Statistical Learning Homework 1

110024516 邱繼賢

```
library(dplyr)
library(psych)
library(latex2exp)
library(knitr)
library(tibble)
library(summarytools)
```

Problem 1.

(a) Exploratory data analysis (EDA) among 4 variables

```
data1 = read.csv("ozone.csv")
data1 = data1[,c(2,3,4,1)]
dim(data1)
```

```
## [1] 111   4
```

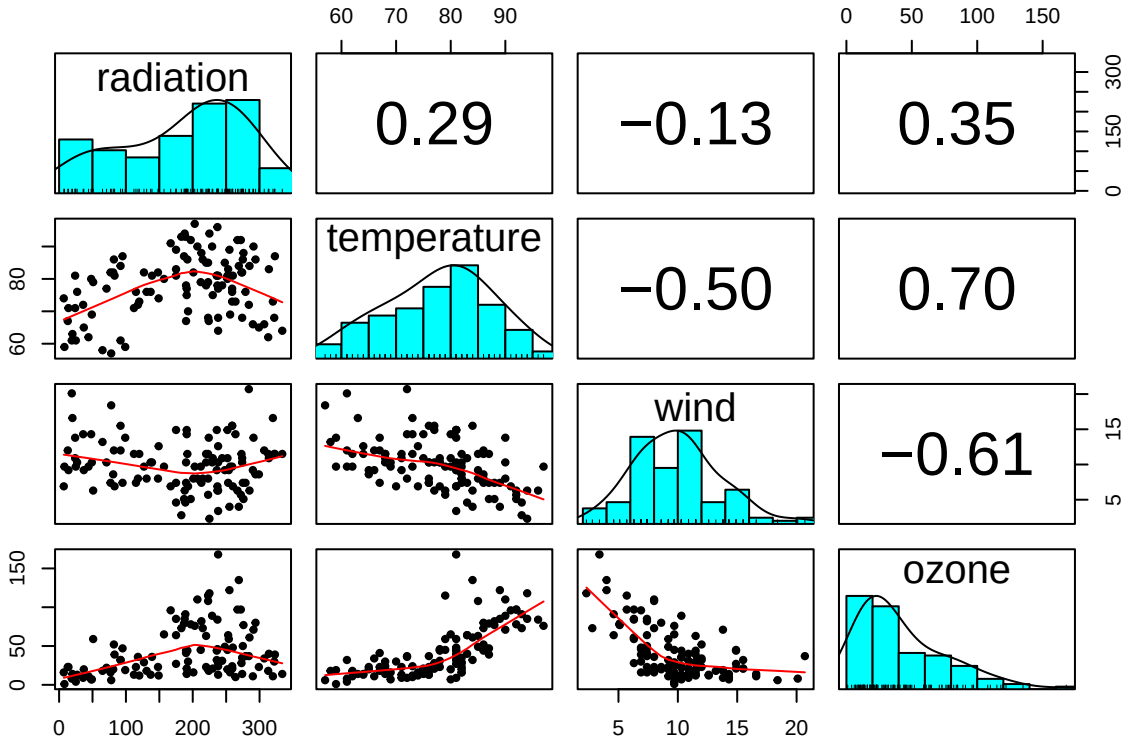
```
summary(data1)
```

```
##    radiation      temperature         wind            ozone      
##  Min.   :  7.0   Min.   :57.00   Min.   : 2.300   Min.   :  1.0  
##  1st Qu.:113.5   1st Qu.:71.00   1st Qu.: 7.400   1st Qu.: 18.0  
##  Median :207.0   Median :79.00   Median : 9.700   Median : 31.0  
##  Mean   :184.8   Mean   :77.79   Mean   : 9.939   Mean   : 42.1  
##  3rd Qu.:255.5   3rd Qu.:84.50   3rd Qu.:11.500   3rd Qu.: 62.0  
##  Max.   :334.0   Max.   :97.00   Max.   :20.700   Max.   :168.0
```

- 此筆資料共 111 個觀測值，4 個變數

- 4 個變數皆為連續型變數
- 粗略觀察各變數級距，並沒有發現明顯離群值 (outlier)

```
pairs.panels(data1, ellipses = F)
```



- 變數 *ozone* 呈現明顯右偏現象
- 變數 *ozone* 和 *temperature* 有較強的正相關，相關係數 = 0.7
- 變數 *ozone* 和 *wind* 有較強的負相關，相關係數 = -0.61
- 變數 *temperature* 和 *wind* 有中等強度的負相關，相關係數 = -0.5，配飾模型時可能要注意此二變數的共線性
- 變數 *ozone* 和 *radiation* 散佈圖呈現些微二次函數的趨勢
- 變數 *ozone* 和 *temperature* 散佈圖呈現類似遞增的二次函數趨勢
- 變數 *ozone* 和 *wind* 散佈圖呈現類似遞減的二次函數趨勢

(b) Regression model fitting and model summaries

配飾模型

$$ozone = \beta_0 + \beta_1\, radiation + \beta_2\, temperature + \beta_3\, wind + \epsilon$$

```r
fit1.1 = lm(ozone ~ radiation + temperature + wind, data1)
summary(fit1.1)
```

```
##
## Call:
## lm(formula = ozone ~ radiation + temperature + wind, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.485 -14.210  -3.556  10.124  95.600
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -64.23208   23.04204  -2.788  0.00628 **
## radiation     0.05980    0.02318   2.580  0.01124 *
## temperature   1.65121    0.25341   6.516 2.43e-09 ***
## wind         -3.33760    0.65384  -5.105 1.45e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.17 on 107 degrees of freedom
## Multiple R-squared:  0.6062, Adjusted R-squared:  0.5952
## F-statistic: 54.91 on 3 and 107 DF,  p-value: < 2.2e-16
```

- 變數 *radiation, temperature, wind* 的效應皆呈現顯著
- 變數 *radiation, temperature, wind* 所對應係數的正負值 $(+,+,-)$，與三變數和 *ozone* 之間的相關係數正負值一致，符合直觀
- $R^2 = 60.62\%$ 模型表現還不是很好

對模型加入各變數的二次項和交互作用項：

```r
fit1.2 = lm(ozone ~ .^2 + I(radiation^2) + I(temperature^2) + I(wind^2), data1)
summary(fit1.2)
```

```
##
## Call:
## lm(formula = ozone ~ .^2 + I(radiation^2) + I(temperature^2) +
```

```
##     I(wind^2), data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.611 -11.455  -2.901   8.548  70.325
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.245e+02  1.957e+02   2.680   0.0086 **
## radiation              2.628e-02  2.142e-01   0.123   0.9026
## temperature           -1.021e+01  4.209e+00  -2.427   0.0170 *
## wind                  -2.802e+01  9.645e+00  -2.906   0.0045 **
## I(radiation^2)        -3.388e-04  2.541e-04  -1.333   0.1855
## I(temperature^2)       5.953e-02  2.382e-02   2.499   0.0141 *
## I(wind^2)              6.173e-01  1.461e-01   4.225 5.25e-05 ***
## radiation:temperature  3.750e-03  2.459e-03   1.525   0.1303
## radiation:wind        -1.127e-02  6.277e-03  -1.795   0.0756 .
## temperature:wind       1.734e-01  9.497e-02   1.825   0.0709 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.77 on 101 degrees of freedom
## Multiple R-squared:  0.7383, Adjusted R-squared:  0.715
## F-statistic: 31.66 on 9 and 101 DF,  p-value: < 2.2e-16
```

- $R^2 = 73.83$ 相較於前一個模型有所上升
- 變數 *radiation* 和 *radiation^2* 的效應皆呈現不顯著，可能是因為兩變數間的共線性造成

將一次和二次項變數皆改成 orthogonal polynomial 的形式，並考慮所有的二階和三階交互作用項，重新配飾模型：

```
fit1.3 = lm(ozone ~ poly(radiation,2)*poly(temperature,2)*poly(wind,2), data1)
summary(fit1.3)
```

```
##
## Call:
## lm(formula = ozone ~ poly(radiation, 2) * poly(temperature, 2) *
```

```
##     poly(wind, 2), data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.399  -6.827  -1.508   6.408  44.300
##
## Coefficients:
##                                                                  Estimate Std. Error
## (Intercept)                                                        38.000      3.294
## poly(radiation, 2)1                                               120.501     32.544
## poly(radiation, 2)2                                               -78.530     32.673
## poly(temperature, 2)1                                             173.184     41.342
## poly(temperature, 2)2                                               3.979     38.786
## poly(wind, 2)1                                                    -52.311     34.299
## poly(wind, 2)2                                                     65.653     58.140
## poly(radiation, 2)1:poly(temperature, 2)1                        1153.262    513.120
## poly(radiation, 2)2:poly(temperature, 2)1                        -394.670    417.559
## poly(radiation, 2)1:poly(temperature, 2)2                         263.417    387.377
## poly(radiation, 2)2:poly(temperature, 2)2                        -490.168    339.717
## poly(radiation, 2)1:poly(wind, 2)1                               -301.387    430.042
## poly(radiation, 2)2:poly(wind, 2)1                                351.981    469.993
## poly(radiation, 2)1:poly(wind, 2)2                               1051.213    609.727
## poly(radiation, 2)2:poly(wind, 2)2                               -485.203    606.135
## poly(temperature, 2)1:poly(wind, 2)1                              150.055    458.170
## poly(temperature, 2)2:poly(wind, 2)1                             1019.189    455.516
## poly(temperature, 2)1:poly(wind, 2)2                              232.678    724.893
## poly(temperature, 2)2:poly(wind, 2)2                            -1060.125    653.960
## poly(radiation, 2)1:poly(temperature, 2)1:poly(wind, 2)1        -3284.863   5202.810
## poly(radiation, 2)2:poly(temperature, 2)1:poly(wind, 2)1          715.698   5175.828
## poly(radiation, 2)1:poly(temperature, 2)2:poly(wind, 2)1          133.667   5750.070
## poly(radiation, 2)2:poly(temperature, 2)2:poly(wind, 2)1         3828.655   5504.955
## poly(radiation, 2)1:poly(temperature, 2)1:poly(wind, 2)2        13359.144   9128.664
## poly(radiation, 2)2:poly(temperature, 2)1:poly(wind, 2)2          360.395   6513.572
## poly(radiation, 2)1:poly(temperature, 2)2:poly(wind, 2)2         2495.430   5818.334
## poly(radiation, 2)2:poly(temperature, 2)2:poly(wind, 2)2        -1027.463   5793.267
##                                                                 t value Pr(>|t|)
```

```
## (Intercept)                                                    11.535  < 2e-16 ***
## poly(radiation, 2)1                                             3.703 0.000381 ***
## poly(radiation, 2)2                                            -2.404 0.018440 *
## poly(temperature, 2)1                                           4.189 6.88e-05 ***
## poly(temperature, 2)2                                           0.103 0.918544
## poly(wind, 2)1                                                 -1.525 0.130974
## poly(wind, 2)2                                                  1.129 0.262021
## poly(radiation, 2)1:poly(temperature, 2)1                       2.248 0.027226 *
## poly(radiation, 2)2:poly(temperature, 2)1                      -0.945 0.347277
## poly(radiation, 2)1:poly(temperature, 2)2                       0.680 0.498374
## poly(radiation, 2)2:poly(temperature, 2)2                      -1.443 0.152775
## poly(radiation, 2)1:poly(wind, 2)1                             -0.701 0.485345
## poly(radiation, 2)2:poly(wind, 2)1                              0.749 0.456006
## poly(radiation, 2)1:poly(wind, 2)2                              1.724 0.088373 .
## poly(radiation, 2)2:poly(wind, 2)2                             -0.800 0.425687
## poly(temperature, 2)1:poly(wind, 2)1                            0.328 0.744097
## poly(temperature, 2)2:poly(wind, 2)1                            2.237 0.027905 *
## poly(temperature, 2)1:poly(wind, 2)2                            0.321 0.749020
## poly(temperature, 2)2:poly(wind, 2)2                           -1.621 0.108747
## poly(radiation, 2)1:poly(temperature, 2)1:poly(wind, 2)1       -0.631 0.529517
## poly(radiation, 2)2:poly(temperature, 2)1:poly(wind, 2)1        0.138 0.890352
## poly(radiation, 2)1:poly(temperature, 2)2:poly(wind, 2)1        0.023 0.981509
## poly(radiation, 2)2:poly(temperature, 2)2:poly(wind, 2)1        0.695 0.488666
## poly(radiation, 2)1:poly(temperature, 2)1:poly(wind, 2)2        1.463 0.147082
## poly(radiation, 2)2:poly(temperature, 2)1:poly(wind, 2)2        0.055 0.956007
## poly(radiation, 2)1:poly(temperature, 2)2:poly(wind, 2)2        0.429 0.669101
## poly(radiation, 2)2:poly(temperature, 2)2:poly(wind, 2)2       -0.177 0.859657
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.29 on 84 degrees of freedom
## Multiple R-squared:  0.8169, Adjusted R-squared:  0.7602
## F-statistic: 14.42 on 26 and 84 DF,  p-value: < 2.2e-16
```

