Something About Prostate Cancer

Andrew Bates

November 15, 2018

Executive Summary

This is the executive summary. Unfortunately, it has to be typed in the .yaml header. The only other thing i can think of is to have a separate file for the abstract. I'm not sure I want to do this. But actually, this may not be so bad. Each section could have a different file. this might make things a bit easier to edit. because then i would only have to look at say, the conclusion file instead of having to scroll all the way down, passing it, srolling up, passing it, etc.

1 Introduction

In this paper, we investigate the relationship between the results of various prostae related exams and whether tumor penetration of the prostatic capsule has occured. The objective of this analysis is to determine if there are any factors that have a particularly influential relationship with prostatic capsule penetration. Additionally, we wish to develop a model to predict capsule penetration so it can be used as a diagnostic tool for future patients.

2 Methods

In this analysis we examine a subset of data collected by the Ohio State University Comprehensive Cancer Center as part of a study to determine the potential of standard exam results to predict whether a tumor will penetrate the prostatic capsule. Out of 380 patients, 153 have experienced capsule penetration and 227 have not. The data set contains six explanatory variables: the patients age and race, results of a digital rectal exam (DRE), whether capsular involvement was detected, Prostatic Specific Antigen (PSA) value, and total Gleason score. For a detailed description of each variable see Table 1. Observe that race is recorded as only Black or White. This may be the reason why three observations contain missing values for race. Perhaps three of the patients were neither Black nor White. However, without access to details of the study and the population considered, we can only speculate. Furthermore, the other variables in these observations do not point to any particular reason as to why race is not recorded[1]. Because of this, we decided to omit these three observations. The data set used for analysis then, consists of 377 observations. Of these, 151 patients have experienced tumor penetration of the prostatic capsule, and 226 have not.

To investigate the relationship between the covariates and tumor penetration status, we use a logistic regression model. The response is especially well balanced with 40% of the observations having capsular penetration and 60% not having penetration. As such, procedures designed to assist with class imbalances, such as downsampling, are not considered in this analysis. In addition to the explanatory variables included in the data set, all two-way interaction terms are examined. Using this as a starting point, the model is chosen via a backwards elimination procedure using Akaike

[1] As in, some have experienced capusle penetration, the ages and PSA scores vary significantly across the observations, etc.

Table 1: Description of variables in the data set.

Name	Description	Details
penetrate	Tumor penetration of prostatic capsule?	Yes, no
age	Patient age	Years
race	Patient race	Black, White
dre	Results of digital rectal exam	No nodule, unilobar left, unilobar right, bilobar
caps	Detection of capsular involvement?	Yes, no
psa	Prostatic Specific Antigen value	mg/ml
gleason	Total Gleason score	0-10

Table 2: Results of digital rectal exam vs. whether tumor pentration of prostatic capsule occured, marginalized by the digital rectal exam results.

	DRE Result			
	no nodule	unilobar left	unilobar right	bilobar
No Penetration	80.8	63.4	47.4	34.6
Penetration	19.2	36.6	52.6	65.4

Information Criteria (AIC) as the model selection criterion. (variable name) is not significant at the 0.05 level so it is subsequently removed, along with its corresponding interaction terms.

Diagnostics take the form of examining explanatory variable patterns. also do blah test. fit same logistic regression model on EVPs weighted by number of original data points in each EVP. Hosmer-Lemeshow test - basically a chi squared goodness of fit test. Note: we do NOT want to reject the null here. the null is that the model fits the data well

log odds ratios were calculated, and predictions were made. predictive power was examined through a confusion matrix. What's more important here, false negatives or false positives?

3 Analysis

3.1 Exploratory Analysis

We begin by inspecting tables of the categorical variables against whether tumor penetration has occured. Rather than examing contingency tables of counts, we find it more informative to view tables marginalized by the covariates. This provides more context by allowing us to see, for example, the proportion of Black patients who had tumor penetration and the proportion who did not. Table 2 displays the proportion of patients who did and did not experience tumor penetration of the prostatic capsule grouped by the results of a digital rectal exam. It shows that the majority of patients who do not have a nodule have not experience tumor penetration. For those patients who did have a nodule, the rate of tumor penetration is reversed. The most pronounced difference is for patients who had a bilobar nodule with 65% experiencing tumor penetration. The relationship we see here tells us that DRE results will likely be a valuable predictor.

The relationship for the remaining categorical variables can be found in Tables 3 to 5 in the appendix.

