Predictions on Whether a Tumor has Penetrated the Prostatic Capsule

Kelso Quan

December 3, 2018

Abstract

Prostate cancer is one of the most common cancers among men. Luckily, prostate cancer is treatable when detected early enough. The Ohio State University Comprehensive Cancer Center wants to know if baseline exam measurements can predict whether a tumor has penetrated the prostatic capsule indicating there is indeed a tumor present. The Prostatic Specific Antigen value (PSA), results of digital exam (dpros), and total Gleason score has the ability of predicting whether a patient has a tumor penetrating the prostatic capsule. A patient that has a right unilobar nodule has up to 4.73 times the odds of having their prostatic capsule penetrated by a tumor compared to patients that has no presence of a nodule. For every increase of total Gleason score, there is an additional increase of 2.71 odds likely of having a tumor puncture the prostatic capsule. With every 1 mg/ml increase of PSA, there will be an increase of 3 % odds in having a tumor penetrate the capsule. For one of the predictions, a patient that has bilobar nodule with PSA of 25 and a total Gleason score of 9 has 2.58 times more likely odds of having a tumor protruding from their prostatic capsule. In the end, this model does not have any interaction terms even though they were considered within the scope of the analysis.

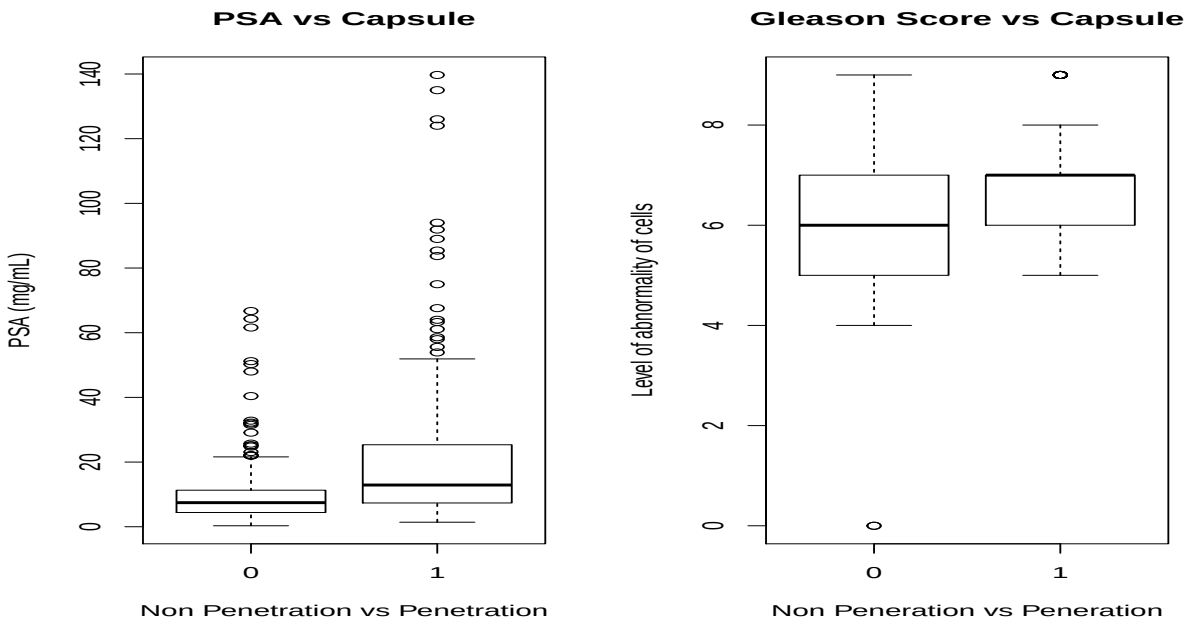
1 Introduction

Since prostate cancer is one of the most common cancer among men, a study was conducted at the Ohio State University Comprehensive Cancer Center tried to determine if a tumor has penetrated the prostatic capsule. This is good news for a majority of men. Prostate cancer which occurs in about 1 out of 7 men, but only 1 in 39 men will die due to this cancer according to webmd. Several factors contribute to cancer tumors penetrating the prostatic capsule which include: age of subject, race, results of digital exam, detection of capsular involvement, Prostatic Specific Antigen value, and total Gleason score. It appears that the Prostatic Specific Antigen value (PSA), results of digital exam, and total Gleason score is able to predict whether a patient has a tumor penetrating the prostatic capsule.

