STAT 512 : Final Project Report
Factors associated with blood PSA levels in prostate cancer.

T. Padma Ragaleena

April 29, 2022

1 Introduction

Prostate Cancer is one of the most common cancer among men in America. According to the American Cancer society, about one in eight American men will be diagnosed with prostate cancer during his lifetime. In fact, prostate cancer is the second leading cause of cancer death among American men. Given the prevalence of prostate cancer, it is clear that a better understanding of risk factors associated with prostate cancer risk is a very important step forward in helping people diagnosed with cancer.

In this report, we analyze a dataset collected on 97 men diagnosed with prostate cancer. Prostate is a gland found in men that is responsible for producing the seminal fluid that nourishes and transports sperms. Therefore, prostate cancer, as the name suggests, is the cancer that occurs in the prostate. A prostate cancer patient either have their cancer confined to the prostate or the cancer spreads to the outer wall of the prostate. The latter scenario is called *capsular penetration*. Doctors sometimes recommend a surgical procedure called *radical prostatectomy* as a treatment option for patients diagnosed with this cancer. However, there is a risk associated with this procedure. Around 10% of the men who undergo this surgery are effected by a condition called *seminal vesicle invasion* (SVI), which is said to occur when cancer spreads to a nearby gland called seminal vesicle. Finally, prostate gland enlargement is also called *Benign Prostatic hyperplasia* (BPH). BPH causes uncomfortable urinary symptoms and is a common condition as men get older.

Amount of a protein called *Prostate Specific Antigen*(PSA) in blood samples of patients is a good indicator of prostate cancer. Men with high levels of PSA are more likely to be diagnosed with prostate cancer. Once a person has been diagnosed with prostate cancer, it is important for doctors to stage and grade the cancer i.e. understand the cancer's growth, spread and how it looks like under a microscope. *Gleason score* is one such cancer staging and grading system. A Lower Gleason score indicates a low grade cancer, i.e. cancer that grows more slowly and less likely to spread.

2 Prostate cancer study design

The prostate cancer study considered in this report is based on data collected on 97 men who were about to undergo radical prostatectomies. The experimenters are interested to understand the association between Prostate specific antigen (PSA) and a number of clinical measurements taken in men with advanced prostate cancer. Each of these 97 patients is associated with an ID number between 1 and 97 to protect their identities. Eight other clinical measurements were taken for each of the 97 patients involved in this study. The response variable of interest is PSA while the independent variables involved are ID, cancer volume, weight of prostate, age, amount of BPH, SVI status, degree of capsular penetration and Gleason score. Table 1 summarizes important information about the dependent and independent variables involved in the study.

As it can be observed in Table 1, there are two factors involved in the study - SVI and Gleason. The factor SVI is a binary factor indicating the presence or absence of SVI, while Gleason is a factor with three levels - 6, 7, 8. Therefore, this is a *two-factor study* with six possible treatment combinations. Each of the 97 patients

Variable name	Variable type	Description
ID	Indexing variable	Identification of patients (1-97)
PSA	quantitative/numerical	Serum PSA level (mg/ml)
Cancer volume	quantitative/numerical	Estimate of prostate cancer volume (cc)
Prostate weight	quantitative/numerical	Prostate weight (gm)
Age	quantitative/numerical	Age of patient (years)
BPH	quantitative/numerical	Amount of BPH (cm^2)
SVI	Factor with levels 0, 1	0 (Absence of SVI), 1 (Presence of SVI)
Capsular Penetration	quantitative/numerical	Degree of capsular penetration (cm)
Gleason score	Factor with levels 6, 7, 8	6 (Low grade cancer), 7 (Medium grade cancer), 8 (high grade cancer)

Table 1: Table summarizing the variables involved in the Prostate cancer data set.

involved in the study are the *experimental units* of this study. Both the factors involved are *observational factors,* as the treatment combination corresponding to each person is not a result of random assignment. Instead, both the factor levels are intrinsic to the patient's condition. Since there is no randomization involved in assigning treatments to experimental units, it is clear that this is an *observational study.* The SVI effects must be modelled as fixed effects, as there are exactly two possible SVI statuses (SVI present or absent). Gleason score, by definition, is an integer score given by pathologists that is one of 6, 7, or 8. Therefore, factor effects associated with Gleason score are fixed.

SVI \Gleason	6	7	8
0	32	34	10
1	1	9	11

Table 2: Number of sample points under each treatment combination

Based on Table 2, we observe that there exists at least one observation under every treatment combination. Therefore, the *factors are crossed*, and we are dealing with a *two-factor study with unequal sample sizes* under each treatment combination.

3 Exploratory data analysis

The Figure 1 gives an initial picture of how the data looks like and what properties it has. Based on plot in Figure 1, we observe that all quantitative variables except that of age have a right skewed kernel density estimate. Age has a left skewed kernel density estimate. Some pairs of quantitative variables appear to be sufficiently correlated. Except for a couple of scatter plots, an obvious pattern is not clear. However, most plots appear to have data points packed densely in one region of the graph. Each plot should be looked at individually for better understanding.

Our aim is to fit a regression model that assumed IID normal errors. This means that the response of interest (here it is PSA) must be approximately normally distributed. Therefore, applying a *log transformation* to the

