Linear Model Assignment 7

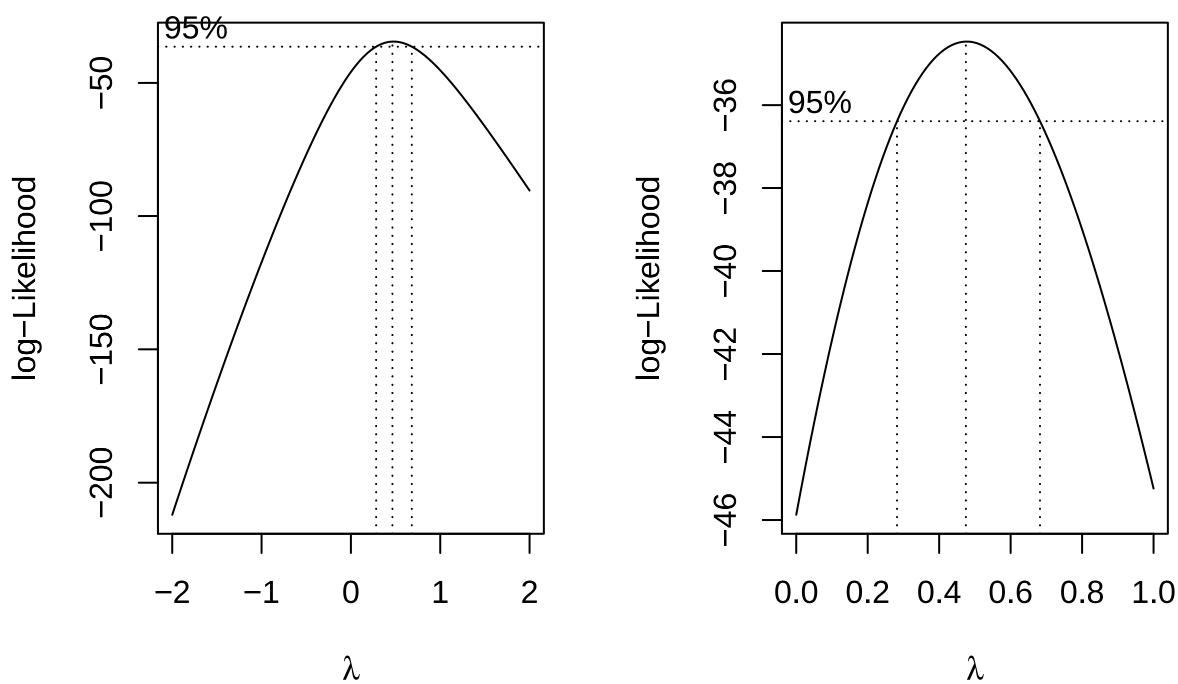
110024516 統研碩一邱繼賢

Problem 1.
a.
匯入資料並建構模型：

$$g_{1.1}\ :\ pasture\ \sim\ arable\ +\ cows\ +\ diff$$

對 response *pasture* 做 Box-Cox transformation 然後繪製其 log-likelihood 圖形：

```
library(MASS)
rent_data = read.table("pasture.txt", skip = 1)
colnames(rent_data) = c("arable", "cows", "diff", "pasture")
g1.1 = lm(pasture ~ arable+cows+diff, data = rent_data)
par(mfrow = c(1,2))
boxcox(g1.1, plotit = T)
boxcox(g1.1, plotit = T, lambda = seq(0,1,by = 0.1)) # take lambda = 0.5
```



可發現 λ 的 95% 信賴區間並沒有包含 1 $\Rightarrow$ 有充分理由對 response *pasture* 做變換，且 $0.4 < \hat{\lambda}_{MLE} < 0.5$，但因為此模型主要目的為 explanation 不是 prediction，所以取 $\lambda = 0.5$，即為將 *pasture* 變換為 $\sqrt{pasture}$

重新建構模型：

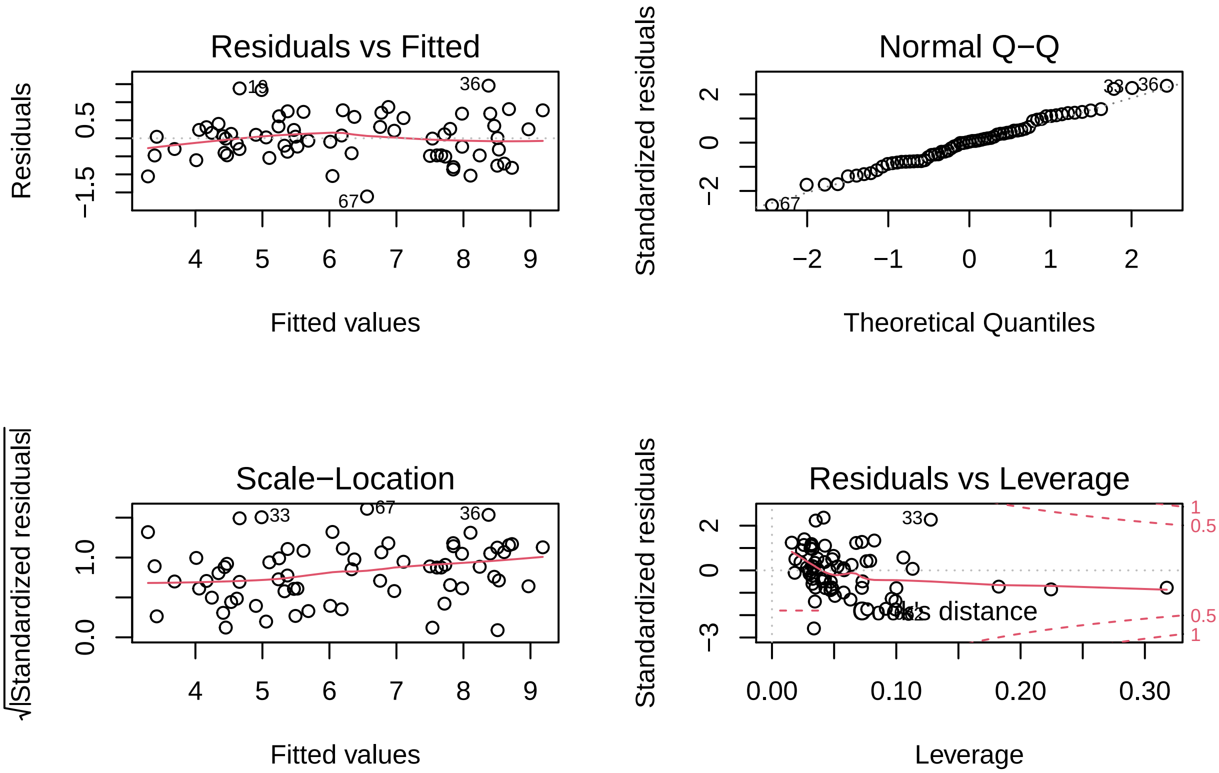
$$g_{1.2} \ : \ \sqrt{pasture} \ \sim \ arable \ + \ cows \ + \ diff$$

```
g1.2 = lm(sqrt(pasture) ~ arable+diff+cows, data = rent_data)
summary(g1.2)
```

```
##
## Call:
## lm(formula = sqrt(pasture) ~ arable + diff + cows, data = rent_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.61163 -0.47176  0.02407  0.33587  1.45740
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.476606   0.267693   9.252 2.38e-13 ***
## arable       0.072483   0.004629  15.660  < 2e-16 ***
## diff        -0.635228   0.798462  -0.796    0.429
## cows         0.035275   0.006531   5.401 1.07e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6317 on 63 degrees of freedom
## Multiple R-squared:  0.8819, Adjusted R-squared:  0.8763
## F-statistic: 156.8 on 3 and 63 DF,  p-value: < 2.2e-16
```

