溫度資料檢覈之離群值偵測方法的回顧與探討

陳翠玲、薛宏宇、鄭克聲、陳雲蘭

October 4, 2016

目的：資料品質檢覈，供給高品質資料

- 資料內容包含：單一測站溫度觀測資料（TT）與克利金內插估計值（UK）、NWP 產生的估計值（MOS、INI、DA、EA1）
- 利用 UK 與 NWP 產生的估計值與 TT 有高度相關的資料，利用線性迴歸來檢覈資料
- 回顧搜尋離群值的統計方法
  - Frequentist（頻率域）：利用殘差值的機率分布發展的方法
  - Bayesian（貝氏域）：利用貝氏理論的估計分布（Predictor Distribution）的方法

What data looks like (hr_tt_03mm_12hr_C0A590)

	date	TT	UK	INI	MOS	DA	EA1
[1,]	20110301	16.2	17.0	16.9	16.9	15.7	17.0
[2,]	20110302	9.9	10.3	12.5	11.8	11.1	12.3
[3,]	20110303	10.6	11.1	12.5	11.5	10.8	12.3
[4,]	20110304	9.0	9.7	11.0	10.1	9.6	11.1
[5,]	20110305	17.9	17.5	17.1	16.6	15.6	17.2
[6,]	20110306	17.4	16.6	20.7	19.8	20.3	21.5
[7,]	20110307	8.4	9.6	10.6	9.5	8.7	10.3
[8,]	20110308	12.5	12.0	13.6	13.4	12.5	14.1
[9,]	20110309	19.2	18.0	17.2	16.9	16.0	17.7
[10,]	20110310	12.5	12.4	14.1	13.0	12.7	14.1
[11,]	20110311	11.7	11.5	13.1	11.8	11.7	13.2
[12,]	20110312	14.1	14.8	17.7	16.5	16.5	18.0
[13,]	20110313	19.8	18.1	20.0	20.0	18.5	20.6
[14,]	20110314	22.5	23.2	23.2	23.3	21.6	23.5
[15,]	20110315	12.6	12.6	14.4	14.1	12.7	14.5
[16,]	20110316	10.2	10.7	11.5	11.0	10.1	12.0

資料中間忽略

[176,]	20160321	15.7	15.7	16.5	14.5	14.5	16.9
[177,]	20160322	20.7	20.7	20.7	20.7	19.0	21.2
[178,]	20160323	15.9	15.8	17.3	15.5	15.7	17.7
[179,]	20160324	10.7	10.7	11.4	9.8	9.6	11.6
[180,]	20160325	9.8	9.7	10.8	9.5	9.3	11.2
[181,]	20160326	13.6	13.6	13.2	13.1	12.2	14.0
[182,]	20160327	12.6	12.5	13.9	13.7	12.9	14.6
[183,]	20160328	16.6	16.6	15.6	15.6	14.8	16.5
[184,]	20160329	21.3	21.3	19.7	20.6	19.0	20.4
[185,]	20160330	21.1	21.1	21.8	21.0	22.1	23.2
[186,]	20160331	20.0	20.0	20.7	20.0	19.6	21.0

- Taking hr_tt_03mm_12hr_C0A590 as example:
  - 03 represents March
  - 12 represents 12 o'clock
  - C0A590 represents Station code
- Having 12X24XN of such data, ranging from 2011 to 2016

What data looks like (hr_tt_03mm_12hr_C0A590)

	date	TT	UK	INI	MOS	DA	EA1
[1,]	20110301	16.2	17.0	16.9	16.9	15.7	17.0
[2,]	20110302	9.9	10.3	12.5	11.8	11.1	12.3
[3,]	20110303	10.6	11.1	12.5	11.5	10.8	12.3
[4,]	20110304	9.0	9.7	11.0	10.1	9.6	11.1
[5,]	20110305	17.9	17.5	17.1	16.6	15.6	17.2
[6,]	20110306	17.4	16.6	20.7	19.8	20.3	21.5
[7,]	20110307	8.4	9.6	10.6	9.5	8.7	10.3
[8,]	20110308	12.5	12.0	13.6	13.4	12.5	14.1
[9,]	20110309	19.2	18.0	17.2	16.9	16.0	17.7
[10,]	20110310	12.5	12.4	14.1	13.0	12.7	14.1
[11,]	20110311	11.7	11.5	13.1	11.8	11.7	13.2
[12,]	20110312	14.1	14.8	17.7	16.5	16.5	18.0
[13,]	20110313	19.8	18.1	20.0	20.0	18.5	20.6
[14,]	20110314	22.5	23.2	23.2	23.3	21.6	23.5
[15,]	20110315	12.6	12.6	14.4	14.1	12.7	14.5
[16,]	20110316	10.2	10.7	11.5	11.0	10.1	12.0

資料中間忽略

[176,]	20160321	15.7	15.7	16.5	14.5	14.5	16.9
[177,]	20160322	20.7	20.7	20.7	20.7	19.0	21.2
[178,]	20160323	15.9	15.8	17.3	15.5	15.7	17.7
[179,]	20160324	10.7	10.7	11.4	9.8	9.6	11.6
[180,]	20160325	9.8	9.7	10.8	9.5	9.3	11.2
[181,]	20160326	13.6	13.6	13.2	13.1	12.2	14.0
[182,]	20160327	12.6	12.5	13.9	13.7	12.9	14.6
[183,]	20160328	16.6	16.6	15.6	15.6	14.8	16.5
[184,]	20160329	21.3	21.3	19.7	20.6	19.0	20.4
[185,]	20160330	21.1	21.1	21.8	21.0	22.1	23.2
[186,]	20160331	20.0	20.0	20.7	20.0	19.6	21.0

