TreatmentTolerability

Achintya Varma

5/17/2019

This Rmarkdown file is to study the effects of several patient characteristics of Head and Neck Cancer Patients and how they tolerate treatment corresponding to the characteristics in question. The investigators of this study are primarily interested in studying the difference of Treatment Tolerability across sites while controlling for factors such as ECOG(clincical score- ordrinal variable), age(coded as binary), treatment type(nominal with two variables combined, total= 2 x 2 levels), CCI, polypharmacy(binary), Potentially Inappropriate Medications(Binary),drinking history(ordinal), tobacco use(ordinal)

General Format of an Rmarkdown file is the statistical analysis or the code is followed by comments from the person creating the file.

We start the analysis by testing whether any of our covariates are correlated to each other, any correlation implies that our the estimation of the effect of our predictors on the the response will be biased and any inferences based on the general linear model are incorrect and needed to be assessed after removal of correlated variables. Since Site is the primary variable of interest clinically, we remove any variables correlated with Site

HND621 <- read\_excel("~/Desktop/CapstoneRevised/HND621.xlsx")  
  
f.test<-function(x){  
 if(length(unique(x[[1]])) >2 | length(unique(x[[2]]))>2){  
 set.seed(1234)  
 f<-fisher.test(x[[1]],x[[2]],simulate.p.value=T)  
 }  
 else{  
 f<-fisher.test(x[[1]],x[[2]])  
 }  
 return(round(f$p.value,4))  
}   
  
names\_add<-function(x){  
 null<-""  
 for(y in x){  
   
 }  
}  
fisher\_names<-function(x){  
 stringr = ""  
 for(i in x){  
   
 stringr = paste(stringr, " ", i)  
 }  
 return(stringr)  
}  
Variables<-as.data.frame(cbind(HND621$record\_id,HND621$site,HND621$ecog,HND621$cci\_excel,HND621$stagefinal8,HND621$interruption,HND621$age\_at\_rt\_start))  
  
colnames(Variables)<-c("rec\_id","site","ecog","cci","stage","interruption","age")  
   
names\_c<-colnames(Variables)  
# code to convert site numbers to site list  
site\_names<-list("1"="Larynx","2"="Lip+Oral","3"="Major\_Salivary\_Glands","4"="Nasopharynx","5"="Oropharynx","6"="Hypopharynx","7"="Nasal\_Sinuses")  
  
Variables<-cbind(Variables,t(as.data.frame(site\_names[Variables$site]))) #uses the list above to convert the "numeric" values of the tumor sites to character type  
   
colnames(Variables)<-c(names\_c,"site\_name")  
#remove Major Salivary Glands and Nasal Sinuse  
Variables1<-Variables[Variables$site\_name!="Major\_Salivary\_Glands",]  
Variables<-Variables1[Variables1$site\_name!="Nasal\_Sinuses",]  
Variables$site\_name<-droplevels(Variables)$site\_name #drops tumor sites we are not interested in analyzing  
Variables$site\_name<-relevel(Variables$site\_name,ref="Oropharynx") #Sets the reference to the biggest site in our dataset  
Variables$cci<-ifelse(Variables$cci>6,7,ifelse(Variables$cci>5,6,5)) #Cci greater than 7 is counted as one category  
Variables$ecog<-ifelse(Variables$ecog>1,2,ifelse(Variables$ecog>0,1,0))#Ecog higher than 2 is counted as a single score  
  
Variables$age\_cat<-ifelse(Variables$age>80,1,0) #Categorizing age into 80+ vs 70-80  
Variables$stage<-ifelse(Variables$stage>1,1,0)  
Variables<-Variables[,c(1,3,4,5,6,8,9)] #Drops "Site" and "age" as we are interested in categorizing age and the character values of site  
head(Variables)

## rec\_id ecog cci stage interruption site\_name age\_cat  
## X1 105 1 7 0 1 Larynx 0  
## X5 113 1 6 0 0 Oropharynx 1  
## X1.1 239 1 6 0 1 Larynx 0  
## X5.1 242 2 7 1 1 Oropharynx 0  
## X1.2 322 1 6 0 0 Larynx 1  
## X5.2 385 1 5 1 0 Oropharynx 0

Variables[,2:ncol(Variables)]<-apply(Variables[,2:ncol(Variables)],2,as.character) #convert all the predictors to characters as it is helpful to have characters to conduct logistic regressions  
  
#a\_tox<-a\_tox[a\_tox$rec\_id %in% Variables$rec\_id,]# Make sure the data frame with all our response variables has the same rows as our dataframes with all the predictors, a\_tox was defined in a different .r file from the same project and is in the workspace

The first correlation is f\_rt\_site, it is testing whether there is an association between a patient having a tumor in any one of the sites listed and the treatment they had to undergo, a p-value of 2.735e-11 is highly significant and shows that the two variables have a very high likelihood of being associated. For the sake of obtaining unbiased estimates of our regression it is advisable not to use both sites and radiation intent in building any models. mAge and Site are not correlated with a p value of 0.3182 for the fisher test.

Site and Ecog have a simulated p value of 0.0109, the p value might not be big enough to warrant adjustment for covariance. Since Fisher’s test is strong in detecting in even minor associations, we could change the cut-off pvalues to a lower value such as 0.01, for the sake of continuity we should consider all assocations to be valid if the p-value is lower than that.

We can utilize the apply family functions to see multiple correlations at the same time in a list. This is for the convenience of the programmer so you can skip reading the following chunk.

cov\_mat<-matrix(replicate(0,36),nrow=6,ncol=6)  
cov\_mat\_names<-matrix(replicate(0,36),nrow=6,ncol=6)  
combos<-combn(Variables[,2:ncol(Variables)],2)  
combos\_df<-as.data.frame(combos)  
combos\_name<-combn(Variables[,2:ncol(Variables)],2,function(x){names(x)})  
combos\_name<-apply(combos\_name,2,fisher\_names)  
#combos\_t<-combn(Variables[,2:ncol(Variables)],5)  
combos\_p<-lapply(combos\_df,f.test)  
names(combos\_p)<-combos\_name  
cov\_mat[lower.tri(cov\_mat)]<-combos\_p  
cov\_mat\_names[lower.tri(cov\_mat)]<-combos\_name  
colnames(cov\_mat)<-c("ecog","cci","stage","interruption","site\_name","age\_cat")  
rownames(cov\_mat)<-colnames(cov\_mat)  
  
  
#var\_mod<-inner\_join(Variables,a\_tox,by="rec\_id")  
#musc\_glm\_mv<-glm(musc~site\_name+age\_cat,data=var\_mod,family="binomial")  
#print(pf\_df)

The Code following prints out the correlations matrix tested with Fisher Test.

We used Fisher’s test to test the associations between different predictors in our analysis as some of the cells might be too sparse to use Chi-Square tests.

## ecog cci stage interruption site\_name age\_cat  
## ecog NULL NULL NULL NULL NULL NULL   
## cci 5e-04 NULL NULL NULL NULL NULL   
## stage 5e-04 0.4988 NULL NULL NULL NULL   
## interruption 0.7221 0.2434 0.0067 NULL NULL NULL   
## site\_name 0.021 0.0665 0.0225 0.009 NULL NULL   
## age\_cat 0.0315 5e-04 0.1189 0.4895 0.3188 NULL