HW 7-Linear model

ID : 111024517　　　Name : 鄭家豪

due on 01/05

Q1

Read Data

```
data <- read.table("http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/aatemp.txt",
                   header = T)
head(data)
```

```
##   year  temp
## 1 1854 49.15
## 2 1855 46.52
## 3 1871 48.80
## 4 1881 47.95
## 5 1882 47.31
## 6 1883 44.64
```

i

建構 simple linear model:

$$\text{temp}_i = \beta_0 + \beta_1 \text{year}_i + \epsilon_i \quad , \text{where } \epsilon_i \sim N(0, \sigma^2)$$

```
fit_1 <- lm(temp~.,data=data)
summary(fit_1)
```

```
##
## Call:
## lm(formula = temp ~ ., data = data)
##
## 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
```

我們得到 fit model:

$$\text{temp}_i = 24.005510 + 0.012237\text{year}_i$$

接著加入 year 的二次項至 model:

```
fit_2 <- lm(temp ~ poly(year,2,raw=T),
            data = data)
summary(fit_2)
```

```
## 
## Call:
## lm(formula = temp ~ poly(year, 2, raw = T), data = data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -4.0412 -0.9538 -0.0624  0.9959  3.5820 
## 
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -2.127e+02  3.837e+02  -0.554    0.580
## poly(year, 2, raw = T)1  2.567e-01  3.962e-01   0.648    0.518
## poly(year, 2, raw = T)2 -6.307e-05  1.022e-04  -0.617    0.539
## 
## Residual standard error: 1.47 on 112 degrees of freedom
## Multiple R-squared:  0.08846,    Adjusted R-squared:  0.07218 
## F-statistic: 5.434 on 2 and 112 DF,  p-value: 0.005591
```

這會使得 year 一次項和二次項的係數皆不顯著，因此不考慮加入二次項。

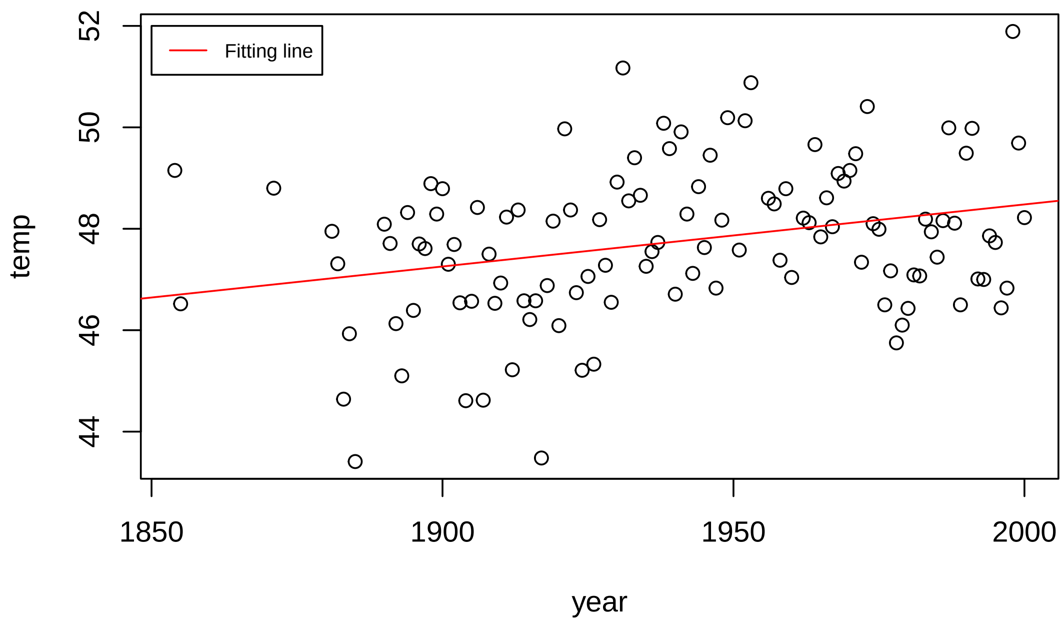
接著觀察一次項模型的 β_1 95% 信賴區間

```
confint(fit_1,level = 0.95)
```

```
##                   2.5 %      97.5 %
## (Intercept) 9.521535277 38.48948531
## year        0.004771599  0.01970293
```

其 95% 信賴區間: (0.004771599,0.01970293)

由於不包含 0 ，因此在顯著水準 0.05 下，有 linear trend 。



ii

使用 package: “nlme” ，進行 fit the model with correlated error following an AR(1) structure :

```
library(nlme)
fit_ar <- gls(temp~year,correlation = corAR1(form = ~year),
              data=data)
intervals(fit_ar,level = 0.95)
```

```
## Approximate 95% confidence intervals
##
##  Coefficients:
##                    lower         est.        upper
```

```
## (Intercept) 7.409192415 25.18407264 42.95895286
## year        0.002474401  0.01164028  0.02080617
##
##  Correlation structure:
##           lower      est.     upper
## Phi1 0.02920118 0.2303887 0.4136364
##
##  Residual standard error:
##    lower     est.    upper
## 1.284091 1.475718 1.695942
```

我們得到 the estimated correlation $\rho = 0.2303887$ ，且在顯著水準 0.05 下，拒絕 $\rho = 0$ 的假設。接著，在此模型下，year 的 95% 信賴區間不包含 0 ，因此不改變 trend 的看法。

iii

建立 year 十次多項式的模型:

```
fit_10 <- lm(temp~poly(year,10),data = data)
summary(fit_10)
```