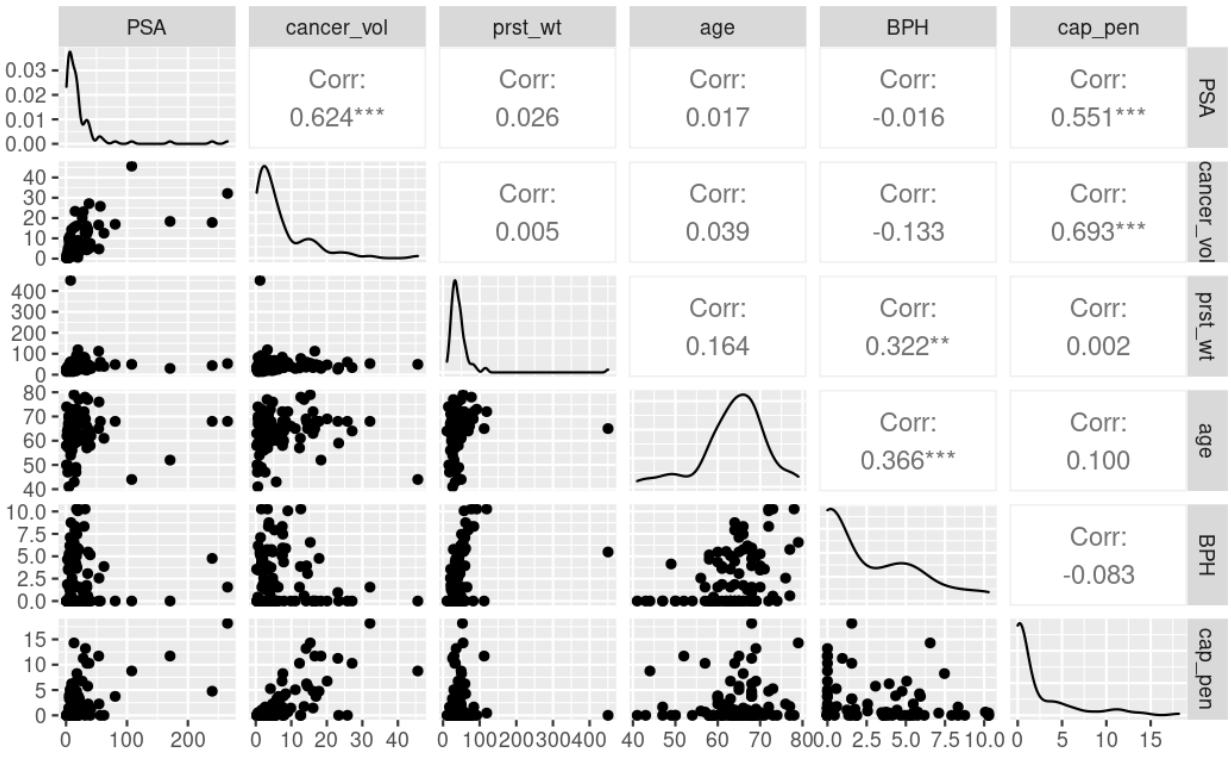
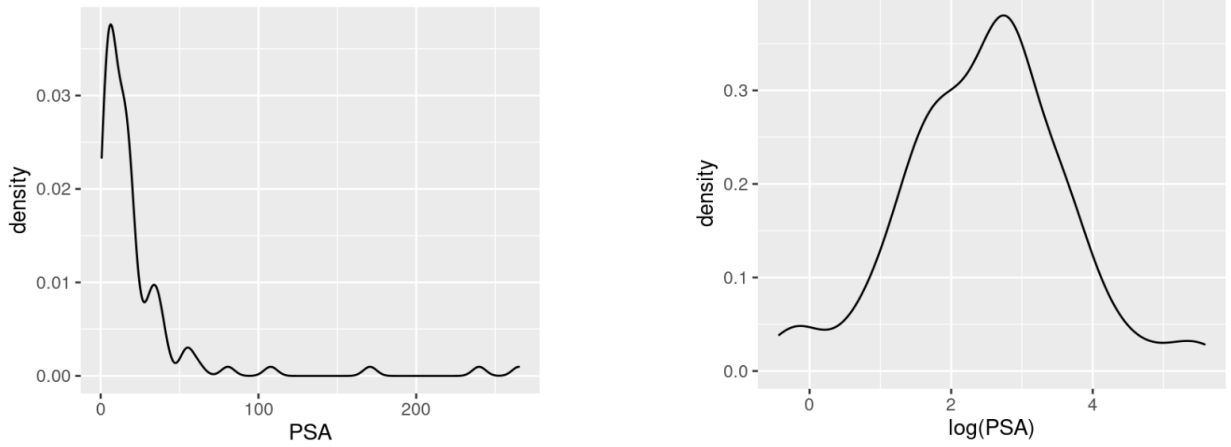


Figure 1: Kernel density estimates and correlation coefficients

response PSA will be appropriate, as it would make the distribution less right skewed and more symmetric. This is what we observe in the plots in Figure 2a and Figure 2b.



(a) KDE of PSA values

(b) KDE of log(PSA) values

Figure 2: Kernel density estimate of PSA before and after log transformation

Figure 3 is a paired plot constructed (similar to Plot in Figure 1) for log transformed PSA values and other untransformed quantitative variables given in Table 1. The data points in plots involving log(PSA) appear to be more spread out now than before. In order to model the other quantitative variables as *covariates*, we need evidence that the quantitative variables vary approximately linearly with the log(PSA) values. Since this is not clear in Figure 3, individual quantitative variables should be examined to see if they must be

