MSDS 6372 Project 3

Logistic Regression of Prostate Capsules

Team members:

Brian Kruse, Christopher Boomhower, Andrew Abbott, Johnny Quick

Date: 12/2/2016

Introduction

Per the Centers for Disease Control and Prevention, prostate cancer is the most common cancer among men in the United Sates with 101.6 per 100,000[i]. There are various tests used to look for warning signs of prostate cancer along with other risk factors such as race, age, and family history. Separately, these indicators are useful but there are questions about the benefits of early screening. There are risks associated with false-positives, false-negatives, and treating cancer that is not dangerous. A more robust model that incorporates the significant available variables could reduce those risks and improve outcomes. The goal of this paper is to use preliminary medical screenings and patient demographic information to better predict prostate cancer.

This paper is organized as follows: the descriptive statistics section introduces the data set. The next section removes missing values, addresses assumptions, conducts variable reduction techniques, and selects the final model. Lastly, the interpretation section contains the statistical and contextual interpretation of the model.

Descriptive Statistics

Hosmer and Lemeshow (2000) Applied Logistic textbooks refer to a Prostate Cancer case study that will be used for this project. Data set details are listed in Table 1 below, which includes 380 observations and nine variables. A Gleason score is given to prostate cancer based upon its microscopic appearance. Cancers with a higher Gleason score are more aggressive and have a worse prognosis. A transrectal ultrasound of the prostate gland is typically used to help diagnose symptoms such as: a nodule felt by a physician during a routine physical or prostate cancer screening exam, an elevated blood test result, or difficulty urinating. Prostate-specific antigen, or PSA, is a protein produced by cells of the prostate gland. The PSA test measures the level of PSA in a man's blood. DCAPS relates to prostate cancer cells having extended into, and possibly through, the prostate capsule or outer lining of the prostate gland. It may be unilateral (one lobe or side) or bilateral (involving both lobes or sides).

