統計諮詢 - 作業 1

國立成功大學統計學系暨數據科學研究所

廖傑恩（RE6094028）

2021-03-10

1 Exercise 4.2

1.1 問題敘述

研究者發現空氣中二氧化硫（SO_2）含量可能與環境或氣候有關，希望藉由分析資料了解二氧化硫含量與環境或氣候因子之間是否有關。另外也想探討環境或氣候因子這些變相彼此間是否有線性相關。

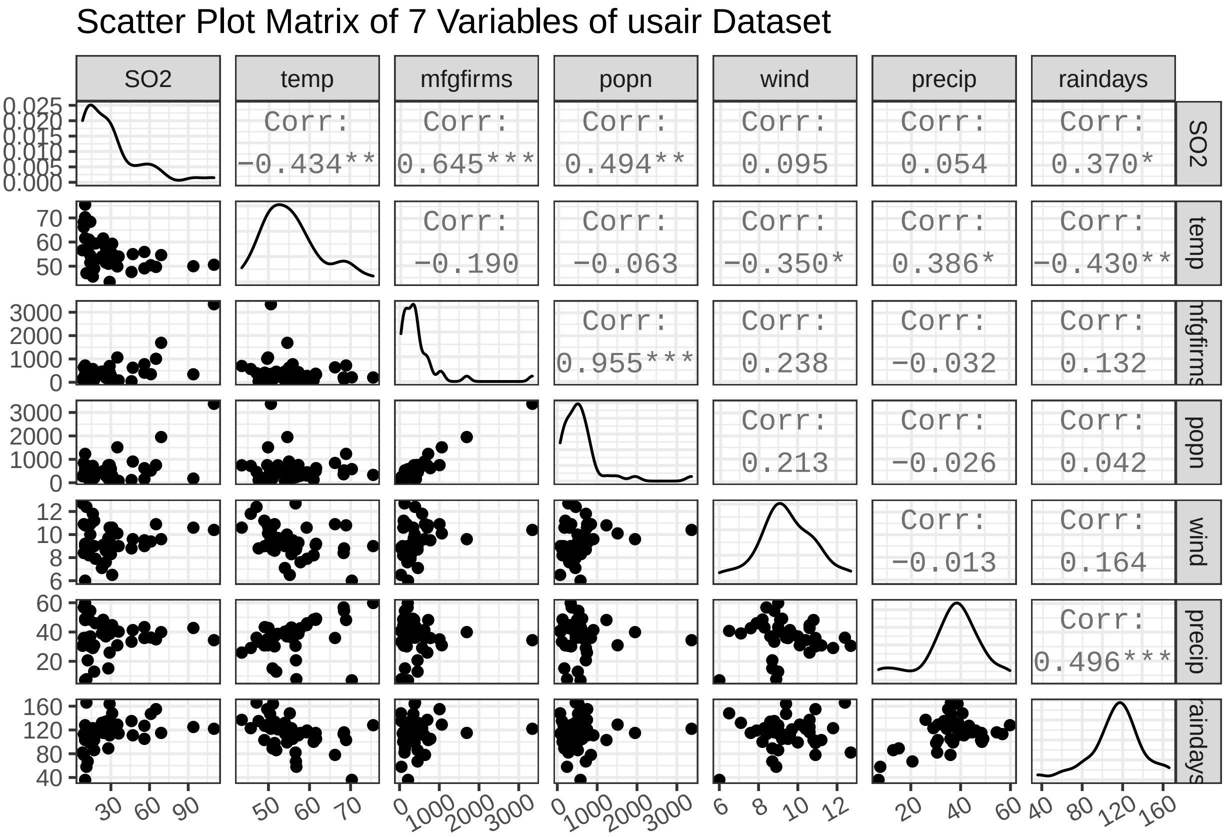
1.2 資料介紹

此份資料由 Sokal and Rohlf (1981) 與 Hand et al. (1994) 等人收集，包含了美國 41 個城市 7 個環境與氣候因子等變項，分別敘述如下：

