Predicting Prostate Specific Antigen Levels from Prognostic Clinical Measurements in Men with Advanced Prostate Cancer.

Lindsey Hornberger

Department of Biostatistics
University of Kansas, USA
July 30, 2023

Title
To determine if there is an association between prostate-specific antigen (PSA level) and prognostic clinical measurements in men with advanced prostate cancer.

Abstract
Prostate cancer is the sixth leading cause of cancer death in men. The prostate specific antigen test is one tool that can be used to predict a patient's prostate cancer risk by detecting the level of PSA in the blood. The objective of this study is to determine if there is an association between PSA (the response variable) and seven different prognostic clinic measurements in men with advanced prostate cancer (the predictor variables). In this study we used linear regression models to determine if an association between variables was present. We tested for potential outliers and removed 17 outliers from the dataset. The best linear regression model indicates there is a correlation between PSA and cancer volume (CV), seminal vesicle invasion (SVI), and Gleason score (GS). These four clinical measurements have a significant correlation to PSA at the 95% level.

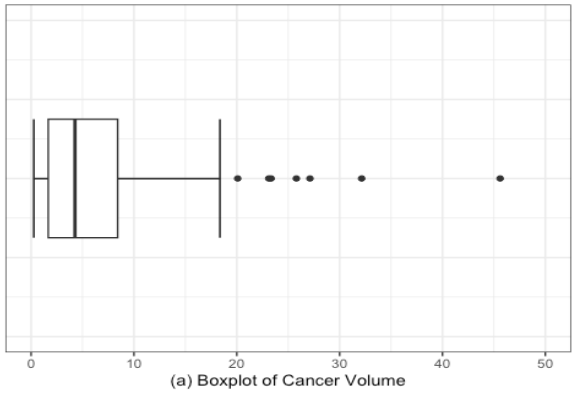
Introduction
The prostate specific antigen is a protein detected in the PSA test is produced by both benign and malignant prostate cells. The results of this test are typically reported in milligrams/milliliter (mg/ml) or nanograms/milliliter (mg/ml). An elevated blood PSA level often indicates the presence of prostate cancer in a patient. An elevated PSA level is defined4 as being above 4.0 ng/ml.[1] While high PSA levels can indicate the presence of prostate cancer, it is not always the case. Some patients with low PSA levels do have prostate cancer and some with high PSA level do not have prostate cancer. While PSA level is an indicator of the presence of prostate cancer, it is not completely accurate and serves as a screening tool. [3] The PSA was approved by the Food and Drug Administration (FDA) in 1986 to monitor the advancing progression of patients who had already been diagnosed with prostate cancer. [2] Shortly after this in 1994, the FDA approved and recommended that the PSA test should be used in addition to a digital rectal exam (DRE) to better diagnose and detect prostate cancer in men 50 years and older. [2] However, many medical professionals encourage men younger than 50 and without prostate cancer to get yearly screenings. However, there are other factors that can increase PSA levels besides prostate cancer. Benign prostate conditions can elevate one's PSA level. The most common of these are prostatitis (inflammation of the prostate) and benign prostatic hyperplasia (BPH) (enlargement of the prostate). [3] There is no evidence to suggest that either one of these benign conditions can lead to the development of prostate cancer, however, a patient can be diagnosed with one of the conditions and still develop prostate cancer.

Materials and Methods
Data Sources
The dataset was obtained online. The data was collected by a university medical center urology group. The data was collected from 97 men who were about to undergo radical prostectomies. Each patient has an associate identification number **(idnum)** and each patient has eight other variables. The response variable is the **prostate-specific antigen (PSA level)** which is a

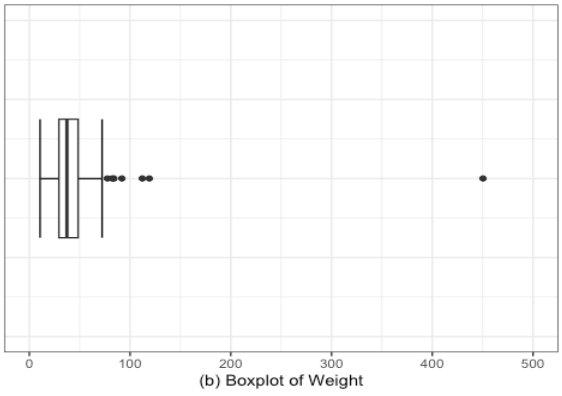
measurement of the serum prostate-specific antigen level measured in mg/ml. The variables in the dataset include **Cancer Volume (CV)** which ranges from 0.2592 - 45.6042 cc. **Prostate Weight (PW)** ranging from 10.67- 450.339 grams, **Patient Age (Age)** ranging from 41-79 years, **Benign Prostatic Hyperplasia (BPH)** ranging from 0-10.2779 cm^2, the presence or absence for **Seminal Vesicle Invasion (SVI)** 1 if yes and 0 if otherwise, the degree **of Capsular Penetration (CP)** with a range of 0- 18.1741cm, and the **Gleason Score (GS)** which is a pathologically determined grade of disease using total score of two patterns with a range of 6-8 (with higher scores indicating a worse prognosis).

Statistical Analysis

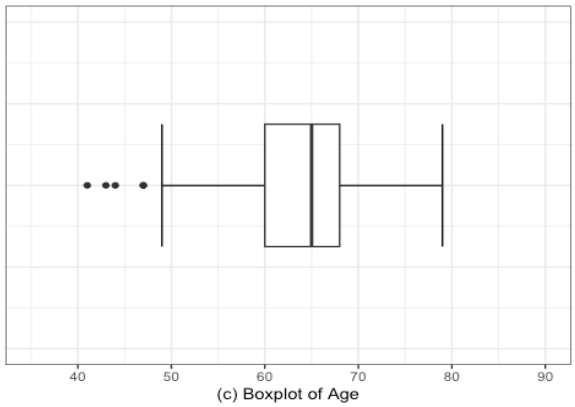
The data was available in .xlsx (excel) format. The data analysis is done using the statistical software R version 4.3.1 (2023-06-16). This project focuses mainly on multiple linear regression. All of the predictor variables are explored individually. In this dataset the sample size is 97 and there are no missing values. The smaller sample size could assume less predictability and a larger sampling variability.