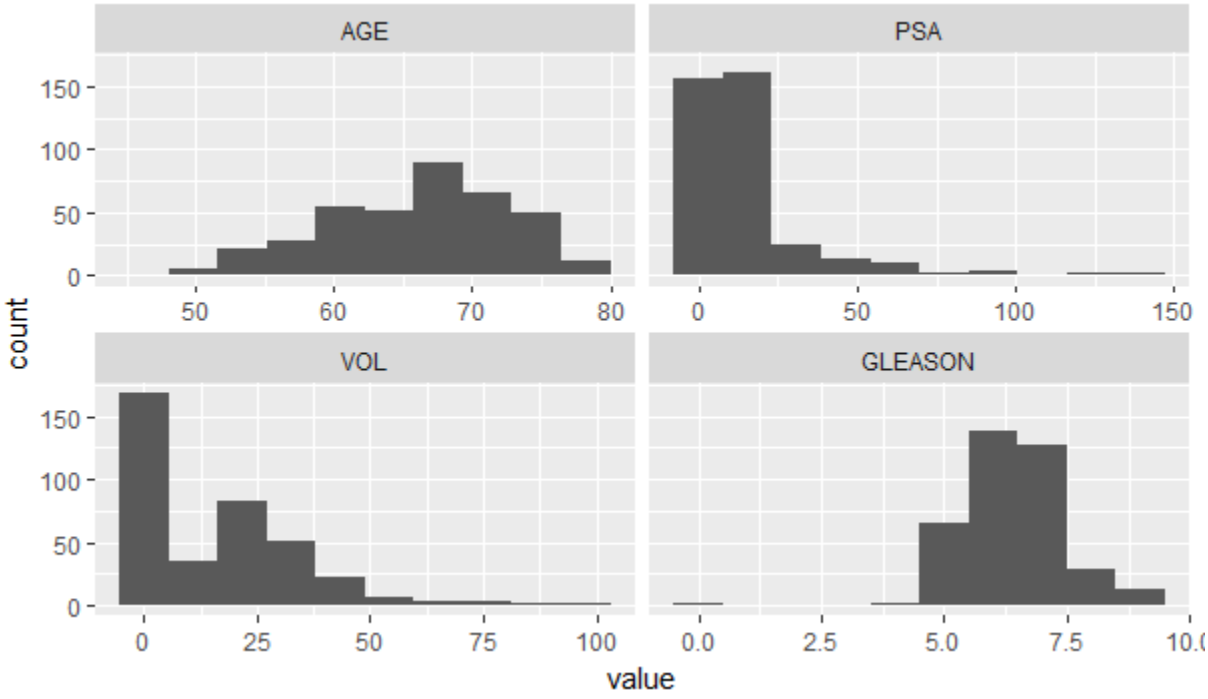
Table 1. Descriptive statistics

Variable	Description	Codes/Values	Name
1	Identification Code	1 - 380	ID
2	Tumor Penetration of Prostatic Capsule	0 = No Penetration,	CAPSULE
		1 = Penetration	
3	Age	Years	AGE
4	Race	1= White, 2 = Black	RACE
5	Results of the Digital Rectal Exam	1 = No Nodule	DPROS
		2 = Unilobar Nodule (Left)	
		3 = Unilobar Nodule (Right)	
		4 = Bilobar Nodule	
6	Detection of Capsular Involvement in Rectal Exam	1 = No, 2 = Yes	DCAPS
7	Prostatic Specific Antigen Value	mg/ml	PSA
8	Tumor Volume Obtained from Ultrasound	cm3	VOL
9	Total Gleason Score	0 - 10	GLEASON

Analysis

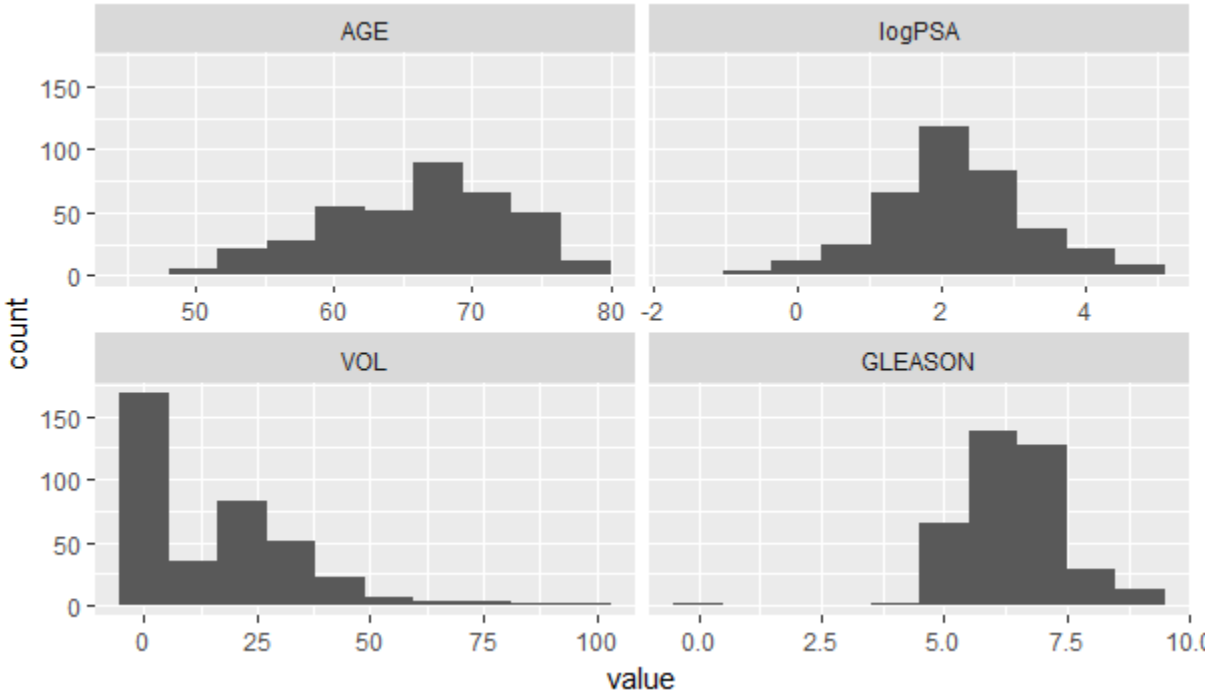
In our review of the data, we found some missing values (1 in the Volume element and 3 in the Race element); we removed these observations. Since one of the assumptions for logistic regression is normality for the continuous variables, we checked them for normality. We observed Volume and PSA were right skewed which can be seen in Figure 1. Most observations for the Volume variable were 0, so transformations on it would be prohibitive.

