Regression analysis 期末專題計畫書

吳翔宇

一、研究動機

隨著科技發展，醫療衛生的進步，人的預期壽命不斷向上提高，但隨著科技進步貧富差距也隨之擴大。這狀況也反應在預期壽命上，有的人能享有世界上最好醫療資源，但也有生個小病都能致命的狀況發生。因應各國不同的醫療政策及經濟環境，不同國家人民的預期壽命也可能天差地遠。

二、研究目的

本次研究目的:

- 探討開發中國家與已開發國家預期壽命是否有顯著差異
- 影響預期壽命的因子與其重要程度

三、資料簡介與篩選

1. 資料來源

本次使用資料為 Kaggle:Life Expectancy (WHO)

2. 資料背景

世界衛生組織 (WHO) 所屬的全球衛生觀察站 (GHO) 在 2015 時發現，過去 15 年由於全球衛生水準迅速提高，與過去 30 年相比，人類死亡路有所提高，在開發中國家尤其明顯，因此在 WHO 的開放資料中收集了 2000 年至 2015 年中 193 個國家的預期壽命與健康因素相關的其他數據，希望藉由此數據來探討影響人們預期壽命的重要因素。

3. 變數

```
data <- read.csv("LifeExpectancyData.csv")[,-c(1,15,18:21)]
colnames(data) <- c("year","status","life","am","ind","alcohol","perexp","HB","measles",
                    "bmi","ufd","polio","texp","hiv","gdp","school")
data$status = ifelse(data$status == "Developing",0,1)
data$status = factor(data$status)
```

變數	定義
year	年份
status	0: 開發中國家 (Developing) 1: 已開發國家 (Developed)
Life	預期壽命
am	15 至 60 歲人民，每 1000 人死亡人數
ind	低於 1 歲之嬰兒，每 1000 人死亡人數
alcohol	15 歲以上人民人均酒精消費量。
perexp	醫療支出佔人均 GDP 百分比