- $R^2 = 81.69$ 此模型解釋能力已相當好
- 有非常多的效應都呈現不顯著，需進一步進行 model selection

(c) Model selection and diagonostics

利用 AIC criterion 進行 model selection：

```
fit1.4 = step(fit1.3)
```

```
## Start:  AIC=642.61
## ozone ~ poly(radiation, 2) * poly(temperature, 2) * poly(wind,
##     2)
##
##                                                       Df Sum of Sq   RSS
## - poly(radiation, 2):poly(temperature, 2):poly(wind, 2)  8      1682 23982
## <none>                                                            22300
##                                                         AIC
## - poly(radiation, 2):poly(temperature, 2):poly(wind, 2) 634.68
## <none>                                                  642.61
##
## Step:  AIC=634.68
## ozone ~ poly(radiation, 2) + poly(temperature, 2) + poly(wind,
##     2) + poly(radiation, 2):poly(temperature, 2) + poly(radiation,
##     2):poly(wind, 2) + poly(temperature, 2):poly(wind, 2)
##
##                                          Df Sum of Sq   RSS    AIC
## - poly(radiation, 2):poly(wind, 2)        4    1486.7 25468 633.36
## <none>                                               23982 634.68
## - poly(radiation, 2):poly(temperature, 2)  4    1904.8 25887 635.17
## - poly(temperature, 2):poly(wind, 2)       4    6934.0 30916 654.87
##
## Step:  AIC=633.36
## ozone ~ poly(radiation, 2) + poly(temperature, 2) + poly(wind,
##     2) + poly(radiation, 2):poly(temperature, 2) + poly(temperature,
##     2):poly(wind, 2)
##
##                                          Df Sum of Sq   RSS    AIC
## <none>                                               25468 633.36
## - poly(radiation, 2):poly(temperature, 2)  4    2395.4 27864 635.34
## - poly(temperature, 2):poly(wind, 2)       4    7773.1 33241 654.92
```

```r
summary(fit1.4)
```

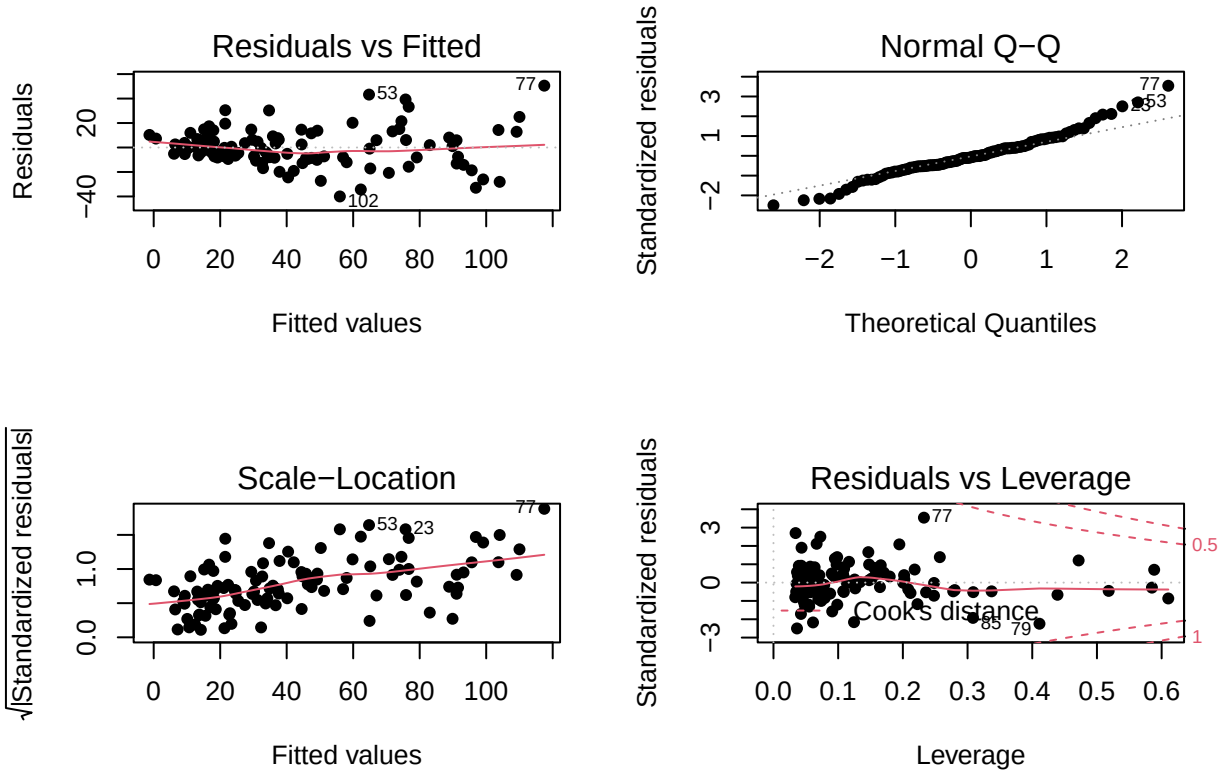
```
##
## Call:
## lm(formula = ozone ~ poly(radiation, 2) + poly(temperature, 2) +
##     poly(wind, 2) + poly(radiation, 2):poly(temperature, 2) +
##     poly(temperature, 2):poly(wind, 2), data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.992  -8.038  -1.103   7.107  50.532
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                                39.055      2.534  15.415  < 2e-16
## poly(radiation, 2)1                       116.429     26.566   4.383 3.00e-05
## poly(radiation, 2)2                       -59.256     23.788  -2.491 0.014453
## poly(temperature, 2)1                     189.145     33.231   5.692 1.36e-07
## poly(temperature, 2)2                      16.096     27.811   0.579 0.564098
## poly(wind, 2)1                            -61.960     30.257  -2.048 0.043311
## poly(wind, 2)2                            120.213     26.666   4.508 1.85e-05
## poly(radiation, 2)1:poly(temperature, 2)1  963.522    328.195   2.936 0.004163
## poly(radiation, 2)2:poly(temperature, 2)1 -565.038    269.240  -2.099 0.038473
## poly(radiation, 2)1:poly(temperature, 2)2  567.361    296.878   1.911 0.058976
## poly(radiation, 2)2:poly(temperature, 2)2 -200.848    239.904  -0.837 0.404560
## poly(temperature, 2)1:poly(wind, 2)1       -88.710    387.627  -0.229 0.819469
## poly(temperature, 2)2:poly(wind, 2)1      1145.305    295.496   3.876 0.000194
## poly(temperature, 2)1:poly(wind, 2)2       699.782    290.530   2.409 0.017921
## poly(temperature, 2)2:poly(wind, 2)2      -652.434    182.630  -3.572 0.000555
##
## (Intercept)                               ***
## poly(radiation, 2)1                       ***
## poly(radiation, 2)2                       *
## poly(temperature, 2)1                     ***
## poly(temperature, 2)2
## poly(wind, 2)1                            *
```

```
## poly(wind, 2)2                             ***
## poly(radiation, 2)1:poly(temperature, 2)1 **
## poly(radiation, 2)2:poly(temperature, 2)1 *
## poly(radiation, 2)1:poly(temperature, 2)2 .
## poly(radiation, 2)2:poly(temperature, 2)2
## poly(temperature, 2)1:poly(wind, 2)1
## poly(temperature, 2)2:poly(wind, 2)1      ***
## poly(temperature, 2)1:poly(wind, 2)2      *
## poly(temperature, 2)2:poly(wind, 2)2      ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.29 on 96 degrees of freedom
## Multiple R-squared:  0.7909, Adjusted R-squared:  0.7604
## F-statistic: 25.94 on 14 and 96 DF,  p-value: < 2.2e-16
```

- 所有的三階交互作用都被移除
- 變數 *radiation* 和 *wind* 之間所有的二階交互作用都被移除
- $R^2 = 79.09\%$ 模型解釋能力雖有所下降，但使用的變數減少非常多

