qbs121_hw8_gibran

Gibran Erlangga

3/9/2022

1 Data Analysis: Propensity Scores

Use the dataset called "Teaching Hospital Outcome.txt". The outcome of interest is unfavorable discharge (discharge of a patient that is not to their home but to a nursing facility). The exposure of interest is whether the patient is in a teaching hospital or not. The covariates (potential confounders) are age, sex, race (white=referent, black, other), hispanic ethnicity, diabetes and hypertension.

```
data <- read.delim("Teaching Hospital Outcomes.txt")
head(data, 3)
```

```
##   UNFAVDX TeachingHospital Age Female Black Race.Other Hispanic DM HTN
## 1       0                1  42      0 FALSE      FALSE     TRUE  0   0
## 2       0                0  85      0 FALSE      FALSE    FALSE  0   1
## 3       0                1  47      0 FALSE      FALSE    FALSE  0   0
```

1. Estimate the odds ratio between unfavorable discharge and teaching hospital controlling for the other variables provided (e.g. potential confounding variables) using a logistic regression model.

```
logit <- glm(UNFAVDX ~ ., family=binomial, data=data)
summary(logit)
```

```
##
## Call:
## glm(formula = UNFAVDX ~ ., family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6748  -0.7856  -0.5782   0.9786   2.6297
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -4.47864    0.12452 -35.969  < 2e-16 ***
## TeachingHospital  0.57692    0.05689  10.141  < 2e-16 ***
## Age               0.04579    0.00192  23.845  < 2e-16 ***
## Female            0.38081    0.04849   7.854 4.02e-15 ***
## BlackTRUE         0.64034    0.09516   6.729 1.71e-11 ***
## Race.OtherTRUE    0.18131    0.09161   1.979  0.04781 *
## HispanicTRUE      0.13930    0.09046   1.540  0.12360
## DM                0.38362    0.06314   6.076 1.23e-09 ***
## HTN               0.16249    0.05309   3.061  0.00221 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11519  on 9999  degrees of freedom
## Residual deviance: 10391  on 9991  degrees of freedom
## AIC: 10409
##
## Number of Fisher Scoring iterations: 4
```

```
# odds ratio
print('OddsRatio')
```

```
## [1] "OddsRatio"
```

```
exp(logit$coefficients)
```

```
##      (Intercept) TeachingHospital              Age           Female
##       0.01134885       1.78054908       1.04684993       1.46347376
##        BlackTRUE   Race.OtherTRUE     HispanicTRUE               DM
##       1.89713242       1.19878415       1.14946549       1.46758554
##              HTN
##       1.17643363
```

2. Develop a prediction model for whether or not patients received care at a teaching hospital using the method of random Forests. Use this model to calculate the propensity of receiving care at a teaching hospital.

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.1.1
```

```
## randomForest 4.7-1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5          v purrr   0.3.4
## v tibble  3.1.5          v dplyr   1.0.7.9000
## v tidyr   1.1.4          v stringr 1.4.0
## v readr   2.0.2          v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.1.1
```

```
## Warning: package 'tibble' was built under R version 4.1.1
```

```
## Warning: package 'tidyr' was built under R version 4.1.1

## Warning: package 'readr' was built under R version 4.1.1

## Warning: package 'stringr' was built under R version 4.1.1

## Warning: package 'forcats' was built under R version 4.1.1

## -- Conflicts --------------------------------------- tidyverse_conflicts() --
## x dplyr::combine()  masks randomForest::combine()
## x dplyr::filter()   masks stats::filter()
## x dplyr::lag()      masks stats::lag()
## x ggplot2::margin() masks randomForest::margin()
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
data %>%
  clean_names() -> data_clean

# calculate propensity of TeachingHospital
rf <- randomForest(as.factor(teaching_hospital) ~ ., data=data_clean)
rf
```

```
##
## Call:
##  randomForest(formula = as.factor(teaching_hospital) ~ ., data = data_clean)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 28.37%
## Confusion matrix:
##   0    1 class.error
## 0 0 2837           1
## 1 0 7163           0
```

```
votes <- rf$votes
estimated <- votes[,2]
estimated %>% summary()
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.6875  0.9892  1.0000  0.9877  1.0000  1.0000
```

3. Create 10 bins of the propensity score (e.g deciles of the propensity scores). Estimate the odds ratio between unfavorable discharge and teaching hospital controlling for this categorical bin variable. Note: The random forest approach may not yield propensities with a alot of distinct values, which may lead to an error when the bins are created.

