# STOR 455 Class 30 Multiple Logistic Regression (Again)

library(readr)  
library(bestglm)  
library(Stat2Data)  
  
insurance <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/insurance.csv")

* Looking at something like mallowCp and see which one is most likely based on teh precitors and a few other things -

1. Only the resposne and possible predictors variables should be withing the datagrame
2. te response variable must be the last column in teh dataframe.

We need to tell R. We need to think what we don’t want in the model ; if we have accept anad acceptance in teh model,then we are going to get some errors abecause they are the same things; there are issues because the logistical model wouldn’t work as well because it would be a straight vertial line.

WE could chosoe teh specific columns we want or choose teh ones we dont’ want with negative index ;

**Example: Predicting Medical School Acceptance** Data: MedGPA  
Accept Status: A=accepted to medical school or D=denied admission Acceptance Indicator for Accept: 1=accepted or 0=denied Sex F=female or M=male BCPM Bio/Chem/Physics/Math grade point average GPA College grade point average VR Verbal reasoning (subscore) PS Physical sciences (subscore) WS Writing sample (subcore) BS Biological sciences (subscore) MCAT Score on the MCAT exam (sum of CR+PS+WS+BS) Apps Number of medical schools applied to

Find the “best” model for Acceptance using some or all of these predictors.

data(MedGPA)  
head(MedGPA)

## Accept Acceptance Sex BCPM GPA VR PS WS BS MCAT Apps  
## 1 D 0 F 3.59 3.62 11 9 9 9 38 5  
## 2 A 1 M 3.75 3.84 12 13 8 12 45 3  
## 3 A 1 F 3.24 3.23 9 10 5 9 33 19  
## 4 A 1 F 3.74 3.69 12 11 7 10 40 5  
## 5 A 1 F 3.53 3.38 9 11 4 11 35 11  
## 6 A 1 M 3.59 3.72 10 9 7 10 36 5

below shows how to set accetance to null, that deltes teh accept vars.

best glm wants the response in the specific part of the dataframe, wants it in teh last section ; if your thing is named soemthign sepciifc, it sometimes acts differently, but mostly this is different.

THe second part of the code below reorders teh columns with teh response value last so that the glm is better. THere are other ways that you can do this, but this is for consistency.

**bestglm for Model Selection** Requirements to use bestglm() 1. Only the response and possible predictor variables should be within the dataframe

MedGPA.1 = within(MedGPA, {Accept = NULL}) #delete Accept variable  
head(MedGPA.1)

## Acceptance Sex BCPM GPA VR PS WS BS MCAT Apps  
## 1 0 F 3.59 3.62 11 9 9 9 38 5  
## 2 1 M 3.75 3.84 12 13 8 12 45 3  
## 3 1 F 3.24 3.23 9 10 5 9 33 19  
## 4 1 F 3.74 3.69 12 11 7 10 40 5  
## 5 1 F 3.53 3.38 9 11 4 11 35 11  
## 6 1 M 3.59 3.72 10 9 7 10 36 5

Above we could have just overwritten the thing; but this is easy to make the running the cell a lot of times and then it will be fine.

using the best glm fucntion; just like when makign teh lienar model, we need to tell it which family of functions to draw from; it’s going to look at a LSRL if we dno’t tell it otherwise

family = binomial tells you to make it logistics.

1. The response variable must be the last column in the dataframe.

MedGPA.2 = MedGPA.1[,c(2:10,1)] #reorder columns with response last  
#bestglm for Model Selection  
head(MedGPA.2)

## Sex BCPM GPA VR PS WS BS MCAT Apps Acceptance  
## 1 F 3.59 3.62 11 9 9 9 38 5 0  
## 2 M 3.75 3.84 12 13 8 12 45 3 1  
## 3 F 3.24 3.23 9 10 5 9 33 19 1  
## 4 F 3.74 3.69 12 11 7 10 40 5 1  
## 5 F 3.53 3.38 9 11 4 11 35 11 1  
## 6 M 3.59 3.72 10 9 7 10 36 5 1