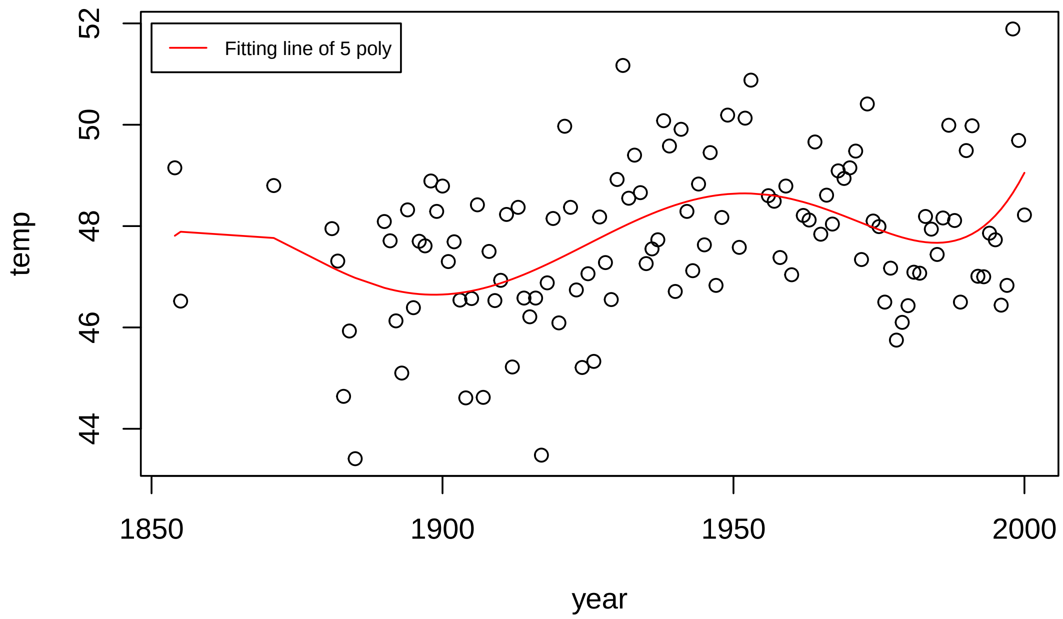
```
##
## Call:
## lm(formula = temp ~ poly(year, 10), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4987 -0.8641 -0.1745  1.1450  3.4255
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       47.7426     0.1319 361.927  < 2e-16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

```
## poly(year, 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
```

可以觀察到，第六項之後的變數皆不顯著，由於 orthogonality ，因此直接將第六項至第十項直接移除，保留前五項的變數再建構模型。

```
fit_5 <- lm(temp~poly(year,5),data=data)
summary(fit_5)
```

```
##
## Call:
## lm(formula = temp ~ poly(year, 5), data = data)
##
## 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, 5)1   4.7616     1.4004   3.400 0.000942 ***
## poly(year, 5)2  -0.9071     1.4004  -0.648 0.518500
## poly(year, 5)3  -3.3132     1.4004  -2.366 0.019749 *
## poly(year, 5)4   2.4383     1.4004   1.741 0.084470 .
## poly(year, 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
```



接著預測 2020 年的 temperature:

```
predict(fit_5,data.frame(year=2020),
                          se=TRUE,
                          interval = "prediction")
```

```
## $fit
##        fit      lwr      upr
## 1 60.07774 49.84092 70.31456
##
## $se.fit
## [1] 4.971514
##
## $df
## [1] 109
##
## $residual.scale
## [1] 1.400373
```

我們得到預測值: 60.07774 ，和 95% 預測區間: (49.84092,70.31456) 。

iv

Define the base function:

$$d(\text{year}) = \begin{cases} 1 & \text{if year} > 1930 \\ 0 & \text{if otherwise} \end{cases}$$

Broken line regression(No continuity)

The Model:

$$\text{temp}_i = \beta_0 + \beta_1 d(\text{year}_i) + \beta_2 \text{year}_i + \beta_3 (\text{year}_i - 1930) d(\text{year}_i) + \epsilon_i \quad \text{, where } \epsilon_i \sim N(0, \sigma^2)$$

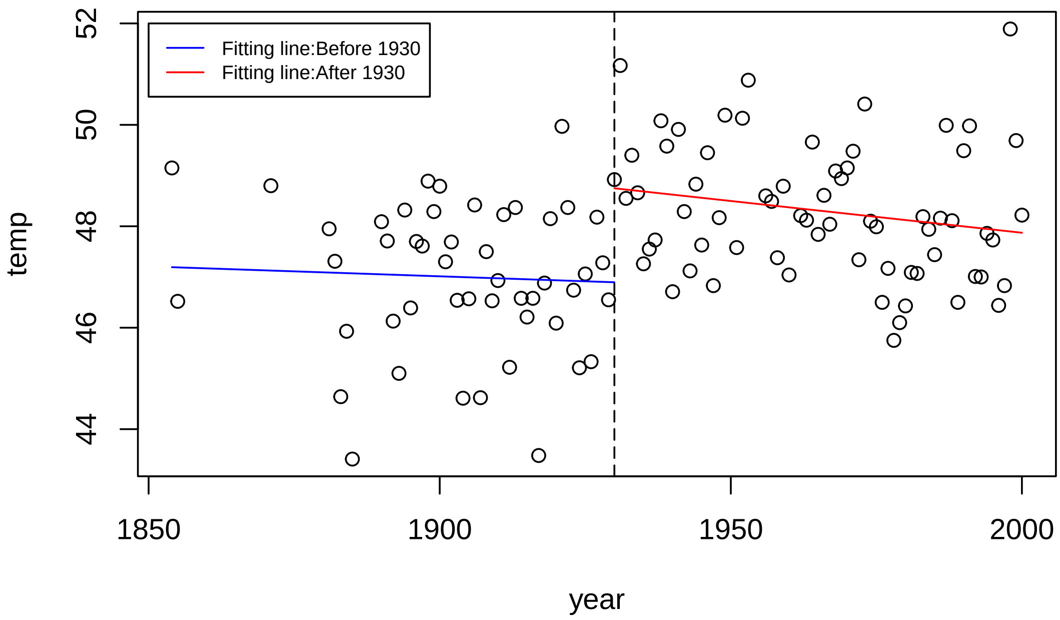
```r
d <- function(x){ifelse(x>1930,1,0)}
model_broken1 <- lm(temp~ d(year) + year + I((year - 1930) * d(year)),
                    data=data)
summary(model_broken1)
```

