Linear Model Assignment 7

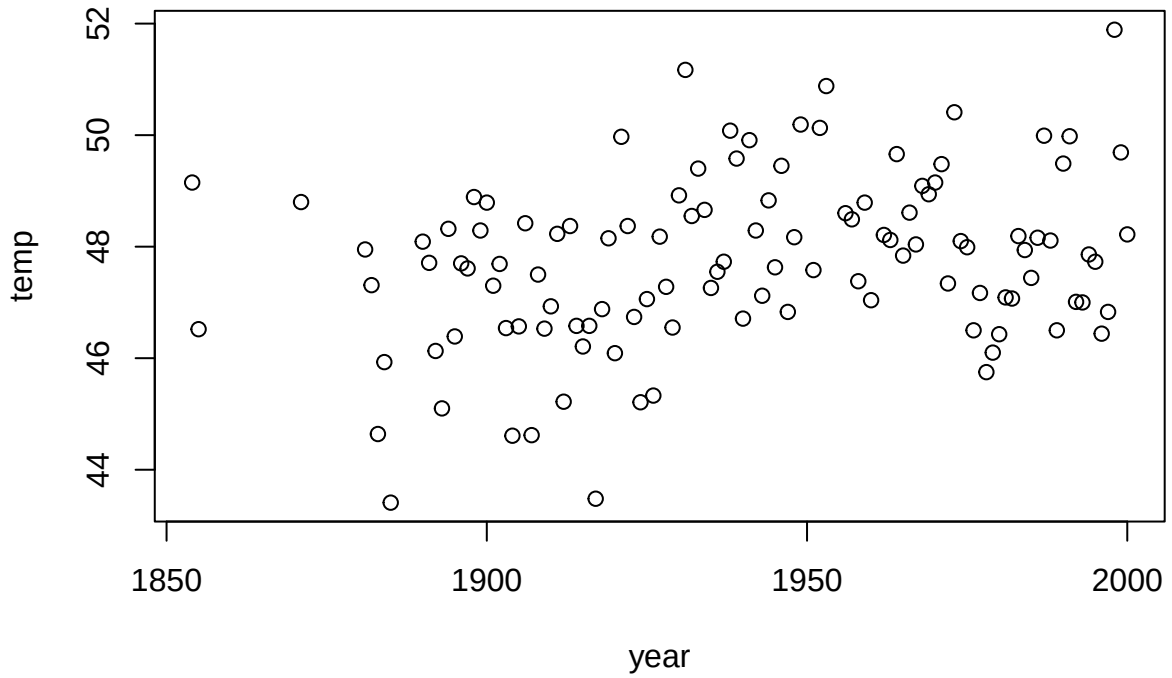
黃晨溦、劉奕宏、鄭雅珊

Problem 1.

(i)

首先對資料繪製散布圖：

```
library(nlme);library(splines)
library(car)
aatemp=read.table("http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/aatemp.txt",
                  header=T)
plot(aatemp$year,aatemp$temp,xlab ="year",ylab="temp")
```



從圖中看起來溫度隨著時間有一個遞增的趨勢。接著對資料配適 linear model:

$$\Omega_1 : \text{temp} = \beta_0 + \beta_1 \text{year} + \epsilon,\ \epsilon \sim N(0, \sigma^2)$$

```
g=lm(temp~year,data=aatemp)
summary(g)
```

```
##
## Call:
## lm(formula = temp ~ year, data = aatemp)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.9843 -0.9113 -0.0820  0.9946  3.5343
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.005510   7.310781   3.284  0.00136 **
## year         0.012237   0.003768   3.247  0.00153 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.466 on 113 degrees of freedom
## Multiple R-squared:  0.08536,    Adjusted R-squared:  0.07727
## F-statistic: 10.55 on 1 and 113 DF,  p-value: 0.001533
```

Fitted model 為 $\hat{\text{temp}} = 24.01 + 0.01\hat{\text{year}},\ \hat{\sigma} = 1.466$，且係數皆顯著大於 0，顯示溫度隨著時間有一個線性遞增的趨勢，平均每年約上升 0.01 degrees Fahrenheit。

(ii)

假設不同年份的溫度之間有關連性，並且假設關聯性為 AR(1)，意即

$$\Omega_2 : \text{temp}_i = \beta_0 + \beta_1 \text{year}_i + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2),\ \text{Cov}(\epsilon_i,\ \epsilon_j) = \sigma^2 \rho^{|\text{year}_i - \text{year}_j|}$$

使用 generalized least squares 配適模型如下：

```
g2=gls(temp~year,data=aatemp, correlation=corAR1(form=~year))
summary(g2)
```

```
## Generalized least squares fit by REML
##   Model: temp ~ year
##   Data: aatemp
##        AIC     BIC    logLik
##   426.5694 437.479 -209.2847
##
## Correlation Structure: ARMA(1,0)
##  Formula: ~year
##  Parameter estimate(s):
##      Phi1
## 0.2303887
##
## Coefficients:
##                Value Std.Error  t-value p-value
## (Intercept) 25.18407  8.971864 2.807006  0.0059
## year         0.01164  0.004626 2.516015  0.0133
##
##  Correlation:
##      (Intr)
## year -1
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.7230803 -0.6321970 -0.0520135  0.6645795  2.3775123
##
## Residual standard error: 1.475718
## Degrees of freedom: 115 total; 113 residual
```

Fitted model 為 $\hat{\text{temp}} = 25.18 + 0.01\hat{\text{year}},\ \hat{\sigma} = 1.476$，相較於 (i) 的配適結果係數並無太大變化。

```
intervals(g2)
```

```
## Approximate 95% confidence intervals
##
##  Coefficients:
##                   lower        est.       upper
## (Intercept) 7.409192415 25.18407264 42.95895286
## year        0.002474401  0.01164028  0.02080617
## attr(,"label")
## [1] "Coefficients:"
##
##  Correlation structure:
##          lower      est.     upper
## Phi1 0.02920118 0.2303887 0.4136364
## attr(,"label")
## [1] "Correlation structure:"
##
##  Residual standard error:
##    lower     est.    upper
## 1.284091 1.475718 1.695942
```

另外 $\hat{\rho} = 0.23$，參數估計值的 95% C.I. 為 $(0.029, 0.41)$ 不包含 0，顯示不同年份的溫度之間的確存在關連性。

(iii)

使用 orthogonal polynomials 配適十次多項式模型如下：

$$\text{temp} = \beta_0 + \sum_{i=1}^{10} \beta_i z_i(\text{year},\ ...,\ \text{year}^i) + \epsilon,\ z_i = \sum_{k=0}^{i} a_{ij}\text{year}^k,\ z_k^T z_l = 0 \text{ for } k \neq l$$

```
summary(lm(temp~poly(year, degree=10), data=aatemp))
```

```
##
## Call:
## lm(formula = temp ~ poly(year, degree = 10), data = aatemp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4987 -0.8641 -0.1745  1.1450  3.4255
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  47.7426     0.1319 361.927  < 2e-16 ***
## poly(year, degree = 10)1      4.7616     1.4146   3.366  0.00107 **
## poly(year, degree = 10)2     -0.9071     1.4146  -0.641  0.52277
## poly(year, degree = 10)3     -3.3132     1.4146  -2.342  0.02108 *
## poly(year, degree = 10)4      2.4383     1.4146   1.724  0.08774 .
## poly(year, degree = 10)5      3.3824     1.4146   2.391  0.01860 *
## poly(year, degree = 10)6      1.2124     1.4146   0.857  0.39337
## poly(year, degree = 10)7     -0.9373     1.4146  -0.663  0.50908
## poly(year, degree = 10)8     -1.1011     1.4146  -0.778  0.43812
## poly(year, degree = 10)9      1.3994     1.4146   0.989  0.32483
## poly(year, degree = 10)10     0.3474     1.4146   0.246  0.80652
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.415 on 104 degrees of freedom
## Multiple R-squared:  0.2165, Adjusted R-squared:  0.1411
## F-statistic: 2.873 on 10 and 104 DF,  p-value: 0.003335
```

由於六次以上的係數皆不顯著，我們可以選擇保留至五次多項式的模型：

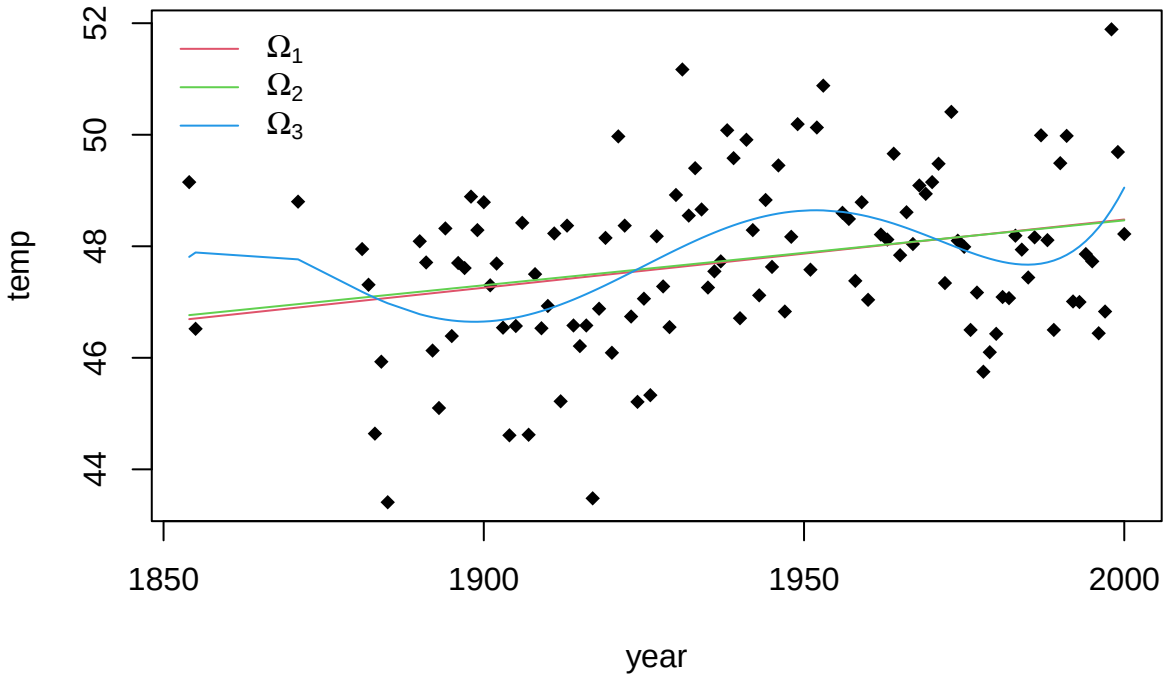
$$\Omega_3 : \text{temp} = \beta_0 + \sum_{i=1}^{5} \beta_i z_i(\text{year},\ ...,\ \text{year}^i) + \epsilon$$

```
g3=lm(temp~poly(year, degree=5), data=aatemp)
summary(g3)
```

```
##
## Call:
## lm(formula = temp ~ poly(year, degree = 5), data = aatemp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7142 -0.9198 -0.1420  0.9903  3.2364
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                47.7426     0.1306 365.604  < 2e-16 ***
## poly(year, degree = 5)1     4.7616     1.4004   3.400 0.000942 ***
## poly(year, degree = 5)2    -0.9071     1.4004  -0.648 0.518500
## poly(year, degree = 5)3    -3.3132     1.4004  -2.366 0.019749 *
## poly(year, degree = 5)4     2.4383     1.4004   1.741 0.084470 .
## poly(year, degree = 5)5     3.3824     1.4004   2.415 0.017384 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.4 on 109 degrees of freedom
## Multiple R-squared:  0.1952, Adjusted R-squared:  0.1583
## F-statistic: 5.289 on 5 and 109 DF,  p-value: 0.0002176
```

