

* 應用迴歸模型辨識歧異的觀測值 一一
一 以氣溫校驗為例 一一

台北市立大學 數學系 李美賢

2018.02.05

*異常的偵測

* What are anomalies?

- The set of data points that are considerably different than the remainder of the data
- 通常異常的資料為離群值(outliers)、極端值(extreme value)，因為在資料的分布中，離群值會遠離其他的資料點。

* 異常偵測 (anomaly detection)

- * 偏差偵測 (deviation detection)
- * 例外探勘 (exception mining)

Tan, P. N., et al. (2018) Introduction to Data Mining 2nd Ed. Addison-Wesley

* 異常的偵測

* Applications:

- * 生態系統干擾(ecosystem disturbances)
- * 故障檢測 (fault detection)
- * 信用卡盜刷偵測 (credit card fraud detection)
- * 電信詐欺偵測 (telecommunication fraud detection)
- * 駭客入侵 (network intrusion detection)

* 異常的偵測

* 異常資料類型與影響

*單變量歧異值的偵測

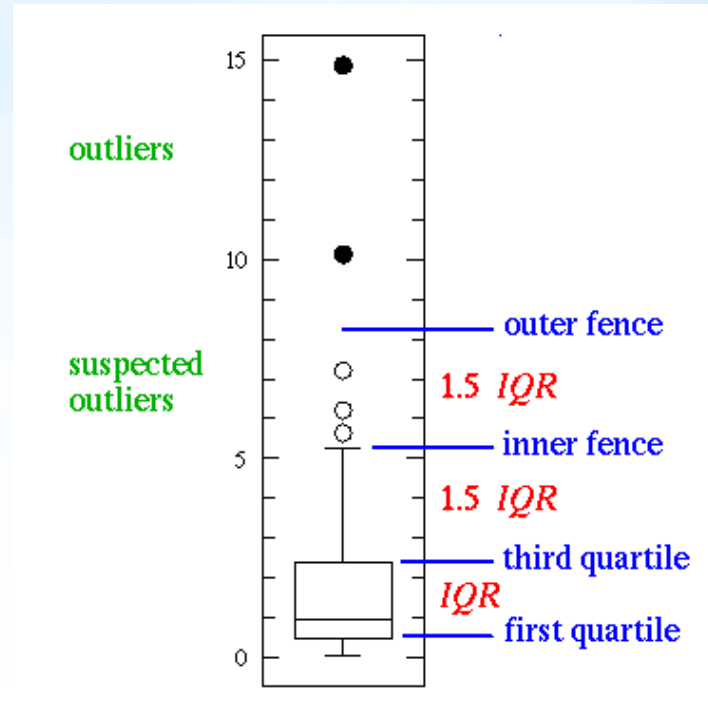
* Box plot (Tukey, 1977)

* Z值法(Younger, 1979)

*影響

統計量的估計值與推論，
如：平均值、變異數...等。

<http://www.physics.csbsju.edu/stats/box2.html>



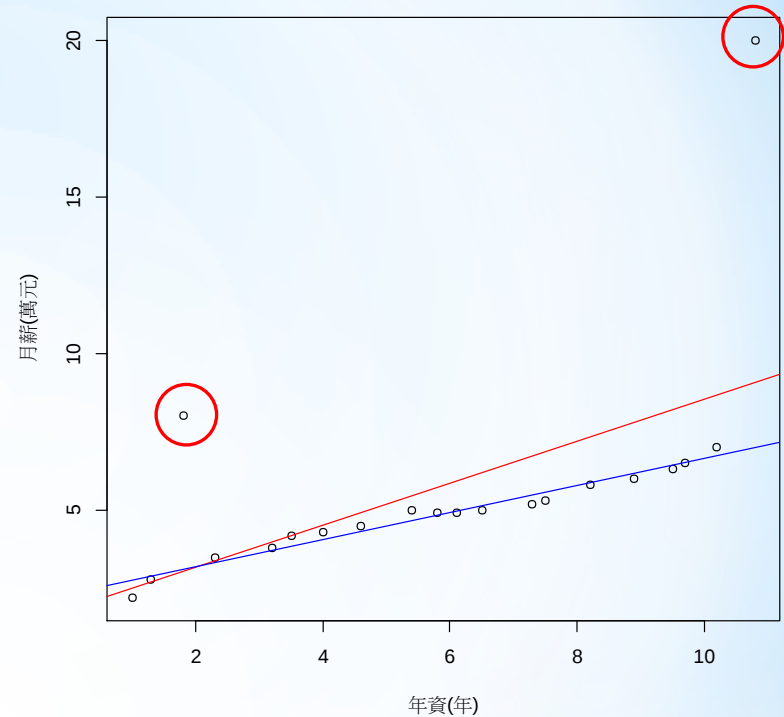
*異常資料類型與影響

* 多變量歧異值的偵測

某一筆觀察值的**組合**與其他觀察值的**組合**有很大的差異。

* 影響

資料關連性、模型推估、迴歸分析中的斜率估計值 (Pedhazur, 1997)