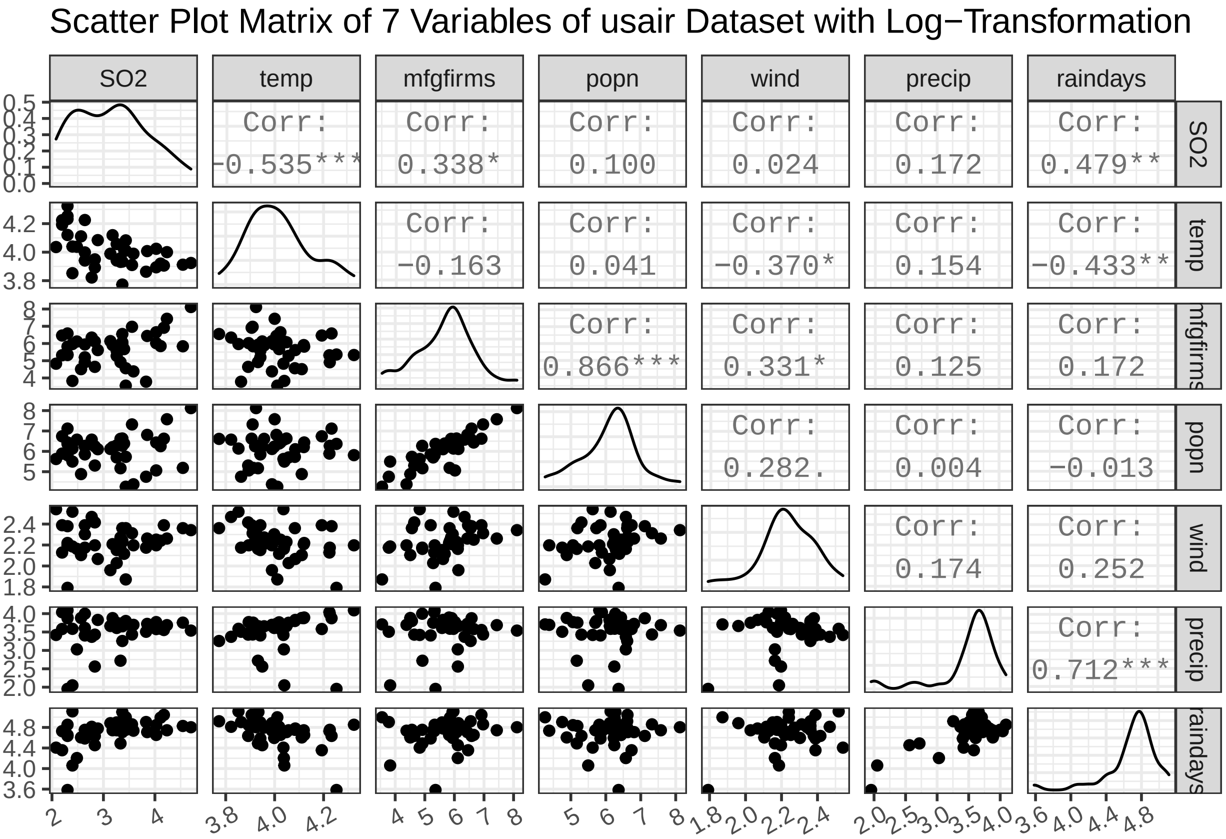
- `SO2`: 二氧化硫含量（單位：微克／立方公尺）
- `temp`: 年平均溫度（單位：華氏溫度）
- `mfgfirms`: 擁有至少 20 名員工之製造業者數量
- `popn`: 1970 年人口普查之人數（單位：千人）
- `wind`: 年平均風速（單位：英里/小時；mph）
- `precip`: 年平均降水量（單位：英吋）
- `raindays`: 年均降水天數（單位: 天）

