統計諮詢 - 作業 2

國立成功大學統計學系暨數據科學研究所

廖傑恩（RE6094028）

2021-03-18

1 Exercise 6.5

1.1 問題敘述

研究者想比較癌症病人在 5 種不同器官接受補充抗壞血酸（supplemental ascorbate）後的存活天數（Cameron & Pauling, 1978; Hand et al., 1994）。這 5 種器官為：胃（Stomach）、支氣管（Bronchus）、結腸（Colon）、卵巢（Ovary）和乳房（Breast）。

1.2 資料介紹

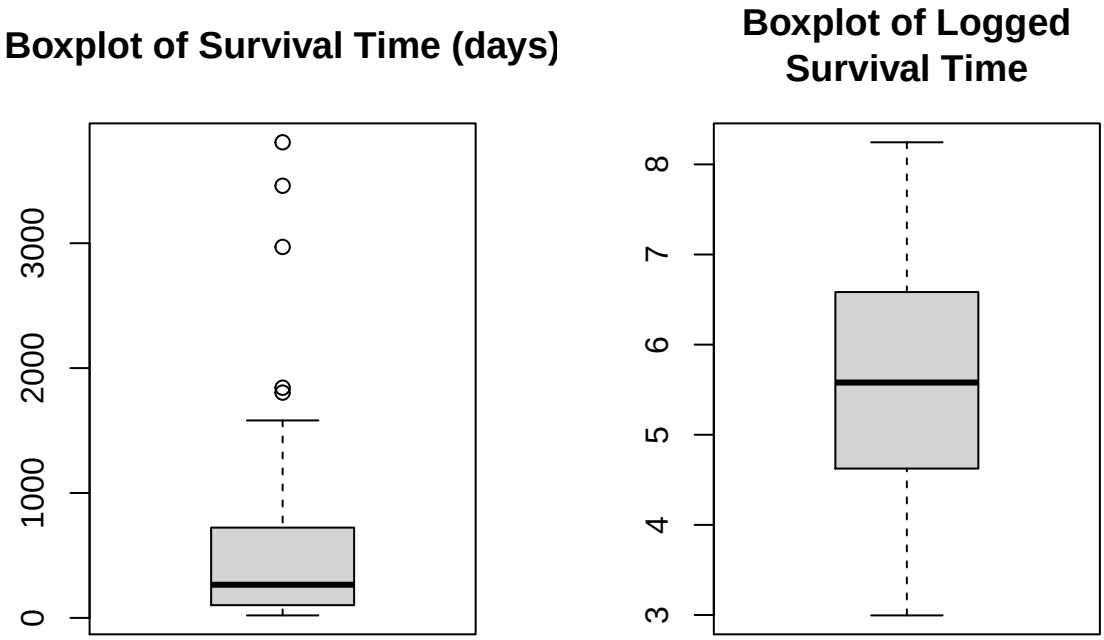
資料共有 64 列、2 個變項，每一列為一名癌症病患的資料。變項說明如下：

- `surv.time`：病患存活天數
- `organ`：接受補充抗壞血酸的器官，有胃、支氣管、結腸、卵巢和乳房 5 種

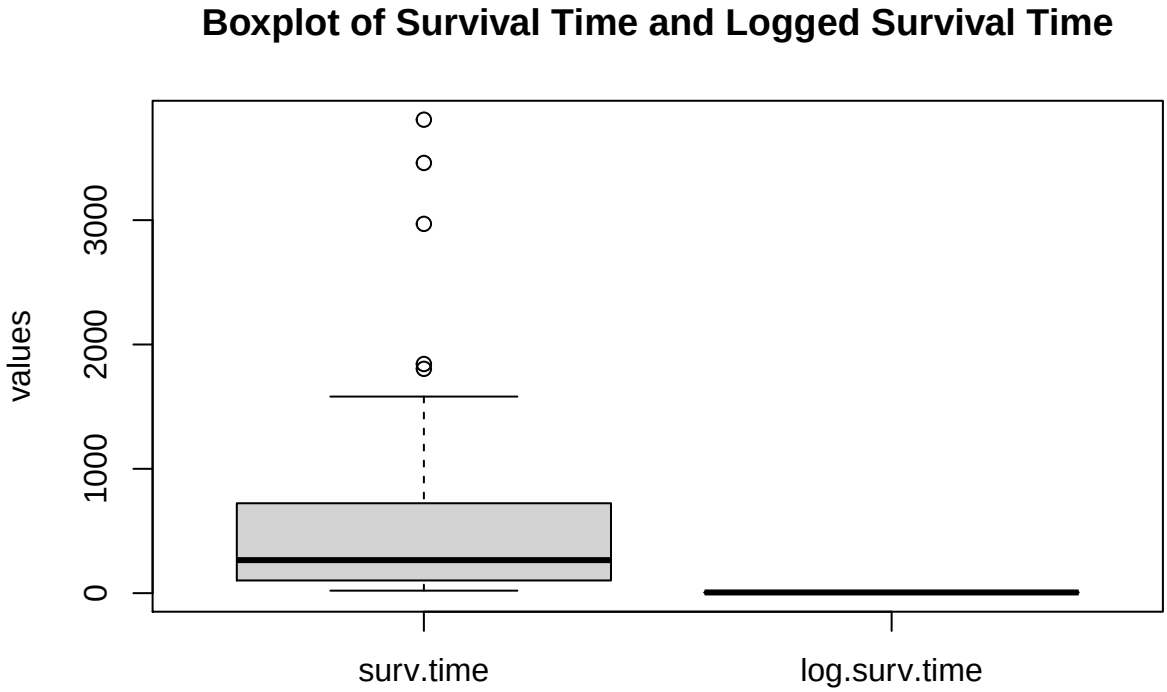
1.3 Exercise 6.5 - (a)

Perform a log transformation of the response days for each of the five levels of the factor site in order to improve conformity with the required assumption that the data be approximately normally distributed with equal within-site variance. Produce and compare box plots to compare the response before and after the transformation.