Figure 1. Initial histograms of continuous variables



We decided to log transform the PSA variable. The histograms after that transformation show the logPSA variable is normally distributed as seen in Figure 2. Histograms of continuous variables after transformationThe Gleason and Age variables are fairly normally distributed.

Figure 2. Histograms of continuous variables after transformation



Another assumption is that the independent variables are not linear combinations of each other. We checked for multicollinearity during our model selection below using the variable inflation factor (VIF) for our independent variables (See Table 8, Table 9, and Table 10 in the Appendix).

We then checked correlations between the variables. The results can be seen in Table 2. We see the Capsule variable is highly correlated with Gleason and logPSA while also slightly correlated with Volume and moderately correlated with DPROS (result of digital rectal exam) and DCAPS (whether capsule was found in digital rectal exam). Age is slightly correlated with Volume. Race is slightly correlated with logPSA. DPROS is moderately correlated with DCAPS and Gleason while also being slightly correlated with logPSA. DCAPS is moderately correlated with Gleason and logPSA while also being slightly correlated with Volume. Finally, Gleason is moderately correlated with logPSA.

Some of these make sense such as Gleason and logPSA because the higher your PSA count, the greater your Gleason score. As stated on the cancer.gov webpage, “The Gleason score indicates how likely it is that a tumor will spread. A low Gleason score means the cancer tissue is similar to normal prostate tissue and the tumor is less likely to spread; a high Gleason score means the cancer tissue is very different from normal and the tumor is more likely to spread.” [ii]

Capsule’s correlation with DPROS, DCAPS, Gleason, and logPSA makes sense because if a capsule is found, then it stands to reason there’s a relation between that and a capsule being found during a digital rectal exam, and also with the Gleason score and logPSA which are indicative of prostate cancer.

DPROS correlation with DCAPS is logical because if a nodule is found in a digital rectal exam, then there’s a greater chance of finding a capsule during that digital rectal exam as well.

Gleason correlation with logPSA is reasonable because, per cancer.gov, "PSA blood levels may be higher than normal in men who have prostate cancer, benign prostatic hyperplasia (BPH), or infection or inflammation of the prostate gland." [iii] Therefore, a higher PSA score is related to the Gleason score which is a score that indicates how likely a tumor will spread as stated previously.

Table 2. Correlation values

	CAPSULE	AGE	RACE	DPROS	DCAPS	VOL	GLEASON	logPSA
CAPSULE	1	-0.040	-0.008	0.318	0.245	-0.115	0.449	0.359
AGE	-0.040	1	-0.045	-0.051	0.008	0.114	0.026	0.008
RACE	-0.008	-0.045	1	0.085	0.064	0.077	0.027	0.162
DPROS	0.318	-0.051	0.085	1	0.244	-0.054	0.253	0.175
DCAPS	0.245	0.008	0.064	0.244	1	-0.109	0.282	0.259
VOL	-0.115	0.114	0.077	-0.054	-0.109	1	-0.059	0.052
GLEASON	0.449	0.026	0.027	0.253	0.282	-0.059	1	0.456
logPSA	0.359	0.008	0.162	0.175	0.259	0.052	0.456	1

Selecting a Model

First, we built a logistic regression model with all variables: Capsule as a response to Age + Race + DPROS + DCAPS + logPSA + Vol + Gleason. Since Race, DPROS, and DCAPS are factors, their coefficients will be viewed based on a comparison against their respective first level (eg. Race = 1, DPROS = 1, and DCAPS = 1). When we checked for large VIF values, no multicollinearity was found (results can be found in Table 8). From Table 3, we can see that the only variables that show statistical significance are DPROS2 (p-value = .036), DPROS3 (p-value <.001), DPROS4 (p-value = .001), logPSA (p-value <.001), and Gleason (p-value <.001). Age had a p-value of 0.54, Race2's p-value was 0.15, DCAPS2's was 0.27, and Volume's was 0.11.