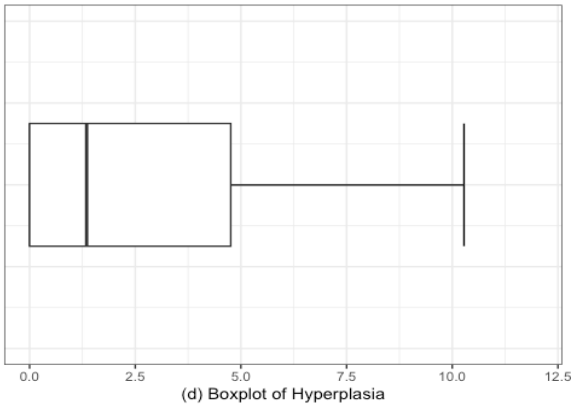
(a) Boxplot of Cancer Volume



(b) Boxplot of Weight



(c) Boxplot of Age



(d) Boxplot of Hyperplasia

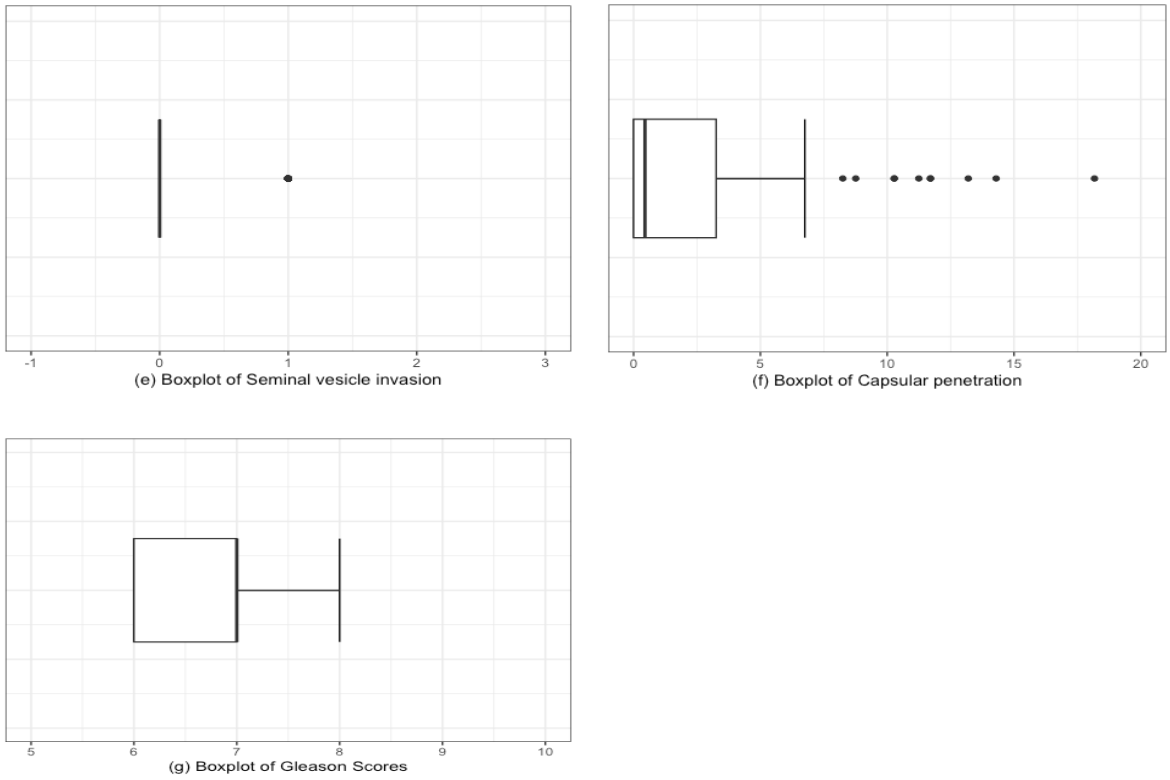


Figure 1: Analysis of the potential predictors

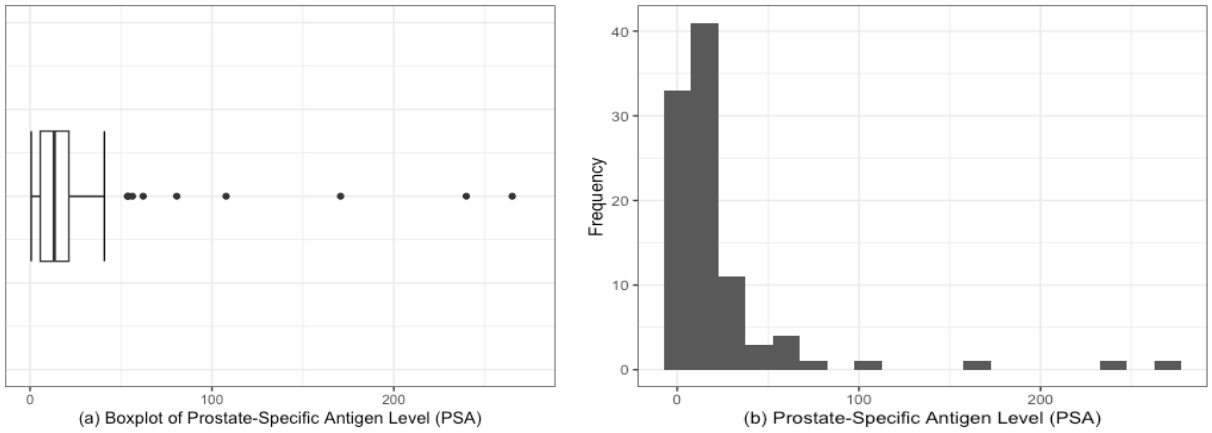
Table 1: Basic Statistics for Dataset Variables

	PSA	CV	PW	Age	BPH	SVI	CP	GS
min	0.651	0.2592	10.697	41.0	0	0	0	6
mean	23.73013	6.998682	45.49136	63.86598	2.534725	0.2164948	2.245367	6.876289
max	265.072	45.6042	450.339	79.0	10.2779	1	18.1741	8
std deviation	40.78292	7.880869	45.70505	7.445117	3.031176	0.4139949	3.783329	0.7396189
variance	1663.247	62.10809	2088.952	55.42977	9.188026	0.1713918	14.31358	0.5470361
coefficient of variation	1.718613	1.12605	1.004697	0.1165741	1.19586	1.912262	1.684949	0.1075608

Analysis of Prostate-Specific Antigen Level (PSA)

Preliminary data analysis shows that the response variable (PSA) is positively skewed. This is illustrated by the longer whisker being on the right of the boxplot in Figure 2a. Also from this data, we see that there are potentially some outliers on the far right with significantly higher PSA levels than most. The basic statistics on PSA data can also be referred from Table 1.

Figure 2: Distribution of Prostate-Specific Antigen Level (PSA)



Effect of Cancer Volume (CV) on Prostate-Specific Antigen (PSA)

Figure 3 illustrates the scatterplot computed by the linear regression model fitted to the data for Cancer Volume (CV) and Prostate-Specific Antigen Level (PSA). This illustration is conducted using the entire dataset for the purpose of preliminary investigation. The resultant predictive model can be written as:

$$Y_i = 1.125 + 3.229X_i \qquad (1)$$

where Y_i represents the prostate-specific antigen and X_i represents the cancer volume. The model states that for every one unit (cc) increase in cancer volume, the prostate-specific antigen level (PSA) increases by 3.228 mg/ml.

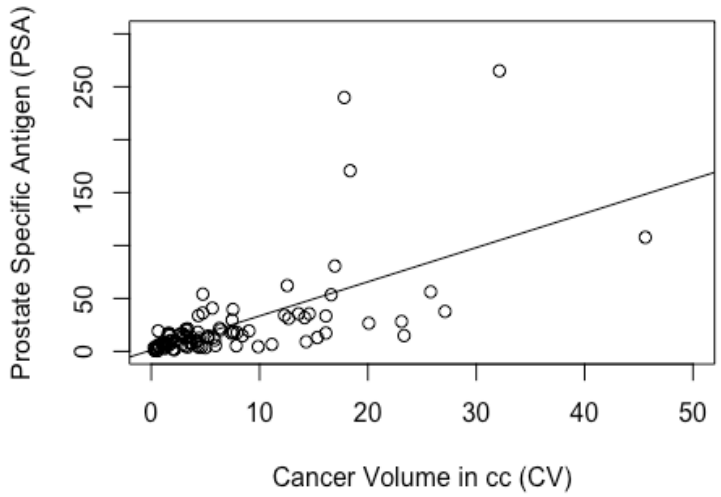


Figure 3: Prediction of PSA (mg/ml) with reference to CV (cc)

Effect of Prostate Weight (PW) on Prostate-Specific Antigen (PSA)

Figure 4 illustrates the scatterplot computed by the linear regression model fitted to the data for Prostate Weight (PW) and Prostate-Specific Antigen Level (PSA). This illustration is conducted using the entire dataset for the purpose of preliminary investigation. The resultant predictive model can be written as:

$$Y_i = 22.666 + 0.023X_i \qquad (2)$$

where Y_i represents the prostate-specific antigen and X_i represents the prostate weight. The model states that for every one unit (gram) increase in prostate weight, the prostate-specific antigen level (PSA) increases by 0.023mg/ml.

