Mike Lehman

IS677 FA2016

Homework #7

#1  
polypharm <- read.csv("polypharm.csv")  
str(polypharm)

## 'data.frame': 3500 obs. of 14 variables:  
## $ id : int 1 1 1 1 1 1 1 2 2 2 ...  
## $ POLYPHARMACY: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ MHV4 : int 0 1 0 1 0 1 2 2 2 3 ...  
## $ INPTMHV3 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ YEAR : int 2002 2003 2004 2005 2006 2007 2008 2002 2003 2004 ...  
## $ GROUP : int 1 1 1 1 1 2 2 2 2 2 ...  
## $ URBAN : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ COMORBID : int 1 1 0 1 1 1 1 1 1 0 ...  
## $ ANYPRIM : int 1 1 0 1 1 1 1 1 1 1 ...  
## $ NUMPRIM : int 1 1 0 1 1 1 1 1 1 1 ...  
## $ GENDER : int 0 0 0 0 0 0 0 1 1 1 ...  
## $ RACE : int 0 0 0 0 0 0 0 1 1 1 ...  
## $ ETHNIC : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ AGE : num 4.67 5.67 6 7.08 8 ...

#2  
polypharm\_data <- polypharm[-1]  
str(polypharm\_data)

## 'data.frame': 3500 obs. of 13 variables:  
## $ POLYPHARMACY: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ MHV4 : int 0 1 0 1 0 1 2 2 2 3 ...  
## $ INPTMHV3 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ YEAR : int 2002 2003 2004 2005 2006 2007 2008 2002 2003 2004 ...  
## $ GROUP : int 1 1 1 1 1 2 2 2 2 2 ...  
## $ URBAN : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ COMORBID : int 1 1 0 1 1 1 1 1 1 0 ...  
## $ ANYPRIM : int 1 1 0 1 1 1 1 1 1 1 ...  
## $ NUMPRIM : int 1 1 0 1 1 1 1 1 1 1 ...  
## $ GENDER : int 0 0 0 0 0 0 0 1 1 1 ...  
## $ RACE : int 0 0 0 0 0 0 0 1 1 1 ...  
## $ ETHNIC : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ AGE : num 4.67 5.67 6 7.08 8 ...

#3  
polypharm\_data$YEAR <- factor(polypharm\_data$YEAR)  
polypharm\_data$ETHNIC <- factor(polypharm\_data$ETHNIC)  
polypharm\_data$GENDER <- factor(polypharm\_data$GENDER)  
polypharm\_data$URBAN <- factor(polypharm\_data$URBAN)  
polypharm\_data$RACE <- factor(polypharm\_data$RACE)  
polypharm\_data$COMORBID <- factor(polypharm\_data$COMORBID)  
polypharm\_data$GROUP <- factor(polypharm\_data$GROUP)  
polypharm\_data$MHV4 <- factor(polypharm\_data$MHV4)  
polypharm\_data$INPTMHV3 <- factor(polypharm\_data$INPTMHV3)  
str(polypharm\_data)

## 'data.frame': 3500 obs. of 13 variables:  
## $ POLYPHARMACY: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ MHV4 : Factor w/ 4 levels "0","1","2","3": 1 2 1 2 1 2 3 3 3 4 ...  
## $ INPTMHV3 : Factor w/ 3 levels "0","1","2": 1 1 1 1 1 1 1 1 1 1 ...  
## $ YEAR : Factor w/ 7 levels "2002","2003",..: 1 2 3 4 5 6 7 1 2 3 ...  
## $ GROUP : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 2 2 2 2 2 ...  
## $ URBAN : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ COMORBID : Factor w/ 2 levels "0","1": 2 2 1 2 2 2 2 2 2 1 ...  
## $ ANYPRIM : int 1 1 0 1 1 1 1 1 1 1 ...  
## $ NUMPRIM : int 1 1 0 1 1 1 1 1 1 1 ...  
## $ GENDER : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 2 2 ...  
## $ RACE : Factor w/ 3 levels "0","1","2": 1 1 1 1 1 1 1 2 2 2 ...  
## $ ETHNIC : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ AGE : num 4.67 5.67 6 7.08 8 ...