繪製前三題的 fitted line 如下圖：

```
x=aatemp$year; y=aatemp$temp
matplot(x,cbind(y,g$fit,g2$fit,g3$fit),type="plll",xlab="year",ylab="temp",pch=18,lty=1)
legend("topleft",col=2:4,lty=rep(1,3),bty="n",
       legend=c(expression(Omega[1]),expression(Omega[2]),
                                    expression(Omega[3])))
```



可發現 Ω_1 和 Ω_2 的配適線很接近，Ω_3 則是上下起伏的曲線。接著預測 2020 年的溫度：

```
predict(g3,newdata=data.frame(year=2020))
```

```
##        1
## 60.07774
```

預測溫度為 60.0774，相較於觀測資料的範圍 (43.41, 51.89) 距離很遠，這和五次多項式模型在接近 2000 年的時候有一向上轉折有關。相較之下線性模型 Ω_1 和 Ω_2 這兩個模型雖然配適較差 ($\hat{\sigma}$ 較大)，但預測結果會比較好。

(iv)

假設有人宣稱溫度在 1930 年前為不變的，而自 1930 年起有一個線性的走勢。我們可以針對這個假設配適 broken stick regression:

$$\Omega_4 : \text{temp} = \beta_0 + \beta_1(\text{year} - 1930)d(\text{year}) + \epsilon,\ \epsilon \sim N(0, \sigma^2)$$

, $d(\text{year}) = 1$ if year ≥ 1930 and $d(\text{year}) = 0$, otherwise.

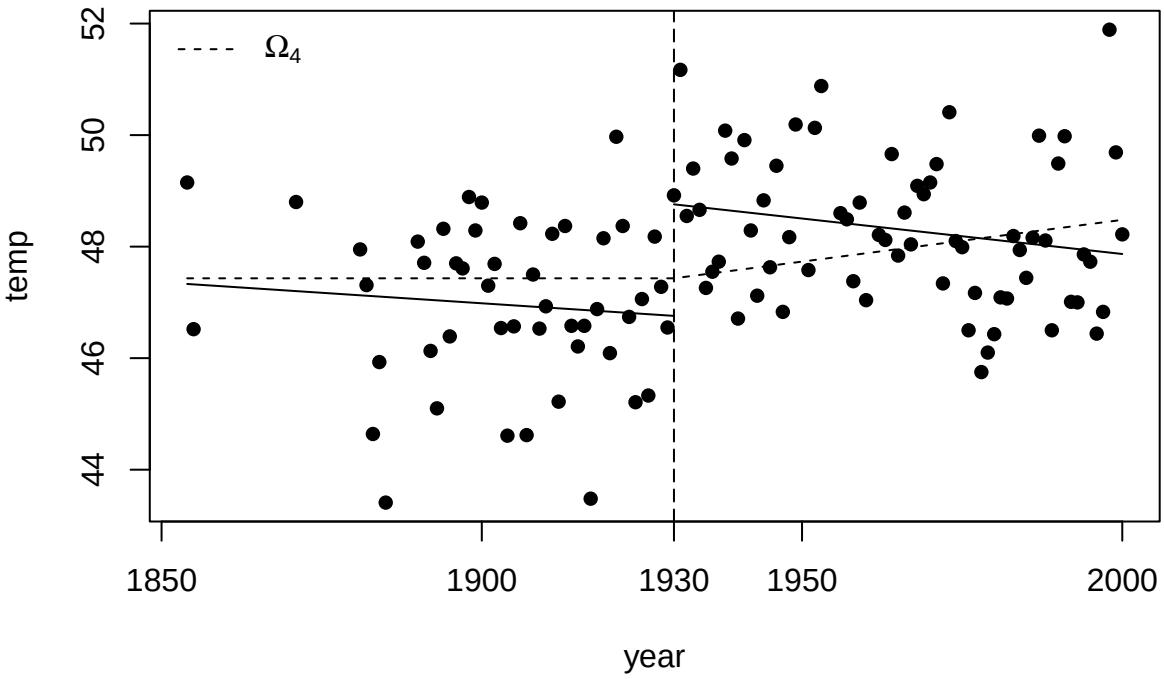
```
d=function(x) ifelse(x<1930, 0, 1)
gb=lm(temp~I((year-1930)*d(year)), data=aatemp)
summary(gb)
```

```
##
## Call:
## lm(formula = temp ~ I((year - 1930) * d(year)), data = aatemp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0222 -0.8872 -0.0355  0.9628  3.7229
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 47.432151   0.184604 256.939   <2e-16 ***
```

```
## I((year - 1930) * d(year))  0.014970   0.005857   2.556   0.0119 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.491 on 113 degrees of freedom
## Multiple R-squared:  0.05465,    Adjusted R-squared:  0.04628
## F-statistic: 6.532 on 1 and 113 DF,  p-value: 0.01192
```

雖然參數估計結果皆顯著，但 R^2 僅 0.055，顯示這可能不是一個很好的配適結果。若我們另外分別對 year < 1930 及 year ≥ 1930 配適 linear model 並繪製 fitted line 於資料散布圖上：

```
g41 <- lm(temp~year, data=aatemp, subset=(year<1930))
g42 <- lm(temp~year, data=aatemp, subset=(year>=1930))
plot(x, y, xlab="year",ylab="temp",pch=16)
abline(v=1930, lty=5)
axis(1,1930,1930)
segments(1854, g41$coef[1]+g41$coef[2]*1854, 1930, g41$coef[1]+g41$coef[2]*1930)
segments(2000, g42$coef[1]+g42$coef[2]*2000, 1930, g42$coef[1]+g42$coef[2]*1930)
py = gb$coef[1]+gb$coef[2]*((x-1930)*d(x))
lines(x, py, lty=2)
legend("topleft",col=1,lty=2,bty="n",
       legend=c(expression(Omega[4])))
```



可見兩條 fitted line (黑色實線) 的斜率皆為負值，且在 1930 年兩條線並無接近或交會，而是有一段落差，因此 broken stick 模型 (Ω_4) 無法解釋這個急遽增加的現象。

(v)

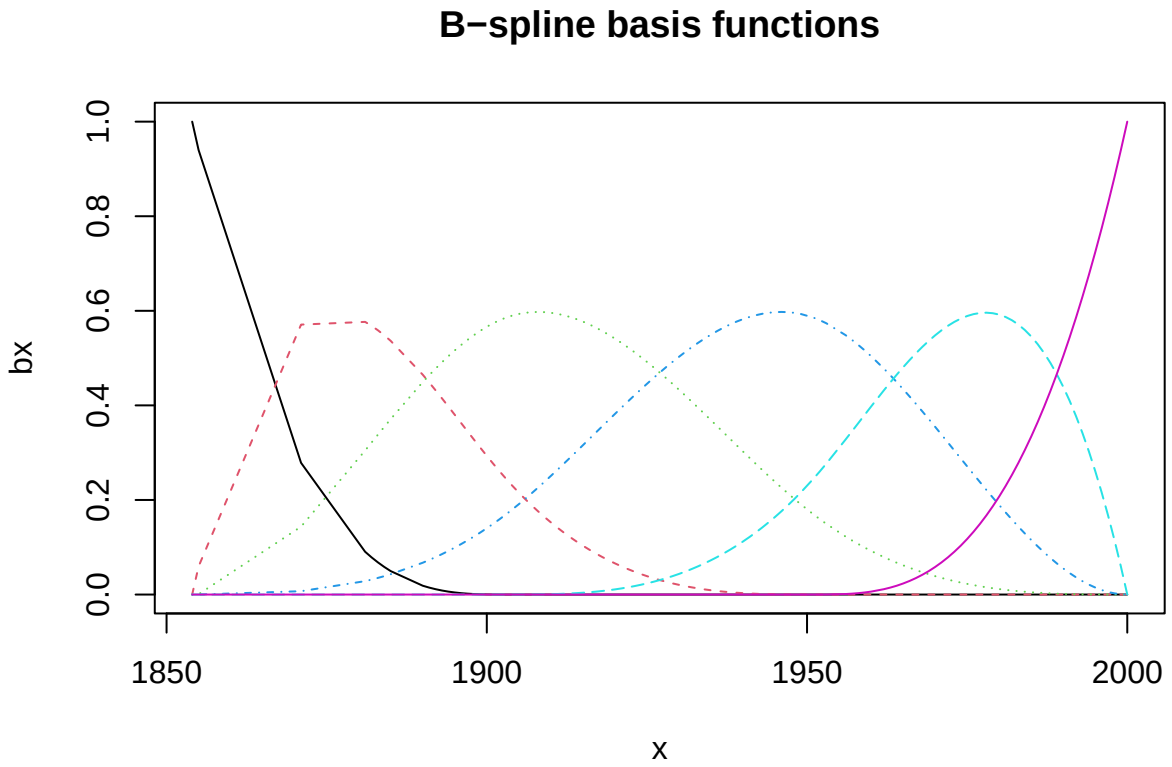
依題意配適 cubic B-spline 模型：

$$\Omega_5 : \text{temp} = \sum_{i=1}^{6} \beta_i g_i(\text{year}) + \epsilon$$

base functions: $g_1, ..., g_6$ defined on an interval [a=1854, b=2000] with knot-points $t_1 \leq ... \leq t_k$, $k = 10 (\because k - 4 = 6)$, $t_1 = 1854$, $t_{10} = 2000$. 因為 year 為整數，等間距取 10 個 knots (四捨五入) 為 (1854, 1854,

