PharmaSUG 2021 - Paper SA-047

Build a model: Introducing a methodology to develop multiple regression models using R in oncology trial analyses.

Girish Kankipati and Jai Deep Mittapalli, Seagen Inc., Bothell WA

ABSTRACT

Regression analysis is a widely used modeling tool to estimate the relationship between two variables in statistics: a predictor variable whose value is observed through experiments and a response variable whose value is derived from the predictor. The general mathematical equation for linear regression is Y = aX + b, where Y is the response or dependent variable, X is the predictor or independent variable, and a and b are constants called coefficients.

If two or more predictor variables have a linear relationship with the dependent variable, the regression is called a multiple linear regression. In R, the lm function is widely used to build such regression models.

This paper discusses a step-by-step process to build a multiple regression model in R using an example oncology trial data set:

a) Get the regression equation on each predictor using the lm function.
b) Check if multicollinearity exists per a pairwise Pearson's coefficient.
c) Select the model using the regsubsets function in the leaps package in R.
d) Check if any data transformation is required and identify any outliers and influential points.
e) Perform residual diagnostics to see whether the predictor variables met all model assumptions such as normality, homoscedasticity, and linearity.

In our sample data set, prostate-specific antigen (PSA) level is the response variable (Y), and prostate cancer volume, prostate weight, and others are the predictor variables (X). Modeling the relationships among these variables in R will be explained thoroughly with the help of the box and other plots using R shiny.

INTRODUCTION

The simple linear regression is used to determine a quantitative outcome 'Y' based on a single predictor variable X. The end goal is to build a formula that defines 'Y' as a function of the X variable. Once a statistical model is built it's possible to use it for future outcomes based on new X values. The mathematical formula of the linear regression can be written as $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon_i$, where β_0 and β_1 are known as the regression beta coefficients or parameters, $\beta_0$is the intercept of the regression line, that is the predicted value when X = 0, β_1 is the slope of the regression line and $\varepsilon_i$is the standard error.

Multiple linear regression is an extension of simple linear regression used to predict an outcome variable (Y) based on multiple distinct predictor variables (X). With three predictor variables (X), the prediction of Y is expressed by the following equation:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon_i$$

The "β" values are called the beta coefficients. They measure the association between the predictor variable and the outcome. The example used to build a model for multiple regression is taken from Kutner et al (2014). A sample size of n = 97 is collected and examined. The data was collected in an Excel spreadsheet. The dataset contains prognostic clinical measurements like Cancer volume, Seminal, Glean Score values, Hyperplasia, Age, Capsular, and Weight . In this example, PSA level and prognostic clinical measurements were checked to determine if an association exists between PSA level and prognostic clinical measurements; specifically, to determine how many factors affect PSA level, the individual effect of these factors, and the combined effect. This is done by analyzing

the individual effect of predictor variables, constructing a model with factors affecting PSA level, analyzing the data, checking for the right model to fit the data, and applying transformations of the data as required.

The data set consists of eight prognostic clinical measurements. The PSA level is considered the response variable for this study. Predictor variables in the data set include cancer volume (cancerv) ranging from 0.2592 to 45.6042 cc, prostate weight (weight) ranging from 10.697 to 450.339 gm, age (age) ranging from 41 to 79 years, hyperplasia (hyperplasia) ranging from 0 to 10.2779 cm, presence of seminal vesicle invasion (seminal) being either 0 (no) or 1 (yes), capsular penetration (capsular) ranging from 0 cm to 18.1741 cm, and Gleason score (score) with a value of 6, 7, or 8 where a higher score indicates a worse prognosis.

The data analysis is done using the statistical software R version 4.0.0 and the project focuses on the multiple linear regression model. Each predictor variable in the data set is explored individually. No missing values are found in this relatively small data set. The automatic model selection method is used to narrow down model selections, and the final model is established using the adjusted-R^2criterion. The model assumptions are assessed and confirmed for the final model.

PRIMARY OBJECTIVE OF THE ANALYSIS:

Exploring the data is very helpful to analyze the effects and identifies skewness in the data if there are any outliers, and how the outliers are affecting the outcome of data. Data exploration also identifies any covariates in the data. Once an initial understanding of the data and its basic properties is in place, an exploration of the data to check the linearity between the predictors and outcome is conducted. This process determines whether any transformations of the data are needed. Next, the best model to fit the data is determined and used to explain the variation in the data.

Regression equation on each predictor

	psa	cancerv	weight	age	hyperplasia	capsular	score
Minimum	0.651	0.259	10.697	41.000	0.000	0.000	6.000
Maximum	265.072	45.604	450.339	79.000	10.278	18.174	8.000
1. Quartile	5.641	1.665	29.371	60.000	0.000	0.000	6.000
3. Quartile	21.328	8.415	48.424	68.000	4.759	3.254	7.000
Mean	23.730	6.999	45.491	63.866	2.535	2.245	6.876
Variance	1663.247	62.108	2088.952	55.430	9.188	14.314	0.547
Stdev	40.783	7.881	45.705	7.445	3.031	3.783	0.740

Table 1 Basic Statistics for Dataset Variables

Analysis of prostate-specific antigen level (PSA level): Analysis of the response variable (PSA level) shows that data for PSA is positively skewed from the box plot with few outliers. The basic statistics are shown in *Table 1*. Below, the bell-shaped graph in *Figure 1* indicates the positive skewness of the PSA data with the curve tending towards the right. A similar pattern can also be seen in the bar graph in *Figure 1*.

