Investigating the Variables Influencing Tumor Penetration of Prostatic Capsule

Blinded

November 12, 2020

Executive Summary: Prostate cancer is a cancer that occurs in the prostate, a small walnut-shaped gland in men that produces the seminal fluid that nourishes and transports sperm. It is one of the most common types of cancer in men. It grows slowly and can often be treated if detected early enough. The Ohio State University Comprehensive Cancer Center want to know if baseline exam measurements can predict whether a tumor has penetrated the prostatic capsule. Penetration means the patient has high-risk of prostate cancer. Then the corresponding screening test and treatment plan can be done as early as enough for those have high-risk patients. Therefore, a logistic regression was selected to check which factors can be used to predict the status of penetration. The final model shows us those significant factors are prostatic specific antigen value (Psa), total gleason score (Gleason), and results of digital rectal exam (Dpros). The predicting accuracy, sensitivity, and specificity of our final model are 0.78, 0.77, and 0.79 at 0.36 cut off. All coefficients of significant predictors are positive, this means that an increase in Psa, (or Gleason or Dpros) is associated with an increase in the probability of being penetration. The results show that 3% more likely to have capsular penetration when Psa increase one unit. 3.22 times more likely to have capsular penetration when gleason increase one unit. Comparing to a patient with no nodule, a patient with bilobar nodule is 4.53 times, a patient with right unilobar nodule is 4.10 times, and a patient with left unilobar nodule is 2.01 times more likely to have capsular penetration respectively.

1 Introduction

Prostate cancer is the second most found cancer in the world and the sixth most common cause of cancer death in men, accounting for 14% (903 500) of total new cancer cases and 6% (258 400) of total cancer deaths in men in 2008 [1]. What exactly causes the initiation and progression of prostate cancer is still unidentified at the moment, but a number of studies have mentioned that genetic, race, diet and environmental factors play important roles [2]-[4]. The objective of this study is to determine if baseline exam measurements can predict whether a tumor has penetrated the prostatic capsule. Penetration means high-risk of prostate cancer.

Therefore, a logistic regression model was built between response variable and six potential predictors to find a model which might be useful to predict the penetration status. The patient may have high-risk of prostate cancer if the tumor has penetrated the prostatic capsule. Then related screening and treatment plan can be performed for those high-risk prostate cancer patients as early as possible. It is meaningful for those patients because prostate cancer can often be treated if detected early enough.

2 Methods

The data set contains 380 observations and 8 columns with 3 missing values in race. So, we will delete those missing values and keep 377 observations for data analysis. Table 1 presents the response variable and six potential predictors. We will randomly split the data into training set (70% of the whole data set) and testing set (30% of the whole data set). Training set which has 263 observations will be used to fit our logistical model, and testing set which has 114 observations will be used to assess predicting performance of our final model.

Table 1: Response variable and six potential predictors

Response Variable	Capsule	Tumor penetration of prostatic capsule
Predictors	Age	Age of subject
	Race	Race of subject
	Dpros	Result of digital rectal exam
	Dcaps	Detection of capsular involvement
	Psa	Prostatic specific antigen value
	Gleason	Total gleason score

Capsule is our binary response variable which denote the status of tumor penetration of prostatic capsule. Age, Psa, and Gleason are continuous predictors. Race, Dpros, and Dcaps are categoric predictors. Race has two levels: white and black. Dpros has four levels: no nodule, left unilobar nodule, right unilobar nodule, and bilobar nodule. Dcaps has two levels: detection of capsular involvement and detection of no capsular involvement.

An exploratory data analysis will be performed to assess each variable individually and the relationship between response and the six potential predictors. A logistic regression model will be built for capsule on six potential predictors on the train data set. The quality and validity of the final module will be evaluated by residual plot and influence plots. The predictive performance will be assessed by ROC curves and corresponding evaluation statistics on the test data set. RStudio Version 3.6.2 will be used for all analysis in this report.

