Mini Project 06
Mini Project Group 31
Group Members
Dhairya Pinakin Desai: DPD140130

Question 1:

A dataset of 97 males with prostate tumours is provided to us. We will now gradually create an appropriate multivariate regression model by removing the irrelevant factors.

```
# 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")

> getwd()

> library(BSDA)

# Exploring our .csv file
> data <- read.csv("prostate_cancer.csv")
```

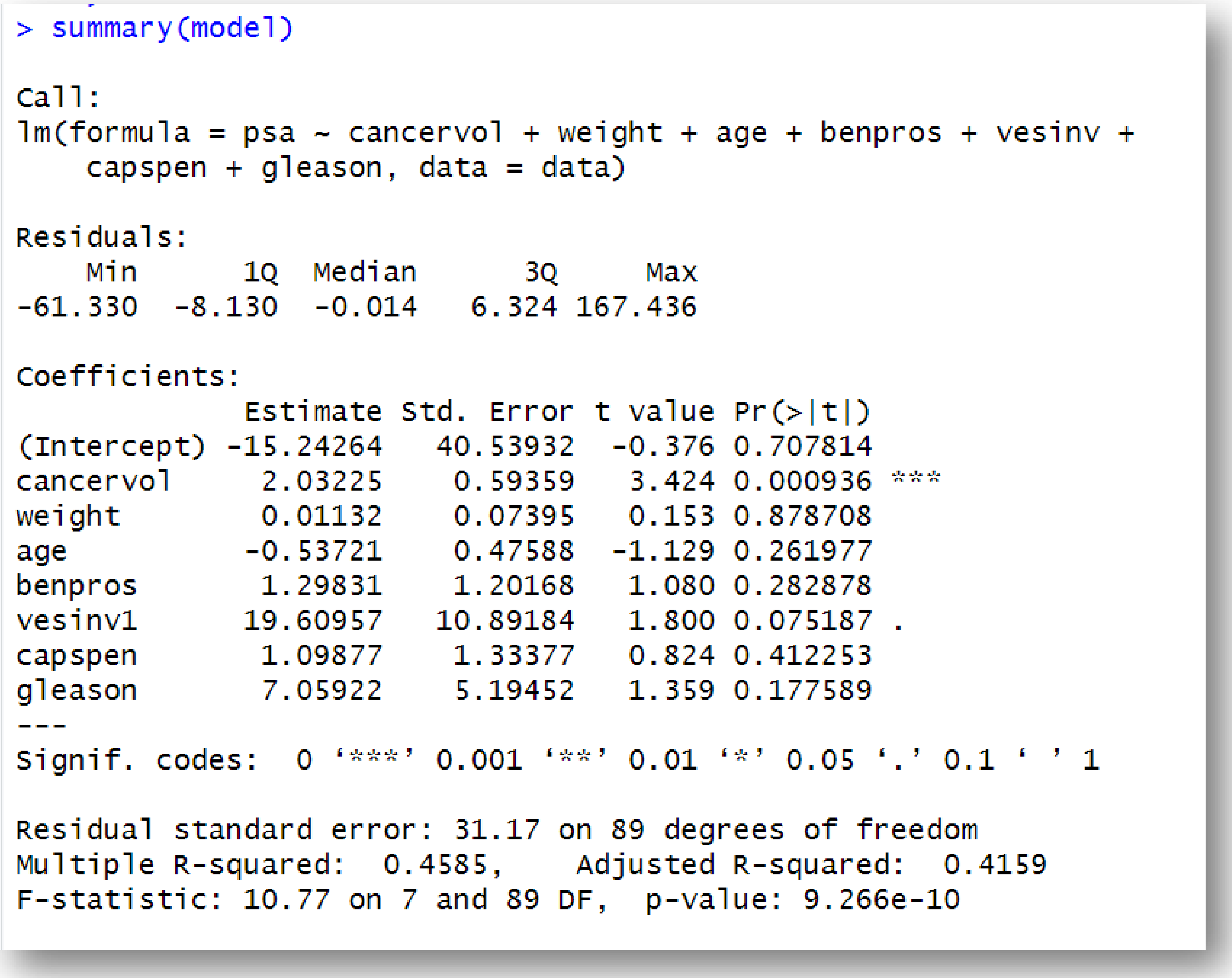
Inferring from our data that we have 6 numerical predictors (CancerVol, Weight, Age, Benpros, Capspen, and Glenson) and 1 categorical predictor (vesinv), we may say that we have these predictors. As a result, our model has 7 + 1 = 8 predictors (p).