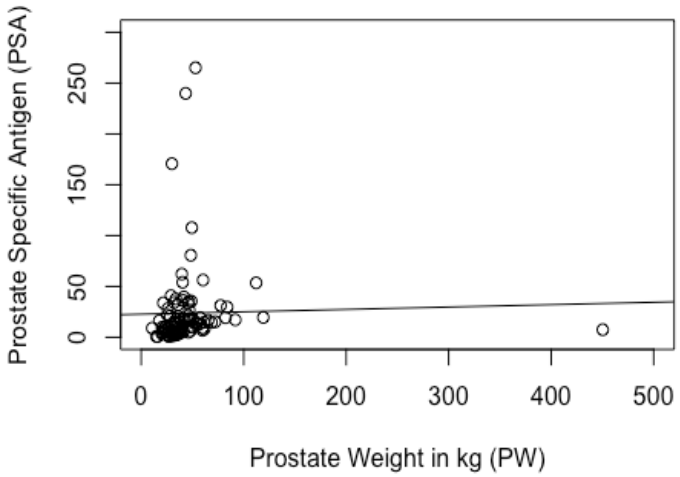


Figure 4: Prediction of PSA (mg/ml) with reference to PW in kg

Effect of Patient Age (PA) on Prostate-Specific Antigen (PSA)

Figure 5 illustrates the scatterplot computed by the linear regression model fitted to the data for Patient Age (PA) and Prostate-Specific Antigen Level (PSA). This illustration is conducted using the entire dataset for the purpose of preliminary investigation. The resultant predictive model can be written as:

$$Y_i = 17.713 + 0.0942X_i \qquad (3)$$

where Y_i represents the prostate-specific antigen and X_i represents the patient age. The model states that for every one unit (year) increase in patient age, the prostate-specific antigen level (PSA) increases by 0.0942mg/ml.

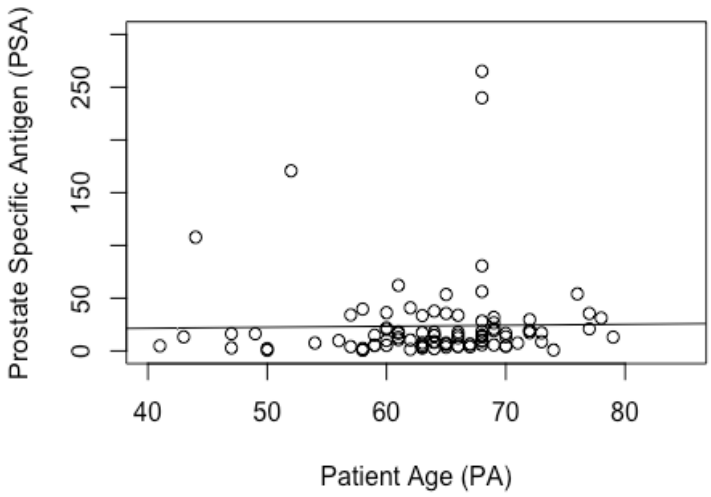


Figure 5: Prediction of PSA (mg/ml) with reference to PA in years.

Effect of Benign Prostatic Hyperplasia (BHP) on Prostate-Specific Antigen (PSA)

Figure 6 illustrates the scatterplot computed by the linear regression model fitted to the data for Benign Prostatic Hyperplasia (BHP) and Prostate-Specific Antigen Level (PSA). This illustration is conducted using the entire dataset for the purpose of preliminary investigation. The resultant predictive model can be written as:

$$Y_i = 24.292 - 0.222X_i \qquad (4)$$

where Y_i represents the prostate-specific antigen and X_i represents the benign prostatic hyperplasia. The model states that for every one unit (year) increase in patient age, the prostate-specific antigen level (PSA) decreases by 0.222mg/ml.

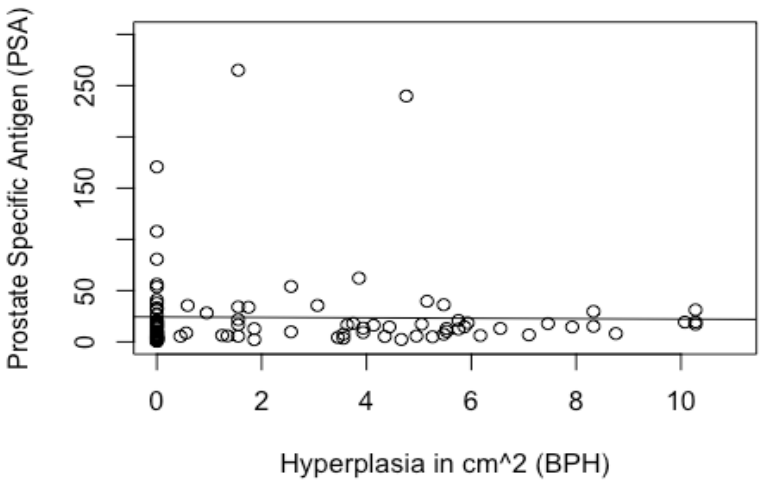


Figure 6: Prediction of PSA (mg/ml) with reference to BPH in cm^2.

Effect of Seminal vesicle invasion (SVI) on Prostate-Specific Antigen (PSA)

Figure 7 illustrates the scatterplot computed by the linear regression model fitted to the data for Seminal vesicle invasion (SVI) and Prostate-Specific Antigen Level (PSA). This illustration is conducted using the entire dataset for the purpose of preliminary investigation. The resultant predictive model can be written as:

$$Y_i = 12.456 + 52.075X_i \qquad (5)$$

where Y_i represents the prostate-specific antigen and X_i represents the seminal vesicle invasion (1 if yes and 0 if other). The model states that if the patient had seminal vesicle invasion, the prostate-specific antigen level (PSA) increases by 57.075mg/ml.

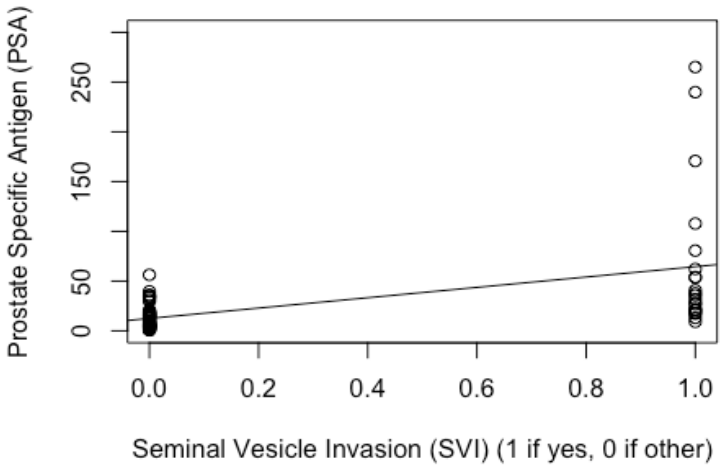


Figure 7: Prediction of PSA (mg/ml) with reference to SVI.

Effect of Capsular Penetration (CP) on Prostate-Specific Antigen (PSA)

Figure 8 illustrates the scatterplot computed by the linear regression model fitted to the data for Capsular Penetration (CP) and Prostate-Specific Antigen Level (PSA). This illustration is conducted using the entire dataset for the purpose of preliminary investigation. The resultant predictive model can be written as:

$$Y_i = 10.399 + 5.937X_i \qquad (6)$$

where Y_i represents the prostate-specific antigen and X_i represents the capsular. The model states that for every one unit (cm) increase in the degree of capsular penetration, the prostate-specific antigen level (PSA) increases by 5.937mg/ml.

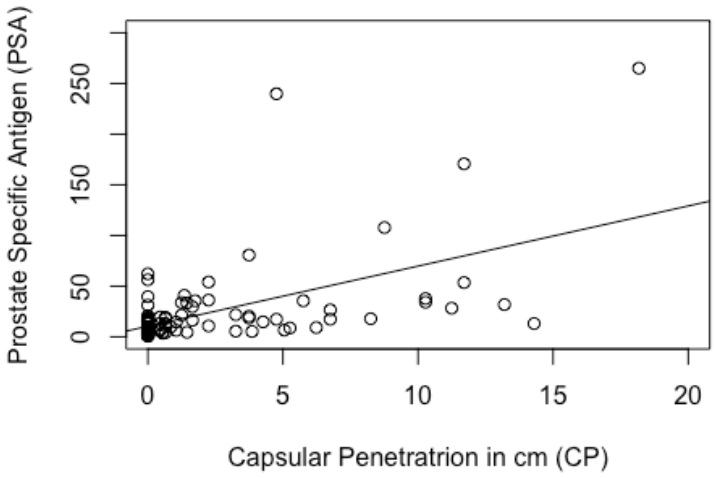


Figure 8: Prediction of PSA (mg/ml) with reference to CP in cm.

