Analysis of Relationship between PSA and Clinical Measurements

Yujun Chen

yjnchen@ucdavis.edu

Siqi Lu

sqlu@ucdavis.edu

Zhen Zhang

ezzhang@ucdavis.edu

December 7, 2015

Abstract

This paper aims to unravel the relationship between serum *prostate−specific antigen level* (PSA) and a number of prognostic clinical measurements in men with advanced prostate cancer by building a general linear model on the data observed on 97 men who were about to undergo radical prostatectomies. After model selection, model diagnosis and cross-validation, a model with *cancer volume*, *weight*, *age*, *benign prostatic hyperplasia*, *seminal vesicle invasion* and *Gleason score* is chosen. The predictor *capsular penetration* is proved to be redundant in this model. The relationship between response variable and each predictor is first-order and linear, and is positive except for *age*. The relationship between PSA and *cancer volume*, *weight*, *seminal vesicle invasion* is highly significant ($p < 0.03$), that between PSA and *age* and *benign prostatic hyperplasia* is somewhat significant ($p \approx 0.15$). The *Gleason score* 7 has the same effect as 6 on PSA ($p = 0.43$), while 8 has a significantly distinct effect on PSA compared with 6 ($p = 0.02$). The total coefficient of determination (R^2) of this model is 0.6385.

1 Introduction

Serum $prostate-specific\ antigen\ level$ (PSA) is a useful index in the diagnosis of prostate cancer and the follow-up observations of the patients who underwent radical prostatectomies. Analysis on the relationship between PSA and several prognostic clinical measurements may simplify the process to diagnose the occurrence and reoccurrence of prostate cancer and thus is a meaningful research. Previous research given by Thomas A. Stamey shows that PSA has a significant relationship with *Gleason score* 7 or greater, *capsular penetration* greater than 1 cm in linear extent, *seminal vesicle invasion* and $volume\ of\ prostate\ cancer$[1]. The following analysis aims to answer these questions: which prediction clinical measurements have effects on the response variable? The relationship is linear or curvilinear? Which predictors have positive effects and which have negative effects? Have we missed some important prediction variables? Does the model suffice? The data used in the analysis was collected on 97 men who were about to undergo radical prostatectomies, for each observation, *patient* ID, $PSA\ level(mb/ml)$, $cancer\ volume(cc)$, $weight(gm)$, $age(years)$, *benign prostatic hyperplasia*(cm^2), *seminal vesicle invasion*, *capsular penetration*(cm) and *Gleasonscore* were measured and recorded.

2 Methods and Results

2.1 Data Exploration

First we read the data into R, we want to decide the type of each variable, namely distinguish between the quantitative and qualitative variables, so we draw the scatterplot (Figure 1):

From the scatterplot, we found that PSA, *cancer volume*, *weight*, *age*, *benign prostate hyperplasia* and *capsular penetration* should be quantitative variables; *seminal vesicle invasion* and *gleason score* should be qualitative variables.

Then we analyze the quantitative variables. To get a first sight of the relationship between the quantitative variables, we calculate the correlation matrix between them: From the table, we find that PSA has a moderate re-

lationship with *cancer volume* and *capsular penetration*. We also draw the histogram of these variables (Figure 2). It is seen that PSA, *cancer volume*, *weight*, *age* and *capsular penetration* have a strong pattern of right skewed. We consider doing some transformation, so doing the Box-Cox procedure (Figure 3): It shows a log transformation should be done with the PSA variable. After that, we want to know whether a transformation is needed for the predictors, so we do another scatterplot (Figure 4). From this figure, we find that there is a logarithm pattern between $\log(PSA)$ and *cancer volume*. So we do a log transformation to *cancer volume*. In addition, we check whether there is multicollinearity between quantitative variables. Since VIF of each variable is between 1 and 1.7, so there is no indication of high multicollinearity.

For the qualitative variables (*seminal vesicle invasion* and *gleason score*), we draw a side-by-side boxplot to see whether they have effects on the response variabls (Figure 5): We find that both of them have an impact on PSA. Moreover, we draw the interaction boxplot, to see whether the interaction term should be included in this model (Figure 6). We find that there is a mild interaction between them.

Based on these preliminary investigations, we decide:

- to use $\log(PSA)$ as the response variable
- to use log(*cancer volume*) instead of *cancer volume*
- not to include the interaction term

Next we should examine whether all predictors are needed, and also whether the first order or the second order should be included in this model.

2.2 Model Selection

To justify whether a model outperforms other models, we need to test their performance on the validation set. Since we have 97 samples in hand, we randomly split them into a train set with 70 samples, and a test set with 27 samples.

2.2.1 First Order Model

First we fit the first order model with all predictors. From the output, we find that some of the coefficients are not significant, so we use stepwise procedure to find the best fist order model with different criteria: AIC and BIC, and different direction, forward and backward.

We anticipate four models, but we find that forward and backward select the same model within each criterion. So we result in these two first order models:

$$\begin{aligned} \log(PSA) &= \beta_0 + \beta_1 * \log(cancer\ volume) + \beta_2 * weight + \\ &\quad \beta_3 * seminal\ vesicle\ invasion1 + \beta_4 * age + \\ &\quad \beta_5 * gleason\ score7 + \beta_6 * gleason\ score8 + \\ &\quad \beta_7 * benign\ prostate\ hyperplasia \end{aligned} \tag{1}$$

$$\begin{aligned} \log(PSA) &= \beta_0 + \beta_1 * \log(cancer\ volume) + \beta_2 * weight + \\ &\quad \beta_3 * seminal\ vesicle\ invasion1 + \beta_4 * age \end{aligned} \tag{2}$$

Model 1 is selected by AIC, and model 2 is selected by BIC. We want to discuss some points here:

- Since BIC puts more penalty on the number of predictors, so BIC tends to select simpler model. Here model 2 is a submodel of model1.
- model1 contains *gleason score* and *benign prostate hyperplasia*, which are not in model2.
- We will reserve both these two models, and we want to find whether second order terms can be added into them.