- length of TT observation started from 2011 to current year 2016
- 簡測溫度觀測值是否有異，從其它可靠的溫度預報方法來檢驗之。
- 使用 regression model 來找 residual 的游離值，進而找出溫度觀測是有有異。
- 終極目標是，利用這 5 個溫度推估法進行的 regression model 所產生的至少 5 組 residuals 值 (偵測游離值不只使用 residual 本身的機率分布一方法)，來偵測 TT 的 outlier

What data looks like (hr_tt_03mm_12hr_C0A590)

	date	TT	UK	INI	MOS	DA	EA1
[1,]	20110301	16.2	17.0	16.9	16.9	15.7	17.0
[2,]	20110302	9.9	10.3	12.5	11.8	11.1	12.3
[3,]	20110303	10.6	11.1	12.5	11.5	10.8	12.3
[4,]	20110304	9.0	9.7	11.0	10.1	9.6	11.1
[5,]	20110305	17.9	17.5	17.1	16.6	15.6	17.2
[6,]	20110306	17.4	16.6	20.7	19.8	20.3	21.5
[7,]	20110307	8.4	9.6	10.6	9.5	8.7	10.3
[8,]	20110308	12.5	12.0	13.6	13.4	12.5	14.1
[9,]	20110309	19.2	18.0	17.2	16.9	16.0	17.7
[10,]	20110310	12.5	12.4	14.1	13.0	12.7	14.1
[11,]	20110311	11.7	11.5	13.1	11.8	11.7	13.2
[12,]	20110312	14.1	14.8	17.7	16.5	16.5	18.0
[13,]	20110313	19.8	18.1	20.0	20.0	18.5	20.6
[14,]	20110314	22.5	23.2	23.2	23.3	21.6	23.5
[15,]	20110315	12.6	12.6	14.4	14.1	12.7	14.5
[16,]	20110316	10.2	10.7	11.5	11.0	10.1	12.0