Effect of Gleason Score (GS) on Prostate-Specific Antigen (PSA)

Figure 9 illustrates the scatterplot computed by the linear regression model fitted to the data for Gleason Score (GS) and Prostate-Specific Antigen Level (PSA). This illustration is conducted using the entire dataset for the purpose of preliminary investigation. The resultant predictive model can be written as:

$$Y_i = -139.150 + 23.687X_i \qquad (7)$$

where Y_i represents the prostate-specific antigen and X_i represents the capsular. The model states that for every one unit increase in the Gleason score, the prostate-specific antigen level (PSA) increases by 23.687mg/ml.

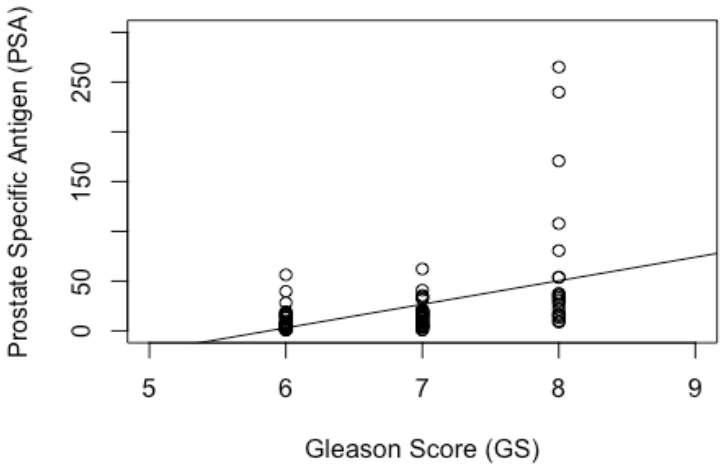


Figure 9: Prediction of PSA (mg/ml) with reference to GS.

Table 2: Summary Statistics for seven separate models: outcome and each of the predictors.

	Dependent Variable: PSA						
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
	lm(psa~cv)	*lm*(psa~pw)	*lm*(psa~age)	*lm*(psa~bph)	*lm*(psa~svi)	*lm*(psa~cp)	*lm*(psa~gs)
CV	3.229***						
PW		0.023					
Age			0.0942				
BPH				-0.222			
SVI					52.075***		
CP						5.937***	
GS							23.687***
Constant	1.125	22.666***	17.713	24.292***	12.456**	10.399*	-139.150***

Observations	97	97	97	97	97	97	97
R^2	0.390	0.00068	0.00029	0.00027	0.27943	0.30337	0.18454
Adjusted R^2	0.383	-0.00983	-0.01023	-0.01025	0.27185	0.29603	0.17595
Residual SE	32.03	40.98	40.99	40.99	34.8	34.22	37.01
F Statistic	60.63	0.06532	0.02811	0.02583	36.84	41.37	21.5
p-value	8.468e-12	0.7988	0.8672	0.8727	2.614e-08	5.055e-09	1.129e-05
*Note: *95 degrees of freedom used and where p<0.1; **p<0.05; ***p<0.01*							

Multicollinearity between predictor variables

Table 3 and Figure 7 show the correlation coefficients r between all the variables present in the dataset. Response variable PSA seems to be correlated to some of the predictor variables such as SVI, CP, and GS. PSA has the strongest correlation to SVI. From Figure 7, we can see that there is some moderate multicollinearity, however it is not very significant.

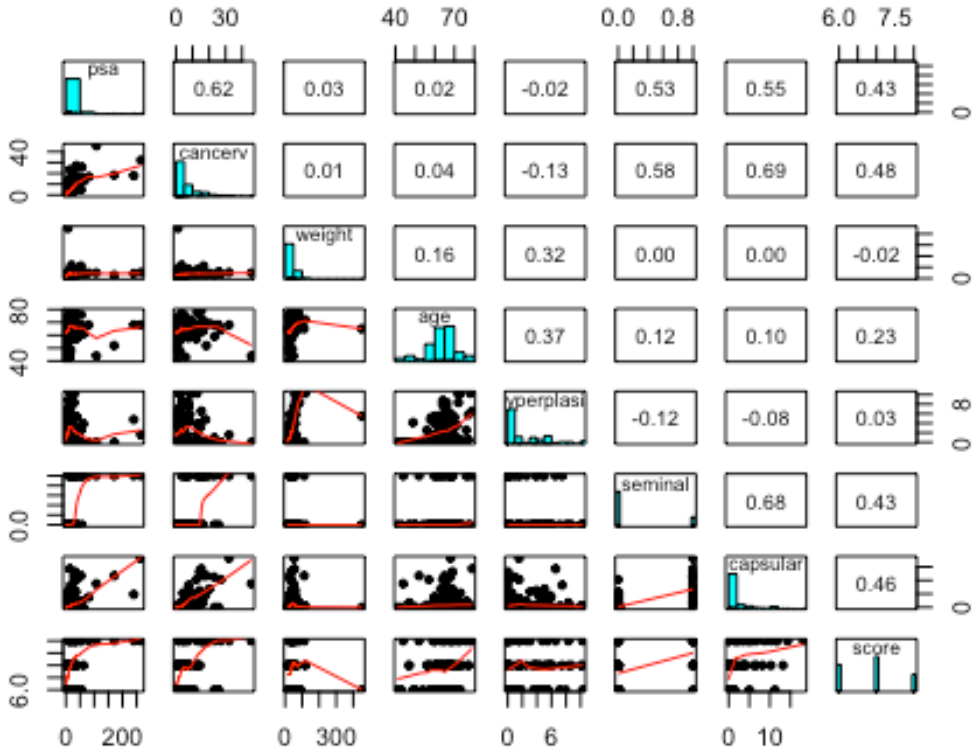


Figure 7: Pairwise Pearson's correlations coefficients for all variables.

Multicollinearity occurs when two or more predictor variables are highly correlated. This can cause problems when it comes to linear regression analysis. The multicollinearity is typically measured by analyzing the coefficient of correlation or Pearson correlation r. Values of r close to ±1 are considered to be highly correlated. Figure 7 and Table 3 show the correlation matrix and pairwise Pearson's correlation coefficients. The highest r value in this dataset is the correlation between CV and CP (0.69) and between SVI and CP (0.68). While neither of these values are very close to ±1, we can still check the multicollinearity by calculating the VIF (variation inflation factor). If VIF exceeds 10, multicollinearity is present.

Table 3: Correlation Matrix for the Dataset

	CV	**PW**	**Age**	**BPH**	**SVI**	**CP**	**GS**	**PSA**
CV	1.00	0.005	0.039	-0.133	0.582	0.693	0.481	0.624
PW	0.005	1.00	0.164	0.322	-0.002	0.002	-0.024	0.026
Age	0.039	0.164	1.00	0.366	0.118	0.099	0.226	0.017
BPH	-0.133	0.322	0.366	1.00	-0.119	-0.083	0.027	-0.017
SVI	0.582	-0.002	0.118	-0.119	1.00	0.680	0.429	0.529
CP	0.693	0.002	0.099	-0.083	0.680	1.00	0.462	0.551
GS	0.481	-0.024	0.226	0.027	0.429	0.462	1.00	0.429
PSA	0.624	0.026	0.017	-0.016	0.529	0.551	0.429	1.00

Table 4: VIF values from the final model.

CV	PW	Age	BPH	SVI	CP	GS
2.622520	1.133358	1.240701	1.430927	2.267358	2.571224	1.596803

Due to the variance inflation factors (VIFs) of the response variables, each value is in the range 1-5 meaning that there is moderate correlation between PSA and any of the predictor variables. However, this correlation is not significant enough to require attention.

Model selection

Automatic variable selection method

Automatic variable selection method helps to eliminate redundant variables from the dataset. For this, the "leap" package is used and more specifically the "regsubsets" function from the package. The best model is selected based on Mallow's C_p, BIC and adjusted R^2 or R_a^2. The selection method selects the model which has least value for Mallow's C_p and BIC. The model

with the largest R_a^2 value is considered the "best" or most accurate. In this case, the best model includes all the seven predictor variables. Refer to Table 5 for a summary of automatic selection method.