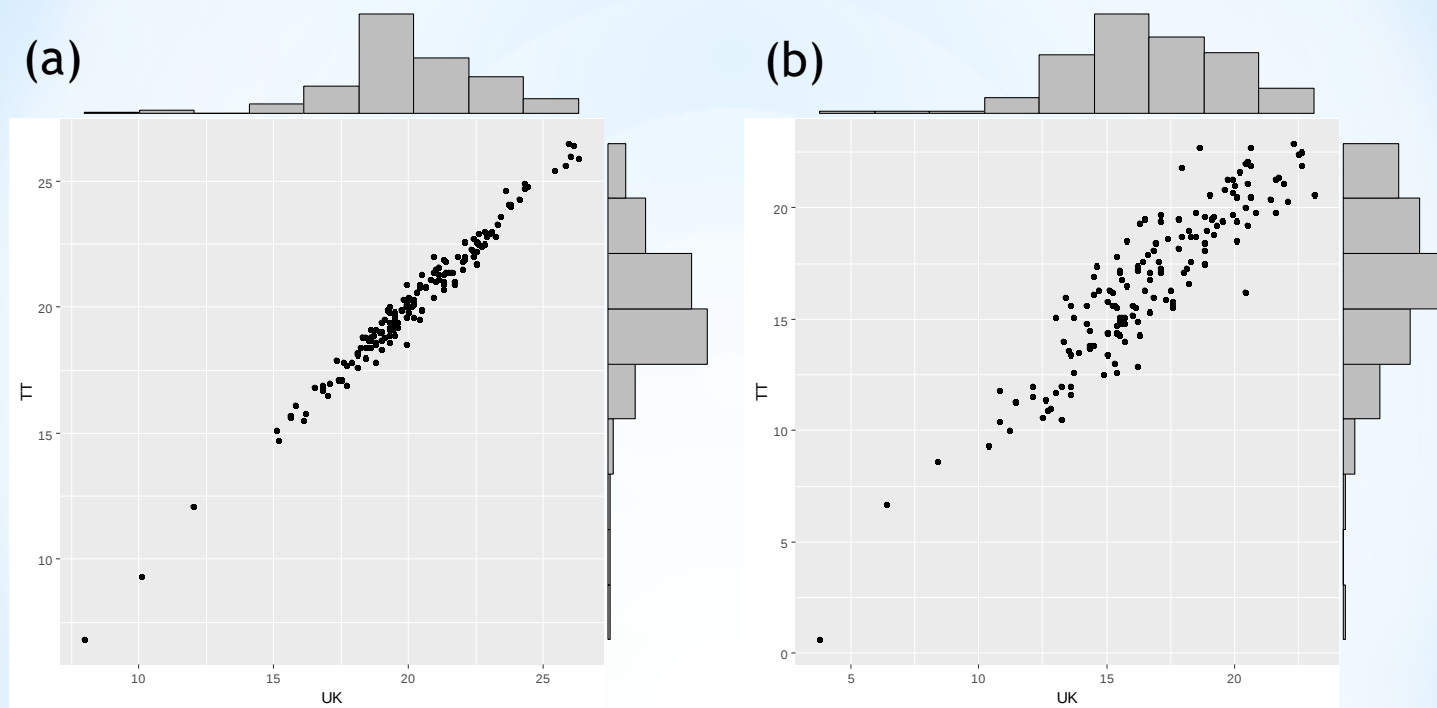
* 異常資料類型與影響

* 異常資料的判斷方法

以氣溫校驗為例

- * 以A (平地) 、B (山地) ，兩測站為例。
- * 擷取2013年至2017年，這五年的一月份、每日上午11點之氣溫觀測值。
- * 兩測站各有155筆資料，各筆資料包含氣溫觀測值(TT)與通用克利金法參考值(UK)兩個變數。
- * 希望透過克利金法參考值(UK)來判斷氣溫觀測值(TT)是否有異常狀況。

* 氣溫校驗資料



氣溫觀測值(TT)與通用克利金法估計值(UK)散佈圖與直方圖
(a) A測站資料，共計155筆資料點；(b) B測站資料，因三個資料點的氣溫觀察值遺漏，僅採用152筆資料。

