Tumor Penetration of Prostatic Capsule with Logistic Regression

San Diego State University

November 14, 2017

Abstract

Prostate cancer is one type of cancer found in males that can be treated if detected in the early stages of development. In this paper, we obtain prostate cancer screening data from Ohio State University and build a logistic regression model to classify the capsule penetration of a tumor. We begin with an exploratory analysis of the data, then build a model with all covariates and interactions considered. From this model, we utilize backward stepwise model selection with AIC. From this model, we obtain covariates that are correlated, so we build two models and proceed with model selecting individually. We compare the two models, and discuss the issues with both, and selecting our final model based on the lowest misclassification rate. We then discuss the diagnostic plots of the model and interpret the odds ratios. This final model contains the result of a digital rectal exam, with four level of nodules, the volume of a tumor, Gleason Score, and an interaction term of the volume of a tumor and the results of the digital rectal exam. This model performs only slightly better than its competitor in the paper, and we discuss some other models to improve prediction accuracy.

I. Introduction

Prostate cancer is a type of cancer found in males that can be treated and possibly cured if detected in early stages of development. If we are able to detect the early stages of prostate cancer, using only the information obtained from an initial screening, we may be able to reduce the mortality rate of the disease. Therefore, the goal of this paper is to obtain a logistic regression model that can predict if a tumor has penetrated the Prostatic capsule. The data is obtained from Ohio State University Comprehensive Cancer Center and contains 380 observations and 9 attributes. There are 4 observations that contain attributes that are not available. Three of these observations are of the variable race, and one observation over the variable volume. We omit these observations to obtain 374 observations. There is 1 attribute that is an identifier, Id, and we also exclude this from our model to obtain a final dimension of 374 observations by 8 attributes. We consider all covariates and interactions. We also utilize backward step-wise model selection with AIC. We split the model into two, and evaluate both. From these two models, we choose the one that obtains a lower misclassification rate. We obtain confidence intervals for the odds ratios and evaluate this model obtained with several diagnostic plots. We consider one threshold, the one that minimizes squared error predicted to observed. We also give some possible justification for the inclusion of some covariates in the final model and the possible exclusion of others.

II. Methods

The data contains information on whether or not a tumor has penetrated the Prostatic Capsule. In this data set, 151 of 376 tumors penetrated the capsule. The data also contains information on the patient's Age in years and if the patient was a white or black American male. It contains the result of a digital rectal exam, and if the result contained a nodule on left, right, or both sides of the prostate. Moreover, there are covariates that determine if capsular involvement was detected, the volume of the tumor in cm^3, Prostatic Specific Antigen value, and Gleason Score. The Prostatic Specific Antigen value (PSA) is a measure of the protein produced by prostate gland cells, and Gleason Score is a scale from 1-10 that measures the abnormality of cells. Large values in both PSA and Gleason Score indicate a higher risk of prostate cancer.

Table 1 indicates the mean and standard deviations of continuous variables and the proportion of total observations and standard deviations for the categorical variables. The mean Age of the data is 66.06 with a standard deviation of 6.43 as seen in green. This is consistent with the notion that prostate cancer affects mostly older men. In blue, we can see that 90.5% of patients are white males. This could possibly introduce some bias in our final model if the final model determines race is a significant attribute to include. In red, we obtain the sample proportion of capsule penetration, and notice majority of the data contains individuals who were screened and tested negative.

Prostate capsule penetration is predicted using a generalized linear model, logistic regression. The methodology is a backward step-wise model selection using AIC, then further evaluation with a Chi-Squared test is used for nested model selection. All analysis is computed in R, with statistical packages and functions defined in the appendix.