2.2.2 Second Order Model

First we fit the second order model with all predictors. Since there are so many second order predictors, we need the stepwise procedure to select a subset of them. Again, we use both the AIC and BIC as the criteria, but now we do not start from the null model or full model. For AIC, we start from model 1, while for BIC, we start from model 2. The reason we start from the corresponding first order model is that we want to know whether

there is some improvement on the first order model when adding second order terms. The improvement can be a decrease either in AIC or BIC. The models are shown below:

$$\begin{aligned}\log(PSA) &= \beta_0 + \beta_1 * \log(cancer\ volume) + \beta_2 * weight + \beta_3 * age + \\ &\quad \beta_4 * benign\ prostate\ hyperplasia * seminal\ vesicle\ invasion1 + \\ &\quad \beta_5 * seminal\ vesicle\ invasion1 + \beta_6 * gleason\ score7 + \\ &\quad \beta_7 * gleason\ score8 + \beta_8 * benign\ prostate\ hyperplasia^2 + \\ &\quad \beta_9 * benign\ prostate\ hyperplasia \qquad (3)\end{aligned}$$

$$\begin{aligned}\log(PSA) &= \beta_0 + \beta_1 * \log(cancer\ volume) + \beta_2 * weight + \\ &\quad \beta_3 * seminal\ vesicle\ invasion1 + \beta_4 * age \qquad (4)\end{aligned}$$

From above:

- The output suggests model 3 has more predictors than model 1, while model 4 is identical with model 2.
- The second order terms have no improvement on BIC criterion.

2.2.3 Model Summary

Now we have three models in hand. The next step is to do model diagnostic and test them on the validation data to select the best model.

2.3 Model Diagnostic and Validation

2.3.1 Model Diagnostic

Now we compare the residual vs fitted value plot and qqplot (Figure 7).

From the residual vs fitted value plots, no abnormal pattern is seen. We observe equal spread of residuals along the x axis, which means there is no indication of unequal variance. The relationship is linear.

From the qqplots, that of model 3 shows a slight pattern of heavy tailed. The other two qqplots shows a straight line pattern.

We think that all of these three models pass the model diagnostic, and we will use model validation to select the best one.

2.3.2 Model Validation

To select the best model, we will calculate the MSPE on the test data:

$$MSPE_v = \frac{\sum_{j=1}^{m}(Y_j - \hat{Y}_j)^2}{m}$$

where $m = 27$, $Y = \log(PSA)$.

The MSPE for model 1 is 2.380, model 2 is 4.996, model 3 is 3.799. We do not compare the fitted coefficients between train and test data within each model, since there are so few samples here, even a good model can have a flexible coefficient. MSPE is much better in this situation.

Model 1 has the smallest MSPE, which means it has the best predict performance. Model 2 has less predictors than model 1, which means that it is underfit, and model 3 has more predictors, which mean it is overfit.

So we choose model 1 as our final model. Next we will fit our final model to the whole data.

2.4 Final Model Analysis

Our final model is:

$$\begin{aligned}\log(PSA) &= 2.124 + 0.504 * \log(cancer\ volume) + 0.002 * weight + \\ &\quad 0.651 * seminal\ vesicle\ invasion1 - 0.015 * age + \\ &\quad 0.128 * gleason\ score7 + 0.640 * gleason\ score8 + \\ &\quad 0.072 * benign\ prostate\ hyperplasia \qquad (5)\end{aligned}$$

The model summary is shown in the output. And we will plot model diagnostic figures (Figure 8). We find that the 32th case has a very high Cook's distance. So we consider remove it from the model and see whether the results get better.

The new model coefficient is shown below:

$$\begin{aligned}\log(PSA) &= 2.042 + 0.489 * \log(\textit{cancer volume}) + 0.011 * \textit{weight} + \\ &\quad 0.628 * \textit{seminal vesicle invasion}1 - 0.017 * \textit{age} + \\ &\quad 0.136 * \textit{gleason score}7 + 0.574 * \textit{gleason score}8 + \\ &\quad 0.043 * \textit{benign prostate hyperplasia} \qquad (6)\end{aligned}$$

We see that the coefficients change a lot, and the adjusted R^2 becomes higher, from 0.6221 to 0.6385. In addition, residual standard error decreases from 0.7092 to 0.6967. We also make model diagnostic here (Figure 9), and the results indicate a good fit.

So model 6 is our final result.

3 Conclusions and Discussion

- The variables log(*cancer volume*), *weight*, *age*, *benign prostate hyperplasia*, *seminal vesicle invasion* and *gleason score* have an effect on the response variable. We do not use *capsular penetration*.
- We derive a linear relationship between predictors and response variables.
- All predictors, excluding the *age* have a positive effect on the response varible, while the *age* has a negative effect on the reponse variable.
- These predictors are adequate since the adjusted R^2 is equal to 0.6385.
- Our results are general since we fit the model on the train data and test it on the validation data.
- The limitation is, it is highly random when we split our data into train and test. Since the size of data is too small, say 97 smaples, the randomness of data will influence our model. The steps to solve this limitation is to either increase the sample size (by bootstrap or have more data), or use cross-validation (LOOCV or 5-fold cross validation) to fit our model.

References

[1] Stamey, Thomas A., et al. "Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients." The Journal of urology 141.5 (1989): 1076-1083.