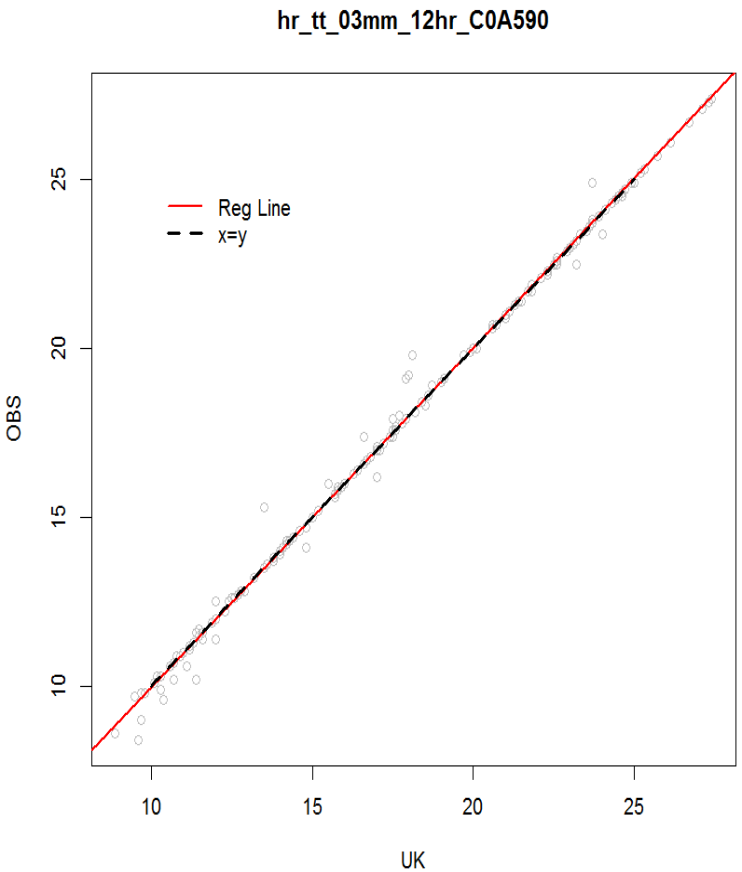
資料中間忽略

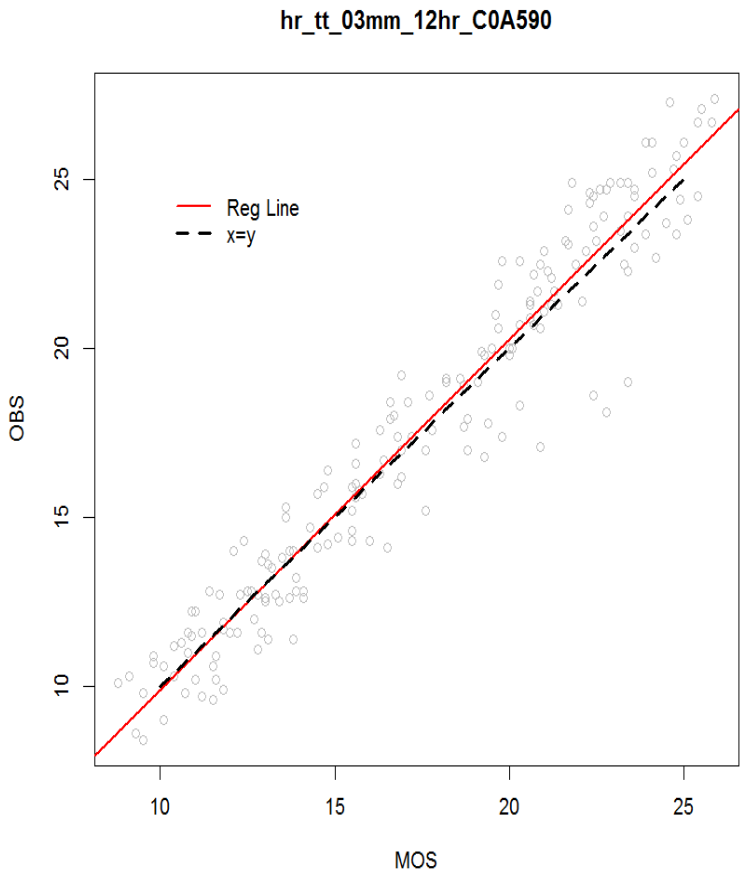
[176,]	20160321	15.7	15.7	16.5	14.5	14.5	16.9
[177,]	20160322	20.7	20.7	20.7	20.7	19.0	21.2
[178,]	20160323	15.9	15.8	17.3	15.5	15.7	17.7
[179,]	20160324	10.7	10.7	11.4	9.8	9.6	11.6
[180,]	20160325	9.8	9.7	10.8	9.5	9.3	11.2
[181,]	20160326	13.6	13.6	13.2	13.1	12.2	14.0
[182,]	20160327	12.6	12.5	13.9	13.7	12.9	14.6
[183,]	20160328	16.6	16.6	15.6	15.6	14.8	16.5
[184,]	20160329	21.3	21.3	19.7	20.6	19.0	20.4
[185,]	20160330	21.1	21.1	21.8	21.0	22.1	23.2
[186,]	20160331	20.0	20.0	20.7	20.0	19.6	21.0

相關於 UK、NWP 方法應用與原理，請參考本研討會文章 A1-25: 應用數值預報模式增強氣溫觀測資料偵錯研判分析 (薛宏宇、呂致穎、陳翠玲)

Taking UK/MOS vs TTobs as example

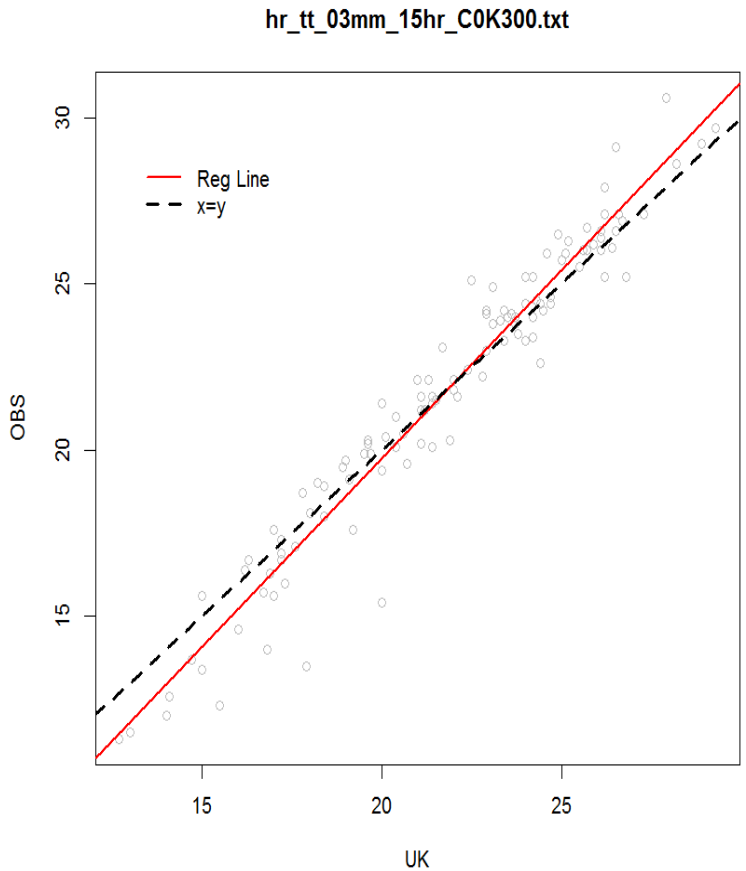
Consider detecting influence point from residual of a linear regression model