*氣溫校驗資料

* $E\{Y_h\}$ 的95%信賴區間：

$$Y_h \pm t_{(0.975, n-2)} \cdot s\{Y_h\}$$

* $Y_{h(\text{new})}$ 的95%預測區間：

$$Y_h \pm t_{(0.975, n-2)} \cdot \left[MSE \left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right] \right]$$

$$Y_h \sim N(\beta_0 + \beta_1 x_h, V(Y_h))$$

$$V(Y_h) = \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]$$

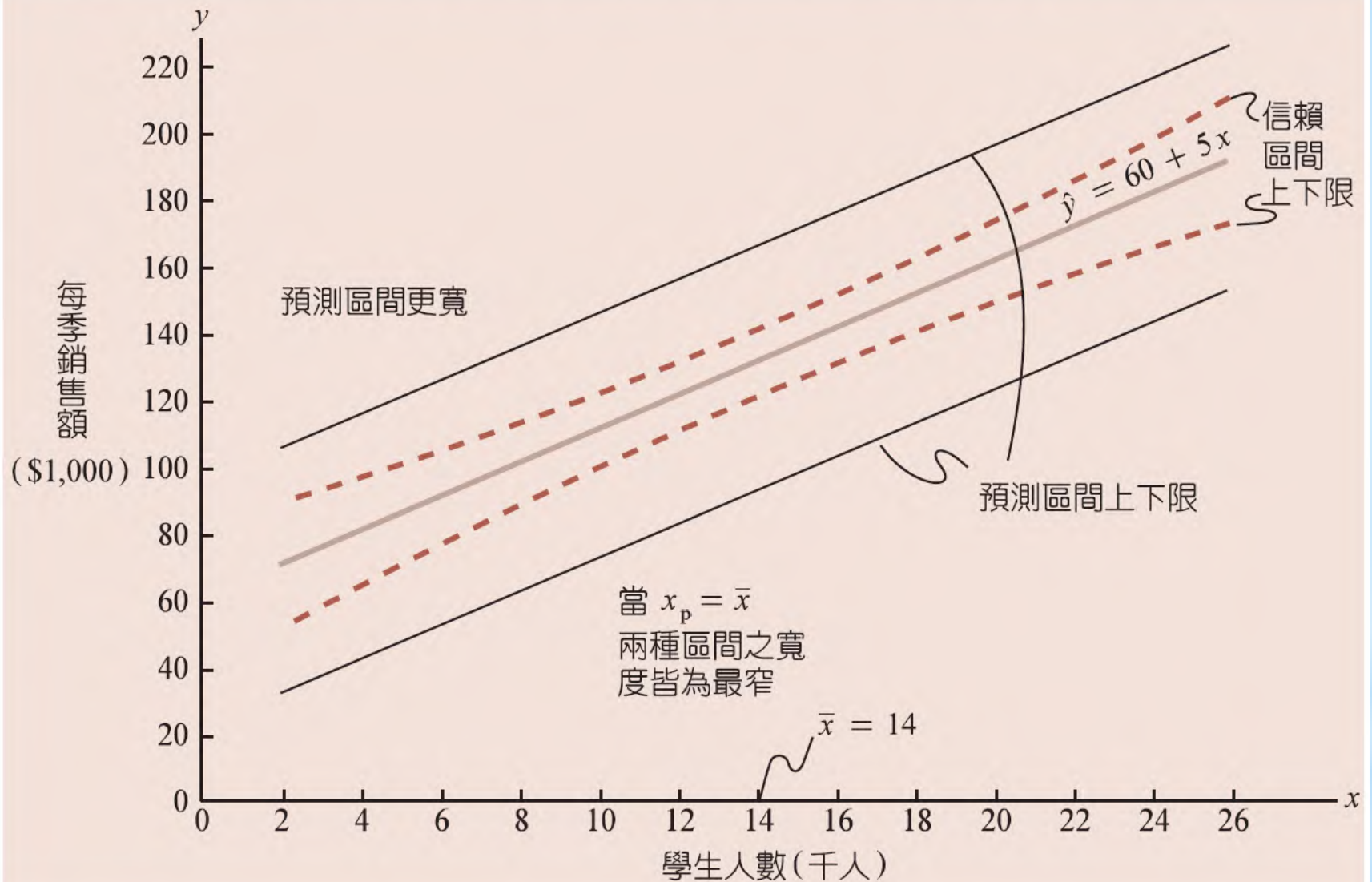
$$Y_{h(\text{new})} \sim N(\beta_0 + \beta_1 x_h, \sigma^2)$$

