Statistical Methods for Data Science
CS 6313.001: Mini Project #5

Due on Thursday April 20, 2017 at 4pm

Instructor: Pankaj Choudhary

Hanlin He / Lizhong Zhang (hxh160630 / lxz160730)

Contents

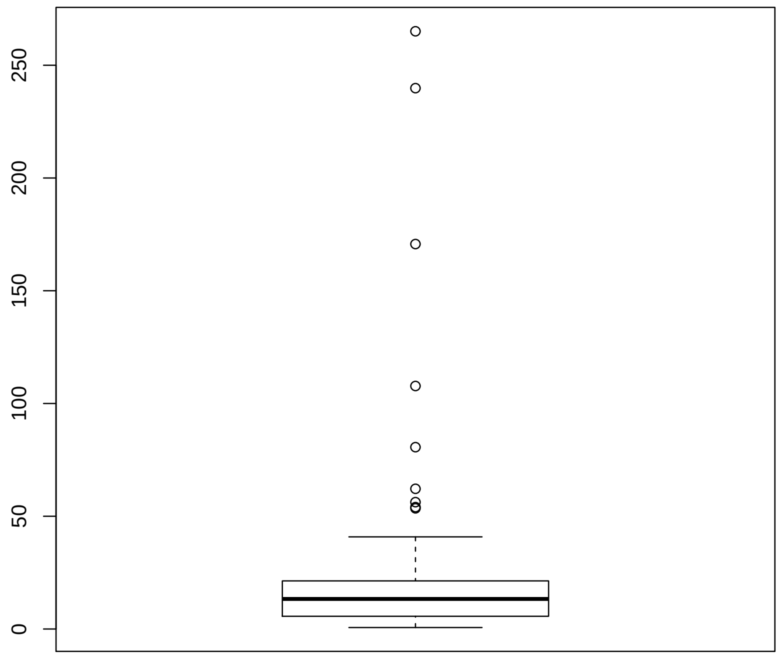
Contribution

Both team members made the same contribution in this project.

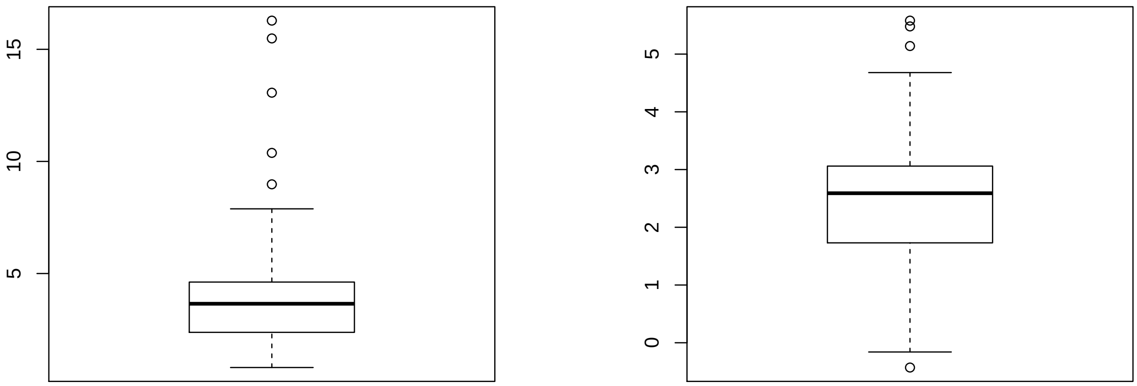
Section 1 Answers

First, we constructed the boxplot of the original PSA level (*psa*) as shown in fig. 1.

Figure 1: Boxplot of the Original PSA Level



From the boxplot, we could obviously identify many outliers. To eliminate these outliers, we transformed the original data to its square root and natural logarithmic value, as shown in fig. 2.



(a) Square Root

(b) Natural Logarithmic Value

Figure 2: Boxplot of the Transformed PSA Level

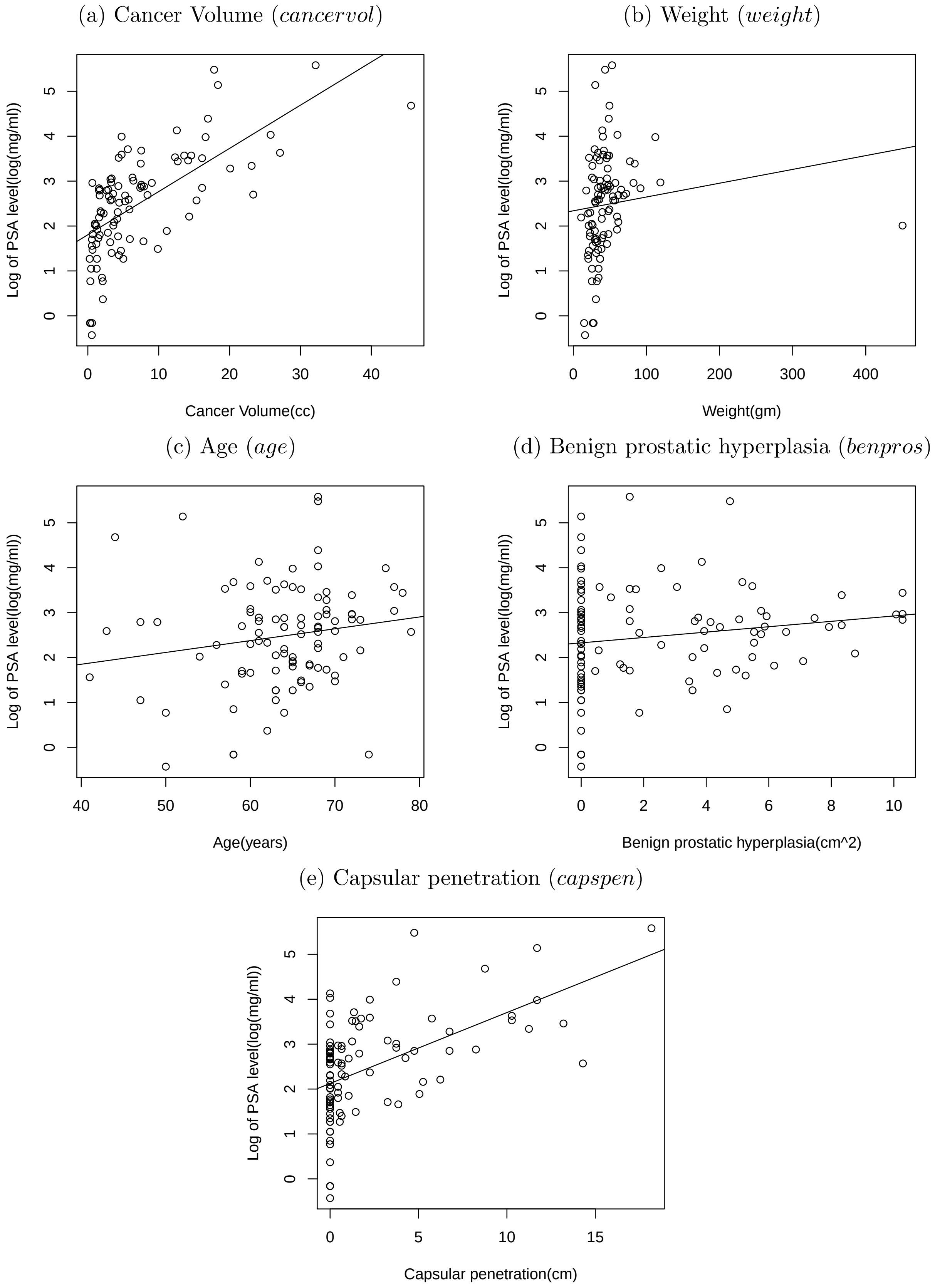
From these boxplots, we could see that the distribution of natural logarithmic value in fig. 2b has less outliers and is closer to a normal distribution than the distribution of the square root in fig. 2a. Thus, the following analysis would based on the natural logarithmic value of the PSA Level.