Tell em about teh BIC and BICQ DO the same thing, but same it as an object

The best nmodels will tell you how many best models there arel

the top rowis the best model that you would like

the next four best models are the other best models

BIC = the baysian information criteria ; we are going to use it like mallowCp

calculated like: klog(n) - 2log(L(alpha)); n = sample size k = number of predictors alpha = set of all paramets L(alpha) = probability of obtraining the data which you have, supposing the modelbeing tested was given

*SMaller values indicate preferred models*

tells you we got teh data ttha we give given the model

there si going to be a best, but there migh tnot be a stat difference between teh things;

it’s saying on ce we take teh neg 2log, that its nmmore likely that it gen teh data thta we got the value is based on teh samp size and num predictors it could still bea g odoo number if we have different predictors

most are within 0-2 BIC, so there are not much difference between them .

there sin’t much difference between teh models its easier to get teh data, but it’s not very stats; E don’t really know that, but these look pretty similar

**bestglm for Model Selection** BIC = Bayesian Information Criteria

bestglm(MedGPA.2, family=binomial)

## Morgan-Tatar search since family is non-gaussian.

## BIC  
## BICq equivalent for q in (0.407407122288894, 0.830512766582046)  
## Best Model:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -39.4708940 12.2144951 -3.231480 0.001231510  
## SexM -2.8403423 1.1580871 -2.452616 0.014182182  
## GPA 5.3344003 2.4807386 2.150327 0.031529326  
## PS 1.0247592 0.4722984 2.169728 0.030027451  
## WS -0.7177605 0.3496614 -2.052730 0.040098780  
## BS 1.7914617 0.6434984 2.783941 0.005370279

**Bayesian Information Criteria** k log(n)- 2log(L(θ̂))

n : sample size k : number of predictors θ : set of all parameters. L(θ̂) :probability of obtaining the data which you have, supposing the model being tested was a given.

Selection criteria, similar to Mallow’s Cp Smaller values indicate preferred models

**Comparing Models by BIC** Change in BIC; Evidence against hiher BIC 0-2; Little 2-6; POsitive 6-10; Strong greater than 10; Very strong

MedGPA.2.bestglm = bestglm(MedGPA.2, family=binomial)

## Morgan-Tatar search since family is non-gaussian.

MedGPA.2.bestglm$BestModels

## Sex BCPM GPA VR PS WS BS MCAT Apps Criterion  
## 1 TRUE FALSE TRUE FALSE TRUE TRUE TRUE FALSE FALSE 51.35809  
## 2 TRUE FALSE TRUE FALSE TRUE FALSE TRUE FALSE FALSE 52.67338  
## 3 TRUE FALSE TRUE TRUE FALSE TRUE FALSE TRUE FALSE 52.81895  
## 4 TRUE FALSE TRUE FALSE FALSE FALSE TRUE FALSE FALSE 52.85687  
## 5 TRUE TRUE FALSE FALSE TRUE TRUE TRUE FALSE FALSE 53.46655

**Example: Predicting Survival** Data: ICU  
ID Patient ID code Survive 1=patient survived to discharge or 0=patient died Age Age (in years) AgeGroup 1= young (under 50), 2= middle (50-69), 3 = old (70+) Sex 1=female or 0=male Infection 1=infection suspected or 0=no infection SysBP Systolic blood pressure (in mm of Hg) Pulse Heart rate (beats per minute) Emergency 1=emergency admission or 0=elective admission

Find the “best” model for Survival using some or all of these predictors.

data("ICU")  
head(ICU)

## ID Survive Age AgeGroup Sex Infection SysBP Pulse Emergency  
## 1 4 0 87 3 1 1 80 96 1  
## 2 8 1 27 1 1 1 142 88 1  
## 3 12 1 59 2 0 0 112 80 1  
## 4 14 1 77 3 0 0 100 70 0  
## 5 27 0 76 3 1 1 128 90 1  
## 6 28 1 54 2 0 1 142 103 1

