Mini Project #6

Group 2.

Group Members:

Manav Gupta [MXG220027]

Shalin Ronakkumar Kaji [SXK220263]

Contribution:

Shalin and Manav equally contributed in the project.

Question :

We are given a dataset of 97 men related to prostate tumour. Now we will develop a suitable Multivariate Regression model step-wise eliminating the insignificant predictors.

```
# STATS-Mini-Project-6

# Setting working directory to Proj-6 folder.

> setwd("C:/Users/Shalin Kaji/Desktop/UT-Dallas-Spr23/STATS-DS-Min.Chen/Mini-Proj-6")

> getwd()

> library(BSDA)

# Exploring our .csv file

> data <- read.csv("prostate_cancer.csv")
```

Observations: On looking at our data we can infer that we have 6 numerical predictors (CancerVol, Weight, Age, benpros, capspen, gleason) and we have 1 categorical predictor – vesinv. Hence, the number of predictors for our model is 7+1 = 8. (p). The categorical variable of vesinv is divided into two dummy variables.

```
> data$vesinv<-as.factor(data$vesinv)
```

This command explicitly states to R to treat vesinv as a categorical variable in our regression model.

Now we create our base regression model without eliminating any predictor variables. We do this, to understand which predictors might be more relevant than others and how we can practise dimensionality reduction for our model.

```
> model <- lm(psa ~ cancervol + weight + age + benpros + capspen + gleason, data = data)
> summary(model)
```

```
> summary(model)

Call:
lm(formula = psa ~ cancervol + weight + age + benpros + vesinv +
    capspen + gleason, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-61.330  -8.130  -0.014   6.324 167.436

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.24264   40.53932  -0.376 0.707814
cancervol     2.03225    0.59359   3.424 0.000936 ***
weight        0.01132    0.07395   0.153 0.878708
age          -0.53721    0.47588  -1.129 0.261977
benpros       1.29831    1.20168   1.080 0.282878
vesinv1      19.60957   10.89184   1.800 0.075187 .
capspen       1.09877    1.33377   0.824 0.412253
gleason       7.05922    5.19452   1.359 0.177589
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.17 on 89 degrees of freedom
Multiple R-squared:  0.4585,    Adjusted R-squared:  0.4159
F-statistic: 10.77 on 7 and 89 DF,  p-value: 9.266e-10
```

The above “model” is our base regression model where we haven’t yet eliminated any insignificant predictors. Based on its summary, the following observations can be made: -

