Stat 794, DAR 2

November 12, 2020

Executive Summary

Prostate cancer is among the most common types of cancer among men. The survival rates for prostate cancer vary dramatically depending on the magnitude of spread, with a near 100 percent five-year survival rate in localized cases and a 31 percent five-year survival rate once the cancer has spread beyond the prostate. Given its wide-spread nature and treatability, it is imperative that we have non-invasive, easily-accessible, and accurate tests to screen for the presence of prostate cancer. In order to help medical professionals make better diagnoses and recomendations, a model was created to illustrate the relationship between a set of demographic data and baseline exam measurements and the odds of prostate cancer. This model was then used on a set of data associated with potential subjects in order to demonstrate its predictive capabilites. From this, we also defined the characteristics associated with subjects at low, medium, and high risk of prostate cancer.

Introduction

The goal for this analysis is to create a model that uses demographic data and baseline exam measurements to predict whether a tumor has penetrated the prostatic capsule. The test to determine if a tumor has penetrated the prostatic capsule is a diagnostic test to ascertain the presence or absence of prostate cancer; comparatively, the baseline exam measurements are less invasive screening tests that allow medical professionals to easily identify subjects who may have prostate cancer, and enable them to take more accurate diagnostic tests.

This model will be constructed by utilizing demographic data and baseline exam measurements from 377 men who participated in the Ohio State University Comprehensive Cancer Center study. Logistic regression will be performed in order to create a model that relates predictor variables from the set of potential predictors listed within the Methods section to the response variable, which denotes the presence or absence of capsular penetration.

Methods

The original dataset from which the model is built contains 380 observations for 380 subjects. The ID variable is an identification code that was not used in our analysis. The capsule variable is a binary response variable, with 1 indicating penetration (i.e. the presence of prostate cancer) and 0 indicating the absence of prostate cancer. The remaining variables–age, race, dpros, dcaps, psa, and gleason–are potential predictor variables. Additional information on descriptions and coding can be found in the Appendix.

Three observations were missing values for race; these observations were excluded because they were incomplete. We also chose to exclude observations 282 and 357–both these observations had a Gleason score of 0. For context, the Gleason score range is typically between 2 and 10, so there is a possibility that these values are an overlooked error. Furthermore, the next smallest Gleason value within the dataset is 4, and we felt that by including the two datapoints with 0, we were giving validity to potential interpretations using Gleason scores of 1 through 3; and this would certainly not be the case. By omitting these two observations, we felt that the final model would ultimately better serve healthcare professionals using Gleason scores

between 4 and 9, as intended. Thus, our final dataset contained 375 observations after omitting both the incomplete cases and two cases with a 0 Gleason score.

In order to use the model for the purposes of prediction, we created a number of test subjects; that is, observations where we defined the values for the predictor variables included within the model, to see how the model itself would assess each test subject's risk of prostate cancer. We also defined measures of low, medium, and high risk based on predictor variables included within the model, so that healthcare professionals might best use baseline exam measurements to inform their practices.

Results

Exploratory Data Analysis

Before diving into model creation, we present a general overview of the information within the data set.

We also wanted to investigate the relationship between potential predictors and the response variable. To do so in a visual manner, we relied on conditional desnity plots, which show the relationship between a potential predictor varaible and the response varaible as a proportion of the response at different levels and values of the predictor.

Figure 1 displays two such conditional density plots. On the left, note the lack of variability of the capsule variable with respect to race; that is, regardless of whether subjects were white or black, the proportion of positive responses was about the same. Thus, we do not expect that knowing the race of a subject to provide us with any information about the likelihood of the presence of prostate cancer for that particular subject.

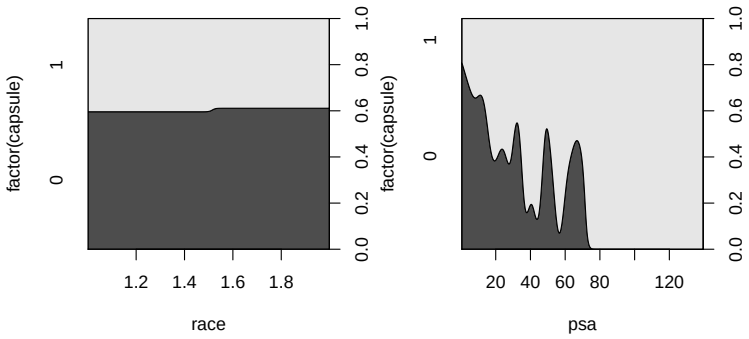

Figure 1: Conditional density plots comparing two potentital predictor variables. We expect that capsule varies dependent on psa, but is invariant with respect to race.