並且檢查 diagnostics：

```
par(mfrow = c(2,2))
plot(g1.2)
```



b.

對 predictor *cows* 取一個夠高的次數，在此設定為五次，並建構模型

$$\sqrt{pasture} \sim arable + diff + cows + cows^2 + cows^3 + cows^4 + cows^5$$

檢定 *cows* 最高次數的變數是否顯著，若不顯著則將其從模型中移除，重複此步驟直到最高次數的變數顯著為止：

```
summary(lm(sqrt(pasture) ~ arable+diff+cows+
               I(cows^2)+I(cows^3)+I(cows^4)+I(cows^5),
           data = rent_data))
```

```
##
## Call:
## lm(formula = sqrt(pasture) ~ arable + diff + cows + I(cows^2) +
##     I(cows^3) + I(cows^4) + I(cows^5), data = rent_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.42677 -0.43133  0.05973  0.39088  1.45484
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.357e+00  6.084e-01   3.874 0.000271 ***
## arable       7.001e-02  4.772e-03  14.672  < 2e-16 ***
## diff        -1.233e+00  8.390e-01  -1.469 0.147066
## cows         2.127e-02  1.926e-01   0.110 0.912417
```

```
## I(cows^2)    8.172e-03  1.953e-02   0.418 0.677201
## I(cows^3)   -4.633e-04  8.328e-04  -0.556 0.580084
## I(cows^4)    9.620e-06  1.553e-05   0.619 0.538023
## I(cows^5)   -6.937e-08  1.047e-07  -0.663 0.510165
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6136 on 59 degrees of freedom
## Multiple R-squared:  0.8956, Adjusted R-squared:  0.8832
## F-statistic: 72.33 on 7 and 59 DF,  p-value: < 2.2e-16
```

```
summary(lm(sqrt(pasture) ~ arable+diff+cows+
               I(cows^2)+I(cows^3)+I(cows^4),
           data = rent_data))
```

```
##
## Call:
## lm(formula = sqrt(pasture) ~ arable + diff + cows + I(cows^2) +
##     I(cows^3) + I(cows^4), data = rent_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.45353 -0.44737  0.06275  0.38596  1.40055
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.092e+00  4.564e-01   4.584 2.36e-05 ***
## arable       6.964e-02  4.716e-03  14.766  < 2e-16 ***
## diff        -1.308e+00  8.273e-01  -1.581    0.119
## cows         1.306e-01  9.900e-02   1.319    0.192
## I(cows^2)   -3.941e-03  6.850e-03  -0.575    0.567
## I(cows^3)    7.558e-05  1.783e-04   0.424    0.673
## I(cows^4)   -6.193e-07  1.541e-06  -0.402    0.689
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6107 on 60 degrees of freedom
## Multiple R-squared:  0.8949, Adjusted R-squared:  0.8843
## F-statistic: 85.11 on 6 and 60 DF,  p-value: < 2.2e-16
```

```
summary(lm(sqrt(pasture) ~ arable+diff+cows+
               I(cows^2)+I(cows^3),
           data = rent_data))
```

```
##
## Call:
## lm(formula = sqrt(pasture) ~ arable + diff + cows + I(cows^2) +
##     I(cows^3), data = rent_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.45194 -0.43327  0.04717  0.41727  1.42438
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.198e+00  3.705e-01   5.932 1.52e-07 ***
## arable       6.992e-02  4.631e-03  15.097  < 2e-16 ***
## diff        -1.288e+00  8.201e-01  -1.570    0.122
## cows         9.581e-02  4.791e-02   2.000    0.050 *
## I(cows^2)   -1.315e-03  2.038e-03  -0.645    0.521
## I(cows^3)    4.598e-06  2.401e-05   0.191    0.849
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6065 on 61 degrees of freedom
## Multiple R-squared:  0.8946, Adjusted R-squared:  0.8859
## F-statistic: 103.5 on 5 and 61 DF,  p-value: < 2.2e-16
```

```
summary(lm(sqrt(pasture) ~ arable+diff+cows+
               I(cows^2),
           data = rent_data))
```

```
##
## Call:
## lm(formula = sqrt(pasture) ~ arable + diff + cows + I(cows^2),
##     data = rent_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.45497 -0.43737  0.04283  0.41838  1.40590
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.2462759  0.2686749   8.361 9.39e-12 ***
## arable       0.0697594  0.0045214  15.429  < 2e-16 ***
## diff        -1.3158695  0.8006583  -1.643  0.10534
## cows         0.0875027  0.0201602   4.340 5.35e-05 ***
## I(cows^2)   -0.0009302  0.0003415  -2.724  0.00838 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6018 on 62 degrees of freedom
## Multiple R-squared:  0.8945, Adjusted R-squared:  0.8877
## F-statistic: 131.4 on 4 and 62 DF,  p-value: < 2.2e-16
```