Variable	*Mean/ Proportion of Total*	*Standard Dev.*	*Min*	*Max*
Age	66.06	6.43	47	79
Volume	15.88	18.41	0	97
Prostatic Apecific Antigen	15.28	19.89	0	139
Gleason Score	6.38	1.09	0	9
Race: White	0.905	0.295		
Black	0.096			
Dpro: No Nodule	0.261	0.086		
Unilobar Nodule (left)	0.348			
Unilobar Nodule (Right)	0.253			
Bilobar Nodule	0.138			
Dcap: No Capsular involvement	0.894	0.557		
Capsular involvement	0.106			
Capsule Penetration	0.4015	0.4908		

Table 1: *The table is split between continuous and categorical variables. The continuous variables contain mean and sd. The categorical variables contain sample proportion and sd. The colors are described in the text.*

III. Results

i. Exploratory Data Analysis

The response, capsule penetration, is a binary variable. Approximately 40.15% of these are positive responses. Figure 1 is a boxplot of the continuous covariates Gleason Score, Prostatic Specific Antigen Value, Age of the patient, and volume of the tumor by capsule penetration. The widths of these boxplots are slightly unequal due to the majority of the sample not being a positive response. The first of these boxplots, Gleason Score by capsule penetration, indicates that a higher Gleason Score is associated with having a positive response of capsule penetration. In the second of these boxplots, Prostatic Specific Antigen Value by capsule penetration, the differences are more pronounced than the first. This plot indicates a skew towards higher Prostatic Specific Antigen Value when there is capsule penetration. In order to justify the significance of these differences, we conducted a t-test found in the appendix. The third and fourth boxplot are of Age and volume respectively. The Age boxplot does not visually indicate any differences among the levels of the response. The volume boxplot does not initially indicate any differences, except for the median capsular penetration volume is almost zero, while the no capsular penetration is nearly 20. Therefore, we do not expect Age to yield very important information on predicting the response, but volume may or may not.

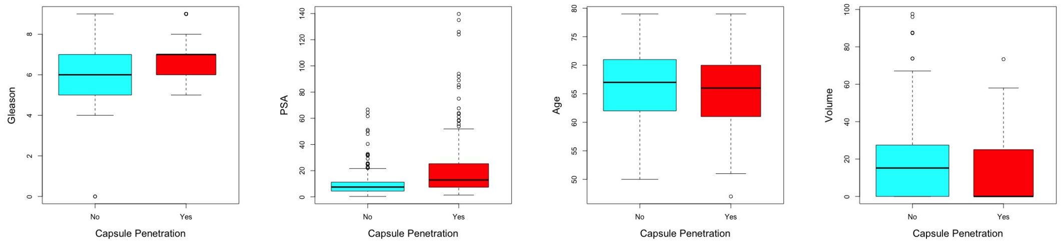

Figure 1: *These boxplot are Gleason Score, Prostatic Specific Antigen Value, patient Age, and volume of a tumor by the response variable capsule penetration. There are differences shown in the plot of Gleason Score and Prostatic Specific Antigen Value; there are no clear differences in Age and volume.*

The categorical variables are seen in Figure 2, plotting the frequency of race, results of the digital rectal exam, and detection of capsular involvement by the response capsular penetration. Although there are more white individuals in the data set, there does not appear to be a proportional difference in capsular involvement among races. Moreover, based on the plot of the result of the digital rectal exam, we observe that a *bilobar nodule* level has the highest capsule penetration rate, while the *no nodule* level has the lowest capsule penetration rate. Finally, the detection of capsular involvement seems to indicate a proportional difference between capsular penetration and the levels *yes or no*. Therefore, before building the logistic regression model, we anticipate that race will not provide as much information as the two other categorical covariates.