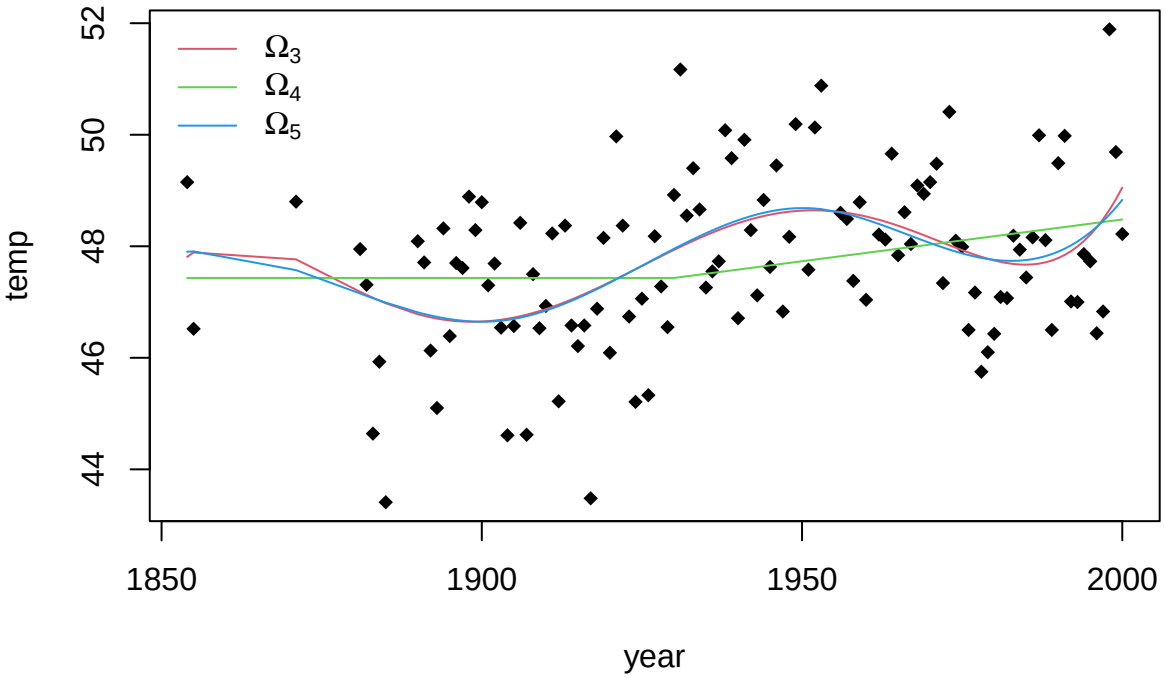
1854, 1854, 1903, 1951, 2000, 2000, 2000, 2000)。B-spline base functions 繪製如下：

```
knots=c(rep(1854,3),round(seq(1854,2000,length.out=4)),rep(2000,3))
bx=splineDesign(knots,x)
matplot(x,bx,type="l",main="B-spline basis functions")
```



接著繪製 $\Omega_i,\ i=3,4,5$ 配適曲線如下圖：

```
gs=lm(aatemp$temp~bx)
matplot(x,cbind(aatemp$temp,g3$fit,gb$fit,gs$fit),type="plll"
        ,xlab ="year",ylab="temp",pch=18,lty=1)
legend("topleft",col=2:4,lty=rep(1,3),bty="n",
       legend=c(expression(Omega[3]),expression(Omega[4]),
                                     expression(Omega[5])))
```



圖中可見五次多項式模型 (Ω_3) 及 cubic B-spline 模型 (Ω_5) 能更貼切描述溫度與時間的關係。但可能無法準確預測未來時間點的溫度。

Problem 2.

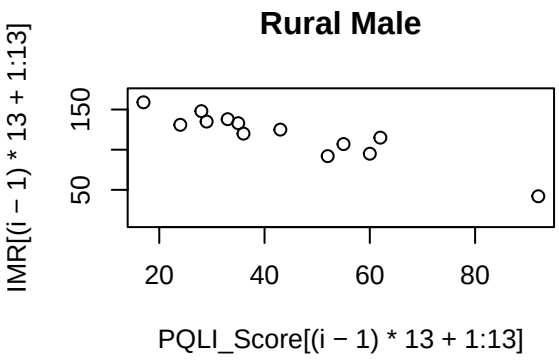
The data contains Infant Mortality Rates (IMR) and Physical Quality of Life Index (PQLI) scores, which is an indicator of average wealth, for selected Indian States. Using the data set, construct a single model for infant mortality rate, using suitably defined dummy variables for rural-urban and male-female distinctions. You should investigate whether there is a male-female and/or rural-urban difference in mortality rate after adjusting for other covariates.

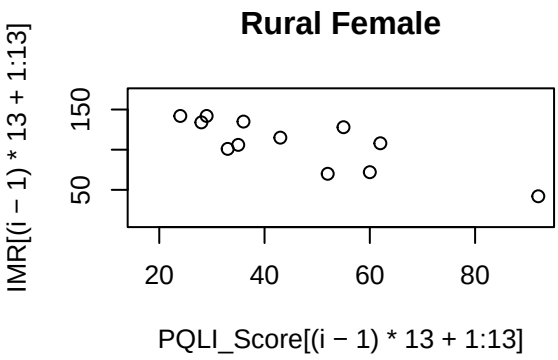
該數據選定印度各邦，收集嬰兒死亡率 (IMR) 和生活質量分數 (PQLI)。在各邦收集 PQLI 分數時，有分別用鄉村/城市的區分，以及農村/城市的區分來收集 PQLI 分數。我們目標是調查 IMR 是否存在男女或城鄉差異，所以需要先重新整理為 newdata。在 newdata 的 IMR 欄位裡，我們放的不再是 combined IMR 而是區分類別時的 IMR，並新增城鄉 (UoR) 和男女 (Sex) 類別欄位。

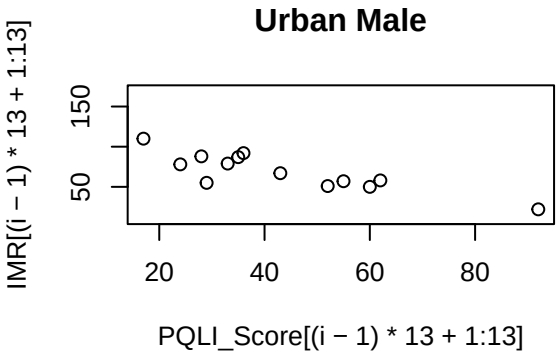
```
library(ggplot2)
library(GGally)
data = read.table("Hw7_data2.txt", header =F,
                  col.names = c("State","PQLI_Score","Combined_IMR",
                                "Rural_Male_IMR","Rural_Female_IMR",
                                "Urban_Male_IMR","Urban_Female_IMR"))
State = rep(data[,1], 4)
PQLI_Score = rep(data[,2], 4)
IMR = unlist(data[,-c(1,2,3)])
UoR = c(rep("R",26),rep("U",26))
Sex = rep(c(rep("M",13),rep("F",13)),2)
newdata = data.frame(State, PQLI_Score, IMR, UoR, Sex)
summary(newdata)
```

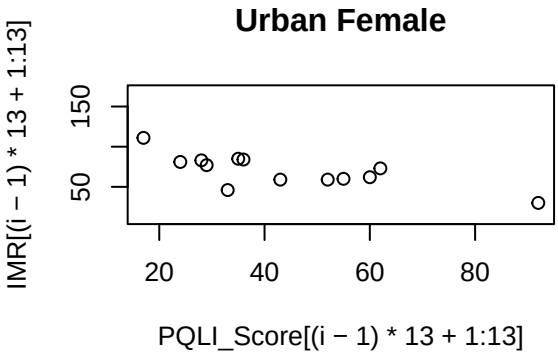
```
##     State              PQLI_Score         IMR             UoR
##  Length:52          Min.   :17.00   Min.   : 22.00   Length:52
##  Class :character   1st Qu.:29.00   1st Qu.: 61.50   Class :character
##  Mode  :character   Median :36.00   Median : 87.50   Mode  :character
##                     Mean   :43.54   Mean   : 92.81
##                     3rd Qu.:55.00   3rd Qu.:121.25
##                     Max.   :92.00   Max.   :187.00
##      Sex
##  Length:52
##  Class :character
##  Mode  :character
##
##
##
```

```
par(mfrow = c(2,2))
# for(i in 1:4) { # 如果要對每個類別畫 fitting line 的話
#   assign(paste0("fit2_", i), lm(IMR ~ PQLI_Score, subset = (i-1)*13+1:13))
# }
for(i in 1:4){
  plot(PQLI_Score[(i-1)*13+1:13], IMR[(i-1)*13+1:13],
       main = c("Rural Male","Rural Female","Urban Male","Urban Female")[i],
       ylim = c(10,170))
  # 如果要對每個類別畫 fitting line 的話
  # abline(c(fit2_1$coefficients,fit2_2$coefficients,
  #          fit2_3$coefficients,fit2_4$coefficients)[(i-1)*2+1:2])
}
```