Thus, the final model can be expressed as:

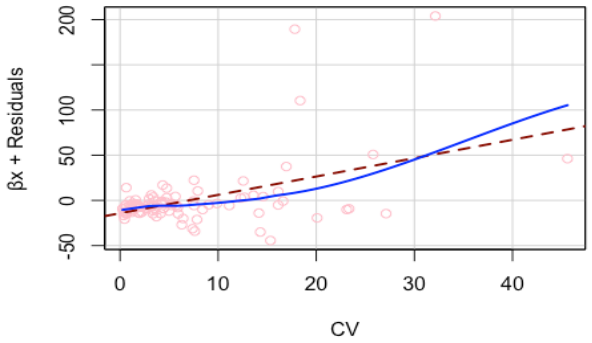
$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + 5_1 X_5 + \beta_6 X_6 + \beta_1 X_7 + \varepsilon_i$$

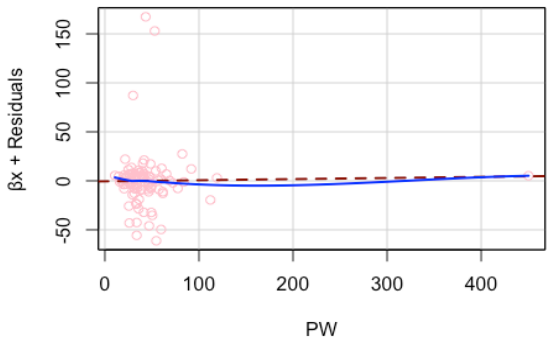
where
Y_i is the prostate-specific antigen (PSA)
X_1 is the cancer volume (CV)
X_2 is the prostate weight (PW)
X_3 is the patient age (PA)
X_4 is the benign prostatic hyperplasia (BPH)
X_5 is the seminal vesicle invasion (SVI)
X_6 is the capsular penetration (CP)
X_7 is the Gleason score (GS)
ε_i is the error term; $\varepsilon_i \sim iidN(0, \sigma^2)$
$i = 1,2,3, \cdots, 97$.
$\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$ and σ^2 are the unknown parameters to be estimated.

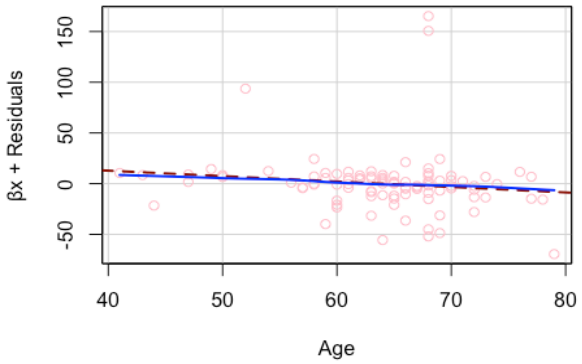
Table 5: Automatic selection method statistics

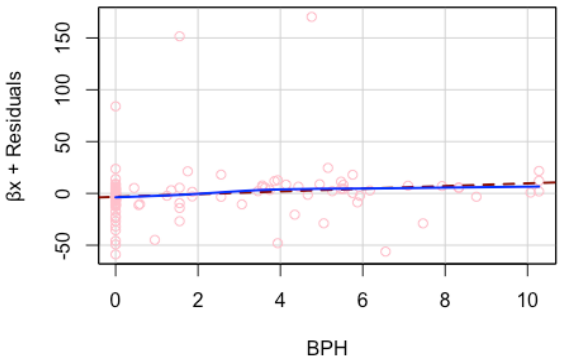
	Dependent Variable: PSA		
	C_p	R_a^2	BIC
CV	7.335852	0.3831383	-38.7281
CV + SVI	2.528792	0.4188708	-40.9679
CV + SVI + GS	2.638356	**0.4244943**	-38.3739
CV + SVI + GS + BPH	3.927756	0.4227501	-34.5543
CV + SVI + GS + BPH + Age	4.700370	0.4242843	-31.2978
CV + SVI + GS + BPH + Age + CP	6.023423	0.4222805	-27.4579
CV + SVI + GS + BPH + Age + CP + PW	8.000000	0.4159430	-22.9088

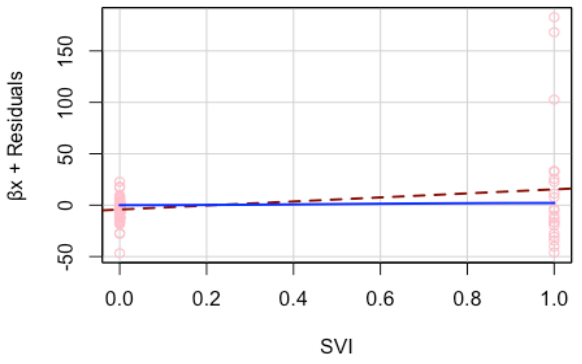
From Table 5, we can see that the third model (CV + SVI + GS) has the highest adjusted R squared value and this is the selected model.

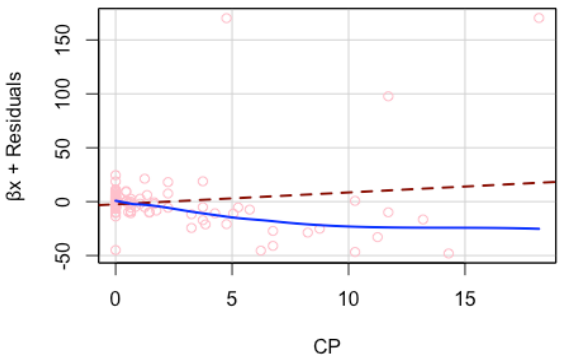

βx + Residuals
-50
0
50
100
200
0
10
20
30
40
CV


βx + Residuals
-50
0
50
100
150
0
100
200
300
400
PW


βx + Residuals
-50
0
50
100
150
40
50
60
70
80
Age


βx + Residuals
-50
0
50
100
150
0
2
4
6
8
10
BPH


βx + Residuals
-50
0
50
100
150
0.0
0.2
0.4
0.6
0.8
1.0
SVI


βx + Residuals
-50
0
50
100
150
0
5
10
15
CP

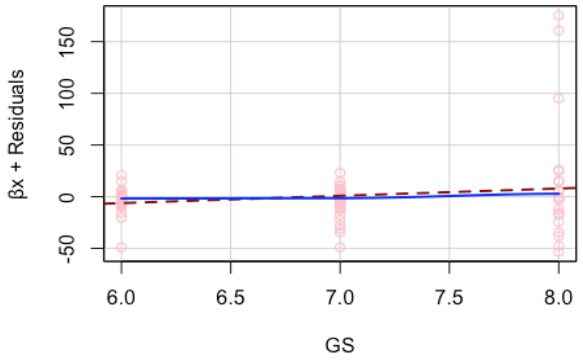


Figure 8: Added variable plots.

These partial residual plots are useful in detecting the significance of a variable in the presence of the other variables. This can also be helpful to detect possible outliers. This also serves as a tool to determine if any transformations are necessary in the predictor models. In this case, Figure 8 shows us that no transformations are required for the current model. From here, we can determine the presence of outliers.