Appendices

A Figures

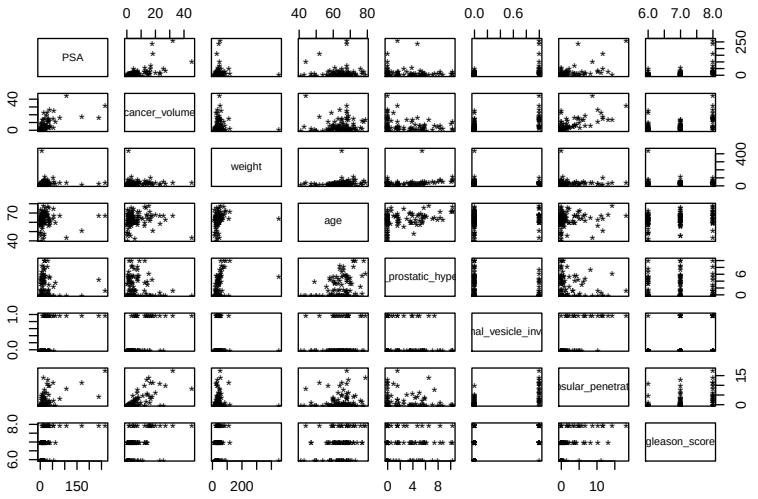

Figure 1: scatter plot of quantitative variables

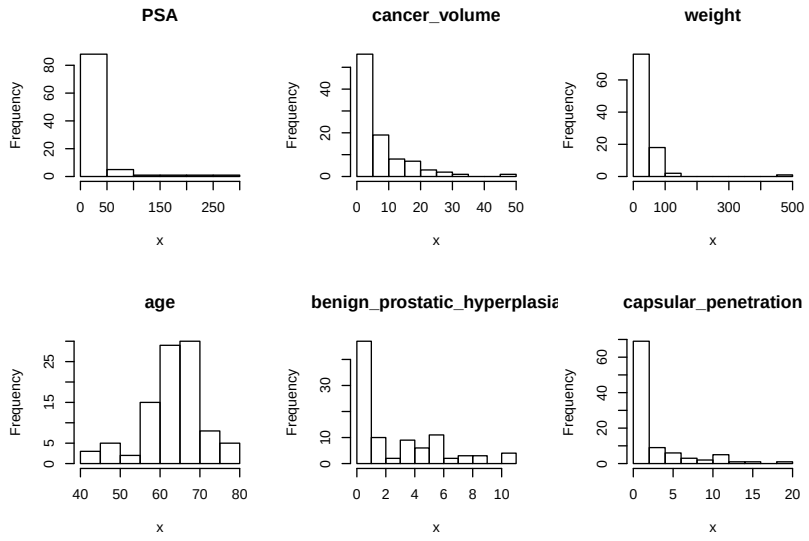

Figure 2: histogram of quantitative variables

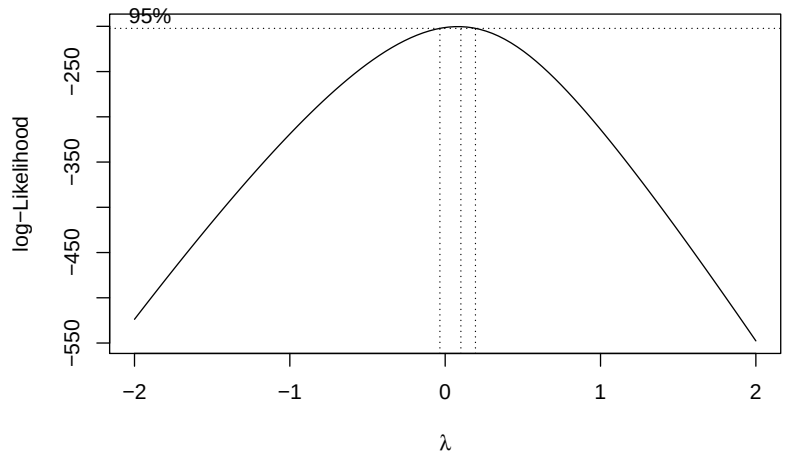

Figure 3: Box-Cox plot of PSA

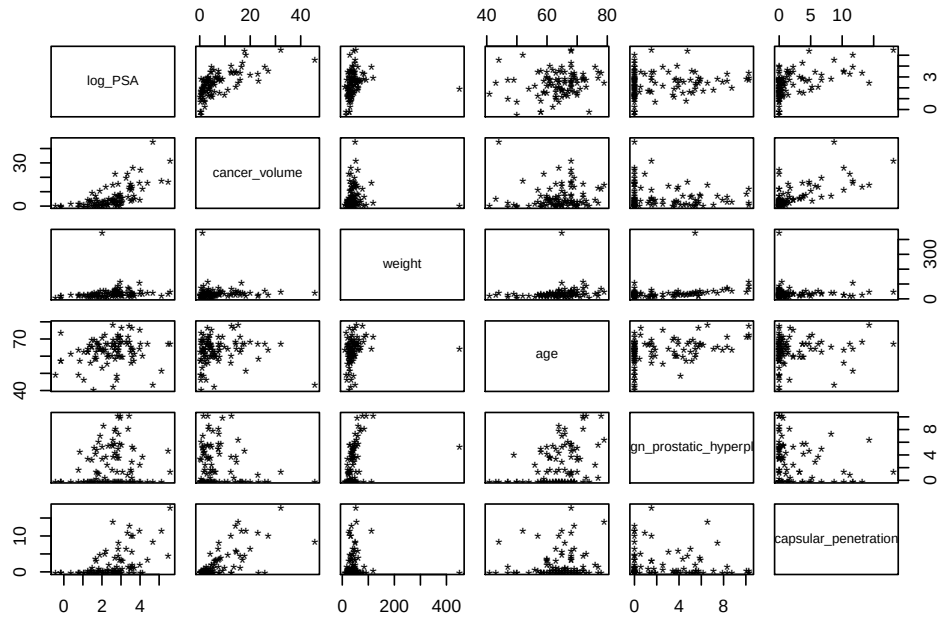

Figure 4: scatter plot of transformed quantitative variables

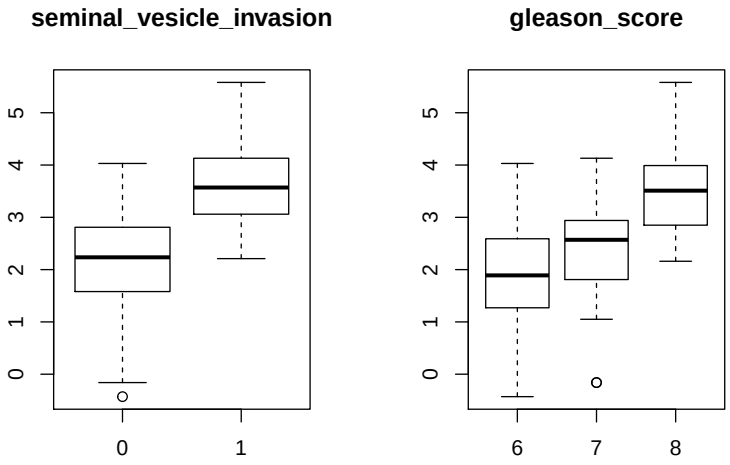