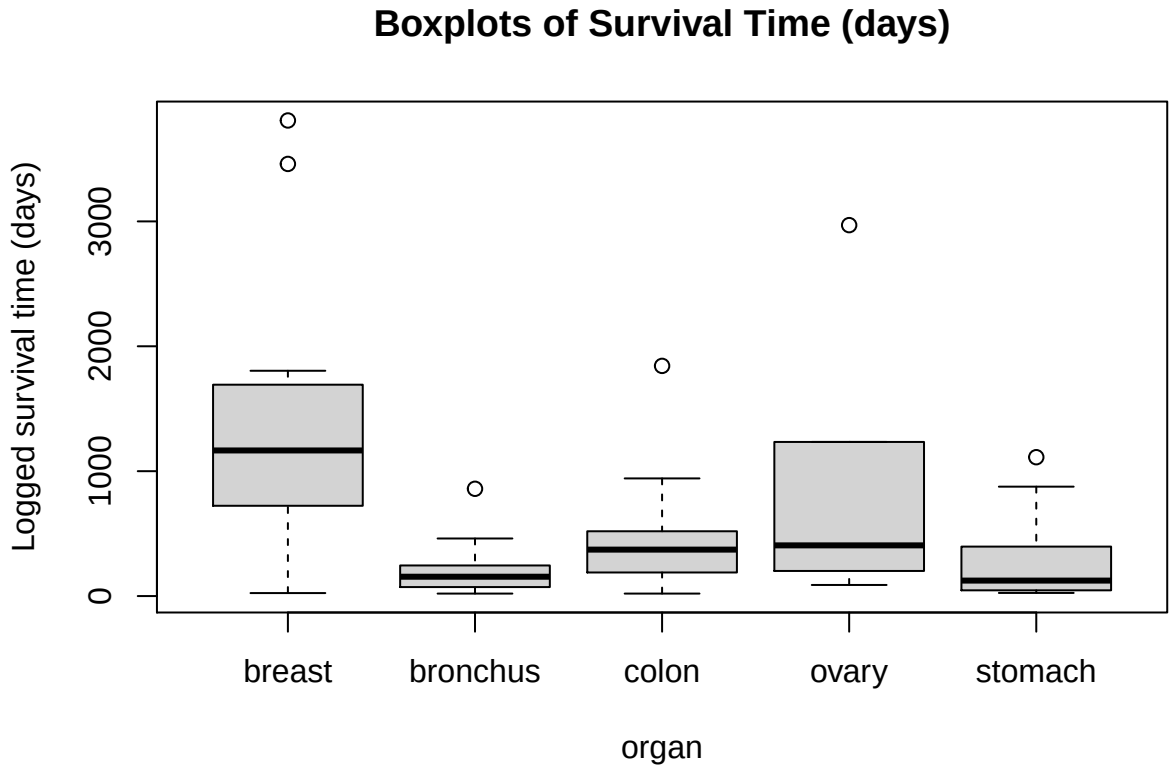
1.3.1 繪製盒鬚圖

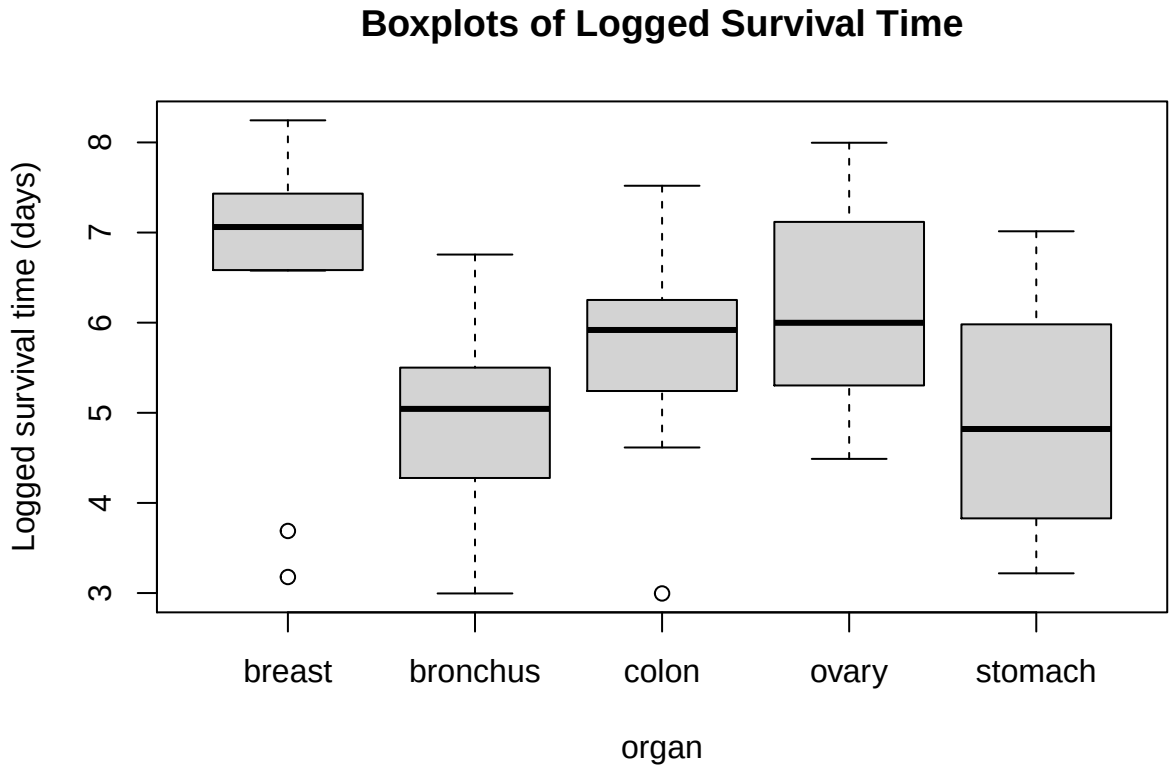


由上兩個盒鬚圖（boxplot）可以發現到原始資料的分布很廣、變異性很大（請注意兩圖的 y 軸不同），而進行 log 轉換後的資料值皆小於 10，變異性低了許多，資料較為集中。下圖將兩個盒鬚圖置於同一 y 軸來比較，可以看出很明顯的差異。



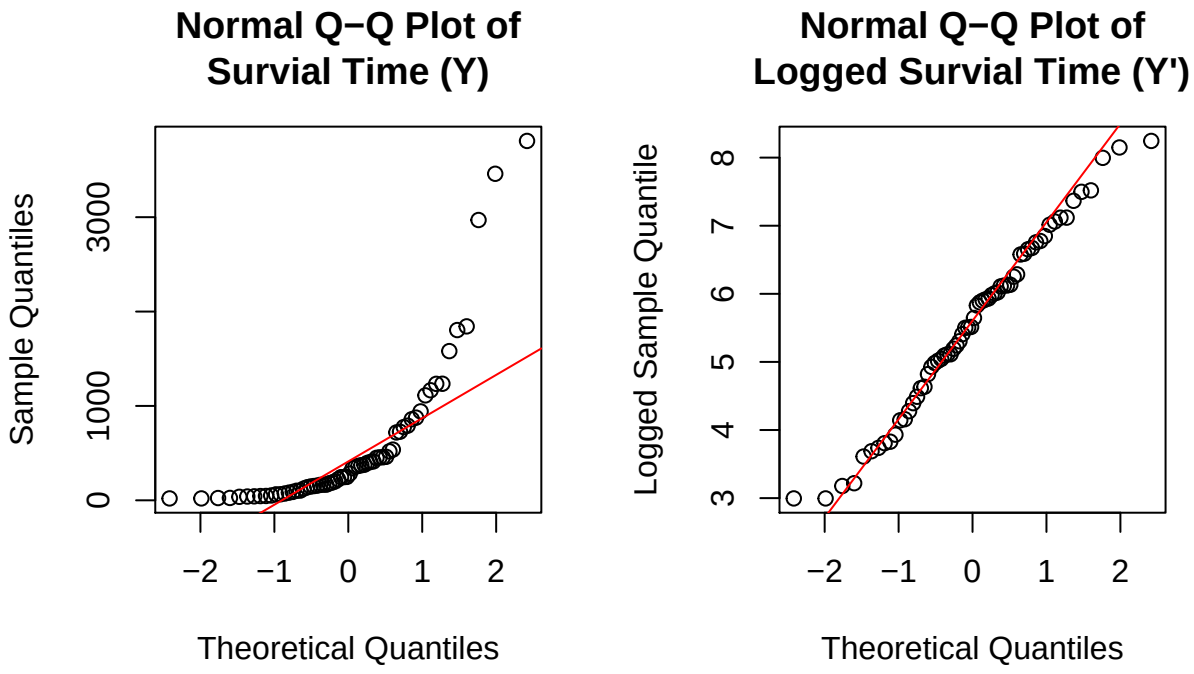
接著我們依據接受補充抗壞血酸的器官分開繪製病患存活天數的盒鬚圖。可以發現，在進行 log 轉換前，各組（不同器官）的存活天數都有特別大的值，而進行 log 轉換後，資料則明顯較為集中。