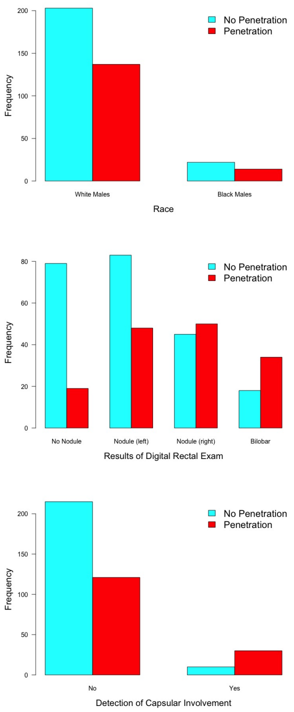

Figure 2: *These barplot displays the categorical variables: race, result of digital rectal exam containing nodules, and if capsular involvement was detected. There are possible proportional differences in the detection of capsular involvement and results of the digital rectal exam. No clear proportional differences among race and capsular involvement.*

The variance-covariance plot found in the appendix indicates little to no correlation among all variables except for Prostatic Specific Antigen Value and Gleason Score. Due to a correlation statistic of 0.38, we decide to remove one of the two variables before building the logistic regression model. This is because a final model selection with both Prostatic Specific Antigen Value and Gleason Score may yield coefficient estimates that are unreliable to interpret within a logistic regression framework. Although this correlation is negligible, we air on the safe side and remove one anyway. One thing to notice is the correlation of volume to capsular penetration is slightly negative. This may be due to individuals who go get a screening for prostate cancer and test negative. That is, most individuals who find a lump and get tested do not actually have capsular penetration, leading to a negative correlation coefficient.

ii. Model Selection

We begin by building a logistic regression model with all covariates and all interactions. We utilize stepwise backward selection AIC to obtain an initial model to tune. Finally, we trim this AIC model selection two times, one without Prostatic Specific Antigen Value and its interactions, and the other without Gleason Score and its interaction. This is because due to a moderate correlation coefficient of as discussed before. From these final models, we determine which has the lowest misclassification rate and proceed with model diagnostics.

The first final model is the one that contains Prostatic Specific Antigen Value where we excluded Gleason Score and its interactions from the output of the AIC model. We then trim the models iteratively, starting with the interaction terms, until a model with significant regression coefficients is obtained. If a main effect is not significant, and there are no higher order terms of the main effect, we exclude this term from our model to test. We test the significance of the model based on a Chi-Squared test for model differences. If we find the nested models are not statistically different from one another, we choose the simplest model. If we find that the models are statistically different, we continue model selection with the higher model. The same model selection technique is applied to the AIC model that contains Gleason Score and not Prostatic Specific Antigen Value.

	Covariates Included	AIC	Cutoff	AUC	Contingency		
Model 1	Prostatic Specific Antigen	435	0.387	0.77		True	False
	Digital Rectal Exam				Pred. True	0.292	0.181
	Tumor Volume				Pred. False	0.109	0.417
Model 2	Gleason Score	396	0.429	0.82		True	False
	Digital Rectal Exam				Pred. True	0.287	0.122
	Tumor Volume				Pred. False	0.114	0.476
	Digital Rectal Exam *Volume						

Table 2: *This table indicates the two models build from AIC and after removing covariates that are not significant base on the Chi-Squared difference of models test. It contains the AIC value, the area under the precision-recall ROC curve, the optimal threshold value based on the minimal sum of squares, and a contingency table based on the threshold as a proportion of total classified. The blue indicates the correctly classified observations of model 1, while the red indicates the correctly classified observations of model 2.*

Table 2 indicates two models. Model 1 contains the covariates Prostatic Specific Antigen Value, results of the digital rectal exam, and the tumor volume. This model yields an AIC value of 435, and optimal probability threshold value of 0.387. The false positive rate for this model is 0.30. The false negative rate for this model is 0.27. The area under the precision-recall ROC curve is 0.77. The blue in the table indicates the correctly classified observation, while those that are not highlighted indicates the misclassified observations for model 1.