Based on the data and its description, we can conclude that the following are quantitative variables:

- Cancer Volume (*cancervol*)
- Weight (*weight*)
- Age (*age*)
- Benign prostatic hyperplasia (*benpros*)
- Capsular penetration (*capspen*)

Seminal vesicle invasion (*vesinv*) and Gleason score (*gleason*) are qualitative variables. We would first build model based on quantitative variables.

We constructed scatter plots based on the quantitative variables, as shown in fig. 3. In the meantime, we calculated the correlation coefficient of each variables as follow:



(a) Cancer Volume (*cancervol*)

(b) Weight (*weight*)

(c) Age (*age*)

(d) Benign prostatic hyperplasia (*benpros*)

(e) Capsular penetration (*capspen*)

Figure 3: Scatterplots for Different Variables

Based on these plots, we may have a guess that the most likely factor are Cancer Volume (*cancervol*), Benign prostatic hyperplasia (*benpros*) and Capsular penetration (*capspen*).

Now we have two model, one contains three variables (*cancervol*, *benpros*and *capspen*), the other has all quantitative variables. And we compare these two model by doing a F test.

```
> # Calculate the first formula.
> fit1 <- lm(psalog ~ cancervol + capspen + weight + age + benpros)
> fit1

Call:
lm(formula = psalog ~ cancervol + capspen + weight + age + benpros)

Coefficients:
(Intercept)    cancervol      capspen       weight          age      benpros
   1.037961     0.088925     0.033572     0.001028     0.007634     0.082325

> fit2 <- lm(psalog ~ cancervol + capspen + benpros)
> fit2

Call:
lm(formula = psalog ~ cancervol + capspen + benpros)

Coefficients:
(Intercept)    cancervol      capspen      benpros
    1.53504      0.08924      0.03544      0.09449

> # Compare first two guess.
> anova(fit2, fit1)
Analysis of Variance Table

Model 1: psalog ~ cancervol + capspen + benpros
Model 2: psalog ~ cancervol + capspen + weight + age + benpros
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     93 63.904
2     91 63.430  2   0.47464 0.3405 0.7123
```

From the result, we can see β_{weight} and $\beta_{age}$are very small, and $\beta_{cancervol}$, $\beta_{capspen}$ and $\beta_{benpros}$ are "acceptably" large, and the p value is also large (0.7123). So we accept the null hypothesis that $\beta_{weight} = 0$ and $\beta_{age} = 0$, and continued with our assumption with three variables.

Now we conduct the stepwise selection to confirm our assumption.

```
> # Apply stepwise selection.
> # Forward selection based on AIC.
> fit3.forward <-
+     step(lm(psalog ~ 1),
+     scope = list(upper = ~ cancervol + capspen + weight + age + benpros),
+     direction = "forward")

```

```
> fit3.forward

Call:
lm(formula = psalog ˜ cancervol + benpros)

Coefficients:
(Intercept)    cancervol      benpros
     1.5309       0.1010       0.0949

> # Backward elimination based on AIC.
> fit3.backward <-
+     step(lm(psalog ˜ cancervol + capspen + weight + age + benpros),
+     scope = list(lower = ˜1),
+     direction = ”backward”)

> fit3.backward

Call:
lm(formula = psalog ˜ cancervol + benpros)

Coefficients:
(Intercept)    cancervol      benpros
     1.5309       0.1010       0.0949

>
> # Both forward/backward.
> fit3.both <-
+     step(lm(psalog ˜ 1),
+     scope = list(lower = ˜1,
+                  upper = ˜ cancervol + capspen + weight + age + benpros),
+     direction = ”both”)

> fit3.both

Call:
lm(formula = psalog ˜ cancervol + benpros)

Coefficients:
(Intercept)    cancervol      benpros
     1.5309       0.1010       0.0949
```