1.3.2 常態檢定

針對原始的存活天數資料（Y）與經過 log 轉換的存活天數資料（Y'）繪製 Normal Q-Q plot（如下兩圖），可以發現 Y 的資料點明顯偏離 45 度線，而 Y' 的資料點則相當符合 45 度線。顯示 Y 可能不服從常態分配（Normal distribution, ND），而 Y' 則可能服從。



除了繪製 Q-Q plot，我們以 Shapiro-Wilk 常態檢定法來檢定原始的存活天數資料（Y）是否服從常態分配，假設如下：

$$\begin{cases} H_0 : Y \sim ND \\ H_1 : Y \ does \ not \sim ND \end{cases}$$

令顯著水準 α 為 0.05，意味著我們進行此檢定犯下型一錯誤（type I error, 當 H_0 為真時拒絕 H_0）的機率為 0.05。

檢定結果如下：檢定統計量 W 為 0.6617，p 值為 $7.2017664 \times 10^{-11}$，小於顯著水準，因此我們拒絕 H_0。在 $1 - \alpha = 95\%$ 的信心水準下，我們認為樣本並不服從常態分配。

我們同樣以 Shapiro-Wilk 法來檢定經過 log 轉換的存活天數資料（Y'）是否服從常態分配，假設如下：

$$\begin{cases} H_0 : Y' \sim ND \\ H_1 : Y' \ does \ not \sim ND \end{cases}$$

令顯著水準 α 為 0.05。檢定結果如下：檢定統計量 W 為 0.9833，p 值為 0.5405，不小於顯著水準，因此我們不拒絕 H_0。在 $1 - \alpha = 95\%$ 的信心水準下，我們並沒有充分的證據認為樣本不服從常態分配。

1.4 Exercise 6.5 - (b)

Perform an analysis to assess differences in mean survival between the different cancer sites.

1.4.1 研究問題

研究問題: 不同癌症發生位置（接補充抗壞血酸的器官受）的平均存活天數是否有所差異。令 μ_x 為某癌症發生位置存活天數的母體平均數，$x = Stomach,\ Bronchus,\ Colon,\ Ovary,\ or\ Breast$，研究假說如下:

$$\begin{cases} H_0 : \mu_{Stomach} = \mu_{Bronchus} = \mu_{Colon} = \mu_{Ovary} = \mu_{Breast} \\ H_1 : Not\ H_0\ (At\ least\ one\ pair\ (x, x')\ s.t.\ \mu_x \neq \mu_{x'}),\ where\ x \neq x' \end{cases}$$

1.4.2 資料分析：ANOVA

我們以單因子變異數分析（one-way ANOVA）來回答這個研究問題。

1.4.2.1 檢驗分析的前提假設是否滿足

One-way ANOVA 有若干前提假設（assumption），在進行檢驗之前，我們先檢查資料是否符合這些假設。

1. 獨變項須為類別變數（categorical variable），依變項必須是連續變數（continuous variable）

此分析獨變項為接受補充抗壞血酸的器官，為含有 5 個類別的類別變數；依變項為病患存活天數（Y），為連續變項。符合。若將依變項進行 log 轉換形成（Y'），仍為連續變項，此前提仍符合。

2. 各組樣本依變項獨立

此分析中，各組依變項為 (1) 在胃接受補充抗壞血酸者的存活天數（或進行 log 轉換的存活天數）、(2) 在支氣管接受補充抗壞血酸者的存活天數（或進行 log 轉換的存活天數）、(3) 在結腸接受補充抗壞血酸者的存活天數（或進行 log 轉換的存活天數）、(4) 在卵巢接受補充抗壞血酸者的存活天數（或進行 log 轉換的存活天數）與 (5) 在乳房接受補充抗壞血酸者的存活天數（或進行 log 轉換的存活天數）。不論是否進行 log 轉換，這五者都互不影響彼此，都符合前提假設。

3. 若要進行事後多重比較，依變項母體必須服從常態分佈（Normal Distribution, ND）

在前面 `Exercise 6.5 - (a)` 時，我們已知當令顯著水準（significant level）為 0.05，我們有充分證據支持依變項 Y 分配不服從常態分佈，而沒有充分證據支持經 log 轉換的依變項 Y' 分配不服從常態分佈。Y' 通過此前提假設，為了有機會進行事後多重比較，以下的檢驗與分析以 Y' 進行。

4. 變異數同質（homogeneity of variance）：各組依變項的變異數必須相等。

我們以 Levene 檢定檢定檢驗變異數同質是否成立。令 σ_x^2 為在 x 器官接受補充抗壞血酸者經 log 轉換存活天數的母體變異數，研究假說如下，並令顯著水準（significant level）為 0.05。

$$\begin{cases} H_0 : \sigma^2_{Stomach} = \sigma^2_{Bronchus} = \sigma^2_{Colon} = \sigma^2_{Ovary} = \sigma^2_{Breast} \\ H_1 : Not\ H_0\ (At\ least\ one\ pair\ (x, x')\ s.t. \sigma_x^2 \neq \sigma_{x'}^2),\ where\ x \neq x' \end{cases}$$

檢定統計量為 W，服從自由度為 $k-1$ 與 $N-k$ 的 F 分配，如下式:

$$W = \frac{(N-k)\sum_{i=1}^{k} n_i(\bar{Z}_{i.}\bar{Z})}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Z_{ij}\bar{Z}_{i.})} \sim F(k-1, N-k)$$

其中 $N = \sum_{i=1}^{k} n_i$ 為總樣本數，$Z_{ij} = |Y_{ij} - \bar{Y_{i.}}|$，$\bar{Y_{i.}} = \frac{1}{n}\sum_{j=1}^{n_j} Y_{ij}$，$\bar{Z_{i.}} = \frac{1}{n_{i.}}\sum_{j=1}^{n_i} Z_{ij}$，$\bar{Z} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Z_{ij}$

檢定結果的檢定統計量 W 為 0.6685，其 p 值為 0.6164，不小於顯著水準，因此我們不拒絕 H_0，意味著我們沒有足夠的證據證明變異數同質性不存在，也就是說我們無法證明至少有一組母體變異數與其他組不同，通過此前提假設。

在常態性成立的情況下，Bartlett 是變異數同質性檢定法中統計檢定力（power）較高者。我們再以 Bartlett 檢定檢驗變異數同質是否成立。研究假說同上，我們同樣令顯著水準為 0.05。