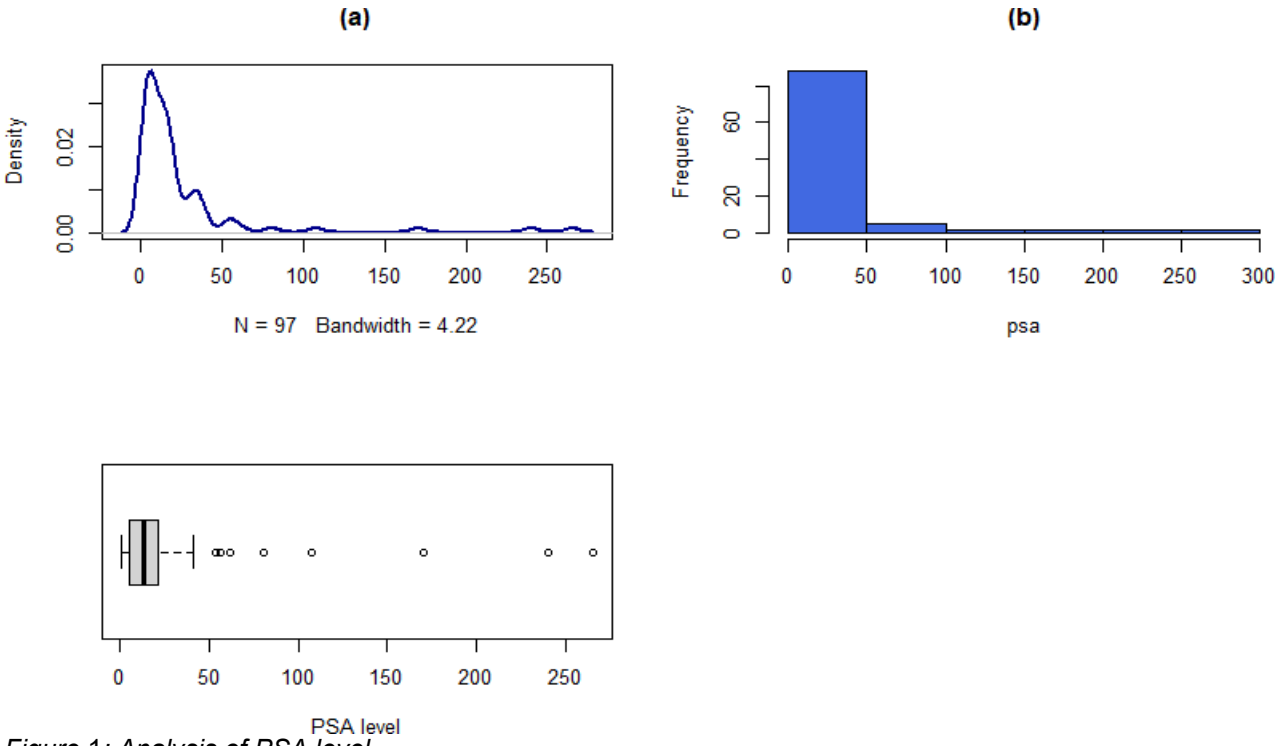

Figure 1*: Analysis of PSA level*

Analysis of cancer volume on PSA level:

Figure 2 indicates the effect of cancer volume on PSA level. The resultant predictive model can be written as:

$$Y_i = 1.1249 + 3.2299\ X_i \qquad 1$$

where Y represents the PSA level and X is the cancer volume (cc). The model states that for each additional rise of 1 cc in cancer volume, PSA level increases by 3.2299 mg/mL.

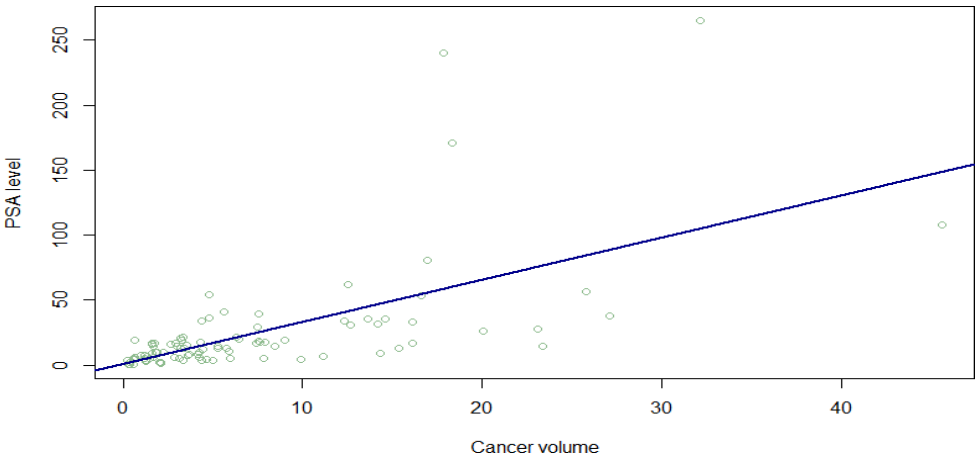

Figure 2: Analysis of cancer volume on PSA level

Analysis of prostate weight on PSA level:

Figure 3 indicates the effect of prostate weight (gm) on PSA level. The resultant predictive model can be written as:

$$Y_i = 22.66607 + 0.02339\ X_i \quad (2)$$

where Y_i represents the PSA level and X_i is the prostate weight. The model states that for each 1 gm rise in prostate weight, PSA level increases by 0.02339 mg/mL.

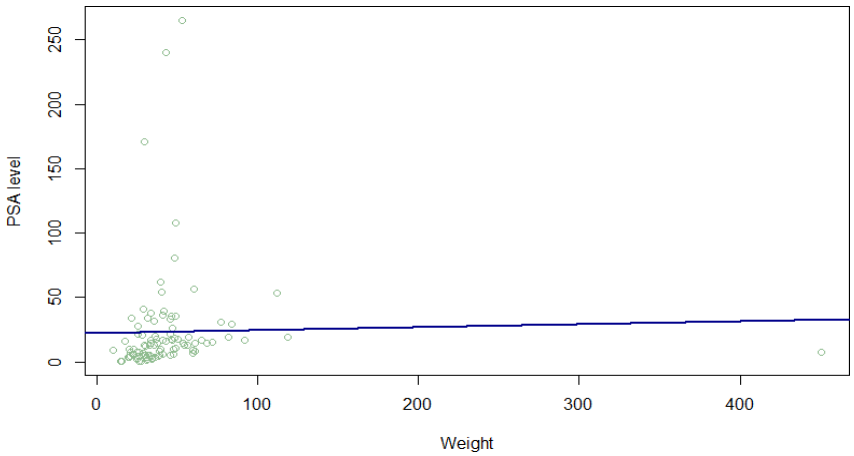

Figure 3: Analysis of prostate weight on PSA level

Analysis of seminal vesicle invasion on PSA level:

A similar analysis was done on age, benign prostatic hyperplasia, Seminal, Capsular and Gleason score. Below are following equations.