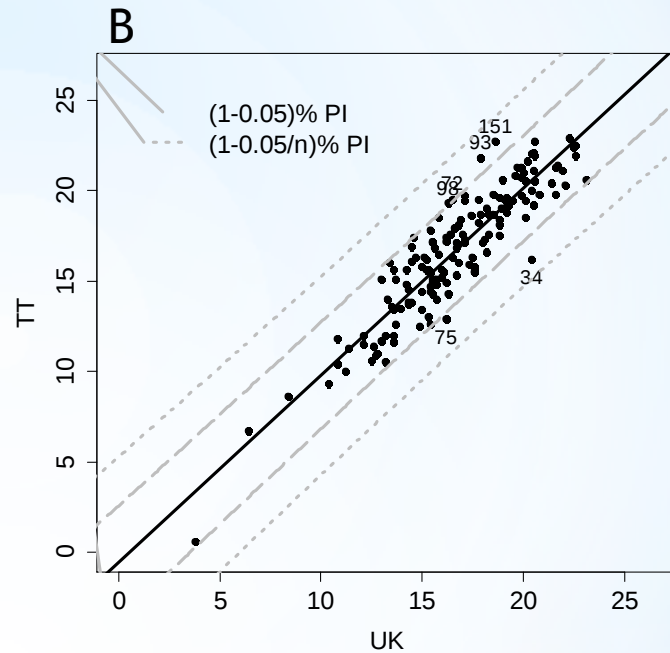
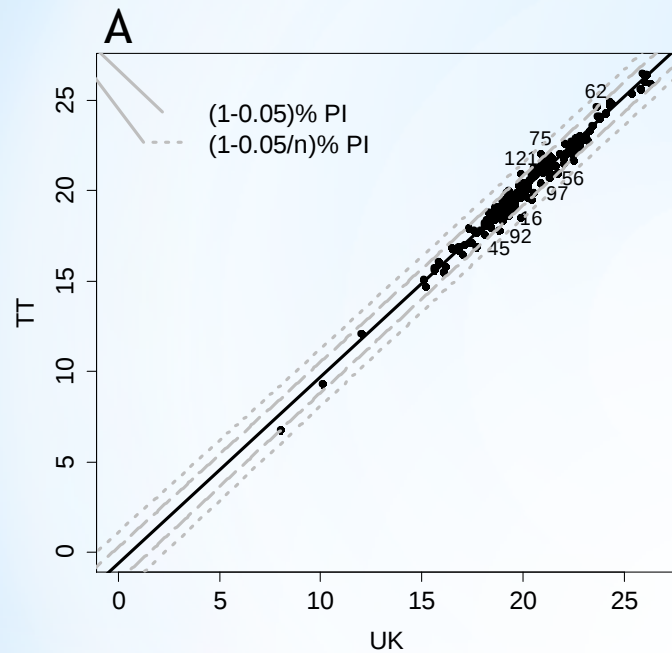
$$Y_{h(\text{new})} - Y_h \sim N(0, \sigma^2 + V(Y_h))$$

$$V(Y_{h(\text{new})} - Y_h) = \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]$$

$$\frac{Y_{h(\text{new})} - Y_h}{s\{Y_{h(\text{new})} - Y_h\}} \sim t_{(n-2)}$$

* 新觀測值的預測區間
(prediction interval)





* 上圖分別顯示A測站和B測站資料，由95%預測區間可偵測出各有8個(第16、45、56、62、75、92、97、121)和6個(第34、72、75、93、98、151)歧異點

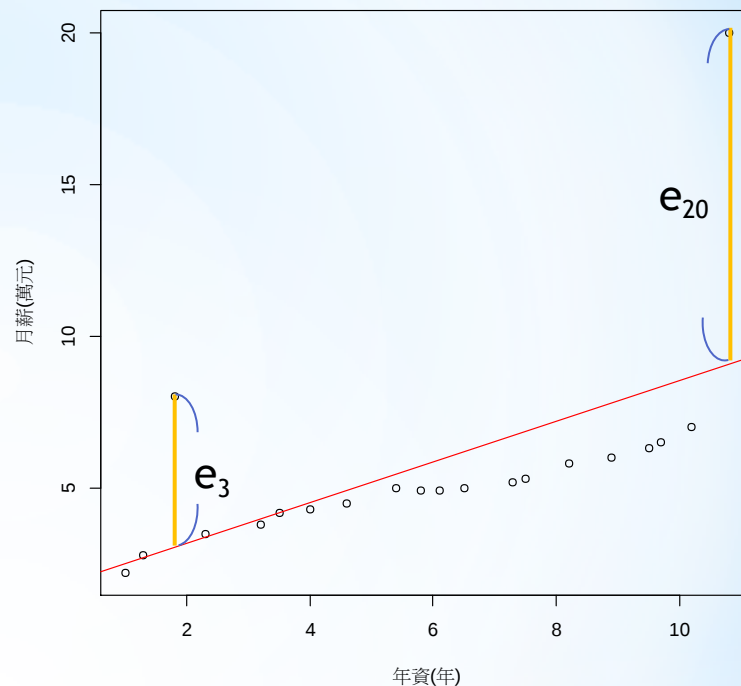
* 殘差(residual)