2 Methods

There was a cancer study on 380 male patients of either white or black race. Patients 1162, 1186, 1392 were excluded because those patients had at least one "na" value listed. For those who do not know about prostate cancer, many of the variables will seem unfamiliar. PSA is a measure of protein produced by prostate gland cells and is measured in mg/mL. Elevated levels of PSA may suggest prostate cancer and is used as a screening test. The total Gleason score is a scale from 1 to 10 measuring the abnormality of cells. Larger values of Gleason score suggest a higher risk of cancer. Race has two factors, whether the patient is black or white. The variable capsule indicates whether the tumor penetrated the prostatic capsule. dpros are the results of the digital rectal exam which can have no nodule, unilobar (left and right) nodule, bilobar

Figure 1: Boxplot of PSA/Gleason vs Capsule Penetration



nodule. dcaps is the detection of capsular involvement. There was 153 of the 380 subjects who had a cancer that penetrated the capsule.

This analysis will not have any transformations, but will consider interaction terms. The analysis will be conducted using R/RStudio.

3 Results

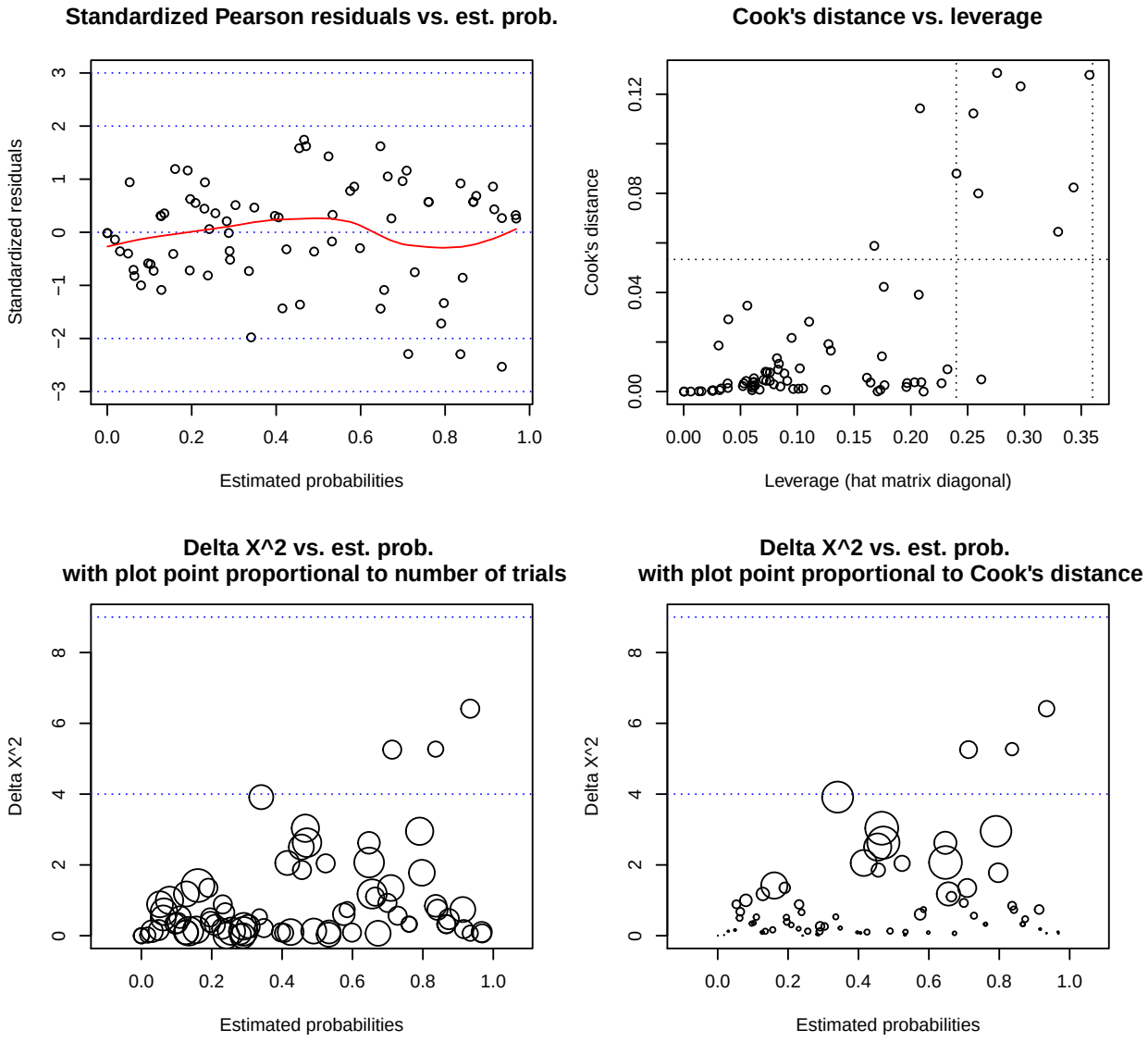
3.1 Exploratory Data Analysis

In figure 1, it shows two things, the amount of PSA when the capsule is penetrated and the distribution of having the capsule penetrated given the level of abnormality in the cells. Both boxplots include patients without a tumor penetrating the capsule. This did show some interesting results. PSA is higher when the capsule is penetrated and the Gleason score is generally higher when there is penetration. These two graphs shows that the doctors and scientists at Ohio State were onto something. In figure 2 shows that there are not any serious violations of homoscedasticity pictured in the Standardized Pearson residual plot. Also, there are very few observations with this model that went beyond three standard deviations. There is a couple of influential points coming from the Cook's distance vs leverage plot that should be discussed. But after examining the possible outliers, these outliers do not significantly impact the model. By applying the HL test, the p-value is 0.246 which means fail to reject the chosen model.

