HomeWork\_6

Joel parker

March 30, 2020

1.) The goal of the analysis is to generate the appropriate model for tumor penetration of the prostatic capsule versus the baseline exam values. Assess whether the assumptions of the logistic regression are met.

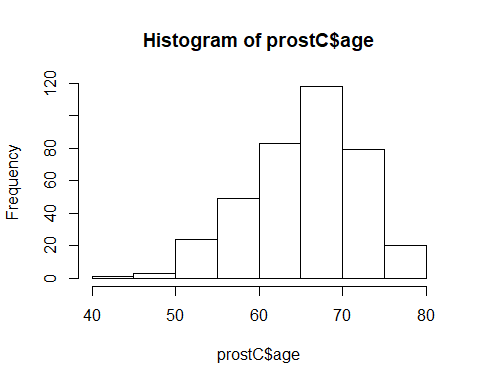
library(haven)  
prostC <- read\_dta('hw6.dta')  
  
  
#### First we will look at discriptive statistics.   
###dependent var  
rbind(table(prostC$capsule),prop.table(table(prostC$capsule))) ### resonable probs

## 0 1  
## [1,] 224.0000000 153.0000000  
## [2,] 0.5941645 0.4058355

summary(prostC$age) ### approximatly normal.

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 43.00 62.00 67.00 66.04 71.00 79.00

hist(prostC$age) ### approximatly normal.



####Race  
rbind(table(prostC$race), prop.table(table(prostC$race))) ### Most of the participants are white.

## 1 2  
## [1,] 338.0000000 36.00000000  
## [2,] 0.9037433 0.09625668

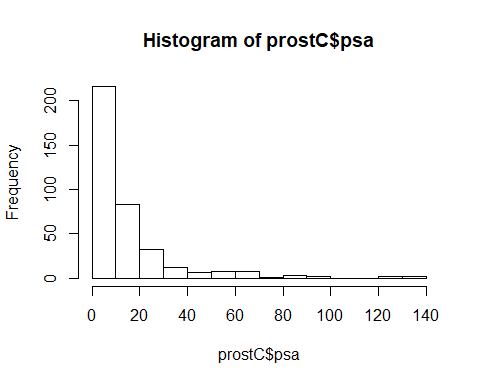
###dpros  
rbind(table(prostC$dpros), prop.table(table(prostC$dpros))) ### 4 may be under represented.

## 1 2 3 4  
## [1,] 96.0000000 132.0000000 96.0000000 53.0000000  
## [2,] 0.2546419 0.3501326 0.2546419 0.1405836

### PSA  
summary(prostC$psa) ## apprears to be very right skewed.

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.30 5.00 8.80 15.47 17.20 139.70

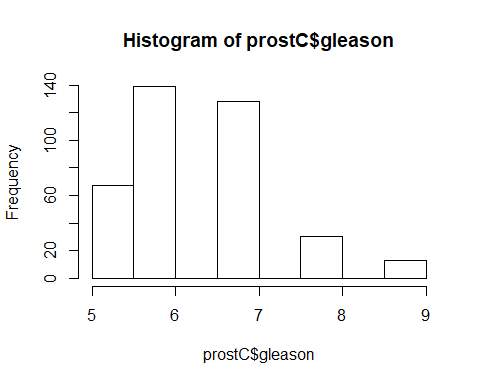
hist(prostC$psa)



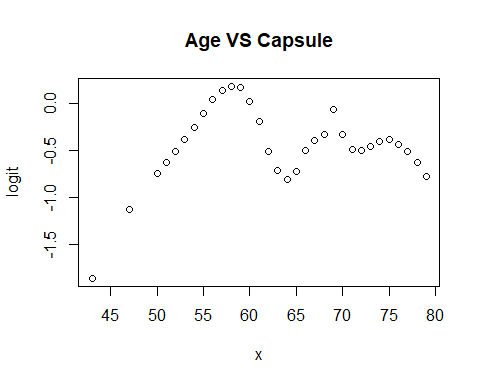
### Gleason Score  
rbind(table(prostC$gleason), prop.table(table(prostC$gleason)))

