Linear Model Assignment 5

黃晨溦、劉奕宏、鄭雅珊

Problem 1.

i.

由於兒子的身高資料為平均身高，不滿足 equal variance 的假設，因此使用 weighted least square 估計參數較為合適。假設每一群的兒子人數和父親人數相等，可使用父親人數作為 weight。配適模型

$$\Omega : \text{height of son} = \beta_0 + \beta_1 \text{height of father} + \epsilon$$

```r
dat1=read.table("http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/height.txt",
                header =F,skip = 2, col.names = c("h_father","avg_son","n"))
g=lm(avg_son~h_father,weights=n,data=dat1)
summary(g)
```

```
##
## Call:
## lm(formula = avg_son ~ h_father, data = dat1, weights = n)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -1.39024 -0.77499  0.04766  1.15672  1.67501
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  32.5820     2.2486   14.49 4.87e-08 ***
## h_father      0.5297     0.0332   15.96 1.93e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.147 on 10 degrees of freedom
## Multiple R-squared:  0.9622, Adjusted R-squared:  0.9584
## F-statistic: 254.6 on 1 and 10 DF,  p-value: 1.926e-08
```

ii.

是否可直接由父親身高預測兒子身高？依題意配適模型如下：

$$\omega : \text{height of son} = \text{height of father} + \epsilon$$

其中父親身高的係數固定為 1，因此可視為 offset，另外此模型不包含截距項。

```
g2=lm(avg_son~offset(h_father)-1,weights=n,data=dat1)
summary(g2)
```

```
## 
## Call:
## lm(formula = avg_son ~ offset(h_father) - 1, data = dat1, weights = n)
## 
## Weighted Residuals:
##    Min     1Q Median     3Q    Max 
## -4.808 -1.721  3.569  7.963  9.700 
## 
## No Coefficients
## 
## Residual standard error: 5.661 on 12 degrees of freedom
```

檢定模型是否可由 Ω 簡化為 ω：

$$H_0 : \omega \text{ vs. } H_1 : \Omega \setminus \omega$$

```
anova(g2,g)
```

```
## Analysis of Variance Table
## 
## Model 1: avg_son ~ offset(h_father) - 1
## Model 2: avg_son ~ h_father
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
## 1     12 384.54                                  
## 2     10  13.17  2    371.37 141.03 4.706e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-statistic=143.01，p-value= 4.706e-08，在顯著水準 $\alpha = 0.05$ 下，有充分證據拒絕 H_0，也就是模型 ω 過於簡化，直接由父親身高預測兒子身高是不適合的。

Problem 2.

i.

```
wd = "http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/pipeline.txt"
data2 = read.table(wd, header=T,fileEncoding = "UTF-8-BOM")
attach(data2)

summary(data2)
```

```
##      Field            Lab            Batch      
##  Min.   : 5.00   Min.   : 4.30   Min.   :1.000  
##  1st Qu.:18.00   1st Qu.:18.35   1st Qu.:2.000  
##  Median :35.00   Median :38.00   Median :3.000  
##  Mean   :33.58   Mean   :39.10   Mean   :3.234  
##  3rd Qu.:46.50   3rd Qu.:55.55   3rd Qu.:5.000  
##  Max.   :85.00   Max.   :81.90   Max.   :6.000  
```

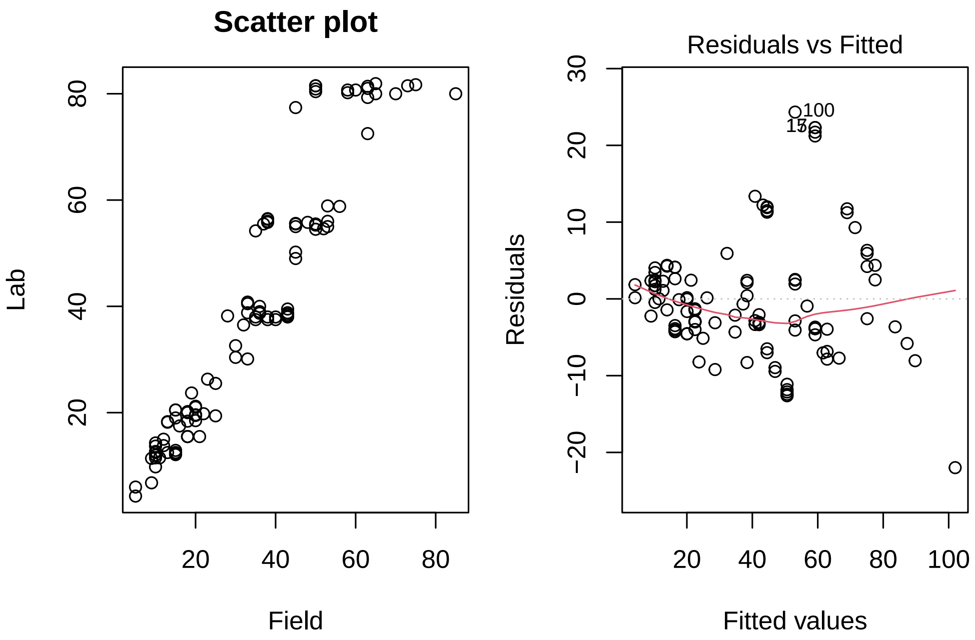
從上面的 summary 可以看到這筆 data 沒有 NA 值。

首先先對 Lab 和 Field，fit 一個 simple linear regression model:

$$\text{Lab} = \beta_0 + \beta_1 \times \text{Field} + \varepsilon.$$

並且看一下 Lab 對 Field 的 scatter plot 和 residual plot，