3 Results

3.1 Exploratory Data Analysis

In order to build out a valid logistic regression model, we will start by performing exploratory data analysis. We will check the distribution of the response variable, relationship between response variable and predictors, and the correlation between all variables.

Response variable capsule

Capsule is our binary response variable which denote the status of tumor penetration of prostatic capsule. In our data set, 149 of 377 subjects had a cancer that penetrated the capsule. It

means that 40% of subjects has penetration and 60% has no penetration. We want to build a logistic regression between capsule and six potential predictors which can be used to predict whether a tumor has penetrated the prostatic capsule. If the predicting result is penetration, it means that person has high-risk of prostate cancer. Then the further screening test and treatment plan can be done as early as enough for a patient who may has prostate cancer.

Relationship between response variable and all predictors

Six potential predictors in our dataset which may be useful for predicting the status of penetration. Those six predictors are: age, race, dpros, dcaps, psa, and gleason. We want to know which predictors might be useful for predicting the status of penetration. So, we will take a look of the relationship between response variable and every individual predictor firstly.

Figure 1 boxplots and Table 4 cohen's d in appendix A show us the relationship between response variable and three continuous predictors. It appears that psa and gleason have medium and strong relationship with capsule. Elevated levels of psa and larger values of gleason suggest higher risk of cancer. We also can see there are many outliers of psa which we may need to be attentive. Relationship between capsule and age is weak.

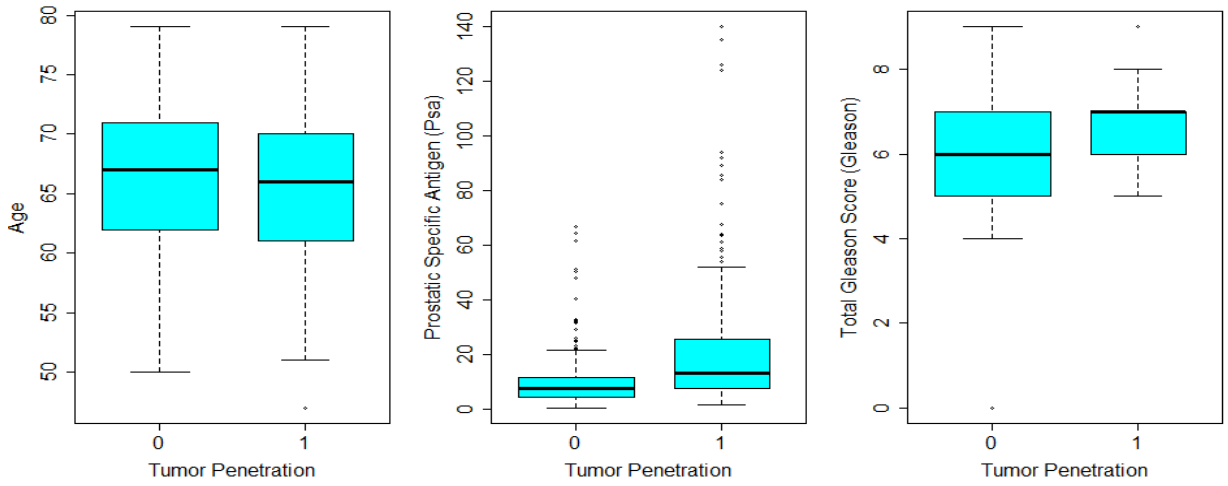

Figure 1: Plot of relationship between capsule and three continuous predictors (Age, Psa, and Gleason) shows there is a strong relationship between capsule and continuous predictors except Age

Table 2 presents us the relationship between response variable and three categorical predictors. We can see that the odds ratio between white and black are almost equal to 1. This means that the association between race and penetration is weak. The odds of penetration are 5 times greater in the detection of capsular involvement (dcaps yes) compared to the detection of no capsular involvement (dcaps no) suggest that the detection of capsular involvement increase the risk of penetration. Chi-square test was used to test the relationship between dpros and penetration (right table of Table 2). Small p-value of chi-square test suggest there is an association between dpros and penetration.