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

With this command, R is explicitly told to consider the vesinv variable in our regression model as a categorical variable.

We are now going to build our fundamental regression model without removing any predictor variables. We do this in order to comprehend which variables could be more important than others and to learn how to reduce the complexity of 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 aforementioned “model” is our first regression model, and no unimportant factors have yet been removed. Its summary allows for the following conclusions to be drawn:-

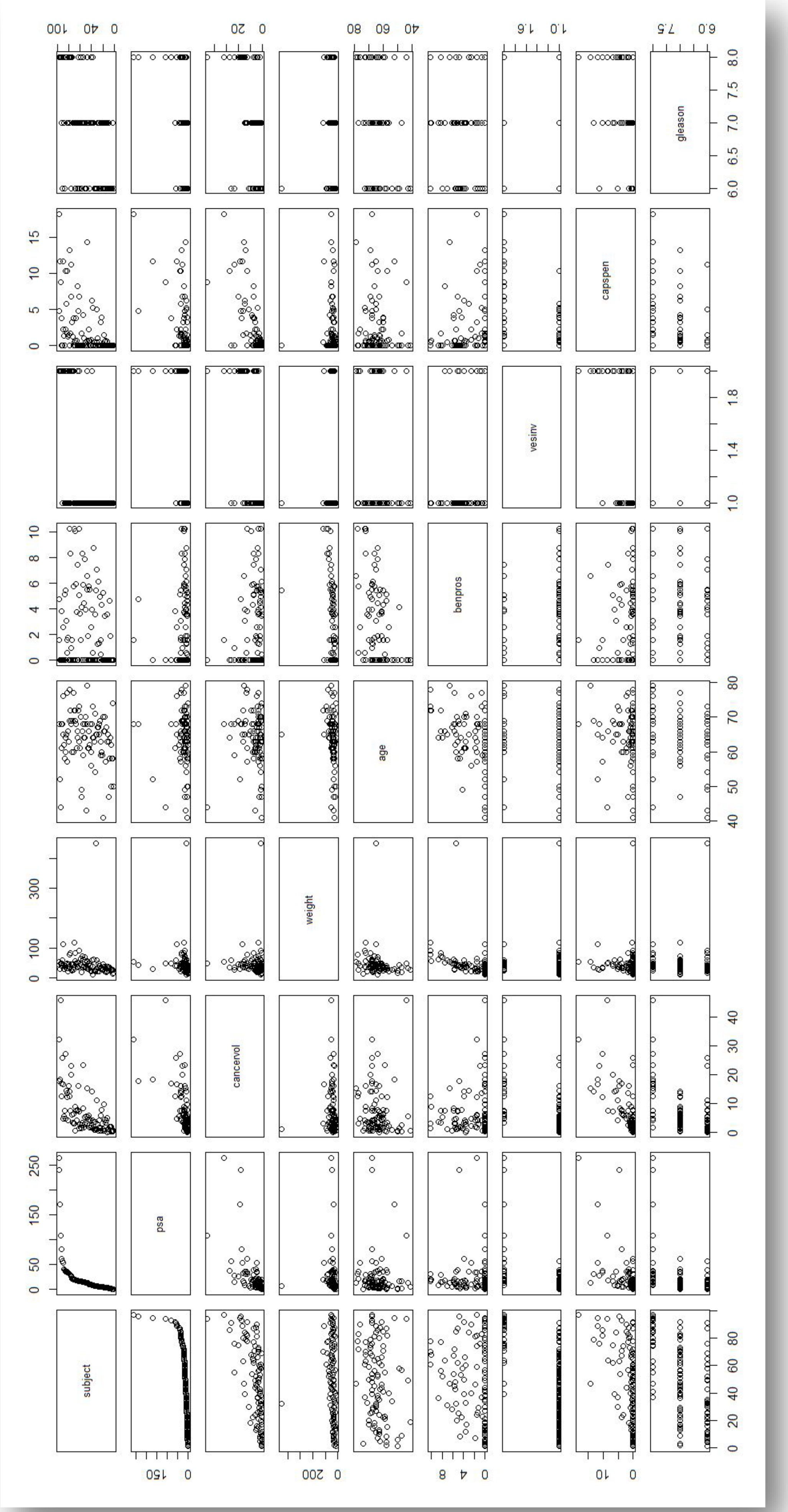
i) An extremely low p-value of 9.266 x 10-9 indicates that the provided predictors are pertinent to our model.