#Requirements to use bestglm()  
#1. Only the response and possible predictor variables should be within the dataframe  
ICU.1 <- within(ICU, {ID = NULL}) #delete ID variable  
  
#delete ID variable  
# WHy do we delete teh ID Variable? We probably don't need it because each row = the incident number  
  
#2. The response variable must be the last column in the dataframe.  
#reorder columns with response last; column 1 is now the survived column because the ID column was deleted.  
ICU.2 = ICU.1[,c(2:8,1)] #reorder columns with response last  
  
# AgeGroup is Treated as Quantitative   
head(ICU.2)

## Age AgeGroup Sex Infection SysBP Pulse Emergency Survive  
## 1 87 3 1 1 80 96 1 0  
## 2 27 1 1 1 142 88 1 1  
## 3 59 2 0 0 112 80 1 1  
## 4 77 3 0 0 100 70 0 1  
## 5 76 3 1 1 128 90 1 0  
## 6 54 2 0 1 142 103 1 1

bestglm(ICU.2, family=binomial)

## Morgan-Tatar search since family is non-gaussian.

## BIC  
## BICq equivalent for q in (0.0293273190867612, 0.516811637275042)  
## Best Model:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 5.3032038 0.9752351 5.437872 5.392065e-08  
## AgeGroup -0.8430258 0.2515652 -3.351123 8.048461e-04  
## Emergency -2.5144865 0.7576616 -3.318746 9.042257e-04

# THis tells me that teh best variable to predict survived is Emergency  
  
bestglm(ICU.2, family=binomial)$BestModels

## Morgan-Tatar search since family is non-gaussian.

## Age AgeGroup Sex Infection SysBP Pulse Emergency Criterion  
## 1 FALSE TRUE FALSE FALSE FALSE FALSE TRUE 183.3483  
## 2 FALSE TRUE FALSE FALSE TRUE FALSE TRUE 183.4829  
## 3 TRUE FALSE FALSE FALSE FALSE FALSE TRUE 183.6723  
## 4 TRUE FALSE FALSE FALSE TRUE FALSE TRUE 183.7191  
## 5 FALSE TRUE FALSE TRUE FALSE FALSE TRUE 186.7208

# The criteria doesn't change very much between teh first three models   
# Criteria is teh BIC; we want this to be low   
  
#THe data is teaching Age group as a numerical verabiel, we need to cahnge it to a cateorical variable if we want to look at each age group

*BElow is how to make agegroup a categorical variable* We are reassingin tee variable age group as teh factor of age group, so this breaks it into whatever age groups that are under agegroup category.

ICU\_factor\_AgeGroup = ICU.2   
ICU\_factor\_AgeGroup$AgeGroup = factor(ICU\_factor\_AgeGroup$AgeGroup)  
  
head(ICU\_factor\_AgeGroup)

## Age AgeGroup Sex Infection SysBP Pulse Emergency Survive  
## 1 87 3 1 1 80 96 1 0  
## 2 27 1 1 1 142 88 1 1  
## 3 59 2 0 0 112 80 1 1  
## 4 77 3 0 0 100 70 0 1  
## 5 76 3 1 1 128 90 1 0  
## 6 54 2 0 1 142 103 1 1

below is running the log model on the log model, but wiht age group sections differentiated.

bestglm(ICU\_factor\_AgeGroup, family=binomial)

## Morgan-Tatar search since family is non-gaussian.

## Note: factors present with more than 2 levels.

## BIC  
## Best Model:  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Age 1 1.149 1.1486 8.004 0.00515 \*\*   
## Emergency 1 2.581 2.5811 17.987 3.42e-05 \*\*\*  
## Residuals 197 28.270 0.1435   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

now it’s not using age group; but it’s giving more datapoints there is a change in the amount of predictirs we dont expect the below to be the same as the above ones, because we added more varaibles by levling teh age group

this goes with us a warning: “Factors rpesent with more than 2 level” it’s saying one thing is more than 2 levels; we its telling us taht there are more to teh columns that they give us