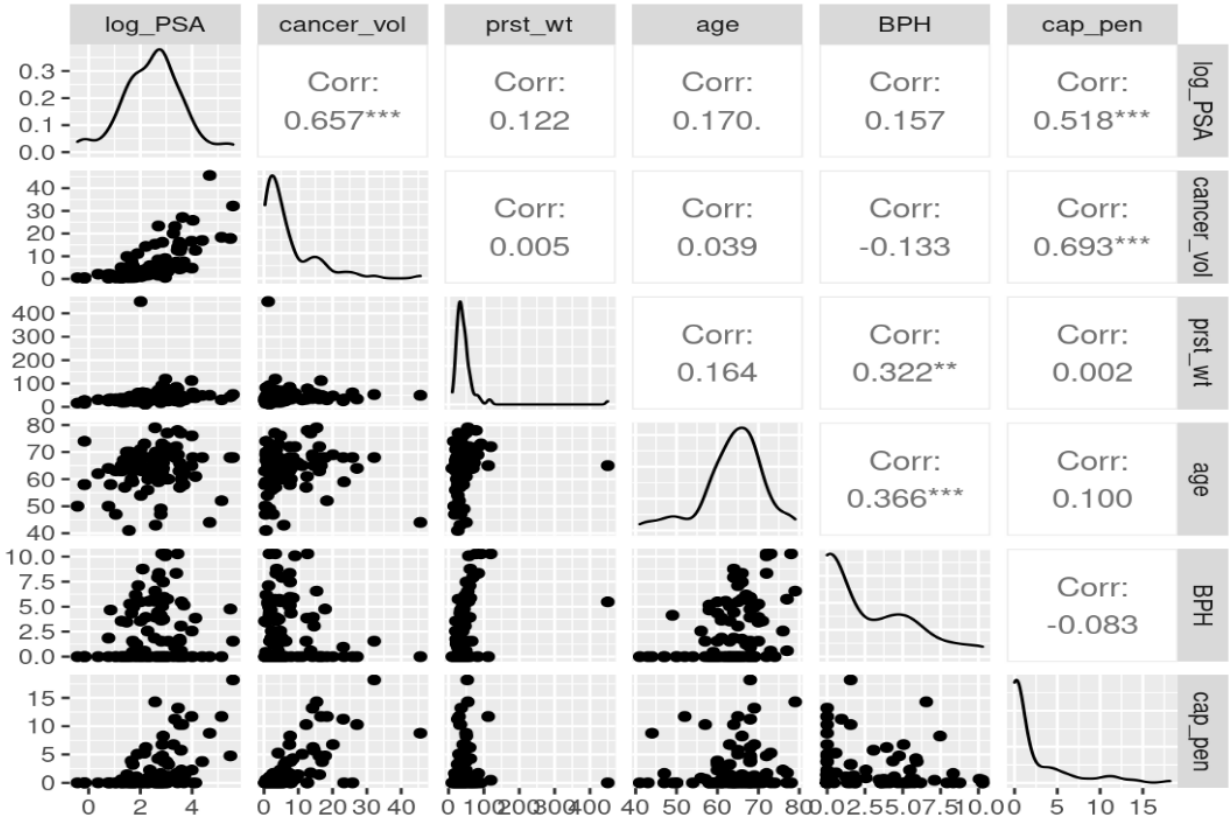
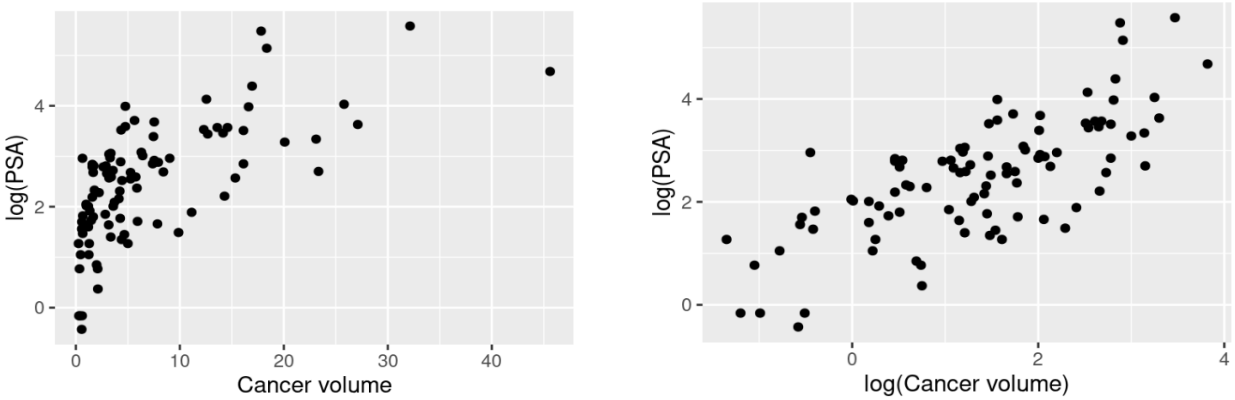


Figure 3: Kernel density estimates and correlation coefficients for log(PSA) and other quantitative

transformed as well.

The scatter plot of log(PSA) against cancer volume is given in Figure 4a. The plot looks like a random scatter around the curve $y = \log(x)$. Therefore, applying a log transformation on the predictor can result in an approximate linear relationship between response and predictor. This is exactly what we observe in the second plot of Figure 4b.

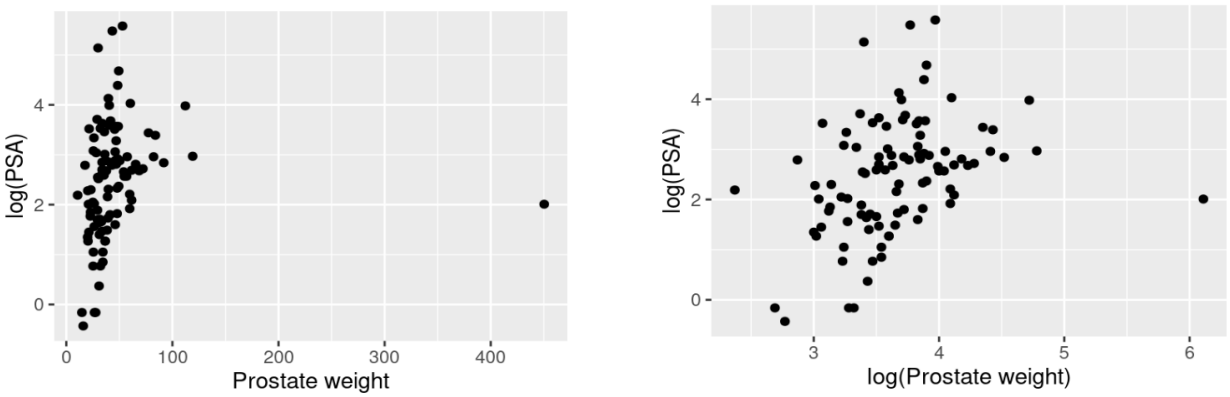


(a) log(PSA) against cancer volume

(b) log(PSA) against log(cancervolume)

Figure 4: Scatter-plot of log(PSA) against untransformed and log transformed cancer volume values

In the scatter plot of log(PSA) against prostate weight in Figure 5a, we observe that the values of log(PSA) increase rapidly with respect to prostate weight in a way that is similar to the behavior of $y = x$. There is one point in Figure 5a that doesn't quite fit with the trends of the rest of the data points. For the prostate weight variable, the log transform is again recommended for two reasons. First, x tends to infinity faster than $\log(x)$. Therefore, the prostate weight data points are more likely to spread out after applying a log transform. This might help avoid scatter around a line that is almost vertical. Second, since the $\log(x)$ function increases at a slower rate than x, transformation will drag the extreme point closer to the remaining points. This is what we observe in Figure 5b. In fact, the plot in Figure 5b exhibits approximate linear trend.