$$e_i = TT_i - \hat{TT}_i$$

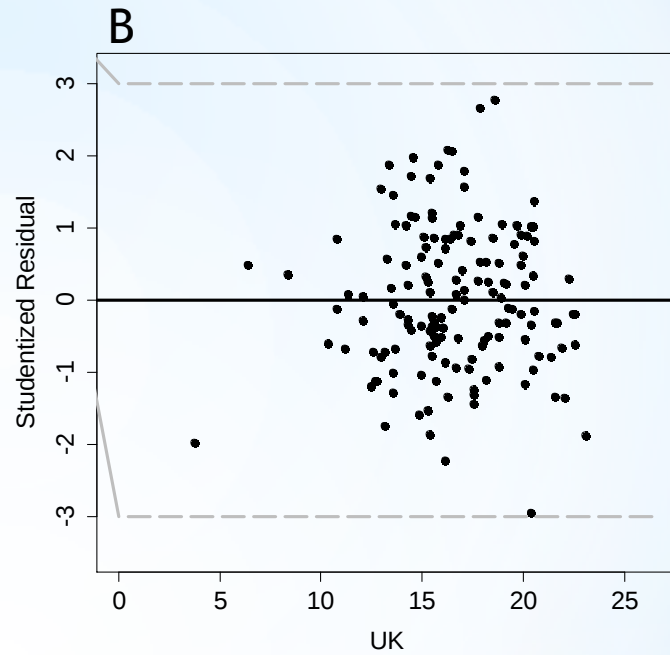
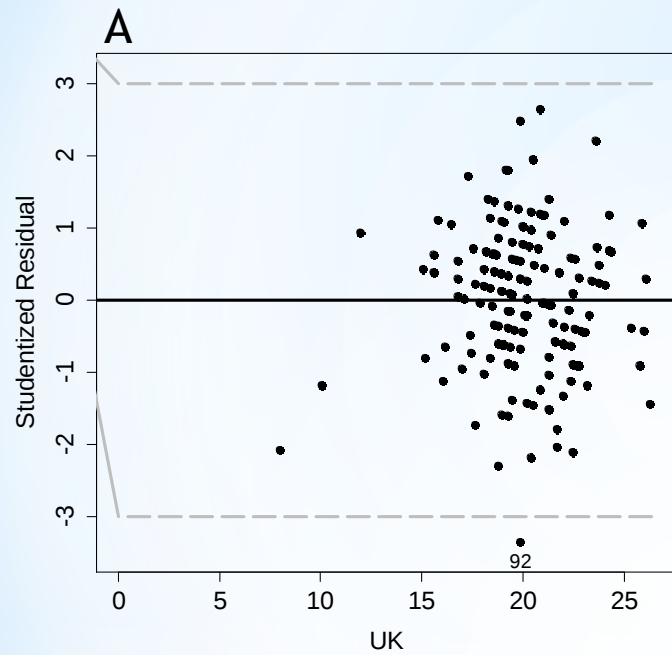
* Student 化殘差

$$r_i = \frac{e_i}{s(e_i)} = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$$