對模型進行診斷：

```
par(mfrow = c(2,2))
plot(fit1.4, pch = 16)
```



- Normal Q-Q plot 大致呈現一直線，代表此模型 residual 的 normality assumption 成立
- 也沒有出現特別明顯的 outlier 或 influential observation
- 但是 residual 的 variance 有隨著 fitted value 變大而上升，呈現出 non-constant variance 的現象

將 response variable *ozone* transform 成 $\sqrt{ozone}$，重新配飾模型：

```
fit1.5 = update(fit1.4, sqrt(ozone)~.)
summary(fit1.5)
```

```
##
## Call:
## lm(formula = sqrt(ozone) ~ poly(radiation, 2) + poly(temperature,
##     2) + poly(wind, 2) + poly(radiation, 2):poly(temperature,
##     2) + poly(temperature, 2):poly(wind, 2), data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1523 -0.7111 -0.1231  0.6369  2.5644
```

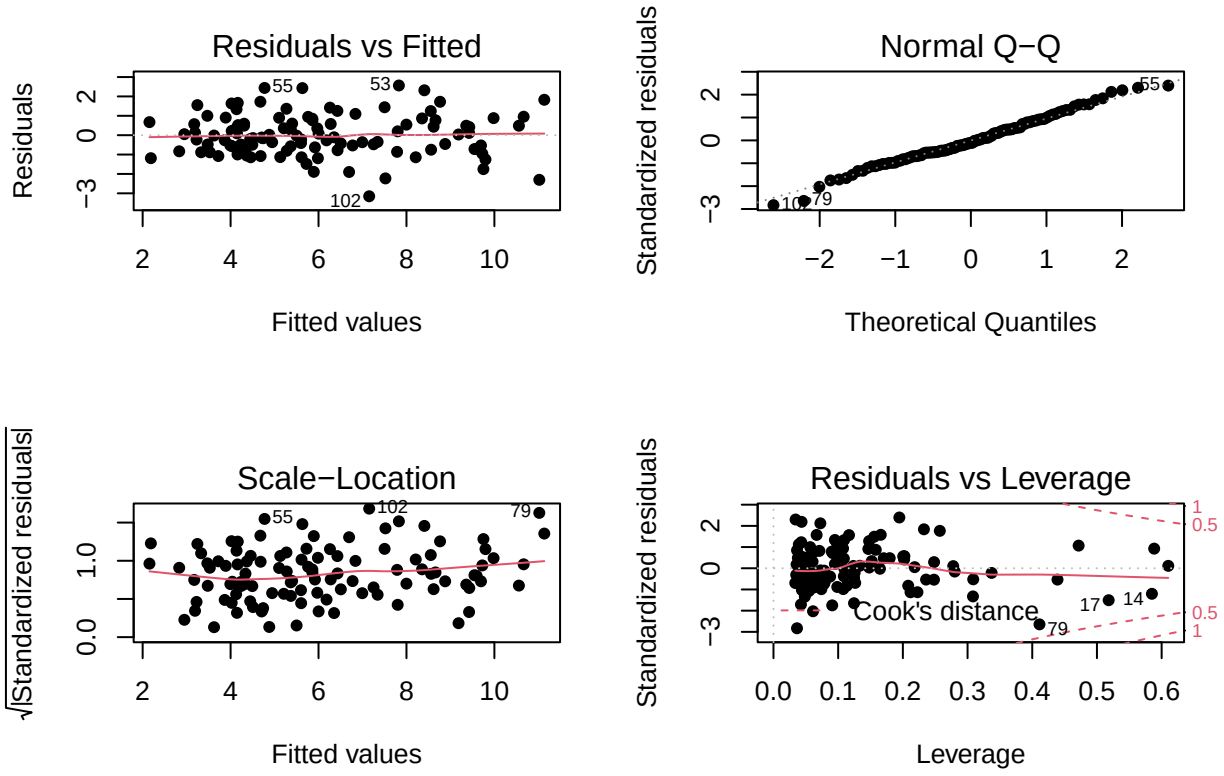
```
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                                5.8634     0.1765  33.229  < 2e-16
## poly(radiation, 2)1                        9.4329     1.8502   5.098 1.72e-06
## poly(radiation, 2)2                       -4.9498     1.6567  -2.988 0.003568
## poly(temperature, 2)1                     14.3668     2.3144   6.208 1.37e-08
## poly(temperature, 2)2                      1.0530     1.9369   0.544 0.587940
## poly(wind, 2)1                            -4.2777     2.1073  -2.030 0.045124
## poly(wind, 2)2                             7.4595     1.8572   4.017 0.000117
## poly(radiation, 2)1:poly(temperature, 2)1  62.7634    22.8575   2.746 0.007206
## poly(radiation, 2)2:poly(temperature, 2)1 -36.9303    18.7515  -1.969 0.051783
## poly(radiation, 2)1:poly(temperature, 2)2  40.6480    20.6764   1.966 0.052197
## poly(radiation, 2)2:poly(temperature, 2)2 -24.4523    16.7084  -1.463 0.146604
## poly(temperature, 2)1:poly(wind, 2)1       -0.2024    26.9967  -0.007 0.994033
## poly(temperature, 2)2:poly(wind, 2)1       80.8609    20.5801   3.929 0.000161
## poly(temperature, 2)1:poly(wind, 2)2       42.0057    20.2343   2.076 0.040570
## poly(temperature, 2)2:poly(wind, 2)2      -43.5418    12.7194  -3.423 0.000911
##
## (Intercept)                               ***
## poly(radiation, 2)1                       ***
## poly(radiation, 2)2                       **
## poly(temperature, 2)1                     ***
## poly(temperature, 2)2
## poly(wind, 2)1                            *
## poly(wind, 2)2                            ***
## poly(radiation, 2)1:poly(temperature, 2)1 **
## poly(radiation, 2)2:poly(temperature, 2)1 .
## poly(radiation, 2)1:poly(temperature, 2)2 .
## poly(radiation, 2)2:poly(temperature, 2)2
## poly(temperature, 2)1:poly(wind, 2)1
## poly(temperature, 2)2:poly(wind, 2)1      ***
## poly(temperature, 2)1:poly(wind, 2)2      *
## poly(temperature, 2)2:poly(wind, 2)2      ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.134 on 96 degrees of freedom
## Multiple R-squared:  0.8114, Adjusted R-squared:  0.7839
## F-statistic:  29.5 on 14 and 96 DF,  p-value: < 2.2e-16
```

- $R^2 = 81.14\%$ 有所上升
- 呈現顯著的變數和前一個模型相差不大

一樣對此模型進行診斷：

```
par(mfrow = c(2,2))
plot(fit1.5, pch = 16)
```



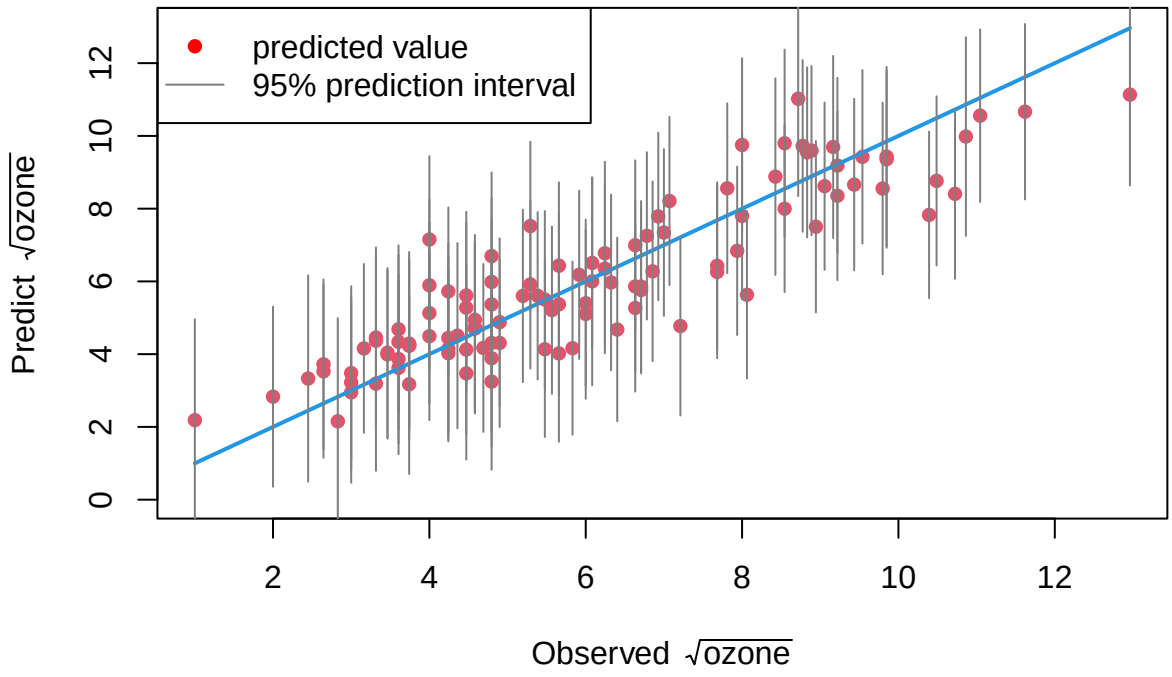
- 此模型的 residual 不再呈現如上一個模型 non-constant variance 的現象

故此模型為最終決定的配飾模型。

(d) Comments on your prediction results and scientific findings

觀察所有預測值 $\sqrt{\hat{ozone}}$ 和實際值 $\sqrt{ozone}$ 的關係

```r
fit1.5_pred = predict(fit1.5, newdata = data1[,-4], se.fit=TRUE, interval="prediction", level=0.95)
plot(sqrt(data1$ozone), fit1.5$fitted.values, ylim = c(0,13), col=2, pch=16, ylab=TeX("Predict\\ $\\sqr
curve(x^1, from=min(sqrt(data1$ozone)), to=max(sqrt(data1$ozone)), col=4, lwd=2, add=T)
for (i in 1:111){
    lines(rep(sqrt(data1$ozone[i]),2), fit1.5_pred$fit[i,2:3], col="gray50", lwd=1)
}
legend("topleft", legend=c("predicted value", "95% prediction interval"), col=c("red","gray50"), lty=c(
```



- 所有的 $\left(\sqrt{ozone}\ ,\ \sqrt{o\hat{z}one}\right)$ 都落在直線 $y = x$ 附近
- 每一個預測值的 95% prediction interval 幾乎都有覆蓋到其所對應的觀測值，代表我們模型的預測效果不錯