```
##
## Call:
## lm(formula = temp ~ d(year) + year + I((year - 1930) * d(year)),
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6618 -0.9456 -0.0876  0.9908  3.9925
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                54.452092  21.267390   2.560 0.011800 *
## d(year)                     1.853081   0.490983   3.774 0.000259 ***
## year                       -0.003915   0.011168  -0.351 0.726576
## I((year - 1930) * d(year)) -0.008603   0.013903  -0.619 0.537319
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.387 on 111 degrees of freedom
## Multiple R-squared:  0.1963, Adjusted R-squared:  0.1745
## F-statistic: 9.035 on 3 and 111 DF,  p-value: 2.102e-05
```

得到模型:

$$\text{temp}_i = 54.452092 + 1.853081 d(\text{year}_i) - 0.003915 \text{year}_i - 0.008603 (\text{year}_i - 1930) d(\text{year}_i)$$

由以上模型，繪製出其 fitting line：



觀察這張圖來判斷 Claim 是否合理，由圖和模型斜率係數的顯著性來看，1930 年之後的斜率變化似乎沒有很顯著，因此這個 Claim 似乎不正確。

Broken line regression(continuity)

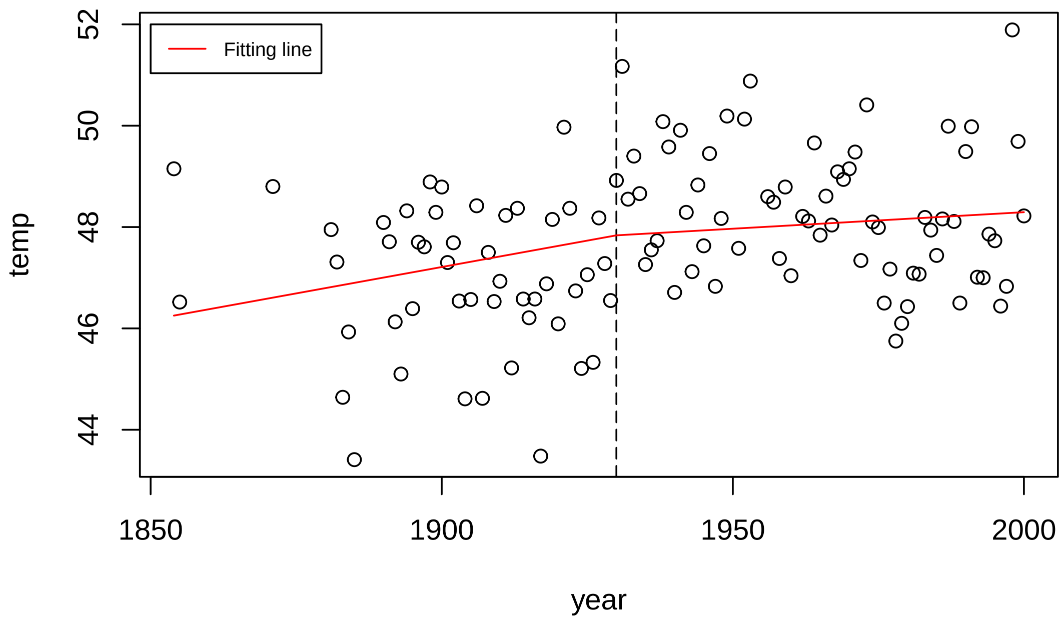
為了在 year = 1930 時連續，模型修改為:

$$\text{temp}_i = \beta_0 + \beta_1 \text{year}_i + \beta_2(\text{year}_i - 1930)d(\text{year}_i) + \epsilon_i \quad , \text{where } \epsilon_i \sim N(0, \sigma^2)$$

```
model_broken2 <- lm(temp~ year + I((year - 1930) * d(year)),
                    data=data)
summary(model_broken2)
```

```
##
## Call:
## lm(formula = temp ~ year + I((year - 1930) * d(year)), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0855 -0.9492 -0.0380  1.0289  3.6096
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  7.619376  18.264975   0.417   0.6774
```