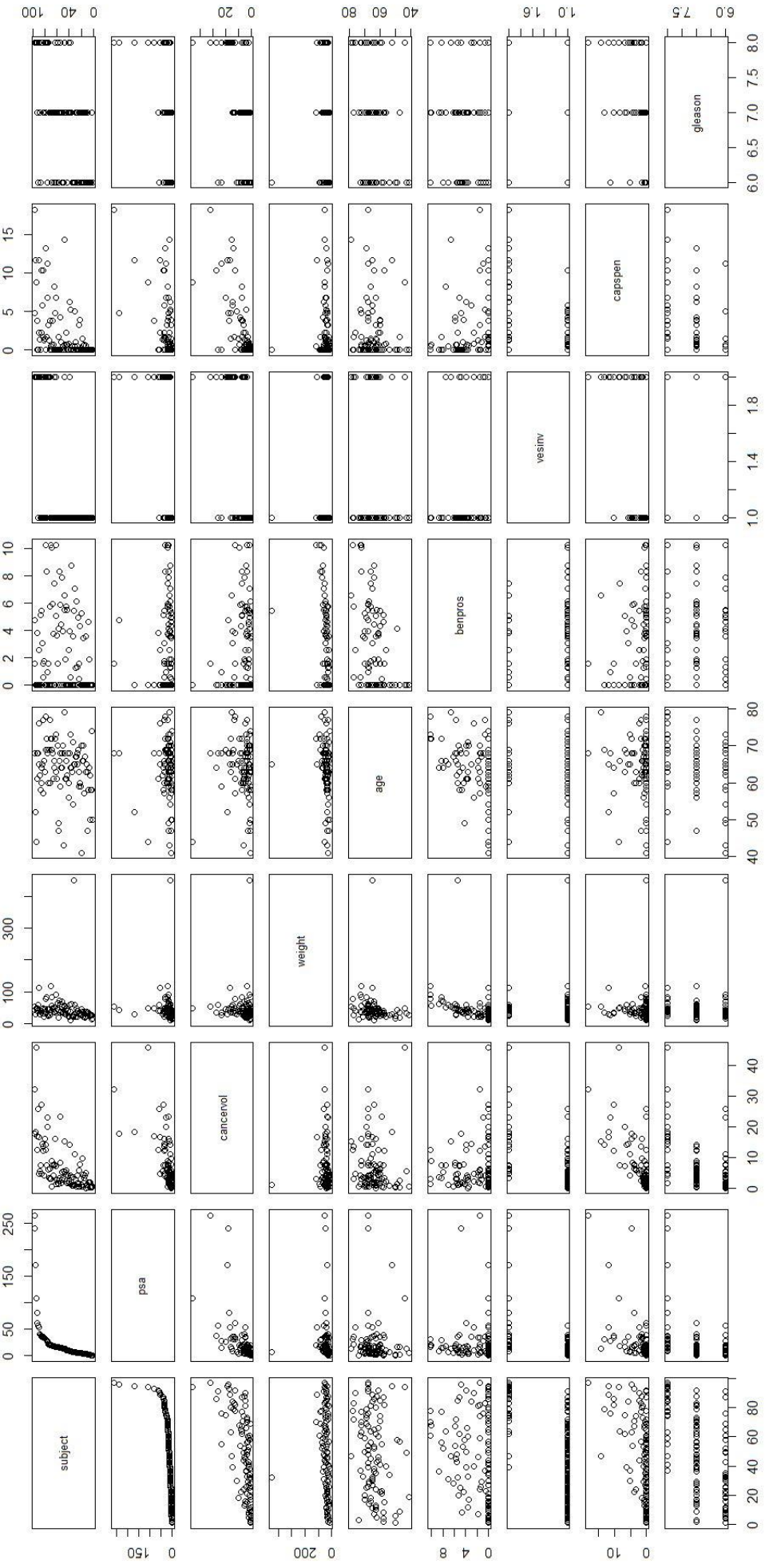
i) p-value = 9.266 x 10^-9 is very small which signifies that the given predictors are relevant to our model.

ii) most significant factor is cancervol (***), while vesinv1 is a significant one.

iii) R-squared (Coefficient of Determination): 0.4585 signifies that only 46% of variation in the data can be explained by the regression model.

First, we looked for any apparent association between any variables of the given data frame. The "plot" function of R and the original data frame provided in the question set are used for this.

```
> plot(data)
```


subject
psa
cancervol
weight
age
benpros
vesinv
capspen
gleason

Refinement Round 1:

We now try to refine our model by eliminating certain insignificant predictors and then will verify our assumptions using ANOVA function.

Predictor eliminated: Weight, reason: p-value too high (0.8)

```
> ModRef1 <- update (model, .~. -weight)
```

```
> anova(ModRef1,model)
```

```
> ModRef1 <- update (model, .~. -weight)
> anova(ModRef1,model)
Analysis of Variance Table

Model 1: psa ~ cancervol + age + benpros + vesinv + capspen + gleason
Model 2: psa ~ cancervol + weight + age + benpros + vesinv + capspen +
    gleason
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     90 86480
2     89 86457  1    22.754 0.0234 0.8787
>
```

As we can see that the p-value = 0.8787 which leads to accepting the Null Hypothesis H_O : weight is insignificant predictor assuming others remain.

Refinement Round 2:

We now try to refine our model by eliminating certain insignificant predictors and then will verify our assumptions using ANOVA function.