```r
fit2 <- lm(Lab ~ Field)
par(mfrow = c(1,2), mar = c(5,4,3,1))
plot(Field, Lab, main = "Scatter plot", xlab = "Field", ylab = "Lab")
plot(fit2, which = 1)
```



從上面兩張圖可以看到 Lab 的 variance 隨著 Field 數值越大而越大，residual 的分布也是隨著 fitted values 越大也越大，所以應該存在 non-constant variance 的問題。

也可以用 NCV-test 來檢測 non-constant variance 的性質，

$$H_0 : \text{error terms have constant variance.} \quad H_1 : \text{error terms have non-constant variance.}$$

```r
library(car)
```

```
## 載入需要的套件：carData
```

```r
ncvTest(fit2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 29.58568, Df = 1, p = 5.3499e-08
```

可以看到 NCV-test 的 p-value 很小，在 significant level $\alpha = 0.05$ 的情況下，拒絕 H_0，該模型存在 non-constant variance 的問題。

ii.

```
i <- order(data2$Field)
npipe <- data2[i,]
ff <- gl(12,9)[-108]
meanfield <- unlist(lapply(split(npipe$Field,ff),mean))
varlab <- unlist(lapply(split(npipe$Lab,ff),var))
```

假設 var(Lab) 對 Field 的連結為:

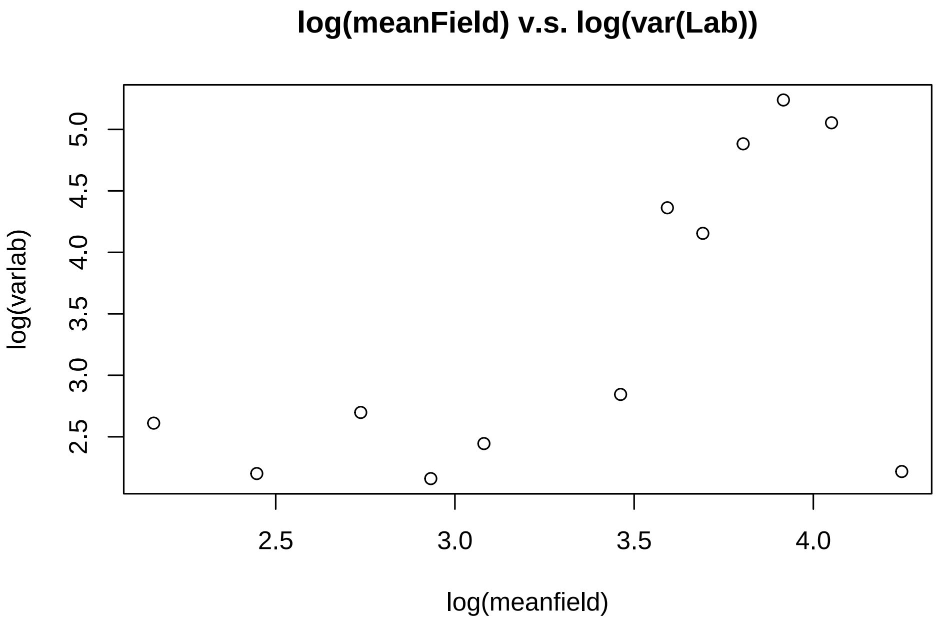
$$var(\text{Lab}) = a_0 \times \text{Field}^{a_1}$$

對兩邊取 log，

$$\begin{aligned} \log(var(\text{Lab})) &= \log\left(a_0 \times \text{Field}^{a_1}\right) \\ &= \log(a_0) + a_1 \times \log(\text{Field}). \end{aligned}$$

首先看一下 log(var(Lab)) 對 meanField 的 scatter plot:

```
plot(log(meanfield), log(varlab), main = "log(meanField) v.s. log(var(Lab))")
```



可以看到除了最後一個點以外，其他的點皆呈線性關係，所以可以拔掉最後一個點在進行配適模型。

```
fit2_2 <- lm(log(varlab) ~ log(meanfield), subset = -12)
summary(fit2_2)
```

```
##
## Call:
## lm(formula = log(varlab) ~ log(meanfield), subset = -12)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00477 -0.42268  0.05989  0.37854  0.93815
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.9352     1.0929  -1.771 0.110403
## log(meanfield)   1.6707     0.3296   5.070 0.000672 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.657 on 9 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7118
## F-statistic:  25.7 on 1 and 9 DF,  p-value: 0.0006723
```

從 summary 中可以看到，$\hat{a}_0 = \exp(-1.9352) = 0.1444,\ \hat{a}_1 = 1.6707$，所以 WLS regression 的 *weights* $= \frac{1}{\hat{a}_0 \text{Field}^{\hat{a}_1}}$。

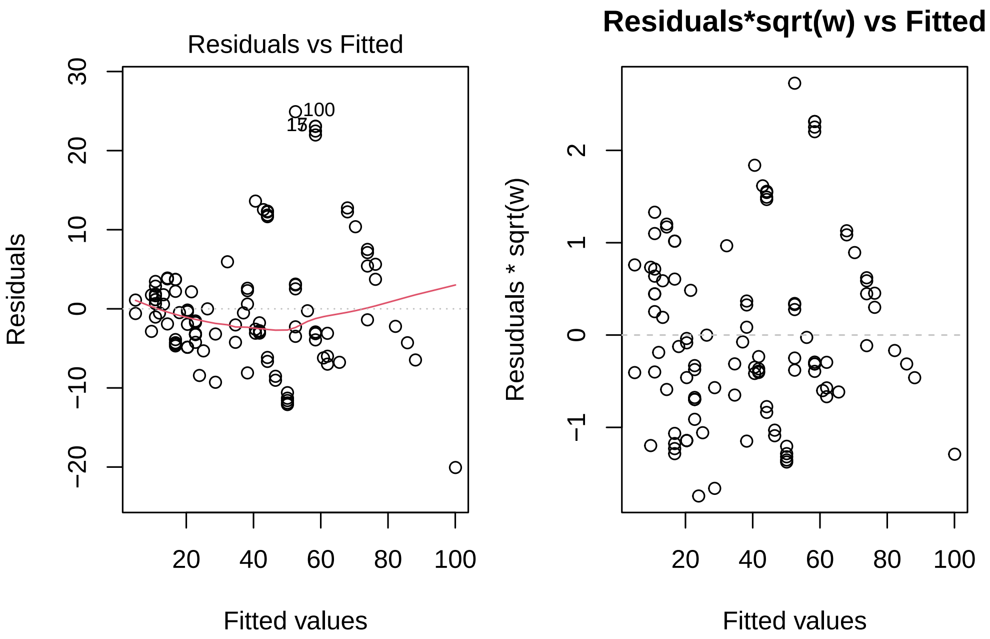
對 Lab 和 Field，fit 一個 WLS regression:

```r
w <- 1/(exp(fit2_2$coefficients[1])*Field^fit2_2$coefficients[2])
fit2_3 <- lm(Lab ~ Field, weights = w)
summary(fit2_3)
```

```
##
## Call:
## lm(formula = Lab ~ Field, weights = w)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7432 -0.6719 -0.2493  0.5967  2.7275
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.05530    0.69765  -1.513    0.133
## Field        1.18963    0.03401  34.984   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9846 on 105 degrees of freedom
## Multiple R-squared:  0.921,  Adjusted R-squared:  0.9202
## F-statistic:  1224 on 1 and 105 DF,  p-value: < 2.2e-16
```

接著一樣去畫 residual plot，