```
## year                          0.020838   0.009559   2.180   0.0314 *
## I((year - 1930) * d(year)) -0.014308   0.014615  -0.979   0.3297
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.467 on 112 degrees of freedom
## Multiple R-squared:  0.09312,    Adjusted R-squared:  0.07693
## F-statistic:  5.75 on 2 and 112 DF,  p-value: 0.004195
```



由這張圖和模型 summary 來看，由於 (year - 1930) * d(year) 項的係數不顯著，因此不太能接受這個 Claim 是正確的。

v

根據 LNp.8-8 的規則，選取 6+4 個 knots:

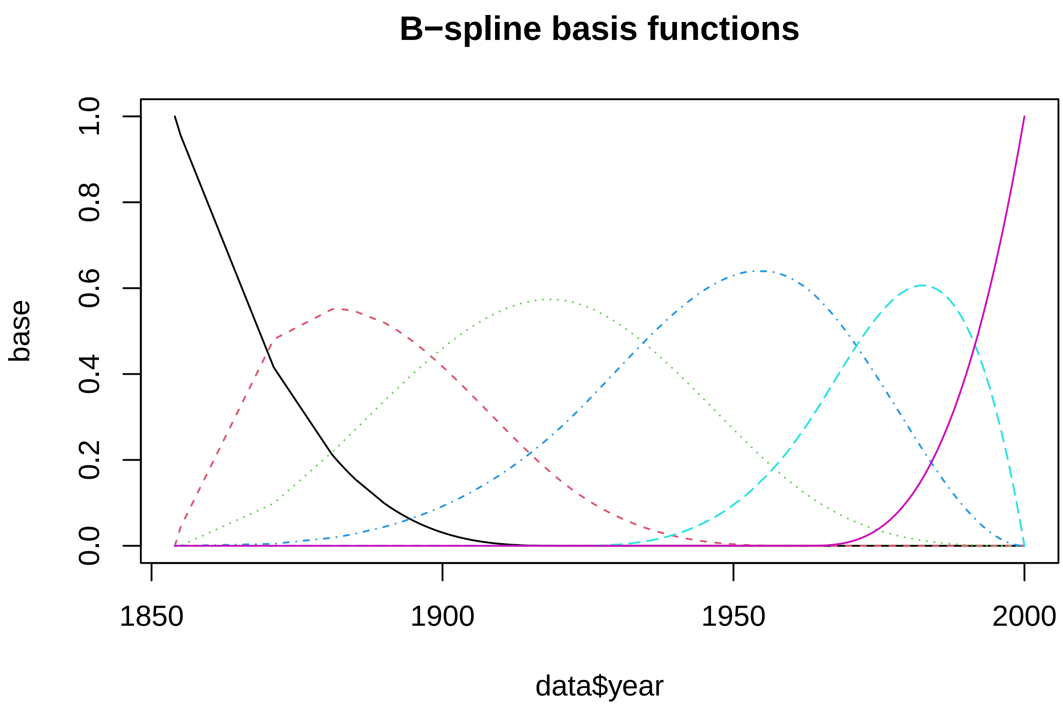
```
knots <- c(1854,1854,1854,1854,1921,1962,2000,2000,2000,2000)
```

接著使用 package: "splines"，進行 cubic spline fit:

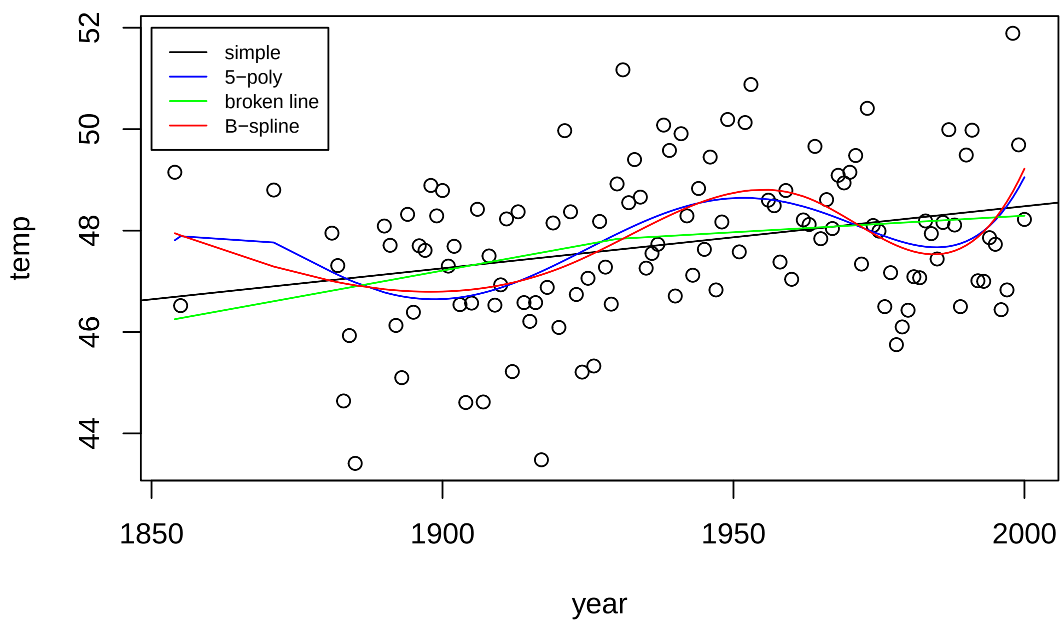
```
library(splines)
base <- splineDesign(knots,data$year)
model_cubic <- lm(temp~base,data=data)
summary(model_cubic)
```