(a) log(PSA) against Prostate weight

(b) log(PSA) against log(Prostate weight)

Figure 5: Scatter-plot of log(PSA) against untransformed and log transformed Prostate weight values

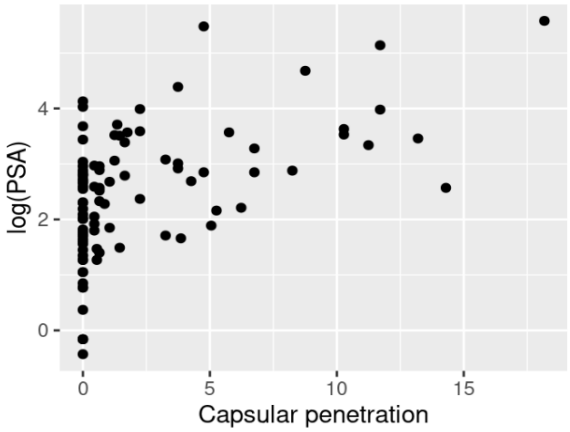
When we plot log(PSA) against capsular penetration, the plot looks approximately linear, as shown in Figure 6a. There are several observations that take 0 capsular penetration values. Therefore, applying a log transform to this variable is not appropriate. A plot of log(PSA) against non-zero log transformed capsular penetration values is in Figure 6b. It is clear that the plot in Figure 6b shows greater linear trend than plot in Figure 6a. One option would be to add a small error term to each of the 0 capsular penetration observations so that the log transform can be applied. But adding noise this way would mean that all the zero capsular penetration values are transformed to large negative values, while the non-zero capsular penetration values transform to value closer to zero. This means that log transformed data is far more spread out than the actual data. This change in the nature of the data might hide some useful information from the untransformed data. Therefore, capsular penetration should be left untransformed.

The plots of, log(PSA) against age and BPH, is given in Figure 7. Except for the two points highlighted in orange, all the other points in log(PSA) against age plot of Figure 7a seem to follow an approximate linear trend. Therefore, no transformation is required for the variable age. All points, with non-zero BPH values, do not appear to follow any particular trend. BPH appears to be a random scatter around a line parallel to the x-axis (a scatter with non-constant variance). The number of available observations reduces as the value of BPH increases. Since no non-linear trend is observed, there is no need for a transformation on BPH.

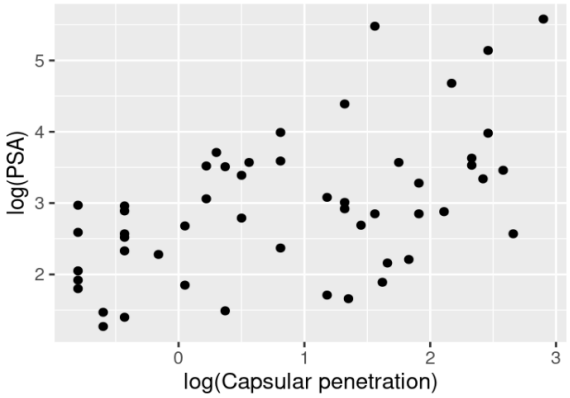
Table 3 summarizes the response and predictor transformations that were concluded to be appropriate based on graphical analysis done so far.

Therefore, the final set of quantitative variables that we are going to work with are log(PSA) (response), log(Cancervolume), log(Prostateweight), Age, BPH, and Capsular penetration. Based on the plots constructed so far, it is clear that we observe approximate linear relationship between the response and other quantitative variables. If it can be seen through graphical evidence that different treatment combination regression lines have equal slopes, then the variables log(Cancervolume), log(Prostateweight), Age, BPH, and Capsular penetration can be modelled as *covariates*.

Based on the plots in Figure 8, it is reasonable to say that regression lines corresponding to different treatments can be approximated as parallel lines for all the quantitative variables. Therefore, the quantitative

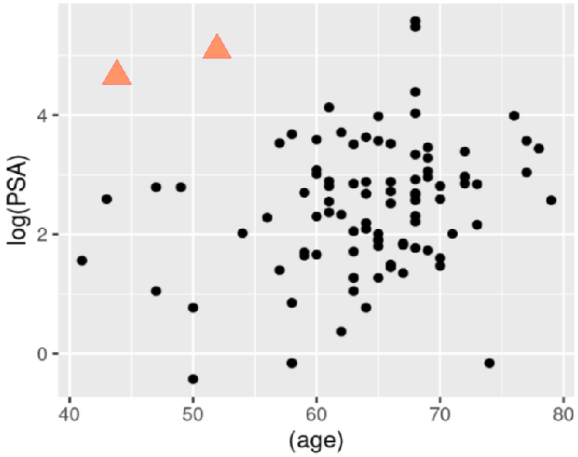


(a) log(PSA) against all capsular penetration values

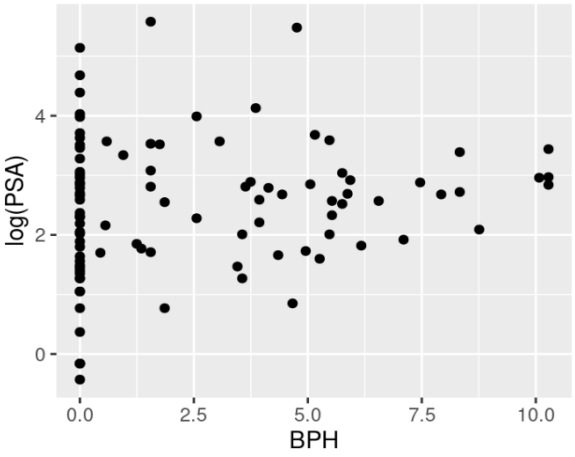


(b) log(PSA) against log transformed non-zero capsular penetration values.