ICU\_factor\_AgeGroup\_bestglm = bestglm(ICU\_factor\_AgeGroup, family=binomial)

## Morgan-Tatar search since family is non-gaussian.

## Note: factors present with more than 2 levels.

ICU\_factor\_AgeGroup\_bestglm$BestModels

## Age AgeGroup Sex Infection SysBP Pulse Emergency Criterion  
## 1 TRUE FALSE FALSE FALSE FALSE FALSE TRUE 183.6723  
## 2 TRUE FALSE FALSE FALSE TRUE FALSE TRUE 183.7191  
## 3 FALSE TRUE FALSE FALSE FALSE FALSE TRUE 187.0545  
## 4 FALSE TRUE FALSE FALSE TRUE FALSE TRUE 187.3861  
## 5 TRUE FALSE FALSE TRUE FALSE FALSE TRUE 187.4172

Below is making the age groups, assigning numbers; so if tha agegroup was 2, then put a 1, if it was 3, then put a 1, then the last code removes the agegroup column because we don’t need age group anymore since we included teh dummy predictors in the first two lines of code below.

below is what bestglm is doing. This looked at the data tiself. IT didn’t look at atransformation if we ignore tranformation, then we have the ebst model here.

But should we ignore tranofmromatio?

NOt always.

#Requirements to use bestglm()  
# 3. Create dummy variables for non binary categorical variables.  
  
ICU.2$AgeGroup2 = ifelse(ICU.2$AgeGroup==2,1,0)  
ICU.2$AgeGroup3 = ifelse(ICU.2$AgeGroup==3,1,0)  
ICU.3 <- within(ICU.2, {AgeGroup = NULL}) #delete AgeGroup variable  
ICU.4 = ICU.3[,c(1:6,8,9,7)] #reorder columns with response last  
  
head(ICU.4)

## Age Sex Infection SysBP Pulse Emergency AgeGroup2 AgeGroup3 Survive  
## 1 87 1 1 80 96 1 0 1 0  
## 2 27 1 1 142 88 1 0 0 1  
## 3 59 0 0 112 80 1 1 0 1  
## 4 77 0 0 100 70 0 0 1 1  
## 5 76 1 1 128 90 1 0 1 0  
## 6 54 0 1 142 103 1 1 0 1

bestglm(ICU.4, family=binomial)

## Morgan-Tatar search since family is non-gaussian.

## BIC  
## BICq equivalent for q in (0.0343073257857045, 0.505855168373752)  
## Best Model:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 5.50876248 1.03351106 5.330144 9.813508e-08  
## Age -0.03401617 0.01069436 -3.180759 1.468899e-03  
## Emergency -2.45353515 0.75256981 -3.260209 1.113300e-03

# Comparing Models by BIC  
ICU.4.bestglm = bestglm(ICU.4, family=binomial)

## Morgan-Tatar search since family is non-gaussian.

ICU.4.bestglm$BestModels

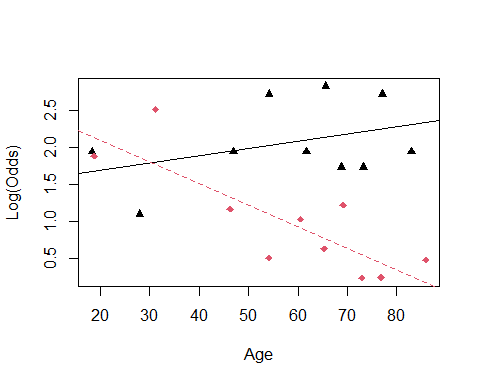
## Age Sex Infection SysBP Pulse Emergency AgeGroup2 AgeGroup3 Criterion  
## 1 TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE 183.6723  
## 2 TRUE FALSE FALSE TRUE FALSE TRUE FALSE FALSE 183.7191  
## 3 FALSE FALSE FALSE FALSE FALSE TRUE TRUE TRUE 187.0545  
## 4 FALSE FALSE FALSE TRUE FALSE TRUE TRUE TRUE 187.3861  
## 5 TRUE FALSE TRUE FALSE FALSE TRUE FALSE FALSE 187.4172

We are assuming that age has teh same impact on surivial as old people; so age in general causes teh same surivial results.