```r
par(mfrow = c(1,2), mar = c(5,4,3,1))
plot(fit2_3, which = 1)
plot(fit2_3$fitted.values, fit2_3$residuals*sqrt(w), main = "Residuals*sqrt(w) vs Fitted",
     xlab = "Fitted values", ylab = "Resuduals * sqrt(w)")
abline(a = 0, b = 0, lty = "dashed", col = "#c4c4c4")
```



左圖可以看到從 residual plot 來看新模型的 residuals 依然隨著 fitted values 變大而變大，但在加入 weight 的模型下，$Var(y) = \sigma^2/w = \sigma^2 * a_0\text{Field}^{a_1}$ 這樣的趨勢是合理的。若想從 residuals 的散佈情形直接看 non-constant variance 問題得到緩解，需要了解的是 $w * Var(y) = \sigma^2$ 的情況。此時，可將 residuals 乘上 sqrt(weight) 對 Fitted value 作出如右圖的結果，觀察發現 residuals*sqrt(weight) 沒有隨著 fitted values 變大而變大，non-constant variance 的問題得到解決，最後，看一下 NCV-Test，

```
ncvTest(fit2_3)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.9203408, Df = 1, p = 0.33739
```

```
detach(data2)
```

從 NCV-Test 的結果中可以看到 $p - value = 0.33739 > 0.05$，所以不拒絕 H_0，也就是沒有足夠的證據說明此模型存在 non-constant variance 的問題。

Problem 3.

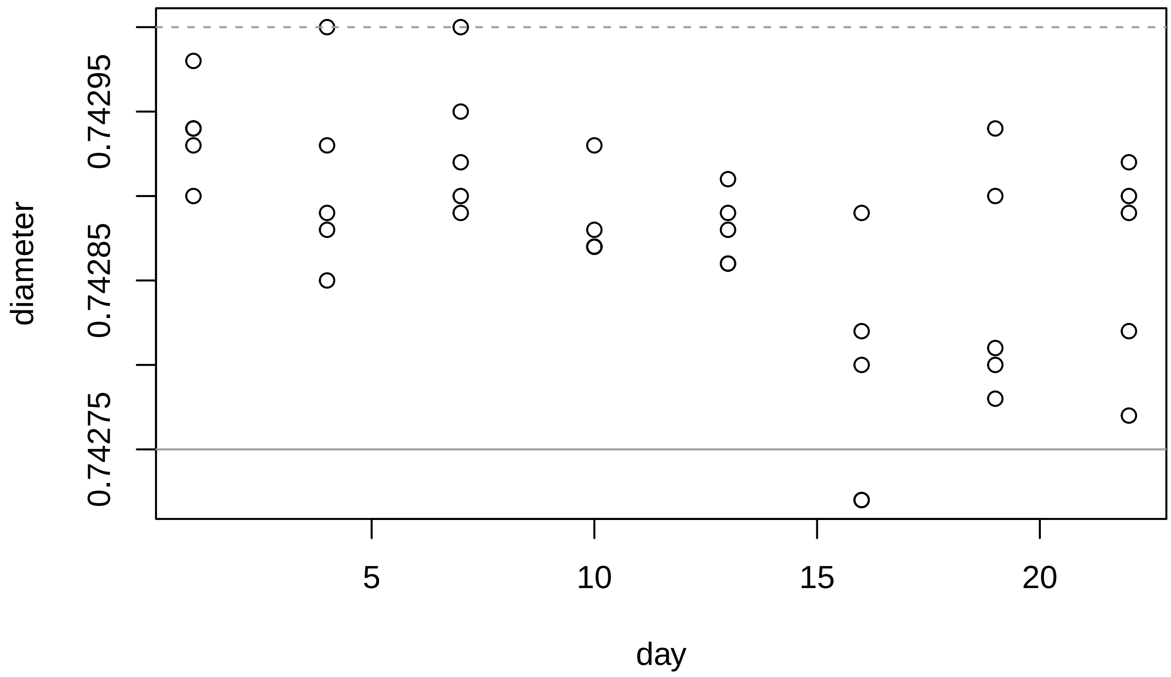
首先對資料中 diameter 變數做轉換：

$$\text{diameter(new)} = \text{diameter(original)} \times 0.00001 + 0.742$$

```
dat3=read.table("http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/crank.txt",
                header=T,fileEncoding = "UTF-8-BOM")
dat3$diameter <- dat3$diameter*1e-5+0.742
```

接著繪製 diameter 對 day 的散布圖：