Figure 6



(a) log(PSA) against age



(b) log(PSA) against BPH

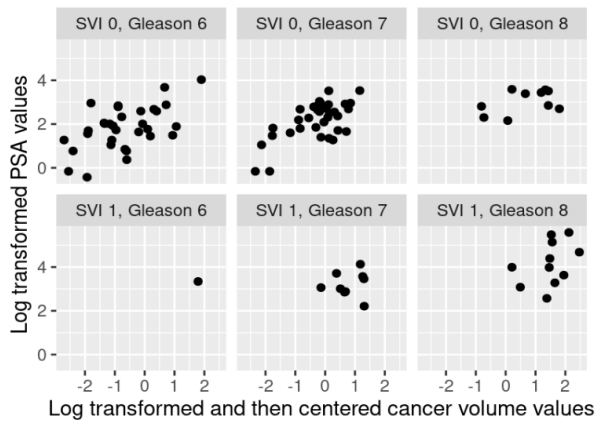
Figure 7: the general caption

Variable	Transformation required (if any)
PSA	log(PSA)
Cancer volume	log(Cancervolume)
Prostate weight	log(Prostateweight)
Age	No transformation applied
BPH	No transformation applied
Capsular penetration	No transformation applied

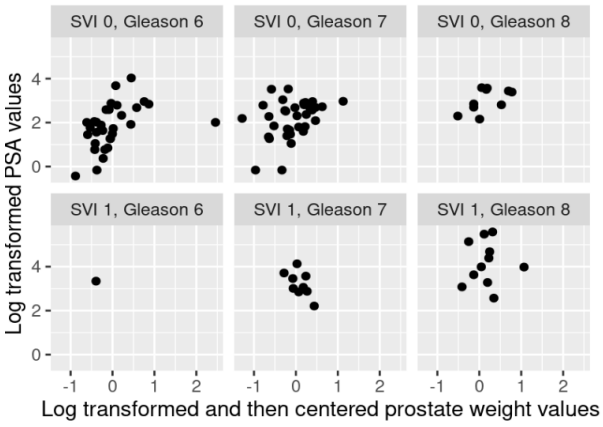
Table 3: Summary of transformations required (if any) on quantitative variables.

variables can be modeled as covariates. Although this seems like a fair assumption to make, one can observe from Figure 8 that the number of observations in some treatment combinations is too low to predict a trend in data.

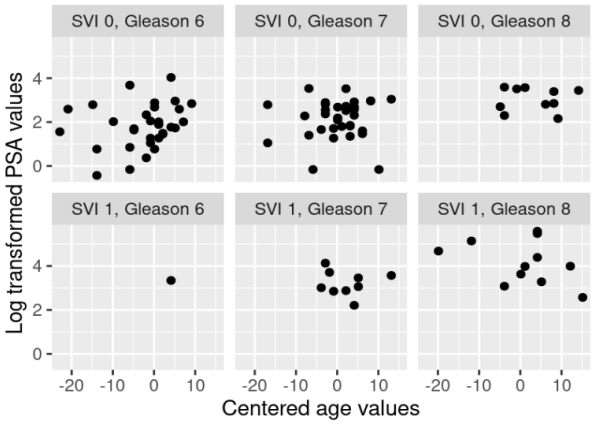
Our regression model includes two factor variables - SVI and Gleason score. Graphical evidence for whether the factors interact can be obtained through an interaction plot. Based on the interaction plot in Figure 9,



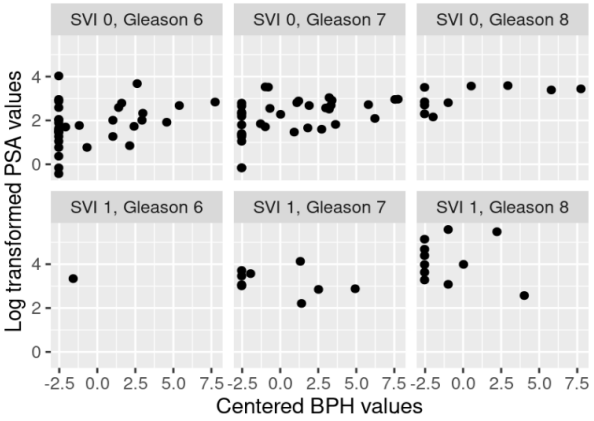
(a) log(PSA) against log(cancervolume)



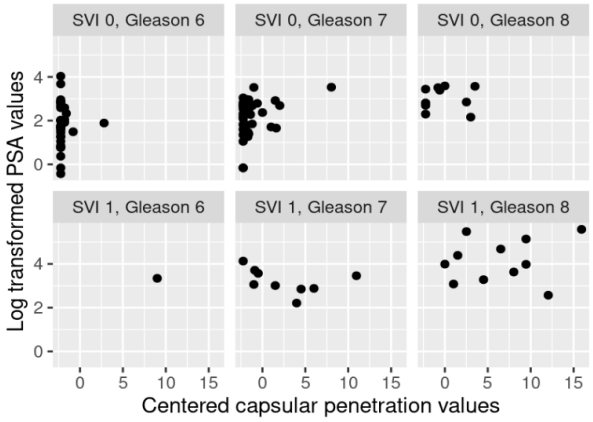
(b) log(PSA) against log(prostateweight)



(c) log(PSA) against age



(d) log(PSA) against BPH



(e) log(PSA) against capsular penetration

Figure 8: Plots of log(PSA) against different quantitative variables grouped by treatment combination.

there appears to be little to no interaction between the factors SVI and Gleason score.