From the AIC value (which is omitted above) and recommended formula, we now have a new formula:

```
> # Model selected.
> fit3 <- lm(formula = psalog ˜ cancervol + benpros)

> summary(fit3)

Call:
```

```
lm(formula = psalog ~ cancervol + benpros)

Residuals:
     Min       1Q   Median       3Q      Max
-2.01672 -0.55101  0.06457  0.56870  1.75415

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.53090    0.13940  10.982  < 2e-16 ***
cancervol    0.10105    0.01085   9.314 5.29e-15 ***
benpros      0.09490    0.02821   3.364  0.00111 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8303 on 94 degrees of freedom
Multiple R-squared:  0.4928,    Adjusted R-squared:  0.482
F-statistic: 45.67 on 2 and 94 DF,  p-value: 1.389e-14
```

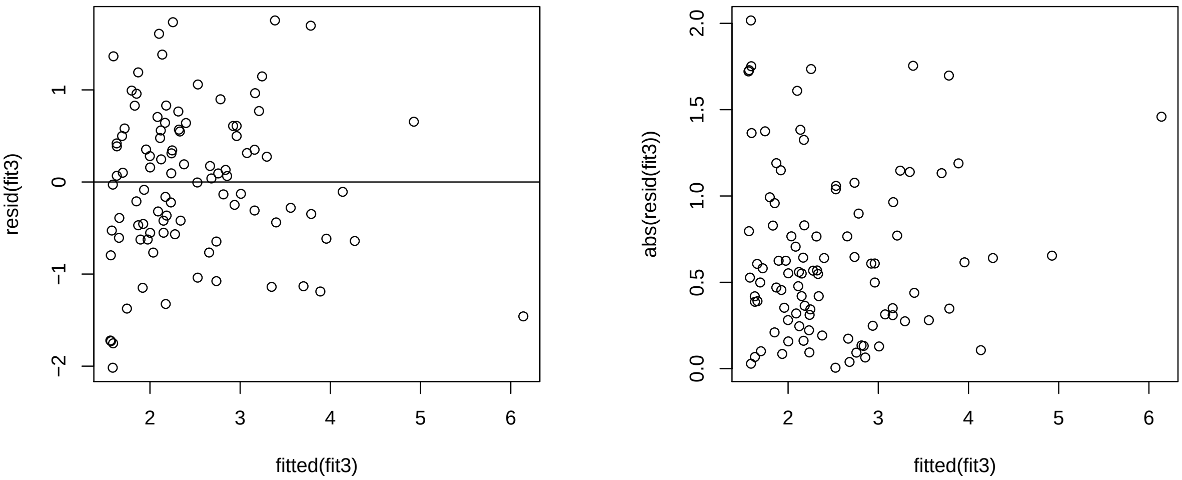
We can compare it with our previous guess one:

```
> # Compare the model with the guess one.
> anova(fit3, fit2)
Analysis of Variance Table

Model 1: psalog ~ cancervol + benpros
Model 2: psalog ~ cancervol + capspen + benpros
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     94 64.802
2     93 63.904  1   0.89737 1.3059 0.2561
```

The p value is large (0.2561), we could accept the null hypothesis that $\beta_{capspen} = 0$. So we could furthermore throw away variable *capspen*.

Now we get the 'final' model with quantitative variables only.

The residual graph for the model is shown in fig. 4a. The absolute residual of the model is shown in fig. 4b.

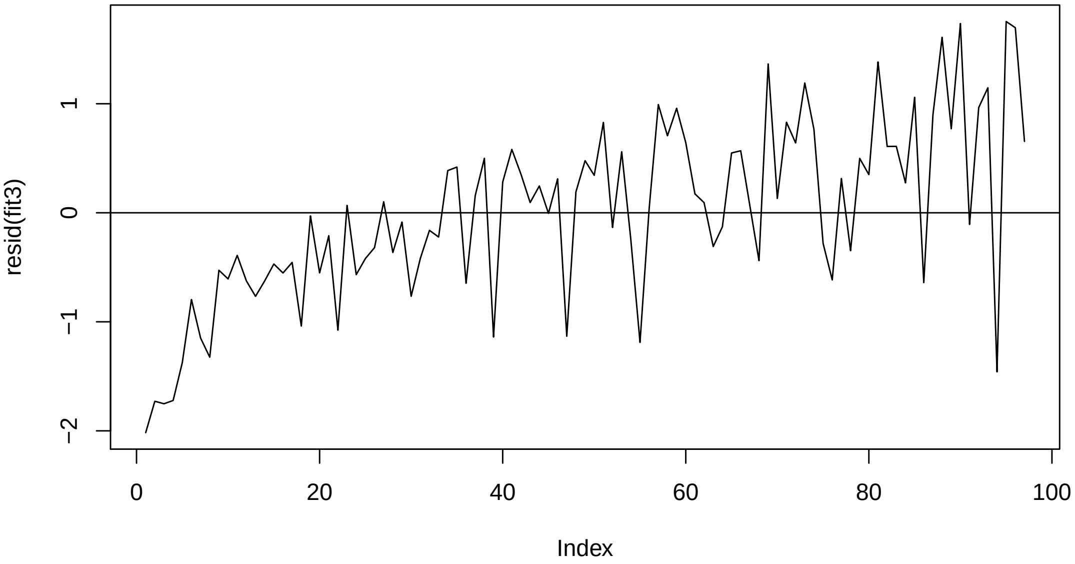

(a) Residual Graph

(b) Absolute Residual Graph

Figure 4: Residual and Absolute Residual Graph for Model with Quantitative Variables Only