Figure 5: side by side boxplot of qualitative variables

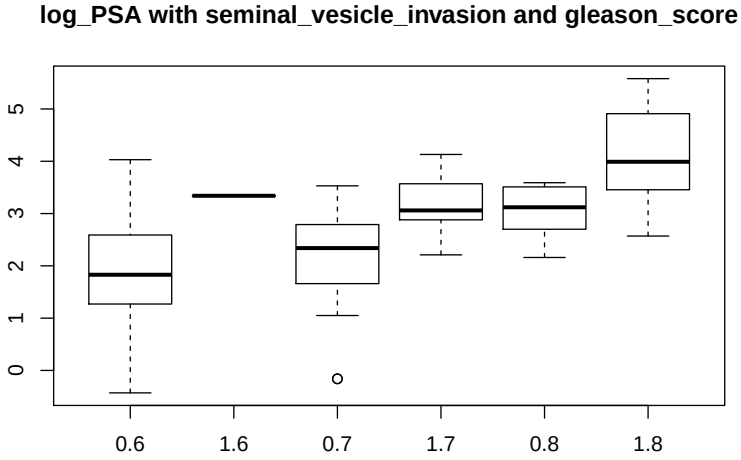

Figure 6: side by side boxplot of interaction qualitative variables

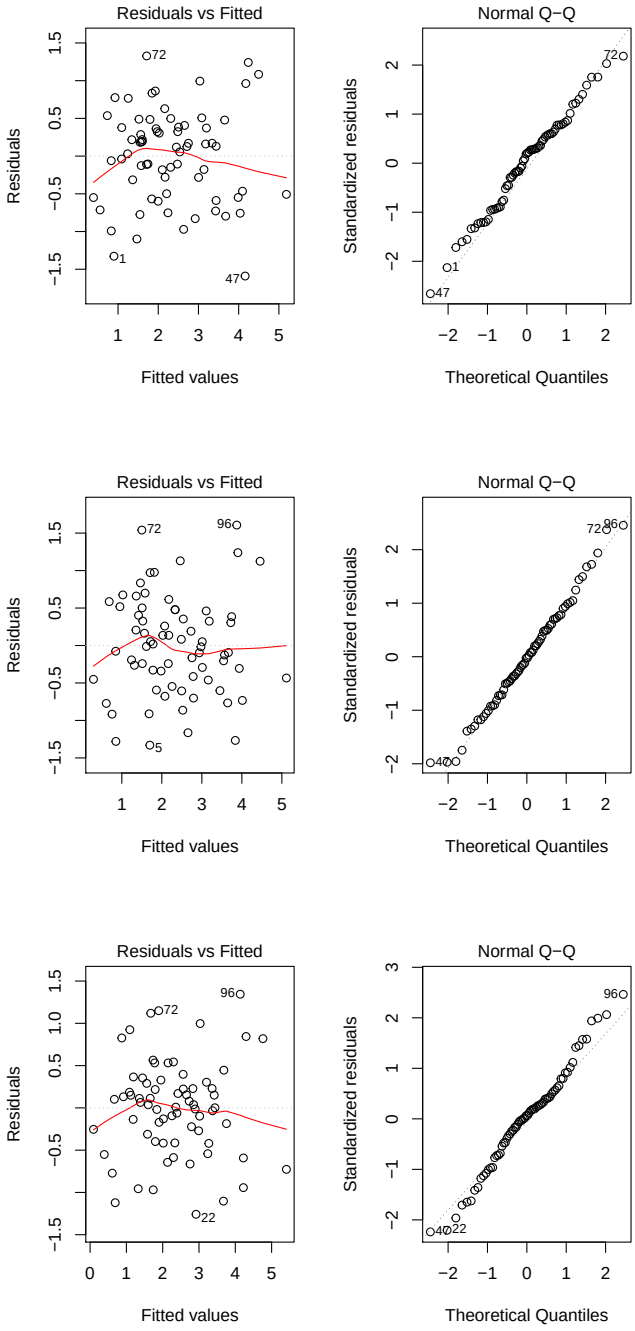

Figure 7: residual plots and qqplots of model 1, 2, 3

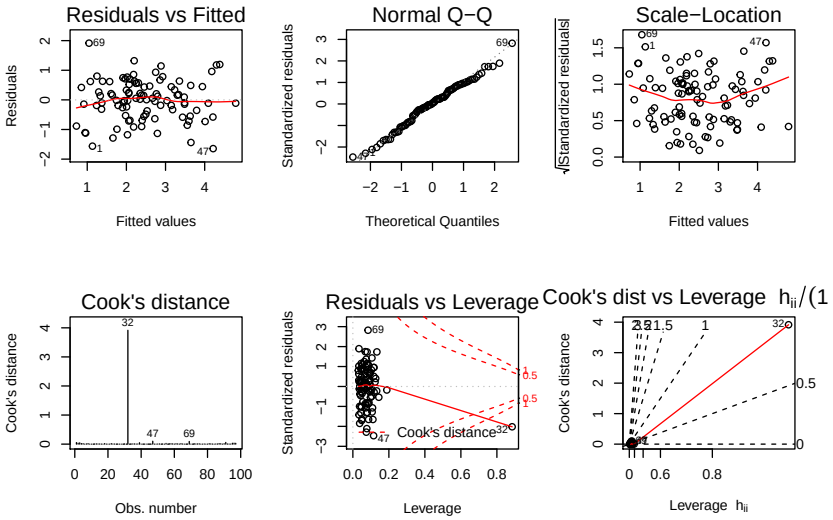

Figure 8: model diagnostics of final model

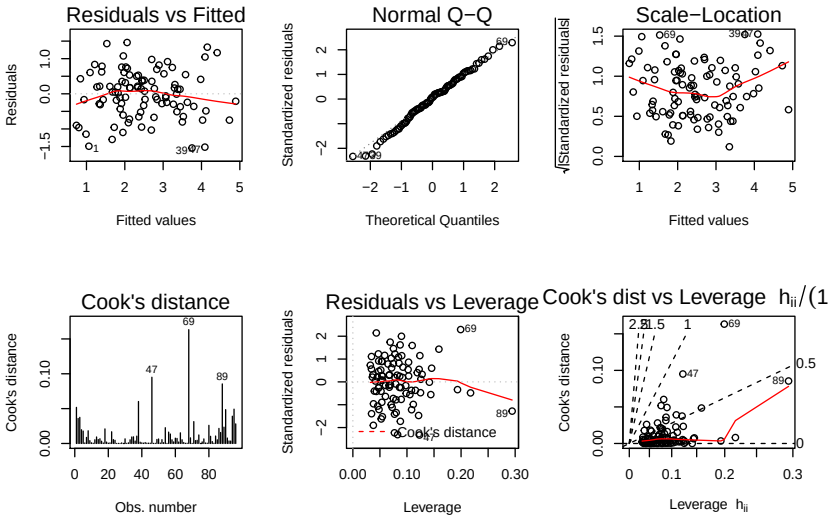

Figure 9: model diagnostics of final model (with 32th observation deleted)

