## SEX2 0.32423 0.04886 6.637 3.21e-11 \*\*\*  
## AGEG.F72 0.08962 0.12111 0.740 0.459345   
## AGEG.F73 0.07980 0.11747 0.679 0.496915   
## AGEG.F74 0.05414 0.11162 0.485 0.627633   
## AGEG.F75 -0.17727 0.10859 -1.633 0.102565   
## AGEG.F76 -0.26486 0.11072 -2.392 0.016746 \*   
## AGEG.F77 -0.49502 0.12315 -4.020 5.83e-05 \*\*\*  
## X\_RACEGR32 0.49190 0.09032 5.446 5.14e-08 \*\*\*  
## X\_RACEGR33 -0.17365 0.12164 -1.428 0.153416   
## X\_RACEGR34 0.24931 0.11564 2.156 0.031099 \*   
## X\_RACEGR35 0.27850 0.07968 3.495 0.000473 \*\*\*  
## X\_RACEGR39 0.38452 0.19882 1.934 0.053114 .   
## EDUCAL2 -0.73916 0.63099 -1.171 0.241430   
## EDUCAL3 -0.26962 0.61645 -0.437 0.661841   
## EDUCAL4 -0.02128 0.61078 -0.035 0.972202   
## EDUCAL5 0.08729 0.61095 0.143 0.886390   
## EDUCAL6 0.08683 0.61133 0.142 0.887056   
## EDUCAL9 0.09631 0.86554 0.111 0.911401   
## X\_INCOMG2 0.03643 0.07844 0.464 0.642304   
## X\_INCOMG3 0.09781 0.09200 1.063 0.287693   
## X\_INCOMG4 -0.11957 0.08780 -1.362 0.173238   
## X\_INCOMG5 -0.01487 0.07428 -0.200 0.841339   
## X\_INCOMG9 -0.11147 0.09198 -1.212 0.225537   
## BRONCH2 -0.11131 0.05074 -2.194 0.028269 \*   
## BRONCH7 -0.32169 0.25218 -1.276 0.202097   
## BRONCH9 -0.42949 0.90079 -0.477 0.633515   
## DUR.30D10 0.25994 0.10053 2.586 0.009719 \*\*   
## DUR.30D11 -0.16578 0.13290 -1.247 0.212242   
## DUR.30D12 -0.10877 0.08429 -1.290 0.196904   
## DUR.30D2 -0.29540 0.10645 -2.775 0.005522 \*\*   
## DUR.30D6 -0.64531 0.39324 -1.641 0.100795   
## DUR.30D7 -0.76771 0.50346 -1.525 0.127298   
## DUR.30D77 -0.25590 0.21537 -1.188 0.234760   
## DUR.30D9 -11.54098 229.52158 -0.050 0.959897   
## DUR.30D99 -0.05268 0.74209 -0.071 0.943406   
## INCINDT2 0.16096 0.18116 0.888 0.374276   
## INCINDT3 0.74473 0.16095 4.627 3.71e-06 \*\*\*  
## INCINDT7 -0.18192 0.59398 -0.306 0.759394   
## INCINDT9 0.97427 0.91090 1.070 0.284813   
## LAST.MD5 -0.32490 0.06911 -4.701 2.59e-06 \*\*\*  
## LAST.MD6 -0.52735 0.10781 -4.892 1.00e-06 \*\*\*  
## LAST.MD7 -0.81515 0.09477 -8.601 < 2e-16 \*\*\*  
## LAST.MD77 -0.94184 0.35074 -2.685 0.007247 \*\*   
## LAST.MD88 -0.46658 0.35894 -1.300 0.193643   
## LAST.MD99 -0.59672 1.15431 -0.517 0.605190   
## LAST.MED2 -0.27670 0.08505 -3.254 0.001140 \*\*   
## LAST.MED3 -0.46020 0.07684 -5.989 2.11e-09 \*\*\*  
## LAST.MED4 -0.56460 0.08990 -6.280 3.38e-10 \*\*\*  
## LAST.MED5 -0.44271 0.10144 -4.364 1.27e-05 \*\*\*  
## LAST.MED6 -0.36147 0.13604 -2.657 0.007883 \*\*   
## LAST.MED7 -0.26209 0.11119 -2.357 0.018410 \*   
## LAST.MED77 -0.59640 0.34346 -1.736 0.082486 .   
## LAST.MED99 -11.25426 142.91341 -0.079 0.937232   
## LAST.SYMP2 0.14554 0.08156 1.785 0.074335 .   
