Modeling Prostate-Specific Antigen Levels in Cancer Patients: An Empirical Approach to Building Multiple Linear Regression Models

Yang Wu

Math Department
Kenyon College

May 2021

Contents

1 Introduction

This report summarizes all of the modeling and analysis results associated with the study of the statistical association between prostate-specific antigen levels (PSA) and various prognostic clinical measurements in men with advanced prostate cancer— cancer volume, prostate weight, age, benign prostatic hyperplasia, seminal vesicle invasion, capsular penetration, and Gleason score. The purpose of this report is to document the best linear regression model obtainable from these clinical measurements and all corresponding inferences during the subsequent statistical analyses. We understand that a useful model is conducive to the success of your research team, and we are confident that the following results and recommendations address your need.

The remainder of this report is organized as follows. Section 2 describes the data in detail, including variable definition, descriptive statistics, and graphical exploration of the variables. Section 3 presents the models in their equation forms, details the specifics of their development, and assesses the model assumptions. Section 4 report the findings of the regression models and the robustness of these findings. Next, Section 5 tackles inferences concerning the regression parameters and constructs interval estimates for these model parameters. Lastly, Section 6 leverages the results of our analyses and concludes with our recommendations on the "best" model for your research team.

2 Data Preparation

The sample of data is provided by our client, the University Medical Center Urology Group. In this observational study, prostate-specific antigen levels (PSA) are recorded for a sample of 97 male patients with advanced prostate cancer. Also recorded are a myriad of other patient measurements. Table 1 provides the definitions of the variables and their units of measurement.

Table 1: Variable Definitions

Variable	**Variable Code**	**Variable Definition**
PSA level	PSA	Serum Prostate-specific antigen level (mg/ml)
Cancer volume	volume	Estimate of prostate cancer volume (cc)
Weight	weight	Prostate Weight (gm)
Age	age	Age of patient (years)
Benign prostatic hyperplasia	hyperplasia	Amount of benign prostatic hyperplasia (cm^2)
Seminal vesicle invasion	invasion	Presence or absence of seminal vesicle invasion: 1 if yes, 0 otherwise
Capsular penetration	capsular	Degree of capsular penetration (cm)
Gleason score	gleason	Pathologically determined grade of disease (summed scores were either 6, 7, or 8 with higher scores indicating worse prognosis)

2.1 Descriptive Statistics

Table 2: Summary Statistics

Variable	**Max**	**Mean**	**Median**	**Min**	**Pctl(25)**	**Pctl(75)**	**St.Dev.**
PSA	265.072	23.730	13.330	0.651	5.641	21.328	40.783
volume	45.604	6.999	4.263	0.259	1.665	8.415	7.881
weight	450.339	45.491	37.338	10.697	29.371	48.424	45.705
age	79	63.866	65	41	60	68	7.445
hyperplasia	10	2.535	1.3	0	0	4.8	3.031
invasion	1	0.216	0	0	0	0	0.414
capsular	18	2.245	0.4	0	0	3.3	3.783
gleason	8	6.876	7	6	6	7	0.740

In Table 2, the max and min columns provide us with a range of validity for our regression analysis. It can be inferred from the mean and median columns that some of the continuous variables— PSA, Volume, weight, hyperplasia, capsular— are positively skewed, which is a common phenomenon in data that arise in a healthcare setting. Before we turn to graphical explorations, we provide preliminary univariate analysis about each of the variables with regards to outliers, particularly those that are highly skewed.

2.2 Univariate Analysis

A procedure for detecting outliers is carried out for each of the continuous variables listed in Table 2. The results are tabulated below. We note that values above $(Q3 + 1.5 \cdot \text{IQR})$ or below $(Q1 - 1.5 \cdot \text{IQR})$ are considered as outliers; values above $(Q3 + 3 \cdot \text{IQR})$ or below $(Q1 - 3 \cdot \text{IQR})$ are considered as extreme outliers. Also note that extreme outliers is a subset of the set of outliers.

Table 3: Outlying Values for Continuous Predictors

Variable	**Outlier**	**Outlier ID**	**Extreme Outlier**	**Outlier ID**	**Total Outliers**
PSA	9	89, 90, 91, 92, 93, 94, 95, 96, 97	5	93, 94, 95, 96, 97	9
volume	7	75, 76, 55, 91, 86, 97, 94	2	97, 94	7
weight	7	78, 69, 77, 61, 89, 70, 32	3	89, 70, 32	7
age	5	19, 49, 94, 57	0	No case	5
hyperplasia	0	No case	0	No case	0
capsular	10	64, 94, 82, 86, 76, 95, 89, 79, 47, 97	3	79, 47, 97	10

As can be seen from Table 3, the response variable and few of the predictor variables have values that are far outlying. While it is useful to understand these characteristics of the sample, the multiple regression models that will be developed in this report make no assumptions about the distributions of these variables. The models do, however, assume that the errors are independent normal random variables with constant variance, which we will assess by analyzing the diagnostic plots. With this being said, we now proceed to graphical exploration of the variables.