B Code and Output

```
# ==================== 1. Data
# Exploration
# ====================
prostate <- read.table("~/Desktop/prostate.txt")

prostate <- prostate[, -1]
names(prostate) <- c("PSA", "cancer_volume",
    "weight", "age", "benign_prostatic_hyperplasia",
    "seminal_vesicle_invasion",
    "capsular_penetration", "gleason_score")

# Use pairwise scatter plot to
# distinguish quantitative
# variables and qualitative
# variables
pairs(prostate, pch = "*")

quantitative_vars <- c("PSA", "cancer_volume",
    "weight", "age", "benign_prostatic_hyperplasia",
    "capsular_penetration")
qualitative_vars <- c("seminal_vesicle_invasion",
    "gleason_score")
prostate$seminal_vesicle_invasion <-
as.factor(prostate$seminal_vesicle_invasion)
prostate$gleason_score <- as.factor(prostate$gleason_score)

# Draw the correlation matrix
# for the quantitatives
cor(prostate[quantitative_vars])
```

```
##                                       PSA
## PSA                            1.00000000
## cancer_volume                  0.62415059
## weight                         0.02621343
## age                            0.01719938
## benign_prostatic_hyperplasia -0.01648649
## capsular_penetration           0.55079252
##                                cancer_volume
## PSA                            0.624150588
## cancer_volume                  1.000000000
## weight                         0.005107148
## age                            0.039094423
## benign_prostatic_hyperplasia  -0.133209431
## capsular_penetration           0.692896688
##                                    weight
## PSA                            0.026213430
## cancer_volume                  0.005107148
## weight                         1.000000000
```

```
## age                           0.164323714
## benign_prostatic_hyperplasia 0.321848748
## capsular_penetration         0.001578905
##                                      age
## PSA                           0.01719938
## cancer_volume                 0.03909442
## weight                        0.16432371
## age                           1.00000000
## benign_prostatic_hyperplasia 0.36634121
## capsular_penetration         0.09955535
##                              benign_prostatic_hyperplasia
## PSA                                            -0.01648649
## cancer_volume                                  -0.13320943
## weight                                          0.32184875
## age                                             0.36634121
## benign_prostatic_hyperplasia                    1.00000000
## capsular_penetration                           -0.08300865
##                              capsular_penetration
## PSA                                     0.550792517
## cancer_volume                           0.692896688
## weight                                  0.001578905
## age                                     0.099555351
## benign_prostatic_hyperplasia           -0.083008649
## capsular_penetration                    1.000000000
```

```
# Plot histograms for each
# quantitative variables
par(mfrow = c(2, 3))
invisible(Map(function(x, y) hist(x,
    main = y), prostate[quantitative_vars],
    quantitative_vars))

# Use Box-Cox procedure to find
# the best transformation of
# the response variable
fit1 <- lm(PSA ~ ., data = prostate[quantitative_vars])
library(MASS)
par(mfrow = c(1, 1))
boxcox(fit1)

# Do a log transformation on
# the response variable
prostate$log_PSA <- log(prostate$PSA)
prostate <- prostate[c("log_PSA",
    "cancer_volume", "weight",
    "age", "benign_prostatic_hyperplasia",
    "seminal_vesicle_invasion",
    "capsular_penetration", "gleason_score")]
quantitative_vars <- c("log_PSA",
    "cancer_volume", "weight",
```

```
    "age", "benign_prostatic_hyperplasia",
    "capsular_penetration")


# Plot a new pairwise scatter
# plot and histograms to see
# whether transformations are
# needed for predictors
pairs(prostate[!(names(prostate) %in%
    qualitative_vars)], pch = "*")

# From the scatter plot, it
# seems that the predictor
# cancer volume needs a log
# transformation
prostate$log_cancer_volume <- log(prostate$cancer_volume)
prostate <- prostate[c("log_PSA",
    "log_cancer_volume", "weight",
    "age", "benign_prostatic_hyperplasia",
    "seminal_vesicle_invasion",
    "capsular_penetration", "gleason_score")]
quantitative_vars <- c("log_PSA",
    "log_cancer_volume", "weight",
    "age", "benign_prostatic_hyperplasia",
    "capsular_penetration")

# Check multicollinearity
solve(cor(prostate[c("log_cancer_volume",
    "weight", "age", "benign_prostatic_hyperplasia",
    "capsular_penetration")]))
```

```
##                               log_cancer_volume
## log_cancer_volume                    1.69716394
## weight                               0.06112201
## age                                 -0.27296326
## benign_prostatic_hyperplasia        -0.04594090
## capsular_penetration                -1.02625401
##                                      weight
## log_cancer_volume                 0.06112201
## weight                            1.12144849
## age                              -0.06615563
## benign_prostatic_hyperplasia     -0.34328385
## capsular_penetration             -0.06147763
##                                         age
## log_cancer_volume                -0.27296326
## weight                           -0.06615563
## age                               1.22488972
## benign_prostatic_hyperplasia     -0.41982333
## capsular_penetration              0.01211048
##                                benign_prostatic_hyperplasia
```

```
