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

this seems too short. maybe comibne w/methods again? call it something like Intro and Methods or Intro and Overview?

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 may 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, prostate specific antigen (PSA) value, and total Gleason score. Also included in the data set was an identification code which is not used in this analysis as it simply identifies the patient. For a detailed description of each variable see Table 1. There are two continuous variables, PSA and age, and the remaining variables are categorical. 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

[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	Prostate 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

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 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.

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

say something about the variable types and removing of missing values.

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.

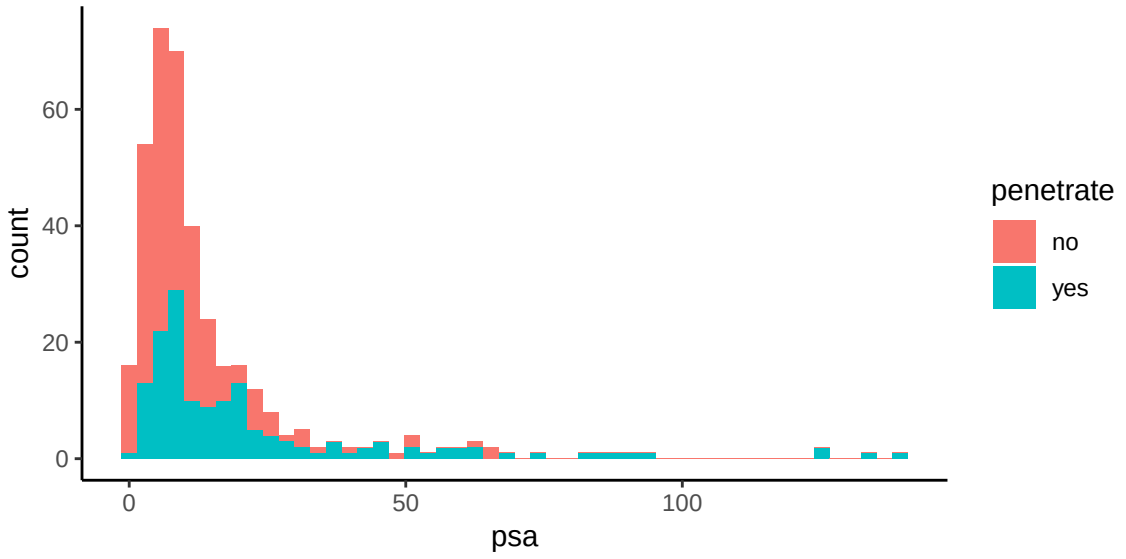

Figure 1: Histogram of prostate specific antigen value colored by whether tumor penetration of prostatic capsule occured.

The relationship for the remaining categorical variables can be found in Tables 4 to 6 in the appendix. 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 7 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.

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 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 9. 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 are large, we remove these 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

To assess the model we use diagnostic tools on explanatory variable patterns[2] (EVP). We collapse the orginal data into groups where each group shares the same covariate values. Recall that one of the variables, PSA, is continous so we first cut the values into nonoverlapping bins. Then we fit the model to the EVPs using the number of original observations in each EVP as weights. Standardized Pearson residuals, Cook's distance, and leverage from the EVP model are what we use to perform our diagnostics. A plot of the residuals vs. fitted probabilites is shown in Figure 2. The dotted horizontal lines indicate the standard cut off values of -3, -2, 2, and 3. There are no discernible patterns in the residuals and they are scattered roughly equally about zero. For larger fitted probabilities there is a bit more variability in the residuals but nothing too concerning. One EVP (lower right hand corner of plot) is slightly smaller than the lowest cut off bound, having a value of -3.4. This is only a minor violation and the cut off values are simply rules-of-thumb so we consider this EVP satisfactory.

Further diagnostic plots can be found in Figure 4 in the appendix. Only one other EVP is of any concern as its Cook's distance and leverage values are both outside the standard cut off values. However, this EVP is also only faintly outside the boundaries[3]. Twenty two of the original observations are contained in this EVP. Given that the Cook's distance and leverage for this EVP are just barely outside their respective cut off, this is not enough justification to remove this many data points.