Table 3. Coefficients of initial logistic regression model

Coefficients:					
	Estimate	Std. Error	z value	Pr(>|z|)	Stat sign
(Intercept)	-7.370998	1.625158	-4.536	5.75E-06	***
AGE	-0.012134	0.01994	-0.609	0.54282	
RACE2	-0.657363	0.46223	-1.422	0.154981	
DPROS2	0.751471	0.358281	2.097	3.60E-02	*
DPROS3	1.500152	0.376022	3.99	6.62E-05	***
DPROS4	1.426709	0.458169	3.114	0.001846	**
DCAPS2	0.506565	0.461195	1.098	0.27204	
logPSA	0.540757	0.160353	3.372	7.45E-04	***
VOL	-0.012413	0.007839	-1.583	0.113332	
GLEASON	0.906996	0.168539	5.382	7.39E-08	***

Due to the results above, we checked for interactions between Age and Race, and between DPROS and DCAPS. We did this by building another logistic regression model with all the variables in the first model along with Age*Race and DPROS*DCAPS. When we checked VIF values, there was evidence of multicollinearity found with Race and Age*Race (results can be found in Table 9). In Table 4, we can see that the interactions Age*Race2 (p-value = 0.39), DPROS2*DCAPS2 (p-value = 0.57), DPROS3*DCAPS2 (p-value = 0.48), and DPROS4*DCAPS2 (p-value = 0.92) are not statistically significant.

Table 4. Coefficients for logistic regression model with interactions

Coefficients:					
	Estimate	Std. Error	z value	Pr(>|z|)	Stat sign
(Intercept)	-7.840149	1.690777	-4.637	3.53E-06	***
AGE	-0.004516	0.021129	-0.214	0.830736	
RACE2	3.603743	4.853004	0.743	0.457736	
DPROS2	0.730376	0.372481	1.961	0.049897	*
DPROS3	1.445704	0.388758	3.719	0.0002	***
DPROS4	1.592763	0.492077	3.237	0.001209	**
DCAPS2	-0.078725	1.447713	-0.054	0.956634	
logPSA	0.535918	0.162068	3.307	9.44E-04	***
VOL	-0.013517	0.007909	-1.709	0.087425	.
GLEASON	0.905555	0.170711	5.305	1.13E-07	***
AGE:RACE2	-0.065137	0.075158	-0.867	0.386127	
DPROS2:DCAPS2	0.964108	1.693173	0.569	0.569079	
DPROS3:DCAPS2	1.18702	1.694971	0.7	0.483728	
DPROS4:DCAPS2	-0.160989	1.653421	-0.097	0.922435	

Next, we built our final logistic regression model which excludes Age, Race, DCAPS, and Volume. When we checked VIF values again, no multicollinearity was found (results can be found in Table 10). Table 5 shows us that all variables are statistically significant with all but DPROS2 having a p-value < 0.001 and DPROS2 having p-value = 0.026. The coefficients will be interpreted in the Interpretation section of this paper.

Table 5. Coefficients for final logistic regression model

Coefficients:					
	Estimate	Std. Error	z value	Pr(>|z|)	Stat sign
(Intercept)	-8.5673	1.0428	-8.215	< 2.00E-16	***
DPROS2	0.7896	0.3544	2.228	2.59E-02	*
DPROS3	1.5424	0.3699	4.17	3.04E-05	***
DPROS4	1.4723	0.4433	3.321	8.96E-04	***
logPSA	0.4947	0.1532	3.229	0.001244	**
GLEASON	0.9479	0.1632	5.809	6.29E-09	***

Since our final logistic regression model was generated, we next tested for goodness of fit. In the various models, the null deviance shows how well the response variable is predicted if the model only includes the intercept. The residual difference shows how well the response variable is predicted with the model used. The difference in these two should be smaller to determine which model is best.

For our initial model, the difference in deviance is 131.25 and the difference in degrees of freedom between the null model and model used is 7. The p-value is 3.437318e-25; since this is less than 0.001, this tells us that the model fits significantly better than an empty model. Furthermore, the log likelihood of the model is -187.6665 with 8 degrees of freedom, and AIC is 391.33 and BIC is 422.77.

Next, our model with interactions has a difference in deviance of 132.41 and a difference in degrees of freedom of 9. The p-value is 3.787802e-24; since this is less than 0.001, this model also fits significantly better than an empty model. The log likelihood of this model is -187.0891 with 10 degrees of freedom, and AIC is 394.18 and BIC is 433.47.