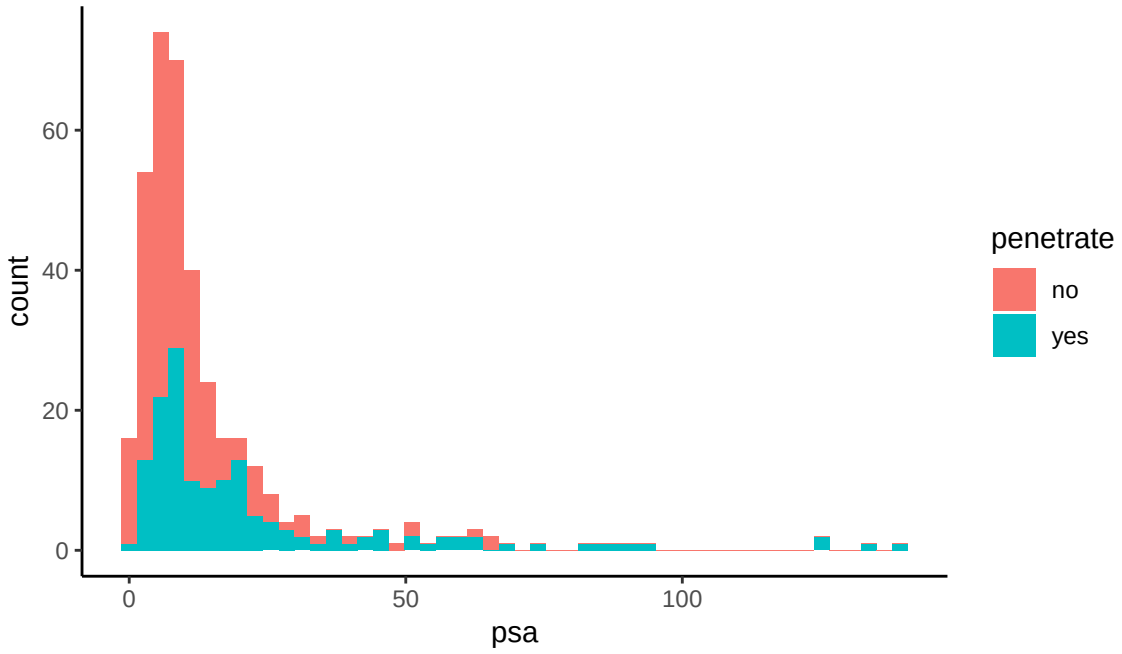

Figure 1: Histogram of prostatic specific antigen value colored by whether tumor penetration of prostatic capsule occured.

For detection of capsular involvment and Gleason score, we find a similar story as DRE results. For patients where capsular involvement was detected, tumor penetration occured at more than twice the rate than patients where involvement was not detected (35% vs. 75%). As Gleason score increases from zero to nine, the tumor penetration rate increases from zero to 92%. Where we do not see a significant change in the rate of tumor penetration is with race. Both Black and White patients have a tumor penetration rate of approximately 40%.

For the numeric variables, we examine summary statistics and plots of age and PSA grouped by whether prostatic capusle penetration has occured. Summary statistics and a box plot for age can be found in Table 6 and Figure 3. There is virtually no difference in the age distribution of patients who have experience tumor penetration and those who have not. As such, we do not expect it to make a significant contribution to our model. On the other hand, PSA value seems like it may be an important predictor. Figure 1 is a histogram of PSA value where color indicates tumor penetration of the prostatic capsule. The distribution has a similar shape for low PSA values however, for PSA values above 67 mg/ml there are no cases where tumor penetration has occured.

Before we move on to the modeling phase of our analysis, it might be beneficial to summarize what we have found so far.

3.2 Modeling

To model the relationship between tumor penetration and the predictor variables, we use a logistic regression model. As most of the covariates are tests covering different aspects of of the prostate, it is not hard to imagine there some of them connected in their relationship with tumor penetration. For this reason, the initial model is fit using all first-order interaction terms. This serves as the starting point for model selection which is done via stepwise regression with AIC as the selection