Table 2: Contigency table between capsule and three categorical predictors (Race, Dpros, and Dcaps)

	Penetration		
Race	No	Yes	Total
White	204	137	341
Black	22	14	36
Total	226	151	377

	Penetration		
Dcaps	No	Yes	Total
No	216	121	337
Yes	10	30	40
Total	226	151	377

	Penetration		
Dpros	No	Yes	Total
No nodule	80	19	99
Left unilobar nodule	83	48	131
Right unilobar nodule	45	50	95
Bilobar nodule	18	34	52
Total	226	151	377

Correlation of all variables

To look for potential problems of multicollinearity, a correlation plot was made for all continuous variables. Correlation matrix plot 4 in appendix A shows us all continuous variables are not highly correlated. Relationship between all continuous variables and categorical variables were checked by boxplot. Figure 5 in appendix A show us that dcaps is correlated with psa and gleason. Therefore, dcaps may not be needed if psa and gleason in our final model. VIF will also be utilized to check our final model to avoid multicollinearity issue when performing the model diagnostics.

3.2 Model Selection

At very first, all possible main effects and two-way interactions are taken into consideration. Stepwise model selection using AIC (StepAIC) is used to check what factors might be significant for predicting tumor penetration of prostatic capsule.The model suggested by StepAIC includes five main effects (age, race, dpros, psa, gleason) and four interactions (age:race, age:dpros, race:dpros, race:gleason). However, we found that only psa and gleason are significant. So, we fit the model with psa and gleason and get the model with AIC value 272.25.

Here, as we have a small number of predictors (n=6), we will try select manually the most significant predictors to compare with the result of StepAIC. We fit the model with all main effects and found that three main effects (psa, gleason, dpros) are significant. This matches our previous exploratory data analysis. The AIC value of this model is 266.1. Then, we fit the model with those three main effects and all two-way interactions of those three main effects into the model. However, all interactions are not significant.

Comparing above two models, we choose the model which has smaller AIC value 266.1. Therefore, three predictors psa, dpros, and gleason are kept in our model. The coefficients of all three predictors are positive. This means they are positive associated with penetration, and an increase of any one of them will lead to an increase of probability of penetration. The coefficients and corresponding confidence interval can be seen in 3.5 table 3. In the following section, we will check the quality and validity of this model by performing model diagnostics and assessing predicting performance.

3.3 Model Diagnostics

In this section, analysis of residuals and the identification of influential outliers for the fitted model are performed for model diagnostics. Figure 2 standardized Pearson residuals against model estimated probability pi.hat, the lowess smooth of the plot result approximately in a horizontal line with zero intercept, which suggests the fitted model is an adequate model. Figure 2 cook's distance against leverage show us most of cook's distance and leverage are in the acceptable range except six outliers. Figure 2 delta chi square is almost below 9. This suggest our fitted model are acceptable except several outliers.

When we go back to our original data, those six outliers are: observation 25, 248, 338, 187, 192, and 217. Almost all of them have large psa and have the status of penetration. However, Cook's distance and leverage of those outliers are not extremely large. Therefore, we keep all those observations in our model. We also checked the VIF of our fitted model. There is no multicollinearity issue because all VIF are almost 1.