## 5 6 7 8 9  
## [1,] 67.0000000 139.0000000 128.0000000 30.0000000 13.00000000  
## [2,] 0.1777188 0.3687003 0.3395225 0.0795756 0.03448276

hist(prostC$gleason)



#Create logit loess function  
logitloess <-function(x, y, s, p) {  
 logit <-function(pr) {log(pr/(1-pr))}  
 if(missing(s)) {locspan <- 0.8}  
 else{locspan <- s}  
 loessfit <-predict(loess(y~x,span=locspan))  
 pi <-pmax(pmin(loessfit,0.9999),0.0001)  
 logitfitted <-logit(pi)  
 plot(x, logitfitted, ylab="logit", main = p)}  
  
  
  
## age vs Capsule  
logitloess(prostC$age, prostC$capsule, .4, "Age VS Capsule")



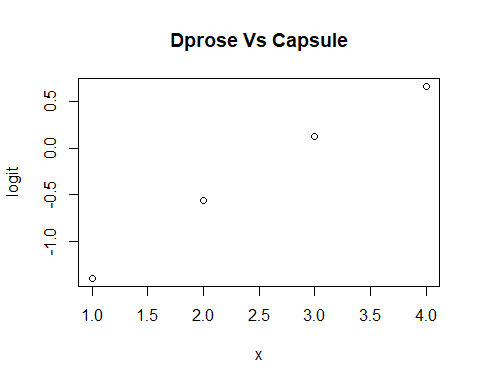
#prostC\_r <- prostC[!is.na(prostC$race),]  
#logitloess(prostC\_r$race, prostC\_r$capsule, .4, "Race vs Capsule")  
  
  
logitloess(prostC$dpros, prostC$capsule, .4, "Dprose Vs Capsule")

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : pseudoinverse used at 0.985

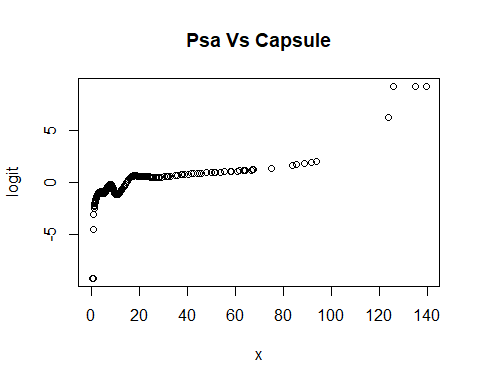
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : neighborhood radius 1.015

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : reciprocal condition number 0

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : There are other near singularities as well. 4.0602



logitloess(prostC$psa, prostC$capsule, .4, "Psa Vs Capsule")



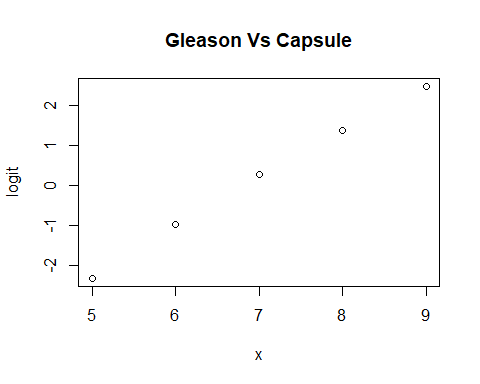
logitloess(prostC$gleason, prostC$capsule,.4, "Gleason Vs Capsule")

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : pseudoinverse used at 4.98

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : neighborhood radius 1.02

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : reciprocal condition number 6.4515e-031

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : There are other near singularities as well. 4.0804



* **capsule** The probeblity of peretration is not to extreme.
* **age** Looking at the hist. Age appears to be approximatly normal. However, looking at our lowess scatter plot age does not have linearity of the log odds ration. We would need to have have a transformation of age.
* **race** 90% of the participants are white.
* **Dprose** does have a linear relationship.
* **PSA** does not have a linear relationship and will need a transoformation.
* **gleason** has a linear relationship. But has a low percentage of participants in catagories 5 and 8.
* **independence** All study participantes are independent.