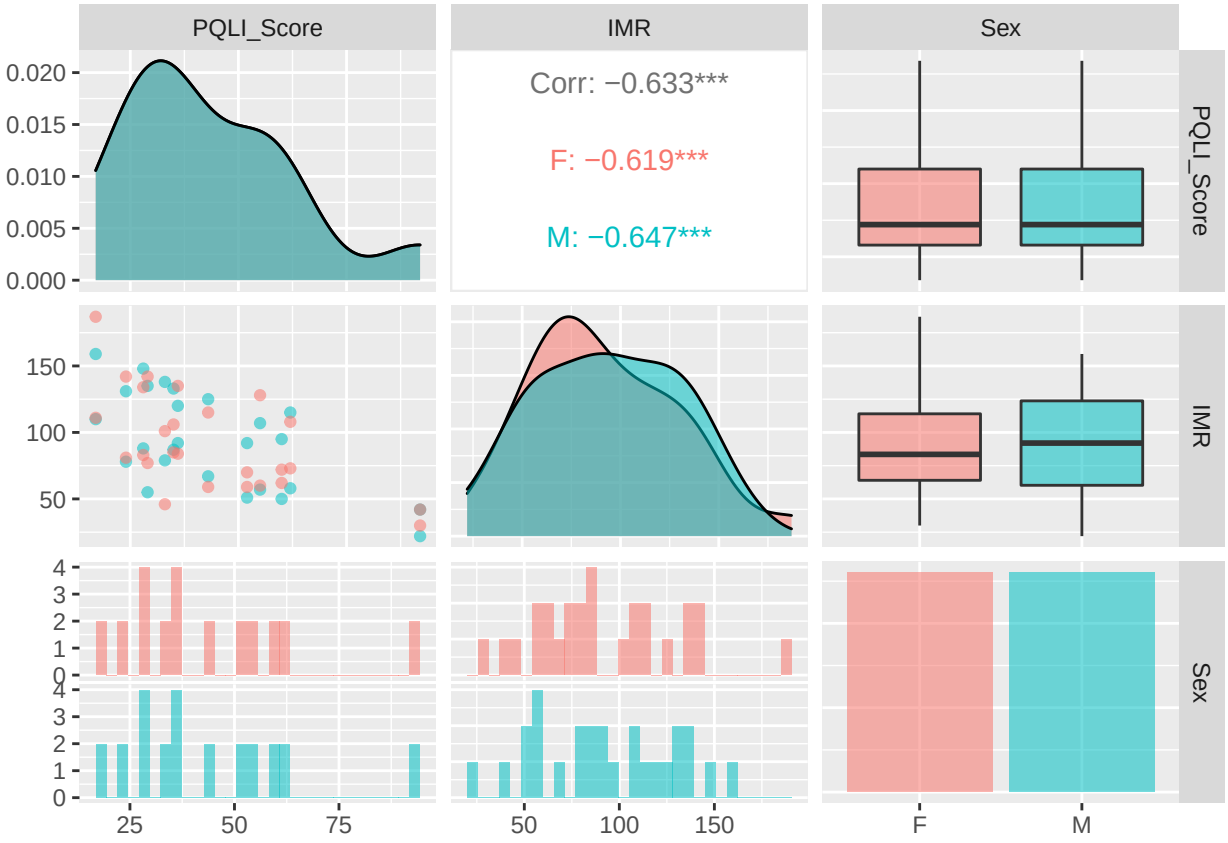




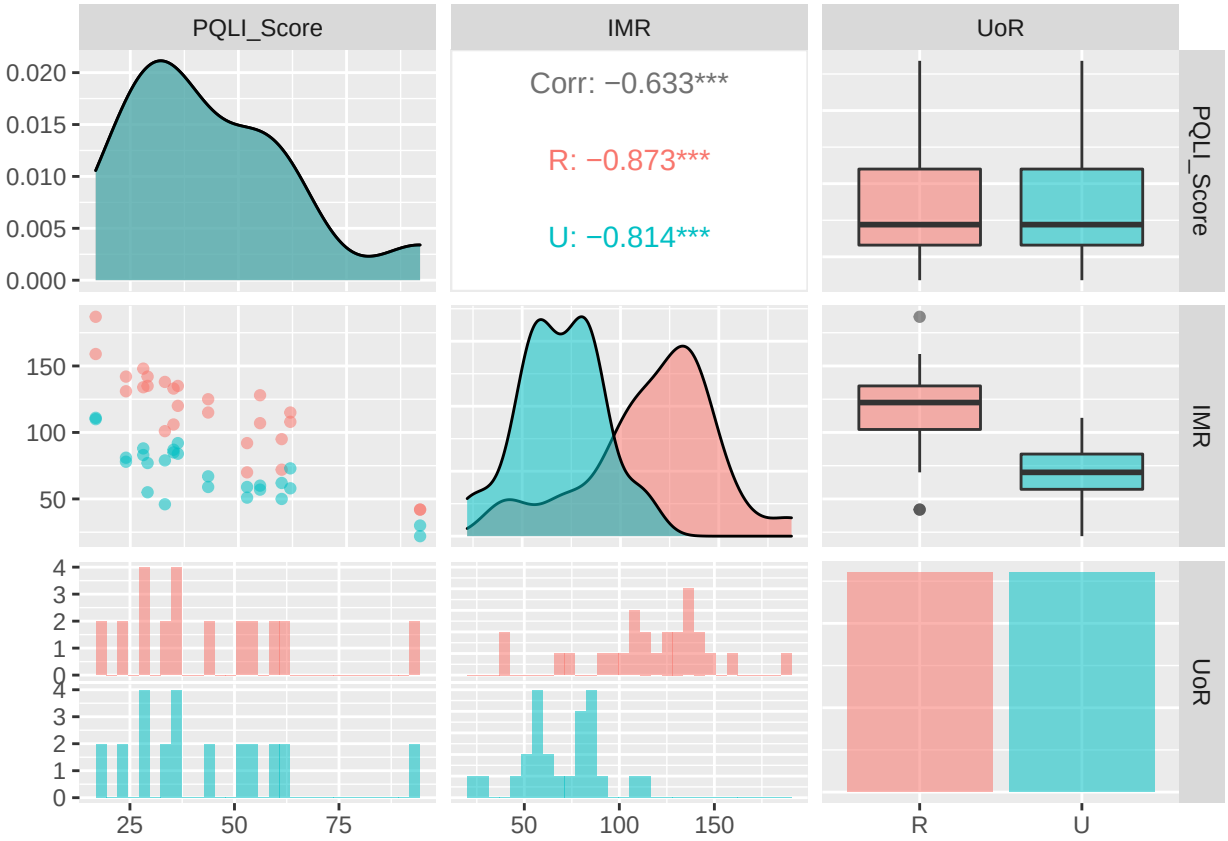





```
ggpairs(newdata[,c("PQLI_Score","IMR", "Sex")], aes(color = Sex, alpha = 0.5))
```



```
ggpairs(newdata[,c("PQLI_Score","IMR", "UoR")], aes(color = UoR, alpha = 0.5))
```



首先分別在不同類別下對資料進行 EDA：
(1) 從個別繪製的散佈圖可以觀察到在不同類別下，PQLI 對 IMR 的關係呈線性，無明顯非線性 pattern。四者趨勢都是 PQLI 越高，IMR 越低，並且又以 Urban 組 IMR 較 Rural 組低。此外，我們觀察到各圖最右下角的點，和其他點的距離較遠，可能是潛在的 outlier。
(2) 從對類別著色的散佈圖可以觀察到，PQLII 對 IMR 的關係存在城鄉 (UoR) 差異，在性別 (Sex) 差異上則較不明顯。

我們首先配適 full model。從 model.matrix 指令可以看到類別變數以 (0,1)-coding 的方式被包含在模型裡。

$$\Omega : \text{IMR} = \text{PQLI} + \text{Urban} + \text{Male} + \text{PQLI:Urban} + \text{PQLI:Male} + \text{Urban:Male} + \text{PQLI:Urban:Male} + \epsilon$$

```
fit2_1 = lm(IMR ~ PQLI_Score * UoR * Sex, data = newdata)
summary(fit2_1)
```

```
##
## Call:
## lm(formula = IMR ~ PQLI_Score * UoR * Sex, data = newdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.110  -5.603   0.007   7.546  31.882
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)          181.4581    10.3419  17.546  < 2e-16 ***
## PQLI_Score            -1.5494     0.2168  -7.147 6.96e-09 ***
## UoRU                 -77.9537    14.6257  -5.330 3.22e-06 ***
## SexM                  -2.4329    14.6257  -0.166   0.8687
## PQLI_Score:UoRU        0.7799     0.3066   2.544   0.0146 *
## PQLI_Score:SexM        0.1584     0.3066   0.517   0.6081
## UoRU:SexM             10.5945    20.6838   0.512   0.6111
## PQLI_Score:UoRU:SexM  -0.3741     0.4336  -0.863   0.3929
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.24 on 44 degrees of freedom
## Multiple R-squared:  0.8498, Adjusted R-squared:  0.8259
## F-statistic: 35.56 on 7 and 44 DF,  p-value: 4.329e-16
```

```
# model.matrix(fit2_1) 檢查 coding
```

可以看到有關 Sex 的預測變數都不顯著，這和我們 EDA 裡的觀察相符。於是我們嘗試拿掉 Sex 變數，聚焦在鄉村差異，配適 reduced model 1。此外，為了確認 PQLI 對 IMR 的趨勢，在沒有鄉村變數下，確實也不存在性別差異，我們配適 reduced model 2。我們也將 reduced model 1 和 2 的結果，分別繪製在以城鄉為區分或以性別為區分的散佈圖上視覺化不同類別標記下的差別。

$$\omega_1 : \text{IMR} = \text{PQLI} + \text{Urban} + \text{PQLI:Urban} + \epsilon$$

$$\omega_2 : \text{IMR} = \text{PQLI} + \text{Sex} + \text{PQLI:Sex} + \epsilon$$

```
fit2_2 = lm(IMR ~ PQLI_Score * UoR, data = newdata)
summary(fit2_2)
```

```
##
## Call:
## lm(formula = IMR ~ PQLI_Score * UoR, data = newdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.790  -5.237   0.175   7.759  31.752
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     180.2417     7.1090  25.354  < 2e-16 ***
## PQLI_Score       -1.4702     0.1490  -9.866 3.93e-13 ***
## UoRU            -72.6564    10.0536  -7.227 3.30e-09 ***
## PQLI_Score:UoRU   0.5928     0.2107   2.813  0.00709 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.82 on 48 degrees of freedom
## Multiple R-squared:  0.8451, Adjusted R-squared:  0.8355
## F-statistic: 87.32 on 3 and 48 DF,  p-value: < 2.2e-16
```

```
anova(fit2_2, fit2_1)
```

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

```
## Model 1: IMR ~ PQLI_Score * UoR
## Model 2: IMR ~ PQLI_Score * UoR * Sex
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     48 10538
## 2     44 10222  4    316.23 0.3403 0.8493
```

