Prostate Data Analysis

Intro

Prostate cancer is a significant health concern, particularly among men, with advanced stages demanding surgical intervention for effective management. In this data analysis, we explore the relationship between various predictor variables and the severity of prostate cancer, as measured by its grade, among a cohort of 97 men about to undergo surgery. Grade, representing the severity of prostate cancer, is categorized into values 6, 7, and 8, with higher grades indicating worse conditions.

The predictor variables considered in our analysis include prostate-specific antigen level, tumor volume, prostate gland weight, patient age, benign prostate hyperplasia, invasion of the seminal vesicle, and degree of capsular penetration. These variables encompass crucial aspects of the disease, ranging from biochemical markers like PSA to anatomical characteristics such as tumor volume and gland weight. The primary objective is to identify the model that best elucidates the relationship between these predictor variables and the severity of prostate cancer grade.

Problems Encountered

The first problem encountered during the analysis was during the initial exploratory analysis portion, involving inspecting the variables present in the dataset. In this portion, histograms were made for each of the variables in the dataset. It was apparent that some of the variables needed to be treated as factors when currently they were stored as numeric variables. These variables that needed to be converted to factors were “inv”, whether the tumor has invaded the seminal vesicle, and “grade”, a measure of the severity of prostate cancer. The latter was converted to an ordered factor as there is a natural ordering to the levels of the “grade” variable. The decision to change these to factors was also based on the context of what the variables represented.

A second problem faced was issues with multicollinearity between two of the predictors in what was thought to be the final model. This model contained the variables psa, wt, age, cap, and the interaction between weight and cap. In this model the interaction term had a very high VIF value which shows strong evidence of multicollinearity. To fix this issue, the interaction term was removed.

Some more general issues that were faced during the analysis were issues with compatibility with R functions for checking VIFs, influential observations, and overdispersion. A cumulative logit proportional odds model was used for this data, this specification was implemented using the VGAM package in R. The object that the model was stored in using the vglm function is not compatible with a large amount of the functions in the car package. Because of this, an alternative method was used to check these issues. This involved fitting binomial logit

models on the data, grouping 2 of the levels of the response variable "grade" into one. This unlocked compatibility with the functions within the car package for checking the model issues. This also involved looking at every combination of level groupings into 2 levels. For example the model containing the logit of cancer severity level 6 and 7 versus 8, 6 and 8 vs 7, and 7 and 8 vs 6.

GLM Description

The final model is a cumulative logit, proportional odds model. Since we have 3 levels in the response variable, there are 2 cumulative logits, they are for severity level 6 and severity level 7. For severity level 6 the logit is as follows:

$$\text{logit}[P(y_i \leq 6)] = \log\frac{P(yi \leq 6)}{P(yi > 6)}$$

And for severity level 7:

$$\text{logit}[P(y_i \leq 7)] = \log\frac{P(yi \leq 7)}{P(yi > 7)}$$

A logit for severity level 8 is not needed because we can obtain the value by subtracting the severity level 7 logit from 1. Since this is a proportional odds model we are assuming the association between the predictor variables and the response is constant across logits. However, we will have 2 intercepts, one for each of the logits.

Predictors

The predictors used in the final model are psa, age, and cap. There are two intercepts, one for each of the logits. The linear predictor is as follows:

$$\text{logit}[P(y_i \leq j)] = \alpha_j - 0.052*\text{psa} - 0.066*\text{age} - 0.185*\text{cap}$$

Where $\alpha_6 = 4.515$ and $\alpha_7 = 7.289$

Interpretations

- α_6 **:** The estimated probability of a 0 year old man with advanced prostate cancer about to undergo surgery with antigen level 0 mg/ml and 0 degree of capsular penetration being in cancer severity level 6 is .989
- α_7 **:** The estimated probability of a 0 year old man with advanced prostate cancer about to undergo surgery with antigen level 0 mg/ml and 0 degree of capsular penetration being in cancer severity level 7 or lower is .999
- **psa:** An increase in 1 mg/ml in antigen level is associated with multiplying the

cumulative odds of being in a cancer severity level or lower by .949 for men with advanced prostate cancer about to undergo surgery, after adjusting for age and degree of capsular penetration.

- **age:** An increase in 1 year in age is associated with multiplying the cumulative odds of being in a cancer severity level or lower by .936 for men with advanced prostate cancer about to undergo surgery, after adjusting for antigen level and degree of capsular penetration.
- **cap:** An increase in 1 degree of capsular penetration is associated with multiplying the cumulative odds of being in a cancer severity level or lower by .831 for men with advanced prostate cancer about to undergo surgery, after adjusting for antigen level and age.

Remaining Issues

A possible concern with the model is that the multicollinearity and influential observations that were checked involved grouping the severity levels into two groups. This means the form of the data that we fit the model is different from what we used to test for these model issues. So, it is possible that the results of these checks are slightly compromised and if that is the case then the results of the final model could also be compromised. If there was multicollinearity present then the process of the variable reduction could have resulted in a different outcome for the predictors in the final model. The standard errors of the coefficients could have been inflated and therefore the assessment of significance of predictors inaccurate. For the same reason the analysis of influential observations could have changed.

It also be noted that these results must only be generalized to men similar to those in the study. That is, men with advanced prostate cancer soon to undergo surgery.

Prediction Table

The first 3 columns contain the predictor values and the last 3 contain the probability of being in each of 3 severity levels

	psa	age	cap	grade=6	grade=7	grade=8
1	0.65	41	0.00	0.8571	0.1326	0.0103
2	13.33	41	0.00	0.7573	0.2231	0.0196
3	265.07	41	0.00	0.0000	0.0001	0.9999
4	0.65	65	0.00	0.5539	0.3982	0.0479
5	13.33	65	0.00	0.3925	0.5194	0.0881
6	265.07	65	0.00	0.0000	0.0000	1.0000
7	0.65	79	0.00	0.3314	0.5567	0.1119
8	13.33	79	0.00	0.2050	0.6000	0.1950
9	265.07	79	0.00	0.0000	0.0000	1.0000
10	0.65	41	0.45	0.8466	0.1422	0.0112
11	13.33	41	0.45	0.7417	0.2370	0.0213
12	265.07	41	0.45	0.0000	0.0001	0.9999
13	0.65	65	0.45	0.5333	0.4149	0.0518
14	13.33	65	0.45	0.3728	0.5321	0.0951
15	265.07	65	0.45	0.0000	0.0000	1.0000
16	0.65	79	0.45	0.3132	0.5664	0.1205
17	13.33	79	0.45	0.1917	0.5999	0.2084
18	265.07	79	0.45	0.0000	0.0000	1.0000
19	0.65	41	18.17	0.1727	0.5970	0.2302
20	13.33	41	18.17	0.0980	0.5370	0.3650
21	265.07	41	18.17	0.0000	0.0000	1.0000
22	0.65	65	18.17	0.0414	0.3676	0.5909
23	13.33	65	18.17	0.0220	0.2428	0.7352
24	265.07	65	18.17	0.0000	0.0000	1.0000
25	0.65	79	18.17	0.0170	0.1995	0.7835
26	13.33	79	18.17	0.0089	0.1168	0.8743
27	265.07	79	18.17	0.0000	0.0000	1.0000

Final Model Output

	Estimate	Std Error	z-value	P(>|z|)
Intercept:1	4.515	2.014	2.424	.025
Intercept:2	7.289	2.112	3.450	.001
psa	-0.052	0.016	-3.212	.001
age	-0.066	.032	-2.082	.037
cap	-0.185	0.076	-2.430	0.015