MiniProj5

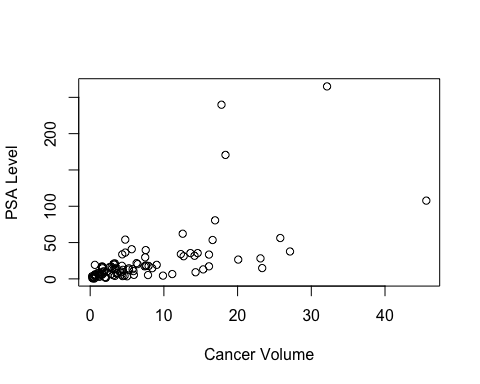
Prashant Prakash

November 29, 2015

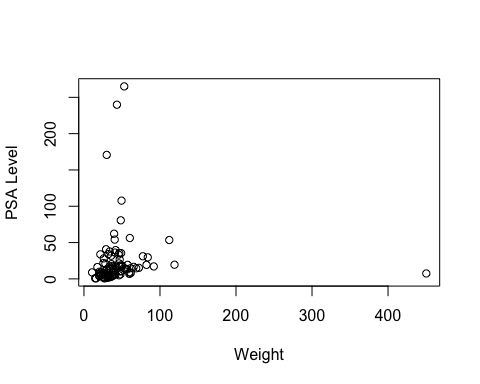
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

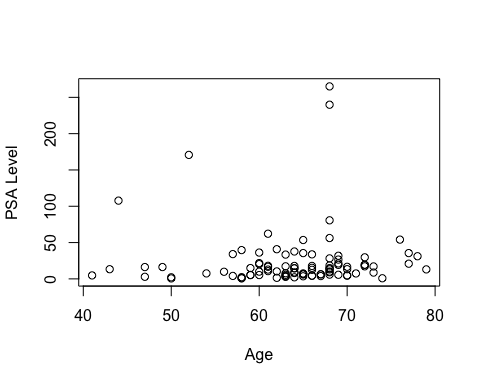
# read data from csv file   
data = read.csv("/Users/Prashant/Downloads/prostate\_cancer.csv",header = TRUE , sep=",")  
  
# Plot scatter plot of PSA level with other variables   
  
# with Cancer volume   
plot(data$cancervol, data$psa, xlab="Cancer Volume", ylab="PSA Level")



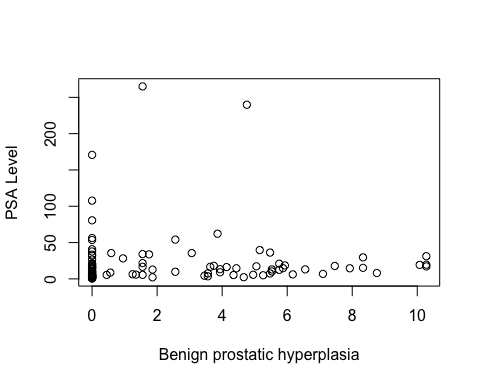
# with weight   
plot(data$weight, data$psa, xlab="Weight", ylab="PSA Level")



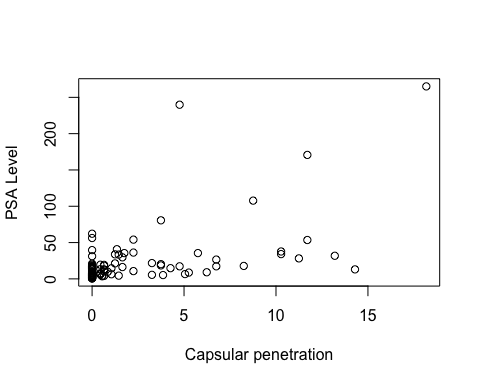
# with Age  
plot(data$age, data$psa, xlab="Age", ylab="PSA Level")



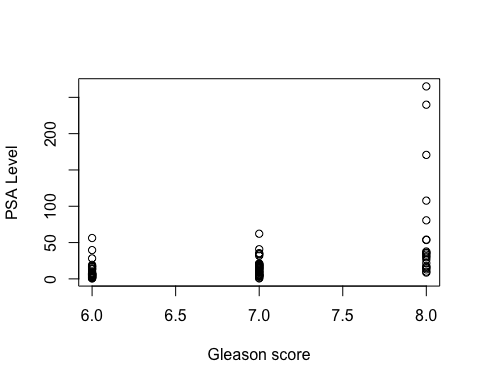
# with Benign prostatic hyperplasia  
plot(data$benpros, data$psa, xlab="Benign prostatic hyperplasia", ylab="PSA Level")



# with Capsular penetration  
plot(data$capspen, data$psa, xlab="Capsular penetration", ylab="PSA Level")



# with Gleason score its a categorical variable is this required ???  
plot(data$gleason, data$psa, xlab="Gleason score", ylab="PSA Level")



# capturing correaltion coefficient to decide which varible is good for deciding PSA level   
  
cor(data$psa,data$cancervol ,use ="everything", method =c("pearson"))

## [1] 0.6241506

cor(data$psa,data$weight ,use ="everything", method =c("pearson"))

## [1] 0.02621343

cor(data$psa,data$age ,use ="everything", method =c("pearson"))

## [1] 0.01719938

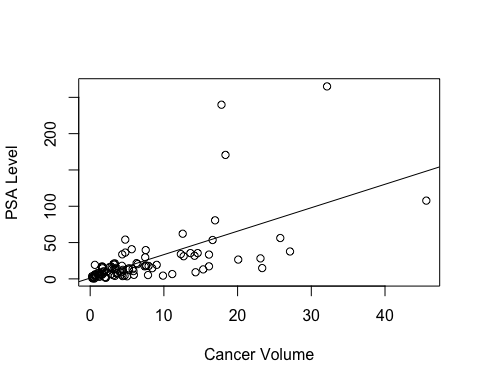
cor(data$psa,data$benpros ,use ="everything", method =c("pearson"))

## [1] -0.01648649

# clearly using correleation it's obvious Cancer volume is the candidate to do a better regression for PSA level   
  
x <- data$cancervol  
y <- data$psa  
  
(psa.reg <- lm(y ~ x))

##   
## Call:  
## lm(formula = y ~ x)  
##   
## Coefficients:  
## (Intercept) x   
## 1.125 3.230

plot(data$cancervol, data$psa, xlab="Cancer Volume", ylab="PSA Level")  
abline(psa.reg)



confint(psa.reg)

## 2.5 % 97.5 %  
## (Intercept) -7.530000 9.779701  
## x 2.406407 4.053461

summary(psa.reg)

##   
## Call:  
## lm(formula = y ~ x)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -61.619 -9.023 -1.586 3.151 181.183   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.1249 4.3596 0.258 0.797   
## x 3.2299 0.4148 7.786 8.47e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 32.03 on 95 degrees of freedom  
## Multiple R-squared: 0.3896, Adjusted R-squared: 0.3831   
## F-statistic: 60.63 on 1 and 95 DF, p-value: 8.468e-12

x.new <- data.frame(x = median(data$cancervol))  
predict(psa.reg, newdata=x.new)

## 1   
## 14.89438

You can also embed plots, for example:

Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.