Compare that to the conditional density plot on the right of Figure 1 of, which displays the distribution of capsule based on the potential predictor psa. Here, note the increasely large proprotion of positive responses (i.e. capsule = 1) as psa increases. So in this case, we expect that greater PSA values are, to some degree, indicative of prostate cancer; this squares with our prior knowledge that elevated PSA levels may suggest prostate cancer.

Similar conditional density plots were generated for the remaining potential predictors, and can be found within the Appendix. Age produced a conditional density plot that indicated no particular trend towards the presence or absence of prostate cancer at any particular value. Dpros, dcaps, and gleason all produced conditional density plots that indicated some level of trend within the data, with larger values of all three potential predictors being associated with a higher probability of the presence of prostate cancer.

We also computed the Cohen's d for the relationships between each of the potential predictor variables and the capsule response variable. The values indicated a negligable relationship between age and capsule, and race and capsule, and a moderate relationship between the capsule response variable and each of psa, gleason, dpros, and dcaps.

Given the trends displayed within our conditional density plots and our subsequent analysis of Cohen's d values, it seems that the exam measurements used as screenings tests are far more predictive for determining prostate cancer status than demographic information, like age and race.

Model Fitting and Diagnostics

Our initial goal was to create a logistic regression model using all potential predictor variables, regardless of our intuitions or conclusions about their viability as predictors. Thus, we considered all potential predictor variables included within the study, as well any potential interaction effects. In selecting a final model, we considered AIC, AICc, and BIC, as well as relying on our own judgements and the principle of parsimony. In general, we will be defaulting to using a simpler model over a more complex one when our criteria indicate that they will perform about the same.

A stepwise selection performed on the full model indicated a reduced model that included age race, dpros, psa, gleason, and the interactions bewteen age and dpros, age and gleason, and race and psa. We note here that while we expected race, one of the demographic identifiers in the original dataset, to not play an important predictive, role, it is still included within this reduced model. This is primarly because of the inclusion of the interation effect between race and psa–we cannot remove any variable involved in any of the interactions without first eliminating the interaction itself.

A closer examination of the suggested reduced model indicated a number of variables that could be potentially excluded. We chose to be aggressive in eliminating predictor variables with large associated p-values, particularly with coefficients associated with interaction effects, which are more complex to include and interpret than straightfoward associations. This left us with a potential model that included only psa, gleason, and dpros. We then chose to compare the two potential models using the aforementioned criteria of AIC, AICc, and BIC.

	AIC	AICc	BIC
Reduced Model	390.87	391.88	441.99
Final Model	393.22	393.45	416.81

Table 1: AIC, AICc and BIC associated with both the reduced model chosen by stepwise selection and the final model, which includes only dpros, gleason, and psa.

Table 1 gives an overview of how our two potential models performed under the three criteria. Note that lower values are desirable. The reduced model chosen by the stepwise selection procedure has slightly lower AIC and AICc values compared to our hand-selected model; our hand-selected model had a lower BIC value. Because the performance of our two models was generally similar, we chose to proceed with the simpler model, in order to make the interpretations more straightfoward, the recommendations to any medical professionals more actionable.

As is the case for any logistic model, before we make our inferences and predictions, we want to make sure that our model is a valid one. In order to check the standard assumption of homoskedasticity of the residuals, we relied on using explanatory variable patterns to bin the data. Thus, the plots shown in Figure 2 have 75 data points, to represent the 75 explanatory variable patterns that the data fell into. The first plot in Figure 2 indicates a constant varaince with no identifiable patterns, and standaraized residuals that generally fall within acceptable regions.

The second plot in Figure 2 is a diagnostic plots shwoing the Cook's distance and leverage for the different explanatory variable patterns. A number of data points fall above the cutoffs for Cook's distance and/or leverage; however, because each bins contains a significant number of trials, it would be foolish to remove any one of them and thereby greatly reduce our dataset. For instance, one explanatory variable pattern has a leverage of 0.28, which is large because of the large number of trials–21–within that particular explanatory variable pattern. Thus, we saw no convincing reason to remove any datapoint or set of datapoints, and left all observations within the model.

Running a Hosmer and Lemshow goodness-of-fit test with the default 10 bins resulted in a relatively high p-value of 0.5883, and expected values that generally matched well with the empirical values for prostate cancer in each of the bins. Given the concordence with model assumptions demonstrated in the residual plot and the goodness-of-fit test, we felt confident moving on to interpretation and predictions using the logistic model with psa, gleason and dpros.

