HW 5-Linear Model

ID : 111024517 Name : 鄭家豪

due on 12/01

Problem 1

Read data

```
data <- read.table('http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/height.txt',
                   header = FALSE,skip = 2)
colnames(data) <- c("HF","Av_HS","NumF")
kable(t(data))
```

HF	62.0	63.0	64.0	65.0	66.0	67.0	68.0	69.0	70.0	71.0	72.0	73
Av_HS	65.5	66.5	66.8	66.8	67.6	67.8	68.6	69.1	69.5	70.6	70.3	72
NumF	2.0	6.0	12.0	19.0	27.0	26.0	26.0	26.0	20.0	15.0	8.0	5

(HF:Height of Father ; Av_HS:Average Height of Son ; Numf : number of Fathers)

(i)

Model: Av_HS $= \beta_0 + \beta_1 \times$ HF $+ \epsilon$, where $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2/n_{s_i}$.
考量到這組數據只提供父親的個數 (n_i) 但並沒有提供每個兒子身高平均數的個數 (n_{s_i})，這裡應使用 Weighted least square(WLS) 進行分析，其權重 (weight)，由這組數據提供的資訊，只能假設父親個數與兒子個數成比例 $(n_{s_i} \propto n_i)$，考慮使用每個身高的父親個數來當作 weight:

$$\mathrm{w}_i = \mathrm{n}_i$$

```
weight <- data$NumF
model <- lm(Av_HS ~ HF , data=data,weights = weight)
summary_model <- summary(model)
summary_model
```

```
##
## Call:
## lm(formula = Av_HS ~ HF, data = data, weights = weight)
##
## 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 ***
## HF            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
```

由此模型的 summary 結果，我們得到 $\widehat{\beta}_1 = 0.5297$，當父親的身高增加一單位，兒子的平均身高會增加 0.5297 單位，代表身高比較高的父親，其兒子平均身高會比較高。

因此，該題所求的模型為: $\text{Av\_HS} = 32.5820 + 0.5297 \times \text{HF}$。

(ii)

依照題意，需要檢定

$$\text{H}_0 : \beta_0 = 0, \beta_1 = 1 \quad \text{v.s.} \quad \text{H}_1 : \beta_0 \neq 0 \text{ or } \beta_1 \neq 1$$

這裡使用 anova() 指令來作檢定:

```r
model_ii <- lm(Av_HS ~ HF -1 ,offset = HF,data = data,weights=weight)
anova(model_ii,model)
```

```
## Analysis of Variance Table
##
## Model 1: Av_HS ~ HF - 1
## Model 2: Av_HS ~ HF
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     11 289.608
## 2     10  13.166  1    276.44 209.96 4.873e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

因為 p-value $= 4.873 \times 10^{-8} < 0.05$，所以 reject H_0 with significant level $\alpha = 0.05$。這意味著不適合直接用爸爸身高來預測兒子平均身高。

Problem 2

Read data

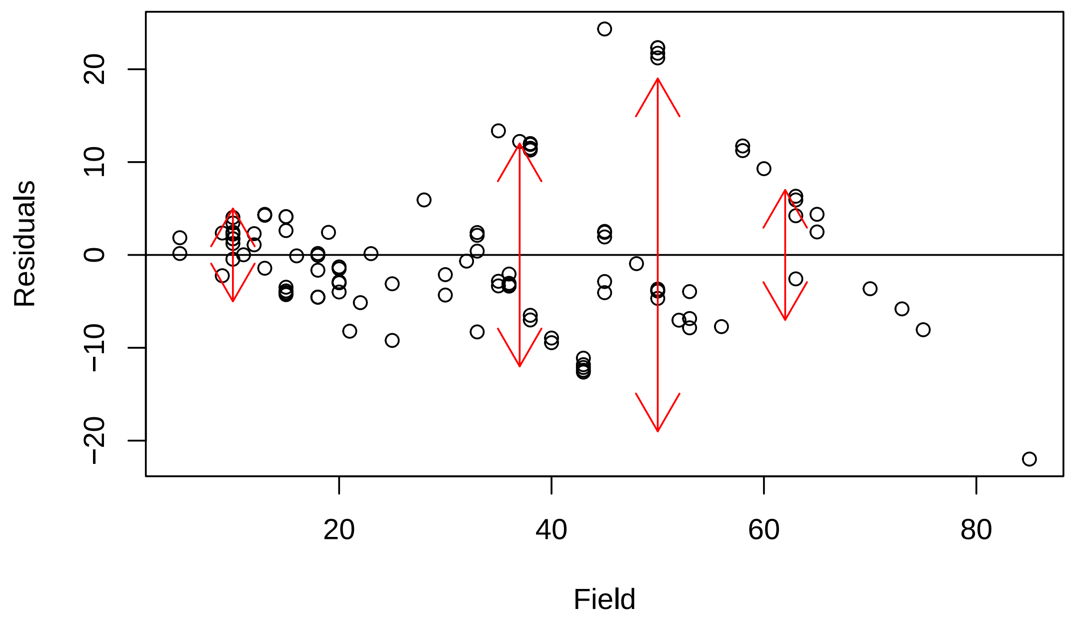
```
pipe <- read.table("http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/pipeline.txt",
                   header=TRUE)
```

(i)

Model: $\text{Lab} = \beta_0 + \beta_1 \times \text{Field} + \epsilon$