## LAST.SYMP3 0.27305 0.07816 3.493 0.000477 \*\*\*  
## LAST.SYMP4 NA NA NA NA   
## LAST.SYMP5 0.47650 0.38975 1.223 0.221498   
## LAST.SYMP6 0.61704 0.39993 1.543 0.122865   
## LAST.SYMP7 0.46331 0.39139 1.184 0.236513   
## LAST.SYMP77 0.01545 0.22635 0.068 0.945595   
## LAST.SYMP88 NA NA NA NA   
## LAST.SYMP99 1.37275 0.87098 1.576 0.115005   
## COMPASTH11 -0.47530 0.07722 -6.155 7.49e-10 \*\*\*  
## COMPASTH2 -0.31109 0.10496 -2.964 0.003039 \*\*   
## COMPASTH3 -0.24747 0.07705 -3.212 0.001320 \*\*   
## COMPASTH4 -1.49834 1.06773 -1.403 0.160530   
## COMPASTH6 NA NA NA NA   
## COMPASTH7 -1.04763 0.46884 -2.235 0.025449 \*   
## COMPASTH9 12.56006 324.74375 0.039 0.969148   
## WORKTALK2 -0.65194 0.05140 -12.685 < 2e-16 \*\*\*  
## WORKTALK6 -0.67839 0.17437 -3.891 0.000100 \*\*\*  
## WORKTALK7 -0.74954 0.28321 -2.647 0.008131 \*\*   
## WORKTALK8 -0.19499 0.58715 -0.332 0.739815   
## WORKTALK9 -0.35493 0.87129 -0.407 0.683741   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 13851 on 11463 degrees of freedom  
## Residual deviance: 12961 on 11392 degrees of freedom  
## AIC: 13105  
##   
## Number of Fisher Scoring iterations: 11

We look at the odd ratio off each variable

## (Intercept) SEX2 AGEG.F72 AGEG.F73 AGEG.F74 AGEG.F75   
## 6.195289e-01 1.382967e+00 1.093754e+00 1.083075e+00 1.055636e+00 8.375516e-01   
## AGEG.F76 AGEG.F77 X\_RACEGR32 X\_RACEGR33 X\_RACEGR34 X\_RACEGR35   
## 7.673137e-01 6.095579e-01 1.635420e+00 8.405917e-01 1.283135e+00 1.321147e+00   
## X\_RACEGR39 EDUCAL2 EDUCAL3 EDUCAL4 EDUCAL5 EDUCAL6   
## 1.468912e+00 4.775156e-01 7.636697e-01 9.789408e-01 1.091211e+00 1.090709e+00   
## EDUCAL9 X\_INCOMG2 X\_INCOMG3 X\_INCOMG4 X\_INCOMG5 X\_INCOMG9   
## 1.101100e+00 1.037105e+00 1.102759e+00 8.873024e-01 9.852412e-01 8.945189e-01   
## BRONCH2 BRONCH7 BRONCH9 DUR.30D10 DUR.30D11 DUR.30D12   
## 8.946646e-01 7.249265e-01 6.508437e-01 1.296853e+00 8.472303e-01 8.969357e-01   
## DUR.30D2 DUR.30D6 DUR.30D7 DUR.30D77 DUR.30D9 DUR.30D99   
## 7.442365e-01 5.245013e-01 4.640768e-01 7.742219e-01 9.723382e-06 9.486825e-01   
## INCINDT2 INCINDT3 INCINDT7 INCINDT9 LAST.MD5 LAST.MD6   
## 1.174640e+00 2.105865e+00 8.336673e-01 2.649229e+00 7.226022e-01 5.901639e-01   
## LAST.MD7 LAST.MD77 LAST.MD88 LAST.MD99 LAST.MED2 LAST.MED3   
## 4.425730e-01 3.899083e-01 6.271459e-01 5.506147e-01 7.582794e-01 6.311567e-01   
## LAST.MED4 LAST.MED5 LAST.MED6 LAST.MED7 LAST.MED77 LAST.MED99   
## 5.685889e-01 6.422908e-01 6.966515e-01 7.694391e-01 5.507925e-01 1.295202e-05   
## LAST.SYMP2 LAST.SYMP3 LAST.SYMP4 LAST.SYMP5 LAST.SYMP6 LAST.SYMP7   
## 1.156665e+00 1.313960e+00 NA 1.610422e+00 1.853437e+00 1.589321e+00   
## LAST.SYMP77 LAST.SYMP88 LAST.SYMP99 COMPASTH11 COMPASTH2 COMPASTH3   
## 1.015566e+00 NA 3.946170e+00 6.216968e-01 7.326472e-01 7.807759e-01   
## COMPASTH4 COMPASTH6 COMPASTH7 COMPASTH9 WORKTALK2 WORKTALK6   
## 2.235013e-01 NA 3.507681e-01 2.849488e+05 5.210336e-01 5.074334e-01   
## WORKTALK7 WORKTALK8 WORKTALK9   
## 4.725842e-01 8.228418e-01 7.012206e-01