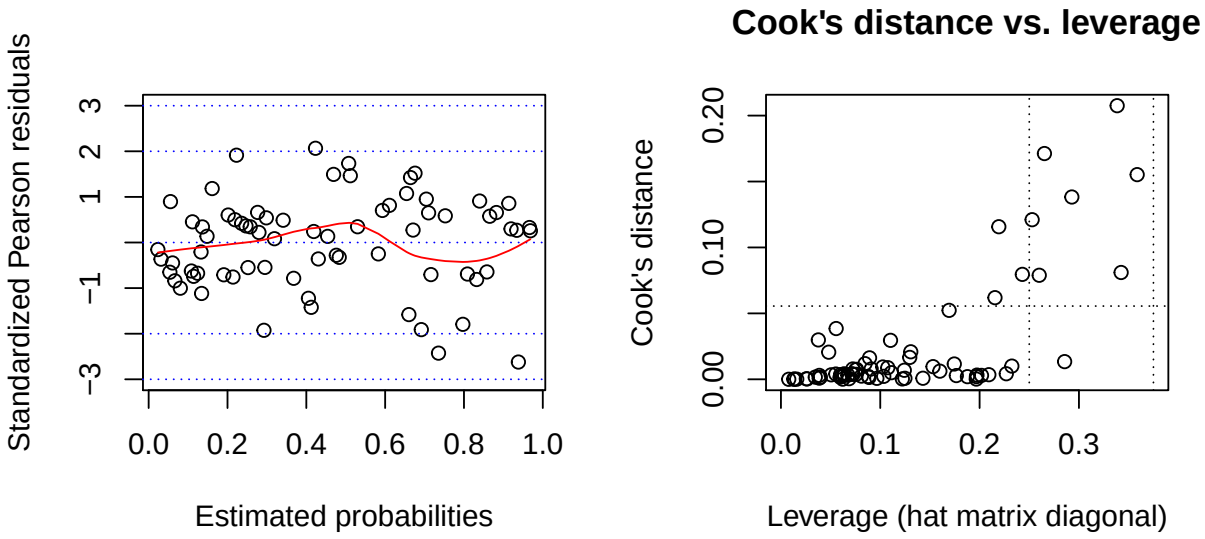

Figure 2: Diagnostic plots for the reduced model. The first plot indicates no violation of homoskedasticity; the second displays leverage and Cook's distance of different explanatory variable patterns.

Inferences

Table 2 lists the point estimates and confidence intervals for all predictor variables included in the final model.

	Coefficient	SE	p-value	95% CI
Intercept	-8.14	1.06	0.0000	(-10.31 , -6.16)
PSA	0.03	0.01	0.0036	(0.01 , 0.05)
Gleason	1.00	0.16	0.0000	(0.69 , 1.32)
DPROS 2	0.77	0.36	0.0300	(0.09 , 1.49)
DPROS 3	1.55	0.37	0.0000	(0.84 , 2.30)
DPROS 4	1.43	0.45	0.0015	(0.56 , 2.33)

Table 2: Regression coefficients and associated standard errors, p-values, and 95 percent confidence intervals for all predictor variables included in the final model.

Here, note that there are regression coefficients associated with the variables "DPROS 2", "DPROS 3", and "DPROS 4". Because the dpros variable was a factor with four levels, each regression coefficients relates a dpros state other than 1 to the state of dpros = 1, i.e. we are comparing each level with a baseline level of 1. Since the other variables are continuous, no levels are created within our model.

For the purposes of illustration, we will use the predictor variable of gleason as an example of interpretation within the model, and will be using the concept of odds to make this interpretation. The odds of success (perhaps a unwise term within the context of this analysis–here, a success indicates the presence of prostate cancer) is given by the following equation:

$$Odds_x = \frac{P(Y = 1|X = x)}{P(Y = 0|X = x)} \tag{1}$$

Thus, the odds of a subject having prostate cancer, given a particular value for x is the ratio of the

probability that the subject has prostate cancer over the probability that the subject does not have prostate cancer.