WE can guess tho; if older peopple come in that is going to be different than if younger people are going in for an enermcy. WE can do that with an emperical logit plot.

THis logitplot will help us split by a factor for those brough tin with emergency and not emergency. if you run into errors with emplogitplot, then you can just factor the variables and sometimes that helps. Factor the last variable only, if that doesnt work, then factor others **bestglm for Model Selection**

emplogitplot2(Survive~Age+factor(Emergency), data=ICU.4, ngroups=10)

 We are assuming that age has teh same impact on surivial as old people; so age in general causes teh same surivial results.

WE can guess tho; if older peopple come in that is going to be different than if younger people are going in for an enermcy. WE can do that with an emperical logit plot.

THis logitplot will help us split by a factor for those brough tin with emergency and not emergency. if you run into errors with emplogitplot, then you can just factor the variables and sometimes that helps. Factor the last variable only, if that doesnt work, then factor others

ICU.4$EMAGE = ICU.4$Age\*ICU.4$Emergency  
head(ICU.4)

## Age Sex Infection SysBP Pulse Emergency AgeGroup2 AgeGroup3 Survive EMAGE  
## 1 87 1 1 80 96 1 0 1 0 87  
## 2 27 1 1 142 88 1 0 0 1 27  
## 3 59 0 0 112 80 1 1 0 1 59  
## 4 77 0 0 100 70 0 0 1 1 0  
## 5 76 1 1 128 90 1 0 1 0 76  
## 6 54 0 1 142 103 1 1 0 1 54

ICU.5 = ICU.4[,c(1:8,10,9)] # THis moves surive to teh end of the columns, so that wecan keep doing the code with bestgml.   
head(ICU.5)

## Age Sex Infection SysBP Pulse Emergency AgeGroup2 AgeGroup3 EMAGE Survive  
## 1 87 1 1 80 96 1 0 1 87 0  
## 2 27 1 1 142 88 1 0 0 27 1  
## 3 59 0 0 112 80 1 1 0 59 1  
## 4 77 0 0 100 70 0 0 1 0 1  
## 5 76 1 1 128 90 1 0 1 76 0  
## 6 54 0 1 142 103 1 1 0 54 1

x = bestglm(ICU.5, family=binomial)

## Morgan-Tatar search since family is non-gaussian.

x$BestModels

## Age Sex Infection SysBP Pulse Emergency AgeGroup2 AgeGroup3 EMAGE  
## 1 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE  
## 2 FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE TRUE  
## 3 FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE TRUE  
## 4 TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE  
## 5 FALSE FALSE FALSE TRUE TRUE FALSE FALSE FALSE TRUE  
## Criterion  
## 1 179.1958  
## 2 179.2383  
## 3 183.0363  
## 4 183.6723  
## 5 183.6742

Someitmes best subsets isn’t as useful as we think so; for example: when you have categorical variables, soemtimes they are not immediately reflected through the best mdoels

*Use bestglm when you have binary categorical variables and when you have quantitative variables* IF you want to add interections adn transformations, then it will cause issues

THe ICU dataset was really nice, ti was really clean and easy to work with, but the below is less clean; insurance. How much you pay is based on a huge amount of htings; WHo elese do we haev data on that is like you and how much do we think that you and them are going toet in an accident and cost us moeny

index is just a number, there are 8k people; target flag = accident or no ltager amount= insurance costs - the first 6 rows, the first frow only 1 there are a lot of other types of things; red care = more insuance; previous thing; own a home, etc. so much we could deal with when making this.