```
plot(dat3$day,dat3$diameter,xlab="day",ylab="diameter")#,axes = F
abline(h=0.74275,col=8); abline(h=0.743,col=8,lty=2)
```



從圖中可見：

1. 大部分的 diameter 數據都落在資料範圍的中間值 (0.7425+0.743)/2=0.74275 以上
2. 隨著時間 (day) 增加，diameter 越靠近 0.74275 水平線

顯示製程可能沒有 under control。進一步配適模型

$$\omega : \text{diameter} = \beta_0 + \beta_1 \text{day} + \epsilon$$

```
g3 <- lm(diameter~day,data = dat3)
summary(g3)
```

```
##
## Call:
## lm(formula = diameter ~ day, data = dat3)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.369e-04 -3.207e-05 -5.000e-06  3.886e-05  9.857e-05
##
## Coefficients:
```

```
##                Estimate Std. Error   t value Pr(>|t|)
## (Intercept)  7.429e-01  1.720e-05 43186.390  < 2e-16 ***
## day         -5.143e-06  1.284e-06    -4.005 0.000278 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.582e-05 on 38 degrees of freedom
## Multiple R-squared:  0.2968, Adjusted R-squared:  0.2783
## F-statistic: 16.04 on 1 and 38 DF,  p-value: 0.000278
```

製程 under control 須滿足以下兩個條件

1. Average size of the crankpins produced fall near the middle of the specified range
2. Average size of the crankpins produced does not depend on time

根據所配適的模型 ω，可將檢定問題視為

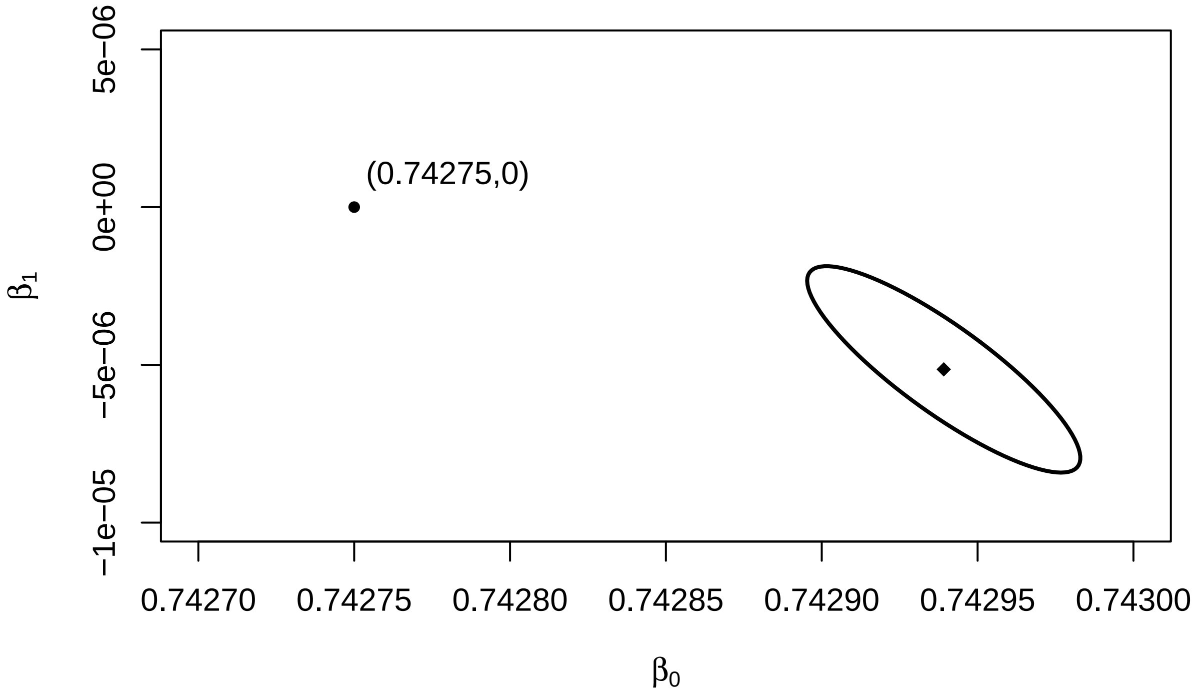
$$H_0 : \beta_0 = 0.74275 \text{ and } \beta_1 = 0 \text{ vs. } H_1 : \text{ not } H_0$$

繪製 β_0 及 β_1 的 95% confidence region 如下：