變數	定義
HB	1 歲孩童 B 型肝炎疫苗覆蓋率
measles	麻疹病例數
bmi	體重 (kg) 除身高 (m) 的平方 $\frac{kg}{m^2}$
ufd	5 歲以下，每 1000 人死亡人數
polio	1 歲一下小兒麻痺症疫苗覆蓋率
texp	政府醫療支出佔總開支百分比
hiv	0 至 4 歲孩童，每 1000 人死於愛滋病毒/愛滋病人數
gdp	人均 GDP(美元)
school	受教育年數 (年)

```
str(data)
```

```
## 'data.frame':    2938 obs. of  16 variables:
##  $ year   : int  2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...
##  $ status : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ life   : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
##  $ am     : int  263 271 268 272 275 279 281 287 295 295 ...
##  $ ind    : int  62 64 66 69 71 74 77 80 82 84 ...
##  $ alcohol: num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...
##  $ perexp : num  71.3 73.5 73.2 78.2 7.1 ...
##  $ HB     : int  65 62 64 67 68 66 63 64 63 64 ...
##  $ measles: int  1154 492 430 2787 3013 1989 2861 1599 1141 1990 ...
##  $ bmi    : num  19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...
##  $ ufd    : int  83 86 89 93 97 102 106 110 113 116 ...
##  $ polio  : int  6 58 62 67 68 66 63 64 63 58 ...
##  $ texp   : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...
##  $ hiv    : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
##  $ gdp    : num  584.3 612.7 631.7 670 63.5 ...
##  $ school : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
```

```
summary(data)
```

```
##       year      status        life             am           ind        
##  Min.   :2000   0:2426   Min.   :36.3   Min.   :  1   Min.   :   0.0  
##  1st Qu.:2004   1: 512   1st Qu.:63.1   1st Qu.: 74   1st Qu.:   0.0  
##  Median :2008            Median :72.1   Median :144   Median :   3.0  
##  Mean   :2008            Mean   :69.2   Mean   :165   Mean   :  30.3  
##  3rd Qu.:2012            3rd Qu.:75.7   3rd Qu.:228   3rd Qu.:  22.0  
##  Max.   :2015            Max.   :89.0   Max.   :723   Max.   :1800.0  
##                          NA's   :10     NA's   :10                    
##     alcohol          perexp             HB          measles            bmi       
##  Min.   : 0.01   Min.   :    0   Min.   : 1.0   Min.   :     0   Min.   : 1.0  
##  1st Qu.: 0.88   1st Qu.:    5   1st Qu.:77.0   1st Qu.:     0   1st Qu.:19.3  
##  Median : 3.76   Median :   65   Median :92.0   Median :    17   Median :43.5  
##  Mean   : 4.60   Mean   :  738   Mean   :80.9   Mean   :  2420   Mean   :38.3  
##  3rd Qu.: 7.70   3rd Qu.:  442   3rd Qu.:97.0   3rd Qu.:   360   3rd Qu.:56.2  
##  Max.   :17.87   Max.   :19480   Max.   :99.0   Max.   :212183   Max.   :87.3  
##  NA's   :194                     NA's   :553                     NA's   :34    
##       ufd           polio           texp            hiv              gdp         
##  Min.   :   0   Min.   : 3.0   Min.   : 0.37   Min.   : 0.10   Min.   :     2  
##  1st Qu.:   0   1st Qu.:78.0   1st Qu.: 4.26   1st Qu.: 0.10   1st Qu.:   464  
##  Median :   4   Median :93.0   Median : 5.76   Median : 0.10   Median :  1767  
##  Mean   :  42   Mean   :82.5   Mean   : 5.94   Mean   : 1.74   Mean   :  7483  
```

```
##  3rd Qu.:  28   3rd Qu.:97.0   3rd Qu.: 7.49   3rd Qu.: 0.80   3rd Qu.:  5911
##  Max.   :2500   Max.   :99.0   Max.   :17.60   Max.   :50.60   Max.   :119173
##                 NA's   :19     NA's   :226                     NA's   :448
##      school
##  Min.   : 0.0
##  1st Qu.:10.1
##  Median :12.3
##  Mean   :12.0
##  3rd Qu.:14.3
##  Max.   :20.7
##  NA's   :163
```

可以看出資料內含有許多遺漏值，且因為是 WHO 的資料，經檢查過後並沒有特別異常的值。

4. 資料概述與清洗

```
dim(data)
```

```
## [1] 2938   16
```

```
sum(is.na(data))
```

```
## [1] 1657
```

此次使用之資料共有 18 個變數每個變數共有 2938 列。其中共有缺失值 1657 筆。

(1) 刪除資料

```
# 刪除缺失值大於等於 5 筆的列
tem = is.na.data.frame(data)
dat = data[-which(rowSums(tem) >= 5),]
dim(data)[1] - dim(dat)[1]
```

```
## [1] 12
```

```
colSums(is.na(dat))
```

```
##    year  status    life      am     ind alcohol  perexp      HB measles     bmi
##       0       0       9       9       0     183       0     542       0      22
##     ufd   polio    texp     hiv     gdp  school
##       0       8     215       0     439     162
```

共刪除 209 筆缺失值大於等於 5 筆的列，刪除後缺失值數量為 2100 筆。

(2) 補值法

- life: 預期壽命

```
# 因 life 為反應變數，因此有缺失的列直接刪除
dat = dat[!is.na(dat$life),]
dim(dat)
```

```
## [1] 2917   16
```

```
# 查看各變數下缺失值數量。
colSums(is.na(dat))
```

```
##    year  status    life      am     ind alcohol  perexp      HB measles     bmi
##       0       0       0       0       0     182       0     542       0      21
##     ufd   polio    texp     hiv     gdp  school
##       0       8     215       0     435     160
```