```
##
## Call:
## lm(formula = temp ~ base, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6499 -0.9081 -0.2034  0.9433  3.3305
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  49.2196     0.6813  72.240  < 2e-16 ***
## base1        -1.2715     1.2120  -1.049  0.29646
## base2        -2.2249     1.1449  -1.943  0.05457 .
## base3        -3.4016     1.2520  -2.717  0.00767 **
## base4         1.1949     0.8534   1.400  0.16433
## base5        -3.1265     1.2629  -2.476  0.01484 *
## base6             NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.392 on 109 degrees of freedom
## Multiple R-squared:  0.2044, Adjusted R-squared:  0.1679
## F-statistic: 5.601 on 5 and 109 DF,  p-value: 0.0001242
```

$R^2 = 0.2044$，比較 i 的模型 $R^2 = 0.08536$，B-spline 會比 simple linear model 還好些。



- Plot the fit in comparison to the previous fits:



將前面所得到的 fitting line(除了 10-poly. 和 unconutious broken regression)，都繪至在同一張圖上，可發現 5-polynomial model 和 Cubic B-spline model 的表現會其他模型還要好，其中這意味著溫度會隨著時間變化而有所變化。

Q2

Read data

在進行讀取前，需要對資料做調整方便讀取:

State	PQLI Score	Combined IMR	Rural Male IMR	Rural Female IMR	Urban Male IMR	Urban Female IMR
UTTAR PRAD.	17	167	159	187	110	111
MADHYA PRAD.	28	135	148	134	88	83
ORISSA	24	133	131	142	78	81
RAJASTHAN	29	129	135	142	55	77
GUJARAT	36	118	120	135	92	84
ANDHRA_PRAD.	33	112	138	101	79	46
HARYANA	55	109	107	128	57	60
ASSAM	35	118	133	106	87	85
PUNJAB	62	103	115	108	58	73
TAMILNADU	43	103	125	115	67	59
KARNATAKA	52	75	92	70	51	59
MAHARASHTRA	60	75	95	72	50	62
KERALA	92	39	42	42	22	30

$\Rightarrow$

State	PQLI Score	Combined IMR	Rural Male IMR	Rural Female IMR	Urban Male IMR	Urban Female IMR
UTTAR_PRAD.	17	167	159	187	110	111
MADHYA_PRAD.	28	135	148	134	88	83
ORISSA	24	133	131	142	78	81
RAJASTHAN	29	129	135	142	55	77
GUJARAT	36	118	120	135	92	84
ANDHRA_PRAD.	33	112	138	101	79	46
HARYANA	55	109	107	128	57	60
ASSAM	35	118	133	106	87	85
PUNJAB	62	103	115	108	58	73
TAMILNADU	43	103	125	115	67	59
KARNATAKA	52	75	92	70	51	59
MAHARASHTRA	60	75	95	72	50	62
KERALA	92	39	42	42	22	30

```
data <- read.table("E1.20.txt",skip =3)
colnames(data) <- c("state","PQLI","Comb.IMR",
                    "Rur.M.IMR","Rur.F.IMR",
                    "Urb.M.IMR","Urb.F.IMR")


head(data)
```

```
##          state PQLI Comb.IMR Rur.M.IMR Rur.F.IMR Urb.M.IMR Urb.F.IMR
## 1  UTTAR_PRAD.   17      167       159       187       110       111
## 2 MADHYA_PRAD.   28      135       148       134        88        83
## 3       ORISSA   24      133       131       142        78        81
## 4    RAJASTHAN   29      129       135       142        55        77
## 5      GUJARAT   36      118       120       135        92        84
## 6 ANDHRA_PRAD.   33      112       138       101        79        46
```

我們的目的是想研究 IMR 與性別和區域的關係，由於原始資料是將性別區域 IMR 合併在一起，分成 “Rur.M.IMR” 、“Rur.F.IMR” 、“Urb.M.IMR” 、“Urb.F.IMR” ，為了便於分析，定義兩個 dummy variable 來拆成三個 column ，再加入對應的 PQLI，整合成新的資料 (名稱為”data_combin”) :

$$d_1(\text{gender}) = \begin{cases} 1 & \text{if gender is Male} \\ 0 & \text{if gender is Female} \end{cases}, \quad d_2(\text{Area}) = \begin{cases} 1 & \text{if area is urban} \\ 0 & \text{if area is rural} \end{cases}$$

```
data_combin <- data.frame("MIR" = c(data$Rur.M.IMR,data$Rur.F.IMR,data$Urb.M.IMR,data$Urb.F.IMR),
                    "Gender" = c(rep("Male",13),rep("Female",13),rep("Male",13),rep("Female",13)),
                    "Area" = c(rep("Rural",26),rep("urban",26)),
                    "PQLI"= rep(data$PQLI,4))
dim(data_combin)
```

```
## [1] 52  4
```

```r
names(data_combin)
```

```
## [1] "MIR"    "Gender" "Area"   "PQLI"
```

利用這資料來做分析:

ANCOVA

```r
fit1 <- lm(MIR ~  Gender + Area ,data = data_combin)
fit2 <- lm(MIR ~  Gender + Area + PQLI,data = data_combin)
anova(fit1,fit2)
```

```
## Analysis of Variance Table
##
## Model 1: MIR ~ Gender + Area
## Model 2: MIR ~ Gender + Area + PQLI
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     49 39485
## 2     48 12241  1     27244 106.83 8.497e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obviously,as p-value is small than 0.05 ,the quantiative predictor PQLI is covariate.