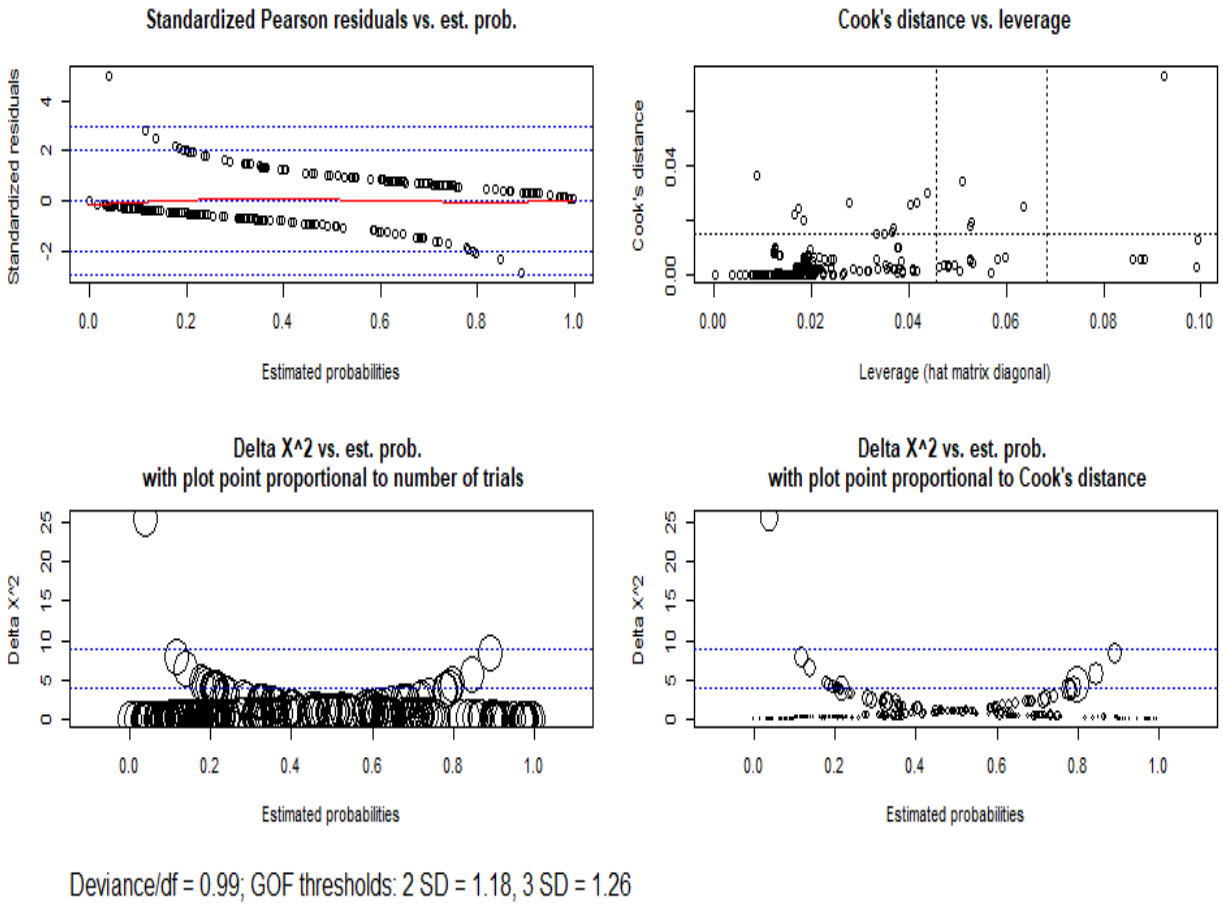

Figure 2: Residual and influence plot of our final logistical regression model

3.4 Model prediction performance

Based on above analysis, our fitted model looks good. Then we will make predictions using the testing data to evaluate predicting performance of our logistic regression. Figure 3 ROC curve located above 45-degree diagonal and AUC value is 0.7914 suggest that our final can be used to predict. The corresponding accuracy, sensitivity, and specificity we got are 0.78, 0.77, and 0.79 at 0.36 cut off. It means that 78% of observations that have been correctly classified. 77% of those are truly penetrated, and 79% of those are not penetrated have been identified. This is good although the predicting performance is not that high.