- alcohol:15 歲以上人民人均酒精消費量。

```
# 檢查 alcohol 分別在開發中與已開法國家缺失值數量
sum(is.na(dat$alcohol)[dat$status == 0])
```

```
## [1] 154
```

```
sum(is.na(dat$alcohol)[dat$status == 1])
```

```
## [1] 28
```

```
# alcohol 在開發中與已開法國家缺失值數量各為 154 與 28，並分別用其 median 補值
alcohol_median_status0 = median(dat$alcohol[dat$status == 0], na.rm = TRUE)
alcohol_median_status1 = median(dat$alcohol[dat$status == 1], na.rm = TRUE)
dat$alcohol[dat$status == 0 & is.na(dat$alcohol)] = alcohol_median_status0
dat$alcohol[dat$status == 1 & is.na(dat$alcohol)] = alcohol_median_status1
# 檢查 alcohol 的缺失值
sum(is.na(dat$alcohol))
```

```
## [1] 0
```

- HB:1 歲孩童 B 型肝炎疫苗覆蓋率

```
# 檢查 HB 分別在開發中與已開法國家缺失值數量
sum(is.na(dat$HB)[dat$status == 0])
```

```
## [1] 369
```

```
sum(is.na(dat$HB)[dat$status == 1])
```

```
## [1] 173
```

```
# HB 在開發中與已開法國家缺失值數量各為 369 與 173，並分別用其 median 補值
HB_median_status0 = median(dat$HB[dat$status == 0], na.rm = TRUE)
HB_median_status1 = median(dat$HB[dat$status == 1], na.rm = TRUE)
dat$HB[dat$status == 0 & is.na(dat$HB)] = HB_median_status0
dat$HB[dat$status == 1 & is.na(dat$HB)] = HB_median_status1
# 檢查 HB 的缺失值
sum(is.na(dat$HB))
```

```
## [1] 0
```

- bmi: 體重 (kg) 除身高 (m) 的平方 $\frac{kg}{m^2}$

```
# 檢查 bmi 分別在開發中與已開法國家缺失值數量
sum(is.na(dat$bmi)[dat$status == 0])
```

```
## [1] 21
```

```
sum(is.na(dat$bmi)[dat$status == 1])
```

```
## [1] 0
```

```
# bmi 的 21 筆缺失值都在開發中國家，並依照開發中國家的 mean 補值
bmi_mean_status0 = mean(dat$bmi[dat$status == 0], na.rm = TRUE)
dat$bmi[is.na(dat$bmi)] = bmi_mean_status0
# 檢查 bmi 的缺失值
sum(is.na(dat$bmi))
```

```
## [1] 0
```

- polio:1 歲一下小兒麻痺症疫苗覆蓋率

```r
# 檢查 polio 分別在開發中與已開法國家缺失值數量
sum(is.na(dat$polio)[dat$status == 0])
```

```
## [1] 8
```

```r
sum(is.na(dat$polio)[dat$status == 1])
```

```
## [1] 0
```

```r
# polio 的 8 筆缺失值都在開發中國家，並依照開發中國家的 median 補值
polio_median_status0 = round(median(dat$polio[dat$status == 0], na.rm = TRUE),0)
dat$polio[is.na(dat$polio)] = polio_median_status0
# 檢查 polio 的缺失值
sum(is.na(dat$polio))
```

```
## [1] 0
```

- texp: 政府醫療支出佔總開支百分比

```r
# 檢查 texp 分別在開發中與已開法國家缺失值數量
sum(is.na(dat$texp)[dat$status == 0])
```

```
## [1] 183
```

```r
sum(is.na(dat$texp)[dat$status == 1])
```

```
## [1] 32
```

```r
# texp 在開發中與已開法國家缺失值數量各為 183 與 32，並分別用其 median 補值
texp_median_status0 = round(median(dat$texp[dat$status == 0], na.rm = TRUE),2)
texp_median_status1 = round(median(dat$texp[dat$status == 1], na.rm = TRUE),2)
dat$texp[dat$status == 0 & is.na(dat$texp)] = texp_median_status0
dat$texp[dat$status == 1 & is.na(dat$texp)] = texp_median_status1
# 檢查 texp 的缺失值
sum(is.na(dat$texp))
```

```
## [1] 0
```