The odds ratio gives us a measure of how much the odds of a subject having prostate cancer changes when the value of x changes, where the odds ratio (OR) is defined by the following equation:

$$OR = \frac{Odds_{x+c}}{Odds_x} = exp(c\beta_1) \tag{2}$$

Then for our example variable of Gleason, since the point estimate for the coefficient is 1.00, we expect that a $c = 1$, or one-unit, increase in Gleason value is associated with an $e^{1.00} = e$ increase in the odds of a subject having prostate cancer. On a more general scale, a c-unit increase in Gleason value is associated with an e^c increase in the odds of a subject having prostate cancer. The 95 percent confidence interval of $(0.69, 1.32)$ indicates bounds of $e^{0.69} = 2$ and $e^{1.32} = 3.7$ for our estimate of the increase in odds ratio for a 1-unit increase in Gleason score.

Predictions

In order to make predictions with our model, we created data for a set of imaginary patients, and assigned to them Gleason, PSA, and DPROS values that spanned the breadth of the original dataset. No values outside of the ranges within the original dataset were used, as this would consitute an attempt at extrapolation that would likely not be particularly informative or accurate.

To more fully ascertain the risk associate with different scores and levels for the predictor variables, we created a test subject for every combination of Gleason score, DPROS score, and quantile of PSA. We assigned the median value for PSA for each quantile to each test subject. Note here that we elected to exclude a number of combinations in the cases where our original data set had 0 or only 1 case displaying that particular combination, as trying to make claims from such limited data seemed unwise.

The data and predicted odds of prostate cancer for ten of the patients within the set are summarized in Table 3.

Odds of Prostate Cancer	Lower CI	Upper CI	Gleason	DPROS	PSA
0.05	0.02	0.09	5.00	1.00	2.60
0.11	0.06	0.20	5.00	2.00	5.85
0.26	0.17	0.42	6.00	2.00	2.60
0.63	0.40	1.01	6.00	3.00	5.85
0.78	0.49	1.24	7.00	2.00	5.85
2.14	1.32	3.47	7.00	3.00	14.15
3.25	1.57	6.74	7.00	4.00	33.95
5.80	2.99	11.26	8.00	3.00	14.15
8.80	3.89	19.89	8.00	4.00	33.95
23.81	8.77	64.67	9.00	4.00	33.95

Table 3: Predicted odds of prostate cancer, along with 95 percent confidence intervals, for a set of ten potential patients that span the extent of the values of the original data set.

As we expect, as values for Gleason score, DPROS and PSA increase, the odds of prostaet cancer increase as well. For instance, a patient with a Gleason score of 5, a DPROS score of 1, and a PSA level of 2.60 has a predicted odds of prostate cancer of 0.05–that is, this patient is about 20 times more likely to not have prostate cancer than to have it. The associated 95 percent confidence intervals indicate the odds are most likely between 0.02 and 0.09, i.e. the patient is between about 10 and 50 times as likely to not have prostate cancer than to have it.

Comparatively, a patient with a Gleason score of 9, a DPROS score of 4, and a PSA level of 33.95 has a predicted odds of prostate cancer of 23.81, indicating that this patient is 23.81 times as likely to have prostate than to not have it.

For the purposes of this analysis, we defined the cutoff boundries for low-risk to be an odds of prostate cancer between 0 and 0.1, medium-risk to be an odss between 0.1 and 1, and high risk to be an odds of prostate cancer greater than one. These boundries were chosen arbitrarily, and should be assessed by healthcare professionals to fall in line with current medical standards. Table **??** displays the combinations of Gleason score, DPROS score, and PSA level that result in classification as either low-, medium-, or high-risk.

(Note: I'm planning on finding a way to format everything in a table, but still not sure how to go about it. This is a placeholder for now, sorry!)

Conclusion

Based on the constructed model, we have defined the combinations of Gleason score, PSA value, and DPROS score that should result in the classification of subjects as low-, medium-, and high-risk for prostate cancer. To reiterate, the chosen cutoffs for each level of risk are subject to change and should be evaluated by those in the healthcare field. Our model indicates that those subjects with a low Gleason score, PSA value, and DPROS score can generally be categorized as low-risk for prostate cancer, and those with high Gleason score, PSA value, and DPROS score can generally be thought of as being high-risk for prostate cancer. Our model indicates that demographic information like age and race was not informative in assessing risk.

Previous research has indicated that African-American men are more likely to have prostate cancer than men of other races, so we were suprised by both the exploratory data analysis and eventual model, both of which supported the take that race was not an important factor in likelihood of prostate cancer. Of pontential concern is that the sample size contained few black men, and while this may be representative of the demographics of the surrounding area and greater US population, it does introduce potential uncertainty on the basis of a relatively small sample size. Further, as the demographic makeup of the US exapands and changes, it is important that scientists include people from a variety of ethnic backgrounds when conducting their research, as the findings here exclude, for instance, Latino and Asian men.