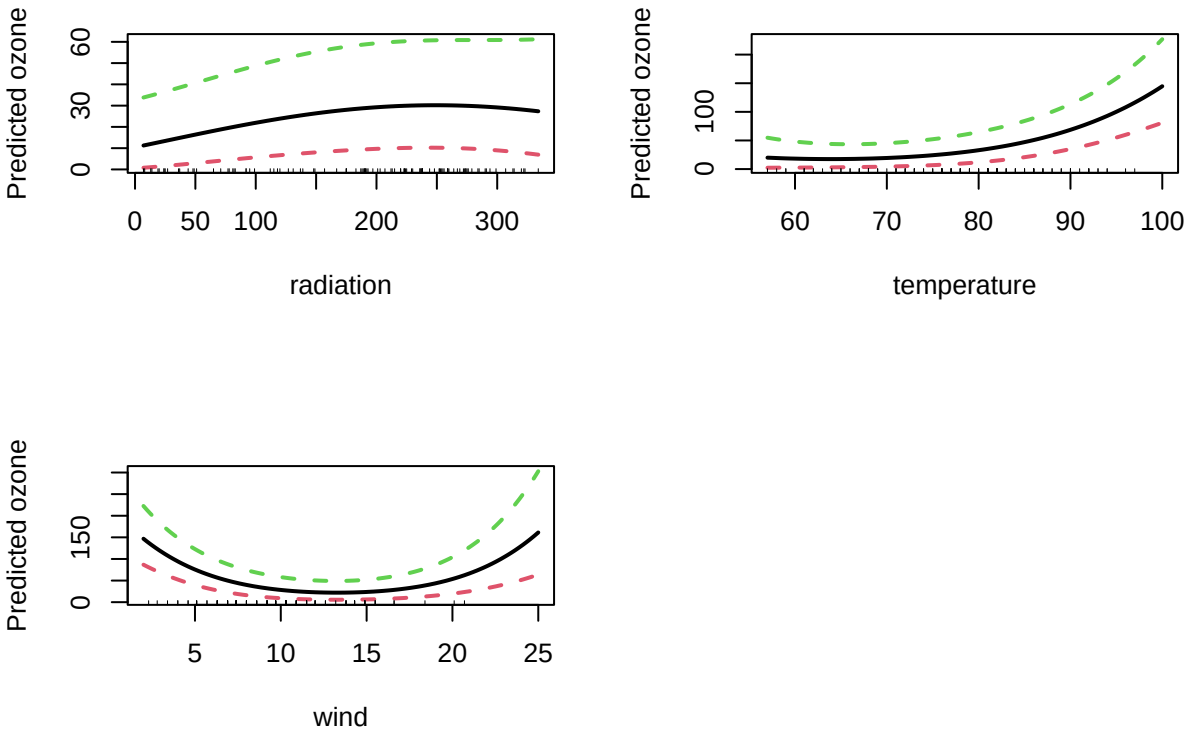
將變數 *ozone* 的預測值對變數 *radiation, temperature, wind* 作圖：

```r
par(mfrow = c(2,2))
grid1 = seq(7, 334, 0.5)
pred1 = predict(fit1.5, data.frame(radiation=grid1, temperature=mean(data1$temperature), wind=mean(data
                se.fit=TRUE, interval="prediction", level=0.95)
```

```
matplot(grid1, pred1$fit^2, lty=c(1,2,2), lwd = 2, type = "l", xlab = "radiation",ylab = "Predicted ozo
rug(data1$radiation)

grid2 = seq(57, 100, 0.5)
pred2 = predict(fit1.5, data.frame(radiation=mean(data1$radiation), temperature=grid2, wind=mean(data1$
                se.fit=TRUE, interval="prediction", level=0.95)
matplot(grid2, pred2$fit^2, lty=c(1,2,2), lwd = 2, type = "l", xlab = "temperature",ylab = "Predicted o
rug(data1$temperature)

grid3 = seq(2, 25, 0.1)
pred3 = predict(fit1.5, data.frame(radiation=mean(data1$radiation),
                                   temperature=mean(data1$temperature),wind=grid3),
                se.fit=TRUE, interval="prediction", level=0.95)
matplot(grid3, pred3$fit^2, lty=c(1,2,2), lwd = 2, type = "l", xlab = "wind",ylab = "Predicted ozone")
rug(data1$wind)
```



- 隨著 *radiation* 數值上升，*ozone* 的預測值也隨之上升，但上升的幅度會逐漸減小

- 隨著 *temperature* 數值上升，*ozone* 的預測值隨之上升，且幅度逐漸變大
- 隨著 *wind* 數值上升，*ozone* 的預測值先降後升

Problem 2.

```
data2 = read.csv("prostate.csv")
data2_train = data2 %>% filter(train.idx == 1) %>% select(-train.idx)
data2_val = data2 %>% filter(train.idx == 0) %>% select(-train.idx)
```

(a) EDA

```
dim(data2_train)
```

```
## [1] 70  9
```

```
data2_train$svi = as.factor(data2_train$svi)
summary(data2_train)
```

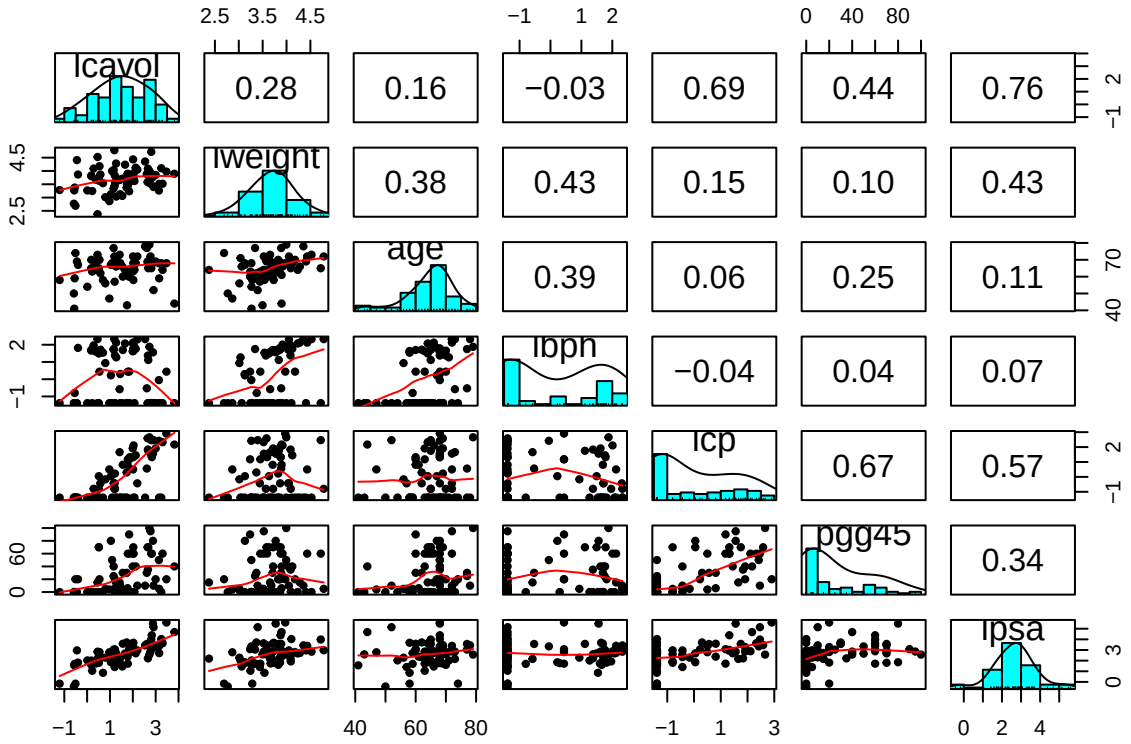
```
##      lcavol           lweight           age              lbph         svi
##  Min.   :-1.2040   Min.   :2.375   Min.   :41.00   Min.   :-1.3863   0:51
##  1st Qu.: 0.5523   1st Qu.:3.381   1st Qu.:61.25   1st Qu.:-1.3863   1:19
##  Median : 1.4940   Median :3.715   Median :66.00   Median : 0.2616
##  Mean   : 1.4841   Mean   :3.684   Mean   :64.69   Mean   : 0.1614
##  3rd Qu.: 2.4528   3rd Qu.:3.961   3rd Qu.:69.00   3rd Qu.: 1.6863
##  Max.   : 3.8210   Max.   :4.780   Max.   :79.00   Max.   : 2.3263
##       lcp              gleason        pgg45             lpsa
##  Min.   :-1.38629   Min.   :6.0   Min.   :  0.00   Min.   :-0.4308
##  1st Qu.:-1.38629   1st Qu.:6.0   1st Qu.:  0.00   1st Qu.: 2.0082
##  Median :-0.43078   Median :7.0   Median : 15.00   Median : 2.6980
##  Mean   :-0.01172   Mean   :6.8   Mean   : 26.57   Mean   : 2.6518
##  3rd Qu.: 1.32176   3rd Qu.:7.0   3rd Qu.: 50.00   3rd Qu.: 3.2252
##  Max.   : 2.90417   Max.   :9.0   Max.   :100.00   Max.   : 5.5829
```

- Training data 一共 70 比觀測值，9 個變數
- 粗略觀察各變數級距，並無發現明顯離群值 (outlier)
- 各變數類型如下：

變數名稱	變數類型	變數解釋
lcavol	continuous	log cancer volume
lweight	continuous	log prostate weight
age	approxiate continuous	age
lbph	continuous	log 良性前列腺增生量
svi	factor variable {0,1}	seminal vesicle invasion
lcp	continuous	log of capsular penetration
gleason	ordinal variable	Gleason score
pgg45	continuous	percent of Gleason scores 4 or 5
lpsa	continuous	log of prostate-specific antigen

將連續型的變數計算相關係數及繪製 pairwise scatter plots：

```
pairs.panels(data2_train[,-c(5,7)], ellipses = F)
```



- 反應變數 *lpsa* 本身分布大致對稱，並無左右偏移
- 反應變數 *lpsa* 和解釋變數 *lca, lcavol* 之間有較強的正相關，相關係數分別為 0.57, 0.76

- 反應變數 *lpsa* 和解釋變數 *lcavol* 的散佈圖有正斜率的線性關係
- 解釋變數 *lcp* 和 *lcavol, pgg45* 之間有較強的正相關，相關係數分別為 0.69, 0.67，配飾模型時可能會有共線性的情況發生