2.) Generate the approprate model for tumor penetration of prostatic capsule versus the baseline exam values. Consider the purpose of your model to determine the most appropriate method to determine which variables should be included.

mod\_1 <- glm(capsule~age+as.factor(race)+as.factor(dpros)+ psa+ gleason, data = prostC, family = binomial(link = 'logit'))  
  
  
lo\_select <- step(mod\_1, direction = "backward")

## Start: AIC=394.38  
## capsule ~ age + as.factor(race) + as.factor(dpros) + psa + gleason  
##   
## Df Deviance AIC  
## - age 1 379.04 393.04  
## <none> 378.38 394.38  
## - as.factor(race) 1 380.68 394.68  
## - psa 1 389.75 403.75  
## - as.factor(dpros) 3 400.14 410.14  
## - gleason 1 424.54 438.54  
##   
## Step: AIC=393.04  
## capsule ~ as.factor(race) + as.factor(dpros) + psa + gleason  
##   
## Df Deviance AIC  
## <none> 379.04 393.04  
## - as.factor(race) 1 381.17 393.17  
## - psa 1 390.57 402.57  
## - as.factor(dpros) 3 401.91 409.91  
## - gleason 1 424.75 436.75

summary(lo\_select)

##   
## Call:  
## glm(formula = capsule ~ as.factor(race) + as.factor(dpros) +   
## psa + gleason, family = binomial(link = "logit"), data = prostC)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.3979 -0.7722 -0.4428 0.9001 2.4517   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -8.113290 1.065909 -7.612 2.71e-14 \*\*\*  
## as.factor(race)2 -0.660431 0.459387 -1.438 0.150537   
## as.factor(dpros)2 0.784124 0.357241 2.195 0.028167 \*   
## as.factor(dpros)3 1.577349 0.372719 4.232 2.32e-05 \*\*\*  
## as.factor(dpros)4 1.497965 0.452761 3.309 0.000938 \*\*\*  
## psa 0.030178 0.009832 3.069 0.002145 \*\*   
## gleason 0.990689 0.162550 6.095 1.10e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 504.53 on 373 degrees of freedom  
## Residual deviance: 379.04 on 367 degrees of freedom  
## (3 observations deleted due to missingness)  
## AIC: 393.04  
##   
## Number of Fisher Scoring iterations: 5

* the only variable removed was age. Our model will be.

3.) Some studies in the literature have identified an interaction between psa and gleason score. Assess whether an interaction term is needed in the model you selected above.

prostC$inter <- prostC$psa \* prostC$gleason  
mod\_iter <- glm(capsule~as.factor(race)+as.factor(dpros)+psa+gleason+ inter, data = prostC, family = binomial(link = "logit"))  
  
library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

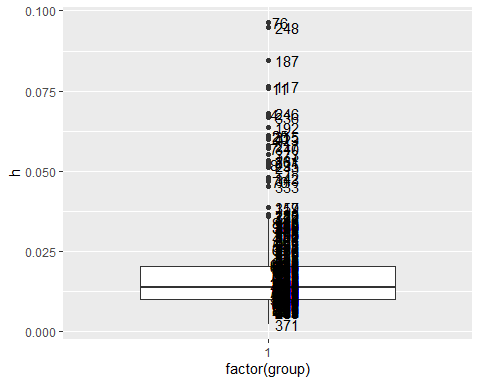
## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

lrtest(lo\_select, mod\_iter)

## Likelihood ratio test  
##   
## Model 1: capsule ~ as.factor(race) + as.factor(dpros) + psa + gleason  
## Model 2: capsule ~ as.factor(race) + as.factor(dpros) + psa + gleason +   
## inter  
## #Df LogLik Df Chisq Pr(>Chisq)  
## 1 7 -189.52   
## 2 8 -189.49 1 0.0746 0.7848

* Looking at the likelyhood ration between the two models (with and without the interaction term) : There is no differenct between models and there is a difference between models. we get a pvalue of .78. Thus there is not enough evidence to suggest that an interaction term is needed.