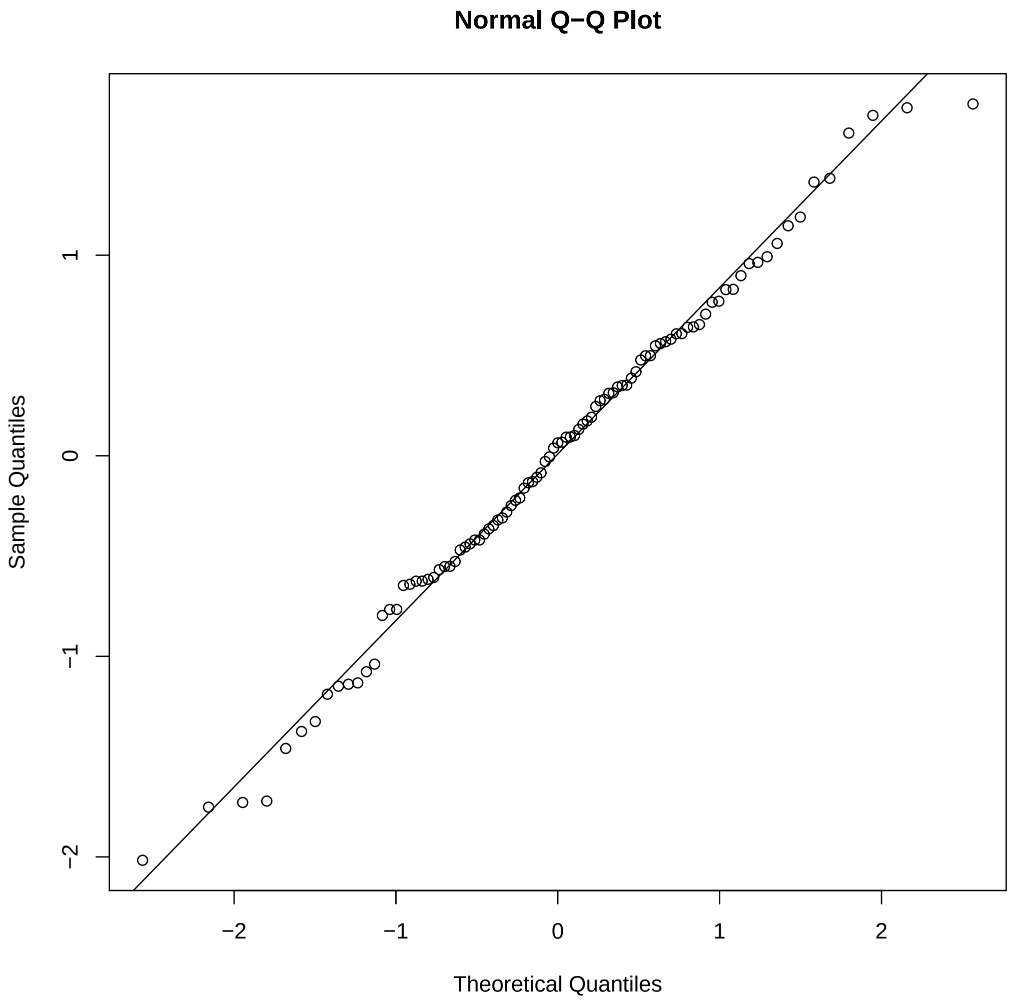
The time series plot of the model is shown in fig. 5.

Figure 5: Time Series Plot for Final Model



The normal QQ plot of the model is shown in fig. 6.

Figure 6: Normal QQ Plot for Model with Quantitative Variables Only



The model seems quite reasonable. Now we consider the qualitative(categorical) variables. We first add two variables to the model separately and

```
> # Consider the categorical variables.
> fit4 <- update(fit3, . ˜ . + factor(vesinv))
> fit5 <- update(fit3, . ˜ . + factor(gleason))
```

Then we compare them with the original model.

```
> # Comparing two categorical variables.
> summary(fit5)

Call:
lm(formula = psalog ˜ cancervol + benpros + factor(gleason))
```

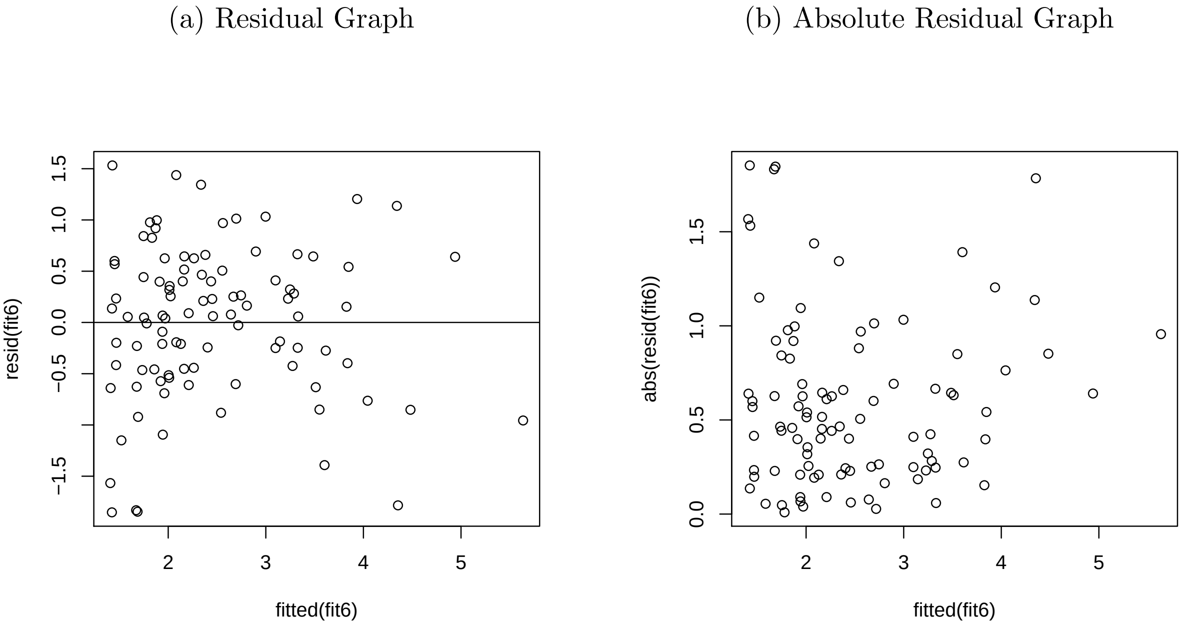
```

Residuals:
     Min       1Q   Median       3Q      Max 
-1.92886 -0.59159  0.04246  0.56555  1.56306 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       1.34533    0.16164   8.323 7.63e-13 ***
cancervol         0.08095    0.01259   6.430 5.62e-09 ***
benpros           0.08622    0.02722   3.167  0.00209 ** 
factor(gleason)7  0.37475    0.18572   2.018  0.04652 *  
factor(gleason)8  0.84137    0.26303   3.199  0.00189 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7942 on 92 degrees of freedom
Multiple R-squared:  0.5458,    Adjusted R-squared:  0.5261 
F-statistic: 27.64 on 4 and 92 DF,  p-value: 4.467e-15

> anova(fit3, fit5)
Analysis of Variance Table

Model 1: psalog ~ cancervol + benpros
Model 2: psalog ~ cancervol + benpros + factor(gleason)
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)   
1     94 64.802                                
2     92 58.032  2    6.7695 5.3659 0.006249 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(fit4)

Call:
lm(formula = psalog ~ cancervol + benpros + factor(vesinv))

Residuals:
    Min      1Q  Median      3Q     Max 
-1.9867 -0.4996  0.1032  0.5545  1.4993 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      1.51484    0.13206  11.471  < 2e-16 ***
cancervol        0.07618    0.01256   6.067 2.78e-08 ***
benpros          0.09971    0.02674   3.729 0.000331 ***
factor(vesinv)1  0.82194    0.23858   3.445 0.000858 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7861 on 93 degrees of freedom
```

```
Multiple R-squared:  0.5502,    Adjusted R-squared:  0.5357
F-statistic: 37.92 on 3 and 93 DF,  p-value: 4.247e-16

> anova(fit3, fit4)
Analysis of Variance Table

Model 1: psalog ~ cancervol + benpros
Model 2: psalog ~ cancervol + benpros + factor(vesinv)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     94 64.802
2     93 57.468  1    7.3339 11.868 0.0008583 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can conclude from the result that these two variables are definitely significant in the model. So we add these two variables to the formula, which results in our final model.