Age	$Y_i = 17.71301 + 0.09421\ X_i$	3
Benign Prostate Hyperplasia	$Y_i = 24.2924 - 0.2218\ X_i$	4
Seminal	$Y_i = 12.456 + 52.075\ X_i$	5
Capsular	$Y_i = 10.3986 + 5.9373\ X_i$	6
Gleason score	$Y_i = -139.150 + 23.687\ X_i$	7

	Dependent Variable						
	PSA level						
	Model 1 lm (psa ~ cancerv)	Model 2 lm (psa ~ seminal)	Model 3 lm (psa ~ score)	Model 4 lm (psa ~ hyperplasia)	Model 5 lm (psa ~ age)	Model 6 lm (psa ~ capsular)	Model 7 lm (psa ~ weight)
CANCERV	3.230 $8.47e^{-12}$						
SEMINAL		52.075 $2.61e^{-08}$					
SCORE			23.687 $1.13e^{-05}$				
HYPERPLASIA				-0.222 0.873			
AGE					0.092 0.867		
CAPSULAR						5.937 $5.06e^{-09}$	
WEIGHT							0.023 0.799
Constant	1.130	12.456	-139.150	24.292	17.713	10.937	22.666
Observations	97	97	97	97	97	97	97
R^2	0.390	0.279	0.184	0.0002	0.0002	0.303	0.0006
Adjusted R^2	0.383	0.271	0.176	-0.010	-0.010	0.296	-0.0098
Residual Std. Error (df=95)	32.03	34.8	37.02	40.99	40.99	34.22	40.98
F Statistic (df=1 and 95)	60.63	36.84	21.5	0.026	0.028	41.37	0.065

Table 2 : Summary statistics of seven separate models, outcome and each of the predictors

Multicollinearity between predictor variables

Both *Table 3* and *Figure 4* indicate the correlation coefficients, r, between the predictors. Based on the Pearson's coefficients values, response variable PSA level shows an almost identical correlation to the other predictors. A high correlation exists between PSA level and cancer volume. Other predictor variables also show good correlation except for weight and age.

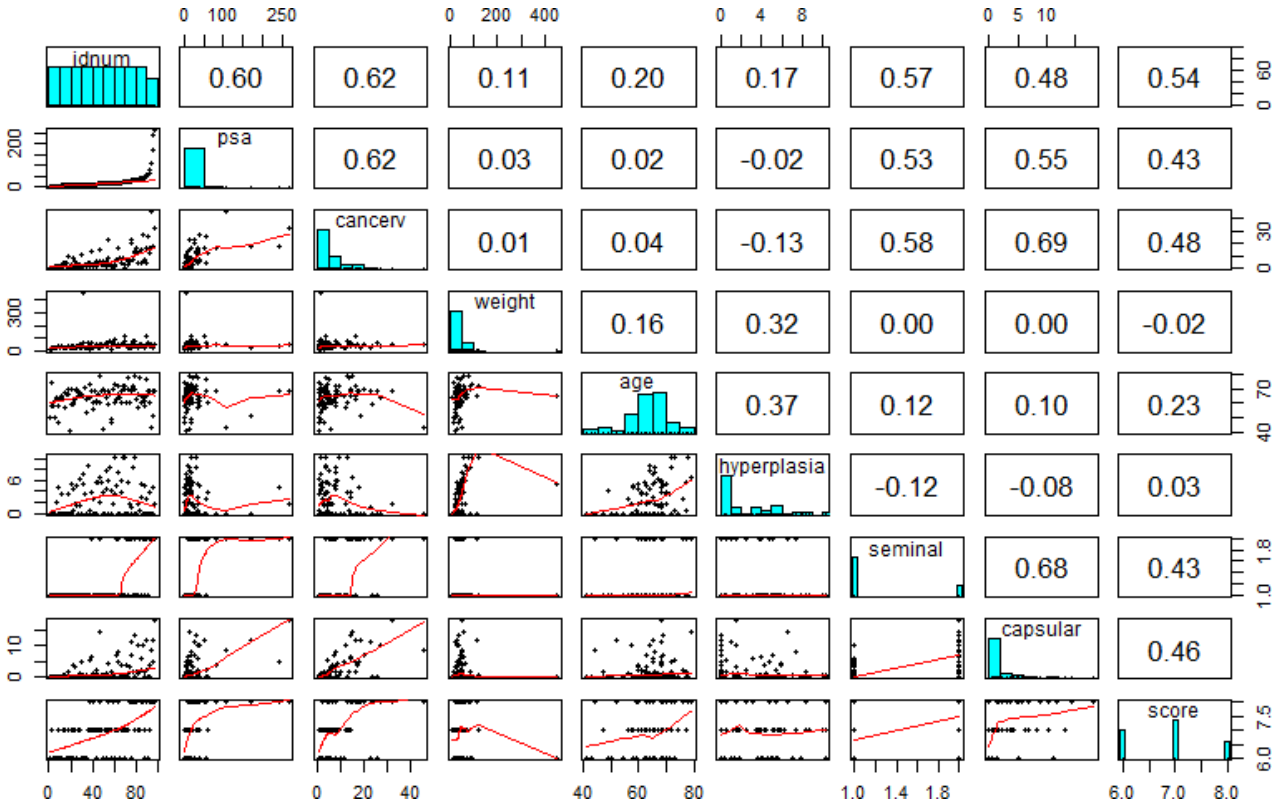

Figure 4: *Pairwise Pearson's correlations coefficients for the ambient variables*

The existence of multicollinearity among predictor variables makes a model more complex. To test this, the Pearson's correlation, r, is used. *Figure 4* shows pairwise Pearson's correlation coefficients. A correlation between two pairs with an r value nearly equal to 1 indicates a good correlation. From *Figure 3*, PSA versus cancerv, seminal, and capsular r values is more than 0.5 and indicates a good correlation, whereas weight and age show less correlation with r values of 0.03 and 0.02, respectively. Multicollinearity is also tested by checking the variance inflation factor (or VIF). A VIF value that exceeds 5 indicates a problematic amount of collinearity. *Table 4* indicates all predictors have VIF with less than 5. Thus, multicollinearity is certainly absent in the final model.