3.2 Model Fitting/Inferences

By using the stepAIC on all possible models containing interaction terms, the model with the least of amount of penalties still had non significant terms. Manually, the terms that were not significant by p-values were dropped one-by-one out of the model. Eventually, the model contained three predictor variables that were not highly correlated with each other shown in figure 3 in the appendix. just at a glance, the correlation

Figure 2: Diagnostic Plots of the model: Capsule = PSA + Gleason + Dpros



Deviance/df = 1.03; GOF thresholds: 2 SD = 1.35, 3 SD = 1.52

plot also shows that the total Gleason score is moderately correlated to a tumor penetration of the prostatic capsule.

Interpreting an odds ratio is is not too difficult. Any odds ratio at 1 does not influence the response. Odds ratios that are between 0 and 1 are (1 - odds ratio) * 100% less odds of having the capsule penetrated by a tumor. But with an odds ratio greater than 1 and less than 2, then the patient has (odds ratio - 1) * 100% more odds of having the capsule penetrated by a tumor. For example, PSA has a coefficient of .03 and has 1.03 odds ratio which means that with 1 mg/mL increase in PSA while all other variables are held constant means a patient has 3% increase in odds of having the capsule penetrated by a tumor for every 1 mg/mL increase in PSA according to table 1. With an odds ratio greater than 2, that odds ratio becomes a factor increase. The odds ratio of total Gleason score is 2.71 which means the patient's odds of having their capsule penetrated by a tumor is increased by a factor 2.71 times for every 1 additional Gleason score added. The rest of the odds ratio can be interpreted in a similar manner, but take caution when interpreting the intercept or any of the dpros predictors. The intercept should be seen as no penetration or dpros1, whereas dpros2 and dpros 3 are directional. In addition, there are confidence intervals associated with their respective odds ratio.