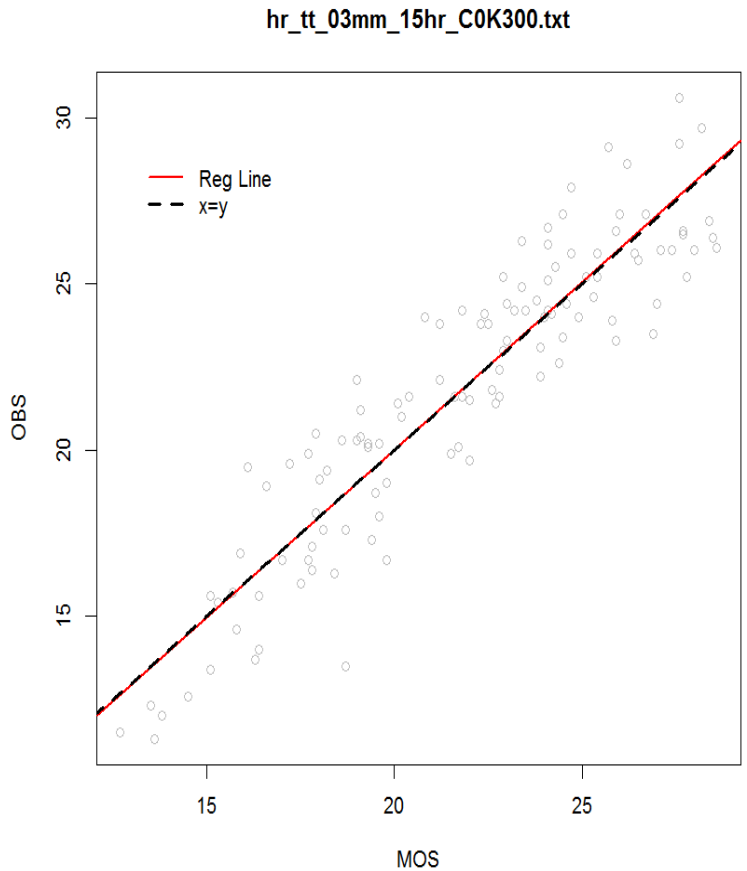




Another example with different station

Consider detecting influence point from residual of a linear regression model





Conventional outlier methods : Frequentist and Bayesian

Frequentist

F1 standardized residuals $r_i = \hat{\varepsilon}_i / s\sqrt{1 - h_{ii}}$

F2 predictive interval (also could be categorized in Bayesian)

F3 externally standardized residual $t_i = \hat{\varepsilon}_i / s_{(i)}\sqrt{1 - h_{ii}}$

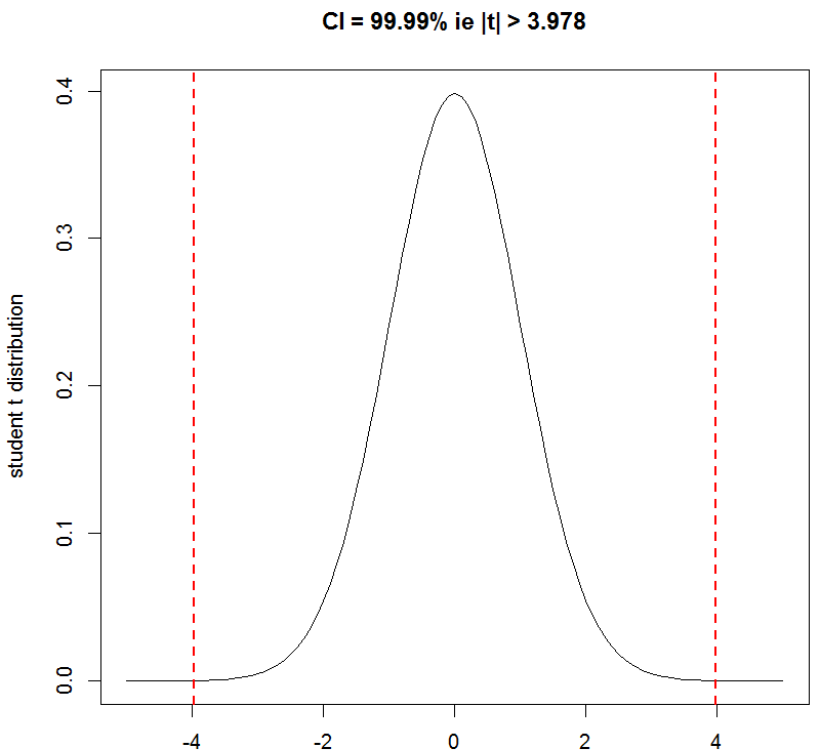
Bayesian

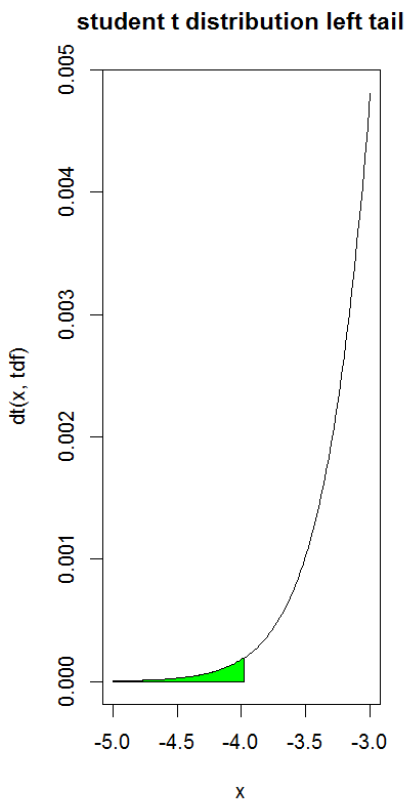
B1 predictive method based on predictive distribution, $c_i = p(y_i | y_{(i)})$