2.3 Graphical Exploration

Figure 1: Scatter Plot and Correlation Matrix

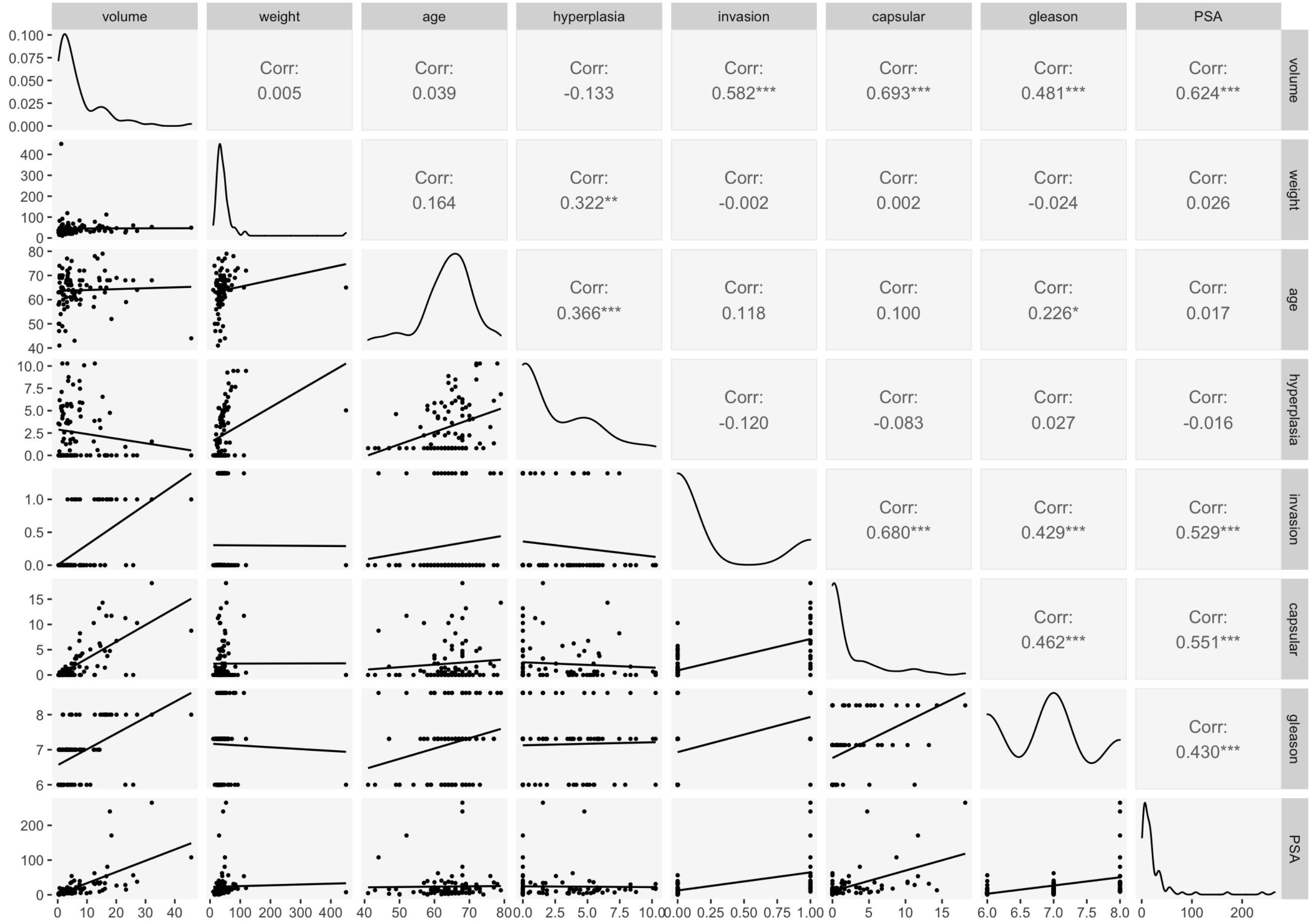


Figure 1 presents the scatter plot matrix and the correlation matrix of all the variables in the data set. The diagonal entries of Figure 1 contain the density plots of the variables. As can be seen, PSA, volume, weight, hyperplasia, and capsular are indeed right-skewed. As expected, gleason and invasion are multimodal in their distributions, as both are discrete. The age variable is less skewed, but the Anderson-Darling test ($p = 0.0003773$) and the Lilliefors test ($p = 0.003615$) both reject the null hypothesis that age is normally distributed in this sample (it need not be for the linear regression model to be a sound modeling technique). Overall, our graphical analyses confirm our conclusions about the skewness of the distributions of the variables from Table 2 and Table 3.

The entries above and below the diagonal of Figure 1 display a multitude of scatter plots and coefficients of simple correlation. The features of interest to our analysis are the last row and last column of the matrix, which show the scatter plots of the response variable against each of the predictor variables and their respective correlations. From the scatter plots, PSA appears to vary positively with volume, capsular, gleason, and invasion. The strengths of the bivariate relationships are mixed, with volume, capsular, and gleason appearing to be correlated with the response variable, PSA. Hyperplasia is negatively correlated with PSA, however, the strength of this bivariate relationship is weak. Some other noteworthy features are that, among all the predictor variables, 1) hyperplasia tends to be correlated with weights ($r = 0.322$) and age ($r = 0.366$), 2) invasion is highly correlated with capsular ($r = 0.680$) and gleason ($r = 0.429$), and 3) capsular is correlated with Gleason scores ($r = 0.462$). Thus, it behooves us to investigate the presence of interaction effects and multicollinearity when we fit the linear regression models.

2.4 Second Order and Interaction Terms

To investigate the functional forms in which the predictor variables should enter the regression model, we fit the second order models for each of the continuous predictor variables:

$$\ln(PSA)_i = \beta_0 + \beta_1 X_i + \beta_{11} X_i^2 + \varepsilon_i$$

where

- X_i are the orthogonal polynomials of the X predictor[1]

Then, we conduct the partial F-test to check if the first-order model is adequate. To avoid the risk of over-fitting, we use $\alpha = 0.01$; that is, we only include the higher order term if the evidence against the null hypothesis is strong. Based on these results,[2] we determine that the squared terms of the weight and volume variables should be included in the model but capsular, hyperplasia, and age should remain linear in their specifications. Thus, we also include the following predictors in the full data set.[3]

- Weight (Orthogonal Polynomials)^2
- Volume (Orthogonal Polynomials)^2

[1] Orthogonal polynomials are employed to mitigate the effects of serious multicollinearity even after centering the variables, which is identified by the Variance Inflation Factor (VIF) procedure.

[2] We use a hierarchical approach; that is, we fit the second-order model for each variable and test if the the first-order model is adequate. These results are included in the R script and can be provided upon request.

[3] We note that these orthogonal polynomials (orthogonal) are simply linear combinations of the original linear and squared variables (raw). The two quadratic models (orthogonal and raw) have the same fitted values and only differ in terms of parameterizations.

Next, on the basis of our graphical exploration (that is, this is not based on *a priori* knowledge), potential interaction effects will need to be investigated. Our approach for detecting interaction effect involves the residual plots and partial F tests. Once we obtain the candidate *additive* models, we will check for any model that contains both components of any of the cross-product terms listed below.