Table 1: Odds Ratio with 95% CI on Odds Ratio

	Coefficient	Odds Ratio	SE	p-value	95% CI on OR
intercept	-8.14	0.00	1.06	0.0000	(0.00 , 0.00)
dpros2	0.77	2.17	0.36	0.0300	(1.09 , 4.43)
dpros3	1.55	4.73	0.37	0.0000	(2.32 , 10.00)
dpros4	1.43	4.18	0.45	0.0015	(1.75 , 10.20)
psa	0.03	1.03	0.01	0.0036	(1.01 , 1.05)
gleason	0.99	2.71	0.16	0.0000	(2.00 , 3.76)

With this model, a patient's odds of a tumor penetrating their prostatic capsule can be determined. Based on table 2, there are several values that were simulated to see a patient's odds of whether a tumor has penetrated their prostatic capsule. For an example, take a patient that has a unilobar nodule on his right side, with PSA of 14.1 and a total Gleason score of 7, then that patient has 98% more likely odds of having a tumor protruding into their capsule.

4 Conclusions

The analysis has shown that total Gleason score, Prostatic Specific Antigen value, and results of a digital rectal exam are the best variables that can predict the probability a man will have a tumor penetrating the prostatic capsule. In the table 1, showed that the results from the digital rectal exam had the most influence on whether a male patient had a tumor protruding through the prostatic capsule. A patient that had a right unilobar nodule had up to 4.73 times the odds of having their capsule penetrated compared to patients that had no presence of a nodule. For every 1 level of total Gleason score increase, there is increase of 2.71 odds likely of having a tumor puncture the prostatic capsule. With every 1 mg/ml increase of PSA, there will be an increase about 3% odds in having a tumor penetrate the capsule. Refer to table 2 to see some

Table 2: Predictions on Simulated values

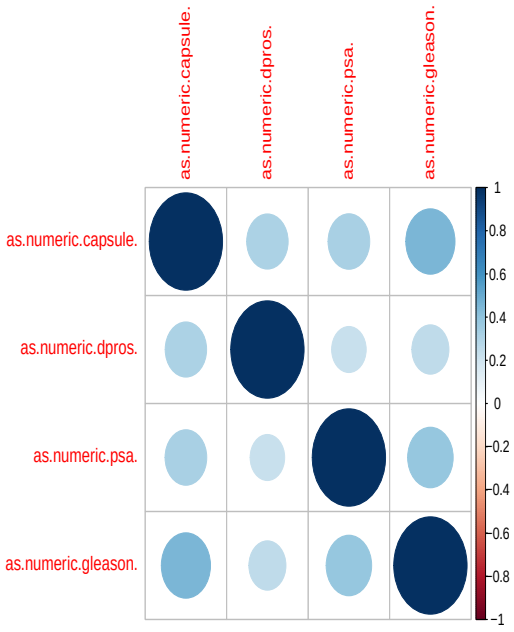
Odds Ratio	dpros	PSA	Gleason
1.98	unilobar nodule (right)	14.1	7
2.23	unilobar nodule (left)	30	8
2.58	bilobar nodule	25	9
1.31	unilobar module (left)	15.25	6

simulated patients and their odds of having a tumor penetrate the prostate capsule. A patient that has a unilobar nodule on his left side, with PSA of 30 and a total Gleason score of 8 means that patient has 2.23 more likely odds of having a tumor protruding into their capsule.

There were several limitations of this study. Age did not play a major factor in the model since the study focused on older men with the youngest man being 47 in the study. If the study included younger men from the ages of 18-35, one can speculate that age will play a huge factor in predicting whether those young men had a tumor penetrate their prostate capsule. In addition, a table showing race would say that only 36 patients were black. That number may or may not be enough to figure out whether this model is works well for black patients, but a simple power calculation would suffice. Furthermore, the results of this study should only apply to those older black or white gentlemen due to the scope of the study. An important question for doctors and scientists is whether prostate cancer can be predicted before a cancerous tumor has penetrated the prostatic capsule. If prostate cancer can be detected even before it has punctured the capsule, then the rate of survival can be raised even higher than the current rate.