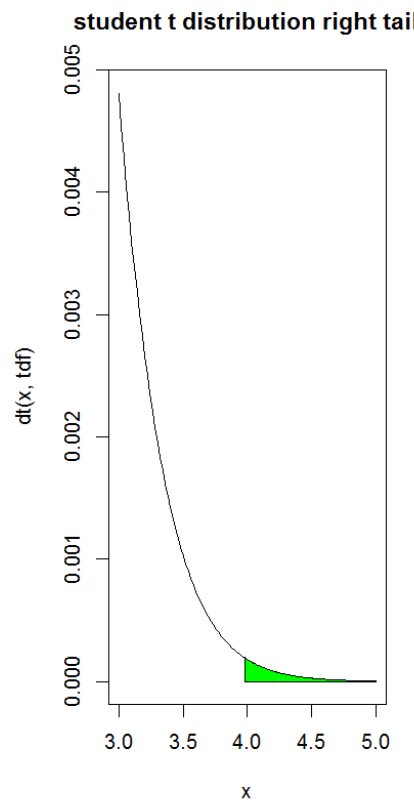
B2 outlier detection based on assuming improper prior distribution of normal-gamma prior, $p_i = pr(|\varepsilon_i| > k\sigma | y)$

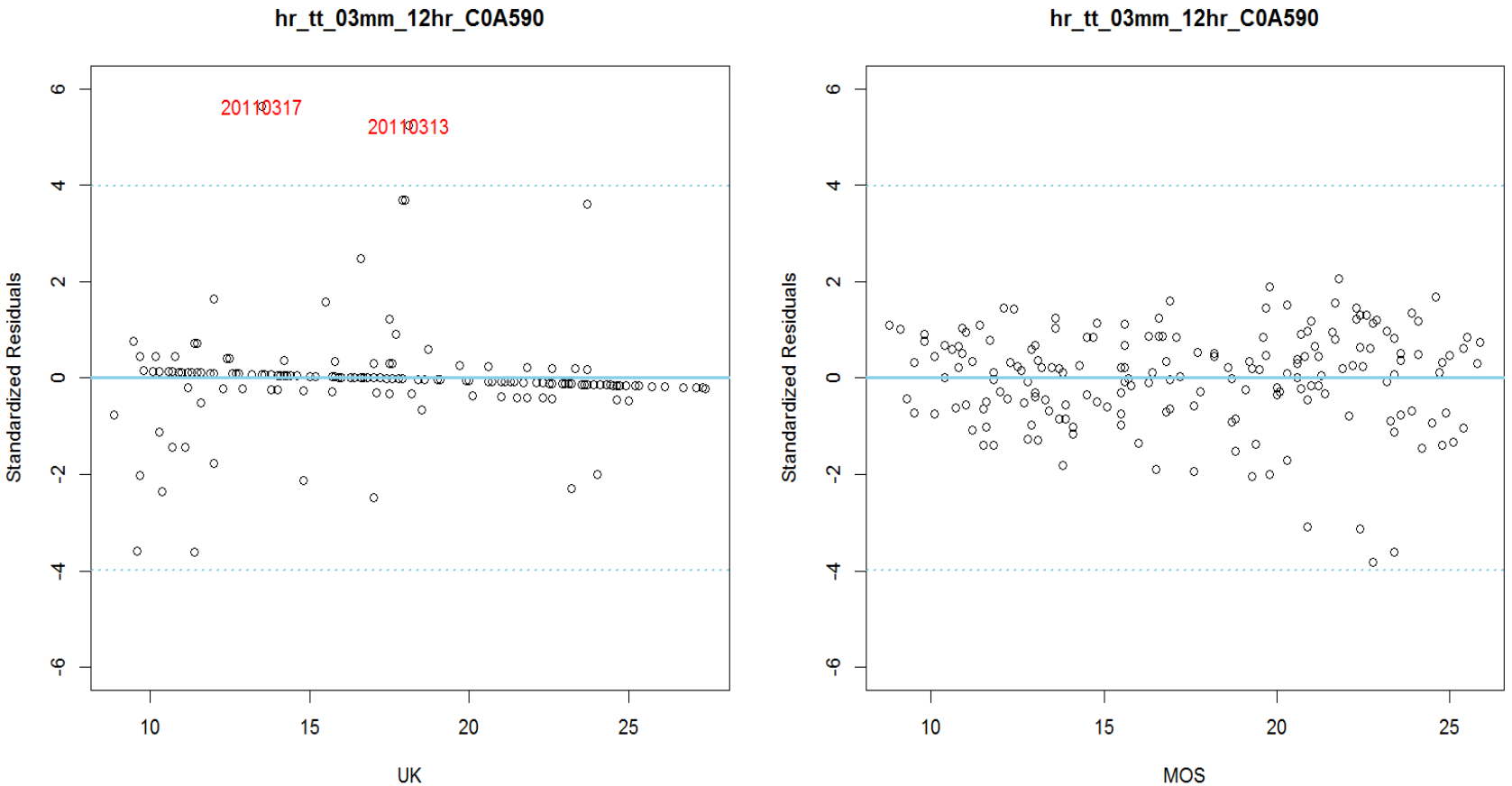
$$r_i \sim t(n-m-1)$$

for n is sample size and m is the number of parameters, in our case, $m = 2$. In this case, we consider anything outside the region of CI $= 99.99\%$ is outlier.