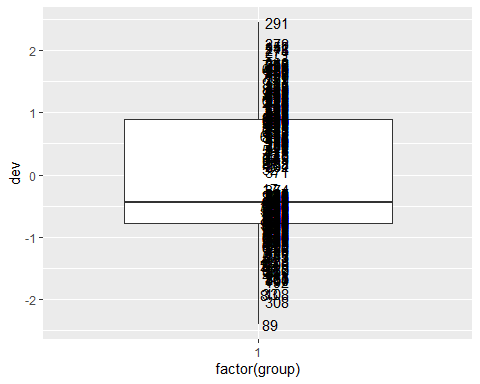
4.) Determine the outliers in X and outliers in Y for you final model. State why they are outliers.



#show h>.06  
list\_id <- hat\_df[hat\_df$h>.06,]$id  
list\_id <- as.numeric(levels(list\_id))[list\_id]  
lever\_df <-cbind(prostC[prostC$id%in%c(list\_id),-8], hat\_df[hat\_df$h>.06,-3])  
lever\_df[order(lever\_df$h),]

## id capsule age race dpros psa gleason id h  
## 20 20 1 67 2 3 8.6 7 20 0.06039812  
## 315 315 1 57 2 3 7.8 7 315 0.06093394  
## 25 25 1 77 1 1 61.1 7 25 0.06100218  
## 192 192 0 61 1 4 61.6 6 192 0.06349416  
## 336 336 1 56 1 2 58.0 6 336 0.06684854  
## 4 4 0 76 2 2 51.2 7 4 0.06757603  
## 246 246 1 70 2 4 13.2 7 246 0.06794961  
## 11 11 1 68 2 4 4.0 7 11 0.07588804  
## 117 117 1 66 2 4 45.3 6 117 0.07637193  
## 187 187 1 68 1 1 85.4 7 187 0.08460282  
## 248 248 1 57 2 1 63.3 7 248 0.09483569  
## 76 76 0 54 2 1 64.3 7 76 0.09643203

* most of the indiduals with high leverage are african american with high psa.



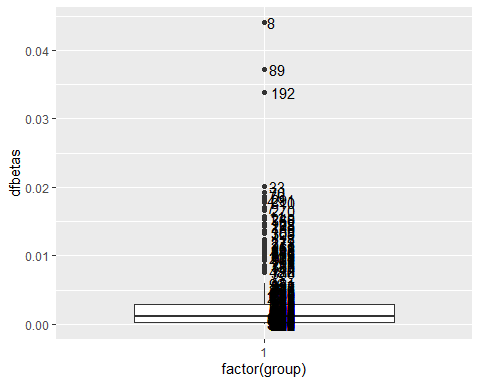
##show dev >2.75  
list\_id <- dev\_df[dev\_df$dev>2.75,]$id  
list\_id <- as.numeric(levels(list\_id))[list\_id]  
prostC[prostC$id %in% c(list\_id),]

## # A tibble: 0 x 8  
## # ... with 8 variables: id <dbl>, capsule <dbl>, age <dbl>, race <dbl>,  
## # dpros <dbl>, psa <dbl>, gleason <dbl>, inter <dbl>

* There are not any outliers in Y.