(b) Determine a good regression model for predicting data

在配飾模型中放入所有解釋變數的 main effects 和 2-factor interaction effects：

```
fit2.1 = lm(lpsa ~ .^2, data2_train)
summary(fit2.1)
```

```
##
## Call:
## lm(formula = lpsa ~ .^2, data = data2_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11728 -0.21846  0.01182  0.17140  1.01617
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -6.301735  26.678899  -0.236  0.81473
## lcavol           5.337644   2.274046   2.347  0.02507 *
## lweight          3.452292   5.306147   0.651  0.51980
## age             -0.403992   0.380320  -1.062  0.29584
## lbph             1.878755   1.791238   1.049  0.30186
## svi1            15.240305   7.720712   1.974  0.05681 .
## lcp             -0.701335   2.910922  -0.241  0.81110
## gleason          3.348595   3.190307   1.050  0.30152
## pgg45           -0.174100   0.125015  -1.393  0.17305
## lcavol:lweight  -0.696380   0.312130  -2.231  0.03259 *
## lcavol:age       0.018101   0.018457   0.981  0.33386
## lcavol:lbph      0.114836   0.086244   1.332  0.19214
## lcavol:svi1     -0.148169   0.912172  -0.162  0.87195
## lcavol:lcp       0.318942   0.321783   0.991  0.32881
## lcavol:gleason  -0.447963   0.336566  -1.331  0.19232
## lcavol:pgg45    -0.005203   0.014715  -0.354  0.72593
```

```
## lweight:age      0.072246   0.050243   1.438  0.15987
## lweight:lbph    -0.197992   0.192082  -1.031  0.31015
## lweight:svi1     1.044442   1.829012   0.571  0.57184
## lweight:lcp      0.261310   0.561076   0.466  0.64447
## lweight:gleason -0.976020   0.608694  -1.603  0.11836
## lweight:pgg45   -0.014663   0.018739  -0.783  0.43950
## age:lbph        -0.001147   0.013973  -0.082  0.93506
## age:svi1        -0.023523   0.085776  -0.274  0.78561
## age:lcp         -0.029423   0.026263  -1.120  0.27067
## age:gleason      0.012138   0.044740   0.271  0.78784
## age:pgg45        0.001345   0.001383   0.972  0.33811
## lbph:svi1       -1.129943   0.415263  -2.721  0.01031 *
## lbph:lcp         0.175476   0.129277   1.357  0.18388
## lbph:gleason    -0.154532   0.229294  -0.674  0.50504
## lbph:pgg45       0.009539   0.004754   2.007  0.05305 .
## svi1:lcp         0.140873   0.733022   0.192  0.84878
## svi1:gleason    -2.708477   0.832211  -3.255  0.00263 **
## svi1:pgg45       0.064183   0.028564   2.247  0.03145 *
## lcp:gleason      0.188853   0.327372   0.577  0.56794
## lcp:pgg45       -0.017915   0.009673  -1.852  0.07299 .
## gleason:pgg45    0.019899   0.013880   1.434  0.16109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5994 on 33 degrees of freedom
## Multiple R-squared:  0.8654, Adjusted R-squared:  0.7185
## F-statistic: 5.893 on 36 and 33 DF,  p-value: 7.105e-07
```