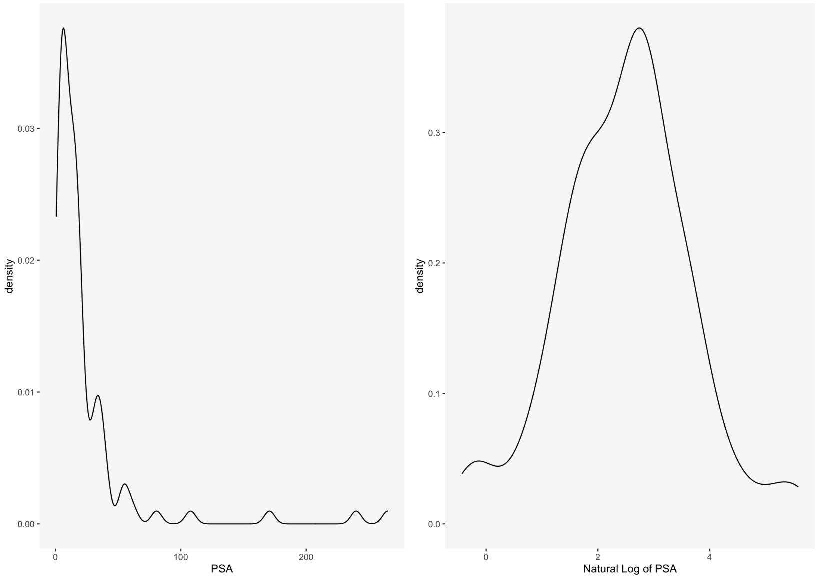
Volume × Invasion
Age × Hyperplasia
Gleason × Invasion
Volume × Capsular
Weight × Hyperplasia
Capsular × Gleason
Volume × Gleason
Capsular × Invasion

Our decision on whether to include an interaction term will the be informed by analyzing the plot(s) of model residuals against these cross-product terms as well as the partial F tests.

2.5 Transformation

Lastly, a Box-Cox procedure suggests $\lambda = 0$, which is by definition the natural logarithmic transformation of the response variable, PSA. Thus, in the remaining sections, we will use the log-transformed PSA as the response variable. Figure 2 below shows the density plots before and after the transformation. As can be seen, the distribution of the natural log of PSA is now fairly symmetric.

Figure 2: Transformation of the Response Variable, PSA



Before proceeding to develop the candidate "best" models, we note that the full data set now contains 18 variables (1 response, 9 potential predictor variables for the additive model, and 8 cross-product terms).

3 Models

3.1 Best Subset

Proceeding with our set of 9 potential predictor variables for the additive model, we employ a few methods for model development. The first is the Best Subset method, which suggests the following model:

$$\ln(PSA)_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_{55} X_{i5}^2 + \beta_6 X_{i6} + \beta_{66} X_{i6}^2 \tag{1}$$

where

- β_0, β_1, β_2, β_3, β_4, β_5, β_{55}, β_6 and β_{66} are the parameters
- $\ln(PSA)_i$ is the natural logarithm of prostate-specific antigen level
- X_{i1} is patient age (years)
- X_{i2} is the amount of benign prostatic hyperplasia (cm^2)
- X_{i3} equals 1 if seminal vesicle invasion is present, 0 otherwise
- X_{i4} is the Gleason scores
- X_{i5} and X_{i5}^2 are the linear and squared terms of cancer volume that are orthogonal to the constant polynomial of degree 0 as well as to each other
- X_{i6} and X_{i6}^2 are the linear and squared terms of prostate weight that are orthogonal to the constant polynomial of degree 0 as well as to each other

We develop the above model using the C_p criterion. In using this criterion, we seek to identify subsets of predictors for which 1) the C_p value is small and 2) the C_p value is near 9, the number of predictors. Model 1 is selected based on these three facts:

1. Among 32 candidate models, model 1 has the highest adjusted coefficient of multiple determination $R_{adj}^2 = 0.588$.

2. Among 32 candidate models, model 1 has the fifth smallest $C_p = 9.21$

3. Among 32 candidate models, model 1 falls closet to the line $C_p = 9$

For model 1, we begin our model refinement by including an interaction term between volume and weight. Then, using $\alpha = 0.01$, we conduct the partial F-test and conclude that the cross-product term should be dropped from the model. Next, we plot the residuals against the cross-products involving age, volume, hyperplasia, weight, gleason, and invasion. The residuals do not vary systematically with these cross-product terms. We further confirm this by conducting the partial F tests to see if the additive model is adequate. For all cross-product terms, we fail to reject the null hypothesis at the 1% significance level, concluding that there is no interaction effect present.

3.2 Backward Elimination & Forward Selection

The second model we will investigate is developed by stepwise procedures such as the forward selection method and the backward elimination method. The forward selection method involves adding candidate variables to the null model with just the intercept that leads to the best Mallow's C_p improvement. Conversely, the backward elimination method removes variables from the full model with all 9 predictors using Mallow's C_p as the criterion for comparing models. For our purposes, both procedures suggest the same model:

$$\ln(PSA)_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_{44} X_{i4}^2 + \beta_5 X_{i5} + \beta_{55} X_{i5}^2 \tag{2}$$

where

- β_0, β_1, β_2, β_3, β_4, β_{44}, β_5 and β_{55} are the parameters
- $\ln(PSA)_i$ is the natural logarithm of serum prostate-specific antigen level
- X_{i1} is the amount of benign prostatic hyperplasia (cm^2)
- X_{i2} equals 1 if seminal vesicle invasion is present, 0 otherwise
- X_{i3} is the Gleason scores
- X_{i4} and X_{i4}^2 are the linear and squared terms of cancer volume
- X_{i5} and X_{i5}^2 are the linear and squared terms of prostate weight

Model 2 differs from model 1 in that the age variable is dropped. Using similar procedures, we conclude that there is no interaction effect present among the variables included in this model.

3.3 Model Comparison

Table 4 summarizes the selection criteria for model 1 and model 2.