5.) Determine the influential points for your final model. State why they are influential.

dfbeta\_df <- cbind(names(lo\_select$fitted.values), cooks.distance(lo\_select))  
dfbeta\_df <- data.frame(dfbeta\_df)  
dfbeta\_df$group <- 1  
colnames(dfbeta\_df)[c(1,2)] <- c("id", "dfbetas")  
dfbeta\_df$dfbetas <- as.numeric(levels(dfbeta\_df$dfbetas))[dfbeta\_df$dfbetas]  
ggplot(dfbeta\_df, aes(x=factor(group), y=dfbetas)) +  
 geom\_boxplot()+  
 geom\_text(aes(label =id), na.rm = TRUE, hjust=-.3)



# show cooks distance greater than .02  
  
list\_id <- dfbeta\_df[dfbeta\_df$dfbetas>.02,]$id  
list\_id <- as.numeric(levels(list\_id))[list\_id]  
prostC[prostC$id%in% c(list\_id),]

## # A tibble: 4 x 8  
## id capsule age race dpros psa gleason inter  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 8 0 61 2 4 66.7 7 467.  
## 2 33 0 77 1 4 11 8 88   
## 3 89 0 68 1 4 17.1 9 154.  
## 4 192 0 61 1 4 61.6 6 370.

6.) Assess the goodness of fit of your final model.

library(ResourceSelection)

## ResourceSelection 0.3-5 2019-07-22

#H0: the model fits the data.   
hoslem.test(prostC[!is.na(prostC$race),]$capsule, fitted(lo\_select), g=10)

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: prostC[!is.na(prostC$race), ]$capsule, fitted(lo\_select)  
## X-squared = 8.1009, df = 8, p-value = 0.4237

* H0: the model fits well. H1: the model does not fit well. Since our p-value is .42 we can conclude our model fits the data.

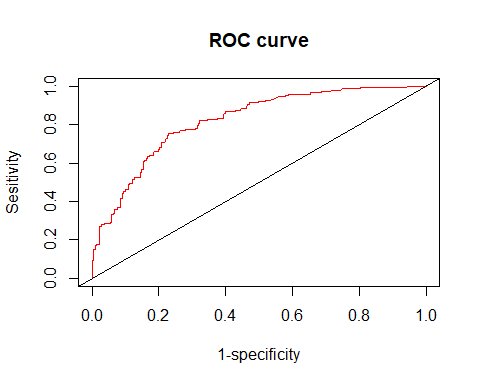
prob <-predict(lo\_select, type = c("response"))  
library(ROCR)

## Loading required package: gplots

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

listR<- prostC[!is.na(prostC$race), ]  
pred <- prediction(prob, as.numeric(listR$capsule))  
perf <- performance(pred, measure = "tpr", x.measure = "fpr")  
plot(perf, col = rainbow(7), main = "ROC curve", xlab= '1-specificity', ylab = 'Sesitivity')  
abline(0,1)



### get AUC  
auc = performance(pred, "auc")  
auc@y.values

## [[1]]  
## [1] 0.8204199

7.) Describe the results of your model in language understandable to a scientific audience.

exp(lo\_select$coefficients)

## (Intercept) as.factor(race)2 as.factor(dpros)2 as.factor(dpros)3   
## 0.0002995319 0.5166284930 2.1904881999 4.8421029192   
## as.factor(dpros)4 psa gleason   
## 4.4725796566 1.0306384043 2.6930906118

* **Summary** We used backwards elimination to determine our final model. We used the criteria of p-value less than .2 for a variable to be included in our model. The first model fit used Age, Race, Results of the Digital Rectal exam, Prostate Specific Anigen Value and Total Gleason Score as predictors of Tumor Penetration prostatic Capsule. Age, PSA, and gleason score were modeled as continuouse variable. Race and Results of Digital Rectal Exam we modeled as categorical variables with white men and No nodule as the referance group respectively. The only variable eliminated from our model was age. African american had an odds ratio of .52 when compared to white men. Uniobar Nodule (Left), Uniobar Nodule (Right) and Biobar Nodule had odds ratios of 2.19, 4.84, and 4.47 resectively when compared to No module. A one unit increase in PSA have an odds ration of 1.03 and a one unit increase in glease score had an odds ratio of 2.69. The model had a good discrimination with an area under the ROC curve of 0.82.