檢定結果的檢定統計量 $Barlett's\ k^2$ 為 4.809，其 p 值為 0.3075，不小於顯著水準，因此我們不拒絕 H_0，意味著我們沒有足夠的證據證明變異數同質性不存在，也就是說我們無法證明至少有一組母體變異數與其他組不同，通過此前提假設。

5. 殘差（residuals）服從常態分配。

待配適完模型後診斷。

以上步驟 1-4 顯示，在我們的資料中，以經 log 轉換的存活天數（Y'）為依變項，ANOVA 的前提假設均滿足，因此我們可以進行 ANOVA。

1.4.2.2 檢定統計量與檢定結果

One-way ANOVA 的檢定統計量為 F 值，其服從自由度為 $k-1$ 與 $N-k$ 的 F 分配，數學式如下:

$$F_{TS} = \frac{explained\ variation}{unexplained\ variation} = \frac{\sum_{j=1}^{k} n_j(\bar{Y'_j} - \bar{Y'})/(k-1)}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y'_{ji} - \bar{Y'_j})^2/(N-k)} \sim F(k-1, N-k)$$

其中，k 為獨變項組別數；n_j 為第 j 組的觀察值（observations）個數；$N = \sum_{j=1}^{k} n_j$，也就是總觀察值數；$\bar{Y'_j}$ 為第 j 組依變項的樣本平均數，$\bar{Y'}$ 為依變項樣本平均數；Y'_{ji} 為第 j 組的第 i 個依變項觀察值。

我們令顯著水準為 0.05，檢定統計量 F_{TS} 為 4.286，p 值為 0.004，小於顯著水準，因此我們可以拒絕 H_0。

最後我們檢查殘差項（ϵ）是否滿足常態假設，研究假說如下：

$$\begin{cases} H_0 : \epsilon \sim ND \\ H_1 : \epsilon \; does \; not \sim ND \end{cases}$$

檢定結果如下：檢定統計量 W 為 0.9651，p 值為 0.067，不小於顯著水準，因此我們不拒絕 H_0，也就是說我們沒有充分證據支持殘差並不服從常態分配，通過常態假設。

我們可以先前 ANOVA 檢定的結果得到結論：我們有足夠的證據支持「不同癌症發生位置（接受補充抗壞血酸的器官）的病患，其平均存活天數有所差異」。

1.4.3 資料分析：事後多重比較

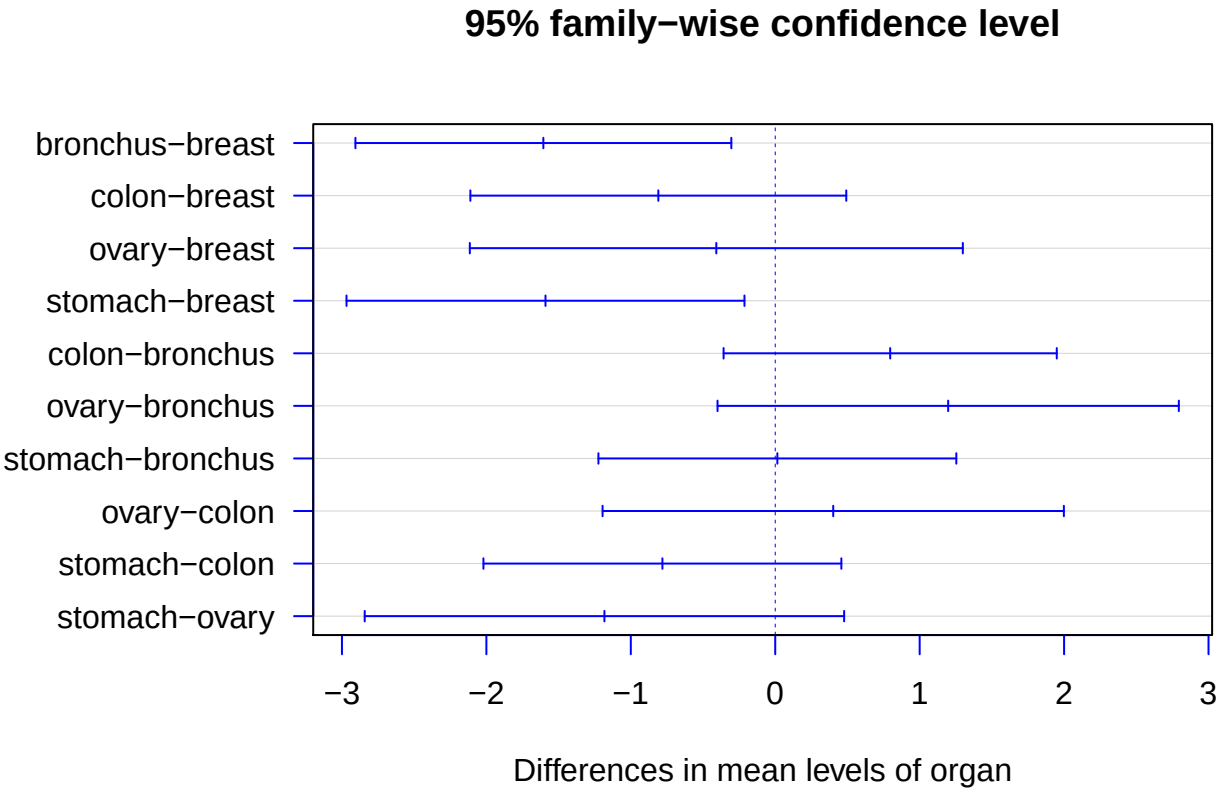
目前的分析只能得到「不同癌症發生位置（接受補充抗壞血酸的器官）的病患，其平均存活天數有所差異」的結論，無法得知哪些接受補充抗壞血酸的病患平均存活天數比較低，如果想得知，必須接著進行事後多重比較，合適的檢定方法有 Scheffé 與 Tukey HSD 等。

1.4.3.1 Tukey HSD 檢定法

我們以 Tukey HSD 來進行 5 個接受補充抗壞血酸器官，其 logged 存活天數的兩兩比較。檢定結果如下表所示，其中的「差異」為組別 2 平均數減去組別 1 平均數，所以負值表示組別 1 的平均數較大。Diff= 兩組差異平均估計值；CI=confident interval；p.adj= 經過 Tukey 方法校正的 p 值。

組別 1	組別 2	Diff	95%CI 下界	95%CI 上界	p.adj	是否顯著
breast	bronchus	-1.6054	-2.9067	-0.3041	0.0083	顯著
breast	colon	-0.8095	-2.1108	0.4918	0.4120	不顯著
breast	ovary	-0.4080	-2.1148	1.2988	0.9620	不顯著
breast	stomach	-1.5907	-2.9684	-0.2130	0.0158	顯著
bronchus	colon	0.7960	-0.3575	1.9494	0.3070	不顯著
bronchus	ovary	1.1974	-0.3995	2.7944	0.2300	不顯著
bronchus	stomach	0.0147	-1.2243	1.2538	1.0000	不顯著
colon	ovary	0.4015	-1.1954	1.9984	0.9540	不顯著
colon	stomach	-0.7812	-2.0202	0.4578	0.3980	不顯著
ovary	stomach	-1.1827	-2.8425	0.4771	0.2760	不顯著