Finally, our final model has a difference in deviance of 123.70 and a difference in degrees of freedom of 3. The p-value is 1.233688e-26; since this is less than 0.001, this model also fits significantly better than an empty model. The log likelihood of this model is - 191.4454 with 4 degrees of freedom, and AIC is 390.89 and BIC is 406.61.

With all the factors mentioned above, the final model is the best model because it has the lowest difference in deviance (higher numbers indicate a worse fit), the lowest p-value, the lowest log likelihood value, the lowest AIC, and the lowest BIC. Our final model is also the simplest of the three discussed.

Interpretation

Now that goodness of fit tests indicate our final model is the most accurate for prostate cancer prediction using preliminary medical screenings, it is appropriate to further discuss coefficient interpretation. Each logistic regression coefficient of Table 5 is representative of the change in log odds of the outcome with each one-unit increase for the respective predictor variable. In other words, it is expected that for every one-unit increase in logPSA, the log odds of having prostate cancer increase by 0.4947 when holding all other explanatory variables constant. As a result, exponentiating this value produces the increase of odds with each unit increase in logPSA: $e^{0.4947} = 1.6400$. This means with each unit increase in logPSA while holding all other explanatory variables constant, we can expect a 64% increase in the odds of having prostate cancer. Similarly, the log odds of having prostate cancer increase by 0.9479 and the odds increase by 158% ($e^{0.9479} = 2.5803$) with each unit increase in GLEASON.

Categorical variable DPROS coefficients may be interpreted similarly level-by-level. Again, DPROS1, the discovery of no nodule, was selected as reference. Therefore, DPROS2, DPROS3, and DPROS4 are all with respect to DPROS1. More specifically, the log odds of prostate cancer increases by 0.7896 with the discovery of a left unilobar nodule (DPROS2), versus no nodule, while holding logPSA and GLEASON constant. This equates to an odds increase of $e^{0.7896} = 2.2025$, or 120%. For the discovery of a right unilobar nodule (DPROS3), versus no nodule, log odds increase by 1.5424 and odds increase by $e^{1.5424} = 4.6758$, or 368%. Finally, for the discovery of a bilobar nodule (DPROS4), versus no nodule, log odds increase by 1.4723 and odds increase by $e^{1.4723} = 4.3593$, or 336%. A summary of these log odds

and odds ratios are provided in the Appendix, Table 7. Final Model Log Odds and Odds Ratios with 95% Confidence Intervals

These same coefficients are what comprise the logistic regression function. Below is the model's logit (log odds) formula (1), followed by odds formula (2), and then logistic regression probability formula (3) using these same coefficient values from Table 6:

$$logit(p) = \log\left(\frac{p}{1-p}\right) = -8.5673 + 0.7896(DPROS2) + 1.5424(DPROS3) + 1.4723(DPROS4) + 0.4947(logPSA) + 0.9479(GLEASON) \quad (1)$$

$$\omega = e^{-8.5673+0.7896(DPROS2)+1.5424(DPROS3)+1.4723(DPROS4)+0.4947(logPSA)+0.9479(GLEASON)} \quad (2)$$

$$p = \frac{e^{a+bx}}{1+e^{a+bx}} = \frac{e^{-8.5673+0.7896(DPROS2)+1.5424(DPROS3)+1.4723(DPROS4)+0.4947(logPSA)+0.9479(GLEASON)}}{1+e^{-8.5673+0.7896(DPROS2)+1.5424(DPROS3)+1.4723(DPROS4)+0.4947(logPSA)+0.9479(GLEASON)}} \quad (3)$$

To put our model to work, we may utilize the above three functions to predict prostate cancer response given values for the explanatory variables. For example, assuming logPSA is 2.203212 (mean logPSA value for these data) and GLEASON is 6.382979 (mean GLEASON for these data), we may estimate the probability of having prostate cancer for each DPROS category. Plugging these values into the equations above (once for each level of DPROS), we obtain the probabilities, odds ratios, and log odds depicted in Table 7. In contextual application, the most important parameter we are interested in is the probability of the presence of prostate cancer. Therefore, these predictions may be interpreted such that for DPROS Level 1 in which no nodule was discovered, for example, given the aforementioned logPSA and GLEASON values, the probability of a patient having prostate cancer is 19.4%. On the other hand, a patient with the same logPSA and GLEASON values who is DPROS Level 4, or who had a bilobar nodule discovered, the probability of having prostate cancer is 51.1%.