	psa	cancerv	weight	age	hyperplasia	capsular	score
psa	1.000	0.624	0.026	0.017	-0.016	0.551	0.430
cancerv	0.624	1.000	0.005	0.039	-0.133	0.693	0.481
weight	0.026	0.005	1.000	0.164	0.322	0.002	-0.024
age	0.017	0.039	0.164	1.000	0.366	0.100	0.226
hyperplasia	-0.016	-0.133	0.322	0.366	1.000	-0.083	0.027
capsular	0.551	0.693	0.002	0.100	-0.083	1.000	0.462
score	0.430	0.481	-0.024	0.226	0.027	0.462	1.000

Table 3: Correlation matrix for the data set

cancerv	seminal1	score	hyperplasia	age	capsular	weight
2.163	2.009	1.459	1.311	1.24	2.516	1.129

Table 4: VIF values from the final model

Model selection

Automatic variable selection method: used as a starting point for model selection. This is completed using the regsubsets function from the leaps package in R. After the model suggestions have been narrowed down, the best model is selected based on Mallow's Cp, BIC and adjusted R^2or R_a^2 criterion. Mallow's C_p criterion selects the model with the value of p closest to the number of variables without exceeding that value. The BIC favors the model which has the smallest value. When using the adjusted-R^2 criterion, the model with largest value of R_a^2 is considered best. In this case, as shown in Table 5, we see that model 5 is considered the best fit. It should be noted that models 4 and 5 are very close on C_p and BIC values, but model 5 was selected to include more predictors so as to not miss any predictors which might be affecting the model.

Thus, the final model can be expressed as:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon_i$$

where

Y_i= PSA level

X_1= Cancer volume

X_2= Prostate weight

X_3= Age

X_4= Benign prostatic hyperplasia

X_5= Seminal vesicle invasion

ε_i is the error term; $\varepsilon_i \sim iidN(0, \sigma^2)$

i = 1,2,3, · · ·, 95.

$\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ and σ^2 are the unknown parameters to be estimated.

		Dependent variable: PSA level		
#Model No	Predictors	C_p	R_a^2	BIC
1	CANCERV	25.222	0.306	-25.878
2	CANCERV+SEMINAL	11.730	0.393	-34.901
3	CANCERV+SEMINAL+SCORE	5.967	0.434	-37.983
4	CANCERV+SEMINAL+SCORE+HYPERPLASIA	3.032	0.459	-38.638
5	CANCERV+SEMINAL+SCORE+HYPERPLASIA+AGE	4.479	0.456	-34.706
6	CANCERV+SEMINAL+SCORE+HYPERPLASIA+AGE+CAPSULAR	6.015	0.453	-30.670
7	CANCERV+SEMINAL+SCORE+HYPERPLASIA+AGE+CAPSULAR+WEIGHT	8.000	0.447	-26.163

Table 5: Automatic selection method statistics

Transformation

It should be noted that PSA level is typically observed in ng/mL and the data set specifies PSA level in mg/mL. Previous model analysis showed large amounts of skewness in PSA level within the data set and values that did not make sense in relation to the expected value of PSA levels in cancer research. To correct for this, a log transformation is applied to the response variable PSA level. By applying the log transformation, the values of the PSA level match what is expected. Diagnostics are done with the log transformed model and all assumptions are met.

After the log transformation was applied to the model, multicollinearity was checked again. As shown in *Figure 5*, the correlation between predictor values increased slightly as compared to the original correlation comparison. However, there are still no issues of multicollinearity present in the model. Since the log transformation is done only on the response variable, there is no change to the VIF values in the model.