```
> # Finalize the model.
> fit6 <- update(fit3, . ~ . + factor(vesinv) + factor(gleason))
>
> summary(fit6)

Call:
lm(formula = psalog ~ cancervol + benpros + factor(vesinv) +
    factor(gleason))

Residuals:
     Min       1Q   Median       3Q      Max
-1.85235 -0.45777  0.06741  0.51651  1.53204

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.38817    0.15609   8.894 5.27e-14 ***
cancervol         0.06241    0.01367   4.566 1.55e-05 ***
benpros           0.09265    0.02627   3.527  0.00066 ***
factor(vesinv)1   0.69646    0.23837   2.922  0.00439 **
factor(gleason)7  0.26028    0.18280   1.424  0.15790
factor(gleason)8  0.70545    0.25712   2.744  0.00732 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7636 on 91 degrees of freedom
Multiple R-squared:  0.5848,    Adjusted R-squared:  0.5619
F-statistic: 25.63 on 5 and 91 DF,  p-value: 4.722e-16
```

The residual graph for the final model is shown in fig. 7a. The absolute residual of the model is shown in fig. 7b.

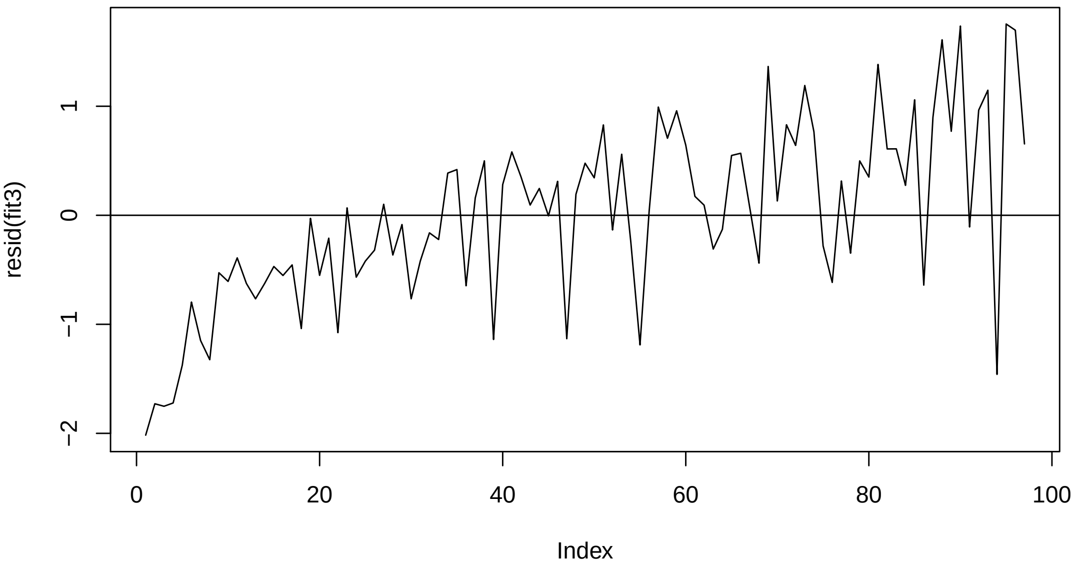

(a) Residual Graph

(b) Absolute Residual Graph

Figure 7: Residual and Absolute Residual Graph for Final Model

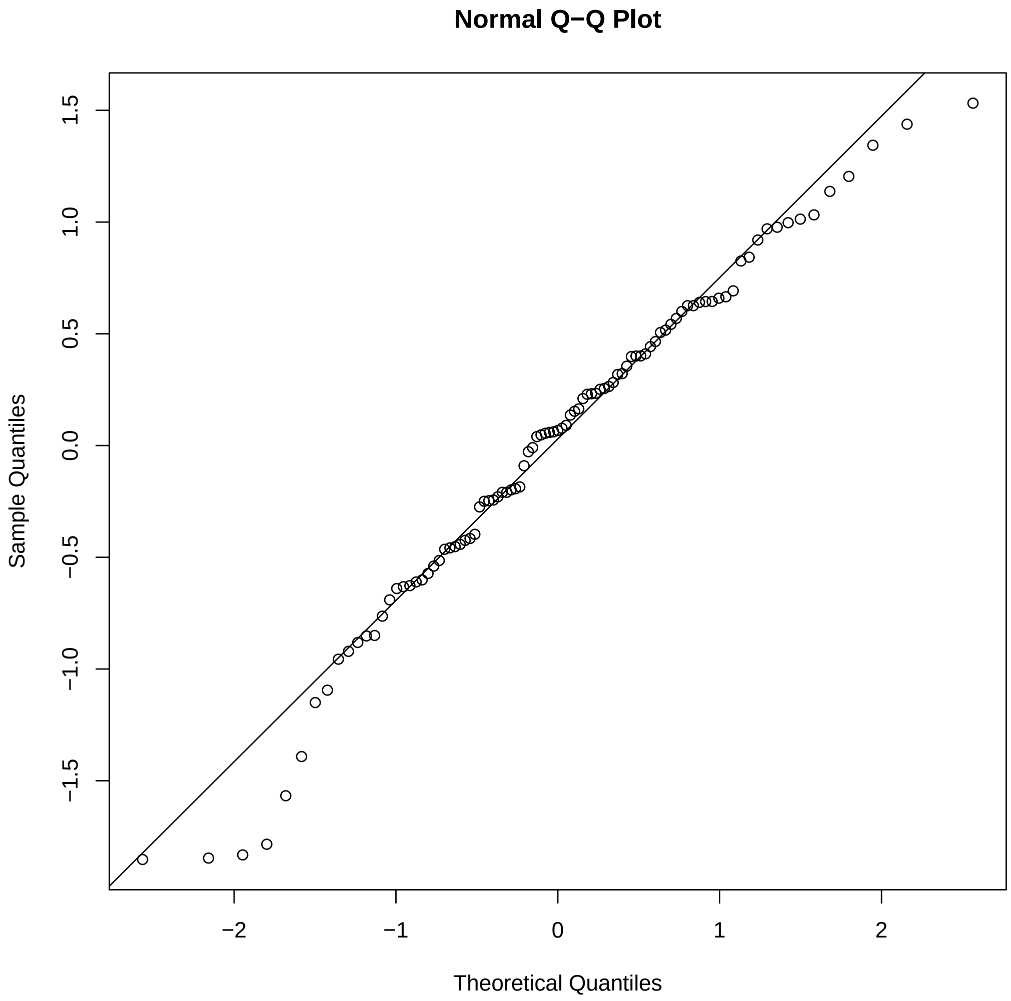
The time series plot of the model is shown in fig. 8.

Figure 8: Time Series Plot for Final Model



The normal QQ plot of the model is shown in fig. 9.

Figure 9: Normal QQ Plot for Final Model



The model seems reasonable in most part but has some outliers.

With this model we can predict the PSA level for a patient whose predictors are at the sample means of the variables as follow:

```
> # Predict the PSA level for sample mean.
> es <- predict(fit6,
+     data.frame(cancervol = mean(cancervol),
+                benpros   = mean(benpros),
+                vesinv    = getmode(vesinv),
+                gleason   = getmode(gleason)))
> exp(es)
       1
10.17628
```

Section 2 R Code

```
# Read data from file.
prostatecancer <- read.table(file="prostate_cancer.csv", sep=",", header=T)

# Create fig folder to store plot.
if(!dir.exists("fig”)) dir.create("fig”)

# Attach data to memory.
attach(prostatecancer)
psalog <- log(psa)

# Box plot for psa.
pdf("fig/boxplotpsa.pdf”, width=7, height=7)
boxplot(psa)
dev.off()

## pdf
##   2

# Box plot for square root of psa.
pdf("fig/boxplotpsasqrt.pdf”, width=5, height=5)
boxplot(sqrt(psa))
dev.off()

## pdf
##   2

# Box plot for logarithm of psa.
pdf("fig/boxplotpsalog.pdf”, width=5, height=5)
boxplot(log(psa))
dev.off()

## pdf
##   2

# Draw scatterplots of each variables with log(psa).
pdf("fig/boxplotcancervol.pdf”, width=5, height=5)
plot(cancervol, psalog,
    xlab="Cancer Volume(cc)”,
    ylab="Log of PSA level(log(mg/ml))”)
abline(lm(psalog ˜ cancervol))
dev.off()
```

```
## pdf
##   2

pdf("fig/boxplotweight.pdf", width=5, height=5)
plot(weight, psalog,
    xlab="Weight(gm)",
    ylab="Log of PSA level(log(mg/ml))")
abline(lm(psalog ~ weight))
dev.off()

## pdf
##   2

pdf("fig/boxplotage.pdf", width=5, height=5)
plot(age, psalog,
    xlab="Age(years)",
    ylab="Log of PSA level(log(mg/ml))")
abline(lm(psalog ~ age))
dev.off()

## pdf
##   2

pdf("fig/boxplotbenpros.pdf", width=5, height=5)
plot(benpros, psalog,
    xlab="Benign prostatic hyperplasia(cm^2)",
    ylab="Log of PSA level(log(mg/ml))")
abline(lm(psalog ~ benpros))
dev.off()

## pdf
##   2

pdf("fig/boxplotcapspen.pdf", width=5, height=5)
plot(capspen, psalog,
    xlab="Capsular penetration(cm)",
    ylab="Log of PSA level(log(mg/ml))")
abline(lm(psalog ~ capspen))
dev.off()

## pdf
##   2
```

```
# Calculate the first formula.
fit1 <- lm(psalog ~ cancervol + capspen + weight + age + benpros)
fit1

##
## Call:
## lm(formula = psalog ~ cancervol + capspen + weight + age + benpros)
##
## Coefficients:
## (Intercept)    cancervol      capspen       weight          age
##    1.037961     0.088925     0.033572     0.001028     0.007634
##     benpros
##    0.082325

fit2 <- lm(psalog ~ cancervol + capspen + benpros)
fit2

##
## Call:
## lm(formula = psalog ~ cancervol + capspen + benpros)
##
## Coefficients:
## (Intercept)    cancervol      capspen      benpros
##     1.53504      0.08924      0.03544      0.09449