- $R^2 = 86.54\%$ 模型可解釋之比例相當高
- 但模型中有非常多不顯著的效應

利用 AIC criterion 進行 model selection：

```
fit2.2 = step(fit2.1)
```

```
## Start:  AIC=-50.31
## lpsa ~ (lcavol + lweight + age + lbph + svi + lcp + gleason +
```

```
##     pgg45)^2
##
##                   Df Sum of Sq    RSS     AIC
## - age:lbph         1    0.0024 11.857 -52.291
## - lcavol:svi       1    0.0095 11.864 -52.249
## - svi:lcp          1    0.0133 11.868 -52.227
## - age:gleason      1    0.0264 11.881 -52.149
## - age:svi          1    0.0270 11.882 -52.146
## - lcavol:pgg45     1    0.0449 11.899 -52.040
## - lweight:lcp      1    0.0779 11.932 -51.846
## - lweight:svi      1    0.1171 11.972 -51.617
## - lcp:gleason      1    0.1195 11.974 -51.603
## - lbph:gleason     1    0.1632 12.018 -51.348
## - lweight:pgg45    1    0.2200 12.075 -51.018
## - age:pgg45        1    0.3394 12.194 -50.329
## <none>                         11.854 -50.305
## - lcavol:age       1    0.3455 12.200 -50.294
## - lcavol:lcp       1    0.3529 12.207 -50.252
## - lweight:lbph     1    0.3817 12.236 -50.087
## - age:lcp          1    0.4509 12.305 -49.692
## - lcavol:gleason   1    0.6364 12.491 -48.645
## - lcavol:lbph      1    0.6369 12.491 -48.642
## - lbph:lcp         1    0.6619 12.516 -48.502
## - gleason:pgg45    1    0.7383 12.593 -48.076
## - lweight:age      1    0.7427 12.597 -48.051
## - lweight:gleason  1    0.9236 12.778 -47.053
## - lcp:pgg45        1    1.2321 13.087 -45.383
## - lbph:pgg45       1    1.4463 13.301 -44.247
## - lcavol:lweight   1    1.7881 13.643 -42.471
## - svi:pgg45        1    1.8137 13.668 -42.339
## - lbph:svi         1    2.6597 14.514 -38.136
## - svi:gleason      1    3.8050 15.659 -32.819
##
## Step:  AIC=-52.29
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45 + lcavol:lweight + lcavol:age + lcavol:lbph + lcavol:svi +
```

```
##     lcavol:lcp + lcavol:gleason + lcavol:pgg45 + lweight:age +
##     lweight:lbph + lweight:svi + lweight:lcp + lweight:gleason +
##     lweight:pgg45 + age:svi + age:lcp + age:gleason + age:pgg45 +
##     lbph:svi + lbph:lcp + lbph:gleason + lbph:pgg45 + svi:lcp +
##     svi:gleason + svi:pgg45 + lcp:gleason + lcp:pgg45 + gleason:pgg45
##
##                   Df Sum of Sq    RSS     AIC
## - lcavol:svi       1    0.0076 11.865 -54.246
## - svi:lcp          1    0.0182 11.875 -54.184
## - age:gleason      1    0.0240 11.881 -54.149
## - age:svi          1    0.0250 11.882 -54.144
## - lcavol:pgg45     1    0.0425 11.899 -54.040
## - lweight:lcp      1    0.0770 11.934 -53.838
## - lweight:svi      1    0.1149 11.972 -53.616
## - lcp:gleason      1    0.1249 11.982 -53.557
## - lbph:gleason     1    0.1829 12.040 -53.220
## - lweight:pgg45    1    0.2233 12.080 -52.985
## - age:pgg45        1    0.3370 12.194 -52.329
## <none>                         11.857 -52.291
## - lcavol:age       1    0.3448 12.202 -52.284
## - lcavol:lcp       1    0.3779 12.235 -52.095
## - lweight:lbph     1    0.4267 12.284 -51.816
## - age:lcp          1    0.4524 12.309 -51.669
## - lcavol:lbph      1    0.6616 12.518 -50.490
## - lbph:lcp         1    0.7624 12.619 -49.928
## - gleason:pgg45    1    0.7810 12.638 -49.825
## - lcavol:gleason   1    0.7994 12.656 -49.724
## - lweight:age      1    0.8300 12.687 -49.555
## - lweight:gleason  1    0.9232 12.780 -49.042
## - lcp:pgg45        1    1.2297 13.087 -47.383
## - lbph:pgg45       1    1.4470 13.304 -46.230
## - lcavol:lweight   1    1.8401 13.697 -44.192
## - svi:pgg45        1    1.9157 13.773 -43.807
## - lbph:svi         1    2.8482 14.705 -39.221
## - svi:gleason      1    3.9587 15.816 -34.125
##
```

```
## Step:  AIC=-54.25
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45 + lcavol:lweight + lcavol:age + lcavol:lbph + lcavol:lcp +
##     lcavol:gleason + lcavol:pgg45 + lweight:age + lweight:lbph +
##     lweight:svi + lweight:lcp + lweight:gleason + lweight:pgg45 +
##     age:svi + age:lcp + age:gleason + age:pgg45 + lbph:svi +
##     lbph:lcp + lbph:gleason + lbph:pgg45 + svi:lcp + svi:gleason +
##     svi:pgg45 + lcp:gleason + lcp:pgg45 + gleason:pgg45
##
##                   Df Sum of Sq    RSS     AIC
## - svi:lcp          1    0.0182 11.883 -56.139
## - age:gleason      1    0.0188 11.883 -56.135
## - age:svi          1    0.0278 11.892 -56.082
## - lcavol:pgg45     1    0.0428 11.907 -55.994
## - lweight:lcp      1    0.1094 11.974 -55.603
## - lweight:svi      1    0.1099 11.975 -55.600
## - lcp:gleason      1    0.1274 11.992 -55.498
## - lbph:gleason     1    0.1838 12.048 -55.170
## - lweight:pgg45    1    0.2310 12.095 -54.896
## <none>                         11.865 -54.246
## - lcavol:age       1    0.3506 12.215 -54.207
## - age:pgg45        1    0.3697 12.234 -54.098
## - age:lcp          1    0.4515 12.316 -53.631
## - lweight:lbph     1    0.4697 12.334 -53.528
## - lcavol:lcp       1    0.5973 12.462 -52.807
## - lbph:lcp         1    0.7707 12.635 -51.840
## - gleason:pgg45    1    0.8075 12.672 -51.637
## - lweight:age      1    0.8372 12.702 -51.473
## - lcavol:lbph      1    0.8516 12.716 -51.394
## - lcavol:gleason   1    0.8943 12.759 -51.159
## - lweight:gleason  1    0.9155 12.780 -51.042
## - lbph:pgg45       1    1.4852 13.350 -47.990
## - lcp:pgg45        1    1.5947 13.459 -47.418
## - lcavol:lweight   1    1.8680 13.732 -46.011
## - svi:pgg45        1    2.2267 14.091 -44.206
## - lbph:svi         1    2.8409 14.706 -41.219
```

```
## - svi:gleason      1    4.1562 16.021 -35.223
##
## Step:  AIC=-56.14
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45 + lcavol:lweight + lcavol:age + lcavol:lbph + lcavol:lcp +
##     lcavol:gleason + lcavol:pgg45 + lweight:age + lweight:lbph +
##     lweight:svi + lweight:lcp + lweight:gleason + lweight:pgg45 +
##     age:svi + age:lcp + age:gleason + age:pgg45 + lbph:svi +
##     lbph:lcp + lbph:gleason + lbph:pgg45 + svi:gleason + svi:pgg45 +
##     lcp:gleason + lcp:pgg45 + gleason:pgg45
##
##                   Df Sum of Sq    RSS     AIC
## - age:gleason      1    0.0126 11.895 -58.065
## - age:svi          1    0.0757 11.959 -57.694
## - lweight:lcp      1    0.0918 11.975 -57.600
## - lcp:gleason      1    0.1096 11.992 -57.496
## - lweight:svi      1    0.1539 12.037 -57.238
## - lbph:gleason     1    0.1660 12.049 -57.168
## - lweight:pgg45    1    0.2157 12.098 -56.880
## - lcavol:pgg45     1    0.2519 12.135 -56.670
## <none>                         11.883 -56.139
## - lcavol:age       1    0.3826 12.265 -55.920
## - age:lcp          1    0.4359 12.319 -55.617
## - lweight:lbph     1    0.4539 12.337 -55.515
## - age:pgg45        1    0.5248 12.408 -55.114
## - lbph:lcp         1    0.7547 12.637 -53.828
## - gleason:pgg45    1    0.8082 12.691 -53.532
## - lweight:age      1    0.8200 12.703 -53.468
## - lcavol:lbph      1    0.8340 12.717 -53.391
## - lcavol:gleason   1    0.9304 12.813 -52.862
## - lweight:gleason  1    1.0329 12.916 -52.304
## - lbph:pgg45       1    1.5720 13.455 -49.442
## - lcavol:lweight   1    1.9161 13.799 -47.674
## - lcp:pgg45        1    2.0610 13.944 -46.943
## - lcavol:lcp       1    2.8230 14.706 -43.218
## - svi:pgg45        1    2.9238 14.806 -42.740
```

```
## - lbph:svi         1    3.1197 15.002 -41.820
## - svi:gleason      1    4.1483 16.031 -37.178
##
## Step:  AIC=-58.06
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45 + lcavol:lweight + lcavol:age + lcavol:lbph + lcavol:lcp +
##     lcavol:gleason + lcavol:pgg45 + lweight:age + lweight:lbph +
##     lweight:svi + lweight:lcp + lweight:gleason + lweight:pgg45 +
##     age:svi + age:lcp + age:pgg45 + lbph:svi + lbph:lcp + lbph:gleason +
##     lbph:pgg45 + svi:gleason + svi:pgg45 + lcp:gleason + lcp:pgg45 +
##     gleason:pgg45
##
##                   Df Sum of Sq    RSS     AIC
## - lweight:lcp      1    0.0856 11.981 -59.563
## - lcp:gleason      1    0.1398 12.035 -59.247
## - age:svi          1    0.1399 12.035 -59.246
## - lbph:gleason     1    0.1651 12.060 -59.100
## - lweight:svi      1    0.1827 12.078 -58.998
## - lcavol:pgg45     1    0.2441 12.139 -58.643
## - lweight:pgg45    1    0.2928 12.188 -58.362
## <none>                         11.895 -58.065
## - lcavol:age       1    0.4052 12.300 -57.720
## - lweight:lbph     1    0.4467 12.342 -57.484
## - age:lcp          1    0.4482 12.344 -57.476
## - lbph:lcp         1    0.7785 12.674 -55.627
## - gleason:pgg45    1    0.8012 12.697 -55.502
## - lweight:age      1    0.8118 12.707 -55.443
## - lcavol:lbph      1    0.8555 12.751 -55.203
## - lcavol:gleason   1    0.9568 12.852 -54.649
## - lweight:gleason  1    1.0977 12.993 -53.886
## - age:pgg45        1    1.5304 13.426 -51.592
## - lbph:pgg45       1    1.5619 13.457 -51.429
## - lcavol:lweight   1    1.9129 13.808 -49.627
## - lcp:pgg45        1    2.0606 13.956 -48.881
## - lcavol:lcp       1    2.8195 14.715 -45.175
## - svi:pgg45        1    2.9137 14.809 -44.728
```

```
## - lbph:svi         1    3.1757 15.071 -43.500
## - svi:gleason      1    4.1385 16.034 -39.166
##
## Step:  AIC=-59.56
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45 + lcavol:lweight + lcavol:age + lcavol:lbph + lcavol:lcp +
##     lcavol:gleason + lcavol:pgg45 + lweight:age + lweight:lbph +
##     lweight:svi + lweight:gleason + lweight:pgg45 + age:svi +
##     age:lcp + age:pgg45 + lbph:svi + lbph:lcp + lbph:gleason +
##     lbph:pgg45 + svi:gleason + svi:pgg45 + lcp:gleason + lcp:pgg45 +
##     gleason:pgg45
##
##                   Df Sum of Sq    RSS     AIC
## - lcp:gleason      1    0.0754 12.056 -61.124
## - lbph:gleason     1    0.1684 12.149 -60.585
## - lweight:pgg45    1    0.2079 12.189 -60.358
## - lcavol:pgg45     1    0.3032 12.284 -59.814
## <none>                         11.981 -59.563
## - age:svi          1    0.3608 12.342 -59.486
## - lweight:lbph     1    0.3777 12.359 -59.390
## - age:lcp          1    0.3785 12.359 -59.386
## - lcavol:age       1    0.3893 12.370 -59.324
## - lweight:age      1    0.7263 12.707 -57.443
## - lcavol:lbph      1    0.8778 12.859 -56.613
## - lcavol:gleason   1    0.8903 12.871 -56.545
## - lbph:lcp         1    0.9491 12.930 -56.226
## - lweight:svi      1    1.0297 13.011 -55.791
## - lweight:gleason  1    1.0712 13.052 -55.568
## - gleason:pgg45    1    1.2033 13.184 -54.863
## - age:pgg45        1    1.4582 13.439 -53.523
## - lcavol:lweight   1    1.8292 13.810 -51.617
## - lbph:pgg45       1    1.8597 13.841 -51.462
## - lcp:pgg45        1    2.2928 14.274 -49.306
## - lcavol:lcp       1    3.0550 15.036 -45.664
## - svi:pgg45        1    3.3789 15.360 -44.172
## - lbph:svi         1    4.0863 16.067 -41.020
```

```
## - svi:gleason      1    4.2299 16.211 -40.397
##
## Step:  AIC=-61.12
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45 + lcavol:lweight + lcavol:age + lcavol:lbph + lcavol:lcp +