As a final diagnostic measure we perform a Homer-Lemeshow goodness-of-fit test using the model fit to the EVPs. The EVP data is separated into bins according ot quantiles of the fitted values. Then observed and expected counts are computed for each bin and a χ^2 test is performed. This test is an overall assessment of how well the model fits the data. The obtained test statistic is 3.9 with a corresponding p-value of 0.42. The p-value may seem unfruitful at first as we are typically looking for small values. However in the situation at hand this is not the case. The null hypothesis in this test is that the model fits the data well so we do *not* want to reject the null in our situation. So based on this test our model seems to fit the data well and we can move on to inferences.

3.4 Inferences

Table 3 contains a summary of the final model used in this analysis. Included in the table are estimates for the coefficients with their associated standard errors and p-values. Also include are the odds ratio for each term and 95% confidence intervals for the odds ratio. The largest odds ratio corresponds to a digital rectal exam where a unilobar nodule is found on the patient's ride side. The odds of experiencing tumor penetration of the prostatic capsule for such patients is 4.73 times

[2] Also called covariate patterns.

[3] Cook's distance = 0.06, leverage = 0.3 with cutoff values of 0.05 and 0.28.

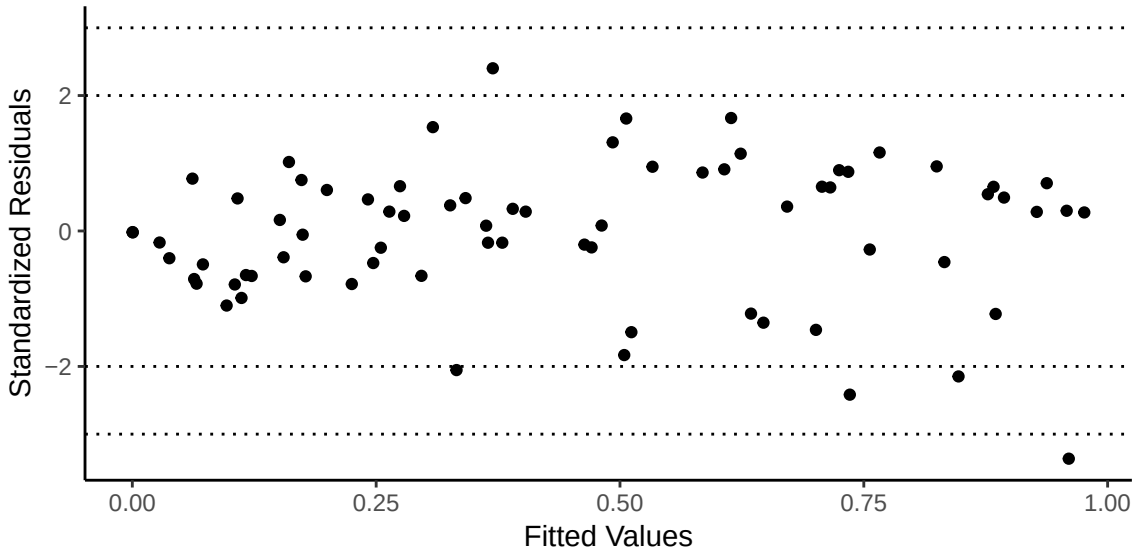

Figure 2: Pearson residuals vs. fitted values via Explanatory Variable Patterns (74). The horizontal reference lines indicate the usual residual cutoffs of +-2 and +- 3.

Table 3: Summary for final model fitting tumor penetration of prostatic capsule with covariates digital rectal exam result, prostate specific antigen value, and total Gleason score. Coefficient estimates with standard errors and p-values are provided along with odds ratios and 95% confidence intervals for odd ratios. AIC for the model is 393.

Term	**Estimate**	**SE**	**p-value**	**Odds ratio**	**OR 95% CI**
Intercept	-8.14	1.06	0.000	0.00	(0.00 , 0.00)
DRE: unilobar left	0.77	0.36	0.030	2.17	(1.09 , 4.43)
DRE: unilobar right	1.55	0.37	0.000	4.73	(2.32 , 10.00)
DRE: bilobar nodule	1.43	0.45	0.001	4.18	(1.75 , 10.25)
Prostate specific antigen value	0.03	0.01	0.004	1.03	(1.01 , 1.05)
Gleason score	1.00	0.16	0.000	2.71	(2.00 , 3.76)

more than the odds for patients which did not have a nodule present. Interestingly, the odds for patients with a unilobar nodule on the left are about half the odds of tumor pentration for patients with either a nodule on the right or a bilobar nodule. The odds ratio for Gleason score indicates that each unit increase in score corresponds to increase in odds of 2.71 times, if all other variables are held constant. Since the odds ratio for PSA is smaller than two, it has a slightly different interpretation. One additional mg/ml of PSA is associated with a 3% increase in odds.