1.3 資料分析

1.3.1 Scatter plot matrices



Scatter Plot Matrix of 7 Variables of usair Dataset



Scatter Plot Matrix of 7 Variables of usair Dataset with Log−Transformation

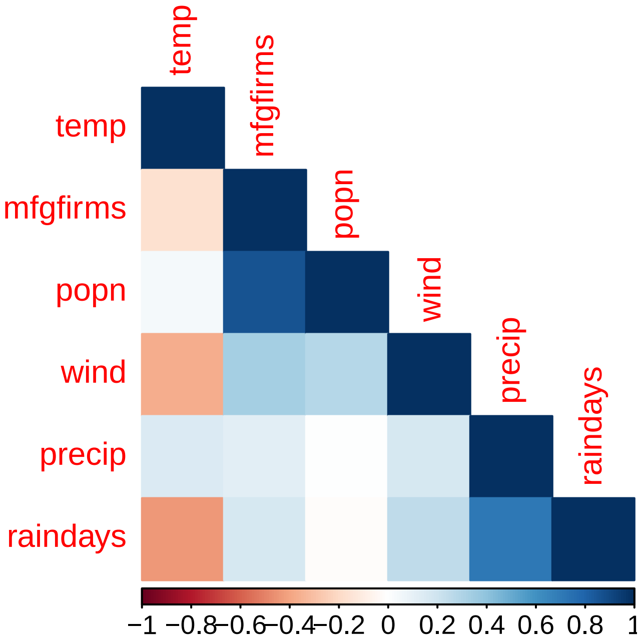
比較兩組 scatter plots，發現原始資料中某些變項尺度較大，有部分值與大部分值差距甚大的觀測點，而這些資料點會大幅影響兩兩變項間的 correlation coefficient 的計算，高估或低估相關性（通常是高估）。而將原始資料進行 log 轉換後，資料尺度縮小，極端值的影響也降低，這時計算出來的兩兩變項間的 correlation coefficient 比較能真實反應變項間的線性關聯，也比較能夠看出線性趨勢與資料散佈情形。這是為何在這份資料中，log 轉換比較適合的原因。此外，這份資料各變項皆為正值的連續變項，所以 log 轉換是合適的。

1.3.2 The most highly correlated predictors with the logged response

	與 logged.response 的相關係數
temp	-0.5465
mfgfirms	0.4843
popn	0.3434
wind	-0.0135
precip	0.0481
raindays	0.4786

上表呈現了 6 個 predictors 與 logged response 的 Pearson 相關係數（取至小數點後四位）。在這 6 個 predictors 中，與 logged response 相關程度最高者為 `temp`（華氏年均溫），呈現中度負相關，顯示溫度較高的城市可能有偏高的二氧化硫濃度。

1.3.3 Pairs of highly correlated logged predictors



上圖呈現了 6 個 predictors 兩兩之間的 Pearson 相關係數大小，顏色越深表示此兩

變項間的線性相關程度高，越接近深藍色表示線性正相關程度越高，越接近深紅色表示線性負相關程度越高。可以發現 popn 與 mfgfirms 以及 raindays 與 precip 這 2 對變項組的線性正相關程度不低，另外也有 3 對變項組呈現線性負相關。

將 15 對 predictors 的 Pearson 相關係數（取至小數點後四位）製作成下表，並依據相關係數絕對值由大到小排列。其中 High 欄標記為 Yes 者，表示其似乎為高度相關者（相關係數絕對值大於 0.7），和在前面的圖片觀察到的相同，有 2 對：popn（人口）與 mfgfirms（具規模之製造業廠商數），相關係數為 0.8659；raindays（一年中平均下雨天數）與 precip（平均年降水量），相關係數為 0.7122。

變項 1	變項 2	相關係數	是否高相關
popn	mfgfirms	0.8659	Yes
raindays	precip	0.7122	Yes
raindays	temp	-0.4331	No
wind	temp	-0.3697	No
wind	mfgfirms	0.3309	No
wind	popn	0.2816	No
raindays	wind	0.2523	No
precip	wind	0.1740	No
raindays	mfgfirms	0.1720	No
mfgfirms	temp	-0.1627	No
precip	temp	0.1536	No
precip	mfgfirms	0.1252	No
popn	temp	0.0410	No
raindays	popn	-0.0135	No
precip	popn	0.0042	No