Although we tried to obtain graphical evidence on how the regression model needs to be like, we need to do hypothesis testing to be more confident about our claims of no interaction and equal slope for all treatment combinations. This has been dealt with in the next section.

Figure 9: Interaction plot

4 Modelling, its adequacy and inference

Based on the exploratory data analysis alone, the following ANCOVA model will be a good fit for the data.

$$Y_{ijk} = \mu_{...} + \alpha_i + \beta_j + \gamma_1 X_{ijk}^{(1)} + \cdots + \gamma_5 X_{ijk}^{(5)} + \epsilon_{ijk} \text{ for } (i,j,k) \in \{1,2\} \times \{1,2,3\} \times \{1, \cdots n_{ij}\} \quad (1)$$

An ANCOVA model implies the assumption that $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$ for all i, j, k. α_i denotes i-th level SVI fixed factor effects subject to the restriction $\alpha_1 + \alpha_2 = 0$. β_j denotes j-th level SVI fixed factor effects subject to the restriction $\beta_1 + \beta_2 + \beta_3 = 0$. The variables $X_{ij}^{(1)}, \cdots, X_{ij}^{(5)}$ represent the five covariates corresponding to log(cancervolume), log(prostateweight), age, BPH and capsular penetration respectively. All covariates are centered.

As mentioned earlier, the model (1) can be claimed to be the final model with more confidence if we can prove two statements, in the order mentioned, through hypothesis testing. First, the assumption that slopes of different treatment regression lines for a given covariate are equal. Another assumption incorporated in (1) is that SVI and Gleason factors do not interact. These two statements are first statistically tested in this section.

When treatments interact with the covariate variable, it results in non-parallel slopes for different treatment combinations. Hence, covariance analysis will be inappropriate. In order to test this assumption, we use the general linear approach where the "full model" is as follows:

$$Y_{ijk} = \mu_{...} + \alpha_1 \mathcal{I}_{ij}^{(1)} + \beta_1 \mathcal{I}_{ij}^{(2.1)} + \beta_2 \mathcal{I}_{ij}^{(2.2)} + (\alpha\beta)_{11} \mathcal{I}_{ij}^{(1)} \mathcal{I}_{ij}^{(2.1)} + (\alpha\beta)_{12} \mathcal{I}_{ij}^{(1)} \mathcal{I}_{ij}^{(2.2)} + \left\{ \sum_{z=1}^{5} \gamma_z X_{ijk}^{(z)} \right\} + \quad (2)$$

$$\left\{ \sum_{z=1}^{5} \delta_z^{(\alpha)} X_{ijk}^{(z)} \mathcal{I}_{ij}^{(1)} \right\} + \left\{ \sum_{z=1}^{5} \delta_z^{(\beta 1)} X_{ijk}^{(z)} \mathcal{I}_{ij}^{(2.1)} \right\} + \left\{ \sum_{z=1}^{5} \delta_z^{(\beta 2)} X_{ijk}^{(z)} \mathcal{I}_{ij}^{(2.2)} \right\} \quad (3)$$

Note that the full model in (2) has, in the order mentioned, SVI treatment effects, Gleason treatment effects, interactions between SVI and Gleason treatments, covariates, interaction between covariates and SVI, and interaction between covariates and Gleason.

Test for equality of slopes is equivalent to the following hypothesis testing:

$$H_0 : \delta_z^{(\alpha)} = \delta_z^{(\beta 1)} = \delta_z^{(\beta 2)} = 0 \forall z \in \{1, \cdots, 5\} \ \text{(vs)} \tag{4}$$

$$H_1 : \text{ at least one of } \delta_z^{(\alpha)}, \delta_z^{(\beta 1)}, \delta_z^{(\beta 2)} \text{ for } z \in \{1, \cdots, 5\} \text{ is non-zero} \tag{5}$$

The reduced model to test the hypothesis (4) using general linear approach is given by:

$$Y_{ijk} = \mu_{...} + \alpha_1 \mathcal{I}_{ij}^{(1)} + \beta_1 \mathcal{I}_{ij}^{(2.1)} + \beta_2 \mathcal{I}_{ij}^{(2.2)} + (\alpha\beta)_{11} \mathcal{I}_{ij}^{(1)} \mathcal{I}_{ij}^{(2.1)} + (\alpha\beta)_{12} \mathcal{I}_{ij}^{(1)} \mathcal{I}_{ij}^{(2.2)} + \left\{ \sum_{z=1}^{5} \gamma_z X_{ijk}^{(z)} \right\} \tag{6}$$

F statistic based on general linear approach is calculated using error sum of squares for the reduced and full models.

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \times \frac{df_F}{SSE(F)} = \frac{42.285 - 37.119}{86 - 71} \times \frac{71}{37.119} = 0.6589 \text{ where } F^* \overset{H_0}{\sim} F(15, 71)$$

Based on the F^* value, the p-value associated with testing the hypothesis (4) is 0.815. Therefore, at 5% significance, we fail to reject the null hypothesis. Using covariance analysis is therefore valid as expected based on initial graphical analysis. We can continue to now work with model (6).

The general linear approach is again used to test for the significance of interaction effects in the model (6). Test for interaction effects is equivalent to the following hypothesis test.

$$H_0 : (\alpha\beta)_{11} = (\alpha\beta)_{12} = 0 \ \ \text{(vs)} \ \ H_1 : \text{ at least one of } (\alpha\beta)_{11}, (\alpha\beta)_{12} \text{ is non-zero.} \tag{7}$$

The full model in this case is (6) while the reduced model is (1).

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \times \frac{df_F}{SSE(F)} = \frac{42.649 - 42.285}{88 - 86} \times \frac{86}{42.285} = 0.3701 \text{ where } F^* \overset{H_0}{\sim} F(2, 86)$$

Based on the F^* value, the p-value associated with testing the hypothesis (4) is 0.692. Therefore, at 5% significance, we fail to reject the null hypothesis i.e. interactions between SVI and Gleason are statistically insignificant.