In addition to making inferences on the odds ratios of each of the variables in our model, we can assess the models potential predictive ability. The primary goal of this analysis is to determine which factors have a particularly strong influence on tumor penetration. So this is not a prediction problem per se and so we did follow the usual paradigm of splitting our data into a training and testing set. This means that the predictions we consider here are in sample predictions and will no doubt have much better performance that what would actually be seen in practice. Nonetheless, we can still get some sense of how well the model performs at predicting tumor penetration. To that end, we plotted a receiver operating characteristic (ROC) curve which we leave in the appendix

(Figure 5) for the curious reader. The curve indicates that our model performs significantly better than the default comparison of randomly assigning whether tumor penetration will occur. The area under the ROC curve (AUC) for our final model was 0.82 which is quite good. The ROC curve and AUC indicate that the model developed here has potential as an overall diagnostic tool that combines the results of a digital rectal exam, prostate specific antigen value, and Gleason score into a single measure of risk, rather than considering each measure individually.

4 Conclusion

Table 4: 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 5: 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

B.3 Inferences

C R Code

Table 6: Total Gleason score vs. whether tumor penetration of prostatic capsule occured, marginalized by Gleason score.

	Gleason Score						
	0	4	5	6	7	8	9
No Penetration	100.0	100.0	91.0	72.5	43.3	20.7	7.7
Tumor Penetration	0.0	0.0	9.0	27.5	56.7	79.3	92.3

Table 7: 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 8: Summary statistics for prostate 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

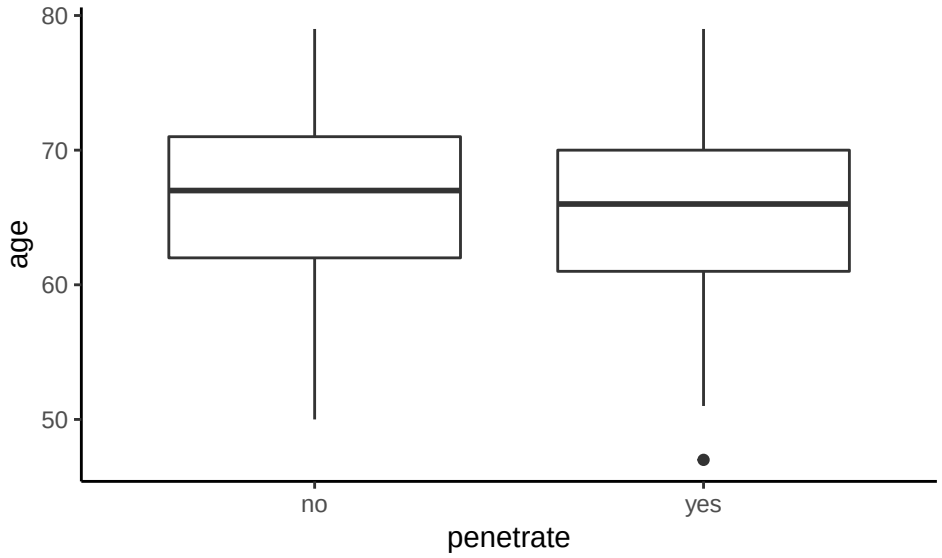

Figure 3: Boxplot of age vs. whether tumor penetration of prostatic capsule occured.