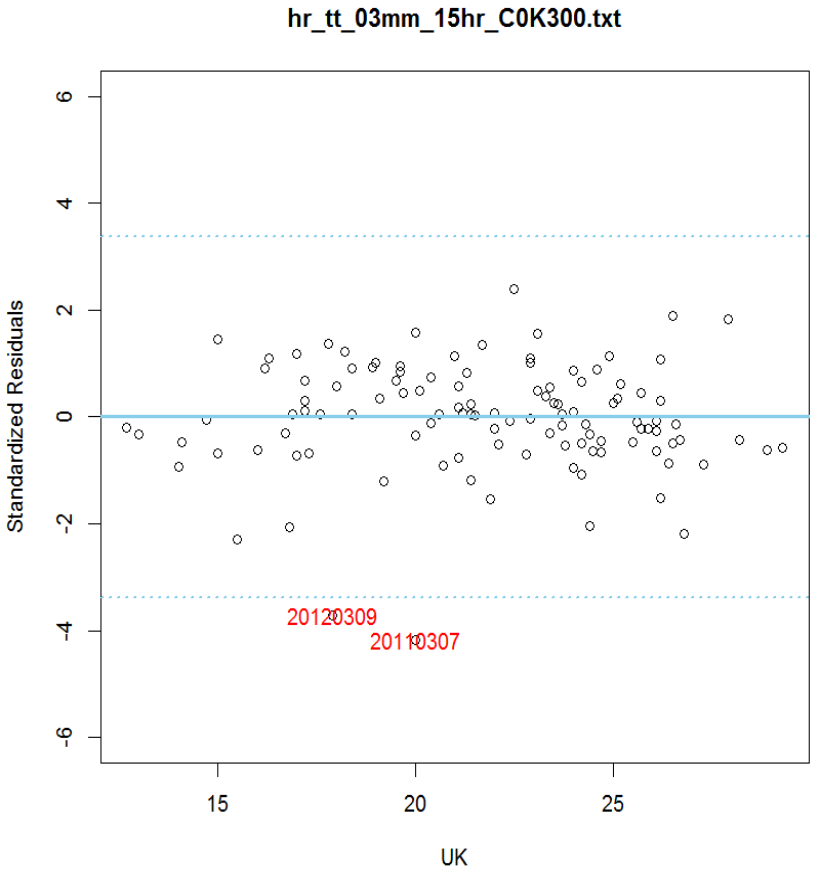


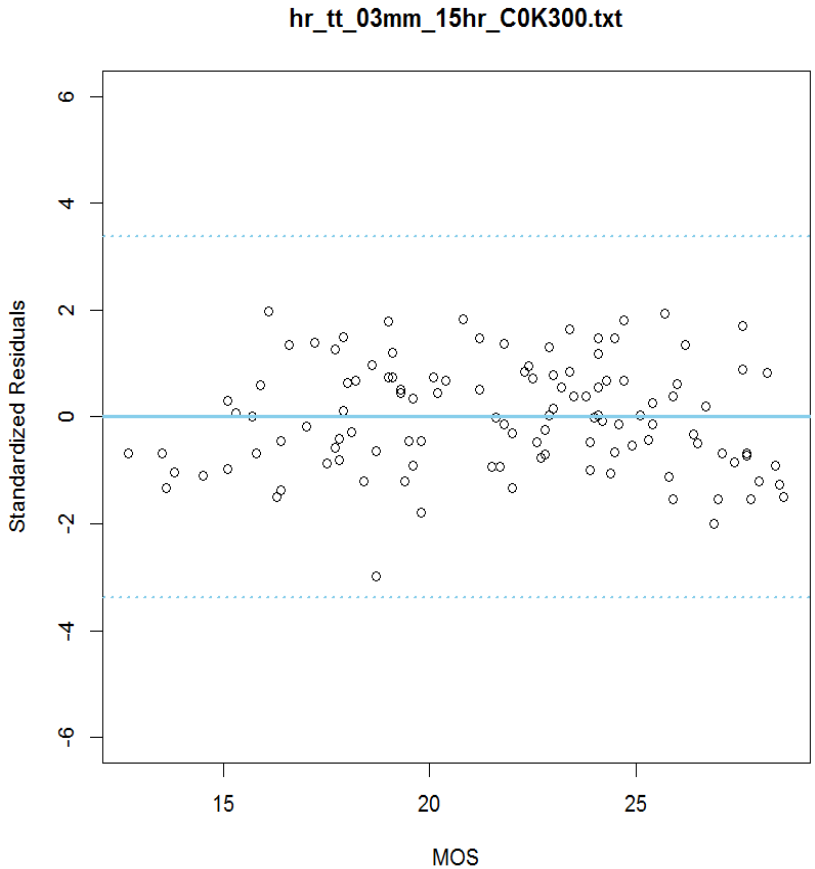






Different TT-predicted models carry out different residual distribution. On this regard, this indicates and highlights the importance of priori.