```
library(ellipse)
```

```
## Warning: ®M¥ó 'ellipse' ¬O¥Î R ª©¥» 4.1.3 ¨Ó«Ø³yªº
```

```
plot(ellipse(g3, c(1,2)), lwd=2, type="l",xlab=expression(beta[0]),
     ylab=expression(beta[1]),xlim=c(0.7427,0.743),ylim=c(-1e-5,5e-6))
points(g3$coef[1], g3$coef[2],pch=18)
se <- sqrt(diag(summary(g3)$cov * (summary(g3)$sigma^2)))
points(0.74275, 0, pch=20)
text(0.74275+3e-5,1e-6,"(0.74275,0)")
```



由於 $(\beta_0,\ \beta_1)=(0.74275,0)$ 並未落在橢圓內，在顯著水準在顯著水準 $\alpha=0.05$ 下，有充分證據拒絕 H_0，也就是製程沒有 under control。

另一種作法為配適模型

$$\omega_2:\text{diameter}=0.74275+\epsilon$$

並和模型 ω 進行比較。

```
g3a <- lm(diameter~offset(rep(0.74275,dim(dat3)[1]))-1,data = dat3)
summary(g3a)
```

```
##
## Call:
## lm(formula = diameter ~ offset(rep(0.74275, dim(dat3)[1])) -
##     1, data = dat3)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0000300  0.0001075  0.0001400  0.0001725  0.0002500
##
## No Coefficients
##
## Residual standard error: 0.0001453 on 40 degrees of freedom
```

檢定問題可視為

$$H_0:\omega_2 \text{ vs. } H_1:\omega\setminus\omega_2$$

```r
anova(g3a,g3)
```

```
## Analysis of Variance Table
## 
## Model 1: diameter ~ offset(rep(0.74275, dim(dat3)[1])) - 1
## Model 2: diameter ~ day
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)    
## 1     40 8.4440e-07                                   
## 2     38 1.1841e-07  2 7.2599e-07 116.49 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

由於 F-test 的 p-value<2.2e-16，在顯著水準 $\alpha = 0.05$ 下，有充分證據拒絕 H_0，顯示 average size of the crankpin 並沒有落在 0.74275 inches 且製程可能有隨著時間偏移。

最後檢定模型 ω 是否有 lack of fit：

$$H_0 : \omega \text{ is correct vs. } H_1 : \omega \text{ is too simple}$$

首先配適 saturated model

$$\Omega : \text{diameter} = \beta_0 + \sum_{i=1}^{7} \beta_i I(\text{day} = 3i + 1) + \epsilon$$

```r
g3b <- lm(diameter~factor(day),data = dat3)
summary(g3b)
```

```
## 
## Call:
## lm(formula = diameter ~ factor(day), data = dat3)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -9.0e-05 -3.3e-05 -6.0e-06  3.0e-05  1.0e-04 
## 
## Coefficients:
##                  Estimate Std. Error   t value Pr(>|t|)    
## (Intercept)     7.429e-01  2.308e-05 32195.321  < 2e-16 ***
## factor(day)4   -2.800e-05  3.263e-05    -0.858  0.39728    
## factor(day)7   -6.000e-06  3.263e-05    -0.184  0.85529    
## factor(day)10  -5.400e-05  3.263e-05    -1.655  0.10776    
## factor(day)13  -5.800e-05  3.263e-05    -1.777  0.08503 .  
## factor(day)16  -1.480e-04  3.263e-05    -4.535 7.63e-05 ***
## factor(day)19  -9.200e-05  3.263e-05    -2.819  0.00819 ** 
## factor(day)22  -7.800e-05  3.263e-05    -2.390  0.02290 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 5.16e-05 on 32 degrees of freedom
## Multiple R-squared:  0.4941, Adjusted R-squared:  0.3834 
## F-statistic: 4.464 on 7 and 32 DF,  p-value: 0.001467
```

```
anova(g3,g3b)
```

```
## Analysis of Variance Table
##
## Model 1: diameter ~ day
## Model 2: diameter ~ factor(day)
##   Res.Df        RSS Df  Sum of Sq     F  Pr(>F)
## 1     38 1.1841e-07
## 2     32 8.5200e-08  6 3.3211e-08 2.079 0.08354 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

由 F-test 結果可知，p-value=0.08354，因此在顯著水準 $\alpha = 0.05$ 下，沒有充分證據拒絕 H_0，也就是用模型 ω 解釋資料是合適的。