```
##
## Call:
## lm(formula = Lab ~ Field, data = pipe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.985  -4.072  -1.431   2.504  24.334
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.96750    1.57479  -1.249    0.214
## Field        1.22297    0.04107  29.778   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.865 on 105 degrees of freedom
## Multiple R-squared:  0.8941, Adjusted R-squared:  0.8931
## F-statistic: 886.7 on 1 and 105 DF,  p-value: < 2.2e-16
```

Fitting model: $\hat{y}_{\text{Label}} = -1.96750 + 1.22297 \times \text{Field}$



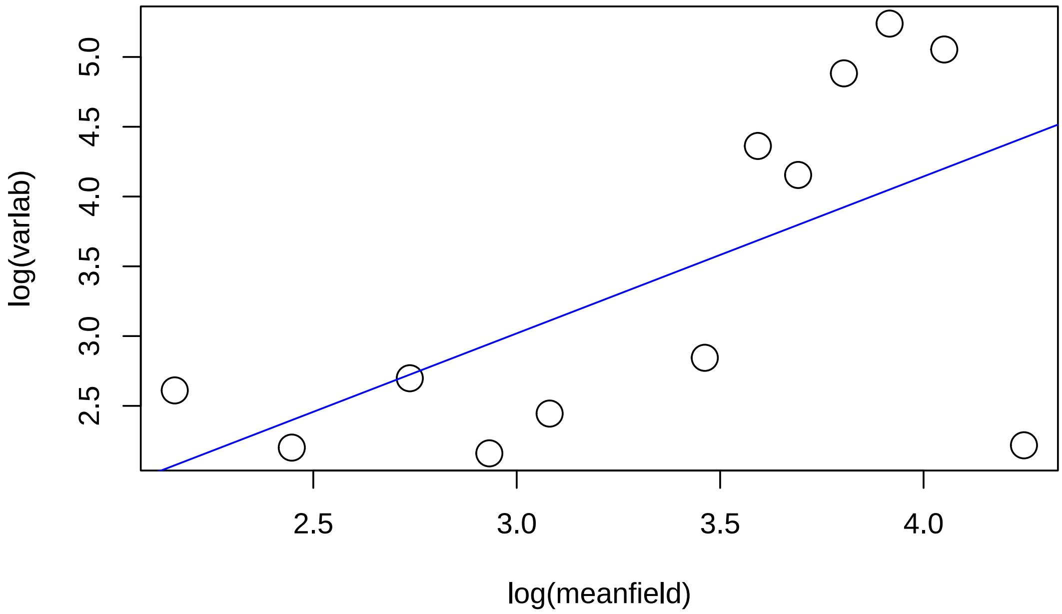