Table 9: Summary of best stepwise fit of tumor penetration of prostatic capsule against age, race, digital rectal exam result, prostate specific antigen value, total Gleason score and the interaction terms age * dre results and race * psa. AIC is 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

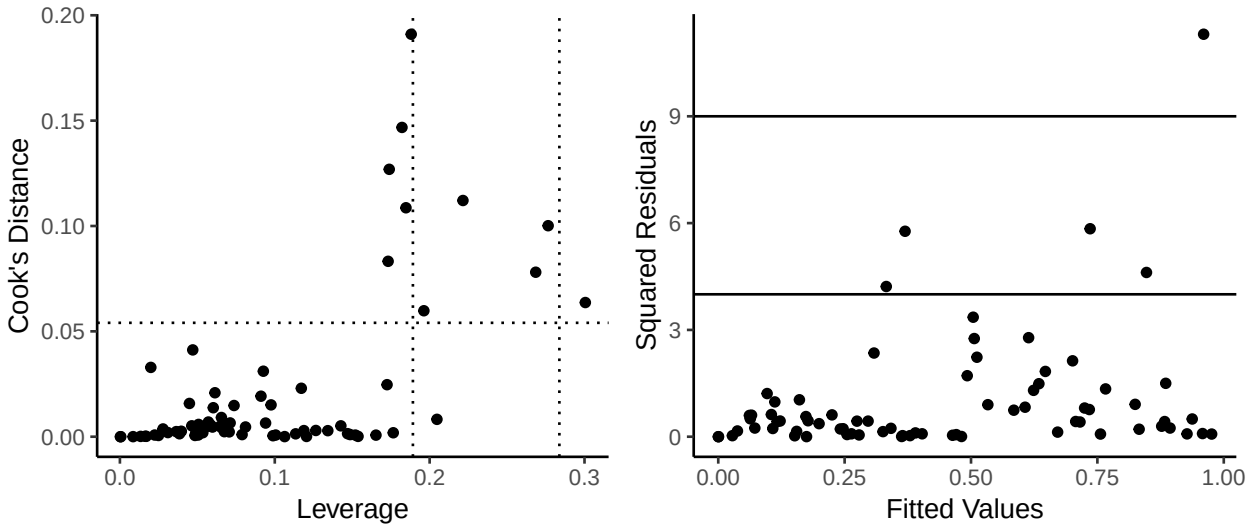

Figure 4: Diagnostic plots for the model fit on explanatory variable patterns. Left: Cook's distance vs. leverage. Right: squared Pearson standardized residuals vs. fitted probabilities

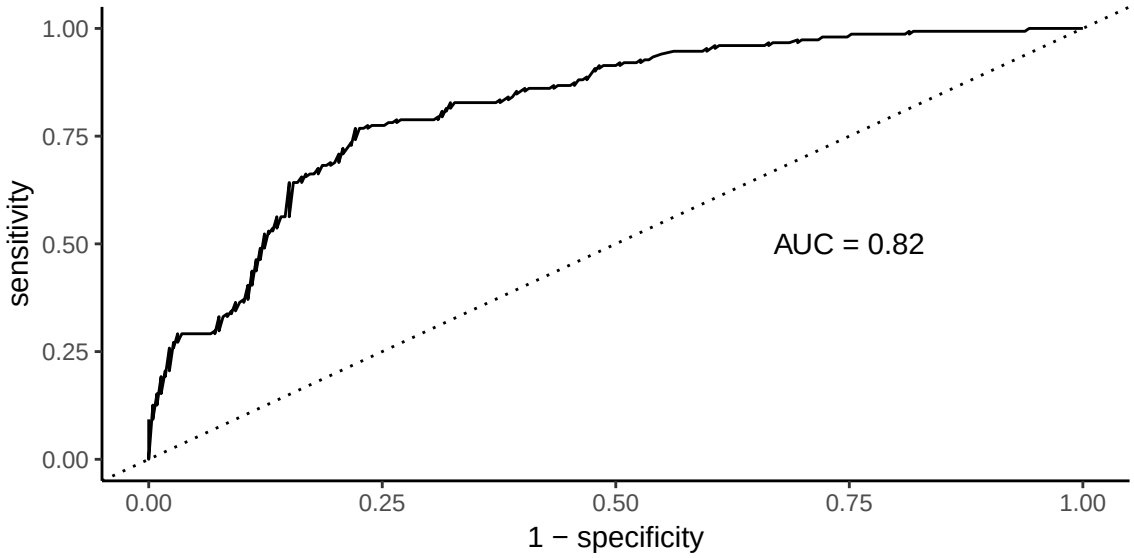

Figure 5: Reciever Operating Characteristic curve for final model fitting tumor penetration of prostatic capsule with covariates digital rectal exam result, prostate specific antigen value, and total Gleason score. The dotted line is a reference line indicating what the curve would be if we randomly decided whether tumor penetration occured.