# *Detection of Prostate Cancer Using Linear Regression Model*

## STAT 840: Fall Class Project, 2019

**Bibekananda Mishra**


Department of Biostatistics and Data Science
University of Kansas, USA
December 11, 2019

# Contents

# Abstract

A university medical center urology group was interested in the association between prostate-specific antigen (PSA level) is the response variable and a number of prognostic clinical measurements in men with advanced prostate cancer. Data were collected on 97 men who were about to undergo radical prostectomies. We fit the data in to various statistical models and choose the best among them. Our analysis shows that 'Estimate of Cancer volume' is the most important paarmeter. However, the final model depends upon all the measurements considered in the data in deciding PSA level. The dependency though happens where we are taking into consideration logarithm of PSA level and logarithm of volume of cancer.

# Introduction

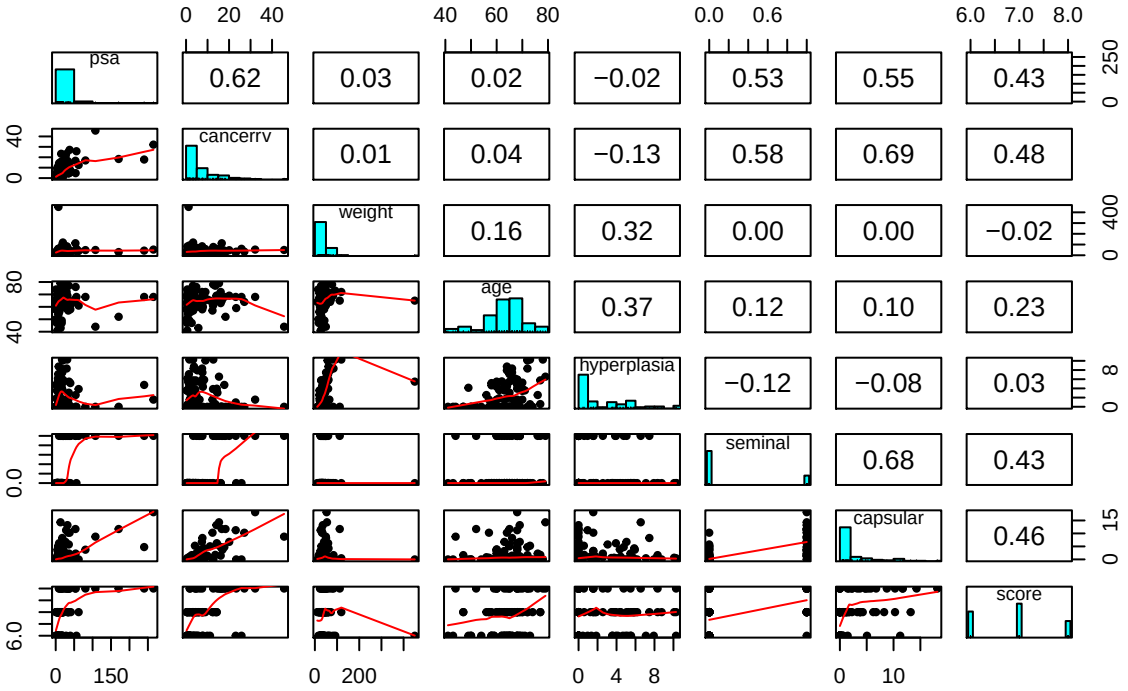
A university medical center urology group collected data on nine measurements from 97 patients who were about to undergo radical prostectomies. The measurements are as follows: prostate-specific antigen level (PSA level), cancer volume, prostate weight, age, amount of benign prostatic hyperplasia, presence of seminal vesicle invasion, degree of capsular penetration and Gleason score. They were interested in finding the association of PSA level with any of the prognostic clinical measurements mentioned above. We fit the data in to various linear regression models, verified the asssumptions involved in the regression, detected and removed outliers and then chose the best model after refitting the data. Our initial analysis shows that 'Cancer volume' and 'Seminal score' are the best indicators in finding PSA level. However, the general model involves log(PSA) and log(cancerrv) apart from the other variables.