ii) Cancervol (***) is the most important predictor, and vesinv1 is also important.

iii) R-squared (Coefficient of Determination): 0.4585 indicates that the regression model can only account for 46% of the variation in the data.

First, we searched for any indication of a relationship between any of the data frame's variables. For this, the original data frame included in the question set and R's "plot" function are utilized.

```
> plot(data)
```


subject
psa
cancervol
weight
age
benpros
vesinv
capspen
gleason
0
20
40
60
80
0
50
150
250
0
10
20
30
40
0
100
300
40
50
60
70
80
0
2
4
6
8
10
1.0
1.4
1.8
0
5
10
15
6.0
6.5
7.0
7.5
8.0
0
40
100
0
150
0
20
0
200
40
60
80
0
4
8
1.0
1.6
0
10
6.0
7.5

Round 1 of Model Refinement:

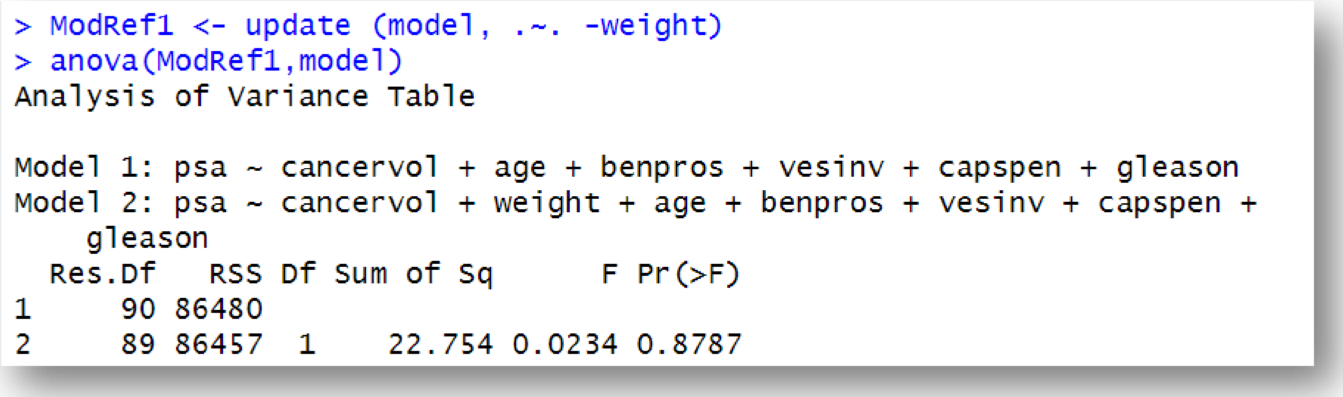
We now attempt to improve our model by removing certain variables that are not significant, and we will then use the ANOVA tool to test our hypotheses.

Removed predictor: Weight

Explanation: p-value is excessive (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, the p-value of 0.8787 leads to the acceptance of the null hypothesis H_0, which states that weight is an inconsequential predictor if others are constant.