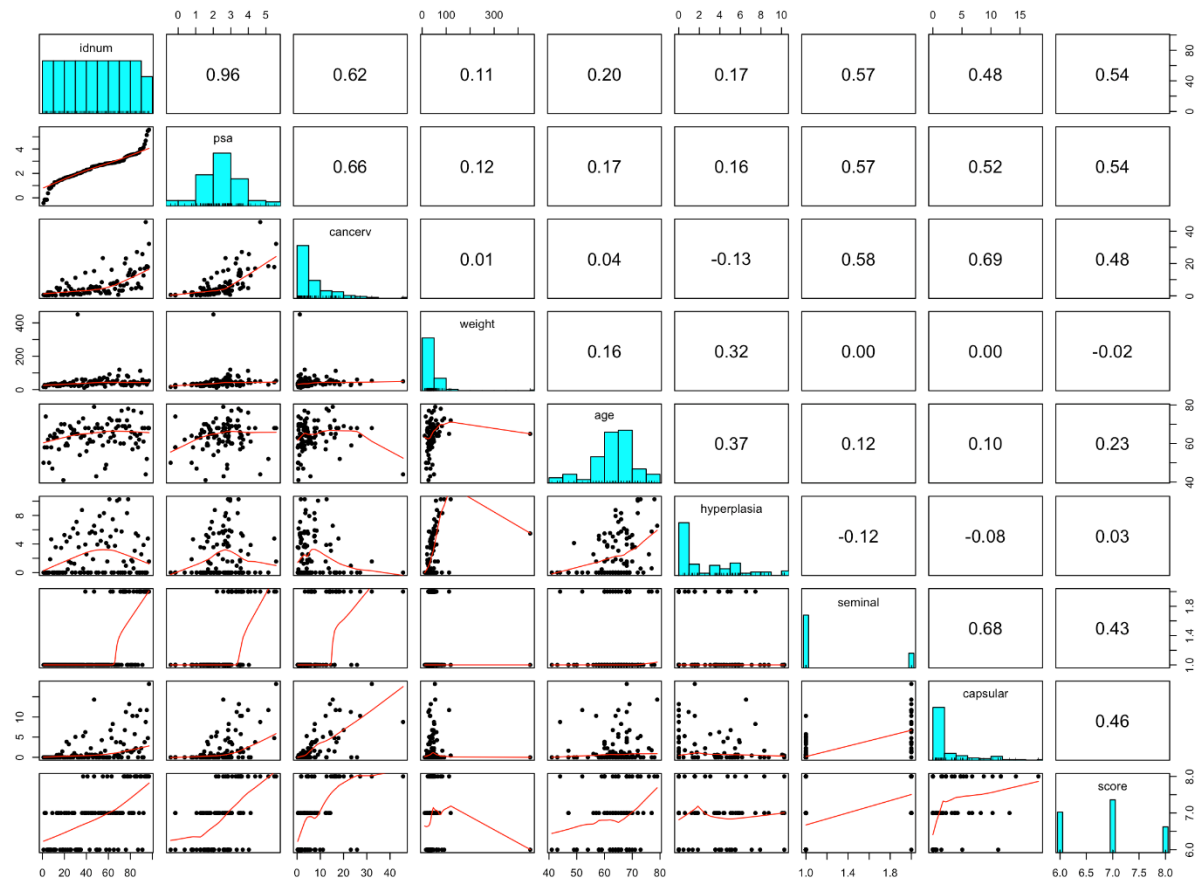

Figure 5: Pairwise Pearson's correlations coefficients for the ambient variables

Added variable plots

Figure 6 shows the significance of each variable when other covariates are present. The added variable plots also indicate if there are any outliers present in the data and if any log transformation is required for any predictor variables. Based on *Figure 6*, no further transformation is required to the covariates.

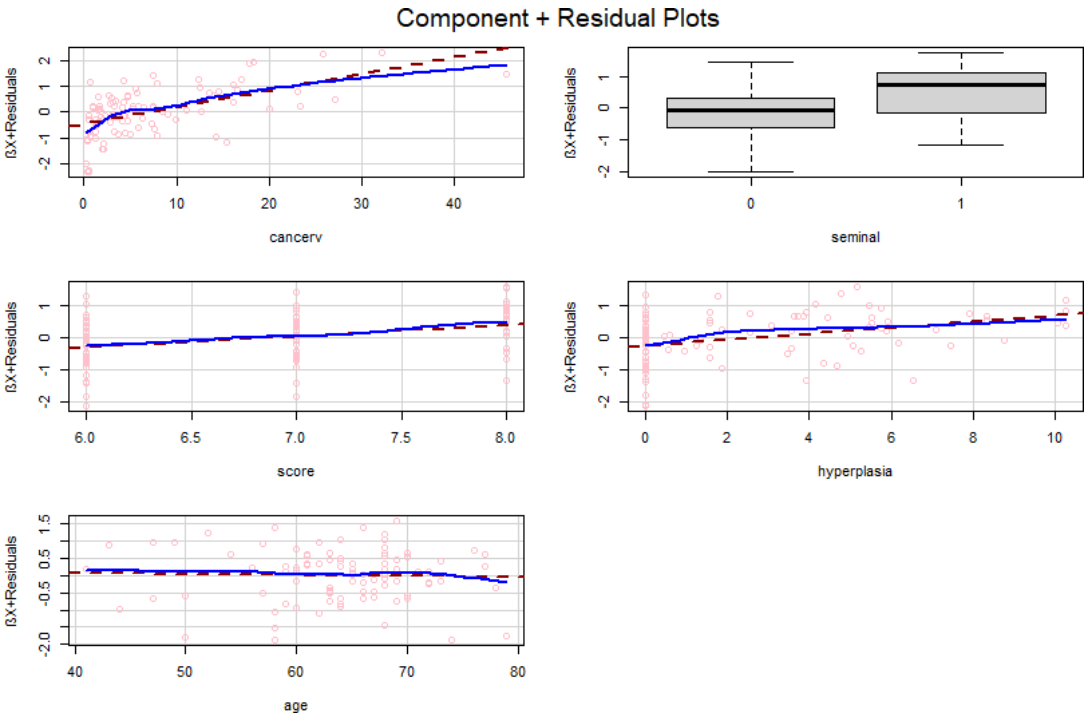

Figure 6: Added variable plots

Outliers and influential points

Figure 7 shows the presence of outliers in the given data. Studentized residual plots are used to determine the outliers. According to the data standards, data points should be within ±3. Available data show 99% of the data distributed within ±3 as per *Figure 7*(a). A total of 28 outlier observations were identified as cases 1, 2, 3, 4, 8, 9, 32, 39, 47, 49, 55, 57, 58, 64, 65, 69, 75, 82, 85, 86, 87, 88, 90, 91, 94, 95, 96, 97 as shown in *Figure 7*(b). According to Cook's Distance, case 32 is identified as a potentially influential case.

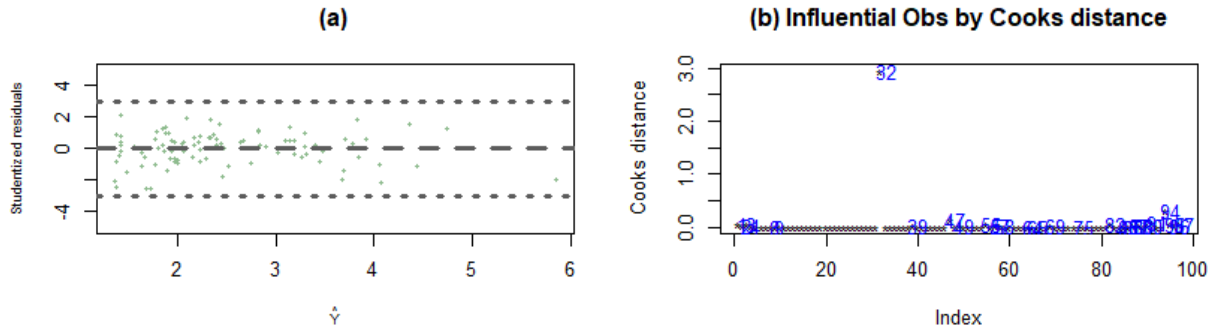

Figure 7: Detecting outliers

Table 6 shows a summary of the model. After removing influential case 32, the p-values for predictors do not show significance in the model. Since there are no major changes when the influential case is added back into the model, case 32 remains in the final model.