Figure 3: Correlation plot among Covariates and Response

```
## corrplot 0.84 loaded
```



Appendix A: Auxiliary Graphics and Tables

Table 3: Race vs Tumor penetration of prostatic capsule

	White	Black
No Penetration	204	22
Penetration	137	14

Table 4: Total Gleason Score vs Tumor penetration of prostatic capsule

	0	4	5	6	7	8	9
No Penetration	2	1	61	100	55	6	1
Penetration	0	0	6	38	72	23	12

Table 5: Digital Rectal Exam Results vs Tumor penetration of prostatic capsule

	No Nodule	Left Nodule	Right Nodule	Bilobar Nodule
No Penetration	80	83	45	18
Penetration	19	48	50	34

Appendix B: R Code

```r
# Read-in data
setwd("~/SchoolWork/Stat696/Prostate")
Prostate = read.table("prostate.txt", header=T)
dim(Prostate); names(Prostate)
#[1] 380 8
sum(is.na(Prostate)) # 3 in race
Prostate <- na.omit(Prostate); dim(Prostate) # 377 8
Prostate$dcaps = as.factor(Prostate$dcaps)
Prostate$dpros = as.factor(Prostate$dpros)
Prostate$race = as.factor(Prostate$race)
Prostate = Prostate[, -1]
attach(Prostate)
head(Prostate)
summary(Prostate)
library(MASS)
library(effsize)
library(VIF)
library(xtable)
library(corrplot)
# A reminder: in this project, we will consider interactions, but not consider
    transformations.
# a) binary response, capsule: consider a simple table of counts
table(capsule)

# b) binary response against categorical covariates: consider tables
#    (can perhaps consider side-by-side bar charts, though I think contingency tables are
    sufficient)
table(capsule, race)
table(capsule, dpros)
table(capsule, age)
# c) binary response against continuous covariates: consider summary tables and box-plots
boxplot(psa ~ capsule, main = "PSA vs Capsule", xlab = "Non Penetration vs Penetration",
    ylab = "PSA (mg/mL)")
boxplot(age ~ capsule, main = "Age vs Capsule", xlab = "Non Peneration vs Peneration", ylab
     = "Age")
boxplot(gleason~capsule, main = "Gleason Score vs Capsule", ylab = "Level of abnormality of
     cells", xlab = "Non Peneration vs Peneration")
```

Table 6: Detection of capsular involvement vs Tumor penetration of prostatic capsule