當變數 $cows$ 最高次數為二次時，才呈現為顯著，故建構模型：

$$g_{1.3}\ :\ \sqrt{pasture}\ \sim\ arable + diff + cows + cows^2$$

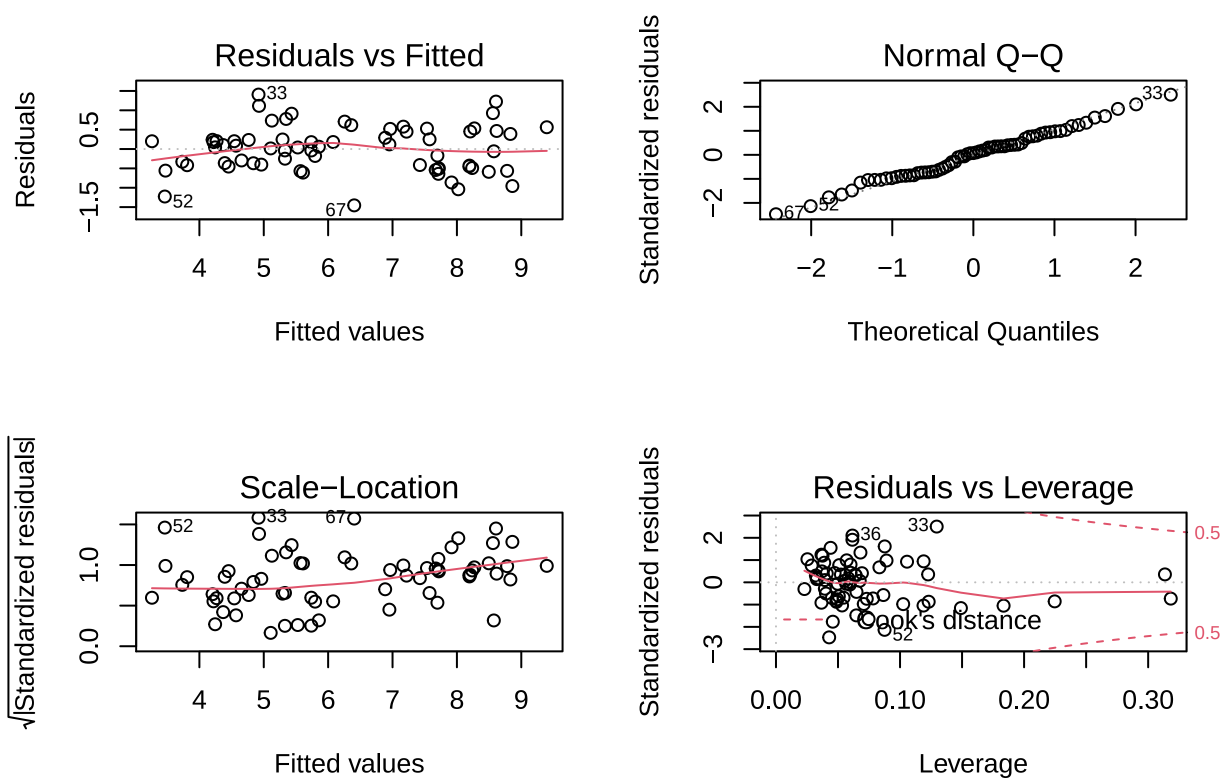
```
g1.3 = update(g1.2, .~. +I(cows^2), data = rent_data)
summary(g1.3)
```

```
##
## Call:
## lm(formula = sqrt(pasture) ~ arable + diff + cows + I(cows^2),
##     data = rent_data)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.45497 -0.43737  0.04283  0.41838  1.40590
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.2462759  0.2686749   8.361 9.39e-12 ***
## arable       0.0697594  0.0045214  15.429  < 2e-16 ***
## diff        -1.3158695  0.8006583  -1.643  0.10534
## cows         0.0875027  0.0201602   4.340 5.35e-05 ***
## I(cows^2)   -0.0009302  0.0003415  -2.724  0.00838 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6018 on 62 degrees of freedom
## Multiple R-squared:  0.8945, Adjusted R-squared:  0.8877
## F-statistic: 131.4 on 4 and 62 DF,  p-value: < 2.2e-16
```

並檢查 diagnostics：

```r
par(mfrow = c(2,2))
plot(g1.3)
```



c.

建構模型：

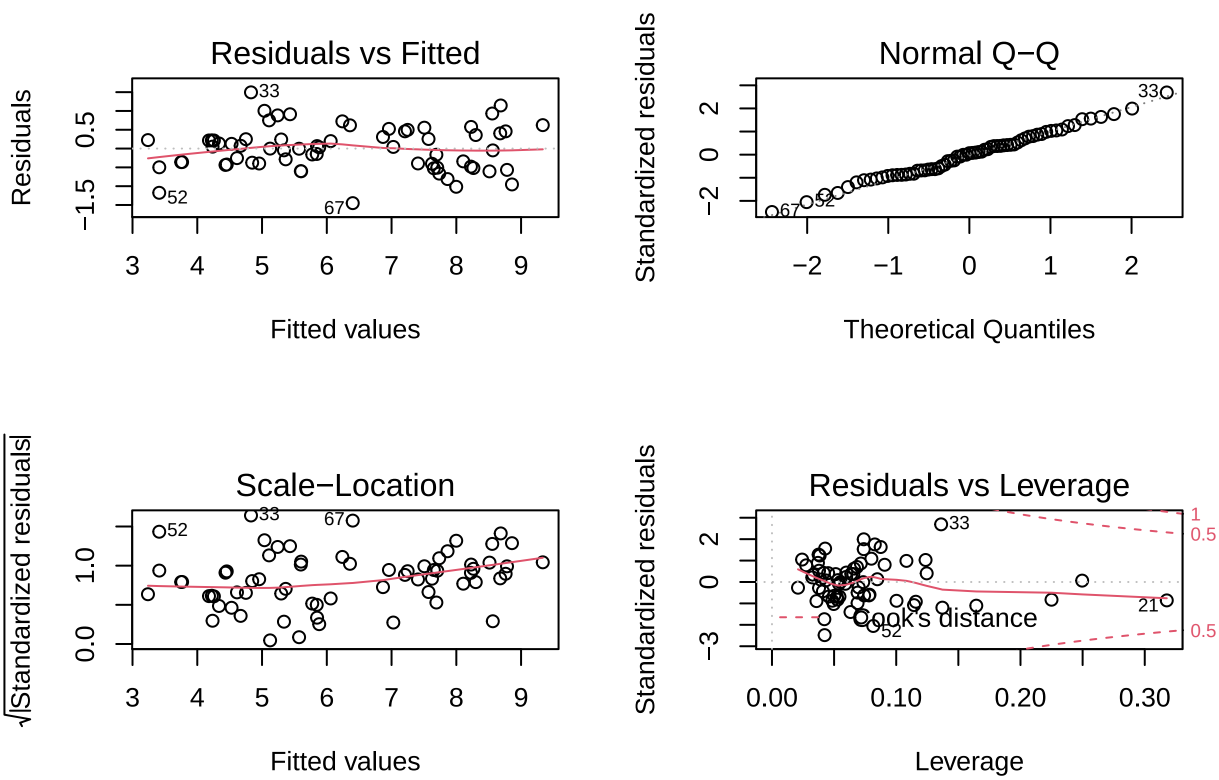
$$g_{1.4}\ :\ \sqrt{pasture}\ \sim\ arable + diff + cows + (cows - 25)\ d_{25}(cows)$$