In Table 2, model 2 contains covariates Gleason Score, results of the digital rectal exam, tumor

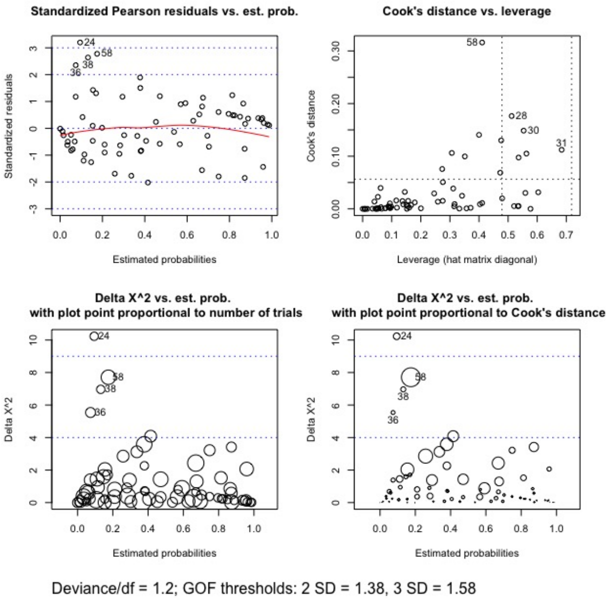

Figure 3: *This figure indicates four diagnostics plots. The top left indicates a plot of Standardized residuals vs. Estimated probabilities. The red line is a fitted "lowess" plot centered around zero. The points that obtain standardized residuals greater than two are numbered. The top right plot indicates the Cook's Distance vs leverage. There is one point 58 that seems to have the most influence. The bottom two plots are Chi-Squared vs. estimated probabilities, the left proportional to the number of trials, the right proportional to Cook's Distance. The size of the points on these plots indicates their influence on the model.*

volume, and an interaction term that includes the results of the digital rectal exam and tumor volume. This model has an AIC value of 396 and is slightly lower than model 1. The optimal probability threshold value is 0.43 and the false positive rate is 0.20. The false negative rate is 0.28. Moreover, the area under the precision-recall ROC curve is 0.82. The red in the table indicates the correctly classified observations, while those that are not highlighted indicates the misclassified observations for model 2. This model has a lower overall misclassification rate than model 1. Based on these results, we choose model 2 as our final model. To see predicted probability density distributions, and ROC curves, please see appendix.

iii. Model Inferences

Figure 3 shows four diagnostics plots of the model chosen. To obtain these plots we first aggregate the data based on volume intervals $(-1, 0], (0, 20], (20, 30],\ (30, 100]$. The top left indicates a plot of Standardized residuals vs. Estimated probabilities. There are four aggregated data points that lie outside of two standard deviations of residuals. These are the points $24, 36, 38$ and 58. The top right plot indicates the Cook's Distance vs leverage of the aggregated data. This plot indicates that point 58 has the highest Cook's distance as a function of leverage, while points $28, 30, 31$ obtain a Cook's distance that indicates moderate influence overall. The bottom two plots are aggregated data points plotted by the change in their Chi-Squared vs. estimated probabilities. The plot on the left bottom is proportional to the number of trials, the plot on the right is proportional to Cook's distance. The size of the points indicates that points $24, 36, 38$ and 58 are the most influential. As seen in the table the influence table in the appendix, these points all have a relatively low number of trials. These 13 points will be kept in the model to maintain consistency.

	Coefficient	OR	SE(OR)	P-value	95% C.I. OR
(Intercept)	-9.23	0.00	0.00	< 0.0001	$(0.0000, 0.0008)$
Nodule (left)	1.28	3.61	1.69	0.0061	$(1.47, 9.34)$
Nodule (right)	1.41	4.10	2.05	0.0048	$(1.57, 11.27)$
Nodule (both)	2.49	12.15	7.86	0.0001	$(3.60, 46.17)$
Gleason Score	1.21	3.35	0.54	< 0.0001	$(2.47, 4.67)$
Volume	0.01	1.01	0.01	0.5554	$(0.98, 1.04)$
Nodule (left)*Volume	-0.04	0.96	0.02	0.0331	$(0.92, 0.99)$
Nodule (right)*Volume	0.00	1.00	0.02	0.9286	$(0.96, 1.04)$
Nodule (both)*Volume	-0.06	0.94	0.03	0.0303	$(0.88, 0.99)$