Table 6. Probability for Prostate Cancer by DPROS Category

DPROS	logPSA	GLEASON	Probability	Odds	Log Odds
1	2.203212	6.382979	0.193551	0.240004	-1.427099
2	2.203212	6.382979	0.345802	0.528590	-0.637542
3	2.203212	6.382979	0.528789	1.122190	0.115282
4	2.203212	6.382979	0.511293	1.046216	0.045180

To further illustrate the manner in which we may predict prostate cancer probability, more explanatory variable values may be passed through the model. Figure 3 displays the predicted probabilities and 95% confidence intervals for each DPROS category while varying logPSA from its minimum to maximum measurement values and holding GLEASON at its average score of 6.382979. Actual cancer outcomes are plotted as well for convenience. This plot clearly shows that for all values of logPSA, discovery of unilobar right nodules and bilobar nodules in patients present the greatest probability for prostate cancer presence whereas the probability for patients who do not have any nodules discovered is

significantly lower, especially as logPSA increases. It is also worth noting that while unilobar right nodule discovery significantly increases the probability of prostate cancer, unilobar left nodule discovery does not portray the same predicted response. This of course would require further investigation beyond the scope of our paper.

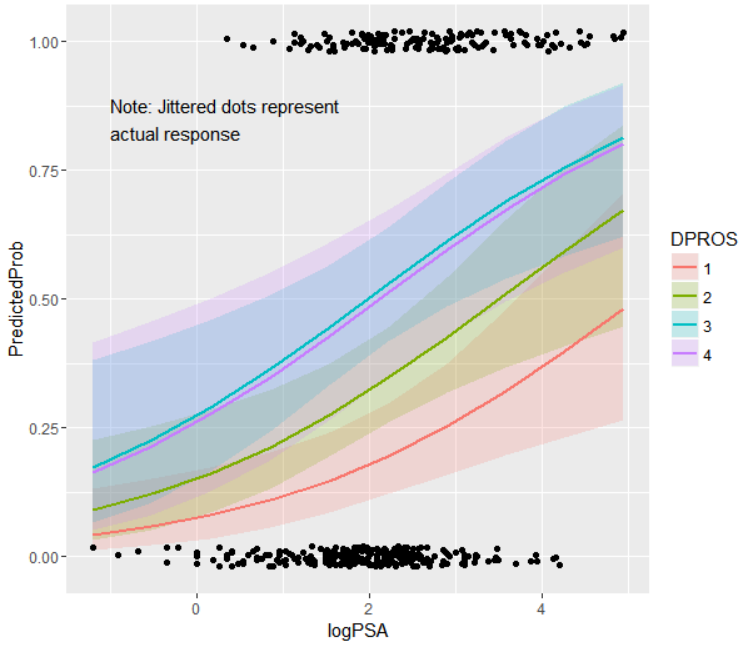


Figure 3. Probabilities and 95% CIs for Prostate Cancer (Holding GLEASON Score at Mean Value)

Conclusion

Logistic regression was performed using the variables from early screening results and patient demographics. Patients with nodules detected show greater odds of developing prostate cancer with those odds increasing as PSA score increases. Age, race, tumor penetration of prostate capsule, and the tumor volume obtained from an ultrasound were not significant. This model is capable of effectively assisting in the correct diagnosis of prostate cancer.

Appendix

Table 7. Final Model Log Odds and Odds Ratios with 95% Confidence Intervals

Parameter	Log Odds	Log Odds Lower CI	Log Odds Upper CI	Odds-Ratio	Odds-Ratio Lower CI	Odds-Ratio Upper CI
(Intercept)	-8.5673	-10.7089	-6.6114	0.0002	2.2345E-05	0.0013
DPROS2	0.7896	0.1067	1.5015	2.2024	1.1126E+00	4.4885
DPROS3	1.5424	0.8328	2.2875	4.6757	2.2998E+00	9.8500
DPROS4	1.4723	0.6152	2.3591	4.3592	1.8500E+00	10.5815
logPSA	0.4947	0.2006	0.8028	1.6399	1.2222E+00	2.2318
GLEASON	0.9479	0.6394	1.2809	2.5802	1.8954E+00	3.5998