1.4 建議

1. 此分析使用到 Pearson 相關係數只能表示兩變項之間的線性相關程度，不具備因果關係的解釋性，也無法捕捉兩變項間非線性的相關程度。

2. 若想進一步以 6 個 predictor 來預測二氧化硫，可以使用多元線性迴歸或其他機器學習方法來操作。

2 Exercise 5.13

The relative rotation angle between the L2 and L3 lumbar vertebrae is defined as the acute angle between posterior tangents drawn to each vertebra on a spinal X-ray. When this angle is too large the patient experiences discomfort or pain. Chiropractic treatment of this condition involves decreasing this angle by applying (nonsurgical) manipulation or pressure. Harrison et al. (2002) propose a particular such treatment. They measured the angle on both pre- and post-treatment X-rays from a random sample of 48 patients.

2.1 Exercise 5.13 - (a)

2.1.1 Problem

Test whether the mean post-treatment angle is less than the mean angle prior to treatment.

2.1.2 Hypotheses

Let μ_{pri} be the mean angle prior to treatment and μ_{post} be the mean post-treatment angle. The hypotheses are:

$$\begin{cases} H_0 : \mu_{post} - \mu_{pri} \geq 0 \\ H_1 : \mu_{post} - \mu_{pri} < 0 \end{cases}$$

2.1.3 Data analysis

We will perform a paired t-test to address this question.

2.1.4 Check assumptions

In a paired sample t-test, the observations are defined as the differences between two sets of values, and each assumption refers to these differences, not the original data values. The paired sample t-test has four main assumptions:

1. The dependent variable must be continuous (interval/ratio).

   - The dependent variable is the difference of relative rotation angle between the L2 and L3 lumbar vertebrae between pre-treatment and post-treatment, which is continuous. This assumption is statisfied.

2. The observations are independent of one another.

   - As the question states, the dataset was derived from a random sample of 48 patients, which leads to independent observations. This assumption is statisfied.

3. The dependent variable should be approximately normally distributed.

- We conduct Shapio-Wilk normally test to check the normality assumption. The hypotheses are H_0: The dependent variable (the angle difference) distributes normally and H_1: The dependent variable does not distribute normally.
- Set $\alpha = 0.05$. The p-value of the normality test is 0.1012, which is not smaller than α. Thus we do not reject the null hypothesis. That is, we do not have enough evidence to claim that the angle difference does not distribute normally. This assumption is statisfied.

4. The dependent variable should not contain any outliers.

- We conduct Grubbs's test to detect whether the highest or lowest value is an outlier. Set the significant level $\alpha = 0.05$.
- Hypotheses for the lowest value test are H_0*: The lowest value is not an outlier* and H_1*: The lowest value is an outlier.* The p-value of the test is 0.0533, which is not lower than the significant level. Thus we do not reject H_0. That is, we have no enough evidence to claim that the lowest value is an outlier.
- Hypotheses for the highest value test are H_0*: The highest value is not an outlier* and H_1*: The highest value is an outlier.* The p-value of the test is 0.4814, which is not lower than the significant level. Thus we do not reject H_0. That is, we have no enough evidence to claim that the highest value is an outlier.
- As two tests demonstrate, we can say that the lowest and the highest value are not outliers. And it also can be believed that the dependent variable does not contain any outliers. This assumption is statisfied.

In conclusion, paired t-test is appropriate for this dataset.