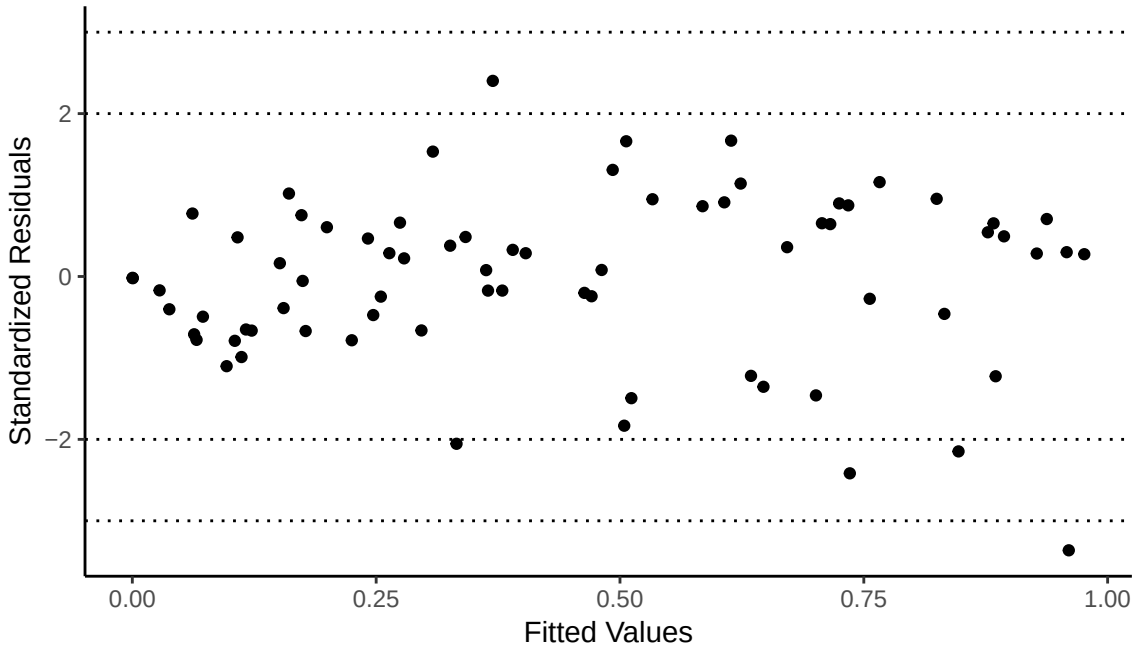

Figure 2: Residual vs. fitted values via Explanatory Variable Patterns (74). The horizontal reference lines indicate the usual residual cutoffs of +-2 and +- 3.

criteria. The model chosen via the stepwise procedure includes race, age, DRE results, PSA value, and Gleason score as well as age interacted with DRE result and race interacted with PSA value. A detailed summary can be found in Table 8. We note here that the AIC is 391 and the p-values for age and race are quite high at 0.11 and 0.99, respectively.

Since some of the p-values of the model obtained via stepwise regression, we remove this terms and refit our model. Specifically, we remove age and race along with interaction terms associated with them. We can not justify keeping variables in interactions that themselves are not included in the model. Note that for this model, removing the relevant interaction terms actually means removing all interaction terms. Once this is done, we are left with a model that relates tumor penetration to just three variables: DRE results, PSA value, and Gleason score.

3.3 Diagnostics

4 Conclusion

Table 3: Race vs. whether tumor penetration of prostatic capsule has occured, marginalized by race.

	Race	
	white	black
No Penetration	59.8	61.1
Penetration	40.2	38.9

Table 4: Detection of capsular involvement vs. whether tumor pentration of prostatic capsule occured, marginalized by capsular involvment.

	Capsular Involvement	
	no	yes
No Penetration	64.1	25.0
Tumor Penetration	35.9	75.0