Fit model

配適一個 MIR ~ Gender * Area * PQLI 的模型:

```r
fit <- lm(MIR ~ Gender*Area*PQLI,data=data_combin)
summary(fit)
```

```
##
## Call:
## lm(formula = MIR ~ Gender * Area * PQLI, data = data_combin)
##
## 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 ***
## GenderMale                -2.4329    14.6257  -0.166   0.8687    
## Areaurban                -77.9537    14.6257  -5.330 3.22e-06 ***
## PQLI                      -1.5494     0.2168  -7.147 6.96e-09 ***
## GenderMale:Areaurban      10.5945    20.6838   0.512   0.6111    
## GenderMale:PQLI            0.1584     0.3066   0.517   0.6081    
## Areaurban:PQLI             0.7799     0.3066   2.544   0.0146 *  
## GenderMale:Areaurban:PQLI -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
```

此時應該要考慮 PQLI 和 Gender 或 Area 之間是否有交互作用效應，使用 anova() 指令觀察:

```
anova(fit)
```

```
## Analysis of Variance Table
## 
## Response: MIR
##                   Df  Sum Sq Mean Sq  F value    Pr(>F)    
## Gender             1    33.9    33.9   0.1460  0.704204    
## Area               1 28529.3 28529.3 122.8063 2.560e-14 ***
## PQLI               1 27243.6 27243.6 117.2720 5.409e-14 ***
## Gender:Area        1   105.3   105.3   0.4533  0.504291    
## Gender:PQLI        1     4.1     4.1   0.0175  0.895326    
## Area:PQLI          1  1737.2  1737.2   7.4779  0.008967 ** 
## Gender:Area:PQLI   1   172.9   172.9   0.7444  0.392935    
## Residuals         44 10221.7   232.3                       
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

由上面結果，只有 Area 和 PQLI 的交互作用項顯著，另外 Gender 項不顯著，於是移除 Gender 項和加入 Area:PQLI 交互作用項然後重配模型:

```
fit_new <- lm(MIR ~ Area + PQLI + Area:PQLI,data=data_combin)
summary(fit_new)
```

```
## 
```

```
## Call:
## lm(formula = MIR ~ Area + PQLI + Area:PQLI, data = data_combin)
##
## 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 ***
## Areaurban        -72.6564    10.0536  -7.227 3.30e-09 ***
## PQLI              -1.4702     0.1490  -9.866 3.93e-13 ***
## Areaurban:PQLI     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
```

我們得到模型:

$$\text{MIR} = 180.2417 - 72.6564 \times d_2(\text{area}) - 1.4702 \times \text{PQLI} + 0.5928 \times (d_2(\text{area}) \times \text{PQLI})$$

我們可以看到，每個項的係數皆是顯著，there exist rural-urban difference in mortality after adjusting for the covariate,PQLI.

Q3

Read data and fit simple linear model

```r
data <- read.table("http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/cornnit.txt",
                   header = T)
head(data)
```

```
##   yield nitrogen
## 1   115        0
## 2   128       75
## 3   136      150
## 4   135      300
```

```
## 5    97         0
## 6   150        75
```

```
fit <- lm(yield~.,data=data)
summary(fit)
```

```
##
## Call:
## lm(formula = yield ~ ., data = data)
##
## 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
```

We have model:

$$\text{yield}_i = 107.43864 + 0.17730(\text{nitrogen}_i)$$

Testing for Lack of fit

用 anova() 指令進行 Testing for Lack of fit:

```
fit_sature <- lm(yield ~ factor(nitrogen),data=data)
anova(fit,fit_sature)
```

```
## Analysis of Variance Table
##
## Model 1: yield ~ nitrogen
## Model 2: yield ~ factor(nitrogen)
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
```

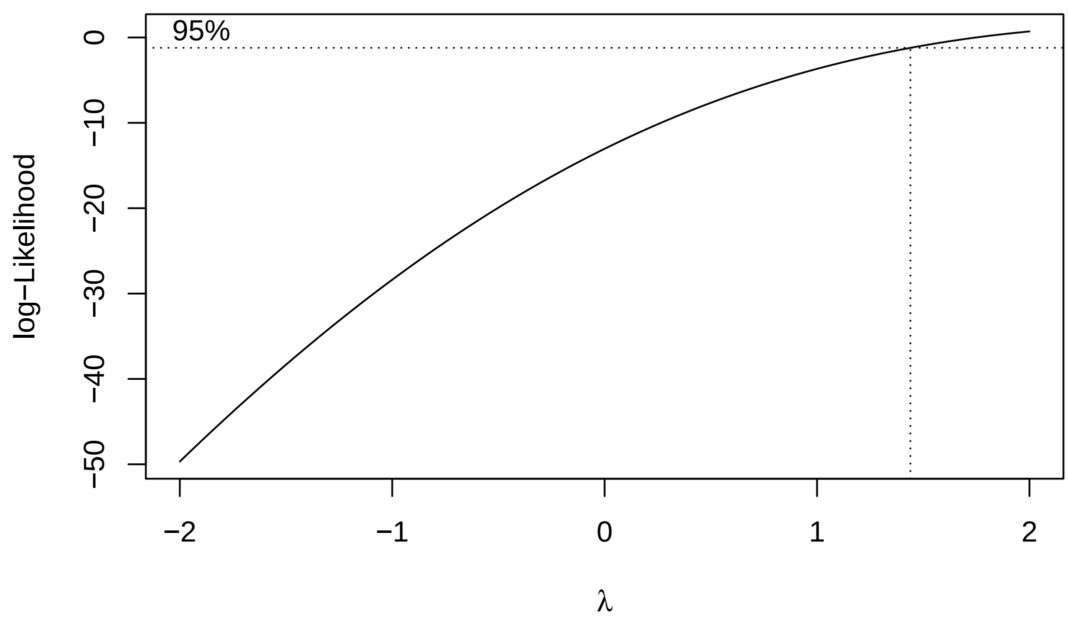
```
## 1     42 17699.2
## 2     37  8186.8  5    9512.4 8.5982 1.774e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