## log_cancer_volume                                -0.0459409
## weight                                           -0.3432838
## age                                              -0.4198233
## benign_prostatic_hyperplasia                      1.2800890
## capsular_penetration                              0.1770058
##                               capsular_penetration
## log_cancer_volume                       -1.02625401
## weight                                  -0.06147763
## age                                      0.01211048
## benign_prostatic_hyperplasia             0.17700579
## capsular_penetration                     1.64821553
```

```r
# Draw side-by-side boxplot to
# see whether the qualitative
# variables have effects on the
# response variable
par(mfrow = c(1, 2))
invisible(Map(function(x, y, z) boxplot(y ~
    x, main = z), prostate[qualitative_vars],
    data.frame(prostate$log_PSA,
        prostate$log_PSA), qualitative_vars))

par(mfrow = c(1, 1))
boxplot(prostate$log_PSA ~
prostate$seminal_vesicle_invasion:prostate$gleason_score,
    main = "log_PSA with seminal_vesicle_invasion and gleason_score")

# ================== 2. Model
# Selection
# ====================

# cross-validation split
set.seed(1)
train <- sample(nrow(prostate),
    70)
test <- setdiff(1:nrow(prostate),
    train)




# 2.1 First Order Model

fit2 <- lm(log_PSA ~ ., data = prostate,
    subset = train)
summary(fit2)
```

```
##
## Call:
## lm(formula = log_PSA ~ ., data = prostate, subset = train)
```

```
##
## Residuals:
##     Min      1Q  Median
## -1.4276 -0.5052  0.1234
##      3Q     Max
##  0.3451  1.3721
##
## Coefficients:
##                               Estimate
## (Intercept)                   2.364026
## log_cancer_volume             0.509867
## weight                        0.016241
## age                          -0.028436
## benign_prostatic_hyperplasia  0.066111
## seminal_vesicle_invasion1     0.907232
## capsular_penetration         -0.031721
## gleason_score7                0.144602
## gleason_score8                0.714372
##                              Std. Error
## (Intercept)                    0.676513
## log_cancer_volume              0.088720
## weight                         0.007795
## age                            0.011084
## benign_prostatic_hyperplasia   0.038260
## seminal_vesicle_invasion1      0.313160
## capsular_penetration           0.035328
## gleason_score7                 0.186805
## gleason_score8                 0.268939
##                              t value
## (Intercept)                    3.494
## log_cancer_volume              5.747
## weight                         2.084
## age                           -2.565
## benign_prostatic_hyperplasia   1.728
## seminal_vesicle_invasion1      2.897
## capsular_penetration          -0.898
## gleason_score7                 0.774
## gleason_score8                 2.656
##                              Pr(>|t|)
## (Intercept)                  0.000891
## log_cancer_volume             3.1e-07
## weight                       0.041394
## age                          0.012778
## benign_prostatic_hyperplasia 0.089063
## seminal_vesicle_invasion1    0.005225
## capsular_penetration         0.372760
## gleason_score7               0.441874
## gleason_score8               0.010069
##
## (Intercept)                  ***
```

```
## log_cancer_volume            ***
## weight                       *
## age                          *
## benign_prostatic_hyperplasia .
## seminal_vesicle_invasion1    **
## capsular_penetration
## gleason_score7
## gleason_score8               *
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01
##   '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6566 on 61 degrees of freedom
## Multiple R-squared:  0.7494, Adjusted R-squared:  0.7165
## F-statistic:  22.8 on 8 and 61 DF,  p-value: 1.161e-15

# Use forward stepwise and
# backward stepwise to find the
# best first-order model under
# AIC
model1aicf <- invisible(stepAIC(lm(log_PSA ~
    1, data = prostate, subset = train),
    scope = list(lower = lm(log_PSA ~
        1, data = prostate, subset = train),
        upper = lm(log_PSA ~ .,
            data = prostate, subset = train)),
    direction = "both"))

model1aicb <- invisible(stepAIC(lm(log_PSA ~
    ., data = prostate, subset = train),
    scope = list(lower = lm(log_PSA ~
        1, data = prostate, subset = train),
        upper = lm(log_PSA ~ .,
            data = prostate, subset = train)),
    direction = "both"))

# Use forward stepwise and
# backward stepwise to find the
# best first-order model under
# BIC
model1bicf <- invisible(stepAIC(lm(log_PSA ~
    1, data = prostate, subset = train),
    scope = list(lower = lm(log_PSA ~
        1, data = prostate, subset = train),
        upper = lm(log_PSA ~ .,
            data = prostate, subset = train)),
    direction = "both", k = log(length(train))))

model1bicb <- invisible(stepAIC(lm(log_PSA ~
```

```
    ., data = prostate, subset = train),
    scope = list(lower = lm(log_PSA ~
        1, data = prostate, subset = train),
        upper = lm(log_PSA ~ .,
            data = prostate, subset = train)),
    direction = "both", k = log(length(train))))

sapply(list(model1aicf = model1aicf,
    model1aicb = model1aicb, model1bicf = model1bicf,
    model1bicb = model1bicb), "[[",
    "call")

## $model1aicf
## lm(formula = log_PSA ~ log_cancer_volume + weight +
seminal_vesicle_invasion +
##     age + gleason_score + benign_prostatic_hyperplasia, data = prostate,
##     subset = train)
##
## $model1aicb
## lm(formula = log_PSA ~ log_cancer_volume + weight + age +
benign_prostatic_hyperplasia +
##     seminal_vesicle_invasion + gleason_score, data = prostate,
##     subset = train)
##
## $model1bicf
## lm(formula = log_PSA ~ log_cancer_volume + weight +
seminal_vesicle_invasion +