上圖為此模型的 residual plot ，可以發現隨著 Field 增加，var(residual) 會跟著增加，不過到後半段又減少了。因此 non-constant width band，代表 non-constant variance.

(ii)

Model: Log var(Lab) = Log $a_0 + a_1$Log Field + ϵ

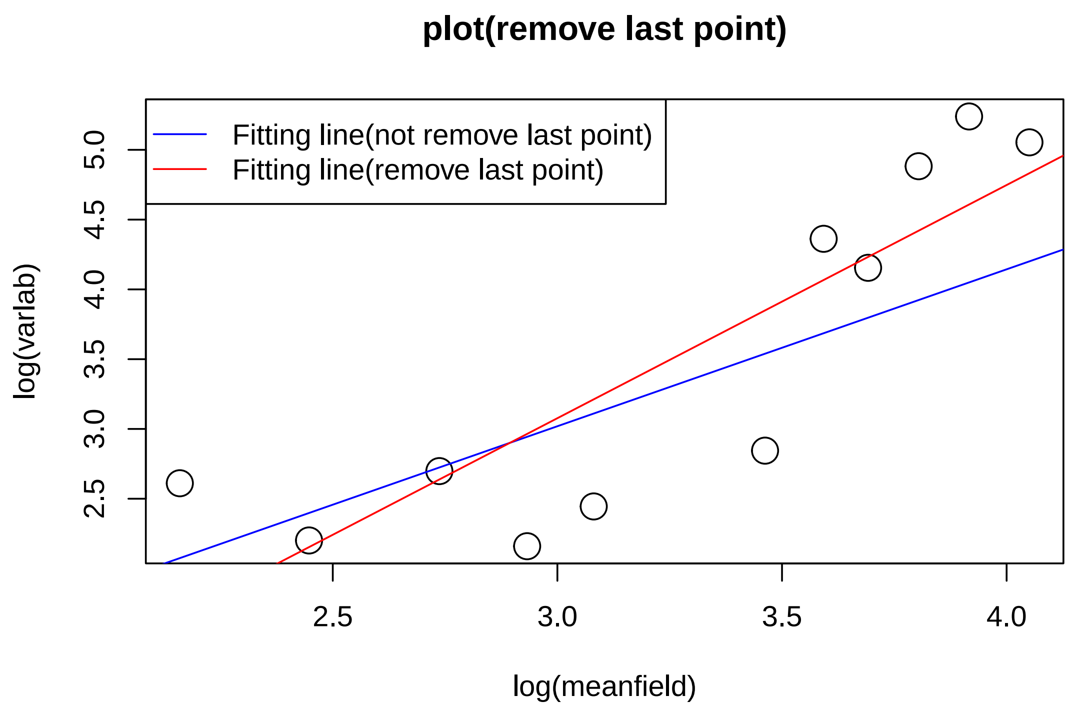
```
##
## Call:
## lm(formula = log(varlab) ~ log(meanfield))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.2038 -0.6729  0.1656  0.7205  1.1891
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.3538     1.5715  -0.225   0.8264
## log(meanfield)    1.1244     0.4617   2.435   0.0351 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.018 on 10 degrees of freedom
## Multiple R-squared:  0.3723, Adjusted R-squared:  0.3095
```