The data was presented in the book "Applied Linear Statistical Models" by Kurtner et al. We have referred the book as well as the lecture notes by Prof. Lazarus Mramba for preparing this report.

## Primary Analysis Objectives

First we would like to know whether there is any significant relationship between any two variables/measurements considered in the study. For this, we compute the correlation coffiecients of the associated paarmeters in the analysis and plot the scatter plot for each variable vs another variable. See figure-1.

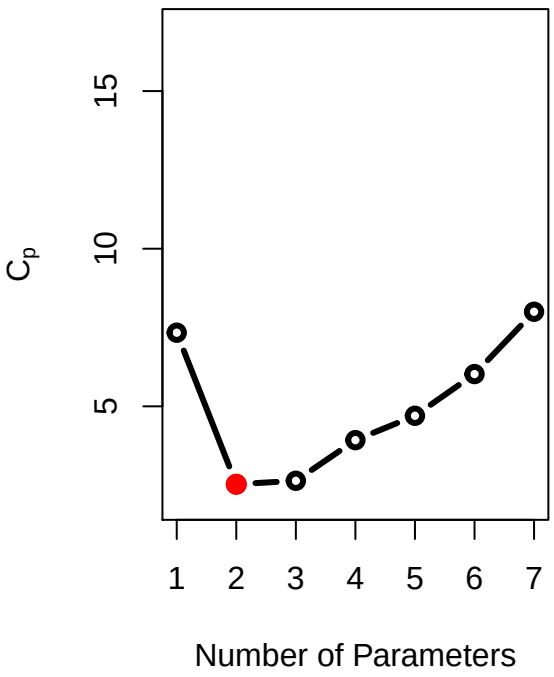
## Figure−1



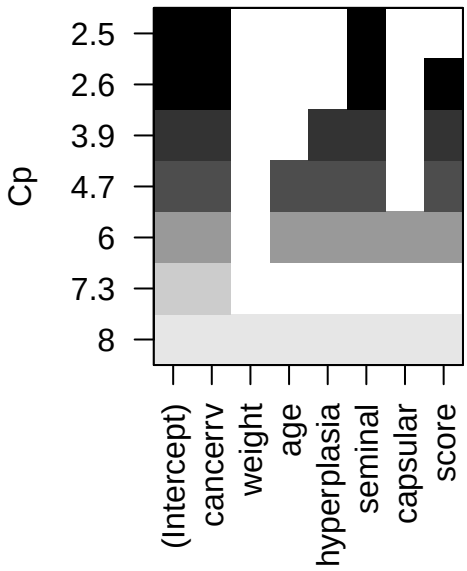

From the correlation cofficients, it is apparent that there is no much relationship between the variables except probably between Cancer volume (cancerrv) vs Capsular penetration (capsular). The scatter plot shows the same conclusion.

Next we compute the Cp value for each model so as to choose the least biased or unbiasedmodel. We have plotted the graph. See figure-2 (a) and (b).