	No	Yes
No Penetration	216	10
Penetration	121	30

```
33
34 # d) standardized mean difference: evaluation of relationships between covariates and the
       response
35 cohen.d(age, capsule)
36 cohen.d(psa, capsule)
37 cohen.d(gleason, capsule)
38 cohen.d(as.numeric(dpros), capsule)
39
40 ## Task 2: Model building
41 # Logistic regression models are fit using the glm function using the logit link:
42 #   e.g., glm(capsule~psa+gleason, family=binomial(link=logit), data=Prostate)
43
44 fit = glm(capsule~ .*., family = binomial(link=logit), data = Prostate)
45 fit1 = glm(capsule ~  dpros + psa + gleason, family = binomial(link = logit), data =
       Prostate)
46
47 provars = data.frame(as.numeric(capsule), as.numeric(dpros), as.numeric(psa), as.numeric(
       gleason))
48 cpv = cor(provars)
49 corrplot(cpv)
50
51 # a) Stepwise model selection: include interactions, consider stepAIC for first pass
52 stepAIC(fit)
53 stepAIC(fit1)
54
55 # b) Parsimonious model: perform backward selection via p-values,
56 #      identify a simpler model by being strict with interaction terms.
57 #      Is it that much worse than best stepwise model?
58 summary(fit)
59 summary(fit1)
60
61 ## Task 3: Model evaluation
62 # Prostate_diagnostics_ClassVersion.R presents sample code for a given model via
       explanatory variable patterns (EVPs).
63 # The components include:
64 # a) Residual plots: use the examine.logistic.reg function provided
65 # b) Outlier detection: evaluate EVPs for potential outlying data points
66 # c) HL test of overall fit: use the HLtest.R function provided; see PK_diagnostics_
       ClassVerion.R
67
68 # Residual plots
69 # Load-in required functions
70 one.fourth.root=function(x){
71   x^0.25
72 }
73 source("examine.logistic.reg.R")
74
75 # Consider model of PSA, Gleason score, and Detection of capsular involvement
76 dat.glm <- glm(capsule ~ psa+gleason+dcaps, family = binomial, data = Prostate)
77 dat.mf <- model.frame(dat.glm)
78 ## Covariate pattern: too many EVPs!
79 w <- aggregate(formula = capsule ~ psa+gleason+dcaps, data = Prostate, FUN = sum)
80 n <- aggregate(formula = capsule ~ psa+gleason+dcaps, data = Prostate, FUN = length)
81 w.n <- data.frame(w, trials = n$capsule, prop = round(w$capsule/n$capsule,2))
82 dim(w.n)
83 #[1] 301 6
84
85 # Create EVPs by binning continuous covariates
86 g = 5 # number of categories
87 psa_interval = cut(psa, quantile(psa, 0:g/g), include.lowest = TRUE)  # Creates factor with
       levels 1,2,...,g
88 levels(psa_interval)
89
90 # Diagnostic plots
91 w <- aggregate(formula = capsule ~ psa_interval+gleason+dcaps, data = Prostate, FUN = sum)
```

```
n <- aggregate(formula = capsule ~ psa_interval+gleason+dcaps, data = Prostate, FUN =
    length)
w.n <- data.frame(w, trials = n$capsule, prop = round(w$capsule/n$capsule,2))
mod.prelim1 <- glm(formula = capsule/trials ~ psa_interval+gleason+dcaps,
                   family = binomial(link = logit), data = w.n, weights = trials)
save1 = examine.logistic.reg(mod.prelim1, identify.points=T, scale.n=one.fourth.root, scale
    .cookd=sqrt)

# Evaluation of EVPs for potential outlying sets of points
w.n.diag1 = data.frame(w.n, pi.hat=round(save1$pi.hat, 2), std.res=round(save1$stand.resid,
     2),
                       cookd=round(save1$cookd, 2), h=round(save1$h, 2))
p = length(mod.prelim1$coef) # number of parameters in model (# coefficients)
ck.out = abs(w.n.diag1$std.res)>2 | w.n.diag1$cookd>4/nrow(w.n) | w.n.diag1$h > 3*p/nrow(w.
    n)
extract.EVPs = w.n.diag1[ck.out, ]
extract.EVPs

# Note: EVPs purely for diagnostics, akin to histogram binning to assess
# distribution shape.  The analysis does not use this EVP binning.

betahat = formatC(signif(fit1$coeff, digits = 3), digits = 2, format = "f", flag = "#")