下圖則為 Tukey HSD 檢定結果的視覺化，縱軸有標示各組兩兩配對，橫軸則為兩組差異，若其信賴區間的 bar 沒有包含到 0，則表示此對的兩組有顯著差異。



根據 Tukey HSD 檢驗，在 95% 信心水準下，我們可以得到以下結論：

- 支氣管（bronchus）接受補充抗壞血酸的病患，其存活天數（經 log 轉換）顯著低於乳房（breast）接受補充抗壞血酸的病患。
- 胃（stomach）接受補充抗壞血酸的病患，其存活天數（經 log 轉換）顯著低於乳房（breast）接受補充抗壞血酸的病患。
- 其餘器官接受補充抗壞血酸的病患存活天數（經 log 轉換）之兩兩比較，皆沒有顯著差異。

2 Exercise 6.9

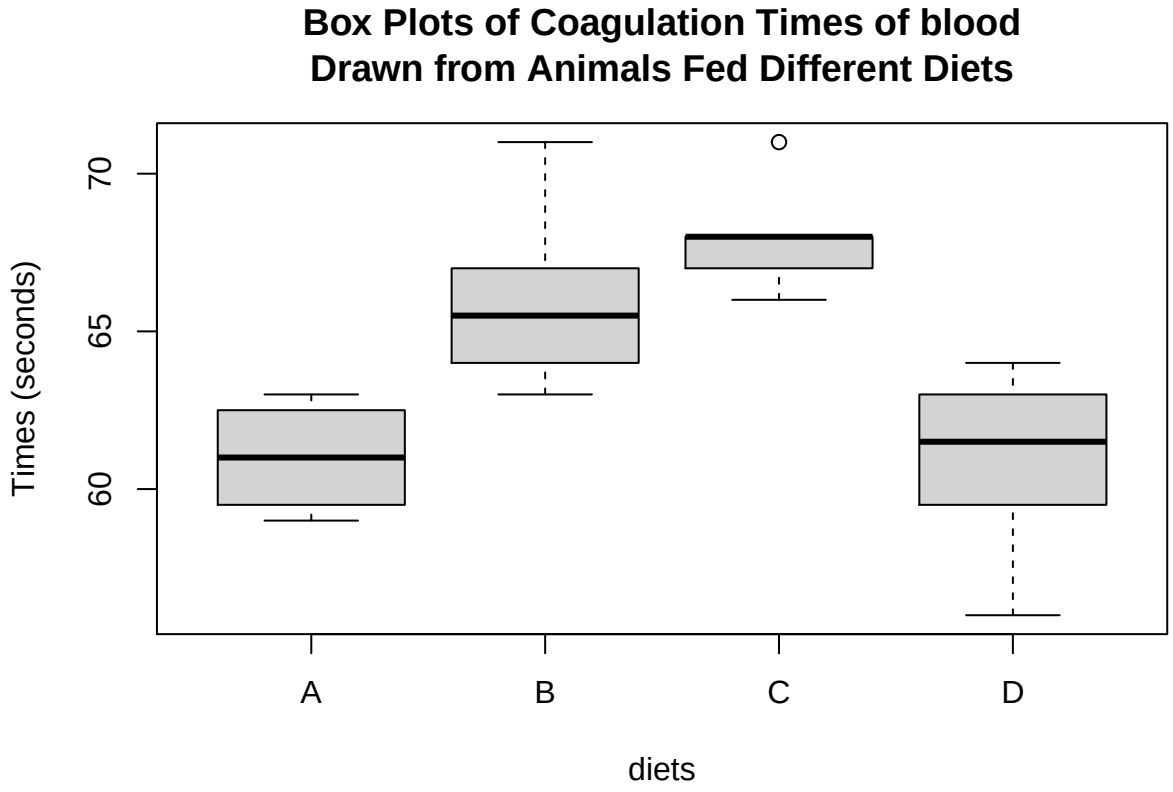
Box 等人（1978）想比較被餵食 4 種不同飲食（A, B, C, D）的動物的血，其凝結時間是否有所差異。

2.1 資料介紹

資料共有 24 筆，每一筆為一單位的動物血，有 2 個變項：

1. `times`: 動物血凝結的時間（單位：秒）
2. `diets`: 動物血取自哪種飲食種類的動物，有 A、B、C 與 D 這 4 種

下兩圖為 4 種飲食種類的動物的血的凝結時間盒鬚圖，不同飲食種類的凝結時間似乎有差異。



2.2 研究假設

令 μ_x 為以 x 飲食的動物的血的平均凝結時間（單位：秒），$x = A, B, C, or\ D$，研究假設如下：

$$\begin{cases} H_0 : \mu_A = \mu_B = \mu_C = \mu_D \\ H_1 : Not\ H_0\ (At\ least\ one\ pair\ (x, x')\ s.t. \mu_x \neq \mu_{x'}),\ where\ x \neq x' \end{cases}$$

2.3 資料分析：One-way ANOVA

我們以單因子變異數分析（one-way ANOVA）來回答這個研究問題。

2.3.1 檢驗分析的前提假設是否滿足

One-way ANOVA 有 5 個前提假設（assumption），在進行檢驗之前，我們先檢查資料是否符合這些假設。

1. 獨變項須為類別變數（categorical variable），依變項必須是連續變數（continuous variable）

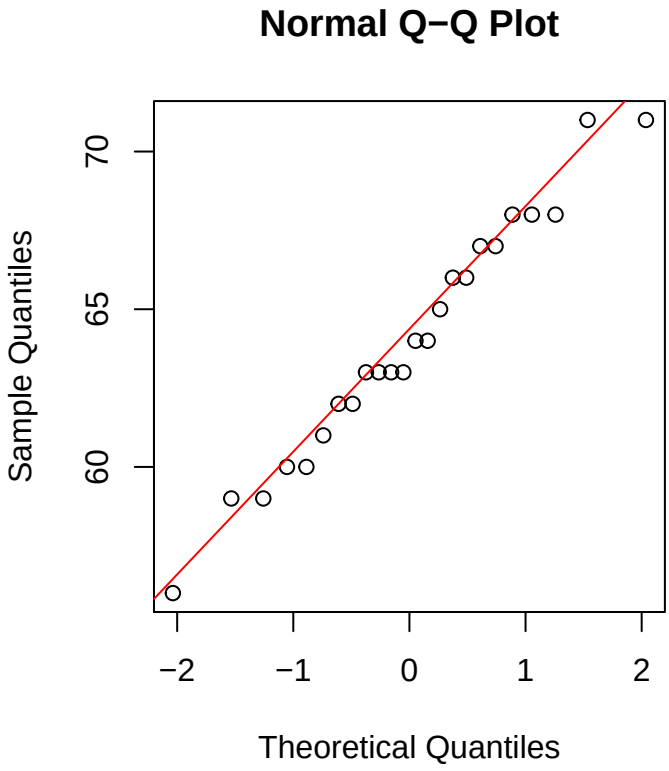
此分析獨變項為飲食種類，為含有 4 個類別的類別變數；依變項為動物血，為連續變項。符合。

2. 各組樣本依變項獨立

四種飲食種類的動物的血的凝結時間都互不影響彼此，符合前提假設。

3. 若要進行事後多重比較，依變項母體必須服從常態分佈（Normal Distribution, ND）

針對動物血的凝結時間（Y）繪製 Normal Q-Q plot（如下圖），可以發現資料點相當符合 45 度線。顯示 Y 可能服從常態分配（Normal distribution, ND）。