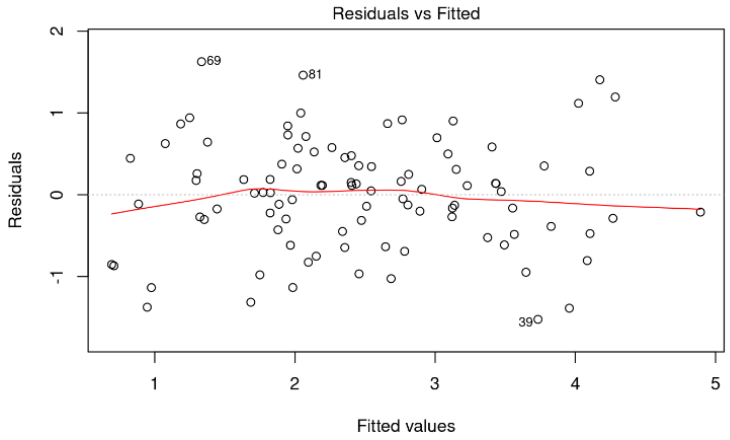
Therefore, a no interaction, equal slope ANCOVA model can be employed for the given data set. The appropriateness of this model can be observed based on plots in Figure 10. There is no evidence of severe deviations from the assumptions of homoscedasticity and normality based on plots in Figure 10a and Figure 10b respectively. The plot in Figure 10c indicates the absence of any influential points in the model.

Finally, since some predictors appeared significantly correlated in the initial exploratory analysis, we check for presence of multi-collinearity using *Generalized variance Inflation Factors* (GVIFs). GVIF values more than five indicate presence of moderate to severe multi-collinearity. Since all GVIF values in Table 4 between 1 and 2.5, there is not enough evidence of multi-collinearity.

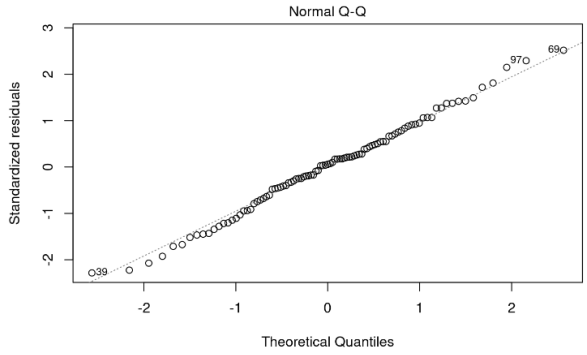
Variable	GVIF
centered log(Cancervolume)	1.930
centered log(Prostateweight)	1.541
centered Age	1.263
centered BPH	1.632
centered Capsular penetration	2.312
SVI	2.022
Gleason	1.624

Table 4: Generalized variance inflation factors of variables involved in model (1).

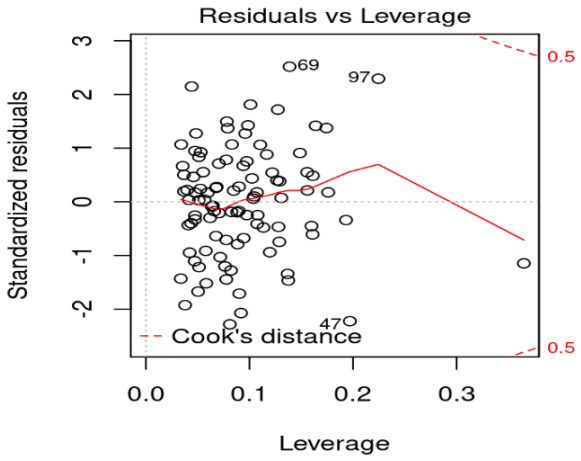
Least square estimates of the fitted model are provided in Table 5. The p-values associated with test for significance of individual coefficients is provided as well.



(a) Residuals vs fitted values plot



(b) Normal Q-Q plot



(c) Standardized residuals vs leverage

Figure 10: Model adequacy checking for the no interaction, equal slope model.

Variable	Least square estimate	p-value
Intercept	2.742	$\approx 10^{-16}$
centered log(Cancervolume)	0.513	$\approx 10^{-8}$
centered log(Prostateweight)	0.390	0.0307
centered Age	-0.017	0.1225
centered BPH	0.047	0.1173
centered Capsular penetration	0.025	0.3785
SVI = 0	-0.362	0.0039
Gleason = 6	-0.269	0.0267
Gleason = 7	0.109	0.2820

Table 5: Least square estimates of the coefficients in model (1).

Based on p-values in Table 5, the covariates log(Cancervolume), log(Prostateweight) are statistically significant while the covariates age, BPH, and capsular penetration are statistically insignificant. The coefficients associated with log(Cancer volume) and log(Prostate weight) are positive, as expected based on plots in Figure 5b and Figure 4b. However, the effect of age is not well captured by the model we have. According to the plot Figure 7a, older men are more likely to have high PSA and therefore are more prone to cancer.

According to the model we have, age is statistically insignificant.

In order to test for the significance of SVI and Gleason treatment effects in model (6), we use the general linear approach again. To test for SVI factor effects, the hypothesis test of interest is:

$$H_0 : \alpha_1 = \alpha_2 = 0 \ \ (\text{vs}) \ \ H_1 : \alpha_1 \neq 0 \text{ or } \alpha_2 \neq 0 \tag{8}$$

The full model to test this hypothesis is model (6) while the reduced model is obtained by setting α_1, α_2 to 0 in model (6).

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \times \frac{df_F}{SSE(F)} = \frac{46.908 - 42.649}{89 - 88} \times \frac{88}{42.649} = 8.788 \text{ where } F^* \overset{H_0}{\sim} F(1, 88)$$

Based on the F^* value, the p-value associated with testing the hypothesis (8) is 0.0039. Therefore, at 5% significance we reject the null hypothesis, i.e. the SVI factor effects are significant. In fact, this is what we observed in the interaction plot in Figure 9 where the estimated mean $\log(PSA)$ value of patients with no seminal vesicle invasion is lower than that of patients with seminal vesicle invasion consistently across all cancer grades.