where

$$d_{25}(cows) \quad = \quad \begin{cases} 1 \ , \ \text{if } cows > 25 \\ 0 \ , \ \text{if } cows \leq 25 \end{cases}$$

然後檢查此模型的 diagnostics：

```
d = function(x) ifelse(x>25, 1, 0)
g1.4 = update(g1.2, .~. +I((cows-25)*d(cows)), data = rent_data)
par(mfrow = c(2,2))
plot(g1.4)
```



比較模型 $g_{1.4}$ 是否 fit 的較模型 $g_{1.2}$ 來得好，即為進行以下檢定：

$$\begin{cases} H_0 & : \ g_{1.2} \text{ fits good enough} \\ H_1 & : \ g_{1.4} \text{ fits significent better} \end{cases}$$

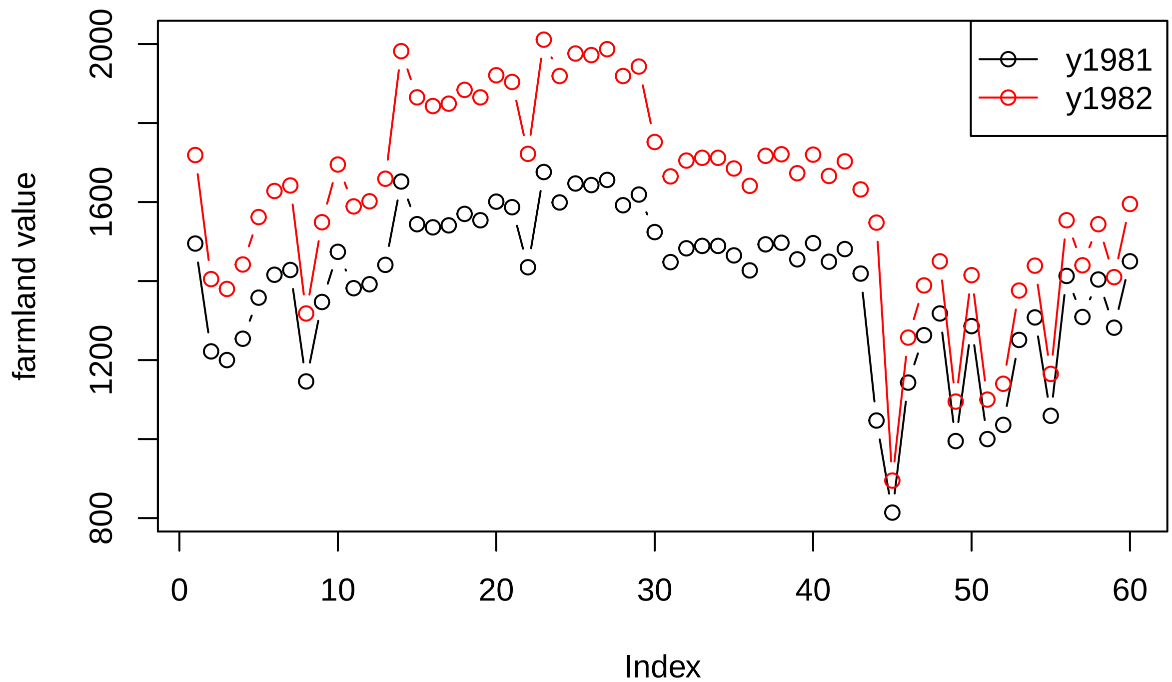
```
anova(g1.2,g1.4)
```

```
## Analysis of Variance Table
##
## Model 1: sqrt(pasture) ~ arable + diff + cows
## Model 2: sqrt(pasture) ~ arable + diff + cows + I((cows - 25) * d(cows))
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     63 25.138
## 2     62 22.145  1    2.9932 8.3803 0.005231 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

pvalue $= 0.005231 < 0.05 \Rightarrow$ reject H_0
$\therefore$ The broken-stick regression improves the fit.

Problem 2.
匯入資料並且將 1981 和 1982 兩年的 farmland value 以 Index 為 x 軸繪製折線圖：

```
assess_data = read.table("assess.txt", header = T)
matplot(cbind(assess_data$y1981, assess_data$y1982),
        type = "b", pch = 1, lty = 1, col = c("black", "red"),
        xlab = "Index", ylab = "farmland value")
legend("topright", c("y1981","y1982"), lty = c(1,1), pch = c(1,1), col = c("black", "red"))
```



以同一組 Index 的兩年 farmland values 的算術平均數 $y = \frac{y1981+y1982}{2}$ 為 response，各組的 $\frac{1}{sample\ variance}$ 為權重，建構模型：

$$g_{2.1} \ : \ y \ \sim \ P \ + \ County \ + \ P : County$$

```
library(dplyr)
assess_data = assess_data %>%
    mutate(y=(y1981+y1982)/2, var=(y1981-y)^2+(y1982-y)^2)
g2.1 = lm(y ~ P*County, weights = 1/var, data = assess_data)
summary(g2.1)
```