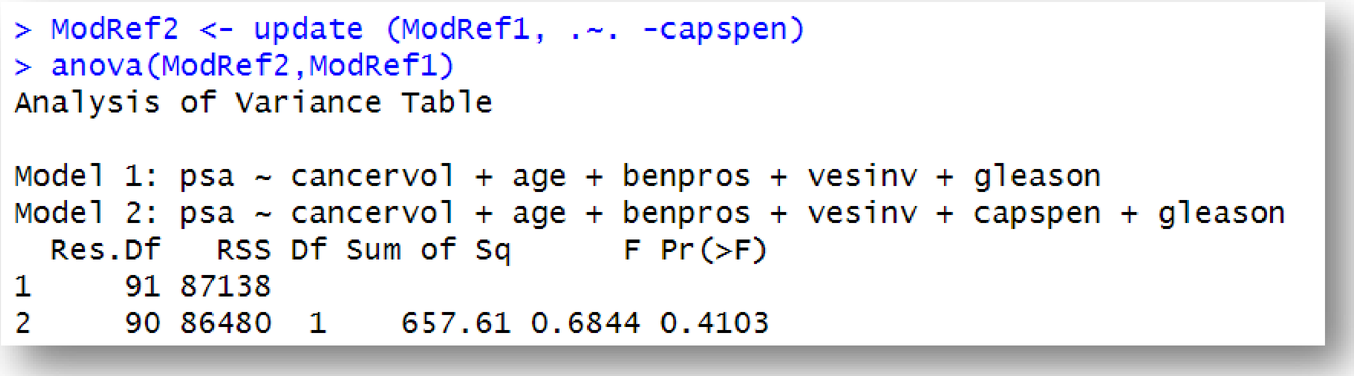
Round 2 of Model Refinement:

By removing certain factors that are not significant, we are now attempting to improve our model. Next, we will use the ANOVA method to test our hypotheses.

Removed predictor: Capspun

Explanation: excessively high p-value (0.4112)

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

> anova(ModRef2,ModRef1)
```



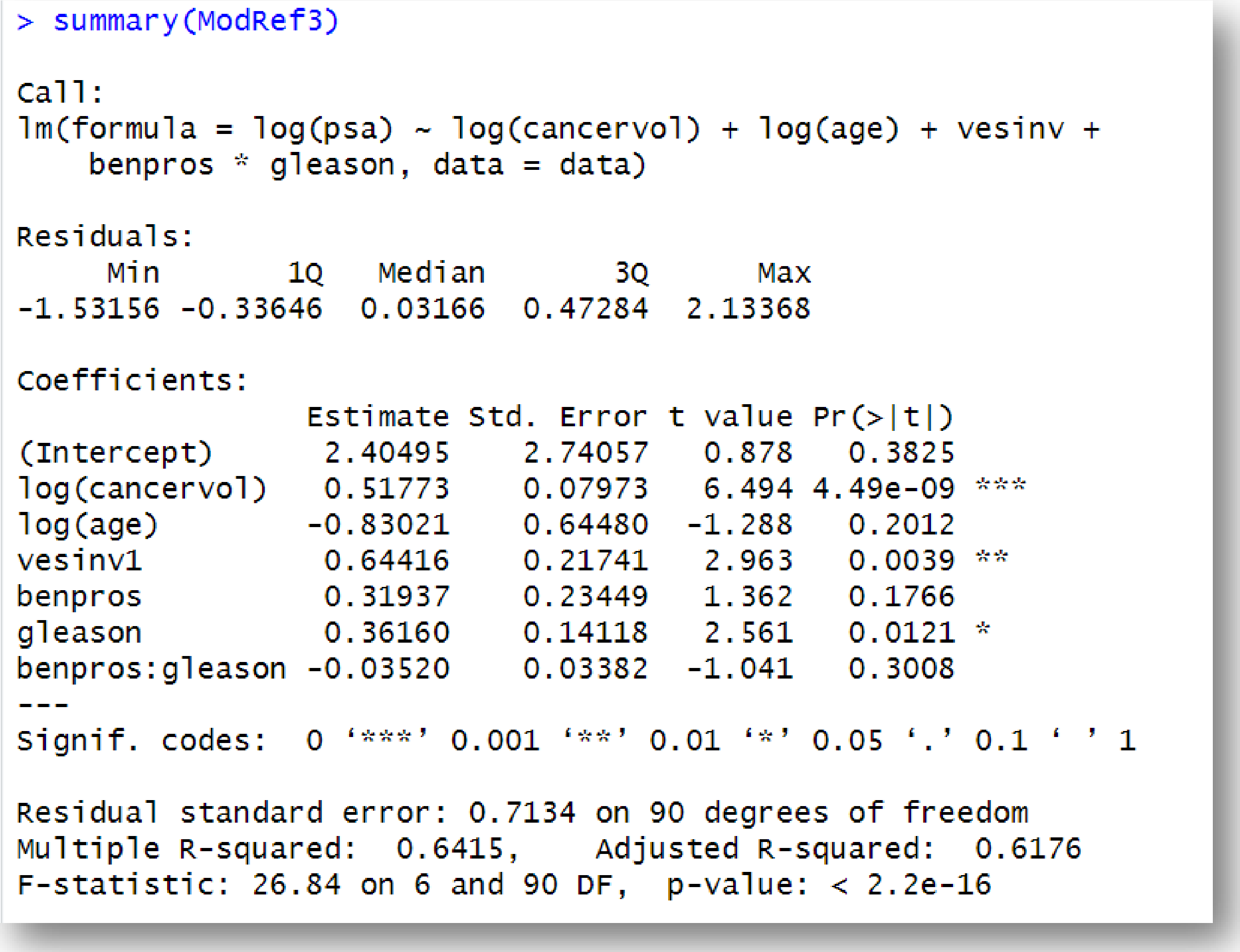
```
> ModRef2 <- update (ModRef1, .~. -capspen)
> anova(ModRef2,ModRef1)
Analysis of Variance Table