```
## F-statistic: 5.931 on 1 and 10 DF,  p-value: 0.03513
```



From the above，由於兩變數的關係主要是隨著 meanfield 增加 varlab 隨之增加，很明顯 log(meanfield) 最大值的點為一個離群點，我們將其移除再做一次 regression:

```
##
## Call:
## lm(formula = log(varlab_remove) ~ log(meanfield_remove))
##
## 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_remove)   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
```



我們會發現，移除離群點後的 regression line 斜率會比原本的 regression line 的斜率還高些，這是因為離群點的 log(varlab) 值太小會使得 regression line 往下移動的效應存在。

對兩條線做比較，會發現紅線的 fitting line 比較貼近大部份的點，因此使用移除離群點後得出的 model 來估計係數:

	coefficient
(Intercept)	-1.935167
log(meanfield_remove)	1.670723

這裡我們得到 a_0 和 a_1 的估計值分別為 $\exp(-1.935167) = 0.1444001$ 和 1.670723.

接著我們進行 WLS fit of Lab on Field: 在 (i) 我們知道 non-constant variance，根據 L.N. p.6-4 ，其 weight=$\frac{1}{\text{var(Lab)}}$ given $(\hat{a}_0 = 0.1444002, \hat{a}_1 = 1.670723)$。

```
w <- 1/(0.1444001*(npipe$Field)^(1.670723))
model2 <- lm(Lab ~ Field , data=npipe,weights = w)
summary(model2)
```

```
##
## Call:
## lm(formula = Lab ~ Field, data = npipe, 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
```

這裡得到相當高的 $R^2 = 0.921$ 和相當小的 $\hat{\sigma} = 0.9846$，與沒給予 weight 的 regression summary 作對比 (其 $R^2 = 0.8941, \hat{\sigma} = 7.865$)，模型結果改善很多。

Problem 3

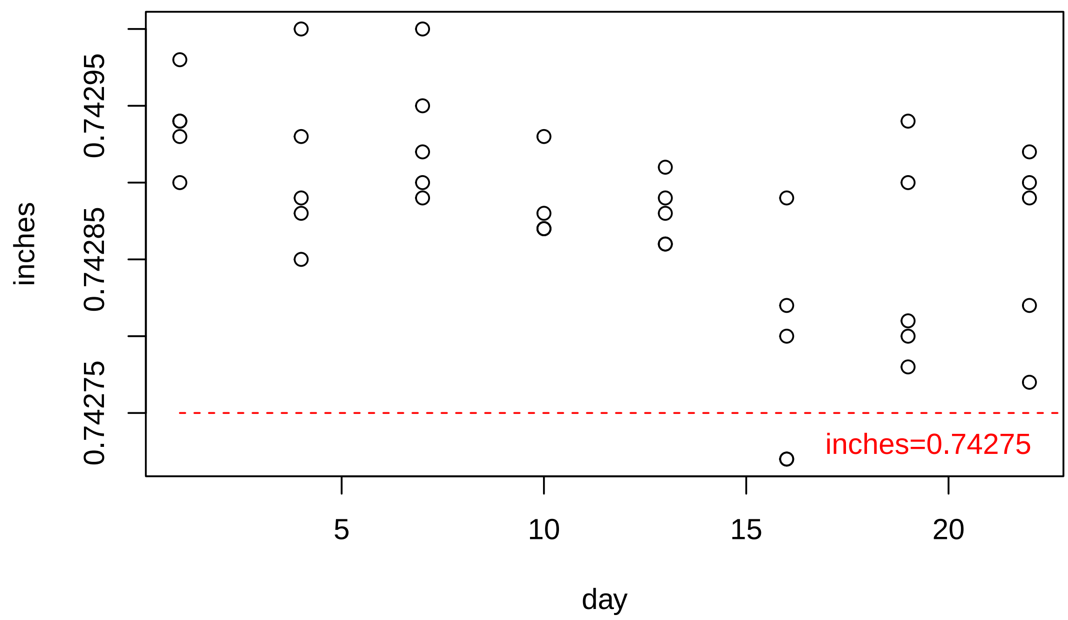
- Read data:

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

這裡先所有曲柄銷 (crankpin) 的外徑長換換成 inches:

```
inches <-  0.00001*data$diameter + 0.742
data[,3] <- inches
names(data)[3] <- "inches"
inches
```