除了繪製 Normal Q-Q plot，我們以 Shapiro-Wilk 常態檢定法來檢定動物血的凝結時間（Y）是否服從常態分配，假設如下：

$$\begin{cases} H_0 : Y \sim ND \\ H_1 : Y \ does \ not \sim ND \end{cases}$$

令顯著水準 α 為 0.05，意味著我們進行此檢定犯下型一錯誤（type I error, 當 H_0 為真時拒絕 H_0）的機率為 0.05。

檢定結果如下：檢定統計量 W 為 0.9776，p 值為 0.8476，不小於顯著水準，因此我們不拒絕 H_0，也就是說我們沒有充分證據支持 Y 不服從常態分配，通過常態假設。

4. 變異數同質（homogeneity of variance）：各組依變項的變異數必須相等。

檢定結果的檢定統計量 W 為 0.6685，其 p 值為 0.6164，不小於顯著水準，因此我們不拒絕 H_0，意味著我們沒有足夠的證據證明變異數同質性不存在，也就是說我們無法證明至少有一組母體變異數與其他組不同，通過此前提假設。

我們以 Levene 與 Bartlett 檢定檢驗變異數同質是否成立。令 σ_x^2 為飲食種類為 x 的動物的血的凝結時間母體變異數，兩檢定的研究假說皆如下所示，顯著水準皆設定為 0.05。

$$\begin{cases} H_0 : \sigma_A^2 = \sigma_B^2 = \sigma_C^2 = \sigma_D^2 \\ H_1 : Not\ H_0\ (At\ least\ one\ pair\ (x, x')\ s.t. \sigma_x^2 \neq \sigma_{x'}^2),\ where\ x \neq x' \end{cases}$$

Levene 的檢定統計量 W 為 0.6492，其 p 值為 0.5926，不小於顯著水準，因此我們不拒絕 H_0，意味著我們沒有足夠的證據證明變異數同質性不存在，也就是說我們無法證明至少有一組母體變異數與其他組不同，通過此前提假設。

檢定結果的檢定統計量 $Barlett's\ k^2$ 為 1.668，其 p 值為 0.6441，不小於顯著水準，因此我們不拒絕 H_0，意味著我們沒有足夠的證據證明變異數同質性不存在，也就是說我們無法證明至少有一組母體變異數與其他組不同，仍然通過此前提假設。

5. 殘差（residuals）服從常態分配。

待配適完模型後診斷。

以上步驟 1-4 顯示，在我們的資料中，以動物血的凝結時間（Y）為依變項，ANOVA 的前提假設均滿足，因此我們可以進行 ANOVA。

2.3.2 檢定統計量與檢定結果

One-way ANOVA 的檢定統計量為 F 值，其服從自由度為 $k-1$ 與 $N-k$ 的 F 分配，數學式如下：

$$F_{TS} = \frac{explained\ variation}{unexplained\ variation} = \frac{\sum_{j=1}^{k} n_j(\bar{Y_j} - \bar{Y})/(k-1)}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ji} - \bar{Y_j})^2/(N-k)} \sim F(k-1, N-k)$$

其中，k 為獨變項組別數；n_j 為第 j 組的觀察值（observations）個數；$N = \sum_{j=1}^{k} n_j$，也就是總觀察值數；$\bar{Y_j}$ 為第 j 組依變項的樣本平均數，$\bar{Y}$ 為依變項樣本平均數；Y_{ji} 為第 j 組的第 i 個依變項觀察值。

我們令顯著水準為 0.05，檢定統計量 F_{TS} 為 13.571，p 值為 4.66×10^{-5}，小於顯著水準，因此我們可以拒絕 H_0。最後我們檢查殘差項（ϵ）是否滿足常態假設，令顯著水準為 0.05，研究假說如下：

$$\begin{cases} H_0 : \epsilon \sim ND \\ H_1 : \epsilon\ does\ not \sim ND \end{cases}$$

檢定結果如下：檢定統計量 W 為 0.9783，p 值為 0.862856，不小於顯著水準，因此我們不拒絕 H_0，也就是說我們沒有充分證據支持殘差並不服從常態分配，通過常態假設。

我們可以先前 ANOVA 檢定的結果得到結論：我們有足夠的證據支持「不同飲食種類的動物的血，其凝結時間有所差異」。

2.3.3 資料分析：事後多重比較

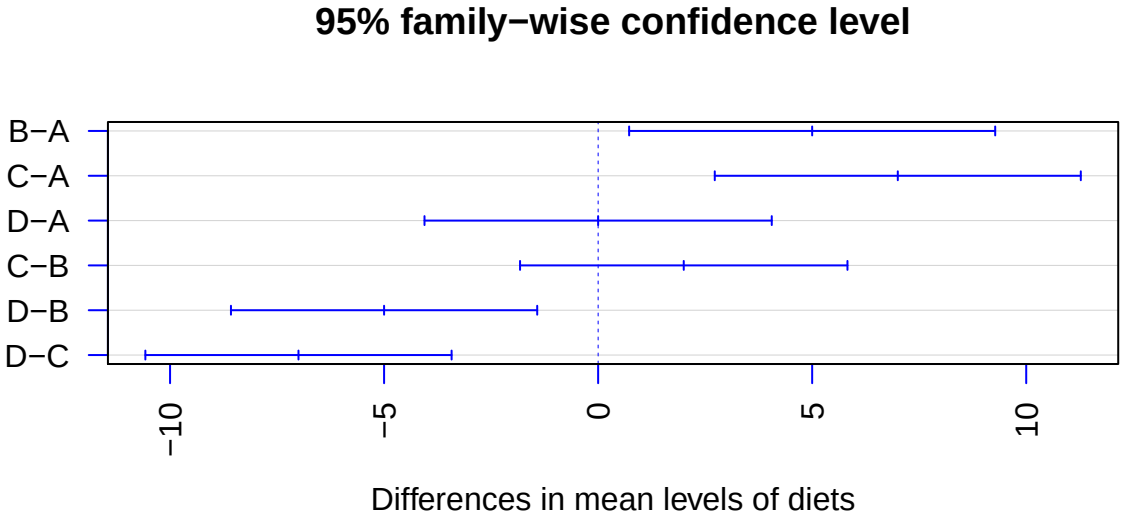
目前的分析只能得到「不同飲食種類的動物的血，其凝結時間有所差異」的結論，無法得知哪些飲食種類的動物的血凝結時間比較短或長，如果想得知，必須接著進行事後多重比較，合適的檢定方法有 Scheffé 與 Tukey HSD 等。

2.3.3.1 Tukey HSD 檢定法

我們 Tukey HSD 來進行 4 個飲食種類的動物的血，其凝結時間的兩兩比較。檢定結果如下表所示，其中的「差異」為組別 2 平均數減去組別 1 平均數，所以負值表示組別 1 的平均數較大。Diff= 兩組差異平均估計值；CI=confident interval；p.adj= 經過 Tukey 方法校正的 p 值。