* h_{ii} 為帽子矩陣 中對角線的第 i 個元素，代表第 i 個資料點的解釋變數值與所有資料點解釋變數平均值間的距離差異。



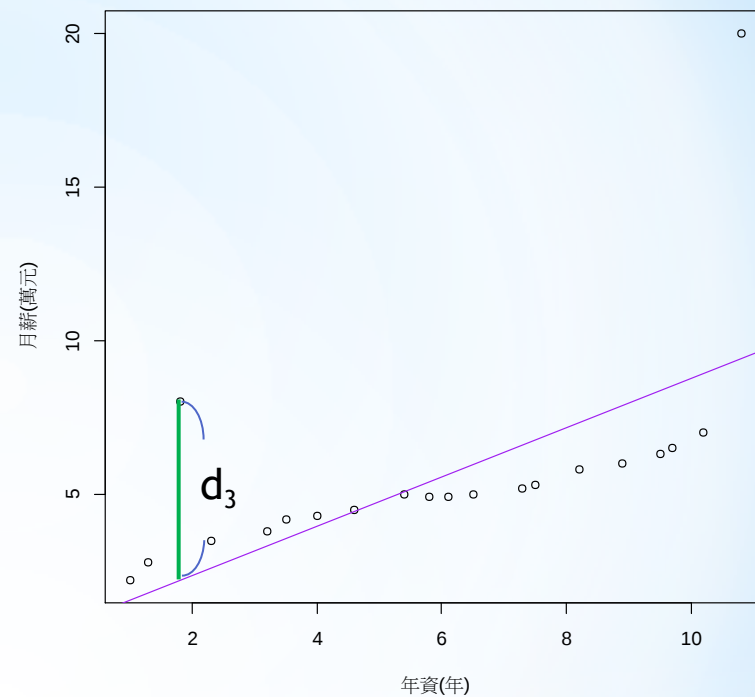
* student 化殘差



* student 化殘差

* 去點殘差(deleted residual)

$$d_i = TT_i - TT_{i(i)}$$



* 去點student 化殘差

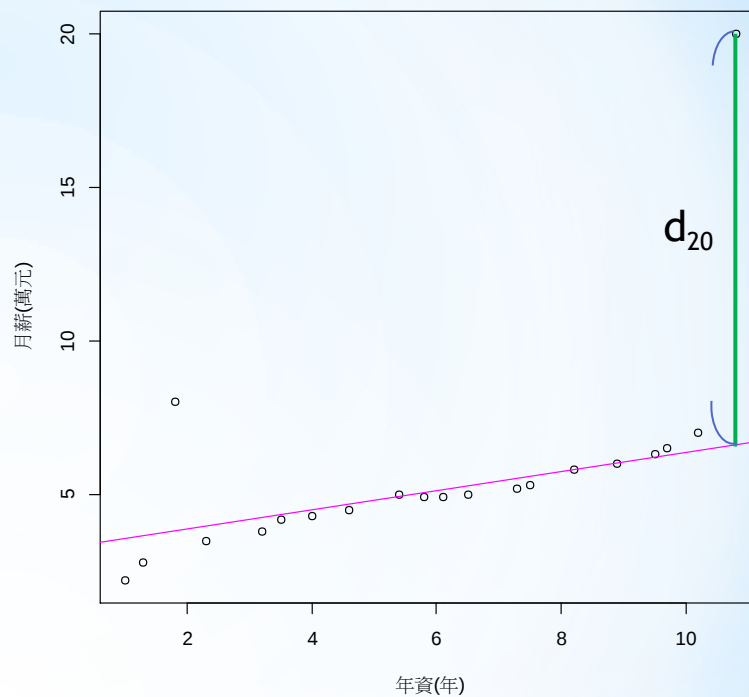
* 去點殘差(deleted residual)