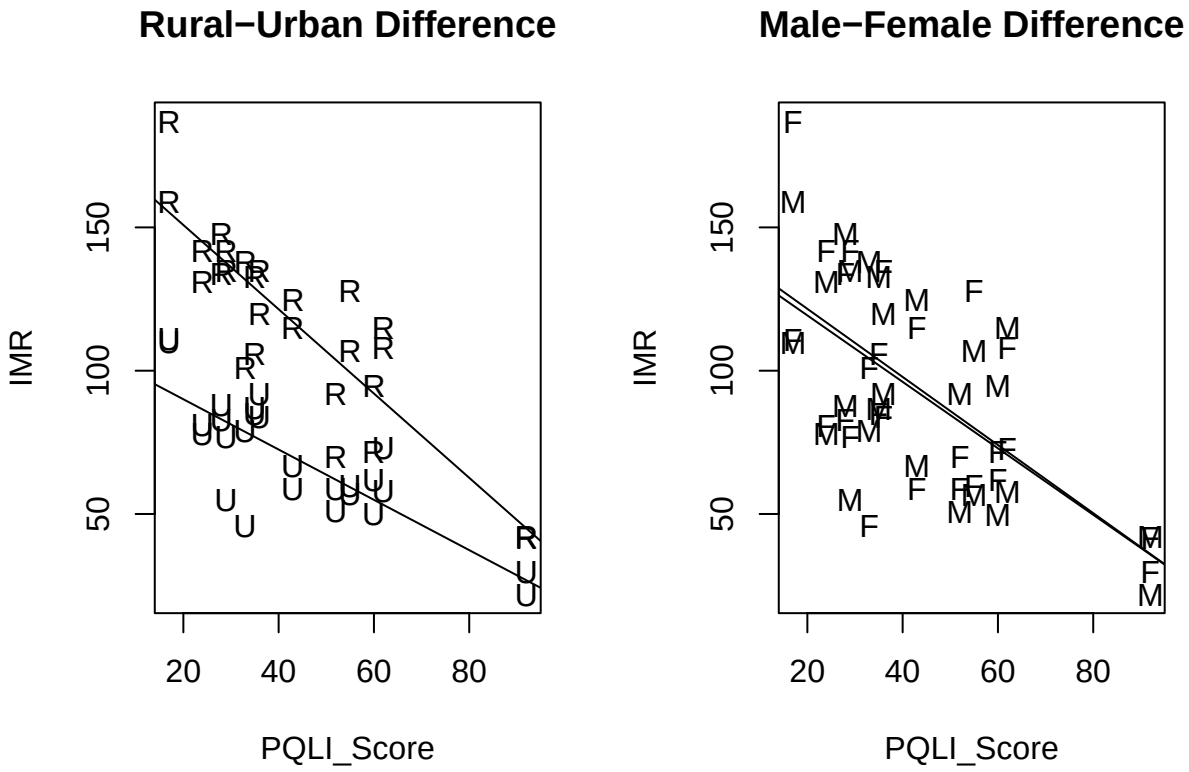
```
fit2_3 = lm(IMR ~ PQLI_Score * Sex, data = newdata)
summary(fit2_3)
```

```
##
## Call:
## lm(formula = IMR ~ PQLI_Score * Sex, data = newdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.219 -23.045  -4.515  24.834  64.230
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     142.48130   13.98240  10.190 1.36e-13 ***
## PQLI_Score       -1.15946    0.29310  -3.956 0.000251 ***
## SexM              2.86438   19.77411   0.145 0.885432
## PQLI_Score:SexM  -0.02869    0.41450  -0.069 0.945111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.14 on 48 degrees of freedom
## Multiple R-squared:  0.4009, Adjusted R-squared:  0.3635
## F-statistic: 10.71 on 3 and 48 DF,  p-value: 1.671e-05
```

```
anova(fit2_3, fit2_1)
```

```
## Analysis of Variance Table
##
## Model 1: IMR ~ PQLI_Score * Sex
## Model 2: IMR ~ PQLI_Score * UoR * Sex
##   Res.Df   RSS Df Sum of Sq     F   Pr(>F)
## 1     48 40766
## 2     44 10222  4     30545 32.87 1.06e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow = c(1,2))
plot(PQLI_Score, IMR, type = "n", main = "Rural-Urban Difference")
text(PQLI_Score, IMR, UoR)
abline(fit2_2$coefficients[1:2])
abline(fit2_2$coefficients[1:2] + fit2_2$coefficient[3:4])
plot(PQLI_Score, IMR, type = "n", main = "Male-Female Difference")
text(PQLI_Score, IMR, Sex)
abline(fit2_3$coefficients[1:2])
abline(fit2_3$coefficients[1:2] + fit2_3$coefficient[3:4])
```



以 reduced model 1 來說，所有變數皆為顯著。使用 anova 比較 reduced model 1 與 full model，得到 p-value > 0.05，代表兩模型無顯著差異，我們可以選擇捨棄性別變數選擇變數較少的 reduced model 1。

以 reduced model 2 來說，性別相關變數皆不顯著。使用 anova 比較 reduced model 2 與 full model，得到 p-value < 0.05，代表兩模型有顯著差異，我們沒辦法捨棄城鄉變數選擇變數較少的 reduced model 2。

在視覺化的結果上也能直觀地了解，城鄉差異較為明顯，而性別差異較不明顯的事實。綜合以上結果，在後續的討論裡我們以 reduced model 1 為主。

而當我們視覺化 reduced model 1 的結果時，發現右下角的資料點相較於其他資料點，PQLI 分數特別高，對於 regression 來說可能是潛在的 influential point，這筆數據來自 KERALA 邦。移除 KERALA 邦再做一次 reduced model 1 的 fitting。

```
newdata$State[which(PQLI_Score == max(PQLI_Score))]
```

```
## [1] "KERALA" "KERALA" "KERALA" "KERALA"
```

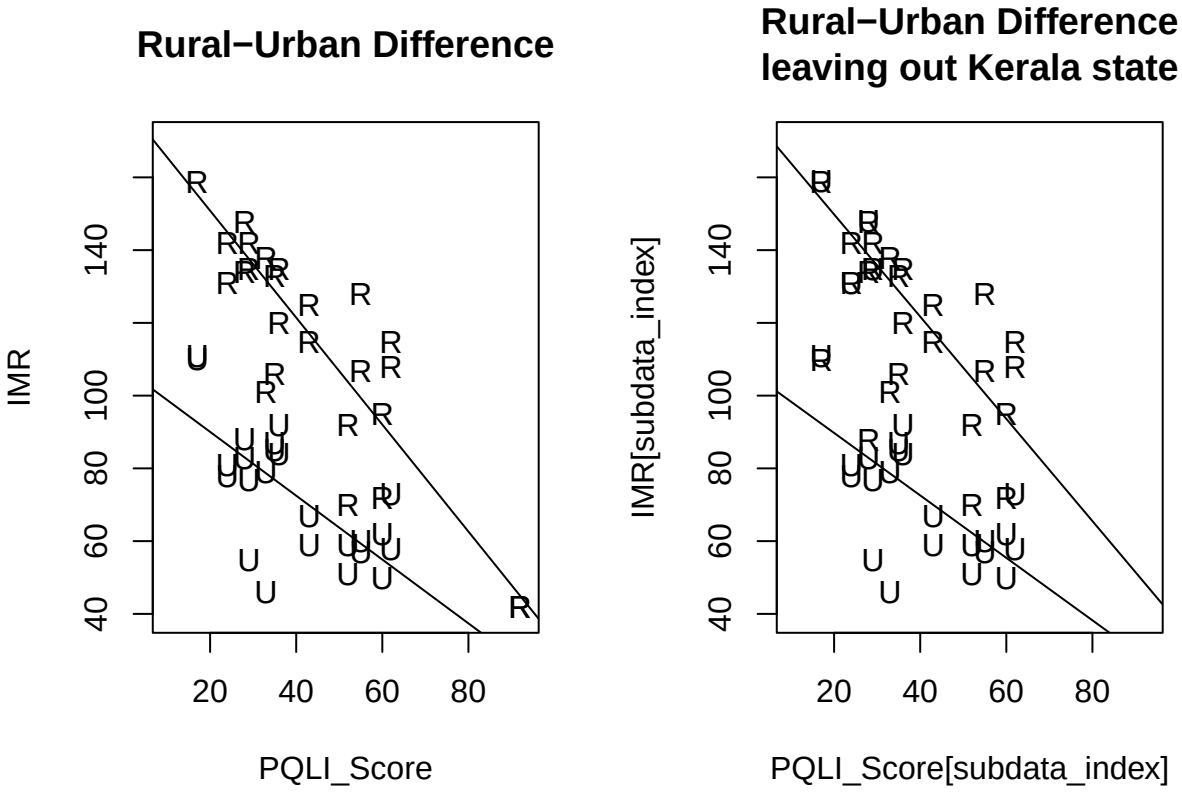
```
subdata_index = rep(T,52)
subdata_index[which(PQLI_Score == max(PQLI_Score))] = F

fit2_4 = lm(IMR ~ PQLI_Score * UoR, data = newdata, subset = subdata_index)
summary(fit2_4)
```

```
##
## Call:
## lm(formula = IMR ~ PQLI_Score * UoR, data = newdata, subset = subdata_index)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.855  -5.884   0.363   7.748  32.972
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)     177.9126     9.3400  19.048  < 2e-16 ***
## PQLI_Score       -1.4050     0.2226  -6.311 1.18e-07 ***
## UoRU            -71.0025    13.2087  -5.375 2.77e-06 ***
## PQLI_Score:UoRU   0.5465     0.3148   1.736   0.0896 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.42 on 44 degrees of freedom
## Multiple R-squared:  0.8017, Adjusted R-squared:  0.7882
## F-statistic: 59.31 on 3 and 44 DF,  p-value: 1.677e-15
```

```
par(mfrow = c(1,2))
plot(PQLI_Score, IMR, type = "n", main = "Rural-Urban Difference",
     xlim = c(10,93), ylim = c(40,170))
text(PQLI_Score, IMR, UoR)
abline(fit2_2$coefficients[1:2])
abline(fit2_2$coefficients[1:2] + fit2_2$coefficient[3:4])
plot(PQLI_Score[subdata_index], IMR[subdata_index], type = "n", main = "Rural-Urban Difference\nleaving
     xlim = c(10,93), ylim = c(40,170))
text(PQLI_Score[subdata_index], IMR[subdata_index], UoR)
abline(fit2_4$coefficients[1:2])
abline(fit2_4$coefficients[1:2] + fit2_4$coefficient[3:4])
```



從視覺化的結果圖可以發現，城鄉間迴歸線的斜率差異減少了；統計結果上對應則是 PQLI:Uraban 交互作用項變得不顯著。移除交互作用項，再配適 reduced model 3。