- gdp: 人均 GDP(美元)

```r
# 檢查 gdp 分別在開發中與已開法國家缺失值數量
sum(is.na(dat$gdp)[dat$status == 0])
```

```
## [1] 371
```

```r
sum(is.na(dat$gdp)[dat$status == 1])
```

```
## [1] 64
```

```r
# gdp 在開發中與已開法國家缺失值數量各為 371 與 64，並分別用其 median 補值
gdp_median_status0 = round(median(dat$gdp[dat$status == 0], na.rm = TRUE),3)
gdp_median_status1 = round(median(dat$gdp[dat$status == 1], na.rm = TRUE),3)
dat$gdp[dat$status == 0 & is.na(dat$gdp)] = gdp_median_status0
dat$gdp[dat$status == 1 & is.na(dat$gdp)] = gdp_median_status1
# 檢查 gdp 的缺失值
sum(is.na(dat$gdp))
```

```
## [1] 0
```

- school: 受教育年數 (年)

```r
# 檢查 school 分別在開發中與已開法國家缺失值數量
sum(is.na(dat$school)[dat$status == 0])
```

```
## [1] 112
```

```r
sum(is.na(dat$school)[dat$status == 1])
```

```
## [1] 48
```

```r
# school 在開發中與已開法國家缺失值數量各為 112 與 48，並分別用其 median 補值
school_median_status0 = round(median(dat$school[dat$status == 0], na.rm = TRUE),1)
school_median_status1 = round(median(dat$school[dat$status == 1], na.rm = TRUE),1)
dat$school[dat$status == 0 & is.na(dat$school)] = school_median_status0
dat$school[dat$status == 1 & is.na(dat$school)] = school_median_status1
# 檢查 school 的缺失值
sum(is.na(dat$school))
```

```
## [1] 0
```