```
##
## Call:
## lm(formula = y ~ P * County, data = assess_data, weights = 1/var)
##
```

```
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8076 -0.1859  0.0623  0.3905  3.0568
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1143.629    305.734   3.741 0.000459 ***
## P                 4.335      4.545   0.954 0.344536
## CountyMcLeod    782.549   1110.939   0.704 0.484325
## CountyMeeker   -349.483    315.252  -1.109 0.272712
## CountySibley    366.654   1064.251   0.345 0.731846
## P:CountyMcLeod   -8.736     14.369  -0.608 0.545835
## P:CountyMeeker    4.366      4.802   0.909 0.367445
## P:CountySibley   -1.623     12.580  -0.129 0.897873
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.12 on 52 degrees of freedom
## Multiple R-squared:  0.7279, Adjusted R-squared:  0.6913
## F-statistic: 19.88 on 7 and 52 DF,  p-value: 1.142e-12
```

可發現各交互作用項的係數皆不顯著，將模型 $g_{2.1}$ 與模型 $g_{2.2}$ ：$y \sim P + County$ 做比較

$$\begin{cases} H_0 & : \ g_{2.2} \text{ fits good enough} \\ H_1 & : \ g_{2.1} \text{ fits significent better} \end{cases}$$

```r
g2.2 = lm(y ~ P + County, weights = 1/var, data = assess_data)
anova(g2.2, g2.1)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ P + County
## Model 2: y ~ P * County
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     55 67.615
## 2     52 65.249  3    2.3656 0.6284 0.5999
```

$\because$ pvalue $= 0.5999 > 0.05 \Rightarrow$ fail to reject H_0
$\therefore$ 模型可以簡化為 $g_{2.2}$

但此題是要探討土壤生產力 P 是否會對土地價值 y 有所影響，故繼續檢定模型 $g_{2.2}$ 是否可以簡化為模型 $g_{2.3}$ ：$y \sim P$

$$\begin{cases} H_0 & : \ g_{2.3} \text{ fits good enough} \\ H_1 & : \ g_{2.2} \text{ fits significent better} \end{cases}$$

```r
g2.3 = lm(y ~ P, weights = 1/var, data = assess_data)
anova(g2.3,g2.2)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ P
## Model 2: y ~ P + County
```

```
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)  
## 1     58 76.221                               
## 2     55 67.615  3    8.6066 2.3336 0.08398 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\because$ pvalue $= 0.08398 > 0.05 \Rightarrow$ fail to reject H_0
$\therefore$ 模型可以簡化為 $g_{2.3}$

```
summary(g2.3)
```

```
## 
## Call:
## lm(formula = y ~ P, data = assess_data, weights = 1/var)
## 
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.5875 -0.2648  0.3454  0.5544  3.1258 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 715.8032    59.0384   12.12  < 2e-16 ***
## P            10.7555     0.9639   11.16 4.58e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.146 on 58 degrees of freedom
## Multiple R-squared:  0.6822, Adjusted R-squared:  0.6767 
## F-statistic: 124.5 on 1 and 58 DF,  p-value: 4.579e-16
```

土地價值評估模型：

$$\hat{y} \;=\; 715.8032 \;+\; 10.7555 \times P$$

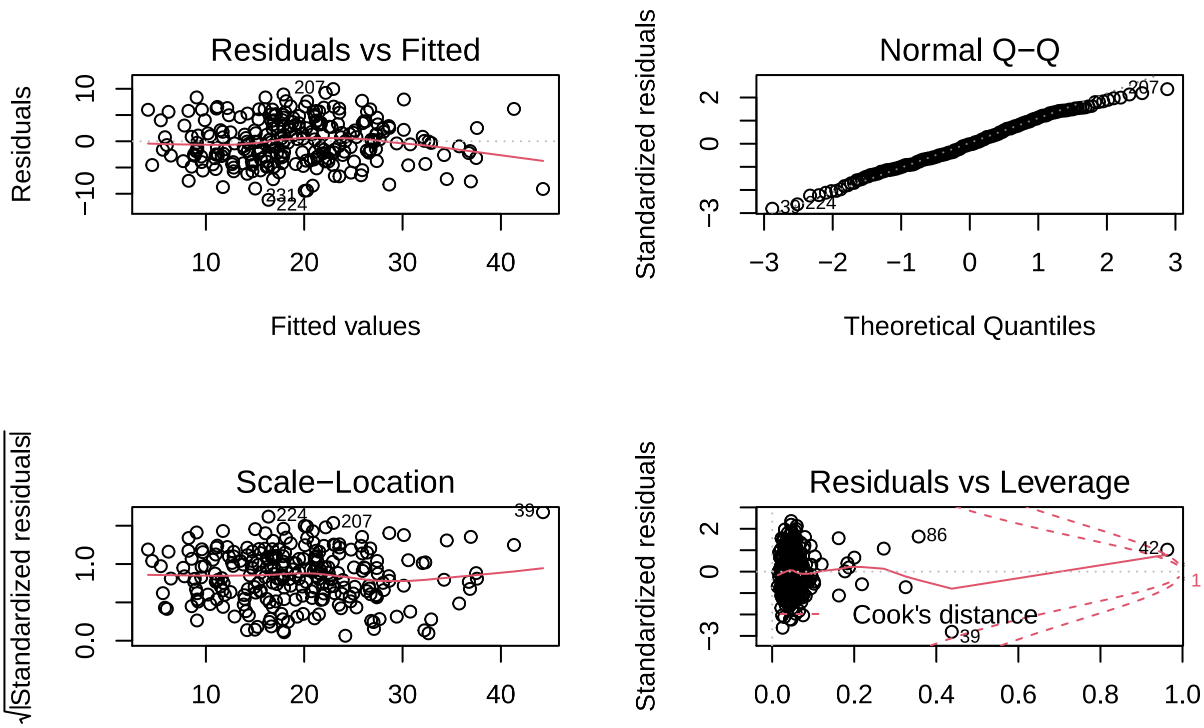
Problem 3.
匯入資料並去除變數 *brozek, density, free*，然後以 *siri* 為 response，其餘 14 個變數為 predictors 建構模型：