THere are some probelms: 1. a lot of the these money variables, are character vectors and not numerical 2. some variabels are not binary, which is okay, but we also see thery’re saved as characters - characters and factors are different, and bestglm doesn’t like characters, they like factors.

WIll look back at this next class

*How to find the variables for the logistical regression models* - Bestglm - backgwards - formard - stepwise

head(insurance)

## # A tibble: 6 x 26  
## INDEX TARGET\_FLAG TARGET\_AMT KIDSDRIV AGE HOMEKIDS YOJ INCOME PARENT1  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr> <chr>   
## 1 1 0 0 0 60 0 11 $67,349 No   
## 2 2 0 0 0 43 0 11 $91,449 No   
## 3 4 0 0 0 35 1 10 $16,039 No   
## 4 5 0 0 0 51 0 14 <NA> No   
## 5 6 0 0 0 50 0 NA $114,986 No   
## 6 7 1 2946 0 34 1 12 $125,301 Yes   
## # ... with 17 more variables: HOME\_VAL <chr>, MSTATUS <chr>, SEX <chr>,  
## # EDUCATION <chr>, JOB <chr>, TRAVTIME <dbl>, CAR\_USE <chr>, BLUEBOOK <chr>,  
## # TIF <dbl>, CAR\_TYPE <chr>, RED\_CAR <chr>, OLDCLAIM <chr>, CLM\_FREQ <dbl>,  
## # REVOKED <chr>, MVR\_PTS <dbl>, CAR\_AGE <dbl>, URBANICITY <chr>

**Issues with Insurance Data for bestglm**

* Stepwise Regression (Linear Regression) Basic idea: Alternate forward selection and backward elimination

1. Use forward selection to choose a new predictor and check its significance.
2. Use backward elimination to see if predictors already in the model can be dropped.

**Is there a package in R to automate this process?** Yes! The stepAIC function in the MASS package can be used. - But we dont learn how to use it yet

Your task is to investigate the stepAIC function to determine how it can be used to determine the best logistic regression model using the insurance data

Currency\_Convert <- function(Field){  
 Field <- as.numeric(gsub("\\$|,","", Field))  
}  
  
#Change factors to numbers  
insurance$HOME\_VAL\_num = Currency\_Convert(insurance$HOME\_VAL)  
insurance$INCOME\_num = Currency\_Convert(insurance$INCOME)  
insurance$BLUEBOOK\_num = Currency\_Convert(insurance$BLUEBOOK)  
insurance$OLDCLAIM\_num = Currency\_Convert(insurance$OLDCLAIM)  
  
#remove unneeded variables  
insurance.1 = within(insurance,   
 {INDEX = NULL  
 TARGET\_AMT = NULL  
 HOME\_VAL = NULL  
 INCOME = NULL   
 BLUEBOOK = NULL  
 OLDCLAIM = NULL})  
  
  
head(insurance.1)

## # A tibble: 6 x 24  
## TARGET\_FLAG KIDSDRIV AGE HOMEKIDS YOJ PARENT1 MSTATUS SEX EDUCATION   
## <dbl> <dbl> <dbl> <dbl> <dbl> <chr> <chr> <chr> <chr>   
## 1 0 0 60 0 11 No z\_No M PhD   
## 2 0 0 43 0 11 No z\_No M z\_High School  
## 3 0 0 35 1 10 No Yes z\_F z\_High School  
## 4 0 0 51 0 14 No Yes M <High School   
## 5 0 0 50 0 NA No Yes z\_F PhD   
## 6 1 0 34 1 12 Yes z\_No z\_F Bachelors   
## # ... with 15 more variables: JOB <chr>, TRAVTIME <dbl>, CAR\_USE <chr>,  
## # TIF <dbl>, CAR\_TYPE <chr>, RED\_CAR <chr>, CLM\_FREQ <dbl>, REVOKED <chr>,  
## # MVR\_PTS <dbl>, CAR\_AGE <dbl>, URBANICITY <chr>, HOME\_VAL\_num <dbl>,  
## # INCOME\_num <dbl>, BLUEBOOK\_num <dbl>, OLDCLAIM\_num <dbl>

insurance.2 = insurance.1[,c(2:24,1)]   
head(insurance.2)