Table 8. VIF for initial model

	GVIF	Df	GVIF^(1/(2*Df))
AGE	1.053353	1	1.02633
RACE	1.091412	1	1.044707
DPROS	1.107219	3	1.01712
DCAPS	1.071492	1	1.035129
logPSA	1.228787	1	1.108507
VOL	1.047841	1	1.023641
GLEASON	1.159648	1	1.07687

Table 9. VIF for interaction model

	GVIF	Df	GVIF^(1/(2*Df))
AGE	1.171181	1	1.0 82211
RACE	115.8975	1	10.76557
DPROS	1.567454	3	1.077786
DCAPS	10.26312	1	3.20361
logPSA	1.260966	1	1.122927
VOL	1.061176	1	1.030134
GLEASON	1.192117	1	1.091841
AGE:RACE	115.1483	1	10.73072
DPROS:DCAPS	14.32001	3	1.558322

Table 10. VIF for final model

	GVIF	Df	GVIF^(1/(2*Df))
DPROS	1.045571	3	1.007455
logPSA	1.16208	1	1.077998
GLEASON	1.125724	1	1.061001

R Code

```
# For reference, see
#http://www.ats.ucla.edu/stat/r/dae/logit.htm
#setwd("C:/Prostate_LogisticRegression/Analysis")
setwd("C:/Users/Owner/Documents/GitHub/MSDS_6372/Prostate_LogisticRegression/Analysis")

#install.packages("aod")
library(aod)

require(reshape2)

#install.packages("ggplot2")
library(ggplot2)
library(Rcpp)

#install.packages("car")
library(car)

#prostatedata <- read.csv("C:/Users/Johnny/Documents/6372/Project3/pros.csv")
prostatedata <- read.csv("pros.csv")

## view the first few rows of the data
head(prostatedata)
summary(prostatedata)
nrow(prostatedata) ##number of rows is 380

## NAs are present in VOL (1) and RACE (3), so we need to remove them
prostatedata.cleaned <- na.omit(prostatedata)
summary(prostatedata.cleaned) ## no NAs are left
nrow(prostatedata.cleaned) ##number of rows is 376

## histograms
# Observe attribute distributions
prostate.subset <- subset(prostatedata.cleaned, select = c(AGE, PSA, VOL, GLEASON))
prostate.melt <- melt(prostate.subset[sapply(prostate.subset, is.numeric)])
ggplot(data = prostate.melt, mapping = aes(x = value)) +
  geom_histogram(bins = 10) + facet_wrap(~variable, scales = 'free_x')

## PSA is right skewed, so log transform it (VOL is but contains many 0s)
prostatedata.cleaned$logPSA <- log(prostatedata.cleaned$PSA)

# Observe attribute distributions after transform
prostate.subset <- subset(prostatedata.cleaned, select = c(AGE, logPSA, VOL, GLEASON))
prostate.melt <- melt(prostate.subset[sapply(prostate.subset, is.numeric)])
```

```
ggplot(data = prostate.melt, mapping = aes(x = value)) +
  geom_histogram(bins = 10) + facet_wrap(~variable, scales = 'free_x')

## Mean, Median, and Standard Deviation
sapply(prostatedata.cleaned, mean)
sapply(prostatedata.cleaned, median)
sapply(prostatedata.cleaned, sd)

#correlation
# Observe correlations between variables
write.csv(cor(prostatedata.cleaned[sapply(prostatedata.cleaned, is.numeric)]), file =
"Prostate_Correlations.csv")

# Set RACE, DPROS, and DCAPS to factors
prostatedata.cleaned$RACE <- factor(prostatedata.cleaned$RACE)
prostatedata.cleaned$DPROS <- factor(prostatedata.cleaned$DPROS)
prostatedata.cleaned$DCAPS <- factor(prostatedata.cleaned$DCAPS)

attach(prostatedata.cleaned)

# First Logit with all variables included
prostatelogit <- glm(CAPSULE ~ AGE + RACE + DPROS + DCAPS + logPSA + VOL + GLEASON, data =
prostatedata.cleaned, family = "binomial")

## check VIF
write.csv(vif(prostatelogit), file = "VIF_InitialModel.csv")

summary(prostatelogit)

prostatelogit.interaction <- glm(CAPSULE ~  AGE*RACE + AGE + RACE + DPROS*DCAPS + DPROS + DCAPS