##     age, data = prostate, subset = train)
##
## $model1bicb
## lm(formula = log_PSA ~ log_cancer_volume + weight + age +
seminal_vesicle_invasion,
##     data = prostate, subset = train)

# 2.2 second order model

# Fit the second-order model
fit3 <- lm(log_PSA ~ .^2 + I(age^2) +
    I(log_cancer_volume^2) + I(weight^2) +
    I(benign_prostatic_hyperplasia^2) +
    I(capsular_penetration^2),
    data = prostate, subset = train)

# Use stepwise to find the best
# second-order model based on
# model1aic under AIC
model2aicaic <- invisible(stepAIC(model1aicb,
    scope = list(lower = lm(log_PSA ~
        1, data = prostate, subset = train),
        upper = fit3), direction = "both"))
```

```r
# Use stepwise to find the best
# second-order model based on
# model1bic under BIC
model2bicbic <- invisible(stepAIC(model1bicb,
    scope = list(lower = lm(log_PSA ~
        1, data = prostate, subset = train),
        upper = fit3), direction = "both",
    k = log(length(train))))

sapply(list(model2aicaic = model2aicaic,
    model2bicbic = model2bicbic),
    "[[", "call")

## $model2aicaic
## lm(formula = log_PSA ~ log_cancer_volume + weight + age +
benign_prostatic_hyperplasia +
##     seminal_vesicle_invasion + gleason_score +
I(benign_prostatic_hyperplasia^2) +
##     benign_prostatic_hyperplasia:seminal_vesicle_invasion, data =
prostate,
##     subset = train)
##
## $model2bicbic
## lm(formula = log_PSA ~ log_cancer_volume + weight + age +
seminal_vesicle_invasion,
##     data = prostate, subset = train)

# ================== 3. Model
# Diagnostic and Validation
# ==========================


# 3.1 model diagnostic
par(mfrow = c(1, 2))
plot(model1aicb, which = c(1, 2))

plot(model1bicb, which = c(1, 2))

plot(model2aicaic, which = c(1,
    2))

# 3.2 Model validation
model1aicb_mspe <- sum((prostate$log_PSA[test] -
    predict(model1aicb, prostate[test,
        ]))^2)/length(test)
model1bicb_mspe <- sum((prostate$log_PSA[test] -
    predict(model1bicb, prostate[test,
        ]))^2)/length(test)
model2aicaic_mspe <- sum((prostate$log_PSA[test] -
    predict(model2aicaic, prostate[test,
        ]))^2)/length(test)
```

```r
# ================ 4. Final
# Model Analysis
# ========================
model_final <- lm(formula = log_PSA ~
    log_cancer_volume + weight +
        age + benign_prostatic_hyperplasia +
        seminal_vesicle_invasion +
        gleason_score, data = prostate)

summary(model_final)
```

```
##
## Call:
## lm(formula = log_PSA ~ log_cancer_volume + weight + age +
benign_prostatic_hyperplasia +
##     seminal_vesicle_invasion + gleason_score, data = prostate)
##
## Residuals:
##      Min       1Q   Median
## -1.64579 -0.48659  0.02146
##       3Q      Max
##  0.48710  1.91447
##
## Coefficients:
##                                Estimate
## (Intercept)                    2.124108
## log_cancer_volume              0.504257
## weight                         0.001849
## age                           -0.014547
## benign_prostatic_hyperplasia   0.072115
## seminal_vesicle_invasion1      0.651022
## gleason_score7                 0.127880
## gleason_score8                 0.639699
##                               Std. Error
## (Intercept)                     0.657133
## log_cancer_volume               0.080054
## weight                          0.001700
## age                             0.010876
## benign_prostatic_hyperplasia    0.027562
## seminal_vesicle_invasion1       0.215670
## gleason_score7                  0.174490
## gleason_score8                  0.240184
##                               t value
## (Intercept)                     3.232
## log_cancer_volume               6.299
## weight                          1.087
## age                            -1.338
## benign_prostatic_hyperplasia    2.616
```

```
## seminal_vesicle_invasion1      3.019
## gleason_score7                 0.733
## gleason_score8                 2.663
##                               Pr(>|t|)
## (Intercept)                    0.00172
## log_cancer_volume             1.11e-08
## weight                         0.27979
## age                            0.18445
## benign_prostatic_hyperplasia   0.01044
## seminal_vesicle_invasion1      0.00331
## gleason_score7                 0.46556
## gleason_score8                 0.00918
##
## (Intercept)                   **
## log_cancer_volume             ***
## weight
## age
## benign_prostatic_hyperplasia  *
## seminal_vesicle_invasion1     **
## gleason_score7
## gleason_score8                **
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01
##   '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7092 on 89 degrees of freedom
## Multiple R-squared:  0.6497, Adjusted R-squared:  0.6221
## F-statistic: 23.58 on 7 and 89 DF,  p-value: < 2.2e-16
```

```
anova(model_final)
```

```
## Analysis of Variance Table
##
## Response: log_PSA
##                               Df
## log_cancer_volume              1
## weight                         1
## age                            1
## benign_prostatic_hyperplasia   1
## seminal_vesicle_invasion       1
## gleason_score                  2
## Residuals                     89
##                               Sum Sq
## log_cancer_volume             68.801
## weight                         1.891
## age                            0.032
## benign_prostatic_hyperplasia   1.796
## seminal_vesicle_invasion       6.618
## gleason_score                  3.868
```

```