**Figure−2(a) Cp variable Selectio**



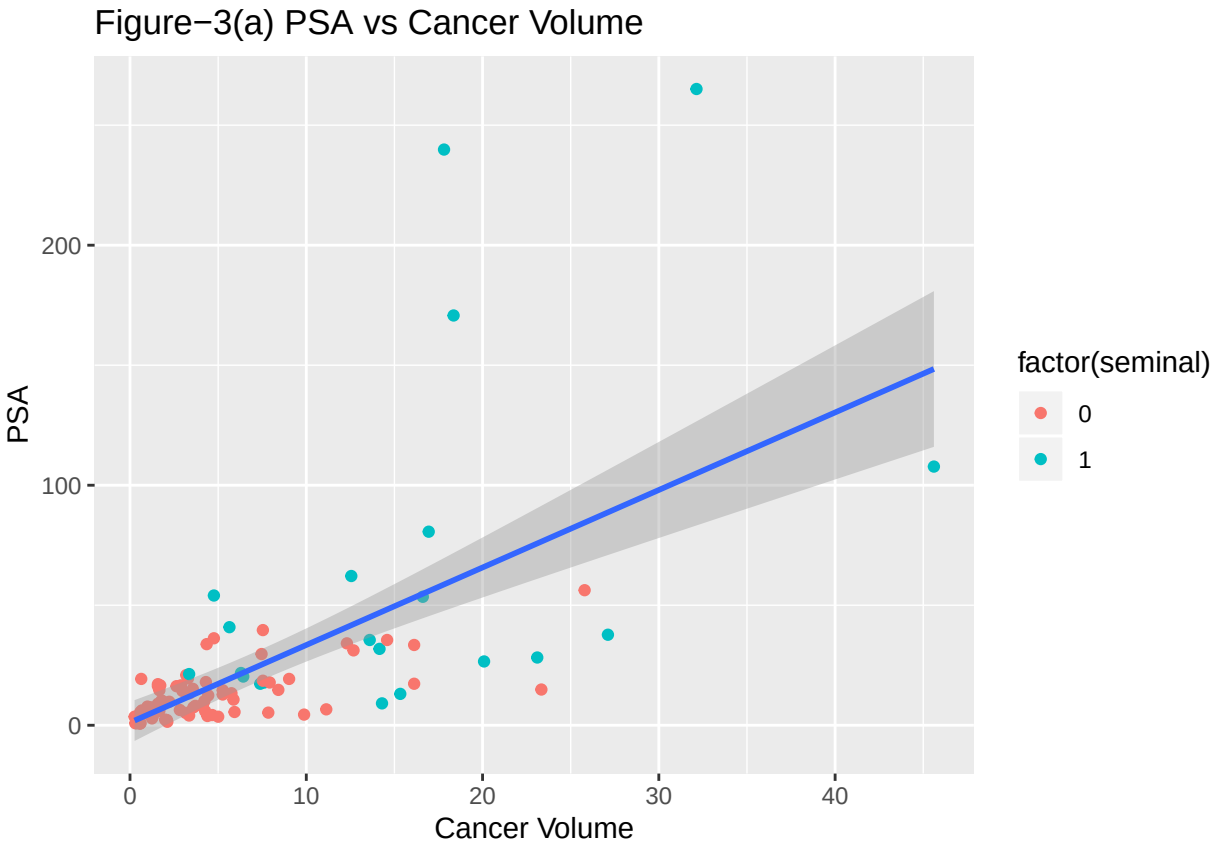

**Figure−2(b)**

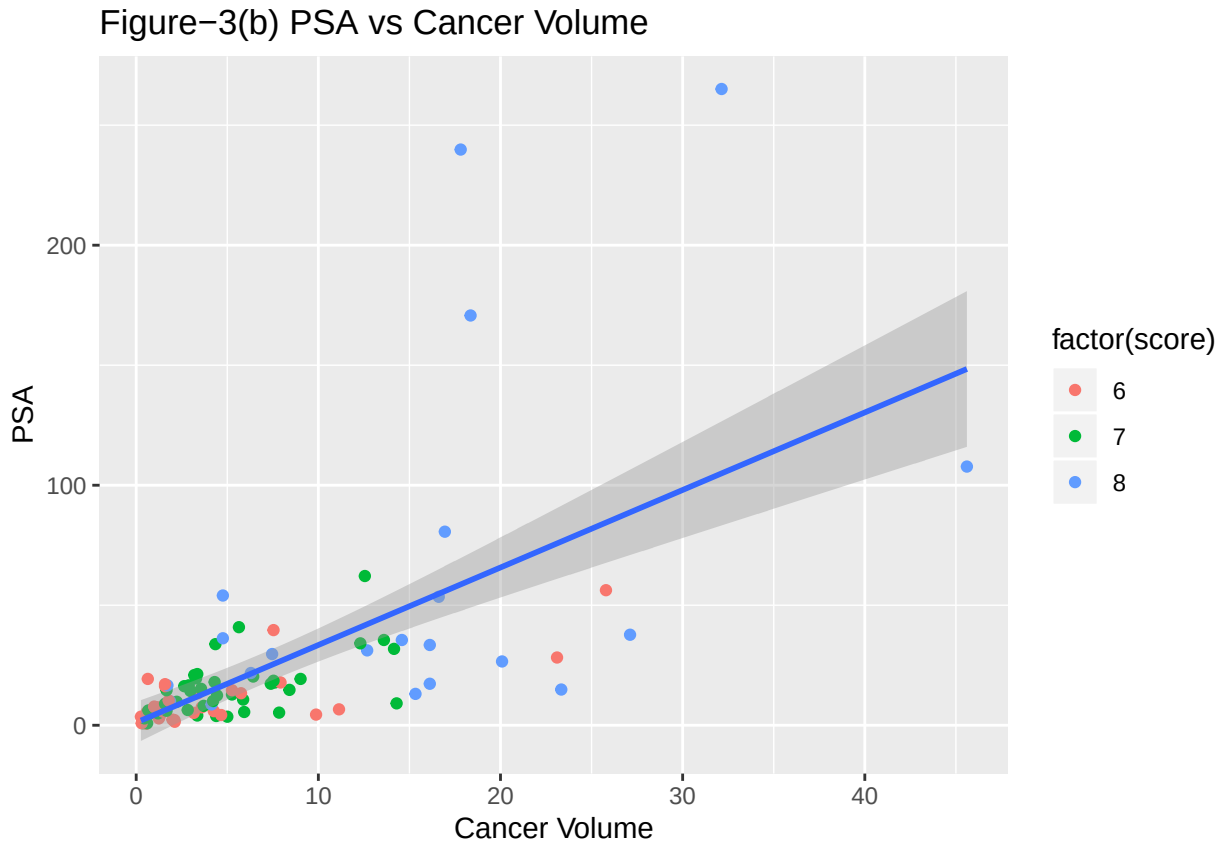



Seeing the plot, we choose the models having Cp values closest to (p+1) so that the resultant models are least biased. We set two models: [1] Model2 having two variables, cancerrv and seminal [2] Model3 having three variables: cancerrv, seminal and score.

We will also consider the full model i.e. Model6 having all the variables.

See the following plots to notice the relationship between cancerrv and psa. The cases in the two plots have been marked with different colors according to seminal and score values respectively.


Figure−3(a) PSA vs Cancer Volume
PSA
200
100
0
0
10
20
30
40
Cancer Volume
factor(seminal)
0
1




It is clear that there is a linear relationship between cancerrv and PSA level. We will test the linearity hypotheis of both the models and compare them with the full models. The F-values for model-2, model-3 and model-6 are as follows:

Model 2: mod2<-lm(psa~cancerrv+seminal, pcancer)
Model 3: mod3<-lm(psa~cancerrv+seminal+score, pcancer)
Model 6: mod6<-lm(psa~., pcancer)

F-values respectively: 3.098e-12, .3063e-12, and .3063e-12.

Since all the values are less than 0.05, we conclude that the models we are considering are the better ones compared to the null hypothesis as stated in the tests above.

## Secondary Analysis Objectives

Next, we verify the assumptions associated with linear regression for each of these three models and check whether there are any outliers in the data.

a] We have plotted sequential plot of the residuals and noted that the terms are broadly independent.

b] We plot the residuals aganist cancerrv to note that the plot is funnel-shaped. The absolute value of residuals plotted against cancerrv also shows that the absolute value

increases as cancerrv increases. So, we have to take some remedial measures. We will consider log(psa) instead of psa and model.

c] The normality plot of residuals shows that there are few cases (namely 95, 96 and 97) which diverge from normality assumption. So we may have to remove these cases.

d] Looking at the residual plots, normal probablity plots, Cook's distance plot and Bonferroni test for each linear model, we find that the cases 96 and 97 are in the outliers.