```
Bin = cut(estimated, unique(quantile(estimated, (0:10)/10, na.rm=TRUE)))

model_binned <- glm(unfavdx ~. + Bin , data=data_clean, family = binomial)
```

```
## Warning in terms.formula(formula, data = data): 'varlist' has changed (from
## nvar=9) to new 10 after EncodeVars() -- should no longer happen!
```

```
summary(model_binned)
```

```
##
## Call:
## glm(formula = unfavdx ~ . + Bin, family = binomial, data = data_clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5153  -0.7835  -0.5698   0.9132   2.7118
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -5.235740   0.171611 -30.509  < 2e-16 ***
## teaching_hospital  0.561015   0.057199   9.808  < 2e-16 ***
## age                0.047166   0.001978  23.847  < 2e-16 ***
## female             0.402021   0.048903   8.221  < 2e-16 ***
## blackTRUE          0.834569   0.102663   8.129 4.32e-16 ***
## race_otherTRUE     0.430668   0.102541   4.200 2.67e-05 ***
## hispanicTRUE       0.284106   0.095127   2.987  0.00282 **
## dm                 0.443393   0.066279   6.690 2.24e-11 ***
## htn                0.161626   0.053518   3.020  0.00253 **
## Bin(0.964,0.984]   0.499703   0.112002   4.462 8.14e-06 ***
## Bin(0.984,0.994]   0.915541   0.113369   8.076 6.70e-16 ***
## Bin(0.994,0.995]   1.089769   0.116726   9.336  < 2e-16 ***
## Bin(0.995,1]       0.607809   0.103379   5.879 4.12e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11519  on 9998  degrees of freedom
## Residual deviance: 10278  on 9986  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 10304
##
## Number of Fisher Scoring iterations: 4
```

```
exp(model_binned$coefficients)
```

```
##       (Intercept) teaching_hospital               age            female
```

```
##       0.005322882       1.752449714       1.048295503       1.494842293
##         blackTRUE    race_otherTRUE      hispanicTRUE                dm
##       2.303821963       1.538285093       1.328573347       1.557984050
##               htn  Bin(0.964,0.984]  Bin(0.984,0.994]  Bin(0.994,0.995]
##       1.175420512       1.648231029       2.498127109       2.973587921
##      Bin(0.995,1]
##       1.836404147
```

4. Calculate inverse propensity weights (after trimming the propensities so none is smaller than 0.01). Estimate the odds ratio between unfavorable discharge and teaching hospital using these weights.

```
# clip lower and upper values
estimated_norm <- ifelse(estimated < 0.1, 0.1, estimated)
estimated_norm <- ifelse(estimated > 0.9, 0.9, estimated)

inverse_prop_weights <- ifelse(data_clean$unfavdx, 1/estimated_norm, 1/(1-estimated_norm))
summary(inverse_prop_weights)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.111   1.111  10.000   7.605  10.000  10.000
```

```
glm <- glm(as.factor(unfavdx) ~ ., family=binomial, weights = inverse_prop_weights, data=data_clean)
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```
summary(glm)
```

```
##
## Call:
## glm(formula = as.factor(unfavdx) ~ ., family = binomial, data = data_clean,
##     weights = inverse_prop_weights)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6641  -0.8890  -0.6161   1.9844   3.5854
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -6.927671   0.104745 -66.139  < 2e-16 ***
## teaching_hospital  0.569818   0.046603  12.227  < 2e-16 ***
## age                0.049826   0.001572  31.692  < 2e-16 ***
## female             0.415184   0.038690  10.731  < 2e-16 ***
## blackTRUE          0.747090   0.071613  10.432  < 2e-16 ***
## race_otherTRUE     0.188440   0.072423   2.602  0.00927 **
## hispanicTRUE       0.222976   0.071344   3.125  0.00178 **
## dm                 0.406064   0.047716   8.510  < 2e-16 ***
## htn                0.166087   0.042341   3.923 8.76e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 24791  on 9999  degrees of freedom
## Residual deviance: 22952  on 9991  degrees of freedom
## AIC: 21221
##
## Number of Fisher Scoring iterations: 5
```

```
exp(glm$coefficients)
```

```
##       (Intercept) teaching_hospital              age           female
##      0.0009802811      1.7679458556      1.0510879485      1.5146492057
##         blackTRUE    race_otherTRUE      hispanicTRUE               dm
##      2.1108491354      1.2073642932      1.2497908567      1.5008987076
##               htn
##      1.1806763745
```