Table 4: Model Selection Criteria

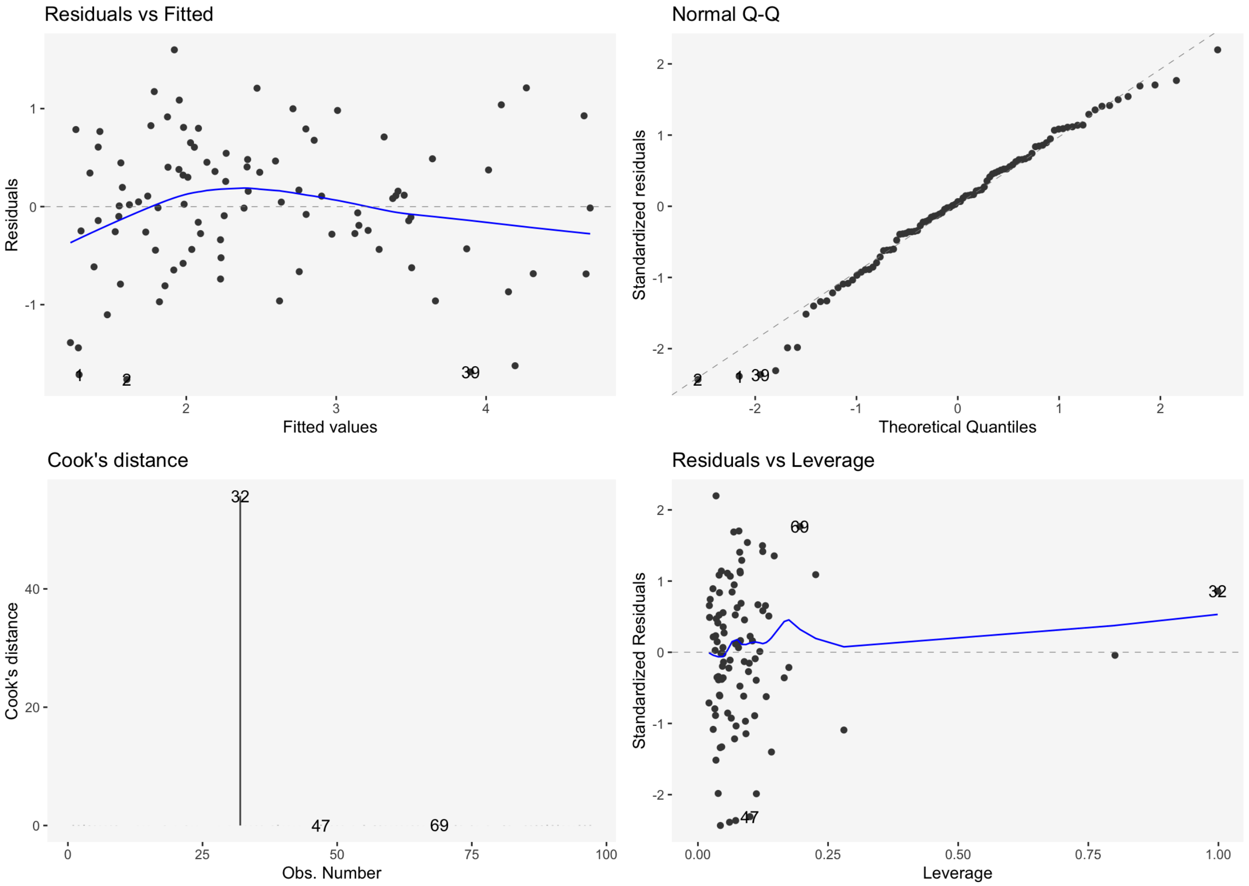
	p	R^2	R^2_{adj}	SSE$_p$	Mallow Cp	AIC$_p$	SBC$_p$	BIC$_p$
Model 1	9	0.6219	0.5876	48.3044	9.2120	-49.6273	-26.4549	253.3939
Model 2	8	0.6158	0.5855	49.0950	8.6559	-50.0525	-29.4548	250.394

In using these selection criteria, we wish to minimize SSE$_p$, Mallow's C_p, AIC$_p$, SBC$_p$, and BIC$_p$, and most of these criteria favor model 2. Compared to model 2, the strengths of model 1 lie in its slightly higher R^2_{adj} value and slightly lower SSE$_p$. In addition, Mallow's C_p for model 1 is closer to 9 than Mallow's C_p is to 8 for model 2, indicating that bias may be smaller for model 1 than for model 2. However, we argue that these differences may not be worth including an additional parameter in the model. To provide further statistical justification for our choice of "best" model, we test the null hypothesis that model 2 is adequate against the alternative hypothesis that the age variable should be included. Based the the partial F test ($p = 0.2333$), we fail to reject the null hypothesis at all significance levels, concluding that the contribution of the age variable is not above and beyond those of the predictors already in model 2. Based on these comparisons and more, we conclude that the more parsimonious model 2 may be desired. With these “best” models, we now turn to model diagnostics.

3.4 Graphical Assessment of Model Assumptions

To examine the aptness of model 1, we present the following panel of plots.

Figure 3: Diagnostic Plots for Model 1



The first panel in Figure 3 plots the residuals against the fitted values. As can be seen, the residuals appear to vary randomly with the fitted values. The regression function is appropriate and there is no concern for non-constancy of error variance. The second panel of Figure 3 shows the normal quantile plot, which indicates that the residuals are reasonably normal. The last two panels provide graphical analysis on two additional measures of model aptness, Cook’s distance and leverage. Cook’s distance considers the influence of the i^{th} case on all 97 fitted values. Crudely speaking, the leverage value of the i^{th} case measures how distant it is from the center of all predictor observations, called the centroid. We care about these measures since they communicate to us whether there are *potentially* influential points that may heavily influence our model results. Evidently, the Cook’s distance plot in Figure 3 flags three such cases, one of which (case 32) is rather extreme. Therefore, it behooves us to explore the influence these values have on our estimated regression coefficients, which we will tackle in the results section. The residual plot against the leverage values shows no cases that are outlying with respect to their response values since there are no standardized residuals with absolute value of three or more. Again, the three cases 32, 47, and 69 are identified as outlying with respect to their predictor values.