由於 p-value <0.05 ，有足夠證據表示這個模型配得不好。

Box-Cox method

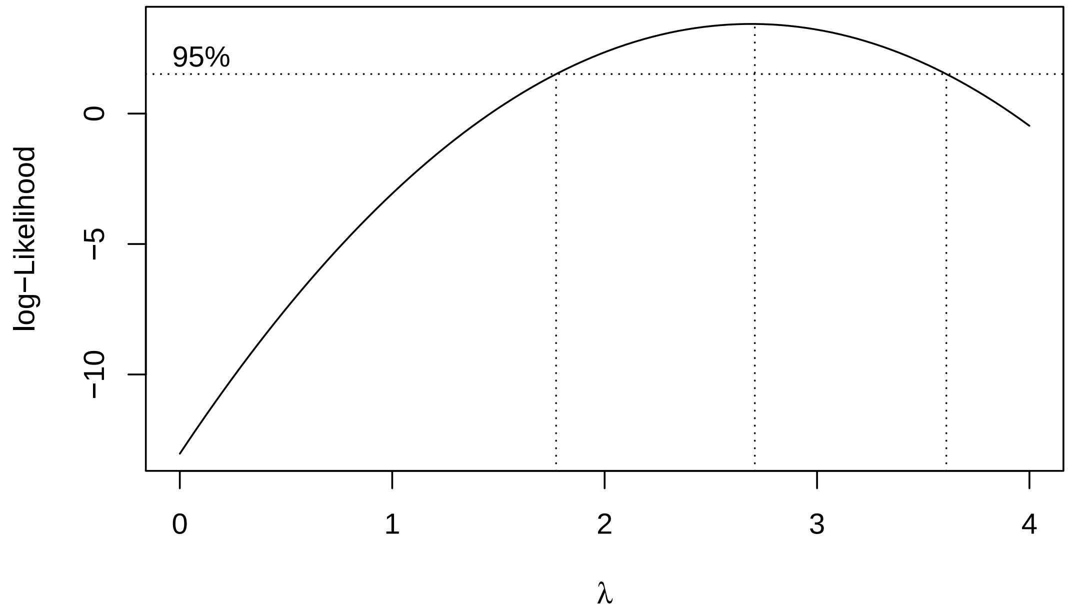
這裡檢查是否適合使用 Box-Cox method for response，Using package：“MASS”。

```
library(MASS)
boxcox(fit,plotit = T)
```



由於 lambda 值似乎超過 1 之後，log-likelihood 還在增加，我們試著把 lambda 的範圍往後拉一點:

```
boxcox(fit,plotit = T,lambda = c(0,4,1/100))
```



可以發現當 $\lambda \in (2,3)$ 時，其 log-likelihood 會達至最大。

我們試著對 response 做 $(y^3-1)/3$ 的 transformation (比較有解釋性且 3 比較靠近最大 log-likelihood 的 λ)，然後 fit model:

```
g1 <- lm(I(yield^3) ~ nitrogen,data=data)
summary(g1)
```

```
##
## Call:
## lm(formula = I(yield^3) ~ nitrogen, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1365038  -570012     3471   525741  1817216
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1468861     186746   7.866 8.63e-10 ***
## nitrogen        7278       1352   5.385 3.03e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 821600 on 42 degrees of freedom
## Multiple R-squared:  0.4084, Adjusted R-squared:  0.3943
```

```
## F-statistic: 28.99 on 1 and 42 DF,  p-value: 3.029e-06
```

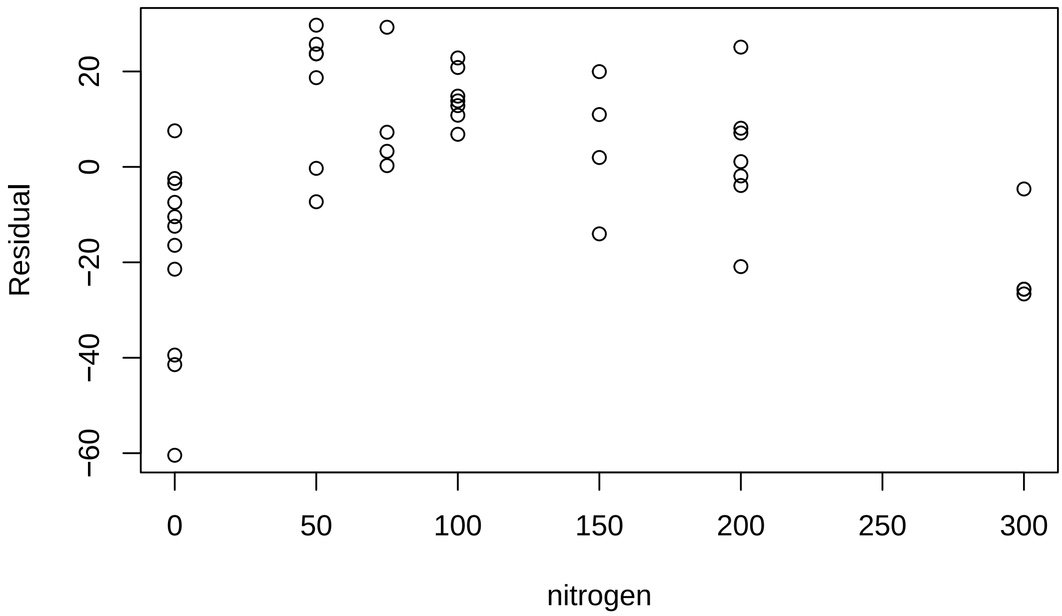
```r
821600^(1/3)
```

```
## [1] 93.65985
```