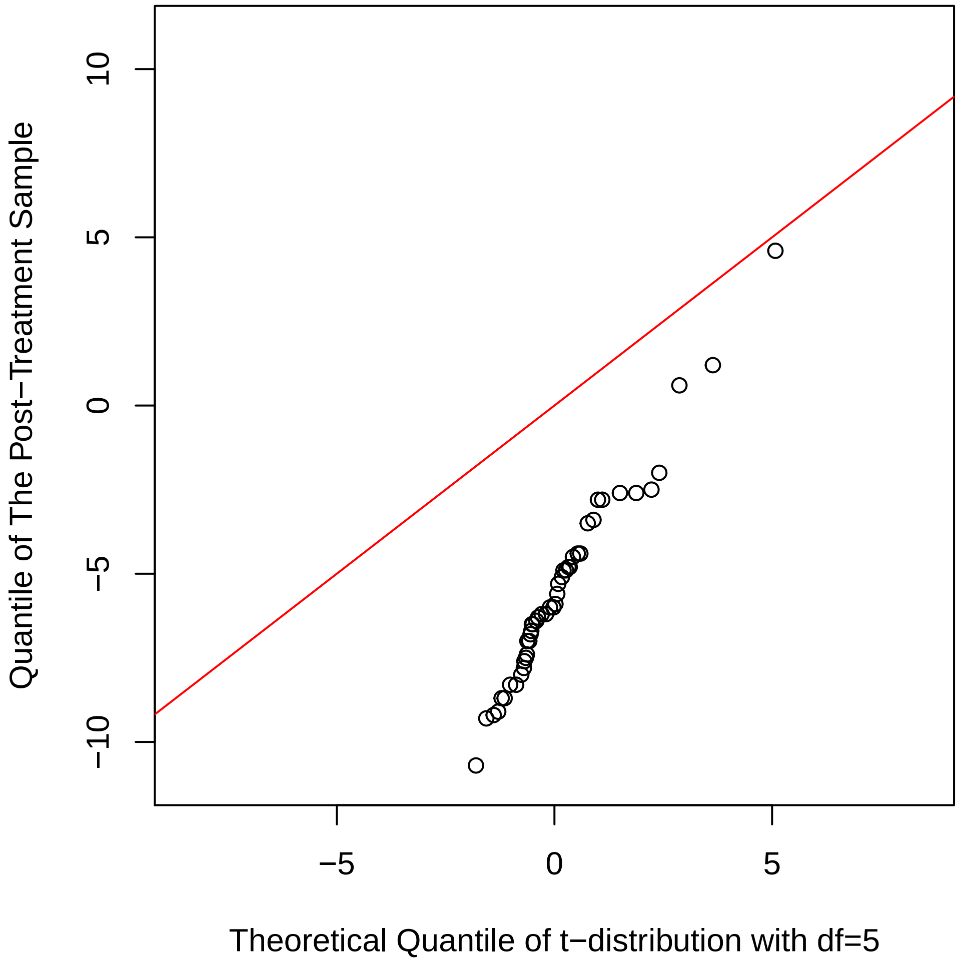
2.1.4.1 Conduct the test

The test statistic is T-value, which is defined as $T = \frac{\bar{D}-\mu_0}{SE(\bar{D})}$, where $\bar{D}$ is the average difference of two samples (i.e., $\sum[x_{post} - x_{pri}]/n$ here), n is the sample size, $SE(\cdot)$ is the standard error term, μ_0 is the testing value (i.e., 0 here). We set the significant level α as 0.05, which means that the probability that we commit type I error (reject H_0 when H_0 is actually true) is 0.05.

The observation test statistic, T_{obs}, is -3.5564 with the degree of freedom 47. The p-value is 4×10^{-4}, which is lower than the significant level. Thus we reject the null hypothesis. The paired t-test showed that the mean post-treatment angle is less than the mean angle prior to treatment.

2.2 Exercise 5.13 - (b)

Construct a quantile plot to assess whether the post-treatment sample is compatible with a t distribution with 5 degrees of freedom.



In a quantile plot, observation points should be on a straight line with the slope of 1 and the intercept of 0 if the sample follow the given distribution. We can see that the observation points do not do that way, indicating that the sample may not follow a t distribution with 5 degrees of freedom.