$$\omega_3 : \text{IMR} = \text{PQLI} + \text{Urban} + \epsilon$$

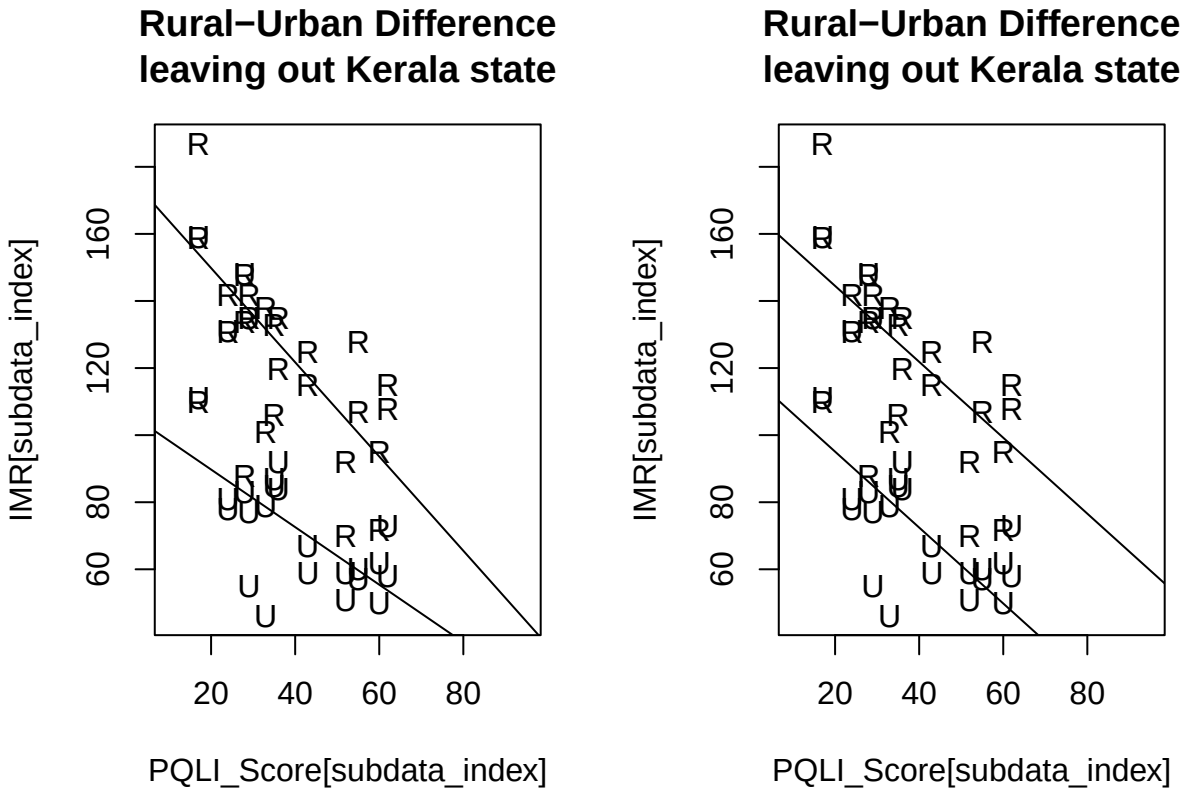
```
fit2_5 = lm(IMR ~ PQLI_Score + UoR, data = newdata, subset = subdata_index)
summary(fit2_5)
```

```
##
## Call:
## lm(formula = IMR ~ PQLI_Score + UoR, data = newdata, subset = subdata_index)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.270  -7.861   1.763   9.296  39.120
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 167.1197     7.1235  23.460  < 2e-16 ***
## PQLI_Score   -1.1317     0.1609  -7.034 9.08e-09 ***
## UoRU        -49.4167     4.5500 -10.861 3.68e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.76 on 45 degrees of freedom
## Multiple R-squared:  0.7882, Adjusted R-squared:  0.7788
## F-statistic: 83.71 on 2 and 45 DF,  p-value: 6.838e-16
```

```
anova(fit2_5, fit2_4)
```

```
## Analysis of Variance Table
##
## Model 1: IMR ~ PQLI_Score + UoR
## Model 2: IMR ~ PQLI_Score * UoR
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1     45 11179
## 2     44 10463  1    716.43 3.0128 0.08961 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow = c(1,2))
plot(PQLI_Score[subdata_index], IMR[subdata_index], type = "n", main = "Rural-Urban Difference\nleaving
text(PQLI_Score[subdata_index], IMR[subdata_index], UoR)
abline(fit2_4$coefficients[1:2])
abline(fit2_4$coefficients[1:2] + fit2_4$coefficient[3:4])
plot(PQLI_Score[subdata_index], IMR[subdata_index], type = "n", main = "Rural-Urban Difference\nleaving
text(PQLI_Score[subdata_index], IMR[subdata_index], UoR)
abline(fit2_5$coefficients[1:2])
abline(fit2_5$coefficients[1:2] + c(fit2_5$coefficient[3],0))
```



使用 anova 比較 reduced model 3 與 reduced model 1，得到 p-value > 0.05，代表兩模型無顯著差異，我們可以捨棄交互作用項選擇變數更少的 reduced model 3。總結來說，在印度 PQLI 越高，IMR 越低，並且此趨勢的城鄉差異據統計顯著性，性別差異不據統計顯著性。而城鄉差距主要造成的是截距不同，如果將 KERALA 邦視為 outlier，則斜率並不會隨著城市鄉村而有顯著不同。

Problem 3.

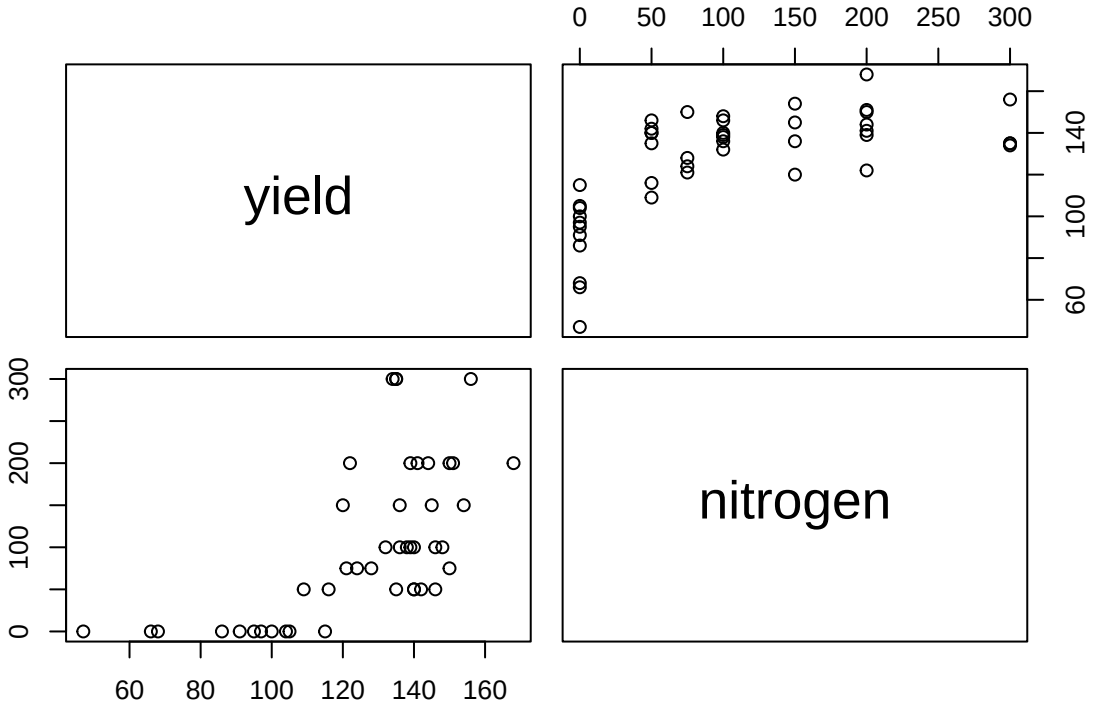
```
wd <- "http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/cornnit.txt"
data3 <- read.table(wd, header=T,fileEncoding = "UTF-8-BOM")
```

我們首先檢查資料有無 NA 值，並觀察資料趨勢。

```
summary(data3)
```

```
##      yield          nitrogen
##  Min.   : 47.0   Min.   :  0.0
##  1st Qu.:113.5   1st Qu.: 37.5
##  Median :135.0   Median : 87.5
##  Mean   :125.8   Mean   :103.4
##  3rd Qu.:142.5   3rd Qu.:162.5
##  Max.   :168.0   Max.   :300.0
```

```
pairs(data3)
```



再來對資料進行建模，以 **yield** 當作我們的 response variable，**nitrogen** 當我們的 predictor 建立一個簡單線性回歸模型。

$$\text{yield} = \beta_0 + \beta_1 \times \text{nitrogen} + \varepsilon.$$

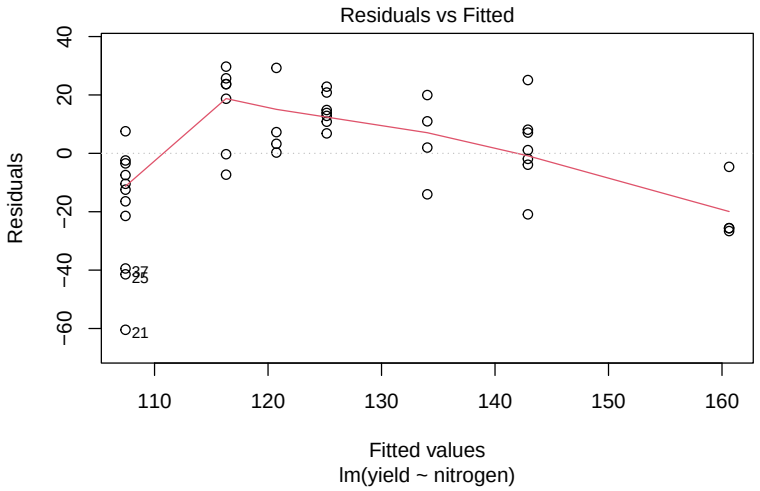
```
fit1 <- lm(yield ~ nitrogen, data = data3)
summary(fit1)
```

```
##
## Call:
## lm(formula = yield ~ nitrogen, data = data3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

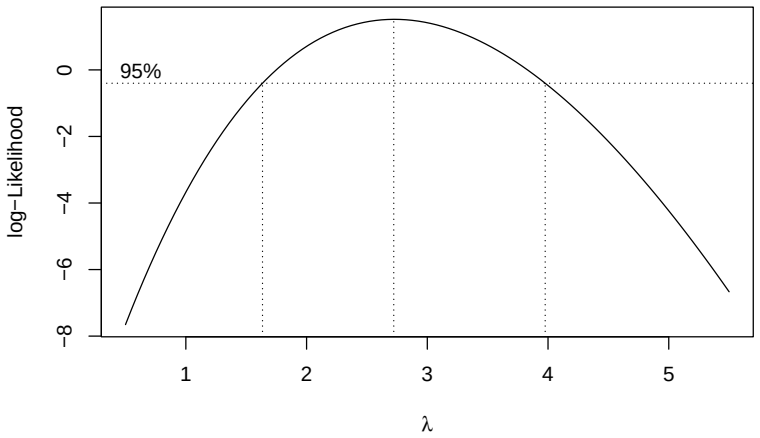
```
## -60.439 -10.939   1.534  14.082  29.697
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 107.43864    4.66622   23.02  < 2e-16 ***
## nitrogen      0.17730    0.03377    5.25 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.53 on 42 degrees of freedom
## Multiple R-squared:  0.3962, Adjusted R-squared:  0.3818
## F-statistic: 27.56 on 1 and 42 DF,  p-value: 4.713e-06
```

可以看到 **nitorgen** 係數顯著，然而看 residual plot 會發現我們並沒有抓到 mean structure 的 pattern，所以嘗試用 boxcox 對 response 進行轉換來解決這個問題。