Predictor eliminated: Capspen, reason: p-value too high (0.4122)

```
> ModRef2 <- update (ModRef1, .~. -capspen)
```

```
> anova(ModRef2,ModRef1)
```

```
> # Refinement - 2
> ModRef2 <- update (ModRef1, .~. -capspen)
> anova(ModRef2,ModRef1)
Analysis of Variance Table

Model 1: psa ~ cancervol + age + benpros + vesinv + gleason
Model 2: psa ~ cancervol + age + benpros + vesinv + capspen + gleason
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     91 87138
2     90 86480  1    657.61 0.6844 0.4103
>
```

As we can see that the p-value = 0.4103 which leads to accepting the Null Hypothesis H_O: “capspen” is insignificant predictor assuming others remain.

Refinement Round 3:

After several trial and errors, we found that the following form of our regression model equation brought us to the multiple-R^2 = 0.6415.

```
> ModRef3 <- lm(log(psa) ~ log(cancervol)+log(age)+vesinv+benpros*gleason, data = data)
```

```
> summary(ModRef3)
```

```
> summary(ModRef3)

Call:
lm(formula = log(psa) ~ log(cancervol) + log(age) + vesinv + 
    benpros * gleason, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.53156 -0.33646  0.03166  0.47284  2.13368 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      2.40495    2.74057   0.878   0.3825    
log(cancervol)   0.51773    0.07973   6.494 4.49e-09 ***
log(age)        -0.83021    0.64480  -1.288   0.2012    
vesinv1          0.64416    0.21741   2.963   0.0039 ** 
benpros          0.31937    0.23449   1.362   0.1766    
gleason          0.36160    0.14118   2.561   0.0121 *  
benpros:gleason -0.03520    0.03382  -1.041   0.3008    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7134 on 90 degrees of freedom
Multiple R-squared:  0.6415,    Adjusted R-squared:  0.6176 
F-statistic: 26.84 on 6 and 90 DF,  p-value: < 2.2e-16

> 
```

The ***p-value < 2.2 x 10^-16*** indicates that the current model when compared to the model: psa ~ 1, the predictors included are significant in calculating the response variable.

Now we predict the response value for the PSA level for a patient whose quantitative predictors are at the sample means of the variables and qualitative

predictors are at the most frequent category. We are taking most frequent category of the variable “vesinv” which is “0” and mean of cancervol, age, benpros and gleason. We stored this data into test_data.

```
> test_data <- data.frame(cancervol=log(mean(data$cancervol)),
age=log(mean(data$age)),vesinv= factor(0),
benpros=mean(data$benpros),gleason=mean(data$gleason))
```

```
> test_data
  cancervol      age vesinv  benpros  gleason
1  1.945722 4.156787      0 2.534725 6.876289
```

We have completed creating our model. Now we used the "predict" function to predict the predict the value of our response variable.

```
> res <- predict(ModRef3, newdata = test_data)
```

```
> res <- predict(ModRef3, newdata = test_data)
> res
       1
4.249206
> exp(res)
       1
70.04974
>
```

Final R code is given on the next page.

R Code:

```
# STATS-Mini-Project-6
# Setting working directory to Proj-6 folder.
setwd("C:/Users/dpd140130.CAMPUS/OneDrive - The University of Texas at Dallas/CS 6313/Projects/06/Shalin/Mini-Proj-6")
getwd()
library(BSDA)

# Exploring our .csv file
data <- read.csv("prostate_cancer.csv")
data
data$vesinv<-as.factor(data$vesinv)
model <- lm(psa ~ cancervol + weight + age + benpros + vesinv + capspen + gleason, data = data)
summary(model)

# refinement - 1
ModRef1 <- update (model, .~. -weight)
anova(ModRef1,model)

# refinement - 2
ModRef2 <- update (ModRef1, .~. -capspen)
anova(ModRef2,ModRef1)

# refinement - 3
ModRef3 <- lm(log(psa) ~ log(cancervol)+log(age)+vesinv+benpros*gleason, data = data)
summary(ModRef3)

#Taking mean of the quantitative predictors and most frequent category of the qualitative predictor
test_data <- data.frame(cancervol=log(mean(data$cancervol)),
age=log(mean(data$age)),vesinv= factor(0),
benpros=(mean(data$benpros)),gleason=(mean(data$gleason)))
test_data

#Making Final Prediction
res <- predict(ModRef3, newdata = test_data)
res
exp(res)
```