We have to recode several of the variables to factors to better model their relationships to the target variable. If they are left as integer variables the model cannot accurately measure the impact that any of the values accurately have on the target variable. Also, several of these variables are inherently nominal and would incorrectly inform the model using values of 1 and 0. A GENDER of 1 is not inherently greater than a GENDER of 0.

For example, by reassigning these variables to factors, we can better understand how a difference of GENDER or URBAN has an effect on POLYPHARMACY.

#4  
set.seed(123)  
train\_sample <- sample(3500,1750)  
  
poly\_train <- polypharm\_data[train\_sample,]  
poly\_test <- polypharm\_data[-train\_sample,]

#5  
y <- "POLYPHARMACY"  
x <- c('MHV4','INPTMHV3','YEAR','GROUP','URBAN','COMORBID','ANYPRIM','NUMPRIM','GENDER','RACE','ETHNIC','AGE')  
poly\_log <- paste(y, paste(x, collapse = "+"), sep = "~")

#6  
model <- glm(poly\_log, family = binomial(link='logit'), data = poly\_train)

#7  
summary(model)

##   
## Call:  
## glm(formula = poly\_log, family = binomial(link = "logit"), data = poly\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6825 -0.7239 -0.4755 -0.2203 2.5478   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.736469 0.436882 -8.553 < 2e-16 \*\*\*  
## MHV41 0.506818 0.277162 1.829 0.06746 .   
## MHV42 1.469296 0.263862 5.568 2.57e-08 \*\*\*  
## MHV43 2.201334 0.259184 8.493 < 2e-16 \*\*\*  
## INPTMHV31 0.577668 0.287286 2.011 0.04435 \*   
## INPTMHV32 0.783335 0.448790 1.745 0.08091 .   
## YEAR2003 0.210381 0.274921 0.765 0.44413   
## YEAR2004 1.072392 0.255358 4.200 2.67e-05 \*\*\*  
## YEAR2005 0.816096 0.266940 3.057 0.00223 \*\*   
## YEAR2006 1.098195 0.275573 3.985 6.74e-05 \*\*\*  
## YEAR2007 0.929077 0.301847 3.078 0.00208 \*\*   
## YEAR2008 1.211335 0.310012 3.907 9.33e-05 \*\*\*  
## GROUP2 0.407796 0.140900 2.894 0.00380 \*\*   
## GROUP3 0.583566 0.184912 3.156 0.00160 \*\*   
## URBAN1 -0.439249 0.144981 -3.030 0.00245 \*\*   
## COMORBID1 -0.195795 0.203547 -0.962 0.33609   
## ANYPRIM 0.199672 0.310249 0.644 0.51984   
## NUMPRIM -0.258505 0.270663 -0.955 0.33954   
## GENDER1 0.487926 0.164194 2.972 0.00296 \*\*   
## RACE1 -0.882788 0.190325 -4.638 3.51e-06 \*\*\*  
## RACE2 -0.810104 0.887184 -0.913 0.36118   
## ETHNIC1 -0.161382 0.689577 -0.234 0.81496   
## AGE 0.003509 0.030987 0.113 0.90984   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1893.5 on 1749 degrees of freedom  
## Residual deviance: 1613.7 on 1727 degrees of freedom  
## AIC: 1659.7  
##   
## Number of Fisher Scoring iterations: 7

coefficients(model)

## (Intercept) MHV41 MHV42 MHV43 INPTMHV31   
## -3.736468648 0.506817568 1.469295568 2.201333821 0.577667803   
## INPTMHV32 YEAR2003 YEAR2004 YEAR2005 YEAR2006   
## 0.783334589 0.210380945 1.072391710 0.816096490 1.098195123   
## YEAR2007 YEAR2008 GROUP2 GROUP3 URBAN1   
## 0.929076893 1.211334589 0.407795665 0.583566490 -0.439249218   
## COMORBID1 ANYPRIM NUMPRIM GENDER1 RACE1   
## -0.195794943 0.199672427 -0.258504995 0.487925877 -0.882788312   
## RACE2 ETHNIC1 AGE   
## -0.810104384 -0.161382193 0.003508936