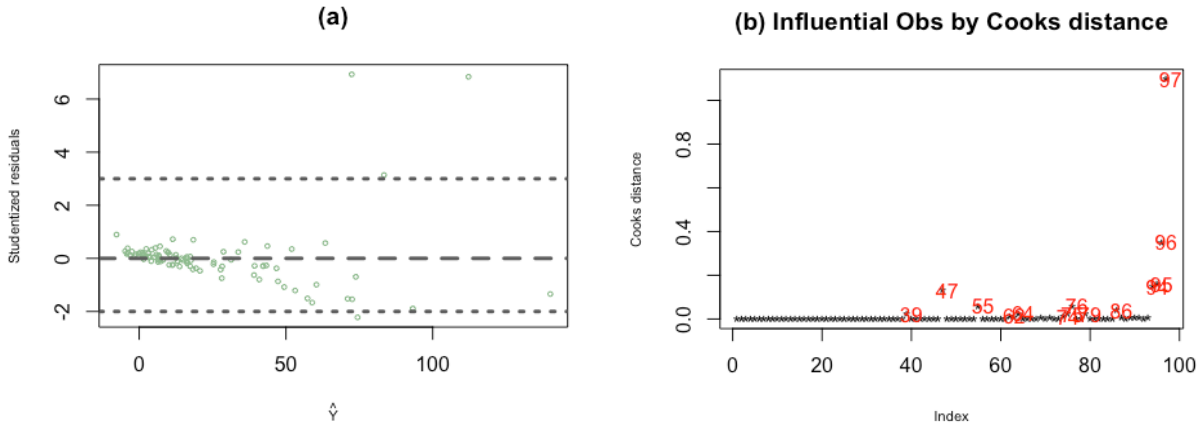
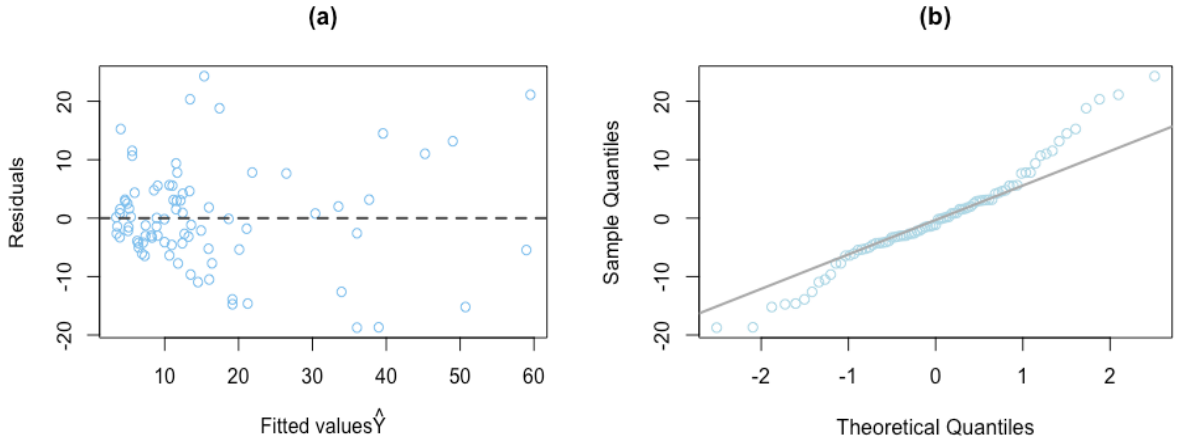


Figure 9: Detecting Outliers

To detect outliers in this study, the studentized residual plot is used. Based on this plot, 90% of the data points should be within the range of ±3 (Figure 9a). Fourteen outliers are detected in the study (Figure 9d). To check if these outliers are influential points or not, a linear model is considered that excludes these data points (obs. 39, 47, 55, 62, 64, 74, 75, 76, 79, 86, 94, 95, 96, 97). We find that the final model does significantly change the model statistics, and these outliers are excluded from the final model.

Table 6: Comparison of model summary with and without outliers

	Dependent variable (PSA):	
	Final model with outliers	Final model without outliers
CV	2.2496 (3.72e-05 ***)	1.638 (7.71e-11***)
SVI	21.8808 (0.0248 *)	22.134 (1.30e-08***)
GS	6.8982 (0.1693)	3.315 (0.0312*)
Constant	-44.1849 (0.1830)	-16.892 (0.0928)
Observations	97	83
R^2	0.4425	0.7028
Adjusted R^2	0.4245	0.6915
Residual SE	30.94 (df = 93)	8.648 (df = 79)
F Statistic	24.6 (df = 3; 93)	62.26 (df = 4; 79)
p-value	8.306e-12	2.2e-16
		*Note: *p<0.1; **p<0.05; ***p<0.01*



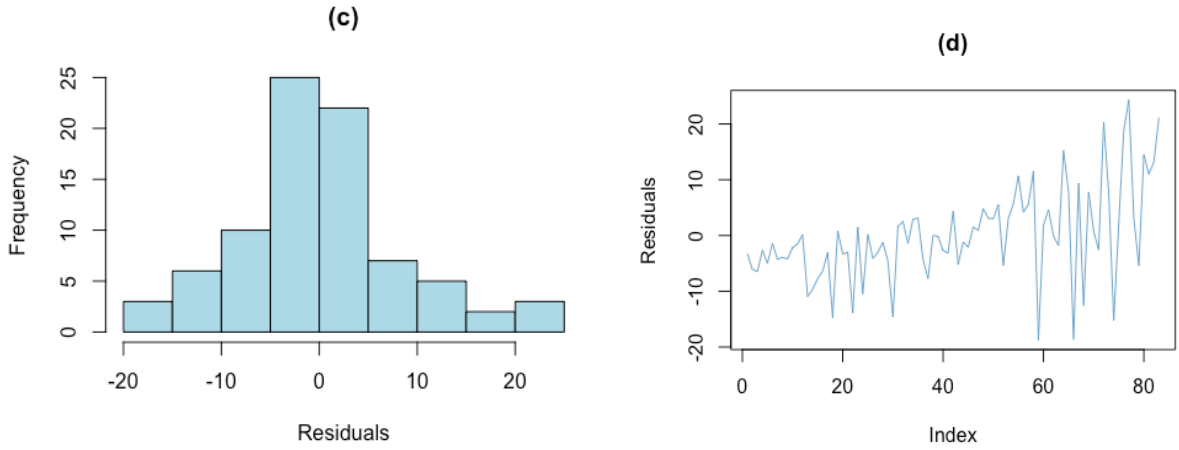


Figure 10: Residuals Diagnostics

Figure 10 shows that all the model assumptions are satisfied by the model. Figure 10a shows the model has equal variance. Figure 10 (b) and (c) confirms that the error terms follow normality. The independence of error terms are also evident in the sequence plot (Figure 10d).

Goodness of Fit Test

To fulfill the objective of this study, scatter plots, tables, linear regression models, t-test, p-value, coefficient of determination or R^2, adjusted R^2 , Mallow's C_p and BIC has been used. All the analyses have been performed using R version 4.3.1 (2023-06-16).

Results

For each of the predictor variables, we test if there is a linear association between them and Prostate-Specific Antigen (PSA). The t test is run under the following conditions:

* **Null hypothesis** :: $H_0 : \beta_1 = 0$
* **Alternative hypothesis** :: $H_1 : \beta_1 \neq 0$.

The decision is taken considering $t^* = (b_1 - \beta_1) / (SE(b_1))$,
where,
t^* is the test-statistics for the t test
b_1 is the observed slope coefficient
β_1 is the expected slope coefficient of the fitted regression model
$SE(b_1)$ is the sampling variability of b_1

The t^* is tested against $t(1 - \alpha/2, df)$,
where,
α is the level of significance = 0.05
df is the degrees of freedom, i.e, *df = no. of observations - no. of estimate parameters = (n − 2)*

If $t^* > t(1 - \alpha/2, df)$, H_0 is rejected. Otherwise, we fail to reject H_0.
The decision rule also considers the p-value and the R^2.
If the p-value $\leq \alpha$, then we reject H_0. Otherwise, we fail to reject H_0.
While considering the coefficient of determination (R^2), if this value is close to 1, then the association between variables is considered strong and the proportion of explained variation within Y is significantly higher than the unexplained variation. If the value is closer to 0, the model is not considered a "good fit" indicating a weak association between the variables and the unexplained variation of Y is significantly high.

Effect of Cancer Volume (CV) on Prostate-Specific Antigen (PSA)

At significance level (α) = 0.05, H_0 is rejected concluding that there is a linear association between CV and PSA. The t test explains that the model (explained in Equation (1)) is able to explain 50.6% of the unexplained variation in PSA while the other 49.4% variation remains unexplained.

Effect of Seminal vesicle invasion (SVI) on Prostate-Specific Antigen (PSA)

At significance level (α) = 0.05, H_0 is rejected concluding that there is a linear association between SVI and PSA. The t test explains that the model (explained in Equation (5)) is able to explain only 41.1% of the unexplained variation in PSA while the other 58.9% variation remains unexplained.