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, the null hypothesis H_0: "capspen" is an inconsequential predictor assuming others remain is accepted due to the p-value of 0.4103.

Round 3 of Model Refinement:

We discovered, via a series of trials and mistakes, that the following formulation of our regression model equation gave us the multiple-$R^2 = 0.6415$.

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

> summary(ModRef3)
```

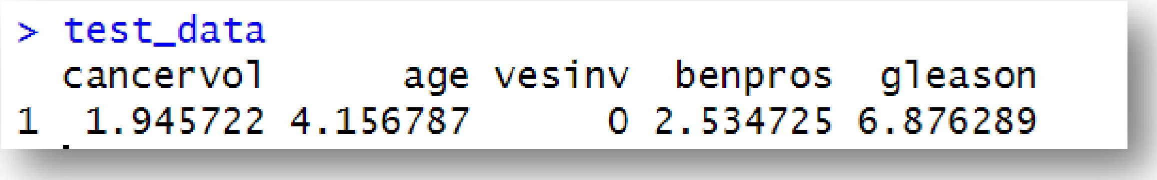


```
> summary(ModRef3)

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 present model's predictors are significant in determining the response variable when compared to the model with psa ~1, according to the p-value < 2.2 x 10^{-16}.

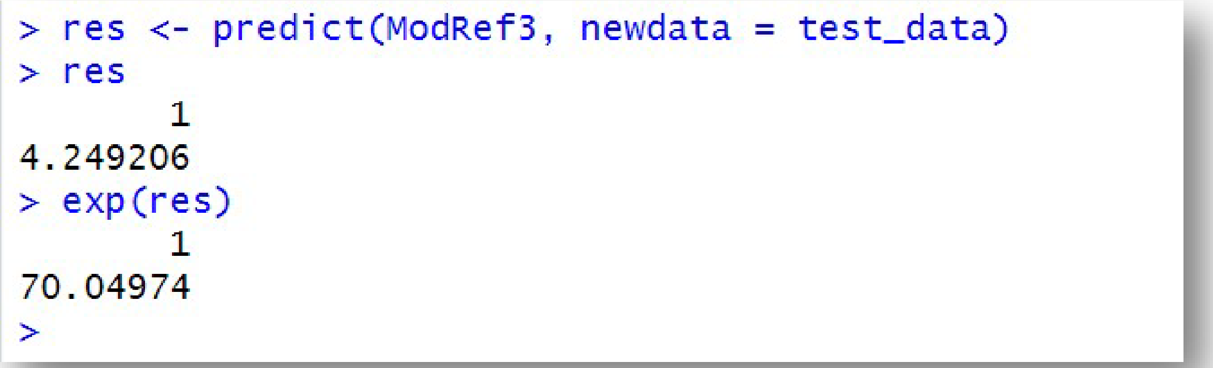
Now, we forecast the PSA level response value for a patient whose quantitative predictors are at the variable sample means and qualitative predictors are at the most prevalent category. We are using the median of cancervol, age, benpros, and gleason as well as the most common category of the variable "vesinv," which is "0." This information was saved in 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 finished building our model. We are currently predicting the value of our response variable using the "predict" function.



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

Final R code is given on the following page.

R Code:

```r
# 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)
```