Figure 4: Diagnostic Plots for Model 2

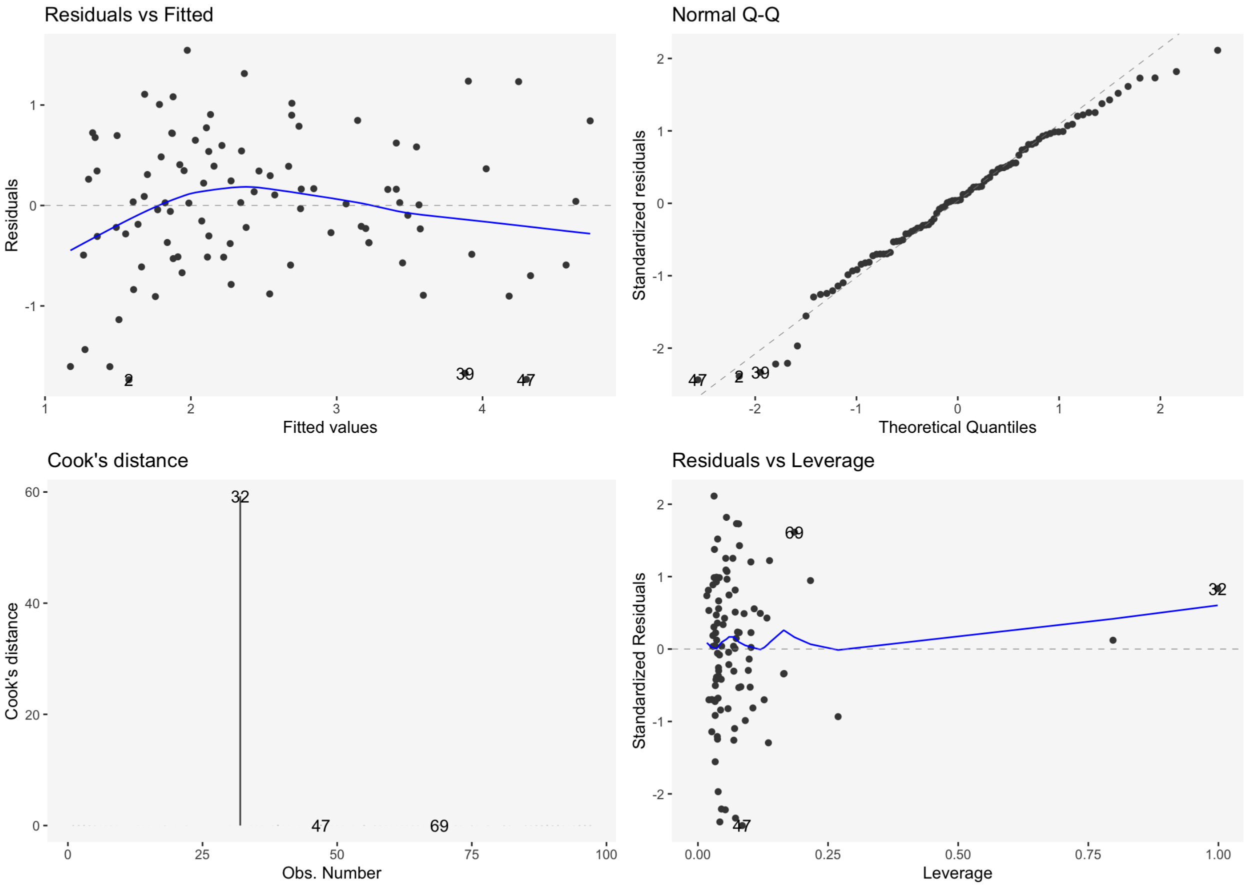


Figure 4 presents the diagnostic plots for model 2. Again, the first two panels show that there is no concern for heteroscedasticity and there is no substantial departure from normality. The regression function appears to be appropriate as the residuals do not vary systematically with the fitted values. Similar to Model 1, there is evidence that the same three cases may be unduly influencing our model and there is no presence of outliers with respect to the response variable.

3.5 Assessment of Model Assumptions Using Formal Tests

Table 5 summarizes the formal test results for model 1. As can be seen, the normality of the error terms is supported by the formal tests. For constancy of error term variance, the Breusch-Pagan test and the Brown-Forsythe test conclude that the assumptions of the normal error model is satisfied.

Table 5: Tests For Model 1

Test	**Hypotheses**	α	**Test Statistic**	**Criteria**	**Conclusion**
Correlation Test	H_0 : Normality H_a : Non-normality	0.1	Pearson's r = 0.9922	critical-value = 0.989	Conclude H_0
Anderson-Darling	H_0 : Normality H_a : Non-normality	0.1	A = 0.3806	p-value = 0.3958	Conclude H_0
Lilliefors	H_0 : Normality H_a : Non-normality	0.1	D = 0.0675	p-value = 0.3407	Conclude H_0
Breusch-Pagan	H_0 : Constant Error Variance H_a : Non-constant Error Variance	0.1	BP = 4.9187	p-value = 0.7662	Conclude H_0
Brown-Forsythe	H_0 : Constant Error Variance H_a : Non-constant Error Variance	0.1	BF = 0.0002	p-value = 0.9901	Conclude H_0
We choose $\alpha = 0.1$ to be more conservative and not hasten to conclude H_0 in support of the model assumptions. For the Brown-Forsythe test, we choose Volume ≥ -0.035 (the median) as the threshold for grouping.					

Table 6 summarizes the formal test results for model 2. Similar to model 1, the assumptions are well satisfied.

Table 6: Tests For Model 2

Test	**Hypotheses**	α	**Test Statistic**	**Criteria**	**Conclusion**
Correlation Test	H_0 : Normality H_a : Non-normality	0.1	Pearson's r = 0.9919	critical-value = 0.989	Conclude H_0
Anderson-Darling	H_0 : Normality H_a : Non-normality	0.1	A = 0.3576	p-value = 0.4477	Conclude H_0
Lilliefors	H_0 : Normality H_a : Non-normality	0.1	D = 0.0497	p-value = 0.8045	Conclude H_0
Breusch-Pagan	H_0 : Constant Error Variance H_a : Non-constant Error Variance	0.1	BP = 5.8639	p-value = 0.5557	Conclude H_0
Brown-Forsythe	H_0 : Constant Error Variance H_a : Non-constant Error Variance	0.1	BF = 0.0342	p-value = 0.8536	Conclude H_0
For the Brown-Forsythe test, we choose Volume ≥ -0.035 (the median) as the threshold for grouping.					

4 Results

4.1 Regression Output

Table 7 displays our full regression output. For both models, we find evidence that, ceteris paribus, cancer volume is positively associated with the response variable, log PSA. Because of our orthogonal transformations, both the linear and the second order terms for the volume variable lose their usual interpretability. However, we may still infer from the signs of the estimated coefficients that cancer volume is positively and significantly associated with PSA levels and at a diminishing rate, as indicated by the negative quadratic effect coefficient on the second order term.

Since model 2 is our preferred model, we look to point out the other notably significant variables— invasion, gleason, and weight squared. From Table 1, we see that the invasion variable is a qualitative predictor; it captures the differential effect of the presence or absence of seminal vesicle invasion. That is, the estimated coefficient indicates how much higher the response function is in the presence of seminal vesicle invasion than the the response function in its absence. Because of the log transformation on the response variable, an interpretation of this coefficient is not straigtforward. Using the formula from Halvorsen and Raymond Palmquist (1980),[4] we provide the following interpretation of the estimated coefficient on invasion— the percent change in PSA levels associated with the presence (that is, switching the dummy variable from 0 to 1) of seminal vesicle invasion is 83.86% for model 1 and 81.27% for model 2.

For Gleason scores, we note again that this is a qualitative predictor variable and the classes employed are elements of the set $\{6, 7, 8\}$. It is important to understand that the allocation of code implies that the mean response changes by the same amount when going from one score to another. This nature of the effect of Gleason scores on the response is the result of code allocation, which assigns equal distances between the three classes of scores. That is, all else equal:

$$E\{\ln(PSA)|\text{A Gleason score of 8}\} - E\{\ln(PSA)|\text{A Gleason score of 7}\} = \beta_3$$
$$E\{\ln(PSA)|\text{A Gleason score of 7}\} - E\{\ln(PSA)|\text{A Gleason score of 6}\} = \beta_3$$

In Table 7, an unbiased estimator for β_3 is 0.302 for model 1 and 0.281 for model 2, both of which are statistically significant at the 5% significance level.