```
##  [1] 0.74293 0.74298 0.74290 0.74294 0.74294 0.74293 0.74300 0.74288 0.74285
## [10] 0.74289 0.74289 0.74290 0.74292 0.74295 0.74300 0.74293 0.74288 0.74287
## [19] 0.74287 0.74287 0.74288 0.74286 0.74291 0.74289 0.74286 0.74282 0.74272
## [28] 0.74280 0.74272 0.74289 0.74281 0.74280 0.74278 0.74294 0.74290 0.74290
## [37] 0.74292 0.74282 0.74277 0.74289
```



檢驗:

- (1) : Responses(average size) fall near the middle of the specified range

- (2) : Responses(average size) should not depend on time
  由上面的 inches-day plot 可以看出有 8 個固定 day 下對應的點有 8 個 group，並且很明顯觀察到每個 group mean(inches(diameter) average per group ,denoted as $\bar{y}_{i\cdot}$) 都在中位數 0.74275 之上，依照圖形來看 (1) 是不滿足的，但為了依統計顯著性說明這件事，需要做檢定。

要檢定 (1)、(2)，等同於 testing :

$$\mathrm{H}_0 : \bar{y}_i = 0.74275 + \epsilon \quad \text{v.s.} \quad \mathrm{H}_1 : \bar{y}_i = \beta_0 + \beta_1 \times \text{unique(day)} + \epsilon$$

這裡算出每一群的 inches(diameter) average:

```
y_bar <- c()
day <- unique(data$day)
for (i in 1:length(day)){
  y_bar[i] <- mean(data$inches[data$day ==day[i]])
}
```

接著使用 anova() 指令來作檢定:

```r
lm <- lm(y_bar ~ day)
lm_null <- lm(y_bar ~ offset(0.74275*rep(1,8))-1)
anova(lm_null,lm)
```

```
## Analysis of Variance Table
##
## Model 1: y_bar ~ offset(0.74275 * rep(1, 8)) - 1
## Model 2: y_bar ~ day
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1      8 1.5184e-07
## 2      6 6.6420e-09  2 1.452e-07 65.579 8.371e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

因為 p-value $= 8.371 \times 10^{-5} < 0.05$，reject H_0 at level $\alpha = 0.05$。不過考量到 response 為每一群的平均值，使得樣本數只有 8，數量太少可能會不夠充分說明檢定結果。這裡用 individual inches(diameter) 作為 response 做檢定:

$$H_0 : \text{inches} = 0.74275 + \epsilon \quad \text{v.s.} \quad H_1 : \text{inches} = \beta_0 + \beta_1 \times \text{day} + \epsilon$$

使用 anova() 指令做檢定:

```r
lm <- lm(inches ~ day , data=data)
lm_null <- lm(inches ~ offset(0.74275*rep(1,40))-1)
anova(lm_null,lm)
```

```
## Analysis of Variance Table
##
## Model 1: inches ~ offset(0.74275 * rep(1, 40)) - 1
## Model 2: inches ~ 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
```

其 p-value < 0.05，reject H_0 at level $\alpha = 0.05$，與使用每一群 inches 的平均數當 response 的推論結果一樣。

總和上述，(1)、(2) 中至少有一個會不成立，因此這個製程不應該是 “under control”。

最後我們來做 lack of fit test:

$$H_0 : \text{inches} = \beta_0 + \beta_1 \times \text{day} + \epsilon \quad \text{v.s.} \quad H_1 : \text{inches} = \beta_0 + \beta_1 \times \text{day} + \epsilon \text{ is too simple}$$

我們使用 anova() 指令做檢定:

```
lm <- lm(inches ~ day , data=data)
lm_sature <- lm(inches ~ factor(day),data=data)
anova(lm,lm_sature)
```

```
## Analysis of Variance Table
##
## Model 1: inches ~ day
## Model 2: inches ~ 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
```

這裡觀察到 p-value $= 0.08354 > 0.05$，代表沒有足夠證據說明此模型有 lack of fit。