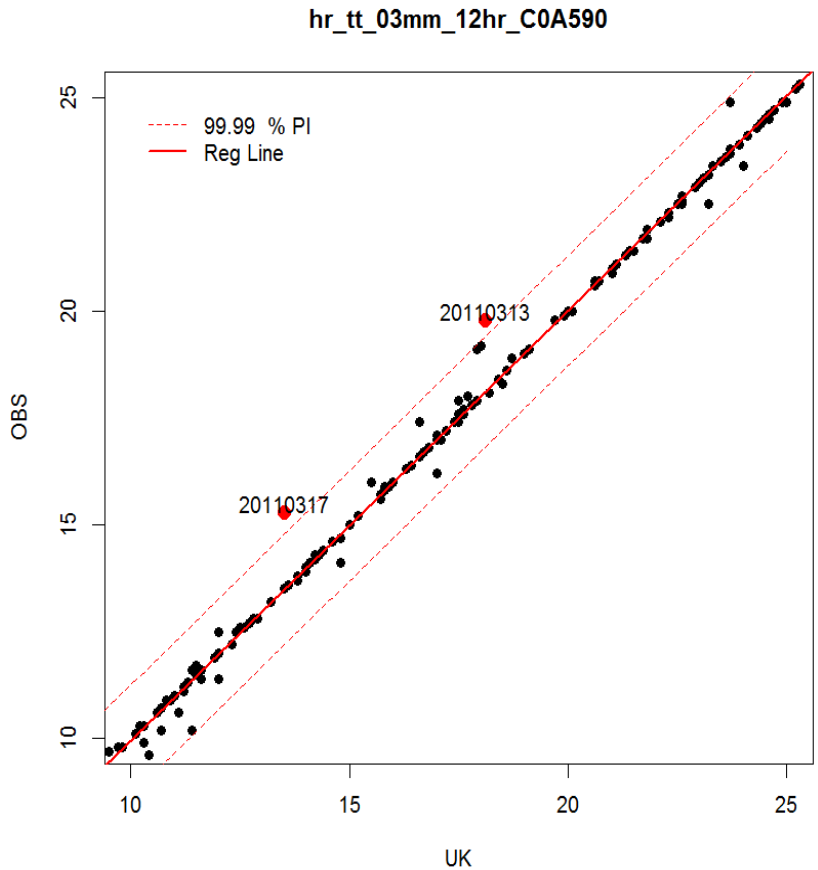
This one use 99.9% CL to detect outliers.

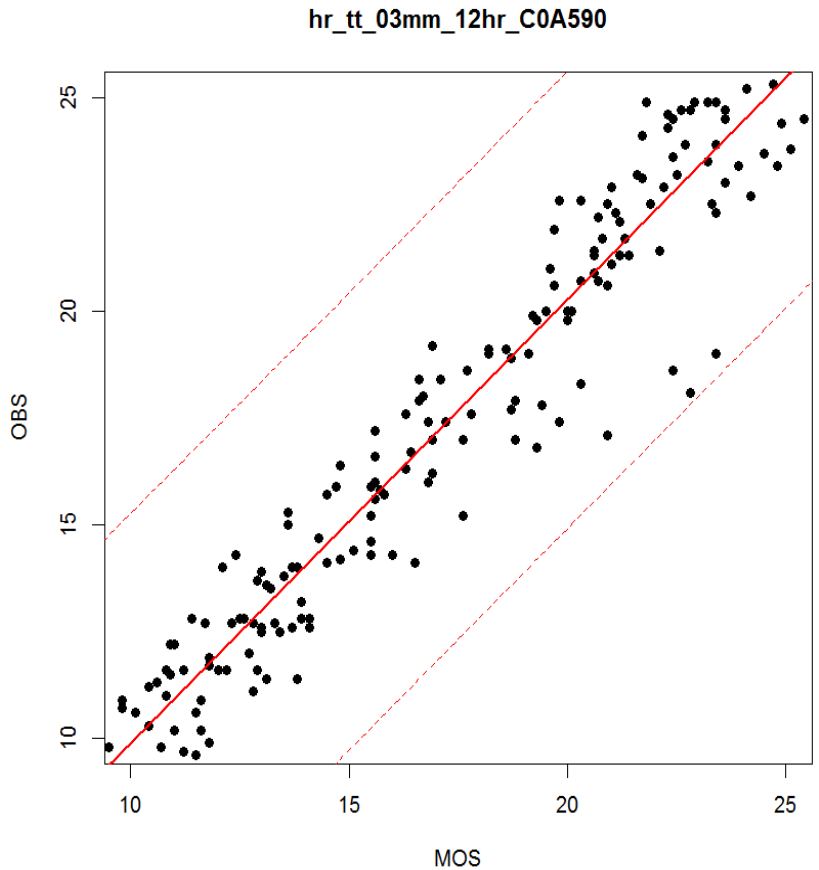
A predictive interval bears the same relationship to a future observation that a Frequentist CI.

- (x_i, y_i) be n observations that follow the normal simple linear regression model
$$y_i|x_i \sim N(a + bx_i, \sigma^2)$$
- the point predictor for a future observation y made at x^* is $\hat{\mu}_{y|x^*} = \hat{a} + \hat{b}x^*$
- $(1-\alpha)100\%$PI is
$$\hat{\mu}_{y|x^*} \pm t_{\alpha/2,n-2} S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)S_x^2}}$$

where
$$s = \sqrt{\frac{1}{n-2}\sum(y_i - \hat{y}_i)^2} = \sqrt{\frac{1}{n-2}\left[\sum(y-\bar{y})^2 - \frac{[\sum(x-\bar{x})(y-\bar{y})]^2}{\sum(x-\bar{x})^2}\right]}$$

Predictive Interval derived from distribution of r_i





Predictive interval using distribution of r_i to obtain the distribution of 'supposed' new data. Therefore, you have exactly the same result from r_i.