$$d_i = TT_i - TT_{i(i)}$$

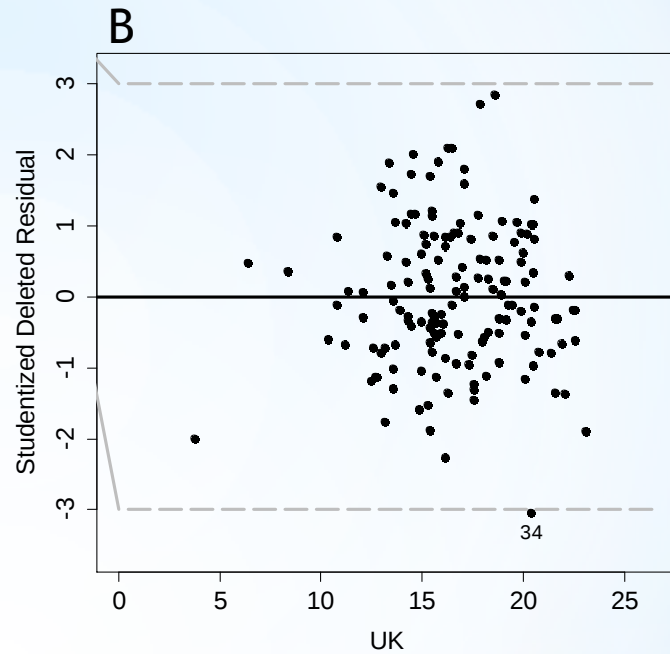
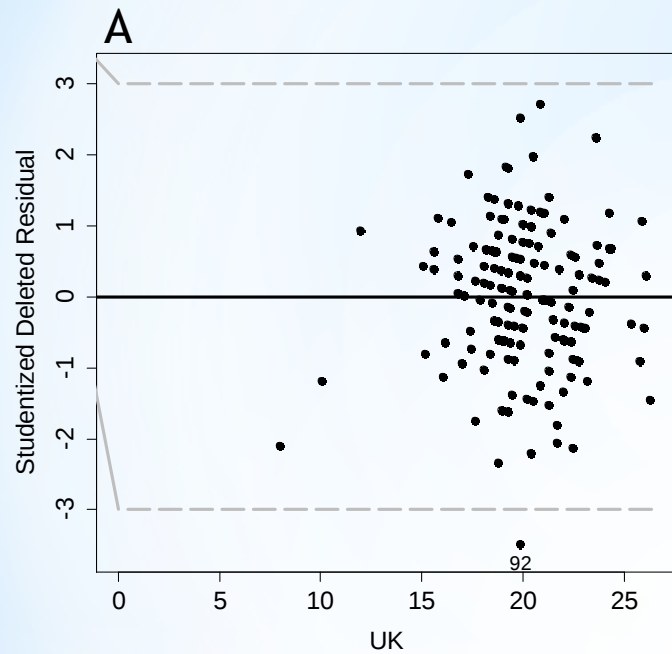
* 去點 student 化殘差

$$t_i = \frac{d_i}{s(d_i)} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}$$

$$t_i \sim t((n-1)-p)$$



* 去點student 化殘差



* 去點student 化殘差

* DFFITS

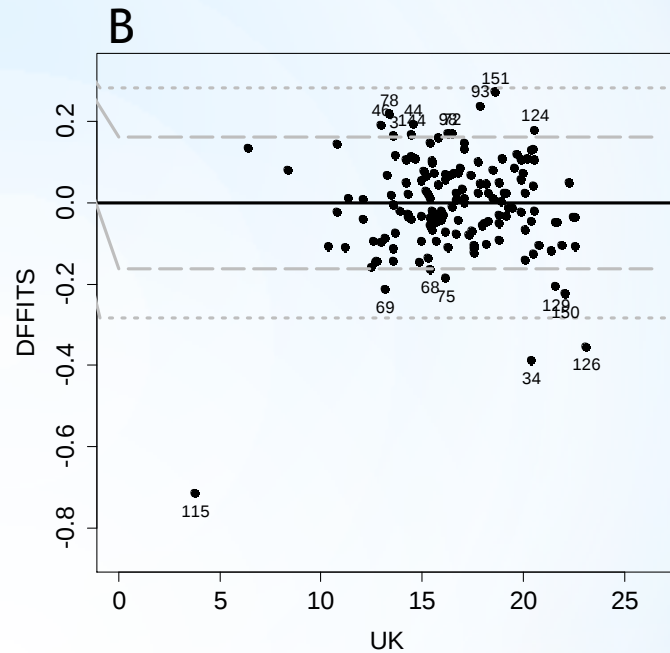
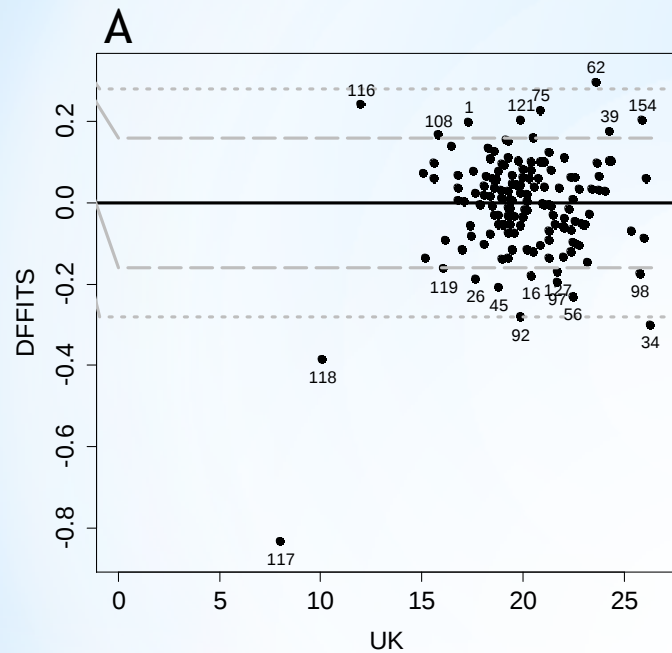
- * 比較剔除與不剔除第*i*個個案，對本身fitted value 是否有影響

$$DFFITS_i = \frac{\hat{T}T_i - \hat{T}T_{i(i)}}{\sqrt{MSE_{(i)} \times h_{ii}}} = t_i \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2}$$