# Compare first two guess.
anova(fit2, fit1)

## Analysis of Variance Table
##
## Model 1: psalog ~ cancervol + capspen + benpros
## Model 2: psalog ~ cancervol + capspen + weight + age + benpros
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     93 63.904
## 2     91 63.430  2   0.47464 0.3405 0.7123

# Apply stepwise selection.
# Forward selection based on AIC.
fit3.forward <-
    step(lm(psalog ~ 1),
    scope = list(upper = ~ cancervol + capspen + weight + age + benpros),
    direction = "forward")
```

```
## Start:  AIC=28.72
## psalog ~ 1
##
##             Df Sum of Sq     RSS      AIC
## + cancervol  1    55.164  72.605 -24.0986
## + capspen    1    34.286  93.482   0.4169
## + age        1     3.688 124.080  27.8831
## + benpros    1     3.166 124.603  28.2911
## <none>                   127.769  28.7246
## + weight     1     1.893 125.876  29.2767
##
## Step:  AIC=-24.1
## psalog ~ cancervol
##
##           Df Sum of Sq    RSS     AIC
## + benpros  1    7.8034 64.802 -33.128
## + age      1    2.6615 69.944 -25.721
## + weight   1    1.7901 70.815 -24.520
## <none>                 72.605 -24.099
## + capspen  1    0.9673 71.638 -23.400
##
## Step:  AIC=-33.13
## psalog ~ cancervol + benpros
##
##           Df Sum of Sq    RSS     AIC
## <none>                 64.802 -33.128
## + capspen  1   0.89737 63.904 -32.480
## + age      1   0.39609 64.406 -31.723
## + weight   1   0.20572 64.596 -31.436
```

```
fit3.forward
```

```
##
## Call:
## lm(formula = psalog ~ cancervol + benpros)
##
## Coefficients:
## (Intercept)    cancervol      benpros
##      1.5309       0.1010       0.0949
```

```
# Backward elimination based on AIC.
```

```
fit3.backward <-
    step(lm(psalog ~ cancervol + capspen + weight + age + benpros),
    scope = list(lower = ~1),
    direction = "backward")
```

```
## Start:  AIC=-29.2
## psalog ~ cancervol + capspen + weight + age + benpros
##
##             Df Sum of Sq    RSS      AIC
## - weight     1    0.1891 63.619 -30.9149
## - age        1    0.2626 63.692 -30.8029
## - capspen    1    0.7963 64.226 -29.9934
## <none>                   63.430 -29.2036
## - benpros    1    4.6231 68.053 -24.3794
## - cancervol  1   24.1971 87.627   0.1424
##
## Step:  AIC=-30.91
## psalog ~ cancervol + capspen + age + benpros
##
##             Df Sum of Sq    RSS     AIC
## - age        1    0.2856 63.904 -32.480
## - capspen    1    0.7869 64.406 -31.723
## <none>                   63.619 -30.915
## - benpros    1    5.6465 69.265 -24.667
## - cancervol  1   24.4216 88.040  -1.401
##
## Step:  AIC=-32.48
## psalog ~ cancervol + capspen + benpros
##
##             Df Sum of Sq    RSS     AIC
## - capspen    1    0.8974 64.802 -33.128
## <none>                   63.904 -32.480
## - benpros    1    7.7334 71.638 -23.400
## - cancervol  1   24.4110 88.315  -3.098
##
## Step:  AIC=-33.13
## psalog ~ cancervol + benpros
##
##             Df Sum of Sq     RSS     AIC
## <none>                    64.802 -33.128
## - benpros    1     7.803  72.605 -24.099
## - cancervol  1    59.802 124.603  28.291
```

```
fit3.backward
```

```
##
## Call:
## lm(formula = psalog ~ cancervol + benpros)
##
## Coefficients:
## (Intercept)    cancervol      benpros
##      1.5309       0.1010       0.0949
```

```
# Both forward/backward.
fit3.both <-
    step(lm(psalog ~ 1),
    scope = list(lower = ~1,
                 upper = ~ cancervol + capspen + weight + age + benpros),
    direction = "both")
```

```
## Start:  AIC=28.72
## psalog ~ 1
##
##             Df Sum of Sq     RSS      AIC
## + cancervol  1    55.164  72.605 -24.0986
## + capspen    1    34.286  93.482   0.4169
## + age        1     3.688 124.080  27.8831
## + benpros    1     3.166 124.603  28.2911
## <none>                   127.769  28.7246
## + weight     1     1.893 125.876  29.2767
##
## Step:  AIC=-24.1
## psalog ~ cancervol
##
##             Df Sum of Sq     RSS     AIC
## + benpros    1     7.803  64.802 -33.128
## + age        1     2.662  69.944 -25.721
## + weight     1     1.790  70.815 -24.520
## <none>                    72.605 -24.099
## + capspen    1     0.967  71.638 -23.400
## - cancervol  1    55.164 127.769  28.725
##
```

```
## Step:  AIC=-33.13
## psalog ~ cancervol + benpros
##
##             Df Sum of Sq     RSS     AIC
## <none>                    64.802 -33.128
## + capspen    1     0.897  63.904 -32.480
## + age        1     0.396  64.406 -31.723
## + weight     1     0.206  64.596 -31.436
## - benpros    1     7.803  72.605 -24.099
## - cancervol  1    59.802 124.603  28.291
```

```
fit3.both
```

```
##
## Call:
## lm(formula = psalog ~ cancervol + benpros)
##
## Coefficients:
## (Intercept)    cancervol      benpros
##      1.5309       0.1010       0.0949
```