Another important finding from Table 7 is that the quadratic effect coefficient for the prostate weight variable is statistically significant while the linear effect coefficient is statistically not different from zero. This finding is consistent across both models. We include the linear term since it is viewed as providing basic information about the shape of the response function while the quadratic term is viewed as providing refinements in the specification of the shape of the response function. Again, due to the orthogonal transformation, the estimated coefficient cannot be interpreted directly. However, we again emphasize that the negative quadratic effect coefficient indicates some important feature of the response function and thus the

[4] https://fvela.files.wordpress.com/2010/11/dummyinterpretation.pdf

Table 7: Regression Output

	Response variable:	
	$\ln(PSA)$	
	(Model 2)	(Model 1)
hyperplasia	0.047	0.055^{*}
	(0.032)	(0.033)
invasion	0.595^{**}	0.609^{**}
	(0.234)	(0.234)
gleason	0.281^{**}	0.302^{**}
	(0.123)	(0.124)
weight	0.974	1.081
	(0.805)	(0.808)
weight squared	-1.641^{*}	-1.807^{*}
	(0.917)	(0.925)
volume	4.879^{***}	4.811^{***}
	(1.002)	(1.001)
volume squared	-1.421^{*}	-1.727^{**}
	(0.775)	(0.814)
age		-0.014
		(0.012)
Constant	0.296	1.057
	(0.842)	(1.052)
Observations	97	97
R^2	0.616	0.622
Adjusted R^2	0.586	0.588
Residual Std. Error	0.743 (df = 89)	0.741 (df = 88)
F Statistic	20.374^{***} (df = 7; 89)	18.096^{***} (df = 8; 88)
Note:		*p<0.1; **p<0.05; ***p<0.01

nature of the relationship between prostate weight and log PSA levels.

Lastly, although age is statistically not different from zero in model 1, its inclusion results in the hyperplasia variable becoming significant at the 10% significance level. It is mildly surprising that the selection methods identify the hyperplasia variable as a predictor for model 2 despite the apparent lack of evidence of its usefulness. Nevertheless, an interpretation for the estimated regression coefficient on hyperplasia in model 1 is as follows— a one unit change in the amount of hyperplasia, approximately speaking, is associated with a 5.5% change in PSA levels. In model 2, the effect size is smaller for the hyperplasia variable and the coefficient is statistically insignificant.

4.2 Sensitivity Analysis

The plots in figure 3 and 4 reveal three outlying cases with respect to their predictor values (henceforth, cases 32, 47, and 69). To the extent that these cases represent legitimate data, we may wish to investigate further the manner in which they influence our regression results.[5] We begin our analysis by assessing whether there are other high leverage cases. Our identification strategy involves the diagonal entries of the hat matrix, h_{ii}, which is a measure of the distance between the predictor values of the i^{th} case and the means of the predictor values for all 97 cases. If the i^{th} case is outlying with respect to its predictor observations and has a large leverage value, it exerts leverage in determining the fitted value $\hat{Y}_i$. Specifically, we consider a leverage value large if it is more than twice as large as the mean leverage value:

$$2\bar{h} = 2[\frac{\sum_i^{97} h_{ii}}{97}] = 2[\frac{8}{97}]$$

Table 8 tabulates the identified outlying cases, including their leverage value and case index. Also included in Table 8 are two additional measures— Cook's distance and DFFITS. The DFFITS measures the distance between the fitted $\hat{Y}_i$ for the i^{th} case when all 97 cases are used and the predicted $\hat{Y}_{i(i)}$ for the i^{th} case when the i^{th} case is omitted from the sample. Together, Cook's distance and DFFITS allow us to ascertain whether an outlying case is influential. For each Cook's distance, we relate it to the $F(8;89)$ distribution and obtain a percentile; we consider an outlying case influential if its percentile value is near 50 percent or more. In addition, a DFFITS value exceeding one in absolute values is also considered influential.

As can be seen in Table 8, there are six cases whose leverage values are considered large relative to the mean leverage value. Notably, case 32 and 69 from Figure 3 and 4 are listed as outlying with respect to their predictor values. Case 47 is missing from the list since its leverage value, 0.0852, is less than twice the mean leverage value. It's Cook's distance, however, is large enough to warrant a flag in the diagnostic plots; this is due to the fact that case 47 has a relatively large residual value, -1.7313, which is factored into the calculation of Cook's distance. What is important is that none of the outlying cases appear to be influential, save case 32. For cases 69, 70, 89, 91, and 94, the percentile values are well below 50% and the DFFITS values

[5] We report the results of the sensitivity analysis for our choice of "Best" model. The same procedure is carried out for model 1 and can be provided upon request.

Table 8: Influential Cases

Case ID	Leverage	Cook's Distance	Percentile Value	DFFITS
32	0.9985	59.2687	100%	21.7374
69	0.1854	0.0741	0.0283%	0.7770
70	0.1655	0.0028	0.00000007%	-0.1497
89	0.2693	0.0402	0.0028%	-0.5664
91	0.2162	0.0308	0.0010%	0.4962
94	0.7975	0.0073	0.000003%	0.2399

do not exceed one in absolute values. To investigate further, we employ the DFBETAS, which measures the difference between the estimated coefficients $\hat{\beta}_k (k = 0, 1, ..., 55)$ based on all 97 cases and those obtained when case 32 is dropped from the sample. The sign of the DFBETAS value indicates whether the inclusion of case 32 increases or decreases the estimated regression coefficient while its absolute magnitude provides information about the size of the difference relative to the standard error of that estimated coefficient. The guideline for identifying influential cases using the DFBETAS measure is to consider a case whose absolute value of DFBETAS exceeds one influential.