最後檢驗

```r
sum(is.na(dat))
```

```
## [1] 0
```

四、建立迴歸模型及檢驗

```r
fit1 = lm(life ~ . - year, data = dat)
summary(fit1)
```

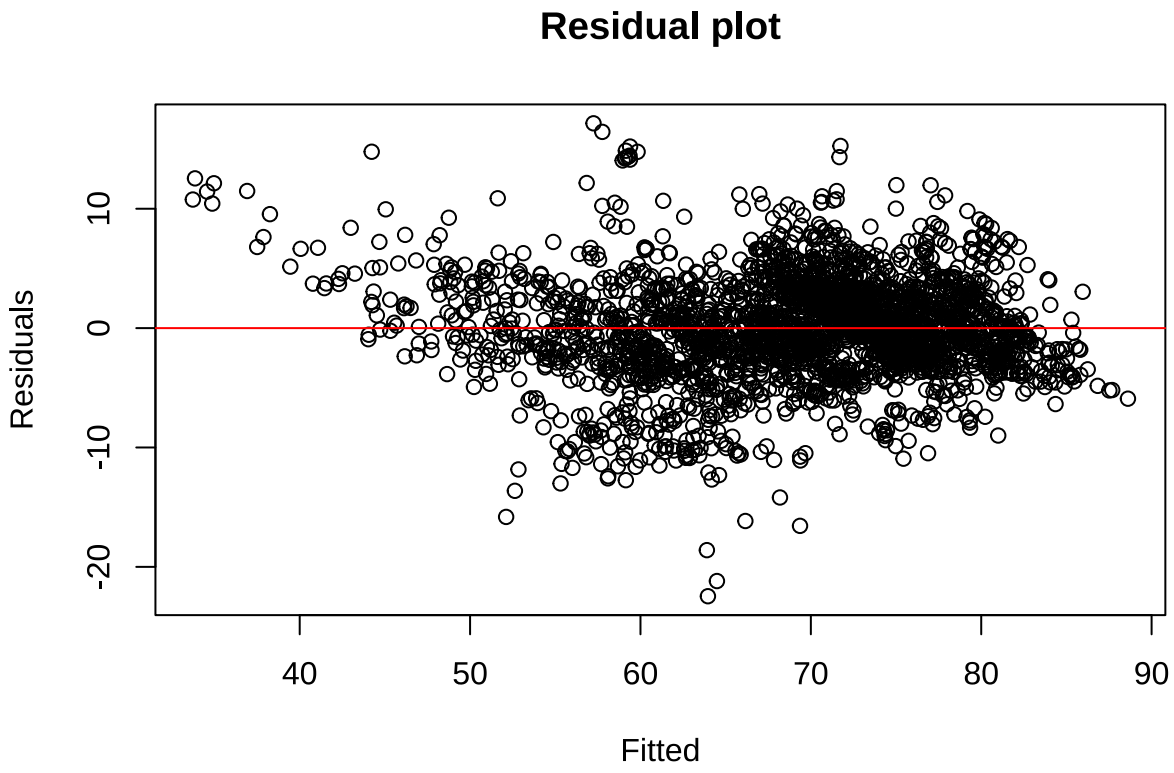
```
##
## Call:
## lm(formula = life ~ . - year, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.459  -2.247   0.014   2.470  17.155
##
## Coefficients:
##                Estimate  Std. Error t value             Pr(>|t|)
## (Intercept) 55.09041132  0.54903952  100.34 < 0.0000000000000002 ***
## status1      1.19753068  0.28582455    4.19             0.000029 ***
## am          -0.02033229  0.00081120  -25.06 < 0.0000000000000002 ***
## ind          0.11193970  0.00833960   13.42 < 0.0000000000000002 ***
## alcohol      0.07936103  0.02644358    3.00              0.00271 **
## perexp       0.00010239  0.00008685    1.18              0.23848
## HB          -0.00631123  0.00362696   -1.74              0.08195 .
## measles     -0.00002099  0.00000779   -2.70              0.00707 **
## bmi          0.04976198  0.00481651   10.33 < 0.0000000000000002 ***
## ufd         -0.08427709  0.00617867  -13.64 < 0.0000000000000002 ***
## polio        0.05023328  0.00387804   12.95 < 0.0000000000000002 ***
## texp         0.02146605  0.03475826    0.62              0.53690
## hiv         -0.48419031  0.01783988  -27.14 < 0.0000000000000002 ***
## gdp          0.00004653  0.00001348    3.45              0.00056 ***
## school       0.98856098  0.03649466   27.09 < 0.0000000000000002 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.13 on 2902 degrees of freedom
## Multiple R-squared:  0.811,  Adjusted R-squared:  0.81
## F-statistic:  890 on 14 and 2902 DF,  p-value: <0.0000000000000002
```

- 殘差檢定

(1) Constant Variance

```
plot(residuals(fit1)~fitted(fit1), xlab = "Fitted", ylab = "Residuals",
     main = "Residual plot")
abline(h = 0, col = "red")
```