$$g_{3.1} \ : \ siri \ \sim \ .$$

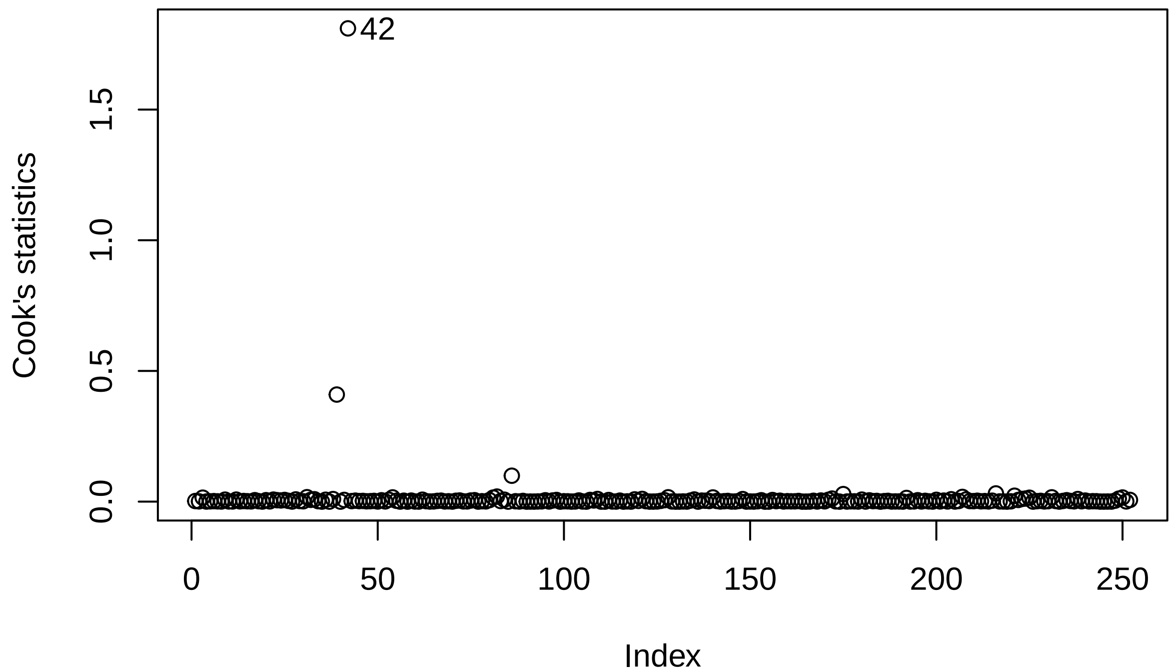
檢查此模型的 diagnostics：

```
library(tibble)
library(dplyr)
fat_data = read.table("fat.txt")
fat_data = as.tibble(fat_data) %>%
  dplyr::select(-brozek, -density, -free)

g3.1 = lm(siri ~. , data = fat_data)
par(mfrow = c(2,2))
plot(g3.1)
```




```
par(mfrow = c(1,1))
cook = cooks.distance(g3.1)
plot(cook, ylab = "Cook's statistics")
text(50,cook[42], "42")
```



可發現第 42 個觀測值的 Cook's distance 遠大於其他觀測值，故推測其為 influential observation，將其移除後重新建構模型：

$$g_{3.2} \ : \ siri \ \sim \ . \quad , \quad subset = (cook < 1)$$

然後以 AIC 的方法做 model selection：

```
g3.2 = lm(siri ~. , subset = (cook<1), data = fat_data)
step(g3.2)
```

```
## Start:  AIC=748.32
## siri ~ age + weight + height + adipos + neck + chest + abdom +
##     hip + thigh + knee + ankle + biceps + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - knee     1      0.07 4390.9 746.32
## - chest    1      3.29 4394.1 746.51
## - ankle    1      7.59 4398.4 746.75
## - height   1     15.12 4406.0 747.18
## - adipos   1     20.10 4410.9 747.46
## - biceps   1     20.17 4411.0 747.47
## <none>                 4390.8 748.32
## - hip      1     45.16 4436.0 748.89
## - thigh    1     50.80 4441.6 749.20
## - weight   1     54.73 4445.6 749.43
## - neck     1     71.04 4461.9 750.35
## - age      1     72.55 4463.4 750.43
```

```
## - forearm  1     83.04 4473.9 751.02
## - wrist    1    171.01 4561.8 755.91
## - abdom    1   1906.72 6297.6 836.84
##
## Step:  AIC=746.32
## siri ~ age + weight + height + adipos + neck + chest + abdom +
##     hip + thigh + ankle + biceps + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - chest    1      3.31 4394.2 744.51
## - ankle    1      8.15 4399.1 744.79
## - height   1     15.23 4406.1 745.19
## - adipos   1     20.04 4410.9 745.46
## - biceps   1     20.10 4411.0 745.47
## <none>                 4390.9 746.32
## - hip      1     45.13 4436.0 746.89
## - weight   1     54.82 4445.7 747.44
## - thigh    1     57.21 4448.1 747.57
## - neck     1     72.39 4463.3 748.43
## - age      1     78.36 4469.3 748.76
## - forearm  1     84.07 4475.0 749.08
## - wrist    1    171.51 4562.4 753.94
## - abdom    1   1906.89 6297.8 834.85
##
## Step:  AIC=744.51
## siri ~ age + weight + height + adipos + neck + abdom + hip +
##     thigh + ankle + biceps + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - ankle    1      9.11 4403.3 743.03
## - height   1     13.88 4408.1 743.30
## - adipos   1     17.37 4411.6 743.50
## - biceps   1     19.54 4413.8 743.62
## <none>                 4394.2 744.51
## - hip      1     41.88 4436.1 744.89
## - weight   1     57.16 4451.4 745.75
## - thigh    1     65.53 4459.7 746.23
## - neck     1     72.00 4466.2 746.59
## - age      1     77.17 4471.4 746.88
## - forearm  1     81.80 4476.0 747.14
## - wrist    1    170.38 4564.6 752.06
## - abdom    1   2025.24 6419.5 837.65
##
## Step:  AIC=743.03
## siri ~ age + weight + height + adipos + neck + abdom + hip +
##     thigh + biceps + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - height   1     16.42 4419.8 741.96
## - biceps   1     17.95 4421.3 742.05
## - adipos   1     20.95 4424.3 742.22
## <none>                 4403.3 743.03
## - hip      1     44.23 4447.6 743.54
## - weight   1     56.79 4460.1 744.25
```

```
## - thigh    1     68.38 4471.7 744.90
## - age      1     74.48 4477.8 745.24
## - forearm  1     80.33 4483.7 745.57
## - neck     1     80.79 4484.1 745.59
## - wrist    1    161.37 4564.7 750.06
## - abdom    1   2040.61 6443.9 836.61
##
## Step:  AIC=741.96
## siri ~ age + weight + adipos + neck + abdom + hip + thigh + biceps +
##     forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - adipos   1      5.63 4425.4 740.28
## - biceps   1     17.31 4437.1 740.95
## <none>                 4419.8 741.96
## - hip      1     40.66 4460.4 742.26
## - thigh    1     62.05 4481.8 743.46
## - age      1     69.67 4489.4 743.89
## - neck     1     87.35 4507.1 744.88
## - forearm  1     92.12 4511.9 745.14
## - weight   1     98.42 4518.2 745.49
## - wrist    1    156.57 4576.3 748.70
## - abdom    1   2128.51 6548.3 838.64
##
## Step:  AIC=740.28
## siri ~ age + weight + neck + abdom + hip + thigh + biceps + forearm +
##     wrist
##
##           Df Sum of Sq    RSS    AIC
## - biceps   1      21.4 4446.8 739.50
## <none>                 4425.4 740.28
## - hip      1      36.7 4462.1 740.36
## - thigh    1      66.9 4492.3 742.05
## - age      1      70.1 4495.5 742.23
## - neck     1      82.3 4507.7 742.91
## - forearm  1      94.9 4520.2 743.61
## - weight   1     102.8 4528.2 744.05
## - wrist    1     159.3 4584.7 747.16
## - abdom    1    3189.9 7615.3 874.53
##
## Step:  AIC=739.5
## siri ~ age + weight + neck + abdom + hip + thigh + forearm +
##     wrist
##
##           Df Sum of Sq    RSS    AIC
## <none>                 4446.8 739.50
## - hip      1      41.5 4488.3 739.83
## - neck     1      74.3 4521.1 741.65
## - age      1      76.2 4523.0 741.76
## - weight   1      88.6 4535.4 742.45
## - thigh    1      94.6 4541.4 742.78
## - forearm  1     138.0 4584.8 745.16
## - wrist    1     158.4 4605.2 746.28
## - abdom    1    3170.1 7616.9 872.58
```