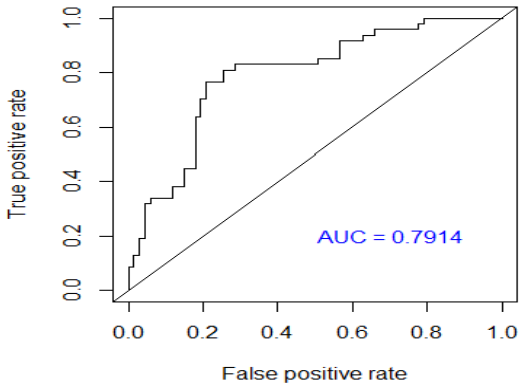

Figure 3: ROC curves and AUC Value of our final model

3.5 Final Model

Above analysis suggest the fitted logistic regression model can be used to predict whether a tumor has penetrated the prostatic capsule, and we will keep that model as our final model in this report. Table 3 presents the inferences of our final model.

Table 3: Inferences of the final logistic regression model of capsule on three predictors

	Coefficient	SE	p-value	95% CI
Intercept	-9.33	1.37	0.0000	(-12.18 , -6.81)
Psa	0.03	0.01	0.0049	(0.01 , 0.06)
Dpros2	0.70	0.45	0.1253	(-0.17 , 1.62)
Dpros3	1.41	0.47	0.0030	(0.51 , 2.38)
Dpros4	1.51	0.60	0.0116	(0.36 , 2.71)
Gleason	1.17	0.21	0.0000	(0.79 , 1.60)

The results show all predictors are positive associated with response variable. We can conclude that 3% more likely to have capsular penetration when Psa increase one unit. We are 95% confident that a unit increase in Psa will lead to a 1% to 6% increase in odds of capsular penetration. 3.22

times more likely to have capsular penetration when gleason increase a unit. We are 95% confident that a unit increase in gleason will lead to a 2.20 to 4.95 times increase in odds of capsular penetration.

Comparing to a patient with no nodule, a patient with bilobar nodule is 4.53 times more likely to have capsular penetration. We are 95% confident that a patient with left unilobar nodule is 1.43 to 15.03 times more likely to have capsular penetration than a patient with no nodule. The confidence interval is rather wide. This makes our result less convincing. A patient with left and right unilobar nodule can be interpretated similarly.

4 Discussion and Conclusions

In this report, logistic regression is applied to estimate the relationship between penetration and six potential predictors (baseline exam measurements). The most significant predictors for penetration in this study are prostatic specific antigen value, results of digital rectal exam, and total gleason score. All predictors are positive associated with the response variable. Model diagnostics is performed using standardized Pearson residuals plot, Cooks' D, Leverage and Pearson Goodness of Fit statistic. Predicting performance is assessed by ROC curves and corresponding evaluation statistics. The predicting accuracy, sensitivity, and specificity of our final model are 0.78, 0.77, and 0.79 at 0.36 cut off. Although the final model can be used to predict the status of penetration, but the predicting performance is not high.

For further study, we may try to increase the predicting performance of our final model. For example, in this study, only the linear terms are discussed. However, quadratic terms, cubic terms can be also considered. Moreover, 263 observations in total for training data set is not enough for getting a good final model. Collecting more observations will help improve the accuracy of fitted logistic regression model. We can also try to fit the model by using all data set and then assess predicting performance by doing k fold cross validation. Lastly, a better logistic regression model could be built by considering additional independent variables (such as genetic, diet factors) that might be associated with prostate cancer.

References

[1] Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin. 2011;61:69–90. [PubMed] [Google Scholar]

[2] Yuri P, Wangge G, Abshari F, Satjakoesoemah AI, Perdana NR, et al. (2015) Indonesian Prostate Cancer Risk Calculator (IPCRC). Acta Med Indones 47: 95-103.

[3] Lee DH, Jung HB, Park JW, Kim KH, Kim J, et al. (2013) Can western based online prostate cancer risk calculators be used to predict prostate cancer after prostate biopsy for the Korean population? Yonsei Med J 54: 665-671.