Table 9: Influence On Regression Coefficeients

Case ID	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_{44}$	$\hat{\beta}_5$	$\hat{\beta}_{55}$
32	0.2056	0.1480	0.0222	-0.0232	18.0871	7.3958	0.1588	-0.0290

From Table 9, we note that $\hat{\beta}_4$ and $\hat{\beta}_{44}$ (weight and weight squared) are severely impacted. The signs suggest that the inclusion of case 32 likely leads to over-estimations of the parameters on weight and weight squared. Furthermore, we report that no other case in the sample has any DFBETAS value that exceeds one in absolute value. In other words, we have strong evidence that case 32 may be an anomaly. We also present graphical representations of DFBETAS values for all variables in model 2 in Figure 5 and 6. Note that the threshold indicated by the red lines may be adjusted to allow for different tolerance levels.[6] This influential point will be brought along in our analysis as we lack any justification for its exclusion. We nevertheless emphasize that there is now a caveat to interpreting the results of our analysis, particularly those pertaining to prostate weight.

[6] The program for these plots is included in the R script accompanying this report. Stakeholders may freely adjust these threshold for further analysis.

Figure 5: DFBETAS Diagnostics

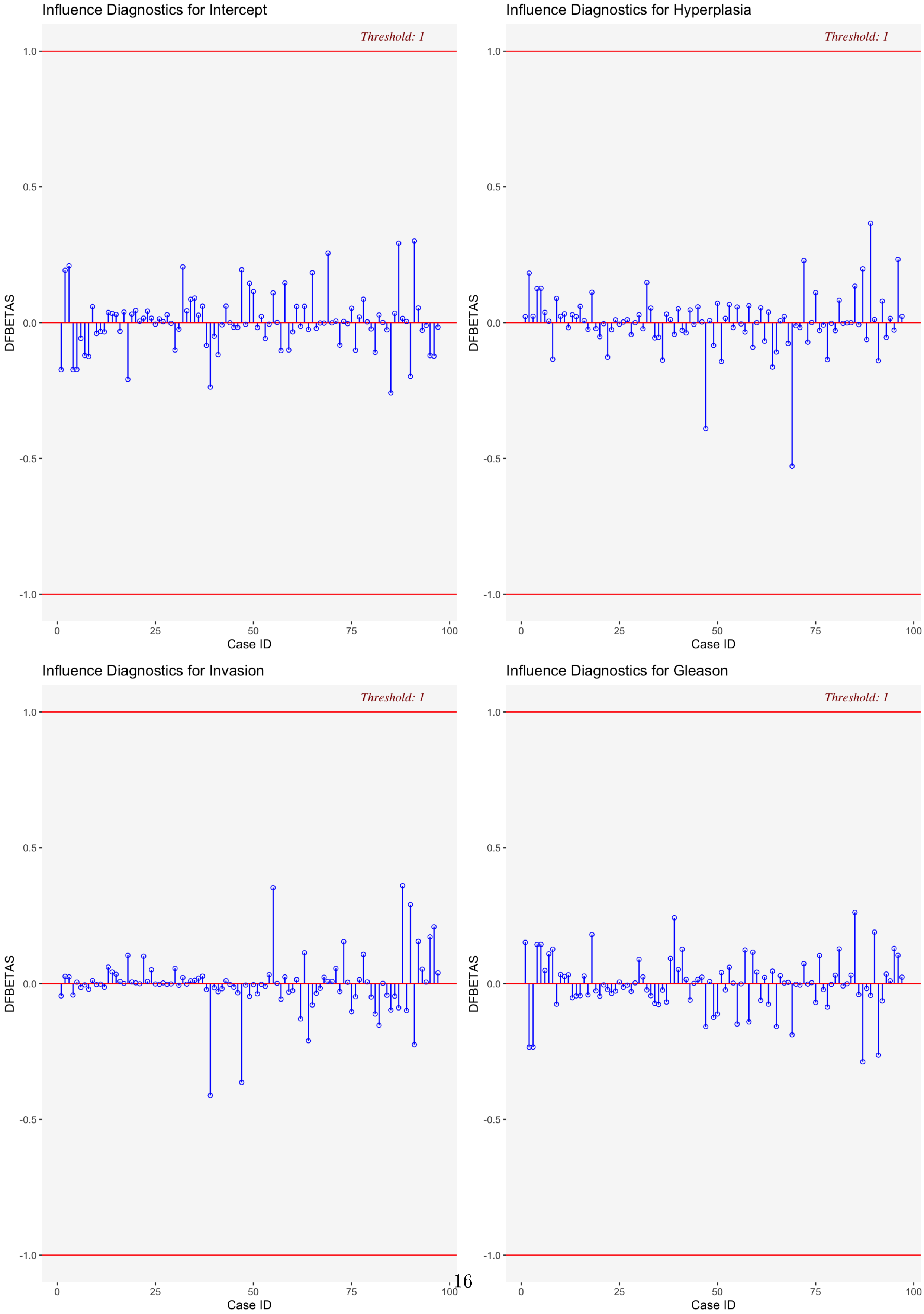
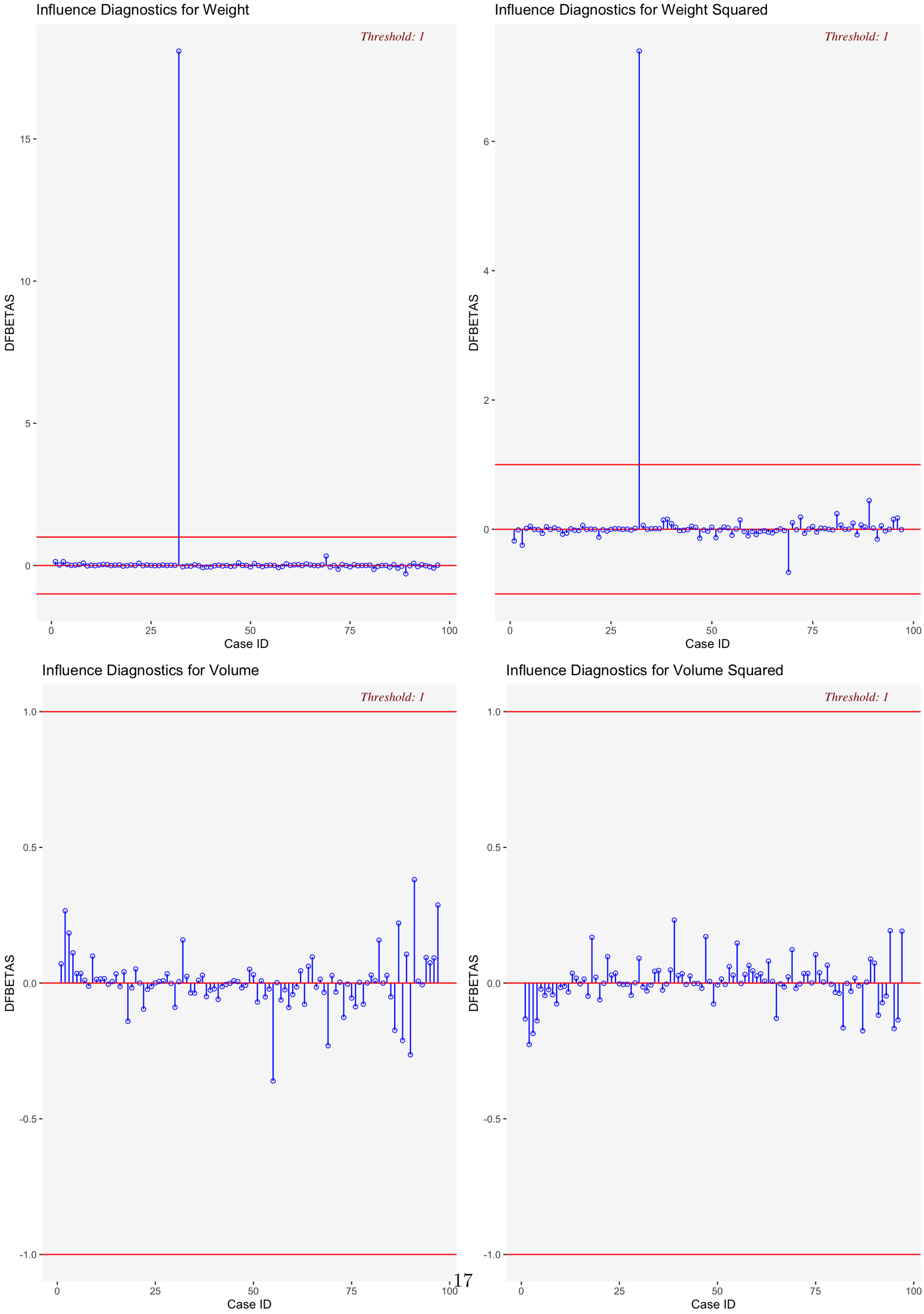