	Dependent Variable
CANCERV	0.0646 2.79e^{-06}
SEMINAL	0.065 0.0.005
SCORE	0.690 0.009
HYPERPLASIA	0.339 0.0.001
AGE	-0.002 0.837
Constant	-0.535
Observations	97
R^2	0.583
Adjusted R^2	0.560
Residual Std. Error	0.765 (df=91)
F Statistic	25.51 (df=5; 91)

Table 6: Model summary

Residual diagnostics

Residual diagnostics were plotted for the model. *Figure 8* shows all the model assumptions are satisfied. Homoscedasticity, or equal variance, is confirmed in *Figure 8*(a), as most of the values appear to be close to the line h=0. *Figure 8*(b) and *Figure 8*(c) confirm that the error terms follow normality. *Figure 8*(c) shows platykurtic behavior. The QQ plot in *Figure 8*(b) shows the residuals follow the slope of the line. There is some deviance in the tails, which is an effect of outliers that were not deemed to be influential enough to remove from the data set. The independence of error terms is evident in the sequence plot shown in *Figure 8*(d). There does not appear to be a pattern in the sequence plot of the residuals. With all model assumptions satisfied, the final model is confirmed.

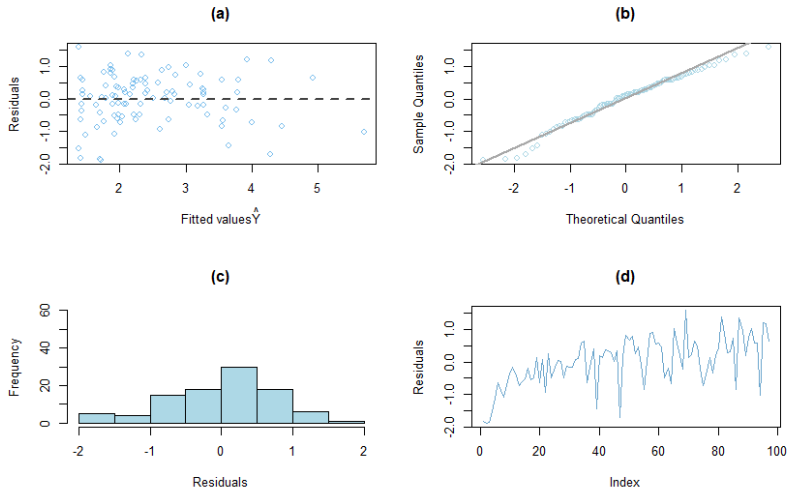

Figure 8: Residuals Diagnostics

Goodness of Fit Test

In this study, tables, scatter plots, linear regression models, Pearson's coefficient determination r, R^2, adjusted R^2, Mallow's C_p, BIC, and the student's t-test were used to test the goodness of fit. All analyses were performed using R version 4.0.0 (2020-04-24).[7]

RESULTS

All the covariate variables have been tested individually to determine if a linear association exists on PSA level using the two-tailed t-test. The t-test considers the following hypothesis:

*__Null hypothesis__: H_0: $\beta_1 = 0$

*__Alternative hypothesis__: H_1: $\beta_1 \neq 0$.

One-sample t-test

We can use this statistic in testing the null hypothesis that the population mean is equal to a specified value μ_0.

The decision is taken considering $t* = \frac{b_1 - \beta_1}{\widehat{SE}\ (b_1)}$,

where

$t*$ = test-statistics for the t-test
b_1 = observed slope coefficient
β_1 = expected slope coefficient of the fitted regression model
$\widehat{SE}\ (b_1)$ = sampling variability of b_1

The calculated t-value is compared to the critical t-value from the t-distribution table with degrees of freedom df = n - 1 and the chosen confidence level. If the calculated t-value is greater than the critical t-value, then we reject the null hypothesis. A coefficient of determination (R^2) close to 1 indicates there is a strong association between X and Y variables, whereas a value close to 0 indicates a weak association between response and predictor variables. A weak association indicates there is no goodness of fit.

Analysis of cancer volume association with PSA level

At significance level α = 0.05, the result of the t-test is to reject H_0, thus concluding that there exists evidence of a linear association between cancer volume and PSA. The p-value seems in agreement with the decision and the coefficient of determination (R^2) shows a positive association. The t-test also illustrates that the model, in equation (1), can explain 39.0% of the unexplained variation in PSA, indicating that the model is a good fit compared to other predictors.

Analysis of prostate weight association with PSA level

At significance level α = 0.05, the result of the t-test is to accept H_0, thus concluding that no evidence exists for a linear association between prostate weight and PSA. The p-value is greater than 0.05 and the coefficient of determination (R^2) value is nearly equal to 0. This indicates no good association exists between response and predictor variable. The t-test also demonstrates that the model, shown in equation (2), can account for only 0.06% of the unexplained variation in PSA fitted by the model. Explained variation is 99.94%, indicating that the model is not a good fit compared to other predictors.

Similar Analysis was done for age, benign prostatic hyperplasia, seminal vesicle invasion, capsular penetration, Gleason score.

	Null Hypothesis (H_0)	Unexplained/Explained Variation	Effect
AGE	Accept	0.02% / 99.98%	None
BENIGN PROSTATIC HYPERPLASIA	Accept	0.02% / 99.98%	None
SEMINAL VESICLE INVASION	Reject	27.9% / 72.1%	Good fit
CAPSULAR PENETRATION	Accept	30.3% / 69.7%	None
GLEASON SCORE	Accept	18.4% / 81.6%	None

Primary Objective Results

The t-test results show the positive association between cancer volume and seminal vesicle invasion. The other positive predictors such as scores are added to the model. Adding age and hyperplasia does not significantly change the results. The strongest or positive association can be decided based on the (R^2). Thus, the final model consists of cancer volume, seminal vesicle invasion, score, age, and hyperplasia.

CONCLUSION

The estimated regression function from this data analysis is as follows:

$$\text{Log}(\widehat{Y_i}) = -0.5355 + 0.0647 * X_1 + 0.3387 * X_2 - 0.0024 * X_3 + 0.0935 * X_4 + 0.6895 * X_5 + \varepsilon$$

Where,

Log(Y_i)= log of prostate-specific antigen level

X_1= Cancer volume

X_2= Gleason Score

X_3= Age

X_4= Benign prostatic hyperplasia

X_5= Seminal vesicle invasion

ε_i is the error term; $\varepsilon_i \sim iidN(0, \sigma^2)$

i = 1,2,3, · · ·, 95.