Appendix

A Exploratory Analysis

A.1 Tables

A.2 Figures

B Model Building and Diagnostics

B.1 Intermediate Models

B.2 Diagnostics

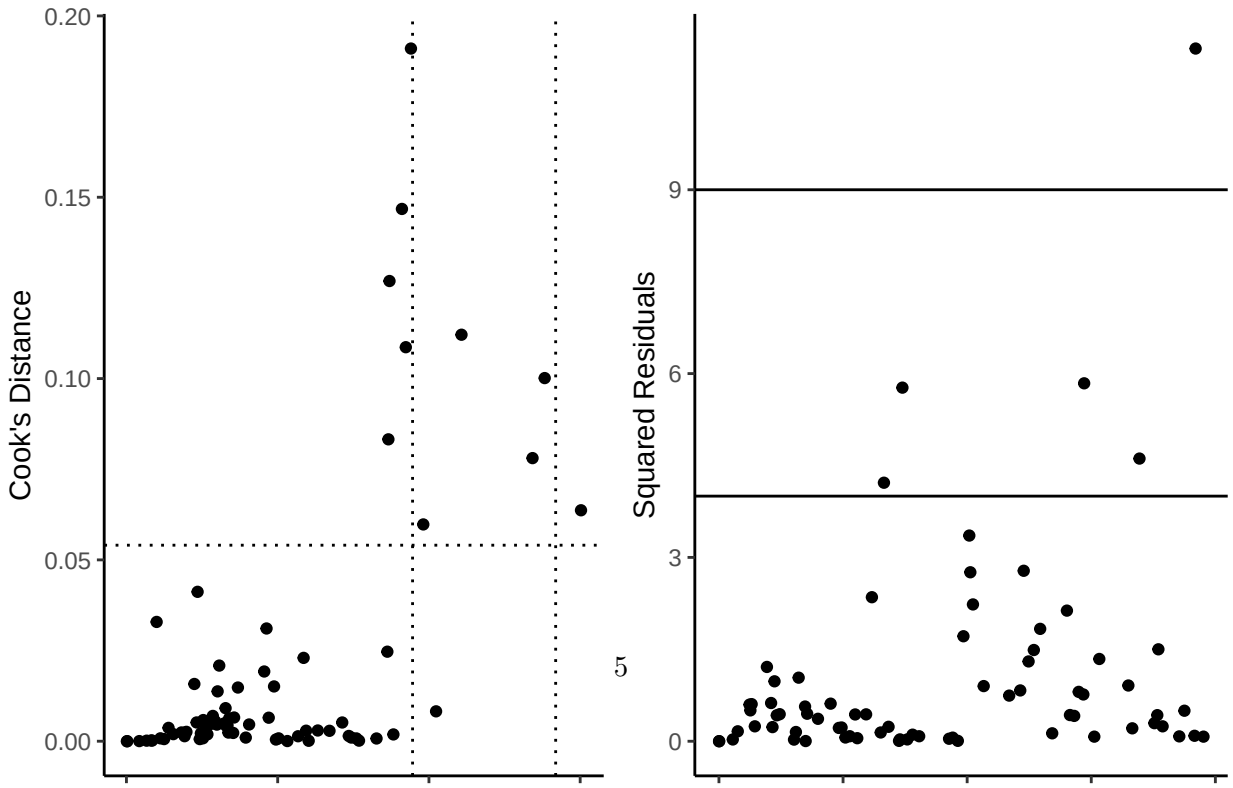

Table 6: Summary statistics of patient age grouped by whether tumor penetration of prostatic capsule occured.

	min	**median**	**mean**	**max**
No Penetration	50.0	67.0	66.3	79.0
Tumor Penetration	47.0	66.0	65.7	79.0

Table 7: Summary statistics for prostatic specific antigen value grouped by whether tumor penetration of prostatic capsule occured.

	min	**median**	**mean**	**max**
No Penetration	0.3	7.5	10.0	66.7
Tumor Penetration	1.4	12.9	23.1	139.7

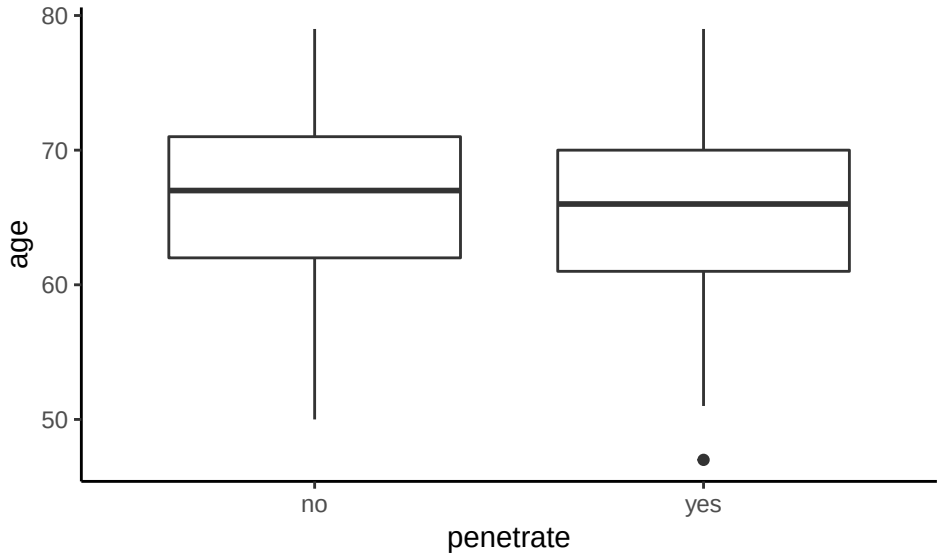

Figure 3: Boxplot of age vs. whether tumor penetration of prostatic capsule occured.

Table 8: Coefficient, standard error, and p-value for model chosen via stepwise procedure. AIC = 391

Term	**Estimate**	**SE**	**p-value**
(Intercept)	-28.15	12.36	0.023
age	0.29	0.18	0.107
raceblack	-0.01	0.66	0.988
dreunilobar left	7.30	4.01	0.069
dreunilobar right	-0.29	4.20	0.944
drebilobar	2.67	5.29	0.614
psa	0.04	0.01	0.001
gleason	3.94	1.84	0.032
age:dreunilobar left	-0.10	0.06	0.095
age:dreunilobar right	0.03	0.06	0.646
age:drebilobar	-0.02	0.08	0.821
age:gleason	-0.04	0.03	0.110
raceblack:psa	-0.03	0.02	0.134

C R Code