組別 1	組別 2	Diff	95%CI 下界	95%CI 上界	p.adj	是否顯著
A	B	5	0.7246	9.2754	0.0183	顯著
A	C	7	2.7246	11.2754	0.0010	顯著
A	D	0	-4.0560	4.0560	1.0000	不顯著
B	C	2	-1.8241	5.8241	0.4770	不顯著
B	D	-5	-8.5771	-1.4229	0.0044	顯著
C	D	-7	-10.5771	-3.4229	0.0001	顯著

下圖則為 Tukey HSD 檢定結果的視覺化，縱軸有標示各組兩兩配對，橫軸則為兩組差異，若其信賴區間的 bar 沒有包含到 0，則表示此對的兩組有顯著差異。



根據 Tukey HSD 檢定，在 95% 信心水準下，我們可以得到以下結論:

- 飲食種類為 B 的動物的血，其凝結時間顯著大於飲食種類為 A 的動物的血。
- 飲食種類為 C 的動物的血，其凝結時間顯著大於飲食種類為 A 的動物的血。
- 飲食種類為 D 的動物的血，其凝結時間顯著小於飲食種類為 B 的動物的血。
- 飲食種類為 D 的動物的血，其凝結時間顯著小於飲食種類為 C 的動物的血。
- 飲食種類為 A 的動物的血以及飲食種類為 D 的動物的血，其凝結時間沒有顯著差異。
- 飲食種類為 B 的動物的血以及飲食種類為 C 的動物的血，其凝結時間沒有顯著差異。

也就是:

$$\mu_B \simeq \mu_C > \mu_A \simeq \mu_D$$

2.4 建議

此資料中觀測值只有 24 筆，某些飲食種類的觀測值個數也偏少，可以增加多一點觀測值，以增加分析結果的可信度。

3 Exercise 8.2

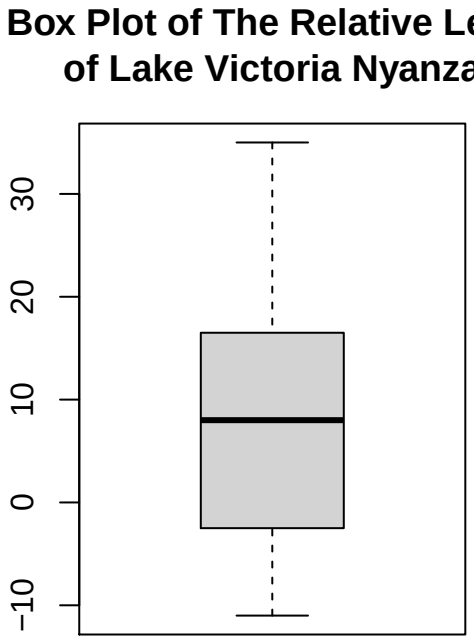
Shaw（1942）以及 Mosteller 與 Tukey（1977）等研究者紀錄了連續 20 年 Lake Victoria Nyanza 的相對水位與太陽黑子（sunspots）的個數。

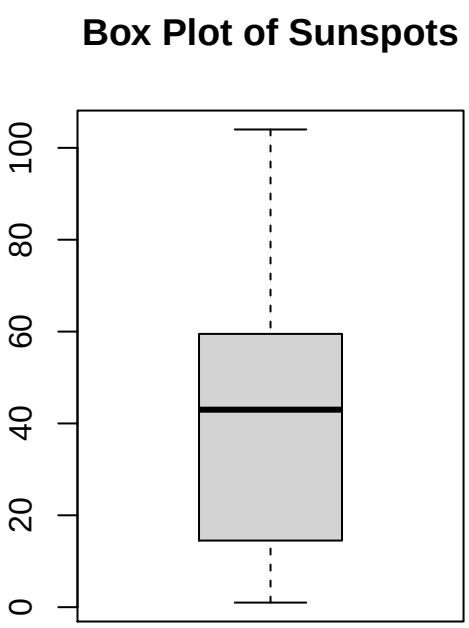
3.1 資料介紹

資料共有 20 筆，每一筆為一年的觀測紀錄，有 2 個變項:

- `level`: Lake Victoria Nyanza 的相對水位
- `sunspots`: 太陽黑子的個數

下兩圖為兩變項的盒鬚圖，可以看出這兩個變項的分佈沒有非常分散，也沒有特別大或特別小的值。





3.2 研究問題

以湖水水位為依變項（Y）、太陽黑子個數為獨變項（X），建立線性迴歸模型：

$$Y_{ij} = \beta_0 + \beta_1 X_i + \epsilon_i,\ where\ i = 1, 2, ..., 20,\ \epsilon_i \sim N(0, {\sigma_\epsilon}^2)$$

其中 β_0 為截距項，β_1 為迴歸線之斜率，ϵ_i 為服從常態分配的殘差項。我們欲探討太陽黑子個數對於湖水水位的預測力是否顯著，也就是要檢定迴歸係數 β_1 是否顯著不為 0，研究假說如下：

$$H_0 : \beta_1 = 0\ v.s.\ H_1 : \beta_1 \neq 0$$

3.3 資料分析

我們進行線性迴歸的配適，並以 t 檢定回答研究問題。

3.3.1 配適線性迴歸

我們以最大概似估計法（maximum likelihood estimation）進行線性迴歸的配適，其估計式如下：

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \quad {\hat{\sigma}_\epsilon}^2 = \frac{SS_E}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}$$

以我們資料估計的結果為：$\hat{\beta}_0$ = -8.0418，$\hat{\beta}_1 = 0.4128$，${\hat{\sigma}_\epsilon}^2 = 41.8065$。