The study shows that cancer volume (cancerv), prostate weight (weight), age (age), benign prostatic hyperplasia (hyperplasia), and seminal vesicle invasion (seminal) all affect the logarithm of prostate-specific antigen (PSA) level. All statistical analysis was conducted at a 95% confidence interval and 0.05 significance level. The average ambient variables were explored individually, and the two-tailed t-test was conducted on each of the fitted linear

regression models. *Table* 7 shows the estimated regression coefficient, the standard error, t-value, and p-value associated with each of the predictors R_a^2, R^2, MSE, and F statistics of the final model. After testing and exploring the test statistics, the strongest association among all predictors and the response was found. The study and model show that cancer volume, seminal vesicle invasion, and age were the major factors affecting prostate-specific antigen levels. It was further inferred that the Gleason Score and benign prostatic hyperplasia were also affecting prostate-specific antigen level, but their effects were not statistically significant.

One limitation of this study is the small data set, which consists only of 97 data points. The model should be validated with another data set and it would be preferred to have a bigger data set to create the best model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.536	0.985	-0.54	0.5881
CANCERV	0.065	0.013	5.00	0.0000 ***
SEMINAL1	0.689	0.239	2.88	0.0049 **
SCORE	0.3386	0.126	2.68	0.0087 **
HYPERPLASIA	0.0935	0.028	3.32	0.0013**
AGE	-0.0024	0.012	-0.21	0.8367
Observation	97			
R^2	0.584			
Adjusted R^2	0.561			
Residual Std. Error	0.765 (df=91)			
F Statistic	25.51 (df=5; 91)			