Fitted value $\hat{T}T_i$
的估計標準差

去點student 化殘差

* DFFITS



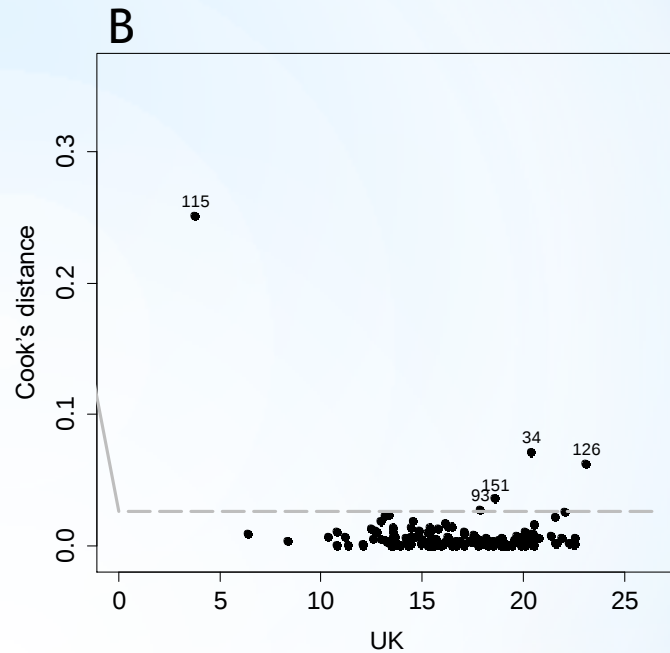
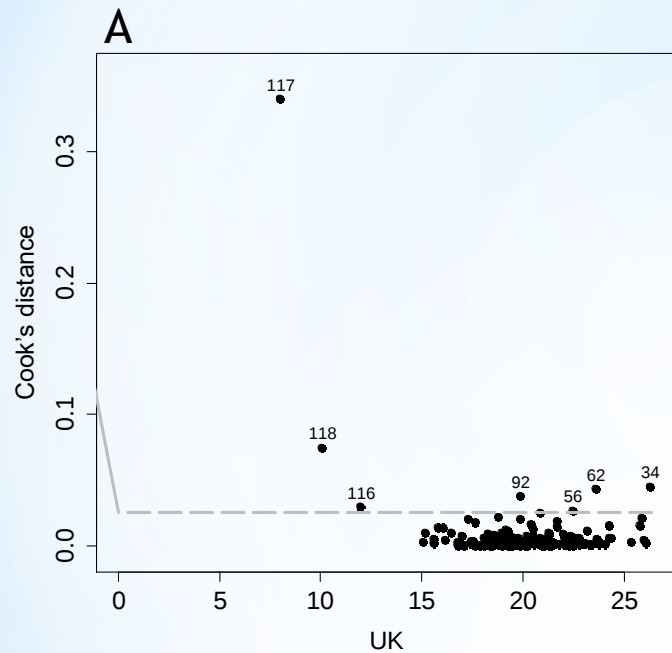
上圖代表A與B兩測站由DFFITS法偵測歧異點的情況，其中長虛線為截斷值正負，分別偵測出20個(第1、16、26、34、39、45、56、62、75、92、97、98、108、116、117、118、119、121、127、154)和18個歧異點(第3、34、44、46、68、69、72、75、78、93、98、115、124、126、129、144、150、151)，短虛線為截斷值正負，分別偵測出5個(第34、62、92、117、118)和3個歧異點(第34、115、126)。

*Cook距離

* 考慮第*i*個個案對所有*n*個個案配適值(fitted)之影響性

$$D_i = \frac{\sum_{j=1}^n \left(\hat{T}T_j - \hat{T}T_{j(i)} \right)^2}{MSE \times p} = \frac{e_i^2}{pMSE} \left(\frac{h_{ii}}{(1-h_{ii})^2} \right)$$

*Cook's distance



- * 上圖分別代表A與B兩測站由Cook距離法截斷值 $4/n$ 偵測歧異點的情況，分別偵測出7個(第34、56、62、92、116、117、118)與5個歧異點(第34、93、115、126、151)，若以截斷值1或F分配自由度 $(k+1, n-k-1)$ 的中位數來判斷，則皆無偵測到歧異點。