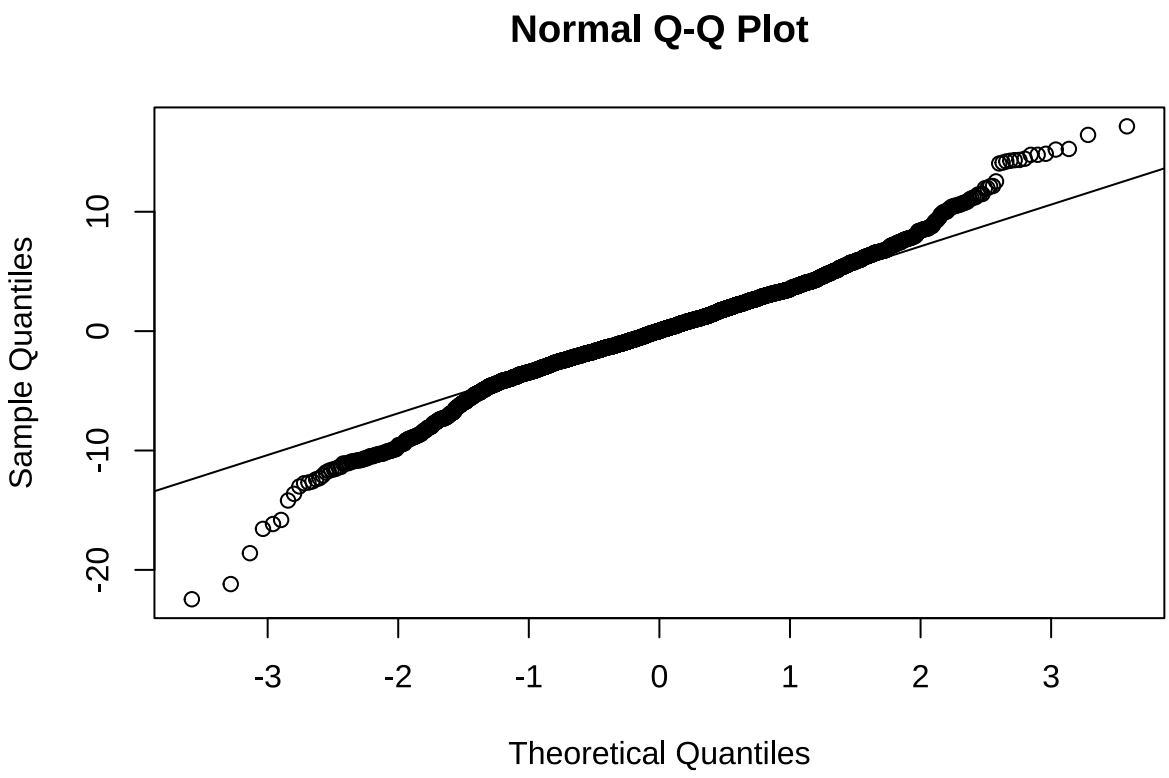

```
library(car)
ncvTest(fit1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 212.9, Df = 1, p = <0.0000000000000002
```