Effect of Gleason Score (GS) on Prostate-Specific Antigen (PSA)

At significance level (α) = 0.05, the result of the t test is to reject H_0, thus, concluding that there is a linear association between GS and PSA. The t test shows that the model (explained in Equation (67) is able to explain 20.9% of the unexplained variation in PSA while the other 79.1% of the variation remains unexplained.

Primary Objective Results

The results from the t test show that a linear relationship exists between PSA and four of the seven predictor variables (CV, SVI, and GS). To find the most effective predictor variable, we can compare the R^2 values for each. This value *100 gives us the percentage of explained variation in PSA. The predictor variable with the highest R^2 value has the strongest linear association with PSA.

In this case, CV has the strongest linear association with PSA with an R^2 value of 0.506 and GS has the weakest association with PSA with an R^2 value of 20.9. In this way, the variable that can predict the maximum percentage of unexplained variation is cancer volume (CV), followed by seminal vesicle invasion (SVI), followed by the Gleason Score (GS).

To confirm that the predictor variables can actually predict the prostate specific antigen (PSA), the predicted values are compared to the actual observed values from the dataset (Table 7).

Table 7: Validating Model Predictions

Observation	CV	SVI	GS	Obs	Pred	Predicted Range

1	5.607	1	7	40.854	37.68678	$19.331 \leq \hat{Y} \leq 56.04$
2	2.637	0	7	16.28	10.63	$-6.745 \leq \hat{Y} \leq 28.01$
3	16.61	1	8	53.51	58.96	$40.47 \leq \hat{Y} \leq 77.46$
4	0.259	0	6	3.56	3.422	$-14.05 \leq \hat{Y} \leq 20.90$
5	7.463	0	8	13.30	12.4241	$-5.060 \leq \hat{Y} \leq 29.9$

Discussions and Conclusions

The estimated regression function from this data would be :

$$\hat{Y}_i = -16.892 + 1.638\, X_1 + 22.134\, X_2 + 3.315 X_3 + \varepsilon_i$$

Where,
Y_i is the prostate-specific antigen (PSA)
X_1 is the cancer volume (CV)
X_2 is the seminal vesicle invasion (SVI)
X_3 is the Gleason score (GS)
ε_i is the error term; $\varepsilon_i \sim iidN(0, \sigma^2)$
$i = 1,2,3, \cdots, 97$.

The $\sigma^2 = MSE = 71.18$
This study shows that the predictor variables: cancer volume (CV), seminal vesicle invasion (SVI), and Gleason Score (GS) can predict the prostate specific antigen level (PSA) of prostate cancer patients. All statistical analysis was conducted at a 95% confidence interval and at 0.05 significance level (α). The average predictor variables were explored individually with two-tailed t tests performed on the linear regression model for each variable. A linear association was found in four of the seven predictor variables as discussed earlier (CV, SVI, and GS).

Table 8 shows the estimated regression coefficient, the standard error, t value, p value associated with each of the predictors, R^2, adjusted R^2, MSE and F statistics of the final model. Table 9 shows the ANOVA table for the final model having SSE, SSR, MSE, MSR, F values and the corresponding p-value. Through analyzing these test statistics, we can conclude that the strongest association is between CV and PSA.

The presence of outliers was detected in the model. Removing the outliers shifted the R^2 value from 0.4425 to 0.7028 and the p value from 8.306e-12 to 2.2e-16. For these reasons, the outliers were trimmed from the dataset. At 95% prediction interval, the prostate specific antigen from the predicted model was compared with the observed values which supports the idea that the clinical diagnostic factors CV, BPH, SVI, and GS can be used to predict prostate specific antigen in men with advanced prostate cancer.

Table 8: Statistics for the Regression Model

	Estimate	Std. Error	t value	Pr(>F)
(Intercept)	-16.892	9.928	-1.701	0.0928
CV	1.638	0.218	7.513	7.71e-11
SVI	22.134	0.019558	3.489	1.30e-08
GS	3.315	0.142923	1.511	0.0312
Observations	83			
R^2	0.7028			
Adjusted R^2	0.6915			
Residual SE	8.648 (df = 79)			
F Statistic	62.26 (df = 3; 79)			

Table 9: ANOVA Table for the Regression Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
CV	1	10055.2	10055.2	134.4534	2.2e-16
SVI	1	3554.1	3554.1	47.5240	1.192e-09
GS	1	360.0	360.0	4.8136	0.03117
Residuals	79	5908.1	74.8	-	-

Appendix: R Code

Boxplots of Predictors

```
# Boxplot of Cancer Volume
ggplot(pcancer, aes(x=CV)) +
  labs(x ="(a) Boxplot of Cancer Volume") +
```

```
  xlim(0,50) +
  ylim(-1,1) +
  theme_bw() +
  theme(axis.ticks.y = element_blank(),
        axis.title.y = element_blank(),
        axis.text.y = element_blank())+
  stat_boxplot(geom ='errorbar') +
  geom_boxplot()

# Boxplot of Weight
ggplot(pcancer, aes(x=PW)) +
  labs(x ="(b) Boxplot of Weight") +
  xlim(0,500) +
  ylim(-1,1) +
  theme_bw() +
  theme(axis.ticks.y = element_blank(),
        axis.title.y = element_blank(),
        axis.text.y = element_blank())+
  stat_boxplot(geom ='errorbar') +
  geom_boxplot()

# Boxplot of Age
ggplot(pcancer, aes(x=Age)) +
  labs(x ="(c) Boxplot of Age") +
  xlim(35,90) +
  ylim(-1,1) +
  theme_bw() +
  theme(axis.ticks.y = element_blank(),
        axis.title.y = element_blank(),
        axis.text.y = element_blank())+
  stat_boxplot(geom ='errorbar') +
  geom_boxplot()

# Boxplot of Hyperplasia
ggplot(pcancer, aes(x=BPH)) +
  labs(x ="(d) Boxplot of Hyperplasia") +
  xlim(0,12) +
  ylim(-1,1) +
  theme_bw() +
  theme(axis.ticks.y = element_blank(),
        axis.title.y = element_blank(),
        axis.text.y = element_blank())+
  stat_boxplot(geom ='errorbar') +
  geom_boxplot()

# Boxplot of Seminal vesicle invasion
ggplot(pcancer, aes(x=SVI)) +
  labs(x ="(e) Boxplot of Seminal vesicle invasion") +
  xlim(-1,3) +
  ylim(-1,1) +
  theme_bw() +
```

```
  theme(axis.ticks.y = element_blank(),
        axis.title.y = element_blank(),
        axis.text.y = element_blank())+
  stat_boxplot(geom ='errorbar') +
  geom_boxplot()

# Boxplot of Capsular penetration
ggplot(pcancer, aes(x=CP)) +
  labs(x ="(f) Boxplot of Capsular penetration") +
  xlim(0,20) +
  ylim(-1,1) +
  theme_bw() +
  theme(axis.ticks.y = element_blank(),
        axis.title.y = element_blank(),
        axis.text.y = element_blank())+
  stat_boxplot(geom ='errorbar') +
  geom_boxplot()

# Bloxplot of Gleason Score
ggplot(pcancer, aes(x=GS)) +
  labs(x ="(g) Boxplot of Gleason Scores") +
  xlim(5,10) +
  ylim(-1,1) +
  theme_bw() +
  theme(axis.ticks.y = element_blank(),
        axis.title.y = element_blank(),
        axis.text.y = element_blank())+
  stat_boxplot(geom ='errorbar') +
  geom_boxplot()
```

Plots of Response Variable (PSA)

```
# Boxplot of Response Variable (PSA)
ggplot(pcancer, aes(x=psa)) +
  labs(x ="(a) Boxplot of Prostate-Specific Antigen Level (PSA)") +
  xlim(0,275) +
  ylim(-1,1) +
  theme_bw() +
  theme(axis.ticks.y = element_blank(),
        axis.title.y = element_blank(),
        axis.text.y = element_blank())+
  stat_boxplot(geom ='errorbar') +
  geom_boxplot()

# plot of PSA
ggplot(pcancer, aes(x=psa)) +
  labs(x ="(b) Prostate-Specific Antigen Level (PSA)", y="Frequency") +
  theme_bw() +
  geom_histogram(binwidth = 15)
```