[4] Yuri P, Hendri AZ (2015) Association between tumor-associated macrophages and microvessel density on prostate cancer progression. Prostate 3: 2-7.

[5] James et al. (2014). An Introducation to Statistical Learning in R. Springer, NY.

[6] Josh Cassidy (August 2013). How to Write a Thesis in LaTeX (Part 3): Figures, Subfigures and Tables

[7] https://www.latex-tutorial.com/tutorials/tables/,LaTeX tables - Tutorial with code examples

Appendix A

Table 4: Cohen's d value measure the strength of the relationship between capsule and three continuous predictors

Predictors	Cohen's d	Strength of relationship
Age	0.08	Negligible
Psa	-0.70	Medium
Gleason	-1.02	Large

(a) (b)

Figure 4: Correlation plot of all continuous variables suggest all continuous variables are not highly correlated

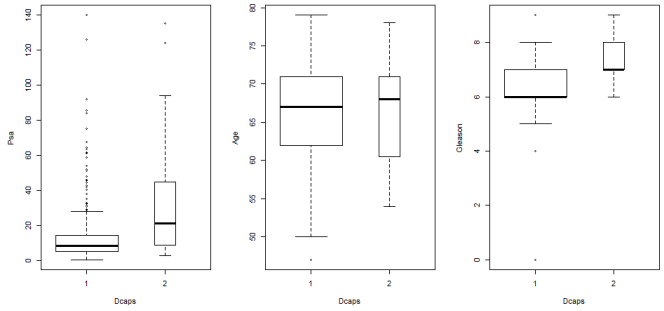

Figure 5: Boxplot between dcaps with all continuous predictors shows us dcaps is correlated with psa and gleason