##     lcavol:gleason + lcavol:pgg45 + lweight:age + lweight:lbph +
##     lweight:svi + lweight:gleason + lweight:pgg45 + age:svi +
##     age:lcp + age:pgg45 + lbph:svi + lbph:lcp + lbph:gleason +
##     lbph:pgg45 + svi:gleason + svi:pgg45 + lcp:pgg45 + gleason:pgg45
##
##                   Df Sum of Sq    RSS     AIC
## - lbph:gleason     1    0.1323 12.188 -62.360
## - lweight:pgg45    1    0.1888 12.245 -62.036
## - age:lcp          1    0.3031 12.359 -61.386
## - lcavol:age       1    0.3175 12.374 -61.304
## <none>                         12.056 -61.124
## - lcavol:pgg45     1    0.3556 12.412 -61.089
## - age:svi          1    0.4308 12.487 -60.666
## - lweight:lbph     1    0.4478 12.504 -60.571
## - lweight:age      1    0.8286 12.885 -58.471
## - lcavol:gleason   1    0.8298 12.886 -58.464
## - lbph:lcp         1    0.8760 12.932 -58.214
## - lcavol:lbph      1    0.9477 13.004 -57.827
## - lweight:svi      1    0.9993 13.056 -57.550
## - lweight:gleason  1    1.0300 13.086 -57.385
## - age:pgg45        1    1.4279 13.484 -55.289
## - lcavol:lweight   1    1.7906 13.847 -53.431
## - lbph:pgg45       1    1.9906 14.047 -52.427
## - lcp:pgg45        1    2.2650 14.321 -51.072
## - lcavol:lcp       1    3.0129 15.069 -47.509
## - gleason:pgg45    1    3.2948 15.351 -46.212
## - svi:pgg45        1    3.5406 15.597 -45.100
## - lbph:svi         1    4.0219 16.078 -42.972
## - svi:gleason      1    4.6750 16.731 -40.185
##
## Step:  AIC=-62.36
```

```
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45 + lcavol:lweight + lcavol:age + lcavol:lbph + lcavol:lcp +
##     lcavol:gleason + lcavol:pgg45 + lweight:age + lweight:lbph +
##     lweight:svi + lweight:gleason + lweight:pgg45 + age:svi +
##     age:lcp + age:pgg45 + lbph:svi + lbph:lcp + lbph:pgg45 +
##     svi:gleason + svi:pgg45 + lcp:pgg45 + gleason:pgg45
##
##                    Df Sum of Sq    RSS     AIC
## - lweight:pgg45     1    0.0978 12.286 -63.800
## - age:lcp           1    0.2199 12.409 -63.108
## - lcavol:pgg45      1    0.2535 12.442 -62.919
## - lcavol:age        1    0.2799 12.469 -62.770
## <none>                         12.188 -62.360
## - lweight:lbph      1    0.3614 12.550 -62.315
## - age:svi           1    0.4763 12.665 -61.676
## - lweight:age       1    0.7099 12.899 -60.397
## - lcavol:gleason    1    0.7316 12.920 -60.279
## - lbph:lcp          1    0.7626 12.951 -60.112
## - lcavol:lbph       1    0.8379 13.027 -59.706
## - lweight:svi       1    0.9618 13.150 -59.043
## - age:pgg45         1    1.3643 13.553 -56.933
## - lcavol:lweight    1    1.7711 13.960 -54.862
## - lbph:pgg45        1    2.1828 14.371 -52.828
## - lweight:gleason   1    2.3262 14.515 -52.133
## - lcp:pgg45         1    2.4296 14.618 -51.636
## - gleason:pgg45     1    3.1626 15.351 -48.211
## - lcavol:lcp        1    3.3118 15.500 -47.534
## - svi:pgg45         1    3.5628 15.751 -46.410
## - lbph:svi          1    4.4957 16.684 -42.382
## - svi:gleason       1    4.5434 16.732 -42.182
##
## Step:  AIC=-63.8
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45 + lcavol:lweight + lcavol:age + lcavol:lbph + lcavol:lcp +
##     lcavol:gleason + lcavol:pgg45 + lweight:age + lweight:lbph +
##     lweight:svi + lweight:gleason + age:svi + age:lcp + age:pgg45 +
```

```
##     lbph:svi + lbph:lcp + lbph:pgg45 + svi:gleason + svi:pgg45 +
##     lcp:pgg45 + gleason:pgg45
##
##                   Df Sum of Sq    RSS     AIC
## - age:lcp          1    0.2112 12.498 -64.607
## - lcavol:age       1    0.2906 12.577 -64.164
## - lcavol:pgg45     1    0.3206 12.607 -63.997
## - lweight:lbph     1    0.3270 12.613 -63.961
## <none>                         12.286 -63.800
## - age:svi          1    0.4112 12.698 -63.496
## - lweight:age      1    0.6493 12.936 -62.195
## - lcavol:gleason   1    0.6585 12.945 -62.146
## - lbph:lcp         1    0.6676 12.954 -62.096
## - lcavol:lbph      1    0.8595 13.146 -61.067
## - lweight:svi      1    0.8750 13.161 -60.985
## - age:pgg45        1    1.2762 13.562 -58.883
## - lcavol:lweight   1    1.7959 14.082 -56.250
## - lbph:pgg45       1    2.0981 14.384 -54.764
## - lcp:pgg45        1    2.3389 14.625 -53.602
## - gleason:pgg45    1    3.0650 15.351 -50.211
## - lcavol:lcp       1    3.3233 15.610 -49.042
## - svi:pgg45        1    3.4813 15.768 -48.337
## - lweight:gleason  1    3.5711 15.857 -47.940
## - svi:gleason      1    4.4472 16.733 -44.176
## - lbph:svi         1    4.5923 16.879 -43.571
##
## Step:  AIC=-64.61
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45 + lcavol:lweight + lcavol:age + lcavol:lbph + lcavol:lcp +
##     lcavol:gleason + lcavol:pgg45 + lweight:age + lweight:lbph +
##     lweight:svi + lweight:gleason + age:svi + age:pgg45 + lbph:svi +
##     lbph:lcp + lbph:pgg45 + svi:gleason + svi:pgg45 + lcp:pgg45 +
##     gleason:pgg45
##
##                   Df Sum of Sq    RSS     AIC
## - lcavol:age       1    0.1482 12.646 -65.782
```

```
## - lcavol:pgg45     1    0.2657 12.763 -65.135
## <none>                         12.498 -64.607
## - lweight:lbph     1    0.4232 12.921 -64.276
## - lbph:lcp         1    0.5353 13.033 -63.672
## - lcavol:gleason   1    0.6692 13.167 -62.956
## - lcavol:lbph      1    0.9223 13.420 -61.624
## - lweight:svi      1    0.9258 13.423 -61.605
## - age:svi          1    1.0002 13.498 -61.218
## - age:pgg45        1    1.0806 13.578 -60.803
## - lweight:age      1    1.1245 13.622 -60.576
## - lcavol:lweight   1    1.7263 14.224 -57.550
## - lbph:pgg45       1    2.4442 14.942 -54.103
## - lcp:pgg45        1    2.7241 15.222 -52.805
## - gleason:pgg45    1    3.1330 15.630 -50.949
## - lcavol:lcp       1    3.2636 15.761 -50.366
## - svi:pgg45        1    3.4864 15.984 -49.384
## - lweight:gleason  1    3.6668 16.164 -48.598
## - lbph:svi         1    4.4805 16.978 -45.160
## - svi:gleason      1    4.5686 17.066 -44.798
##
## Step:  AIC=-65.78
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45 + lcavol:lweight + lcavol:lbph + lcavol:lcp + lcavol:gleason +
##     lcavol:pgg45 + lweight:age + lweight:lbph + lweight:svi +
##     lweight:gleason + age:svi + age:pgg45 + lbph:svi + lbph:lcp +
##     lbph:pgg45 + svi:gleason + svi:pgg45 + lcp:pgg45 + gleason:pgg45
##
##                   Df Sum of Sq    RSS     AIC
## - lcavol:pgg45     1    0.2620 12.908 -66.347
## <none>                         12.646 -65.782
## - lbph:lcp         1    0.4700 13.116 -65.228
## - lcavol:gleason   1    0.5937 13.239 -64.570
## - lweight:lbph     1    0.6973 13.343 -64.025
## - lweight:svi      1    0.7865 13.432 -63.558
## - age:svi          1    0.9325 13.578 -62.802
## - lcavol:lbph      1    1.0120 13.658 -62.393
```

```
## - age:pgg45        1    1.0591 13.705 -62.152
## - lcavol:lweight   1    1.7014 14.347 -58.946
## - lweight:age      1    2.1459 14.792 -56.810
## - lbph:pgg45       1    2.3289 14.975 -55.950
## - lcp:pgg45        1    2.7086 15.354 -54.196
## - lcavol:lcp       1    3.1261 15.772 -52.319
## - gleason:pgg45    1    3.2279 15.874 -51.868
## - svi:pgg45        1    3.4294 16.075 -50.985
## - lweight:gleason  1    3.5202 16.166 -50.591
## - lbph:svi         1    4.3323 16.978 -47.160
## - svi:gleason      1    4.7571 17.403 -45.430
##
## Step:  AIC=-66.35
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45 + lcavol:lweight + lcavol:lbph + lcavol:lcp + lcavol:gleason +
##     lweight:age + lweight:lbph + lweight:svi + lweight:gleason +
##     age:svi + age:pgg45 + lbph:svi + lbph:lcp + lbph:pgg45 +
##     svi:gleason + svi:pgg45 + lcp:pgg45 + gleason:pgg45
##
##                   Df Sum of Sq    RSS     AIC
## <none>                          12.908 -66.347
## - lbph:lcp         1    0.3744 13.282 -66.345
## - lweight:lbph     1    0.6687 13.576 -64.811
## - lweight:svi      1    0.7301 13.638 -64.495
## - age:svi          1    0.8007 13.708 -64.134
## - age:pgg45        1    0.8529 13.761 -63.868
## - lcavol:lbph      1    0.9344 13.842 -63.455
## - lcavol:gleason   1    1.1574 14.065 -62.336
## - lcavol:lweight   1    1.6438 14.552 -59.956
## - lweight:age      1    1.9711 14.879 -58.399
## - lbph:pgg45       1    2.2412 15.149 -57.140
## - lcavol:lcp       1    2.9541 15.862 -53.921
## - gleason:pgg45    1    3.0048 15.912 -53.697
## - lcp:pgg45        1    3.2449 16.153 -52.649
## - lweight:gleason  1    3.2604 16.168 -52.582
## - svi:pgg45        1    3.4784 16.386 -51.644
```

```
## - lbph:svi         1    4.0941 17.002 -49.062
## - svi:gleason      1    5.0222 17.930 -45.342
```

```
summary(fit2.2)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
##     gleason + pgg45 + lcavol:lweight + lcavol:lbph + lcavol:lcp +
##     lcavol:gleason + lweight:age + lweight:lbph + lweight:svi +
##     lweight:gleason + age:svi + age:pgg45 + lbph:svi + lbph:lcp +
##     lbph:pgg45 + svi:gleason + svi:pgg45 + lcp:pgg45 + gleason:pgg45,
##     data = data2_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.16695 -0.28700  0.04097  0.23943  1.10355
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.263e+01  9.990e+00  -1.264 0.212834
## lcavol           5.541e+00  1.752e+00   3.162 0.002836 **
## lweight          3.484e+00  2.878e+00   1.211 0.232462
## age             -2.753e-01  9.761e-02  -2.820 0.007174 **
## lbph             7.691e-01  5.594e-01   1.375 0.176172
## svi1             1.521e+01  4.910e+00   3.097 0.003395 **
## lcp             -1.455e-01  2.288e-01  -0.636 0.528060
## gleason          4.280e+00  1.233e+00   3.473 0.001169 **
## pgg45           -2.387e-01  6.530e-02  -3.655 0.000681 ***
## lcavol:lweight  -5.282e-01  2.231e-01  -2.367 0.022390 *
## lcavol:lbph      1.197e-01  6.709e-02   1.785 0.081207 .
## lcavol:lcp       2.426e-01  7.645e-02   3.173 0.002749 **
## lcavol:gleason  -4.186e-01  2.107e-01  -1.986 0.053250 .
## lweight:age      7.367e-02  2.842e-02   2.592 0.012898 *
## lweight:lbph    -2.099e-01  1.391e-01  -1.510 0.138245
## lweight:svi1     1.118e+00  7.086e-01   1.578 0.121831
## lweight:gleason -1.101e+00  3.302e-01  -3.334 0.001745 **
```

```
## age:svi1        -5.351e-02  3.239e-02  -1.652 0.105641
## age:pgg45        1.008e-03  5.912e-04   1.705 0.095221 .
## lbph:svi1       -8.763e-01  2.346e-01  -3.736 0.000535 ***
## lbph:lcp         9.001e-02  7.967e-02   1.130 0.264699
## lbph:pgg45       7.966e-03  2.882e-03   2.764 0.008307 **
## svi1:gleason    -2.417e+00  5.842e-01  -4.138 0.000156 ***
## svi1:pgg45       5.112e-02  1.485e-02   3.443 0.001272 **
## lcp:pgg45       -1.824e-02  5.484e-03  -3.326 0.001786 **
## gleason:pgg45    2.378e-02  7.429e-03   3.200 0.002548 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5416 on 44 degrees of freedom
## Multiple R-squared:  0.8534, Adjusted R-squared:  0.7701
## F-statistic: 10.25 on 25 and 44 DF,  p-value: 2.608e-11
```