```
# Model selected.
fit3 <- lm(formula = psalog ~ cancervol + benpros)

summary(fit3)
```

```
##
## Call:
## lm(formula = psalog ~ cancervol + benpros)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.01672 -0.55101  0.06457  0.56870  1.75415
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.53090    0.13940  10.982  < 2e-16 ***
## cancervol    0.10105    0.01085   9.314 5.29e-15 ***
## benpros      0.09490    0.02821   3.364  0.00111 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.8303 on 94 degrees of freedom
## Multiple R-squared:  0.4928, Adjusted R-squared:  0.482
## F-statistic: 45.67 on 2 and 94 DF,  p-value: 1.389e-14

# Compare the model with the guess one.
anova(fit3, fit2)

## Analysis of Variance Table
##
## Model 1: psalog ~ cancervol + benpros
## Model 2: psalog ~ cancervol + capspen + benpros
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     94 64.802
## 2     93 63.904  1   0.89737 1.3059 0.2561

# Residual plot of fit3.
pdf("fig/residualplotfit3.pdf", width=5, height=5)
plot(fitted(fit3), resid(fit3))
abline(h = 0)
dev.off()

## pdf
##   2

# Plot the absolute residual of fit3.
pdf("fig/plotfit3abu.pdf", width=5, height=5)
plot(fitted(fit3), abs(resid(fit3)))
dev.off()

## pdf
##   2

# Plot the times series plot of residuals.
pdf("fig/plotfit3times.pdf", width=8, height=5)
plot(resid(fit3), type="l")
abline(h = 0)
dev.off()

## pdf
##   2
```

```
# Normal QQ plot of fit3.
pdf("fig/qqnormplotfit3.pdf", width=8, height=8)
qqnorm(resid(fit3))
qqline(resid(fit3))
dev.off()
```

```
## pdf
##   2
```

```
# Consider the categorical variables.
fit4 <- update(fit3, . ~ . + factor(vesinv))
fit5 <- update(fit3, . ~ . + factor(gleason))

# Comparing two categorical variables.
summary(fit5)
```

```
##
## Call:
## lm(formula = psalog ~ cancervol + benpros + factor(gleason))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.92886 -0.59159  0.04246  0.56555  1.56306
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.34533    0.16164   8.323 7.63e-13 ***
## cancervol           0.08095    0.01259   6.430 5.62e-09 ***
## benpros             0.08622    0.02722   3.167  0.00209 **
## factor(gleason)7    0.37475    0.18572   2.018  0.04652 *
## factor(gleason)8    0.84137    0.26303   3.199  0.00189 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7942 on 92 degrees of freedom
## Multiple R-squared:  0.5458, Adjusted R-squared:  0.5261
## F-statistic: 27.64 on 4 and 92 DF,  p-value: 4.467e-15
```

```
anova(fit3, fit5)
```

```
## Analysis of Variance Table
##
```

```
## Model 1: psalog ~ cancervol + benpros
## Model 2: psalog ~ cancervol + benpros + factor(gleason)
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     94 64.802
## 2     92 58.032  2    6.7695 5.3659 0.006249 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(fit4)
```

```
##
## Call:
## lm(formula = psalog ~ cancervol + benpros + factor(vesinv))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9867 -0.4996  0.1032  0.5545  1.4993
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.51484    0.13206  11.471  < 2e-16 ***
## cancervol        0.07618    0.01256   6.067 2.78e-08 ***
## benpros          0.09971    0.02674   3.729 0.000331 ***
## factor(vesinv)1  0.82194    0.23858   3.445 0.000858 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7861 on 93 degrees of freedom
## Multiple R-squared:  0.5502, Adjusted R-squared:  0.5357
## F-statistic: 37.92 on 3 and 93 DF,  p-value: 4.247e-16
```

```
anova(fit3, fit4)
```

```
## Analysis of Variance Table
##
## Model 1: psalog ~ cancervol + benpros
## Model 2: psalog ~ cancervol + benpros + factor(vesinv)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     94 64.802
## 2     93 57.468  1    7.3339 11.868 0.0008583 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Finalize the model.
fit6 <- update(fit3, . ~ . + factor(vesinv) + factor(gleason))

summary(fit6)

## 
## Call:
## lm(formula = psalog ~ cancervol + benpros + factor(vesinv) + 
##     factor(gleason))
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.85235 -0.45777  0.06741  0.51651  1.53204 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)        1.38817    0.15609   8.894 5.27e-14 ***
## cancervol          0.06241    0.01367   4.566 1.55e-05 ***
## benpros            0.09265    0.02627   3.527  0.00066 ***
## factor(vesinv)1    0.69646    0.23837   2.922  0.00439 ** 
## factor(gleason)7   0.26028    0.18280   1.424  0.15790    
## factor(gleason)8   0.70545    0.25712   2.744  0.00732 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.7636 on 91 degrees of freedom
## Multiple R-squared:  0.5848, Adjusted R-squared:  0.5619 
## F-statistic: 25.63 on 5 and 91 DF,  p-value: 4.722e-16

# Residual plot of fit6.
pdf("fig/residualplotfit6.pdf", width=5, height=5)
plot(fitted(fit6), resid(fit6))
abline(h = 0)
dev.off()

## pdf 
##   2

# Plot the absolute residual of fit3.
pdf("fig/plotfit6abu.pdf", width=5, height=5)
plot(fitted(fit6), abs(resid(fit6)))
dev.off()
```

```
## pdf
##   2

# Plot the times series plot of residuals.
pdf("fig/plotfit6times.pdf", width=8, height=5)
plot(resid(fit3), type="l")
abline(h = 0)
dev.off()

## pdf
##   2

# Normal QQ plot of fit6
pdf("fig/qqnormplotfit6.pdf", width=8, height=8)
qqnorm(resid(fit6))
qqline(resid(fit6))
dev.off()

## pdf
##   2

# Create the mode function.
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

# Predict the PSA level for sample mean.
es <- predict(fit6,
    data.frame(cancervol = mean(cancervol),
               benpros   = mean(benpros),
               vesinv    = getmode(vesinv),
               gleason   = getmode(gleason)))
exp(es)

##        1
## 10.17628
```