Appendix B

R Codes:

```
# Load in required R packages
library(corrplot)
library(MASS) #for stepAIC
library(effsize) # for cohen's d
library(ggplot2)
library(ISLR)
library(ROCR)
library(boot) # for cross-validation
library(xtable) # for LaTex

## Load in prostate data set and take initial looks
rm(list=ls(all=TRUE)) # remove all previous objects from memory
options(warn=-1)  # forces R to ignore all warning messages
Prostate <- read.table("C:/D/SDSU/794/DAR2/prostate.txt", header=T)
dim(Prostate) #[1] 380 8
sum(is.na(Prostate)) # missing data: 3 in race
Prostate<-na.omit(Prostate); dim(Prostate) # 377 8

# as.factor tell R this is factor variables
Prostate$dcaps = as.factor(Prostate$dcaps)
Prostate$dpros = as.factor(Prostate$dpros)
Prostate$race = as.factor(Prostate$race)
attach(Prostate)
View(Prostate)

######Exploratory Data Analysis######
# binary response capsule: 1=penetration, 0=no penetration
table(capsule)
signif(prop.table(table(capsule)), digits = 3) #Table of percentages

# relationship between response and predictors###
# continuous preditors(use boxplot): age, psa, gleason: categorical
but treated as continuous
par(mfrow = c(1,3))
agecap <- plot(factor(capsule), age, col="cyan", varwidth=T,
ylab="Age",xlab="Tumor Penetration", cex.lab =1.5, cex.axis =1.5)

psacap <- plot(factor(capsule), psa, col="cyan", varwidth=T,
ylab="Prostatic Specific Antigen (Psa)",xlab="Tumor Penetration",
cex.lab = 1.5, cex.axis =1.5)

glecap <- plot(factor(capsule), gleason, col="cyan", varwidth=T,
```

```
ylab="Total Gleason Score (Gleason)",xlab="Tumor Penetration",
cex.lab = 1.5, cex.axis =1.5 )

#contigency table and chi-square test between categorical predictors and response
table(race,capsule)
chisq.test(x=race,y=capsule) # high p value suggest there is no association
between race and capsule

table(dpros, capsule)
chisq.test(x=dpros,y=capsule)

table(dcaps, capsule)
chisq.test(x=dcaps,y=capsule)


# correlation plot matrix used to check the correlation of continuous variable
par(mfrow = c(2,1))
Procor <- cor(Prostate[,-c(1,2,4,5,6)])
corrplot(Procor)
corrplot(Procor, method = "number")

#relationship between continuous predictors and categorical predictors
par(mfrow=c(1,3))
plot(dpros, psa, varwidth = T, ylab = "Psa", xlab = "Dpros")
plot(dpros, age, varwidth = T, ylab = "Age", xlab = "Dpros")
plot(dpros, gleason, varwidth = T, ylab = "Gleason", xlab = "Dpros")

par(mfrow=c(1,3))
plot(dcaps, psa, varwidth = T, ylab = "Psa", xlab = "Dcaps")
plot(dcaps, age, varwidth = T, ylab = "Age", xlab = "Dcaps")
plot(dcaps, gleason, varwidth = T, ylab = "Gleason", xlab = "Dcaps")


par(mfrow=c(1,3))
plot(race, psa, varwidth = T, ylab = "Psa", xlab = "Race")
plot(race,age, varwidth = T, ylab = "Age", xlab = "Race")
plot(race, gleason, varwidth = T, ylab = "Psa", xlab = "Race")

# how significant or how much impact a given variable has on penetration
cohen.d(age, factor(capsule))   #0.08424036 (negligible)
cohen.d(race, factor(capsule))  #0.0157124 (negligible)
cohen.d(dpros, factor(capsule)) # -0.6868487 (medium)
cohen.d(dcaps, factor(capsule))  #-0.5159278 (medium)
cohen.d(psa, factor(capsule))    #-0.6968868 (medium)
cohen.d(gleason, factor(capsule)) #-1.026021 (large)
```

```
######Model Building######
# set the data to train/test=7/3
n = dim(Prostate)[1]  # sample size
# Split data into training and testing sets
p = 0.7
set.seed(1)  # set the random number generator seed
train = sample(n, p*n)  # random sample percentage out of n; this creates the index list

fitall <- glm(capsule~.*.,family = binomial(link = logit),
data = Prostate[,-1], subset = train)
summary(fitall)
stepAIC(fitall, direction="both" ) # stepAIC suggest the following model


fitall1 <- glm(capsule ~ age + race + dpros + psa + gleason +
age:race + age:dpros + race:dpros + race:gleason,
family = binomial(link = logit), data = Prostate[,-1], subset = train)
summary(fitall1) # only psa and gleason are significant

# AIC 272.25
fitall2 <- glm(capsule ~ psa + gleason, family = binomial(link = logit),
data = Prostate, subset = train)
summary(fitall2)

# Put in all main effects, only,psa, gleason, dpros are significant
# AIC: 266.1
modfit <- glm(capsule~psa + gleason +  dpros