接著再移除所有不顯著的效應：

```
fit2.3 = update(fit2.2, .~.-lweight-lbph-lcp-lcavol:lbph-lcavol:gleason-lweight:lbph-lweight:svi1-age:s
summary(fit2.3)
```

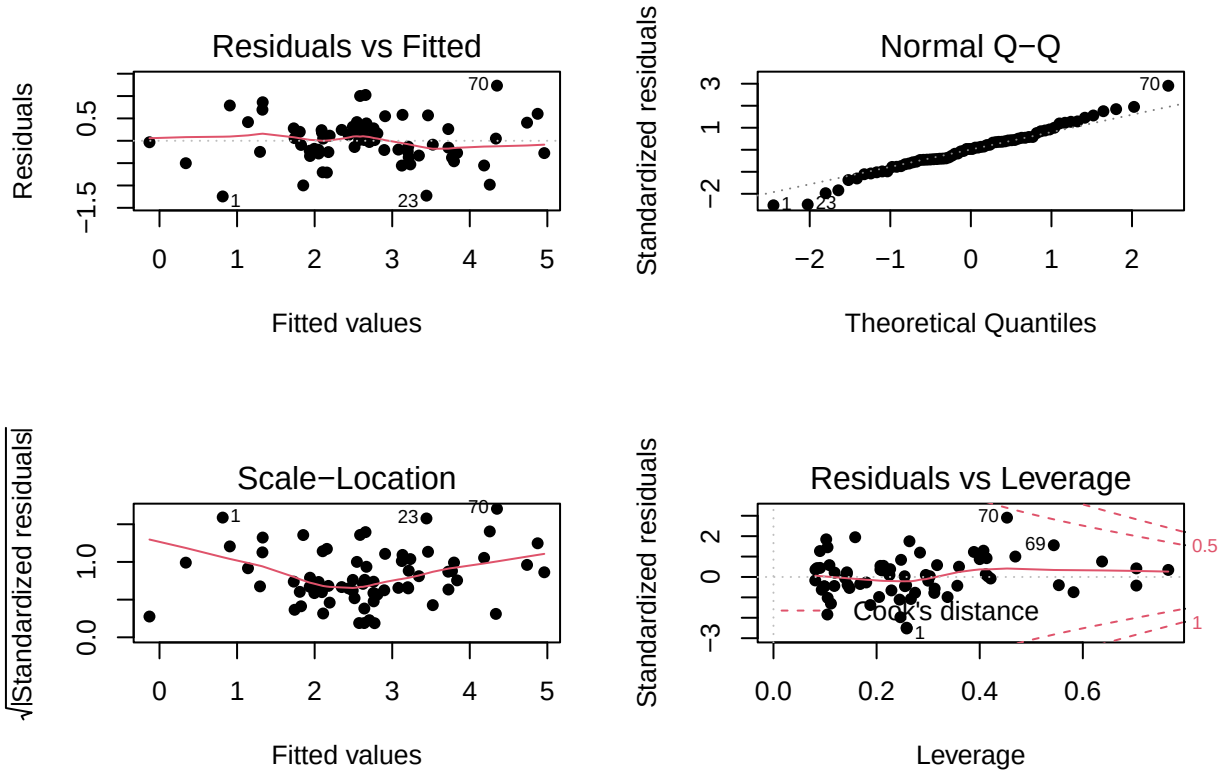
```
##
## Call:
## lm(formula = lpsa ~ lcavol + age + svi + gleason + pgg45 + lcavol:lweight +
##     lcavol:lcp + lweight:age + lweight:svi + lweight:gleason +
##     age:svi + lbph:svi + lbph:pgg45 + svi:gleason + svi:pgg45 +
##     lcp:pgg45 + gleason:pgg45, data = data2_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.24543 -0.26662  0.01743  0.27180  1.23246
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.2185843  1.4729023   2.185 0.033493 *
## lcavol           2.4405125  0.6974674   3.499 0.000978 ***
## age             -0.2613407  0.0714947  -3.655 0.000607 ***
```

```
## svi1            10.2052480  4.3309028   2.356 0.022334 *
## gleason          1.8621723  0.7499019   2.483 0.016349 *
## pgg45           -0.1156917  0.0373583  -3.097 0.003176 **
## lcavol:lweight  -0.4430464  0.1856654  -2.386 0.020770 *
## lcavol:lcp       0.1646786  0.0576901   2.855 0.006216 **
## age:lweight      0.0740703  0.0207266   3.574 0.000780 ***
## svi1:lweight     0.9628775  0.6394177   1.506 0.138271
## gleason:lweight -0.6136378  0.2105451  -2.915 0.005279 **
## age:svi1        -0.0254769  0.0247204  -1.031 0.307590
## svi0:lbph       -0.0009051  0.0763697  -0.012 0.990591
## svi1:lbph       -0.5192987  0.1760883  -2.949 0.004801 **
## pgg45:lbph       0.0092089  0.0025771   3.573 0.000781 ***
## svi1:gleason    -1.8631655  0.5502823  -3.386 0.001373 **
## svi1:pgg45       0.0451765  0.0133551   3.383 0.001386 **
## pgg45:lcp       -0.0192632  0.0047118  -4.088 0.000154 ***
## gleason:pgg45    0.0165695  0.0052710   3.144 0.002781 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5733 on 51 degrees of freedom
## Multiple R-squared:  0.8097, Adjusted R-squared:  0.7425
## F-statistic: 12.05 on 18 and 51 DF,  p-value: 1.338e-12
```

- $R^2 = 80.97\%$ 有所下降，但模型使用的變數減少了很多
- 留下來的變數大部份皆為顯著

對模型進行診斷：

```
par(mfrow = c(2,2))
plot(fit2.3, pch = 16)
```



- Residual plot 沒有明顯 mean curve 和 non-constant variance
- 藉由 normal Q-Q plot 也可得知 residual 服從 norality assumption
- 沒有特別明顯的 outlier 或 influential observation

故此模型即為我們的配飾模型。

(c) Describe the important main effects and interaction effects

藉由 training data 所得的配飾模型如下：

$$\begin{aligned}\widehat{lpsa} =& 3.22 + 2.44\ lcavol - 0.26\ age + 10.21\ svi + 1.86\ gleason - 0.12\ pgg45 - 0.44\ lcavol \times lweight \\ &+ 0.16\ lcavol \times lcp + 0.07\ lweight \times age - 0.61\ lweight \times gleason - 0.52\ lbph \times svi + 0.009\ lbph \times pgg45 \\ &- 1.86\ svi \times gleason + 0.05\ svi \times pgg45 - 0.02\ lcp \times pgg45 + 0.02\ gleason \times pgg45 \\ &+ (\text{unimportant effects}) \ + \ \hat{\epsilon}\end{aligned}$$

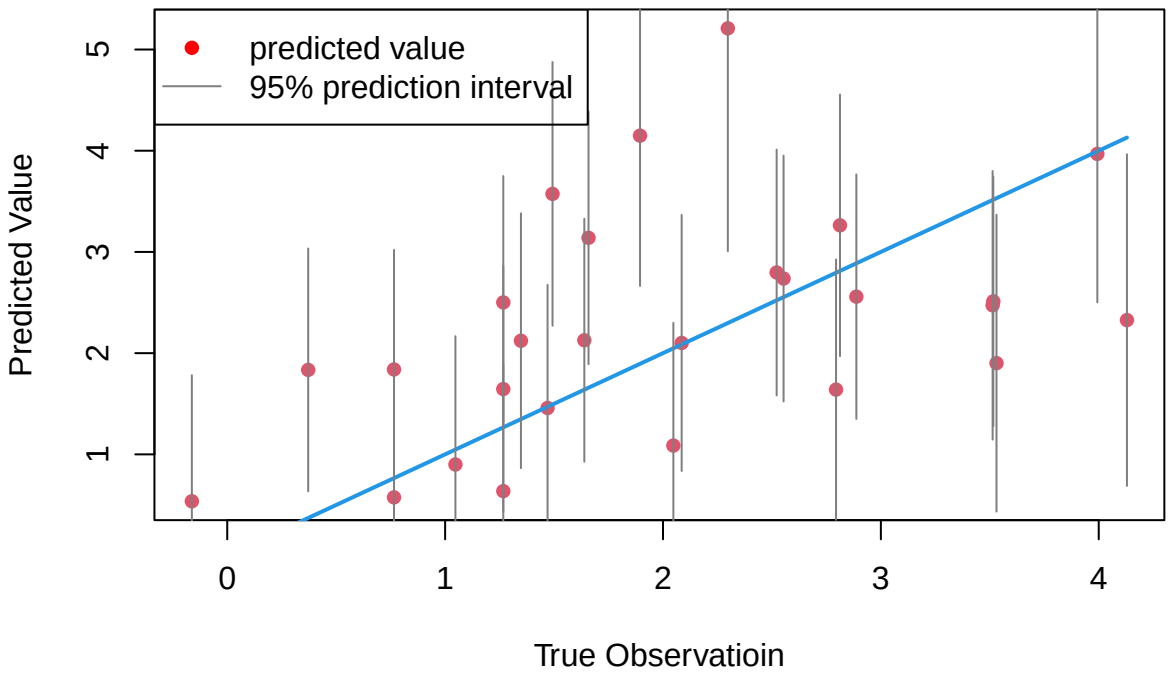
因為模型中有很多顯著的二階交互作用，要描述其中一個解釋變數如何影響反應變數，都必須考慮其他的解釋變數的數值為多少，為方便呈現，整理如下表 (以下結果皆忽略不顯著的各效應)：

每增加 1 單位 _____ 變數	lpsa 會增加 _____ 單位
lcavol	2.44 - 0.44 *lweight* + 0.16 *lcp*
age	-0.26 + 0.07 *lweight*
svi	10.21 - 0.52 *lbph* - 1.86 *gleason* + 0.05 *pgg45*
gleason	1.86 - 0.61 *lweight* - 1.86 *svi* + 0.02 *pgg45*
pgg45	-0.12 + 0.009 *lbph* + 0.05 *svi* - 0.02 *lcp* + 0.02 *gleason*

(d) Predict lpsa for the validation data set based on the fitted model, with their prediction intervals. And compared the prediction results to the true observations. Comment on your model performance.

將 validation data set 的預測值及實際觀測值繪製成 scatter plot：

```
data2_val$svi = as.factor(data2_val$svi)
pred_val = predict(fit2.3, newdata = data2_val[,-9], se.fit = T, interval = "prediction", level = 0.95)
plot(data2_val$lpsa, pred_val$fit[,1], pch = 16, col = 2, xlab = "True Observatioin", ylab = "Predicted
curve(x^1, from = min(data2_val$lpsa), to = max(data2_val$lpsa), col = 4, lwd = 2, add = T)
for (i in 1:27){
    lines(rep(data2_val$lpsa[i],2), pred_val$fit[i,2:3], col = "gray50", lwd = 1)
}
legend("topleft", legend = c("predicted value", "95% prediction interval"), col = c("red", "gray50"), l
```



- 可以看到大部份的點大致落在 $y = x$ 直線兩側
- 不是所有的 95% prediction interval 都能覆蓋住 true observation

分別計算此模型的 95% prediction interval 覆蓋住 training data 和 validation data true observation 的比例

```
pred_train = predict(fit2.2, newdata = data2_train[,-9], se.fit = T, interval = "prediction", level = 0
prob_train = mean(pred_train$fit[,2]<data2_train$lpsa & data2_train$lpsa<pred_train$fit[,3])
prob_val = mean(pred_val$fit[,2]<data2_val$lpsa & data2_val$lpsa<pred_val$fit[,3])
c(prob_train, prob_val)
```

```
## [1] 1.0000000 0.7407407
```

再對兩個 data sets 分別計算 $MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$

```
mse_train = mean((pred_train$fit[,1]-data2_train$lpsa)^2)
mse_val = mean((pred_val$fit[,1]-data2_val$lpsa)^2)
c(mse_train, mse_val)
```

```
## [1] 0.1843963 1.3887819
```

Training data 的預測結果比 validation data 來得好，可能有 overfitting 的現象發生，可以嘗試簡化 training data 的 fitted model，有機會能改善此現象。