## Residuals                     44.763
##                               Mean Sq
## log_cancer_volume              68.801
## weight                          1.891
## age                             0.032
## benign_prostatic_hyperplasia    1.796
## seminal_vesicle_invasion        6.618
## gleason_score                   1.934
## Residuals                       0.503
##                               F value
## log_cancer_volume            136.7917
## weight                         3.7600
## age                            0.0638
## benign_prostatic_hyperplasia   3.5700
## seminal_vesicle_invasion      13.1582
## gleason_score                  3.8451
## Residuals
##                                  Pr(>F)
## log_cancer_volume            < 2.2e-16
## weight                       0.0556592
## age                          0.8012397
## benign_prostatic_hyperplasia 0.0620865
## seminal_vesicle_invasion     0.0004769
## gleason_score                0.0250221
## Residuals
##
## log_cancer_volume            ***
## weight                       .
## age
## benign_prostatic_hyperplasia .
## seminal_vesicle_invasion     ***
## gleason_score                *
## Residuals
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01
##   '*' 0.05 '.' 0.1 ' ' 1
```

```r
par(mfrow = c(2, 3))
plot(model_final, which = 1:6)
```

```r
model_final_del <- lm(formula = log_PSA ~
    log_cancer_volume + weight +
        age + benign_prostatic_hyperplasia +
        seminal_vesicle_invasion +
        gleason_score, data = prostate[-32,
    ])
summary(model_final_del)
```

```
##
## Call:
## lm(formula = log_PSA ~ log_cancer_volume + weight + age +
benign_prostatic_hyperplasia +
##     seminal_vesicle_invasion + gleason_score, data = prostate[-32,
##     ])
##
## Residuals:
##      Min       1Q   Median
## -1.54232 -0.38308  0.05055
##       3Q      Max
##  0.44597  1.46036
##
## Coefficients:
##                                Estimate
## (Intercept)                    2.041883
## log_cancer_volume              0.488958
## weight                         0.011098
## age                           -0.017439
## benign_prostatic_hyperplasia   0.042962
## seminal_vesicle_invasion1      0.627956
## gleason_score7                 0.135673
## gleason_score8                 0.573952
##                               Std. Error
## (Intercept)                     0.646759
## log_cancer_volume               0.078991
## weight                          0.004797
## age                             0.010776
## benign_prostatic_hyperplasia    0.030561
## seminal_vesicle_invasion1       0.212156
## gleason_score7                  0.171449
## gleason_score8                  0.238096
##                               t value
## (Intercept)                     3.157
## log_cancer_volume               6.190
## weight                          2.313
## age                            -1.618
## benign_prostatic_hyperplasia    1.406
## seminal_vesicle_invasion1       2.960
## gleason_score7                  0.791
## gleason_score8                  2.411
##                               Pr(>|t|)
## (Intercept)                    0.00218
## log_cancer_volume             1.86e-08
## weight                         0.02303
## age                            0.10916
## benign_prostatic_hyperplasia   0.16331
## seminal_vesicle_invasion1      0.00395
## gleason_score7                 0.43088
## gleason_score8                 0.01801
```

```
##
## (Intercept)                   **
## log_cancer_volume             ***
## weight                        *
## age
## benign_prostatic_hyperplasia
## seminal_vesicle_invasion1     **
## gleason_score7
## gleason_score8                *
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01
##   '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6967 on 88 degrees of freedom
## Multiple R-squared:  0.6651, Adjusted R-squared:  0.6385
## F-statistic: 24.97 on 7 and 88 DF,  p-value: < 2.2e-16
```

```
anova(model_final_del)
```

```
## Analysis of Variance Table
##
## Response: log_PSA
##                               Df
## log_cancer_volume              1
## weight                         1
## age                            1
## benign_prostatic_hyperplasia   1
## seminal_vesicle_invasion       1
## gleason_score                  2
## Residuals                     88
##                               Sum Sq
## log_cancer_volume             68.720
## weight                         6.439
## age                            0.690
## benign_prostatic_hyperplasia   0.122
## seminal_vesicle_invasion       5.913
## gleason_score                  2.954
## Residuals                     42.710
##                               Mean Sq
## log_cancer_volume              68.720
## weight                          6.439
## age                             0.690
## benign_prostatic_hyperplasia    0.122
## seminal_vesicle_invasion        5.913
## gleason_score                   1.477
## Residuals                       0.485
##                                F value
## log_cancer_volume             141.5895
## weight                         13.2661
```

```
## age                              1.4214
## benign_prostatic_hyperplasia     0.2504
## seminal_vesicle_invasion        12.1839
## gleason_score                    3.0428
## Residuals
##                                    Pr(>F)
## log_cancer_volume               < 2.2e-16
## weight                          0.0004560
## age                             0.2363730
## benign_prostatic_hyperplasia    0.6180218
## seminal_vesicle_invasion        0.0007557
## gleason_score                   0.0527511
## Residuals
##
## log_cancer_volume               ***
## weight                          ***
## age
## benign_prostatic_hyperplasia
## seminal_vesicle_invasion        ***
## gleason_score                   .
## Residuals
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01
##   '*' 0.05 '.' 0.1 ' ' 1
```

```r
par(mfrow = c(2, 3))
plot(model_final_del, which = 1:6)
```