+ logPSA + VOL + GLEASON, data = prostatedata.cleaned, family = "binomial")

## check VIF
write.csv(vif(prostatelogit.interaction), file = "VIF_InteractionModel.csv")

summary(prostatelogit.interaction)

## Remove AGE, RACE, DCAPS, AND VOL as they are not statistically significant
prostatelogit2 <- glm(CAPSULE ~ DPROS + logPSA + GLEASON, data = prostatedata.cleaned, family =
"binomial")

## check VIF
write.csv(vif(prostatelogit2), file = "VIF_FinalModel.csv")
```

```
summary(prostatelogit2)

detach(prostatedata.cleaned)

##goodness of fit
with(prostatelogit, null.deviance - deviance)
with(prostatelogit, df.null - df.residual)
with(prostatelogit, pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = FALSE))
logLik(prostatelogit)
AIC(prostatelogit)
BIC(prostatelogit)

with(prostatelogit.interaction, null.deviance - deviance)
with(prostatelogit.interaction, df.null - df.residual)
with(prostatelogit.interaction, pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = FALSE))
logLik(prostatelogit.interaction)
AIC(prostatelogit.interaction)
BIC(prostatelogit.interaction)

with(prostatelogit2, null.deviance - deviance)
with(prostatelogit2, df.null - df.residual)
with(prostatelogit2, pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = FALSE))
logLik(prostatelogit2)
AIC(prostatelogit2)
BIC(prostatelogit2)

## log odds and odds ratios with 95% CI
logOdds <- cbind(Log.Odds = coef(prostatelogit2), confint(prostatelogit2))
oddsRatio <- exp(cbind(OR = coef(prostatelogit2), confint(prostatelogit2)))
cbind(logOdds, oddsRatio)

## predictions
predprostatedata1 <- with(prostatedata.cleaned,
        data.frame(DPROS = factor(1:4), logPSA = mean(logPSA), GLEASON = mean(GLEASON)))
predprostatedata1

predprostatedata1$DPROSP <- predict(prostatelogit2, newdata = predprostatedata1, type = "response")
predprostatedata1

predprostatedata2 <- with(prostatedata.cleaned,
        data.frame(logPSA = rep(seq(from = min(prostatedata.cleaned$logPSA), to =
max(prostatedata.cleaned$logPSA), length.out = 10), 4),
                GLEASON = mean(GLEASON), DPROS = factor(rep(1:4, each = 10))))
```

```
predprostatedata3 <- cbind(predprostatedata2, predict(prostatelogit2, newdata = predprostatedata2,
type="link", se=TRUE))
predprostatedata3 <- within(predprostatedata3, {
  PredictedProb <- plogis(fit)
  LL <- plogis(fit - (1.96 * se.fit))
  UL <- plogis(fit + (1.96 * se.fit))
})
predprostatedata3

predprostatedata4 <- cbind(predprostatedata1, oddsRatio = -
predprostatedata1$DPROSP/(predprostatedata1$DPROSP-1), logOdds = log(-
predprostatedata1$DPROSP/(predprostatedata1$DPROSP-1)))
predprostatedata4

prostatedata.cleaned$PredictedProb <- prostatedata.cleaned$CAPSULE
ggplot(predprostatedata3, aes(x = logPSA, y = PredictedProb)) +
  geom_ribbon(aes(ymin = LL, ymax = UL, fill = DPROS), alpha = .2) +
  geom_line(aes(colour = DPROS), size=1) +
  geom_jitter(data = prostatedata.cleaned, height = 0.05, width = 0) +
  annotate("text", x = -1, y = 0.85, label = "Note: Jittered dots represent\nactual response", hjust = 0) +
  labs(title = "Predicted Probabilities and 95% CIs for Prostate Cancer")
```

[i] http://www.cdc.gov/cancer/dcpc/data/men.htm
[iii] https://www.cancer.gov/publications/dictionaries/cancer-terms?cdrid=45696
[iii] https://www.cancer.gov/publications/dictionaries/cancer-terms?CdrID=44867