So, we have to remove those two outlier cases and build model again based on the modified data. We apply our regression model to log(psa) and also experiment by adding nonlinear term cancerrv*seminal. So our new models are as follows:

New model 2: modn2<-lm(log(psa)~log(cancerrv)+seminal, pcancer)
New model 3: modn3<-lm(log(psa)~log(cancerrv)+seminal+score, pcancer)
New Model 6: modn6<-lm(log(psa)~log(cancerrv)+.., pcancer)

Comparing the the multiple r-squared values for the three models above, we choose the model having highest r-squared value i.e. model which explains the data variance the most. See the table below. Moreover, for comparing the models, we use anova test. See the table below. Note that the two entries in anova column denote the F-values of 'modn2 vs modn3' and modn3 vs modn6' respectively.

## Table comparing various models

| Model | Num. of predictors | RSE | R-Squared | ANOVA_Fvalue |
|---|---|---|---|---|
| modn2 | 2 | 0.73 | 0.55 | 0.000 |
| modn3 | 3 | 0.72 | 0.57 | 0.039 |
| modn6 | 7 | 0.68 | 0.63 | 0.018 |

Since clearly RSE decreases, R-squared value increases and F-values are less than 0.05, model modn6 is the winner here.

# Results

Our final linear model for this data is:

**Table 2:** Table of model involving log(cancerrv), weight, age, hyperplasia, seminal capsular and score.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 0.46 | 0.92 | 0.50 | 0.62 |
| log(cancerrv) | 0.55 | 0.08 | 6.72 | 0.00 |
| seminal | 0.74 | 0.25 | 3.00 | 0.00 |
| score | 0.29 | 0.12 | 2.48 | 0.01 |
| age | -0.02 | 0.01 | -1.56 | 0.12 |
| weight | 0.00 | 0.00 | 1.39 | 0.17 |
| hyperplasia | 0.06 | 0.03 | 2.37 | 0.02 |
| capsular | -0.05 | 0.03 | -1.50 | 0.14 |

**Table 3:** ANOVA for the model

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| log(cancerrv) | 1 | 56.343 | 56.343 | 120.145 | 0.000 |
| seminal | 1 | 3.404 | 3.404 | 7.259 | 0.008 |
| score | 1 | 2.264 | 2.264 | 4.828 | 0.031 |
| age | 1 | 0.044 | 0.044 | 0.095 | 0.759 |
| weight | 1 | 2.153 | 2.153 | 4.591 | 0.035 |
| hyperplasia | 1 | 2.681 | 2.681 | 5.717 | 0.019 |
| capsular | 1 | 1.061 | 1.061 | 2.263 | 0.136 |
| Residuals | 87 | 40.799 | 0.469 | NA | NA |

# Discussion and Conclusion

Our analysis shows that the PSA leve depends upon all the parameters considered in the experiment, albeiet in a 'logarithmic' way. Since it only explains 63% of the variation in the data, we believe we can improve this by playing some more with the paaremeters by introducing nonlinear interaction terms.

# Appendix: R-code

\begin{lstlisting}

title: “lr_project”
author: “Bibek”
date: “11/30/2019”
output:
pdf_document: default
html_document: default

library(ggplot2)
library(dplyr)
library(readr)
library(choroplethr) library(tinytex) library(lmtest) library(car) library(olsrr)
library(psych) library(faraway) library(leaps)

library(data.table) require(epiDisplay)
pcancer<-read.table(“http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerData/Appendix%20C%20Data%20Sets/APPENC05.txt”, quote=“"”, comment.char="")

pcancer$V1<-NULL

setnames(pcancer, old=c(“V2”,“V3”,“V4”,“V5”,“V6”,“V7”,“V8”,“V9”), new=c(“psa”, “cancerrv”,“weight”,“age”,“hyperplasia”,“seminal”,“capsular”,“score”)) head(pcancer)