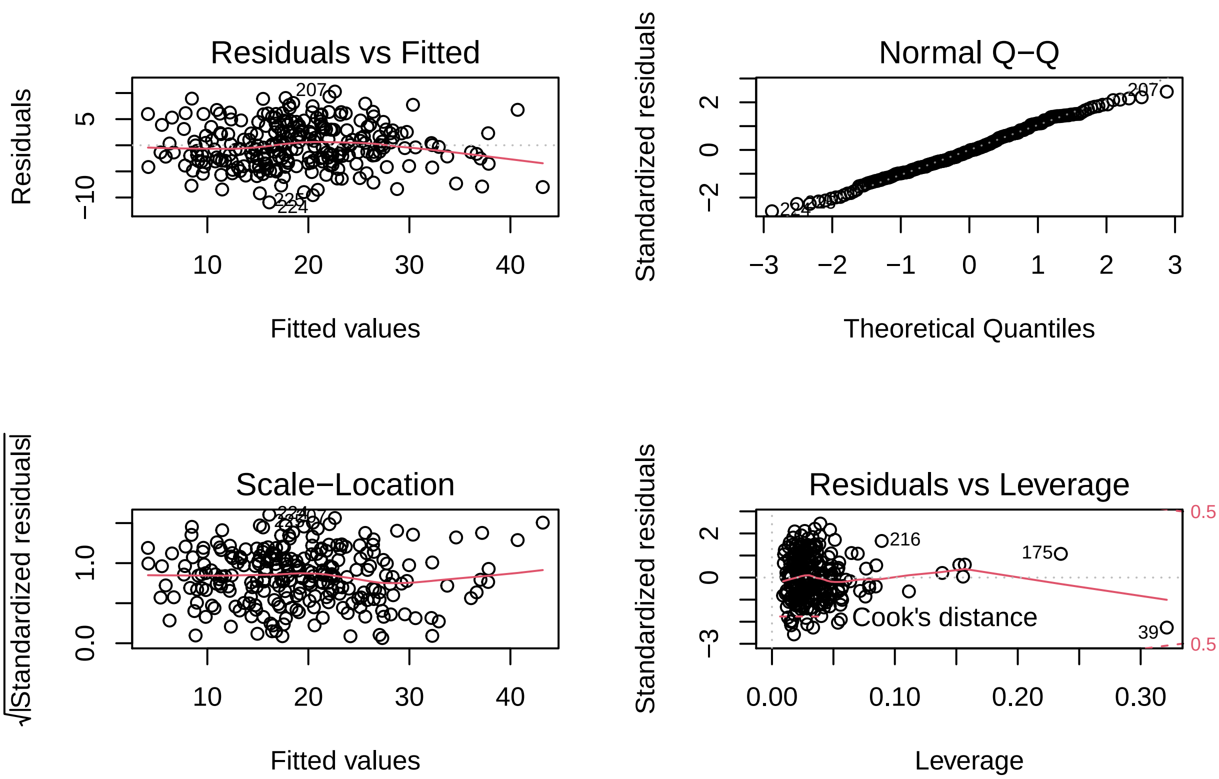
```
##
## Call:
## lm(formula = siri ~ age + weight + neck + abdom + hip + thigh +
##     forearm + wrist, data = fat_data, subset = (cook < 1))
##
## Coefficients:
## (Intercept)          age       weight         neck        abdom          hip
##   -22.11598      0.06322     -0.08796     -0.45374      0.94840     -0.21137
##       thigh      forearm        wrist
##     0.29440      0.51137     -1.50392
```

最後選出的模型為

$$g_{3.3} \ : \ siri \ \sim \ age + weight + neck + abdom + hip + thigh + forearm + wrist$$