PSA vs. Cancer Volume

```
plot(psa~CV, pcancer, xlim=c(0,50), ylim=c(0,300), xlab="Cancer Volume in cc (CV)", ylab = "Prostate Specific Antigen (PSA)")
abline(lm(psa~CV, pcancer))

mod_v <- pcancer %>% lm(psa ~ CV, data = .)
mod_v %>% tidy()
```

PSA vs. Weight

```
plot(psa~PW, pcancer, xlim=c(0,500), ylim=c(0,300), xlab="Prostate Weight in kg (PW)", ylab = "Prostate Specific Antigen (PSA)")
abline(lm(psa~PW, pcancer))

mod_w <- pcancer %>% lm(psa ~ PW, data = .)
mod_w %>% tidy()
```

PSA vs. Age

```
plot(psa~Age, pcancer, xlim=c(40,85), ylim=c(0,300), xlab="Patient Age (PA)", ylab = "Prostate Specific Antigen (PSA)")
abline(lm(psa~Age, pcancer))

mod_a <- pcancer %>% lm(psa ~ Age, data = .)
mod_a %>% tidy()
```

PSA vs. Hyperplasia

```
plot(psa~BPH, pcancer, xlim=c(0,11), ylim=c(0,300), xlab="Hyperplasia in cm^2 (BPH)", ylab = "Prostate Specific Antigen (PSA)")
abline(lm(psa~BPH, pcancer))

mod_hyper <- pcancer %>% lm(psa ~ BPH, data = .)
mod_hyper %>% tidy()
```

PSA vs. Seminal

```
plot(psa~SVI, pcancer, xlim=c(0,1), ylim=c(0,300), xlab="Seminal Vesicle Invasion (SVI) (1 if yes, 0 if other)", ylab = "Prostate Specific Antigen (PSA)")
abline(lm(psa~SVI, pcancer))

mod_sem <- pcancer %>% lm(psa ~ SVI, data = .)
mod_sem %>% tidy()
```

PSA vs. Capsular

```
plot(psa~CP, pcancer, xlim=c(0,20), ylim=c(0,300), xlab="Capsular Penetratrion in cm (CP)", ylab = "Prostate Specific Antigen (PSA)")
abline(lm(psa~CP, pcancer))
```

```
mod_cap <- pcancer %>% lm(psa ~ CP, data = .)
mod_cap %>% tidy()
```

PSA vs. Gleason Score

```
plot(psa~GS, pcancer, xlim=c(5,9), ylim=c(0,300), xlab="Gleason Score (GS)",
ylab = "Prostate Specific Antigen (PSA)")
abline(lm(psa~GS, pcancer))

mod_score <- pcancer %>% lm(psa ~ GS, data = .)
mod_score %>% tidy()
```

Multicolinearity between predictors

```
ppcancer <- pcancer %>% select(-idnum)
pairs.panels(ppcancer, ellipses = FALSE,
density = FALSE)
```

Model Selection

```
avsm <- regsubsets(psa ~ CV + PW + Age + BPH + SVI + CP + GS, data = pcancer)
sum <- summary(avsm)
sum$which

sum$cp

sum$bic

sum$adjr2
```

Plotting Residuals

```
my.lm <- lm(psa ~ CV + PW + Age + BPH + SVI + CP + GS, data = pcancer)

crPlots(my.lm,col = "pink", layout = NA, ylab = "βx + Residuals", col.lines =
c("darkred",
"blue"))
```

Detect presence of outliers

```
plot(fitted(my.lm), rstudent(my.lm),
col = "darkseagreen",
xlab = expression(hat(Y)),
ylab = "Studentized residuals",
main = "(a)",
cex.lab = 0.7, cex = 0.5)
abline(h = 0, lty = 2, lwd = 3,
col = "gray36")
abline(h = 3, lty = 3, lwd = 3,
```

```
col = "gray36")
abline(h = -3, lty = 3, lwd = 3,
col = "gray36")

# cooks distance
cooksd <- cooks.distance(my.lm)
plot(cooksd, pch = "*",
main = "(b) Influential Obs by Cooks distance",
ylab = "Cooks distance", cex.lab = 0.7)
text(x = 1:length(cooksd) + 1,
y = cooksd, labels = ifelse(cooksd >
0.01, names(cooksd), ""),
col = "red")
```

Dataset without outliers

```
no.outs <- pcancer[-c(39, 47, 55, 62, 64, 74, 75, 76, 79, 86, 94, 95, 96, 97)
,]
with.outs.model <- lm(psa ~ CV + PW + Age + BPH + SVI + CP + GS, pcancer)
no.outs.model <- lm(psa ~ CV + PW + Age + BPH + SVI + CP + GS, no.outs)
summary(with.outs.model)

summary(no.outs.model)
```

Residuals Diagnostics

```
plot(fitted(with.outs.model), residuals(with.outs.model),
col = "skyblue2", main = "(a)",
xlab = expression("Fitted values" *
hat(Y)), ylab = "Residuals")
abline(h = 0, col = "gray26", lwd = 2,
lty = 2)

#Normality
qqnorm(residuals(with.outs.model), col = "lightblue",
main = "(b)")
qqline(residuals(with.outs.model), col = "darkgray",
lwd = 2)

hist(residuals(with.outs.model), col = "lightblue", main = "(c)",
xlab = "Residu

als")

# Independence of error terms
plot(residuals(with.outs.model), type = "l",
```

```
col = "skyblue3", main = "(d)",
ylab = "Residuals")

# R squared values and t tests

library(lmtest)
t <- qt(0.95, (97 - 2))
summary(lm(psa~CV, no.outs))

summary(lm(psa~PW, no.outs))

summary(lm(psa~Age, no.outs))

summary(lm(psa~BPH, no.outs))

summary(lm(psa~SVI, no.outs))

summary(lm(psa~CP, no.outs))

summary(lm(psa~GS, no.outs))
```

Validating Model Predictions

```
no.outs.model <- lm(psa ~ CV + PW + Age + BPH + SVI + CP + GS, no.outs)

#1
predict(no.outs.model, data.frame(CV = 4.2631, PW=22.646, Age=68, BPH=1.3499,
SVI=0, CP=0.0000, GS=6), level = 0.95, interval = "prediction")

#2
predict(no.outs.model, data.frame(CV = 2.6379, PW=17.637, Age=47, BPH=0.0000,
SVI=0, CP=1.6487, GS=7), level = 0.95, interval = "prediction")

#3
predict(no.outs.model, data.frame(CV = 16.6099, PW=112.168, Age=65, BPH=0.000
0, SVI=1, CP=11.7048, GS=8), level = 0.95, interval = "prediction")

#4
predict(no.outs.model, data.frame(CV = 0.2592, PW=36.598, Age=63, BPH=3.5609,
SVI=0, CP=0.0000, GS=6), level = 0.95, interval = "prediction")

#5
predict(no.outs.model, data.frame(CV = 7.4633, PW=83.931, Age=72, BPH=8.3311,
SVI=0, CP=1.6487, GS=8), level = 0.95, interval = "prediction")
```

ANOVA

```
anova(no.outs.model)

model.summary <- summary(no.outs.model)
mean(model.summary$residuals^2)
```

References

1. Thompson IM, Pauler DK, Goodman PJ, et al. Prevalence of prostate cancer among men with a prostate-specific antigen level < or =4.0 ng per milliliter. *New England Journal of Medicine* 2004; 350(22):2239–2246. [PubMed Abstract]

2. Barry MJ. Clinical practice. Prostate-specific-antigen testing for early diagnosis of prostate cancer. *New England Journal of Medicine* 2001; 344(18):1373–1377. [PubMed Abstract]

3. Martin RM, Donovan JL, Turner EL, et al. Effect of a low-intensity PSA-based screening intervention on prostate cancer mortality: The CAP randomized clinical trial. *JAMA* 2018; 319(9):883–895. [PubMed Abstract]