在 Non-constant Variance Score Test 中 p-value < 0.05，因此殘差並不符合 constant 的假設。

(2) Normality

```
qqnorm(residuals(fit1))
qqline(residuals(fit1))
```



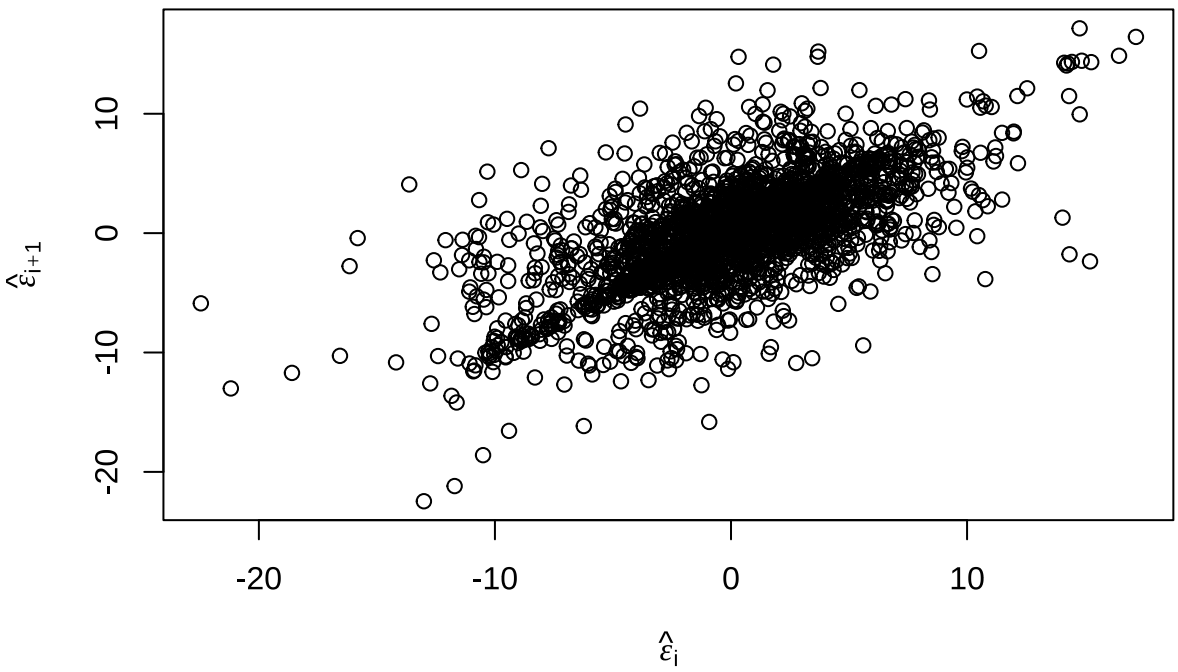

```
# 常態檢定
shapiro.test(residuals(fit1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit1)
## W = 0.98, p-value <0.0000000000000002
```

Shapiro-Wilk 常態檢定中 p-value < 0.05，因此殘差並不符合常態性的假設。

(3) Correlated Errors

```
plot(residuals(fit1)[-length(residuals(fit1))], residuals(fit1)[-1],
     xlab = expression(hat(epsilon)[i]),ylab=expression(hat(epsilon)[i+1]))
library(lmtest)
```




```
dwtest(life ~ . - year, data = dat)
```

```
##
##  Durbin-Watson test
##
## data:  life ~ . - year
## DW = 0.68, p-value <0.0000000000000002
## alternative hypothesis: true autocorrelation is greater than 0
```

Durbin-Watson 檢定中 p-value < 0.05，因此殘差不符合獨立性的假設。

- 共線性

```
library(faraway)
x = model.matrix(fit1)[,-1]
vif(x)
```

```
## status1      am     ind alcohol  perexp      HB measles     bmi     ufd   polio
##   2.019   1.724 166.251   1.899   5.120   1.189   1.370   1.566 168.924   1.404
##    texp     hiv     gdp  school
##   1.190   1.410   5.554   2.345
```

可以看到 ind(低於 1 歲之嬰兒，每 1000 人死亡人數) 與 ufd(5 歲以下人口，每 1000 人死亡人數) 兩個變數 vif > 100，我們將 ind 去除後重新配適模型。

```
fit2 = lm(life ~ . -year -ufd, data = dat)
summary(fit2)
```

```
##
```

```
## Call:
## lm(formula = life ~ . - year - ufd, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.545  -2.284   0.025   2.568  18.123
##
## Coefficients:
##                Estimate  Std. Error t value             Pr(>|t|)
## (Intercept) 53.92883135  0.55941515   96.40 < 0.0000000000000002 ***
## status1      1.16891361  0.29478573    3.97             0.000075 ***
## am          -0.02100118  0.00083512  -25.15 < 0.0000000000000002 ***
## ind         -0.00132880  0.00079235   -1.68               0.0936 .
## alcohol      0.02954275  0.02701197    1.09               0.2742
## perexp       0.00011586  0.00008957    1.29               0.1959
## HB          -0.00785204  0.00373896   -2.10               0.0358 *
## measles     -0.00003196  0.00000799   -4.00             0.000065 ***
## bmi          0.05044324  0.00496738   10.15 < 0.0000000000000002 ***
## polio        0.05862318  0.00394910   14.84 < 0.0000000000000002 ***
## texp         0.02060464  0.03584891    0.57               0.5655
## hiv         -0.49671844  0.01837529  -27.03 < 0.0000000000000002 ***
## gdp          0.00004117  0.00001390    2.96               0.0031 **
## school       1.06166792  0.03723171   28.52 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.26 on 2903 degrees of freedom
## Multiple R-squared:  0.799,  Adjusted R-squared:  0.798
## F-statistic:  888 on 13 and 2903 DF,  p-value: <0.0000000000000002
```