Figure 6: DFBETAS Diagnostics (Continue)



Influence Diagnostics for Weight

Influence Diagnostics for Weight Squared

Influence Diagnostics for Volume

Influence Diagnostics for Volume Squared

5 Interval Estimation

5.1 Model 1: Confidence Interval for Estimated Coefficients

For model 1, we report the Bonferroni 95% family confidence intervals for all statistically significant parameters in the Table 7.

Table 10: Bonferroni 95% Family Confidence Intervals Model 1

			B Multiple = 2.6329			
Predictor	Parameter	Point Estimator	Standard Error	Margin of Error	Lower	Upper
hyperplasia	β_2	0.0553	0.0327	0.0861	-0.0308	0.1414
invasion	β_3	0.6090	0.2341	0.6163	-0.0074	1.2253
gleason	β_4	0.3015	0.1237	0.3257	-0.0242	0.6273
volume	β_5	4.8113	1.001	2.6366	2.1748	7.4479
volume squared	β_{55}	-1.7273	0.8140	2.1432	-3.8706	0.4160
weight squared	β_{66}	-1.8067	0.9246	2.4345	-4.2411	0.6278

These intervals use a 0.95 family confidence coefficient, which means that if repeated samples are selected and interval estimates for these parameters are calculated for each sample, 95% of the independent samples would lead to all correct interval estimates. The error rate is 0.05, which means that there is only a 5% probability that one or all of the interval estimates would be incorrect. We note that many of these intervals contain 0, which may not be useful for your purposes, but the family confidence coefficient may be adjusted accordingly using the programs we provide in the R script. As such, family confidence intervals involving subsets of predictors that are of interest to your research team can be constructed. Table 11 reports the Bonferroni 95% family confidence intervals for all statistically significant parameters in the model 2. The same interpretations apply for these intervals.

Table 11: Bonferroni 95% Family Confidence Intervals Model 2

			B Multiple = 2.6329			
Predictor	Parameter	Point Estimator	Standard Error	Margin of Error	Lower	Upper
invasion	β_2	0.5948	0.2344	0.6169	-0.0221	1.2117
gleason	β_3	0.2812	0.1229	0.3234	-0.0422	0.6046
volume	β_4	4.8790	1.0023	2.6382	2.2408	7.5172
volume squared	β_{44}	-1.7273	0.8140	2.1432	-3.8706	0.4160
weight squared	β_{55}	-1.4212	0.7750	2.0399	-3.4611	0.6187

As can be seen, the only interval estimate that does not contain zero is that for the volume variable. With family confidence coefficient 0.95, we estimate that the true parameter on volume is between 2.2408 and 7.5172. Similar to model 1, the confidence intervals reported may be too wide to be useful. To address these concerns, we have two recommendations: 1) the confidence coefficient may be adjusted to allow for tighter interval estimates, allowing the family error rate to be somewhat higher in exchange for more useful inferences and 2) joint intervals can be constructed for smaller subsets of predictors to provide inferences on a few selected parameters that are most important to your team's research.

6 Conclusion

The evidence in this report is in favor of model 2. In section 3.3, we find that model 2 beats model 1 in 4 of the 7 selection criteria, and the strengths of model 1 lack extremities to outweigh the extra variable and strengths of model 2. These results, coupled with the a partial F test for the inclusion of the additional variable suggested by model 1, strengthen our argument for model 2 significantly. Sections 3.4 and 3.5 indicate that there are no issues with non-constancy of error variance, normality of error terms, or model aptness with either model. We report these results to further argue for model 2's advantages; with no violation of model assumptions, a simpler model is more ideal. In section 4, our regression results show that all but hyperplasia and weight are significant predictor variables in model 2, and the age variable added by model 1 is insignificant. All of these findings culminate to suggest that model 2 is our choice of "best" predictive model.

We address and satisfy a significant number of typically problematic issues in this paper, but there are a few that remain. In section 4.2, we find that case 32 in the sample is influential; however, without the expertise of the University Medical Center to determine the true nature of this observation, it is included in our model and may result in slightly more inefficient parameter estimation and interval estimates. By assessing the model assumptions in sections 3.4 and 3.5, we are able to illustrate that our models have no intrinsic issues. There is always the possibility of omitted variable bias, but with our methods of model construction and considerations of interaction and higher-order terms, such biases may only result from variables beyond scope of the data set provided to us. Therefore, we argue that there are very few analytical problems inherent in our methodology.

To conclude, we have sufficient evidence that model 2 is the "best" linear regression model obtainable from the potential predictor variables in the data set, and we suggest it for use among your medical research team members. We hope you consult KCKC Consulting for any future statistical needs.