```
plot(fit1, which = 1)
```




```
library(MASS)
bc <- boxcox(fit1, lambda=seq(0.5,5.5,by=0.1))
```




```
lambdahat=bc$x[which.max(bc$y)]
lambdahat
```

```
## [1] 2.722222
```

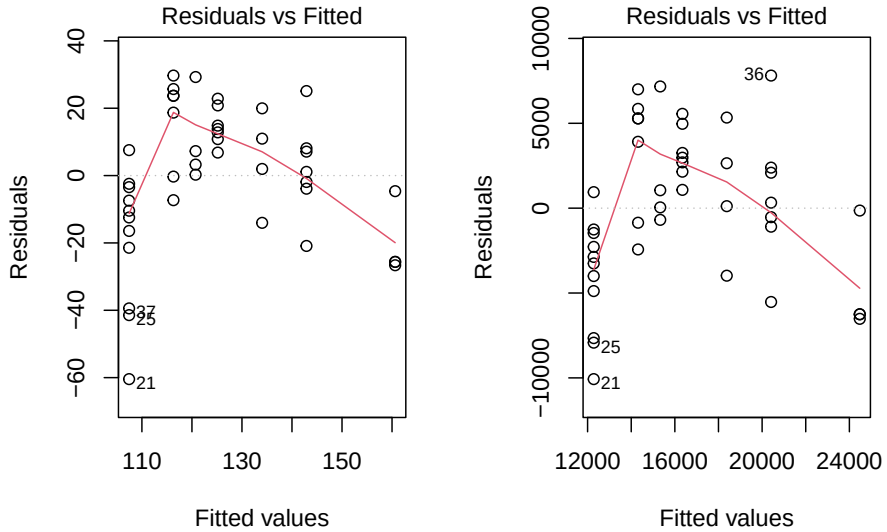
可以看到 $\hat{\lambda} = 2.722 \geq 2$，但由於 boxcox 轉換比較適用於 $\lambda \in (-2, 2)$，剛好圖中 $\hat{\lambda}$ 的 confidence interval

有包含到 2，所以我們可以考慮做 yield^2 的轉換。此外，由於 1 不包含在 $\hat{\lambda}$ 的 confidence interval，所以 response 做二次轉換的 fit2 優於 fit1。(注意：雖然 $R^2_{adj} = 0.3968$ 有所提升，但在兩模型 response 不同的情況下，兩模型的 R^2_{adj} 並不能直接比較。)

```
fit2 <- lm(I(yield^2) ~ nitrogen, data = data3)
summary(fit2)
```

```
##
## Call:
## lm(formula = I(yield^2) ~ nitrogen, data = data3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10072.1  -2968.1     81.4   3044.3   7812.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12281.067   1033.871  11.879 5.16e-15 ***
## nitrogen       40.653      7.483   5.433 2.58e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4548 on 42 degrees of freedom
## Multiple R-squared:  0.4127, Adjusted R-squared:  0.3988
## F-statistic: 29.52 on 1 and 42 DF,  p-value: 2.584e-06
```

```
par(mfrow = c(1, 2))
plot(fit1, which = 1)
plot(fit2, which = 1)
```



左上的圖是 yield ~ nitrogen 的 residual plot、右上的則是 yield^2 ~ nitrogen 的，雖然 $\hat{\lambda}$ 的 confidence interval 告訴我們轉換後的模型較佳，但從 residual plot 可以發現 response 二次轉換的 fit2 模型在描述 mean structure 仍沒有很好的表現，模型還有待改進，所以我們考慮對 predictor 進行 box-cox 轉換。首先，由於 **nitrogen** 裡有 0 的數據，所以我們會先對 **nitrogen** $+1$，**nitrogen** 原本的 range 是 $(0, 300)$，所以 $+1$ 並不會造成太大的影響。

由於我們知道 $x^\lambda \approx x + (\lambda - 1)\, x \log(x)$ (by Taylor's expansion)，所以我們對於我們的模型加入 $x \log(x)$ 項。

$$\text{yield} = \beta_0 + \beta_1 \times \text{nitrogen} + \beta_2 \times (\text{nitrogen} \times \log(\text{nitrogen})) + \varepsilon.$$

由於我們知道此模型是趨近而來的所以我們所想得到的 λ 為

$$\begin{aligned} & \beta^* x^\lambda \approx \beta^*[x + (\lambda - 1)\, x \log(x)] \\ \Rightarrow\ & \hat{\beta}_2 = \hat{\beta}^*(\lambda - 1) \\ \Rightarrow\ & \hat{\lambda} = \frac{\hat{\beta}_2}{\hat{\beta}^*} + 1, \quad \text{where} \ \ \hat{\beta}^* = \hat{\beta}_1 \end{aligned}$$

```
xlogx <- (data3$nitrogen + 1)*log(data3$nitrogen + 1)
fit3 <- lm(yield ~ I(nitrogen+1) + xlogx, data = data3)
summary(fit3)
```

```
##
## Call:
## lm(formula = yield ~ I(nitrogen + 1) + xlogx, data = data3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.275  -7.372  -0.546   9.741  24.725
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     88.30496    4.48708  19.680  < 2e-16 ***
## I(nitrogen + 1)  1.97042    0.28201   6.987 1.72e-08 ***
## xlogx           -0.31711    0.04969  -6.382 1.24e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.72 on 41 degrees of freedom
## Multiple R-squared:  0.6971, Adjusted R-squared:  0.6823
## F-statistic: 47.18 on 2 and 41 DF,  p-value: 2.325e-11
```

可以看到 $x\log(x)$ 係數十分顯著，再藉由上面的公式得到 $\hat{\lambda} = \frac{-0.31711}{1.97042} + 1 = 0.8390648$，並重新建模。

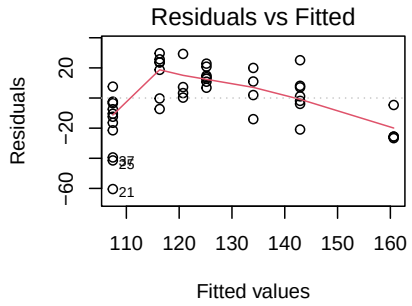
```
fit4 <- lm(yield ~ I(nitrogen^0.8390648), data = data3)
summary(fit4)
```

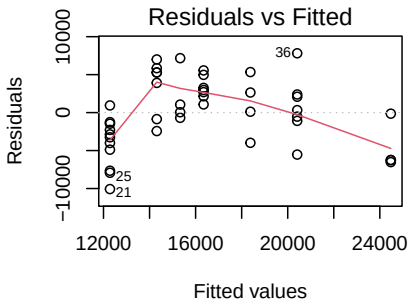
```
##
## Call:
## lm(formula = yield ~ I(nitrogen^0.8390648), data = data3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.243  -8.143   0.504  12.587  29.224
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           104.24330    4.66076  22.366  < 2e-16 ***
## I(nitrogen^0.8390648)   0.47043    0.07912   5.946 4.75e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

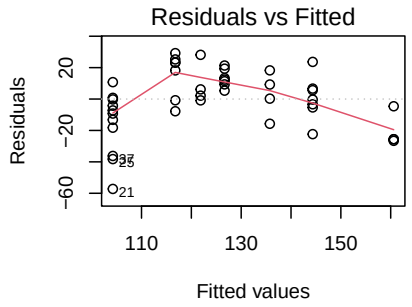
```
## Residual standard error: 19.47 on 42 degrees of freedom
## Multiple R-squared:  0.4571, Adjusted R-squared:  0.4441
## F-statistic: 35.36 on 1 and 42 DF,  p-value: 4.749e-07
```

可以發現 R_{adj} 的值從原本的 0.3818 上升到 0.4441，對於預測方面有提升，但是相反的就比較失去了解釋上的意義。再來比較一下 residual plot。

```
par(mfrow = c(2, 2), mar = c(4.5, 4, 2, 3))
plot(fit1, which = 1)
plot(fit2, which = 1)
plot(fit4, which = 1)
```

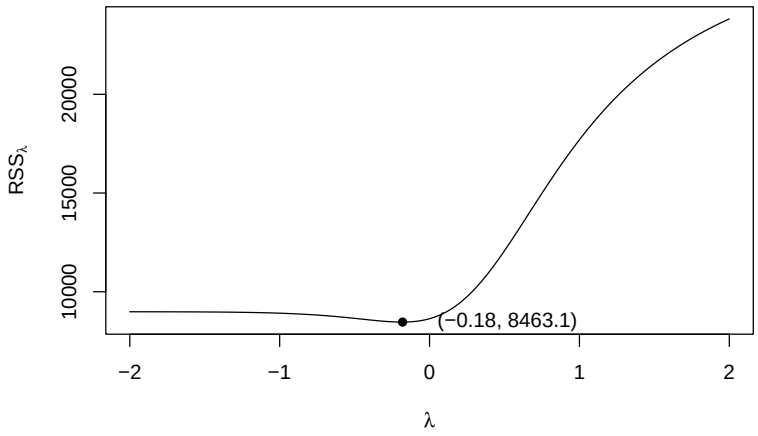






左上的圖是 yield ~ nitrogen 的 residual plot、右上的則是 yield^2 ~ nitrogen 的、左下的圖是 yield ~ $\text{nitrogen}^{0.8390648}$，可以看到左下的圖其實跟左上的沒有太大的差別。那這可能是因為我們在計算 box-cox transformation 所需的 $\hat{\lambda}$ 時，使用了近似。於是，接下來我們不使用近似求 $\hat{\lambda}$，而是直接代入-2 到 2 間的數值作為 $\hat{\lambda}$，確認 λ 為多少時可以達到 minimize RSS_λ 的目的。

```
lv <- seq(-2,2,0.01)
RSS <- numeric()
for(i in 1:length(lv)){
if(lv[i]==0) fit_l <- lm(yield ~ I(log(nitrogen+1)), data = data3)
else fit_l <- lm(yield ~ I(((nitrogen+1)^lv[i]-1)/lv[i]), data = data3)
RSS[i] <- sum(fit_l$residuals^2)
}
plot(lv,RSS,type="l",xlab=expression(lambda),ylab=expression(RSS[lambda]))
points(lv[which.min(RSS)],RSS[which.min(RSS)],pch=16)
text(lv[which.min(RSS)]+0.7,RSS[which.min(RSS)],
paste("(",round(lv[which.min(RSS)],2),", ",round(RSS[which.min(RSS)],2),")",sep=""))
```