Appendix

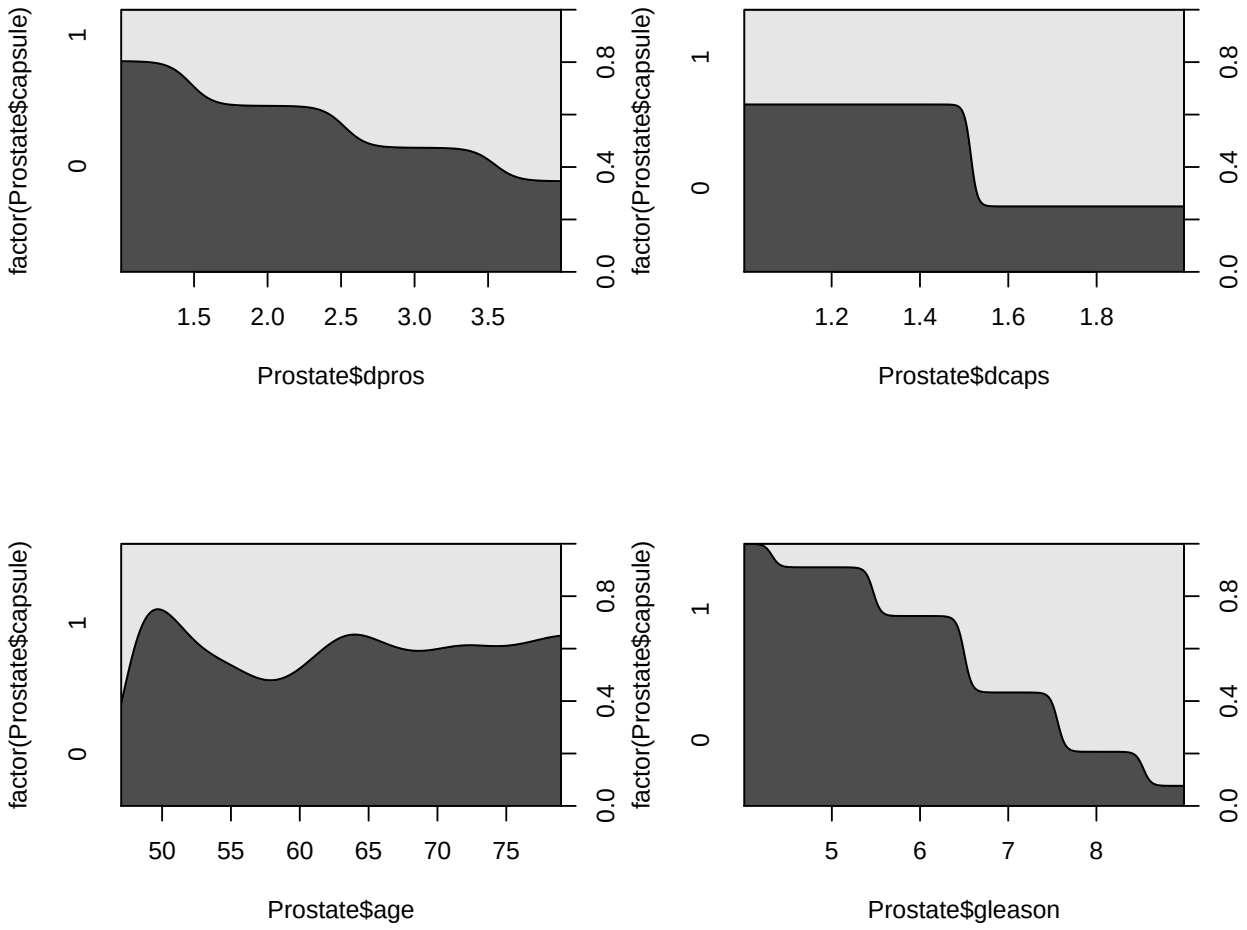

Figure 3: Conditional density plots for all potential predicotr variables.