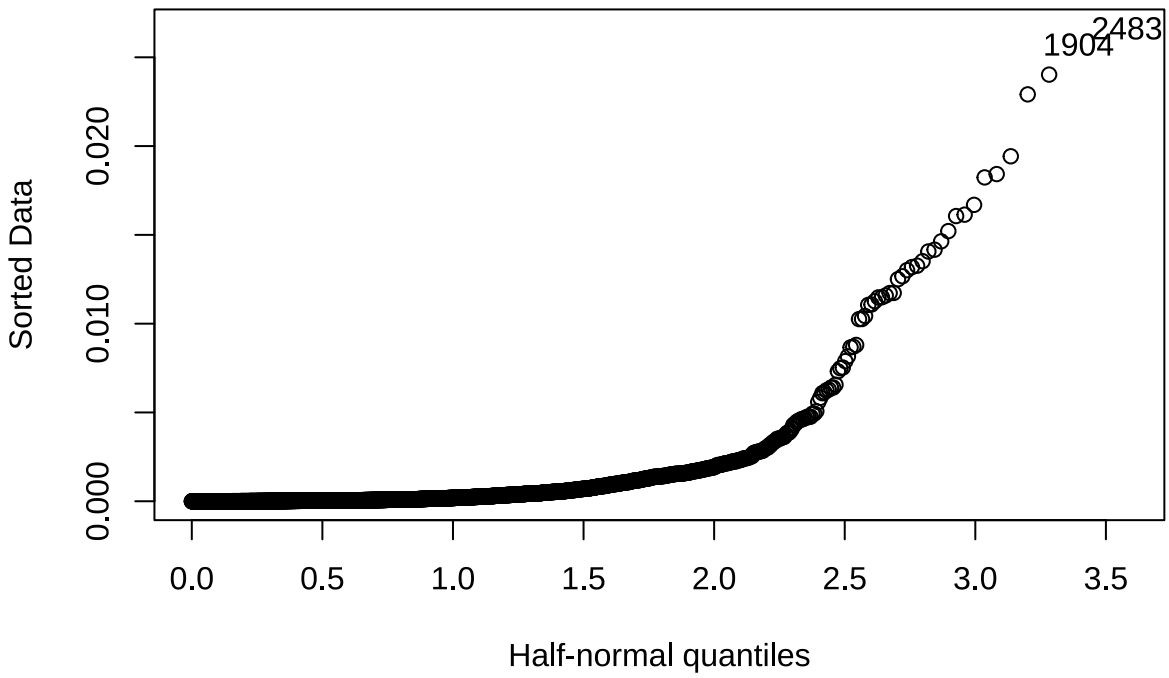
```
x = model.matrix(fit2)[,-1]
vif(x)
```

```
## status1      am     ind alcohol  perexp      HB measles     bmi   polio    texp
##   2.019   1.717   1.411   1.863   5.119   1.188   1.356   1.566   1.369   1.190
##     hiv     gdp  school
##   1.406   5.550   2.295
```

刪除 ind 後模型就不存在共線性問題。

- 離群值 & 影響點偵測

```
cook = cooks.distance(fit2)
halfnorm(cook)
```


2483
1904
Sorted Data
0.000
0.010
0.020
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Half-normal quantiles