## # A tibble: 6 x 24  
## KIDSDRIV AGE HOMEKIDS YOJ PARENT1 MSTATUS SEX EDUCATION JOB TRAVTIME  
## <dbl> <dbl> <dbl> <dbl> <chr> <chr> <chr> <chr> <chr> <dbl>  
## 1 0 60 0 11 No z\_No M PhD Profe~ 14  
## 2 0 43 0 11 No z\_No M z\_High Sc~ z\_Blu~ 22  
## 3 0 35 1 10 No Yes z\_F z\_High Sc~ Cleri~ 5  
## 4 0 51 0 14 No Yes M <High Sch~ z\_Blu~ 32  
## 5 0 50 0 NA No Yes z\_F PhD Doctor 36  
## 6 0 34 1 12 Yes z\_No z\_F Bachelors z\_Blu~ 46  
## # ... with 14 more variables: CAR\_USE <chr>, TIF <dbl>, CAR\_TYPE <chr>,  
## # RED\_CAR <chr>, CLM\_FREQ <dbl>, REVOKED <chr>, MVR\_PTS <dbl>, CAR\_AGE <dbl>,  
## # URBANICITY <chr>, HOME\_VAL\_num <dbl>, INCOME\_num <dbl>, BLUEBOOK\_num <dbl>,  
## # OLDCLAIM\_num <dbl>, TARGET\_FLAG <dbl>

#Sad trombone, because best gml wont run here  
insurance.2 = as.data.frame(insurance.2)  
#bestglm(insurance.2, family=binomial)

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

insurance.2.1 = insurance.2 %>% mutate\_if(is.character, factor)  
head(insurance.2.1)

## KIDSDRIV AGE HOMEKIDS YOJ PARENT1 MSTATUS SEX EDUCATION JOB  
## 1 0 60 0 11 No z\_No M PhD Professional  
## 2 0 43 0 11 No z\_No M z\_High School z\_Blue Collar  
## 3 0 35 1 10 No Yes z\_F z\_High School Clerical  
## 4 0 51 0 14 No Yes M <High School z\_Blue Collar  
## 5 0 50 0 NA No Yes z\_F PhD Doctor  
## 6 0 34 1 12 Yes z\_No z\_F Bachelors z\_Blue Collar  
## TRAVTIME CAR\_USE TIF CAR\_TYPE RED\_CAR CLM\_FREQ REVOKED MVR\_PTS CAR\_AGE  
## 1 14 Private 11 Minivan yes 2 No 3 18  
## 2 22 Commercial 1 Minivan yes 0 No 0 1  
## 3 5 Private 4 z\_SUV no 2 No 3 10  
## 4 32 Private 7 Minivan yes 0 No 0 6  
## 5 36 Private 1 z\_SUV no 2 Yes 3 17  
## 6 46 Commercial 1 Sports Car no 0 No 0 7  
## URBANICITY HOME\_VAL\_num INCOME\_num BLUEBOOK\_num OLDCLAIM\_num  
## 1 Highly Urban/ Urban 0 67349 14230 4461  
## 2 Highly Urban/ Urban 257252 91449 14940 0  
## 3 Highly Urban/ Urban 124191 16039 4010 38690  
## 4 Highly Urban/ Urban 306251 NA 15440 0  
## 5 Highly Urban/ Urban 243925 114986 18000 19217  
## 6 Highly Urban/ Urban 0 125301 17430 0  
## TARGET\_FLAG  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 1

#Sadder trombone; because bestglm won't run   
insurance.2.1 = as.data.frame(insurance.2.1)  
#bestglm(insurance.2.1, family=binomial)