Table 7: Statistics for the regression model

APPENDIX: R CODE

```
# Association of serum prostate-specific antigen
# level and prognostic clinical measurements in men
# with advanced prostate cancer ##########
## Load data into R
setwd("/Users/jay/Documents ")
CCPP <- read_xlsx("pcancer.xlsx", sheet = 1)
attach(CCPP)
CCPP$seminal <- as.factor(CCPP$seminal)
summary(CCPP)
library(xtable)
kable(summary(CCPP))
str(CCPP)
mod <- CCPP[, -c(1, 7)]
tab <- round(basicStats(mod), 3)
kable(tab[-c(1, 2, 8, 9, 10, 11, 12, 15, 16), ], caption = "Basic Statistics for Data Set Variables") %>%
   kable_styling(bootstrap_options = c("striped"))
# Create boxplots and histograms for each variable
par(mfrow = c(2, 2))
boxplot(cancerv, data = CCPP, horizontal = TRUE, xlab = "(a)CANCER VOLUME")
boxplot(weight, data = CCPP, horizontal = TRUE, xlab = "(b)WEIGHT")
boxplot(age, data = CCPP, horizontal = TRUE, xlab = "(c)AGE")
boxplot(hyperplasia, data = CCPP, horizontal = TRUE,
  xlab = "(d)BEGNIN PROSTATIC HYPERPLASIA")

hist(seminal, data = CCPP, xlab = "(e)SEMINAL VESICLE INVASION")
boxplot(capsular, data = CCPP, horizontal = TRUE, xlab = "(f)CAPSULAR PENETRATION")
hist(score, data = CCPP, horizontal = TRUE, xlab = "(g)GLEASON SCORE")
boxplot(psa, data = CCPP, horizontal = TRUE, xlab = "(h)PSA LEVEL")

# Analyzing PSA level
par(mfrow = c(2, 2))
plot(density(psa), col = "darkblue", lwd = 2, main = "(a)")
hist(psa, col = "royalblue", main = "(b)")
boxplot(psa, data = CCPP, horizontal = TRUE, xlab = "PSA level")
## Dependency of PSA level on cancer volume Fitting
## into linear regression model Y_i = beta_0 +
## beta_1 * X + epsilon_i
lm.cancerv <- lm(psa ~ cancerv)
plot(psa ~ cancerv, col = "darkseagreen", xlab = expression("Cancer volume"),
   ylab = "PSA level")
abline(lm.cancerv, col = "darkblue", lty = 1, lwd = 2)

# Dependency of PSA level on weight fitting into
# linear regression model Y_i = beta_0 + beta_1 * X
# + epsilon_i
```

```
lm.weight <- lm(psa ~ weight)
plot(psa ~ weight, col = "darkseagreen", xlab = "Weight",
   ylab = "PSA level")
abline(lm.weight, col = "darkblue", lty = 1, lwd = 2)
## Dependency of PSA level on age fitting into
## linear regression model Y_i = beta_0 + beta_1 * X
## + epsilon_i

lm.age <- lm(psa ~ age)
plot(psa ~ age, col = "darkseagreen", xlab = "AGE",
   ylab = "PSA level")
abline(lm.age, col = "darkblue", lty = 1, lwd = 2)
## Dependency of PSA level on benign prostatic
## hyperplasia fitting into linear regression model
## Y_i = beta_0 + beta_1 * X + epsilon_i
lm.hyperplasia <- lm(psa ~ hyperplasia)
plot(psa ~ hyperplasia, col = "darkseagreen", xlab = "BENIGN PROSTATIC HYPERPLASIA",
   ylab = "PSA level")
abline(lm.hyperplasia, col = "darkblue", lty = 1, lwd = 2)

## Dependency of PSA level on seminal vesicle
## invasion fitting into linear regression model Y_i
## = beta_0 + beta_1 * X + epsilon_i
lm.seminal <- lm(psa ~ as.factor(seminal))
plot(psa ~ seminal, col = "darkseagreen", xlab = "SEMINAL VESICLE INVASION",
   ylab = "PSA level")
abline(lm.seminal, col = "darkblue", lty = 1, lwd = 2)

## Dependency of PSA level on capsular penetration
## fitting into linear regression model Y_i = beta_0
## + beta_1 * X + epsilon_i
lm.capsular <- lm(psa ~ capsular)
plot(psa ~ capsular, col = "darkseagreen", xlab = "CAPSULAR PENETRATION",
   ylab = "PSA level")
abline(lm.capsular, col = "darkblue", lty = 1, lwd = 2)
## Dependency of PSA level on Gleason score fitting
## into linear regression model Y_i = beta_0 +
## beta_1 * X + epsilon_i
lm.score <- lm(psa ~ score)
plot(psa ~ score, col = "darkseagreen", xlab = "Gleason score",
   ylab = "PSA level")
abline(lm.score, col = "darkblue", lty = 1, lwd = 2)
# Summary statistics of seven separate models :
# outcome and each of the predictors
summary(lm.cancerv)
```

```
summary(lm.weight)
summary(lm.age)
summary(lm.hyperplasia)
summary(lm.seminal)
summary(lm.capsular)
summary(lm.score)
# Multicollinearity between predictors
pairs.panels(CCPP, ellipses = FALSE, density = FALSE)

# cor(CCPP)
mod <- CCPP[, -c(1, 7)]
tab <- round(basicStats(mod), 3)
tab2 <- cor(mod)
kable(tab2, caption = "Correlation Matrix for the Data Set",
   digits = 3)
# Checking the Variance Inflation Factors (V.I.F)
lm.all <- lm(psa ~ cancerv + seminal + score + hyperplasia +
   age + capsular + weight, data = CCPP)
summary(lm.all)
tab4 <- vif(lm.all)
t(tab4)
kable(t(tab4), caption = "VIF Values from the Final Model",
   digits = 3)
# Studentized Residuals
par(mfrow = c(2, 2))
plot(fitted(lm.all), rstudent(lm.all), col = "darkseagreen",
   pch = 16, xlab = expression(hat(Y)), ylab = "Studentized residuals",
   ylim = c(-5, 5), 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(lm.all)
plot(cooksd, pch = "*", main = "(b) Influential Obs by Cooks distance",
   xlab = "Index", ylab = "Cooks distance")
text(x = 1:length(cooksd) + 1, y = cooksd, labels = ifelse(cooksd >
   0.01, names(cooksd), ""), col = "blue")

# Model analysis

lm.all <- lm(psa ~ cancerv + seminal + score + hyperplasia +
   age + capsular + weight, data = CCPP)
summary(lm.all)
```

```
# Normality check
par(mfrow = c(2, 2))

plot(lm.all)

# Model selection
asm <- regsubsets(psa ~ cancerv + seminal + score +
   hyperplasia + age + capsular + weight, data = CCPP)
rs <- summary(asm)
names(rs)
rs$cp
rs$bic
rs$adjr2

## Analyzing LM with all the variables to see who has the
## most dependency
model <- lm(psa ~ cancerv + seminal + score + hyperplasia +
   age, data = CCPP)
summary(model)
# Homoscedasticity
plot(fitted(model), residuals(model))
abline(h = 0)  #Homoscedasticity is not met
par(mfrow = c(2, 2))

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

#### Normality check
qqnorm(residuals(model), col = "lightblue", main = "(b)")
qqline(residuals(model), col = "darkgray", lwd = 2)
hist(residuals(model), ylim = c(0, 60), col = "lightblue",
   main = "(c)", xlab = "Residuals") #normality not met

# Independence of error terms
plot(residuals(model), type = "l", col = "skyblue3",
   main = "(d)", ylab = "Residuals")

qqnorm(residuals(model))
qqline(residuals(model))

# Normality and linearity not met: need transformation
CCPP$psa<-log(CCPP$psa)
modelb <- lm(log(psa) ~ cancerv + seminal + score + hyperplasia +
   age, data = CCPP)
```

```
mod<-CCPP[,-c(1,7)]
tab2<-cor(mod)
kable(tab2,caption = "Correlation Matrix for the Dataset", digits = 3)
# Recheck homoscedasticity
plot(fitted(modelb), residuals(modelb))
abline(h=0)
# Recheck normality
qqnorm(residuals(modelb))
qqline(residuals(modelb))
#Recheck independence
plot(residuals(modelb))
abline(h=0)

lm.all <- lm(log(psa) ~ cancerv + seminal + score + hyperplasia +
age + capsular + weight   , data = pcancer)
summary(lm.all)
tab4<-vif(lm.all)
t(tab4)
#Added Variable Plots
crPlots(lm.all, col = "pink", col.lines = c("darkred",
                                "blue"), ylab="βX+Residuals")

# Model with outliers
model <- lm(log(psa) ~ cancerv + seminal + score + hyperplasia +
   age, data = CCPP)
summary(model)

# Model without outliers
model1 <- lm(log(psa) ~ cancerv + seminal + score + hyperplasia +
   age, data = CCPP, subset = -c(32))
summary(model1)
```

REFERENCES

[1] Dean, A., Voss, D., and Draguljić, D. (2017), *Design and Analysis of Experiments*, 2nd ed., Springer.

[2] Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2014), *Applied Linear Statistical Models,* 5th ed., McGraw-Hill Irwin.

[3] Advances in Prostate Cancer Research. Available online. https://www.cancer.gov/types/prostate/research

[4] Screening Tests for Prostate Cancer. Available online. https://www.cancer.org/cancer/prostate-cancer/detection-diagnosis-staging/tests.html#:~:text=About%2015%25%20of%20men%20with,prostate%20cancer%20is%20over%2050%25.

[5] Prostate Cancer—Patient Version. Available online. https://www.cancer.gov/types/prostate

[6] Gleason Score: Prostate Cancer Grading & Prognostic Scoring. Available online. https://www.prostateconditions.org/about-prostate-conditions/prostate-cancer/newly-diagnosed/gleason-score

[7] R Core Team (2018), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/

ACKNOWLEDGMENTS

We would like to acknowledge our Sr Biostatistician Peigen Zhou and Programming Manager Avani Kaja and Programming Director Shefalica Chand for reviewing our paper and providing valuable feedback.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Girish Kankipati
Seattle Genetics, Inc.
21823 - 30th Drive S.E.
Bothell, WA 98021
425-527-2140
gkankipati@seagen.com

Jai Deep Mittapalli
Seattle Genetics, Inc.
21823 - 30th Drive S.E.
Bothell, WA 98021
650-273-1854
jamittapalli@seagen.com