其 $\hat{\sigma} = 821600$，由於單位是 response 單位的 3 次方，算回去原本的單位後，得到 $93.65985 > 20.53$($\hat{\sigma}$ from the model without transformation)。然後 $R^2 = 0.4084$ 相較於沒轉換後的 $R^2 = 0.3962$ ，差不了多少，因此對 response 做 Box-Cox transformation 對於模型沒有改善。

我們來檢驗是否要對 predictor 做 transformation:

先觀察 nitrogen-residual 之間的關係:



由以上的圖，可以觀察出 nitrogen 和 residual 似乎有”凹口向下”的曲線關係 (second derivative is small than 0)，因此試著對 nitrogen 做 Box-Cox transformation(x^λ,$\lambda \in (0, 1)$)。

定義:

$$xlog(x) = \begin{cases} xlog(x) & \text{if x} > 0 \\ 0 & \text{if otherwise} \end{cases}$$

這樣定義的目的是為了使 nitrogen $= 0$ 時有意義 $(0^\lambda = 0)$

接著建構 model:

$$\text{yield}_i = \beta_0 + \beta_1(\text{nitrogen} + (\lambda - 1)\text{nitrogen} \times \log(\text{nitrogen})) + \epsilon_i \quad \text{, where } \epsilon_i \sim N(0, \sigma^2)$$

```r
f = function(x){
  c=c()
  for (i in 1:length(x)){
    if(x[i] == 0){c[i]=0}
    else{c[i]=x[i]*log(x[i])}
  }
  return(c)
}
g2 <- lm(yield ~ nitrogen + f(data$nitrogen),data=data)
summary(g2)
```

```
##
## Call:
## lm(formula = yield ~ nitrogen + f(data$nitrogen), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.159  -7.262  -0.471   9.597  24.841
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      90.15890    4.28892  21.021  < 2e-16 ***
## nitrogen          1.90973    0.27122   7.041 1.44e-08 ***
## f(data$nitrogen) -0.30757    0.04796  -6.413 1.12e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.68 on 41 degrees of freedom
## Multiple R-squared:  0.6986, Adjusted R-squared:  0.6839
## F-statistic: 47.51 on 2 and 41 DF,  p-value: 2.104e-11
```

由於 $x_i log(x_i)$ 項的係數顯著不為 0 ，因此對 nitrogen 做轉換 ($\lambda = \frac{-0.30757}{1.90973} + 1 = 0.8389458$): 轉換完後重新 fit:

```r
fit_trans <- lm(yield ~ I(nitrogen^0.8389458),data = data)
summary(fit_trans)
```

```
##
## Call:
## lm(formula = yield ~ I(nitrogen^0.8389458), data = data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -57.241  -8.142   0.505  12.586  29.224 
## 
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             104.24072    4.66074  22.366  < 2e-16 ***
## I(nitrogen^0.8389458)   0.47076    0.07917   5.947 4.74e-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:  0.4442 
## F-statistic: 35.36 on 1 and 42 DF,  p-value: 4.74e-07
```

轉換完後，$R^2 = 0.4571$ and $\hat{\sigma} = 19.47$ ，與沒轉換的模型做比較，其 $R^2 = 0.3962$ and $\hat{\sigma} = 20.53$ ，模型有得到改善。不過缺點是犧牲兩者變數之間的解釋性。

```r
anova(fit_trans,fit_sature)
```

```
## Analysis of Variance Table
## 
## Model 1: yield ~ I(nitrogen^0.8389458)
## Model 2: yield ~ factor(nitrogen)
##   Res.Df     RSS Df Sum of Sq     F    Pr(>F)    
## 1     42 15914.4                                 
## 2     37  8186.8  5    7727.6 6.985 0.0001105 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

雖然還是有 lack of fit ，但 p-value 相比沒轉換前，有增加很多，這意味著對 predictor 轉換後確實有得到些許改善。

我們最後用 shapiro.test() 來檢定對 predictor 轉換前和轉換後的模型哪個比較符合 normality:

```
## 
##  Shapiro-Wilk normality test
## 
## data:  rstandard(fit)
## W = 0.95056, p-value = 0.05772
```

```
## 
```

```
##  Shapiro-Wilk normality test
##
## data:  rstandard(fit_trans)
## W = 0.95239, p-value = 0.06754
```

若顯著水準 $= 0.05$ ，那這兩個模型都符合 normality，但因為轉換後的模型，p-value 較轉換前的模型還要大，所以轉換後比轉換前更加符合 normality。