R Code

```
library(MASS)
library(leaps)
library(car)
library(effsize)
library(corrplot)
library(stargazer)
library(xtable)
#library(rJava); library(glmulti)
library(effsize)
rm(list=ls(all=TRUE)) # remove all previous objects from memory

# Load data
Prostate = read.table("/Desktop/prostate.txt", header=T)
sum(is.na(Prostate)) # missing data: 3 in race

# Remove cases
Prostate <- Prostate[-c(282, 357),]
Prostate<-na.omit(Prostate)

# Remove ID
Prostate <- Prostate[,-c(1)]

Prostate$dcaps = as.factor(Prostate$dcaps)
Prostate$dpros = as.factor(Prostate$dpros)
Prostate$race = as.factor(Prostate$race)
attach(Prostate)

# conditional density plots

par(mfrow = c(1,2))
cdplot(factor(capsule)~race)
cdplot(factor(capsule)~psa)

# intial fit

fit.all <- glm(capsule~.*., family=binomial(link=logit), data=Prostate)

stepAIC(fit.all)

# reduced fit chosen by stepAIC

reduced.fit <- glm(formula = capsule ~ psa + race + gleason + age + dpros +
    psa:race + gleason:age + age:dpros, family = binomial(link = logit),
    data = Prostate)

# final fit chosen by hand

final.fit <- glm(formula = capsule ~ psa + gleason + dpros, family = binomial(link = logit)
    , data = Prostate)

AICc=function(object){
  n=length(object$y)
  r=length(object$coef)
  AICc=AIC(object)+2*r*(r+1)/(n-r-1) # finite sample size corrected AIC
  list(AIC=AIC(object), AICc=AICc, BIC=BIC(object))
}

reduced.fit.crit <- AICc(reduced.fit)
final.fit.crit <- AICc(final.fit)


dat.glm <- glm(capsule ~ psa + gleason + dpros, family=binomial(link=logit), data=Prostate)

```

```r
one.fourth.root=function(x){
  x^0.25
}

g = 5
psa_interval = cut(psa, quantile(psa, 0:g/g), include.lowest = TRUE)

w <- aggregate(formula = capsule ~ psa_interval+gleason+dpros, data = Prostate, FUN = sum)
n <- aggregate(formula = capsule ~ psa_interval+gleason+dpros, 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+dpros,
                   family = binomial(link = logit), data = w.n, weights = trials)

stand.resid <- rstandard(model = mod.prelim1, type = "pearson")
pred <- mod.prelim1$fitted.values
cookd <- cooks.distance(mod.prelim1)
h <- hatvalues(mod.prelim1)
p <- length(mod.prelim1$coefficients)

par(mfrow = c(1,2))

plot(x = pred, y = stand.resid, ylim = c(min(-3, stand.resid), max(3, stand.resid)), ylab =
     "Standardized Pearson residuals", xlab = "Estimated probabilities")
abline(h = c(3,2,0,-2,-3), lty = "dotted", col = "blue")
ord.pred <- order(pred)
smooth.stand <- loess(formula = stand.resid ~ pred, weights = mod.prelim1$prior.weights)
lines(x = pred[ord.pred], y = predict(smooth.stand)[ord.pred], lty = "solid", col = "red")

plot(x = h, y = cookd, ylim = c(0, max(4/length(cookd), cookd)), xlim = c(0, max(3*p/length
    (h), h)),
xlab = "Leverage (hat matrix diagonal)", ylab = "Cook's distance", main = "Cook's distance
    vs. leverage")
abline(h = c(4/length(cookd), 1), lty = "dotted")
abline(v = c(2*p/length(h), 3*p/length(h)), lty = "dotted")


p1 <- data.frame(gleason = 5, dpros = as.factor(1), psa = 0.2)
p2 <- data.frame(gleason = 5, dpros = as.factor(2), psa = 4.9)
p3 <- data.frame(gleason = 6, dpros = factor(2), psa = 5.8)
p4 <- data.frame(gleason = 6, dpros = factor(3), psa = 11.2)
p5 <- data.frame(gleason = 7, dpros = factor(2), psa = 31.5)
p6 <- data.frame(gleason = 7, dpros = factor(3), psa = 20.6)
p7 <- data.frame(gleason = 7, dpros = factor(4), psa = 40.4)


pred_data <- rbind(p1,p2,p3,p4,p5,p6,p7)

gleason <- rbind(5,5,6,6,7,7,7)
dpros <- rbind(1,2,2,3,2,3,4)
psa <- rbind(1.5, 4.9,5.8,11.2, 31.5, 20.6, 40.4)
predAll <- predict(final.fit, pred_data, interval = "prediction", se.fit = TRUE)
preds <- exp(predAll$fit)
upper <- exp(predAll$fit + 1.96 * predAll$se.fit)
lower <- exp(predAll$fit - 1.96 * predAll$se.fit)
```

Listing 1: Associated R Code for EDA