,family = binomial(link = logit), data = Prostate, subset = train)
summary(modfit)

# put in three main effects and all two way interactions,
all interatctions are not significant
modfitin <- glm(capsule~psa + gleason +  dpros + psa:gleason + psa:dpros + gleason:dpros
,family = binomial(link = logit), data = Prostate, subset = train)
summary(modfitin)

######Diagnostics######
# Kutner et al. (2005; Section 14.8)
# Load-in required functions
one.fourth.root=function(x){
x^0.25
}
source("examine.logistic.reg.R")
```

```
#modfit = glm(capsule ~ psa+gleason+dpros, family = binomial, data = Prostate, subset = train)
save2=examine.logistic.reg(modfit, identify.points=T,
scale.n=one.fourth.root, scale.cookd=sqrt)

modfit.diag1=data.frame(Prostate[train,], pi.hat=round(save2$pi.hat, 2),
 std.res=round(save2$stand.resid, 2),
cookd=round(save2$cookd, 2), h=round(save2$h, 2))
p=length(modfit$coef) # number of parameters in model (# coefficients)
ck.out=abs(modfit.diag1$std.res)>3 | modfit.diag1$cookd>4/nrow(Prostate[train,]) |
 modfit.diag1$h > 3*p/nrow(Prostate[train,])
extract.outliers=modfit.diag1[ck.out, ]
extract.outliers

#check VIF
vif(modfit)


######Assess predicting performance######
n = dim(Prostate)[1]  # sample size
# Split data into training and testing sets
p = 0.7
set.seed(1)  # set the random number generator seed
train = sample(n, p*n)  # random sample percentage out of n; this creates the index list
fit_train1 = glm(capsule ~ psa + gleason + dpros, family = binomial(link = logit), data =
 Prostate, subset = train)
test_probs1 = predict.glm(fit_train1, Prostate, type="response")[-train]


# ROC curve
# Plot function of ISLR
rocplot=function(pred, truth, ...){
predob = prediction (pred, truth)
perf = performance (predob , "tpr", "fpr")
plot(perf ,...)}

# Statistics off the ROC
pred1 = prediction(test_probs1, Prostate[-train,"capsule"])
# calculating AUC
auc1 <- performance(pred1,"auc")
# convert S4 class to vector
auc1 <- unlist(slot(auc1, "y.values"))

# Compute optimal cutoff
opt.cut = function(perf, pred){
cut.ind = mapply(FUN=function(x, y, p){
```

```
d = (x - 0)^2 + (y-1)^2
ind = which(d == min(d))
c(sensitivity = y[[ind]], specificity = 1-x[[ind]],
cutoff = p[[ind]])
}, perf@x.values, perf@y.values, pred@cutoffs)
}
# Present sensitivity and specificity for that optimal cutoff
roc.perf1 = performance(pred1, measure="tpr", x.measure="fpr")



# Roc output
# Confusion matrix at 0.5 cutoff
c = 0.36  # the following code suggest the optimal cut off is 0.36
test_class1 = (test_probs1 > c)
table(test_class1, Prostate$capsule[-train], dnn=c("Predicted", "Model 1 Truth"))
# cross-classification accuracy
paste("Accuracy of Model  is", sum(as.numeric(test_class1) ==
 (as.numeric(Prostate$capsule[-train])))/(n-n*p))


# ROC curves

rocplot(test_probs1, Prostate[-train,"capsule"], main="")
abline(a=0, b=1)
text(0.7, 0.2, paste("AUC =", signif(auc1, digits=4)), col="blue")


# Sensitivity and specificity at an optimal cutoff
print("Sensitivities and specificities at the optimal cutoff:")
signif(opt.cut(roc.perf1, pred1), digits=2)

######Final Model######
# Creating inference table of final model
betahat <- formatC(signif(modfit$coeff,digits=6), digits=2, format="f", flag="#")
SE <- formatC(signif(summary(modfit)$coeff[,2],digits=6), digits=2, format="f", flag="#")
cibounds <- formatC(signif(confint(modfit),digits=6), digits=2, format="f", flag="#")
pval <- formatC(signif(summary(modfit)$coeff[,4],digits=4), digits=4, format="f", flag="#")
# Create column names
colnames(x) <- cbind("Coefficient", "SE", "95% CI")
# Use rownames(x) to change variable names
rownames(x) <- cbind("Intercept", "Psa", "Dpros2", "Dpros3", "Dpros4",
"Gleason")
# Create table matrix
x <- cbind(betahat, SE, pval, matrix(paste("(", cibounds[,1], ",", cibounds[,2], ")")))
```

```
colnames(x) <- cbind("Coefficient", "SE", "p-value", "95% CI")
rownames(x) <- cbind("Intercept", "Psa", "Dpros2", "Dpros3", "Dpros4",
"Gleason")
inftable <- xtable(x, digits=2,
caption="Inferences from final model fit using xtable.", label="modelinf")
align(inftable) <- "|l|rrrr|"
print(inftable)
```