然後檢查此模型的 diagnostics：

```r
g3.3 = lm(formula = siri ~ age + weight + neck + abdom + hip + thigh +
              forearm + wrist, data = fat_data, subset = (cook < 1))
par(mfrow = c(2,2))
plot(g3.3)
```

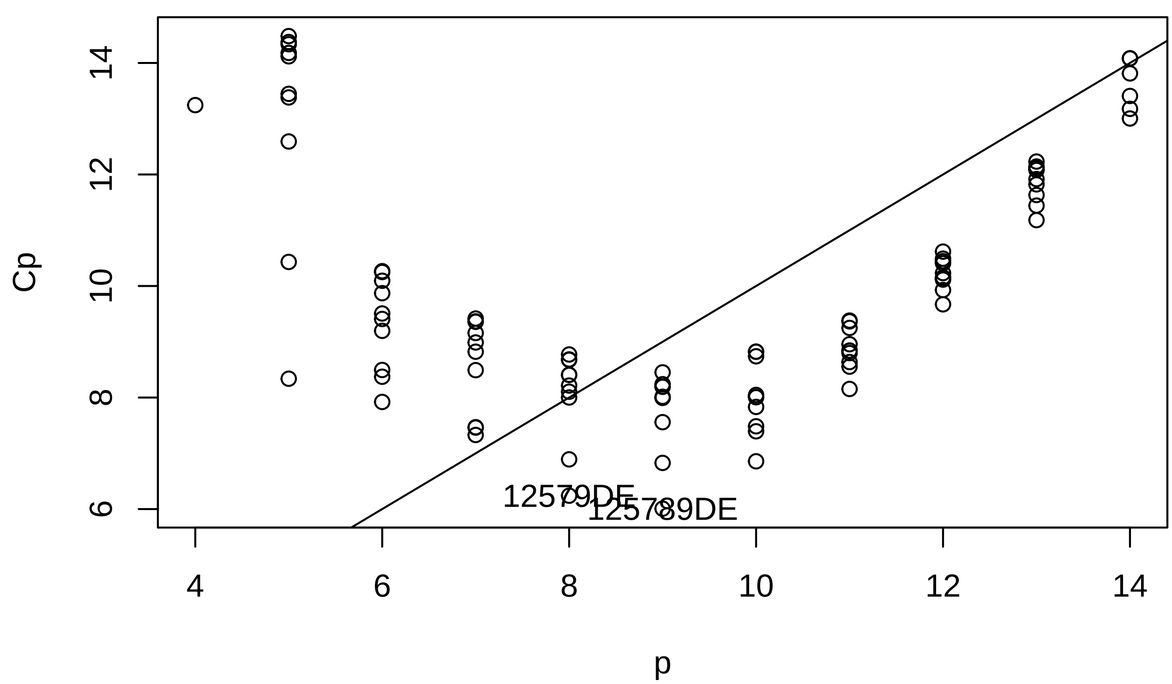


接著嘗試使用 Mallow's Cp statistics：

```
library(leaps)
x = fat_data[,-1][-42,]
y = fat_data[,1][-42,]$siri
gcp = leaps(x,y) # Cp
par(mfrow = c(1,1))
# plot(gcp$size, gcp$Cp, xlab = "p", ylab = "Cp")
small = (gcp$Cp < 15)
plot(gcp$size[small], gcp$Cp[small], xlab = "p", ylab = "Cp")
abline(0,1)

gcp.labels =
  apply(gcp$which, 1,
        function(x) paste(as.character(c(1:9,"A","B","C","D","E")[x]),collapse = ""))
# text(gcp$size[small], gcp$Cp[small], gcp.labels[small])

text(8, min(gcp$Cp[gcp$size==8]),
     gcp.labels[gcp$Cp==min(gcp$Cp[gcp$size==8])])
text(9, min(gcp$Cp[gcp$size==9]),
     gcp.labels[gcp$Cp==min(gcp$Cp[gcp$size==9])]) # the same result as AIC
```



挑選 $C_p \approx p$ or $C_p < p$ 的模型，可發現 p=8,9 兩種情況下，C_p 最小的模型即為 $g_{3.3}$ 和該模型扣除 hip 變數，其餘模型的使用變數較多較複雜，故不考慮選擇。

再來使用 adjusted R^2：

```r
gadjr = leaps(x,y, method = "adjr2") # adjusted R square
gadjr.labels =
  apply(gadjr$which, 1,
        function(x) paste(as.character(c(1:9,"A","B","C","D","E")[x]), collapse = ""))
names(gadjr$adjr2) = gadjr.labels
round(sort(gadjr$adjr2, decreasing = T)[1:8], 4)
```

```
##   125789CDE    125789DE  125789BCDE   125789BDE   1245789DE  1245789CDE
##      0.7360      0.7358      0.7357      0.7354      0.7353      0.7352
##  12345789DE 12345789CDE
##      0.7352      0.7351
```

adjusted R^2 最大的模型為 $g_{3.3}$ 再加上變數 *biceps*，但其 R_a^2 值只比第二大的模型 $g_{3.3}$ 高出 0.0002，且 $g_{3.3}$ 使用的變數更少。

綜合 AIC, C_p, adjusted R^2 三種方法，我會選擇模型

$$g_{3.3} \ : \ siri \ \sim \ age + weight + neck + abdom + hip + thigh + forearm + wrist$$

```r
summary(g3.3)
```

```
##
## Call:
## lm(formula = siri ~ age + weight + neck + abdom + hip + thigh +
##     forearm + wrist, data = fat_data, subset = (cook < 1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.933  -2.995  -0.211   2.981  10.273
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -22.11598   11.75357  -1.882  0.06109 .
## age           0.06322    0.03104   2.037  0.04276 *
## weight       -0.08796    0.04005  -2.196  0.02901 *
## neck         -0.45374    0.22565  -2.011  0.04545 *
## abdom         0.94840    0.07221  13.135  < 2e-16 ***
## hip          -0.21137    0.14058  -1.504  0.13400
## thigh         0.29440    0.12972   2.270  0.02412 *
## forearm       0.51137    0.18662   2.740  0.00660 **
## wrist        -1.50392    0.51221  -2.936  0.00364 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.287 on 242 degrees of freedom
## Multiple R-squared:  0.7443, Adjusted R-squared:  0.7358
## F-statistic: 88.04 on 8 and 242 DF,  p-value: < 2.2e-16
```