Let $y_{(i)}$ be the observations excluding y_i, the external standardized residual is defined by the following:

$$t_i = \hat{\varepsilon}_i / s_{(i)} \sqrt{1 - h_{ii}}$$

where $s_{(i)}$ is estimated variance of $y_{(i)}$. And

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

under null hypothesis of no shift in the mean of the ith observation.

Detect Influential points Result from t_i

For UK
My external standardized residual result with cutoff p-value at 0.00005

date	ti	pvalue	BonferonniP
20110313	5.66	2.9e-8	5.4e-6
20110317	6.20	1.8e-9	3.4e-7

result from R function outlierTest{car}

date	ti	pvalue	BonferonniP
20110313	5.66	5.7e-8	1.06e-5
20110317	6.20	3.7e-9	6.97e-7

For MOS
My external standardized residual result with cutoff p-value at 0.00005

date	ti	pvalue	BonferonniP
20120315	-3.96	5.3e-5	0.0098

result from R function outlierTest{car}
No Studentized residuals with Bonferonni p $<$ 5e-05
Largest |rstudent|:

date	ti	pvalue	BonferonniP
20120315	-3.96	1.06e-4	0.020

External standardized residual t_i doesn't supply anything distinguishably different from r_i

Started from Bayesian...

Assume that if the priori is essential, the joint density for potential data y and parameter θ can be derived from the following:

$$p(y, \theta|A) = p(y|\theta, A)p(\theta|A)$$

The posterior probability of θ is

$$p(\theta|y, A) = \frac{p(y|\theta, A) \cdot p(\theta|A)}{\int p(y|\theta, A) \cdot p(\theta|A)d\theta}.$$

That

$$p(y|A) = \int p(y|\theta, A) \cdot p(\theta|A)d\theta$$

is the predictive distribution.

Suppose $y = (y_1, y_2)$ for subsets y_1 and y_2, then we may have

$$p(y|A) = p(y_2|y_1, A)p(y_1|A)$$

If model is crooked because of some outliers, say y_2, then make probability of y_2 conditional on y_1 to see if the predicted y_2 is adhered to the model based on y_1.
Thus, we may have

$$c_i = \frac{p(y)}{p(y_{(i)})} = p(y_i|y_{(i)})$$

(Box, 1980)

Assume $y \sim N(\theta, \sigma^2)$ with unknown θ and θ has priori with $\theta \sim N(\theta_0, \sigma_\theta^2)$, then

$$p(\theta|y, A) = \frac{1}{\sqrt{2\pi}} \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_\theta^2} \right)^{1/2} \exp \left\{ -\frac{1}{2} \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_\theta^2} \right) (\theta - \bar{\theta})^2 \right\}$$

where

$$\bar{\theta} = \frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\sigma_\theta^2}\theta_0}{\frac{n}{\sigma^2} + \frac{1}{\sigma_\theta^2}}$$

And predictive distribution is

$$p(y|A) = \frac{1}{\sqrt{n}\sqrt{2\pi}^{n}} \frac{1}{\sigma^{n-1}} \frac{1}{\tilde{\sigma}} \exp\left\{ -\frac{1}{2} \left[\frac{(n-1)s^2}{\sigma^2} + \frac{(\bar{y}-\theta_0)^2}{\tilde{\sigma}^2} \right] \right\}$$

where

$$\tilde{\sigma} = \left(\frac{\sigma^2}{n} + \sigma_\theta^2 \right)^{1/2}$$

Outlier detection through predictive distribution(Geisser 1980, 1987)

Given observation $y_{(i)}$ with its predictive distribution $p(y_{(i)}|A)$ and let c_i be the conditional predictive ordinate, which is defined by the following:

$$c_i = p(y_i|y_{(i)}) = \frac{p(y|A)}{p(y_{(i)}|A)}$$

The most discordant having the smallest value of c_i. Recall variance part of $p(y|A)$, let $g(y_d) = \frac{(n-1)s_d^2}{\sigma^2} + \frac{(\bar{y_d}-\theta_0)^2}{\tilde{\sigma}^2}$

$$g(y_d) \sim \chi^2(n)$$

Although this doesn't come with the p-value or any mechanism to identify outlier, we can plunder the variance part of $p(y|A)$ which has χ^2 distribution with degree of freedom, n.

實驗結果: 使用 UK/MOS 來偵測是否有 OBS 錯誤值

Under the confidence level 99.99 no outlier detect
Below are the table of statistics ci who has p-value less than 0.1

date	c_i	$g(y_d)$	p-value*
20110313	3.9e-6	157.4	0.077
20110317	4.1e-7	152.9	0.046

Under the confidence level 99.99 no outlier detect
Below are the table of statistics ci who has the smallest p-value

date	c_i	$g(y_d)$	p-value*
20120315	1.5e-4	169.4	0.243

*p-value is defined with $Pr(\chi^2_n > g(y_d))$.

- How to quantify 5 different estimator of TT
- How to ensemble the residuals from 5 different model estimator
- Different variance determine how residual behave, some are slack some are tense, then how to quantify those information
- Bayesian provides the information concerning predictor future behavior which is more reliable.
- Certain mechanism on information feedback and