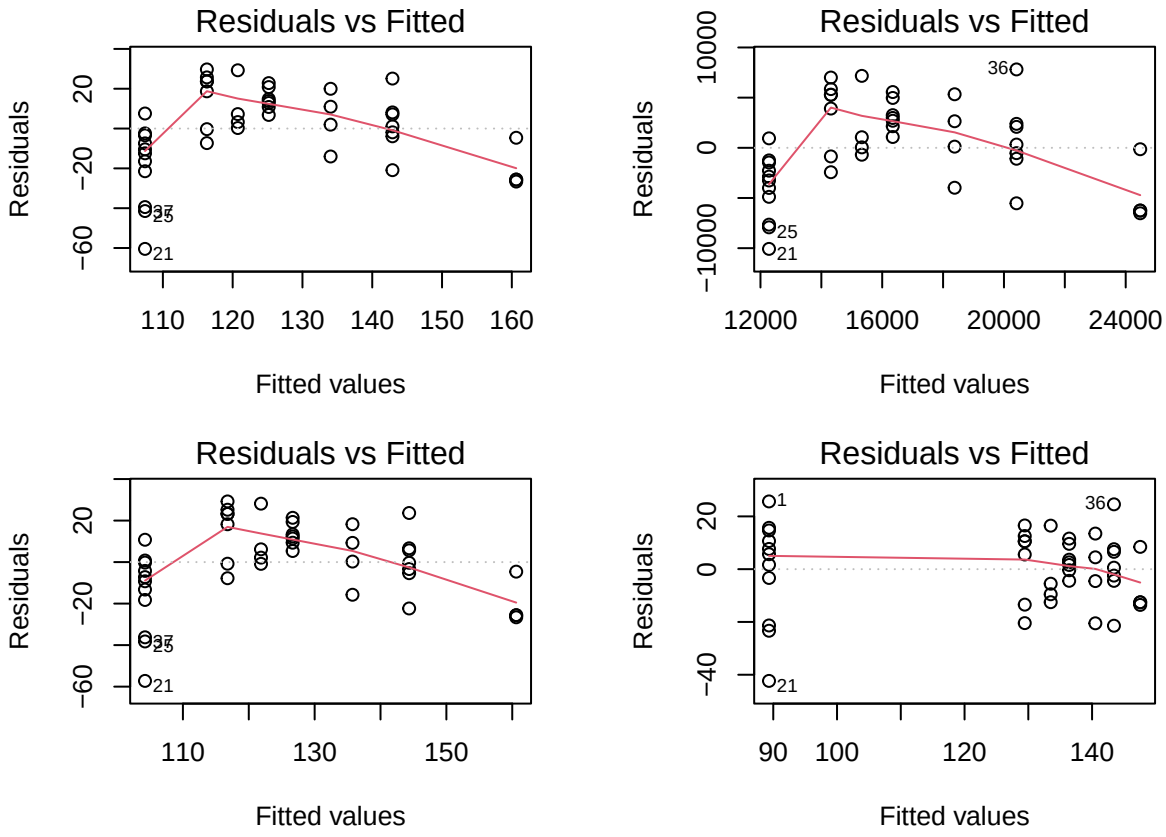
由於 $\hat{\lambda}$ 在 -0.18 有最小值，可考慮在更具有解釋意義的 $\hat{\lambda} = 0$ 對 predictor 轉換以提升模型解釋力。我們同樣檢查 residual plot。

$$\text{yield} = \beta_0 + \beta_1 \times \log(\text{nitrogen}) + \varepsilon.$$

```
fit5 <- lm(yield ~ log(nitrogen+1), data = data3)
summary(fit5)
```

```
##
## Call:
## lm(formula = yield ~ log(nitrogen + 1), data = data3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.335 -10.261   2.126  10.558  25.665
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         89.335      4.227   21.13  < 2e-16 ***
## log(nitrogen + 1)   10.201      1.017   10.03 1.03e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.34 on 42 degrees of freedom
## Multiple R-squared:  0.7055, Adjusted R-squared:  0.6985
## F-statistic: 100.6 on 1 and 42 DF,  p-value: 1.025e-12
```

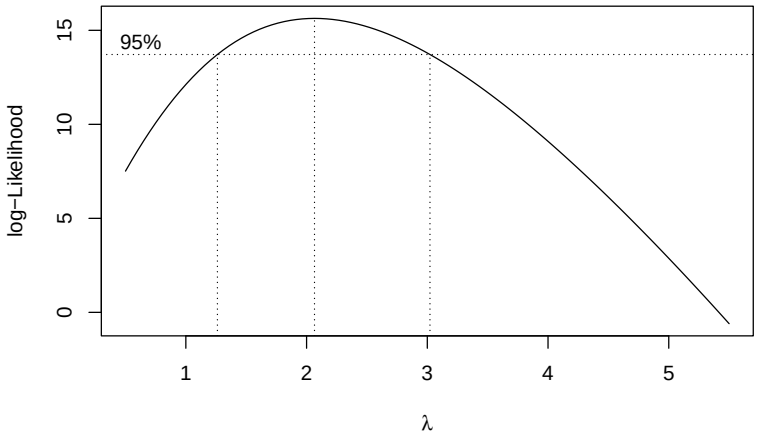
```
par(mfrow = c(2, 2), mar = c(4.5, 4, 2, 3))
plot(fit1, which = 1)
plot(fit2, which = 1)
plot(fit4, which = 1)
plot(fit5, which = 1)
```



可以看到新模型的 $R^2_{adf} = 0.6985$，有不錯的解釋力。比較不同模型的 residual plot，左上的圖是 yield $\sim$ nitrogen 的 residual plot、右上的則是 yield^2 $\sim$ nitrogen 的、左下的圖是 yield $\sim$ nitrogen$^{0.8390648}$、右下的圖是 yield $\sim$ *log*(nitrogen)。可以看到右下角 yield $\sim$ log(nitrogen) 模型對於 mean structure 的 pattern 掌握，比其他模型都好。

最後，我們可以再考慮模型 $y = \log(\text{nitrogen} + 1) + \varepsilon$ 下，是否還可以再對 y 進行 box-cox transformation。

```
bc2 <- boxcox(fit5, lambda=seq(0.5,5.5,by=0.1))
```



由於 $\hat{\lambda}$ 的 confidence interval 不包含 1，所以對我們應該要再對 response 作轉換。而由於 log-likelihood 最高點接近 2，可考慮在 predictoer 有做轉換下，同時也將 y 轉換成 y 的平方，配適模型。

$$\text{yield}^2 = \beta_0 + \beta_1 \times \log(\text{nitrogen}) + \varepsilon.$$

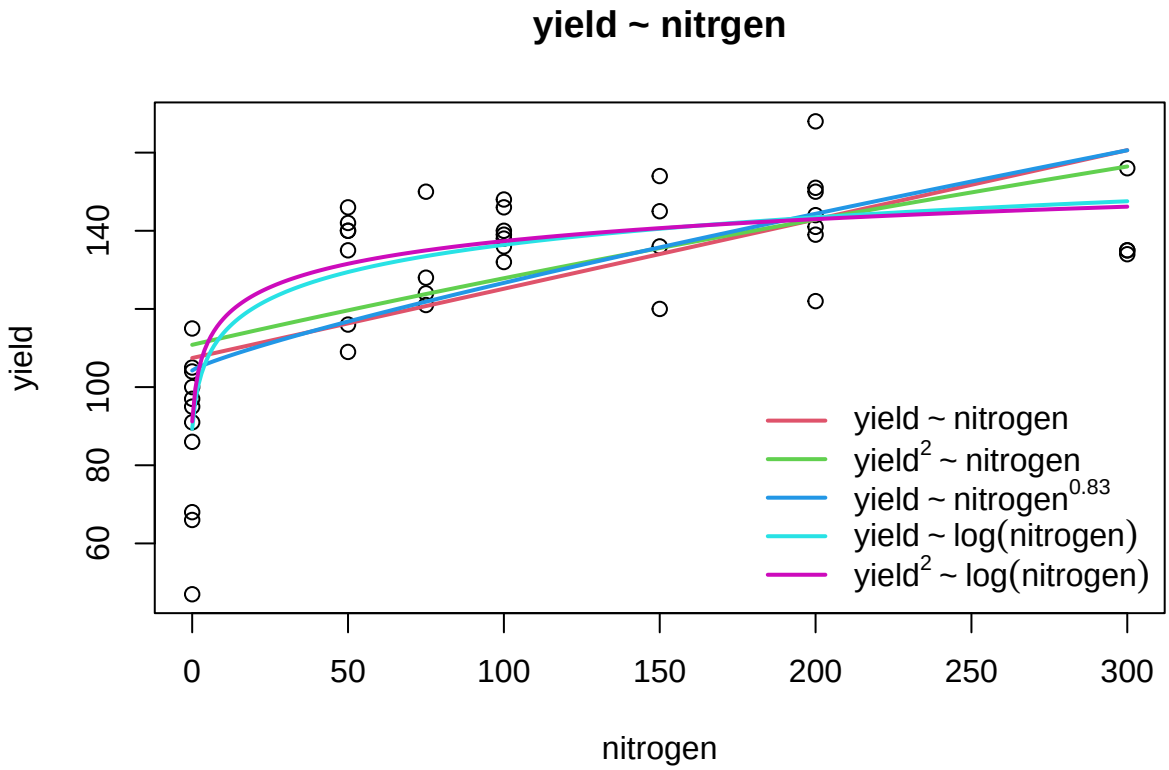
```
fit6 <- lm(I(yield^2) ~ log(nitrogen+1), data = data3)
summary(fit6)
```

```
##
## Call:
## lm(formula = I(yield^2) ~ log(nitrogen + 1), data = data3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6110.8 -2917.4   372.8  2380.8  7781.7
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         8319.8      955.4   8.708 5.89e-11 ***
## log(nitrogen + 1)   2285.8      229.9   9.944 1.33e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3241 on 42 degrees of freedom
## Multiple R-squared:  0.7019, Adjusted R-squared:  0.6948
## F-statistic: 98.89 on 1 and 42 DF,  p-value: 1.325e-12
```

將 fitted line 繪製於 scatter plot 中，我們可以看到淺藍色以及紫色的線明顯比其他好，他們兩個模型的解釋力也都在 0.6 以上。雖然 $\text{yield} \sim \log(\text{nitrogen})$ 和 $\text{yield}^2 \sim \log(\text{nitrogen})$ 不能直接透過 R^2_{adj} 比較，但從 box-cox 的結果可以得知又以 $\text{yield}^2 \sim \log(\text{nitrogen})$ 是更好的選擇。

```
xx <- 0:300
plot(data3$nitrogen, data3$yield, xlab = "nitrogen", ylab = "yield",
     main = "yield ~ nitrgen")
lines(xx, predict(fit1,newdata = data.frame(nitrogen=xx)), col = 2, lwd = 2)
lines(xx, sqrt(predict(fit2,newdata = data.frame(nitrogen=xx))), col = 3, lwd = 2)
lines(xx, predict(fit4,newdata = data.frame(nitrogen=xx)), col = 4, lwd = 2)
lines(xx, predict(fit5,newdata = data.frame(nitrogen=xx)), col = 5, lwd = 2)
lines(xx, sqrt(predict(fit6,newdata = data.frame(nitrogen=xx))), col = 6, lwd = 2)
legend("bottomright",
legend = c(expression(yield %~% nitrogen),
           expression(yield^2 %~% nitrogen),
           expression(yield %~% nitrogen^0.83),
           expression(yield %~% log(nitrogen)),
           expression(yield^2 %~% log(nitrogen))),

col=2:6, lty=1, lwd = 2,bty="n")
```



最後我們做 lack of fit test（goodness of fit test）比較一下 transform 過後的模型和最複雜的模型。

```
fit7 <- lm(I(yield^2) ~ factor(nitrogen), data = data3)
anova(fit6, fit7)
```

```
## Analysis of Variance Table
##
## Model 1: I(yield^2) ~ log(nitrogen + 1)
## Model 2: I(yield^2) ~ factor(nitrogen)
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1     42 441058609
## 2     37 416219682  5  24838928 0.4416 0.8165
```

從結果中可以看到 p-value 都大於 0.05，transform 過後的模型和最複雜的模型沒有顯著差異。總結來說，對 predictor 進行 log 轉換，並對 response 進行二次轉換的模型 $\text{yield}^2 \sim \log(\text{nitrogen})$ 是最適合的。並且它 goodness of fit test 也有過、解釋力也不錯的模型。