pairs.panels(pcancer, density = FALSE, ellipses = FALSE, main = “Scatterplot”) mod0 <- regsubsets(psa ~ ., data = pcancer)

summary(mod0)

summary(mod0)$cp

cp_min <- which.min(summary(mod0)$cp)

par(mfrow = c(1, 2)) plot(1:7, summary(mod0)$cp, type = "b", ylab = expression(C[p]), ylim = c(2, 17), lwd = 3, main = "CpvariableSelection", xlab = "NumberofParameters")points($cp_min$, summary(mod0)cp[cp_min], col = “red”, cex = 2, pch = 20) plot(mod0, scale = “Cp”)

mod2<-lm(psa~cancerrv+seminal, pcancer)

mod3<-lm(psa~cancerrv+seminal+score, pcancer)

mod6<-lm(psa~., pcancer)

linearHypothesis(mod2, c(“seminal=0”, “cancerrv=0”))

linearHypothesis(mod3, c(“seminal=0”, “cancerrv=0”, “score=0”))

linearHypothesis(mod6, c(“seminal=0”, “capsular=0”))

ggplot(pcancer, aes(y= psa, x=cancerrv))+ geom_point(aes(color = factor(seminal))) +labs(x= “Cancer Volume”, y=“PSA”, title = “PSA vs Cancer Volume”)+ geom_smooth(method = “lm”)

ggplot(pcancer, aes(y = psa, x = cancerrv)) + geom_point(aes(color = factor(score)))+labs(x= “Cancer Volume”, y=“PSA”, title = “PSA vs Cancer Volume”)+ geom_smooth(method = “lm”)

#assumption verifications for model-2 ggplot(pcancer, aes(y = residuals(mod2), x = cancerrv)) + geom_point(aes(color = factor(seminal)))+labs(x= “Cancer Volume”, y=“Residuals(model-2)”, title = “Residuals vs Cancer Volume”)+ geom_smooth(method = “lm”)

plot(mod2)

halfnorm(cooks.distance(mod2))

#assumption verifications for model-3 #Verifying independence of error terms plot(residuals(mod3))+labs(title=“Sequence plot of residuals”) #verifying normaility assumption cancer.stdres3<-rstandard(mod3)

qqnorm(cancer.stdres3, ylab=“Sample quantiles”, xlab=“Normal Quantiles”, main=“Normal probability plot of residuals”)

qqline(cancer.stdres3)

#detection of outlying observations by plotting studentised residual and deleted residuals bptest(mod3, ~psa, data = pcancer, studentize = TRUE)

bptest(mod3, ~psa, data = pcancer, studentize = FALSE)

qqPlot(mod3, main=“QQ Plot”)

ols_plot_resid_stud_fit(mod3)

ggplot(pcancer, aes(y = residuals(mod3), x = cancerrv)) + geom_point(aes(color = factor(seminal)))+labs(x= “Cancer Volume”, y=“Residuals(model-3)”, title = “Residuals vs Cancer Volume”)+ geom_smooth(method = “lm”)

plot(mod3)

halfnorm(cooks.distance(mod3))

#assumption verifications for model-6 ggplot(pcancer, aes(y = residuals(mod6), x = cancerrv)) + geom_point(aes(color = factor(seminal)))+labs(x= “Cancer Volume”, y=“Residuals(model-6)”, title = “Residuals vs Cancer Volume”)+ geom_smooth(method = “lm”)

plot(mod6)

halfnorm(cooks.distance(mod6))

#verifying presence of multicolinearity vif(mod2)

#F-value comparing model 2 and model 3 anova(modn2,modn3)[2,6] #F-value comparing model 3 and model 6 anova(modn3,modn6)[2,6]

pcancer_new<-pcancer[1:95,] modn2<-lm(psa~cancerrv+seminal, pcancer_new) modn3<-lm(psa~cancerrv+seminal+score, pcancer_new) modn6<-lm(psa~., pcancer_new)

\end{lstlisting}