1. The variable MHV4 is quite significant. Not only do we see a low p value from the summary output for each level (level 1 is over the 0.05 threshold for general significance but other levels are lower), we can also see that the coefficient is quite high for level three MHV4 and is also noticably high for first and second levels.
2. MHV40 is the refernece level. When we converted MHV4 to a factor, R treated the first level as the reference level. This can also be observed because MHV40 is not listed in the outputs for the model.
3. The odds of POLYPHARMACY for an individual with more than 14 outpatient mental health visits (MHV4) is: 2.201333821 greater than an individual with 0 visits. This can be observed by the MHV43 level coefficient value. Since MHV40 is the reference level it is 0.

#8  
poly\_test$PRED <- predict(model, newdata = poly\_test, type = "response")  
str(poly\_test)

## 'data.frame': 1750 obs. of 14 variables:  
## $ POLYPHARMACY: int 0 0 0 0 0 0 0 1 1 1 ...  
## $ MHV4 : Factor w/ 4 levels "0","1","2","3": 1 1 3 3 3 3 3 4 4 4 ...  
## $ INPTMHV3 : Factor w/ 3 levels "0","1","2": 1 1 1 1 1 1 1 1 3 2 ...  
## $ YEAR : Factor w/ 7 levels "2002","2003",..: 1 5 7 1 2 6 7 1 2 3 ...  
## $ GROUP : Factor w/ 3 levels "1","2","3": 1 1 2 2 2 1 1 2 2 2 ...  
## $ URBAN : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ COMORBID : Factor w/ 2 levels "0","1": 2 2 2 2 2 1 1 1 1 1 ...  
## $ ANYPRIM : int 1 1 1 1 1 0 1 0 1 0 ...  
## $ NUMPRIM : int 1 1 1 1 1 0 1 0 1 0 ...  
## $ GENDER : Factor w/ 2 levels "0","1": 1 1 1 2 2 2 2 2 2 2 ...  
## $ RACE : Factor w/ 3 levels "0","1","2": 1 1 1 2 2 2 2 1 1 1 ...  
## $ ETHNIC : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ AGE : num 4.67 8 10.67 7.58 8.08 ...  
## $ PRED : num 0.0184 0.0539 0.2962 0.0771 0.0936 ...

#9  
# need to pass na.rm = TRUE to sum to account for any NA's  
loglikelihood <- function(y, py){  
 sum(y \* log(py) + (1-y)\*log(1 - py), na.rm = TRUE)  
}

#10  
testy <- as.numeric(poly\_test$POLYPHARMACY)  
testpred <- predict(model, newdata = poly\_test, type="response")  
pnull.test <- mean(testy)  
null.dev.test <- -2\*loglikelihood(testy, pnull.test)  
resid.dev.test <- -2\*loglikelihood(testy, testpred)  
  
null.dev.test

## [1] 1914.834

resid.dev.test

## [1] 1740.852

# difference between null and residual deviances  
delDev <- null.dev.test - resid.dev.test  
  
# difference in degrees of freedom between null and model  
df.null <- dim(poly\_test)[[1]] - 1  
df.model <- dim(poly\_test)[[1]] - length(model$coefficients)  
deldf <- df.null - df.model  
  
# obtain p value using chi squared distribution  
p <- pchisq(delDev, deldf, lower.tail=F)  
p

## [1] 1.280637e-25

The model prediction is better than guessing the mean for all values (null model) as the deviance is smaller.

OPTIONAL: By examining the difference between the null and residual deviances (delDev) and the difference in degrees of freedom between the null value and model (deldf), we can see that we get a low p value when applying chi squared distribution.

With a relatively low p value, we can be reasonably sure that the likelihood of the model performing better than the null hyopthesis was not by chance.