Table 3: *This table indicates the model coefficients, the odds ratios, the standard error of the odds ratios, the p-value of the test performed that $H_0 : OR = 1$ vs $H_0 : OR \neq 1$, and a confidence interval for the odds ratio.*

Based on the final model, we obtain Table 3, we notice immediately that volume is not significant. However, since we decide to keep the higher order terms involving volume, we keep this main effect. The higher order terms, volume to result of the digital rectal exam, is statistically different than the model without this interaction. This may be due to a nodule found on the digital rectal exam implies a positive tumor volume. These terms may not necessarily give very much information and lowers the AIC value by 4 points over the model without the interaction term.

In table 3, we can interpret the odds ratio of the main effect coefficients. The odds of capsule penetration of unilobar nodule on left increases by 3.61× compared to those without a nodule (baseline), assuming all other variables are fixed. The odds of capsule penetration of unilobar nodule on right increases by 4.1× compared to those without a nodule, assuming all other variables are fixed. Moreover, the odds of capsule penetration of bilobar nodule increases by 12.15× compared to those without a nodule (baseline), assuming all other variables are fixed. For every 1 unit increase in Gleason Score (1-10), the odds of capsule penetration increases by 3.34×, assuming all other variables are fixed. Finally, for every 1 unit increase in volume (cm^3), the odds of capsule penetration increases by 0.85% Assuming all other variables are fixed.

In table 3, we can interpret the odds ratio of the interaction effect coefficients. The first interaction is the difference between the odds ratios corresponding to a change in volume by 1 cm^3 amongst the baseline and the odds ratio corresponding to an increase in volume by 1 cm^3 amongst those with a left nodule. The second interaction is the difference between the odds ratios corresponding to a change in volume by 1 cm^3 amongst the baseline and the odds ratio corresponding to an increase in volume by 1 cm^3 amongst those with a right nodule. The last interaction is the difference between the odds ratios corresponding to a change in volume by 1 cm^3 amongst the baseline and the odds ratio corresponding to an increase in volume by 1 cm^3 amongst those with the bilobar nodule.

For example, assume an individual has a volume of 20 cm^3 and a Gleason score of 6. If the result of the digital rectal exam was no nodule, then the probability of capsule penetration is estimated at 0.1382. For an individual who has the same volume and Gleason score, but instead has a digital rectal exam that yields a nodule on the left, the estimated probability of capsule penetration is 0.205. Then there is an estimated probability increase in probability from the baseline group to the first level of 0.0668. From the baseline no nodule to having a nodule on the right, this difference is 0.258. And finally, from the baseline to the third level, nodules on both sides, this difference is 0.507.

IV. Discussion

If we ignore the issue with interpretation and consider a model that includes both Prostatic Specific Antigen value and Gleason Score, we would have concluded with a model without Volume nor its interactions, and in its place would be Prostatic Specific Antigen value. The issue with the model we obtained currently is that it does not obtain the lowest misclassification rate among all models, and the interaction term is almost negligibly significant. The model described above with Prostatic Specific Antigen value, Gleason Score, and results of digital rectal exam obtain a lower misclassification rate. However, we proceeded with this model to sustain the interpretability of the coefficients, even though the interaction term's interpretation was a bit convoluted.

If a patient has undergone a screening, the result of the digital rectal exam and the Gleason Score seem to be the deciding variables. If the result of the digital rectal exam yields a bilobar nodule on the prostate, that individual is 12 times more likely to have capsule penetration of the prostate over those who have no nodule, assuming all other variables are fixed. Moreover, the interaction between volume and the result of the digital rectal exam are only slightly significant, and all of the levels of interaction obtain odds rations nearly one, an indication of only a slight association between the volume of a tumor and the levels of the digital rectal exam.

Further steps may include rebuilding a model above to include the quantitative Prostatic Specific Antigen value. We should do this if the goal is to only increase prediction accuracy and we are not concerned with the interpretability of the model coefficients.