To test for Gleason factor effects, the hypothesis test of interest is:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \ \ (\text{vs}) \ \ H_1 : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or } \beta_3 \neq 0 \tag{9}$$

The full model to test this hypothesis is the model (6) while the reduced model is obtained by setting $\beta_1 = \beta_2 = 0$ in the same model.

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \times \frac{df_F}{SSE(F)} = \frac{46.334 - 42.649}{90 - 88} \times \frac{88}{42.649} = 3.802 \text{ where } F^* \overset{H_0}{\sim} F(2, 88)$$

Based on the F^* value, the p-value associated with testing the hypothesis (9) is 0.0261. Therefore, at 5% significance, we fail to reject the null hypothesis i.e. the factor effects associated with Gleason are significant. This was observed in the interaction plot in Figure 9 as well. PSA levels of patients with higher grade cancer (i.e. higher Gleason score) is elevated. The higher the Gleason score, the higher is the blood PSA level.

Table 6 provides the interval estimates of pairwise comparison of interest. These intervals are obtained using the *Bonferroni procedure*. Therefore, the confidence coefficient is at least 0.95. Tukey procedure is not appropriate in the ANCOVA setting. As shown in Table 6, the number of comparisons of interest are four, which is close to the total number of treatment combinations. In such scenarios, the Bonferroni procedure is better than the Scheffe procedure. In fact, this can be checked by calculating the Scheffe and Bonferroni multipliers, which can be calculated to be 3.4027 and 1.007 respectively. Clearly, Bonferroni procedure will give us tighter confidence intervals.

Comparison	Confidence interval	Conclusion
$\alpha_1 - \alpha_2$	$(-1.0137, -0.4333)$	$\alpha_1 < \alpha_2$
$\beta_1 - \beta_2$	$(-0.3635, 0.0442)$	$\beta_1 = \beta_2$
$\beta_1 - \beta_3$	$(-0.9303, -0.3625)$	$\beta_1 < \beta_3$
$\beta_2 - \beta_3$	$(-0.7384, -0.2351)$	$\beta_2 < \beta_3$

Table 6: A confidence coefficient of at least 0.95 using Bonferroni procedure.

Based on the Table 5, $(-1.0137, -0.4333)$ is the interval estimate for $\alpha_1 - \alpha_2$. This means that the factor effects associated with $SVI = 1$ is more than that of $SVI = 0$. In other words, the effect of having seminal vesicle invasion on $\log(PSA)$ levels is more than that of not having seminal vesicle invasion. Based on Table 5, $\alpha_1 = -0.362$. Hence $\alpha_2 = 0.362$. Therefore, our final model (6) contributes 0.362 to the response every time the patient has seminal vesicle condition. The response is reduced by 0.362 every time the patient does not have SVI. This models what we observed in Figure 8e appropriately. In Figure 8e, we observe that patients without SVI tend to have a response value of 2 or above, while those without SVI appear to have their response values between -1 and 4. A good number of patients have response value less than 2.

Based on Table 5 :

$$\widehat{\beta}_1 = -0.269 \; ; \; \widehat{\beta}_2 = 0.109 \; ; \; \widehat{\beta}_3 = 0.269 - 0.109 = 0.16$$

The confidence interval for $\beta_1 - \beta_2$ in Table 6 suggests that $\beta_1 = \beta_2$ which is in line with what we observe in the interaction plot in Figure 9. However, the estimated coefficients, $\widehat{\beta}_1$ and $\widehat{\beta}_2$, are not quite close.

The confidence interval for $\beta_1 - \beta_3$ and $\beta_2 - \beta_3$ in Table Table 6 suggests that $\beta_1 < \beta_3$ and, $\beta_2 < \beta_3$ respectively. This again aligns with what we observed in the interaction plot Figure 9. Our estimated regression coefficients also follow $\widehat{\beta}_1 < \widehat{\beta}_3$ and $\widehat{\beta}_2 < \widehat{\beta}_3$ as desired. The effect of having a Gleason score of 8 on response is 0.051 more than that of having a Gleason score 7. Although this might appear as a small effect at first, we must note that the response is log transformed PSA values. Therefore, very small changes in $\log(PSA)$ can imply significant changes in PSA values, since most $\log(PSA)$ values in the data set are small.

5 Conclusion

Based on the data analysis so far, it is clear that people with SVI tend to have higher levels of PSA than the people without SVI. This implies that people with SVI are more likely to be diagnosed with prostate cancer than those without SVI. Patients diagnosed with low and medium grade cancer have similar PSA levels in blood. Therefore, one cannot rely on PSA levels as an indicator to distinguish low grade cancer from medium grade cancer. Patients with high grade cancer have higher PSA levels than those of low or medium grade cancer patients, implying that there is a correlation between grade of cancer and blood PSA levels.

Higher prostate weights contributes to higher PSA levels, which in turn imply an increased likelihood of getting diagnosed with prostate cancer. The model proposed does not capture the effect of age on PSA levels. Age, according to Table 5 is statistically insignificant. However, looking at the age of patients in the data set, we observe that most men are age 50 or above implying that older men are more likely to have higher PSA levels (i.e. more likely to be diagnosed with cancer). The current model can be modified by removing the age as a covariate and instead using it as a *blocking variable*.

Based on graph in Figure 6a, there appears to be a positive correlation between capsular penetration and, $log(PSA)$ which is not capture by the model proposed. This is probably because the number of data points with higher capsular penetration values is low. I believe that a square root transformation, instead of the log transform, can be tried. Finally, there seems to be no effect of BPH on patient PSA levels.

A major assumption throughout the report has been that the covariates and treatments do not interact. This condition does not appear to be satisfied in the case of capsular penetration and SVI-Gleason treatment combination. This can be seen in the plot in Figure 8e.

Finally, since this is an observational study, no cause-effect relationships can be inferred from the analysis so far. Further progress can be made by considering designs that let us infer cause effect relationships in observational studies.