OR = formatC(signif(exp(fit1$coeff), digit = 3), digits = 2, format = "f", flag = "#")
SE = formatC(signif(summary(fit1)$coeff[,2], digits = 3), digits = 2, format = "f", flag =
    "#")
cibounds = formatC(signif(exp(confint(fit1)), digits = 3), digits = 2, format = "f", flag =
     "#")
pval = formatC(signif(summary(fit1)$ coeff[,4], digits = 4), format = "f", flag = "#")

x = cbind(betahat, OR, SE, pval, matrix(paste("(", cibounds[,1], "," , cibounds[,2], ")")))
colnames(x) = cbind("Coefficient", "Odds Ratio", "SE", "p-value", "95% CI on OR")
rownames(x) = cbind("intercept", "dpros2", "dpros3", "dpros4", "psa", "gleason")
inftable = xtable(x)
align(inftable) = "|l|ccccc|"
print(inftable)

###EDA###
table(race)
table(capsule)
table(capsule,race)
table(capsule,dcaps)
table(capsule,dpros)

prop.table(table(capsule))
prop.table(table(capsule,race))
prop.table(table(capsule,dcaps))
prop.table(table(capsule,dpros))


boxplot(age~capsule, data = Prostate)
boxplot(psa~capsule, data = Prostate)
boxplot(gleason~capsule, data = Prostate)
plot(dpros, capsule, data = Prostate)

cohen.d(age, capsule)
cohen.d(psa, capsule)
cohen.d(capsule, race)
cohen.d(gleason, capsule)

t.test(age[capsule==0], age[capsule==1])
t.test(psa[capsule==0], psa[capsule==1])
t.test(gleason[capsule==0], gleason[capsule==1])
table(age)
table(race)
## prediction
```

```
new.data = data.frame(dpros = as.factor(3), psa = 14.1, gleason = 7)
new.data2 = data.frame(dpros = as.factor(2), psa = 30, gleason = 8)
new.data3 = data.frame(dpros = as.factor(4), psa = 25, gleason = 9)
new.data4 = data.frame(dpros=as.factor(2), psa=15.25, gleason=6)
predict(fit1, new.data, interval = "prediction", type = "response")
predict(fit1, new.data2, interval = "prediction", type = "response")
predict(fit1, new.data3, interval = "prediction", type = "response")
predict.glm(fit1, new.data4, interval="prediction", type = "response")
chisq.test(x = dpros, y = capsule)

#diagnostics
one.fourth.root=function(x){
  x^0.25
}
source("examine.logistic.reg.R")
# Consider model of PSA, Gleason score, and Results of digital rectal exam
dat.glm <- glm(capsule ~ psa+gleason+dpros, family = binomial, data = Prostate)
dat.mf <- model.frame(dat.glm)
## Covariate pattern: too many EVPs!
w <- aggregate(formula = capsule ~ psa+gleason+dpros, data = Prostate, FUN = sum)
n <- aggregate(formula = capsule ~ psa+gleason+dpros, data = Prostate, FUN = length)
w.n <- data.frame(w, trials = n$capsule, prop = round(w$capsule/n$capsule,2))


# Create EVPs by binning continuous covariates
g = 5 # number of categories
psa_interval = cut(psa, quantile(psa, 0:g/g), include.lowest = TRUE)  # Creates factor with
     levels 1,2,...,g

# Diagnostic plots
v <- aggregate(formula = capsule ~ psa_interval+gleason+dpros, data = Prostate, FUN = sum)
m <- aggregate(formula = capsule ~ psa_interval+gleason+dpros, data = Prostate, FUN =
    length)
v.m <- data.frame(v, trials = m$capsule, prop = round(v$capsule/m$capsule,2))
mod.prelim <- glm(formula = capsule/trials ~ psa_interval+gleason+dpros,
                  family = binomial(link = logit), data = v.m, weights = trials)
save = examine.logistic.reg(mod.prelim, identify.points=T, scale.n=one.fourth.root, scale.
    cookd=sqrt)

v.m.diag = data.frame(v.m, pi.hat=round(save$pi.hat, 2), std.res=round(save$stand.resid, 2)
    ,
                      cookd=round(save$cookd, 2), h=round(save$h, 2))
p = length(mod.prelim$coef) # number of parameters in model (# coefficients)
ck.out = abs(v.m.diag$std.res)>2 | v.m.diag$cookd>4/nrow(v.m) | v.m.diag$h > 3*p/nrow(v.m)
extract.EVP = v.m.diag[ck.out, ]
extract.EVP
source("C:/Users/Kelso Quan/Documents/SchoolWork/Stat696/Prostate/HLTest.R")
HLTest(mod.prelim1, 4)
```

Listing 1: Prostate Cancer