3.3.2 檢定統計量與結果

令 n 為樣本大小，k 為獨變項個數，檢定統計量 t_{β_j}（$j = 1, ..., k$）服從自由度 $n-k$ 的 t 分配，數學式如下：

$$t_{\beta_j} = \frac{\hat{\beta}_j - 0}{Var(\hat{\beta}_j)} \sim t(n-k)$$

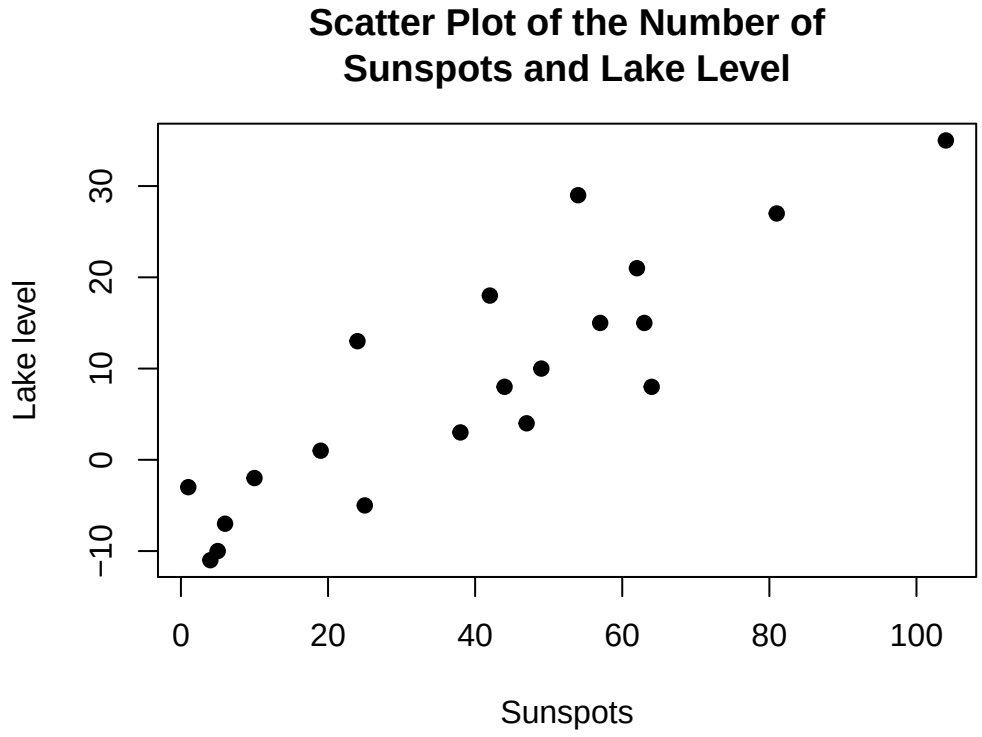
我們令顯著水準 α 為 0.05。檢定結果如下表所示。其中，我們關心的 β_1 之估計值 $\hat{\beta_1}$ 為，$1-\alpha = 1-0.05 = 95\%$ 信賴區間（confident interval, CI）為 [0.302, 0.5236]，統計檢定量 t_{TS} 為 7.8259，其 p 值為 0，小於顯著水準，因此我們拒絕 H_0，表示我們有充分證據支持 $\beta_1 = 0$ 這個宣稱是錯的，也就是說，太陽黑子個數對於湖水水位的預測力顯著。

	估計值	95%CI 下界	95%CI 上界	t 檢定統計量	p 值
截距 (beta_0)	-8.0418	-13.4109	-2.6726	-3.1467	0.0056
sunspots (beta_1)	0.4128	0.3020	0.5236	7.8259	0.0000

3.3.3 診斷前提假設是否滿足

線性迴歸分析有以下四項基本前提假設：

1. 線性關係：依變數和獨變項之間的關係必須是線性。



由以上散佈圖可以看出，依變數（lake level）和獨變項（sunspots）之間的關係是線性的。符合。

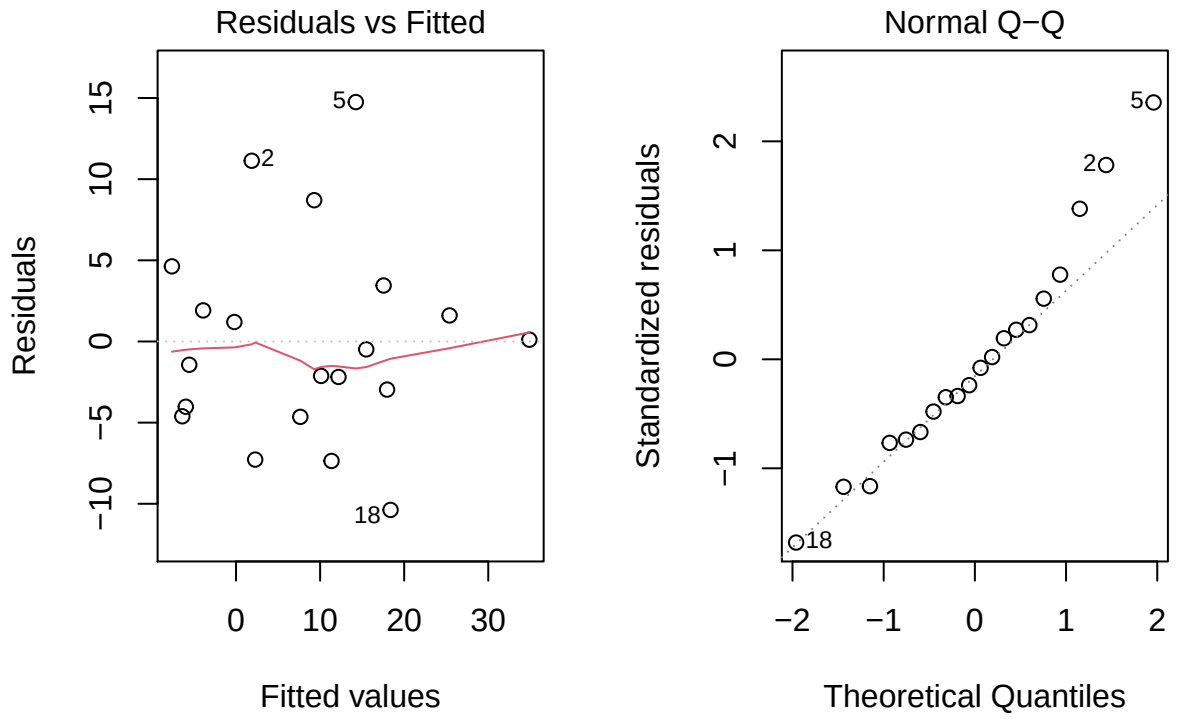
2. 殘差（ϵ）服從常態分配。

3. 殘差具備獨立性。

4. 殘差具備變異數同質性。

2 - 4 這 3 個假設可以下式表示：

$$\epsilon \overset{iid}{\sim} N(0, {\sigma_\epsilon}^2)$$



Normal Q-Q plot（上右圖）中殘差 quantile 資料點大部分落在 45 度線上，顯示殘差可能服從常態分布。而殘差與配適值散佈圖（上左圖）則顯示在各配適值殘差之變異數差異不大。我們以下面檢定來確認這些假設是否成立（顯著水準均設為 0.05）。

- 以 Shapiro-Wilk 檢定檢驗殘差常態假設：

$$\begin{cases} H_0 : \epsilon \sim ND \\ H_1 : \epsilon \; does \; not \sim ND \end{cases}$$

檢定結果如下：檢定統計量 W 為 0.9564，p 值為 0.4749276，不小於顯著水準，因此我們不拒絕 H_0，也就是說我們沒有充分證據支持殘差並不服從常態分配，通過常態假設。

- 以 Durbin-Waston 檢定檢驗殘差獨立性：

$$\begin{cases} H_0 : Residuals\ are\ independent. \\ H_1 : Residuals\ are\ not\ independent. \end{cases}$$

檢定結果如下：檢定統計量 DW 為 1.7077，p 值為 0.1948，不小於顯著水準，因此我們不拒絕 H_0，也就是說我們沒有充分證據支持殘差不具備獨立性，通過獨立假設。

- 以 Brown-Forsythe 檢定檢驗殘差變異同質性：

$$\begin{cases} H_0 : {\sigma_{\epsilon_1}}^2 = {\sigma_{\epsilon_2}}^2 = ... = {\sigma_{\epsilon_i}}^2 \\ H_1 : At\ least\ a\ pair\ (i, i')\ s.t. {\sigma_{\epsilon_i}}^2 \neq {\sigma_{\epsilon_{i'}}}^2 \end{cases}$$

檢定結果如下：檢定統計量 BF 為 0.1507，p 值為 0.7025，不小於顯著水準，因此我們不拒絕 H_0，也就是說我們沒有充分證據支持殘差不具備變異同質性，通過變異同質性假設。

3.4 結論

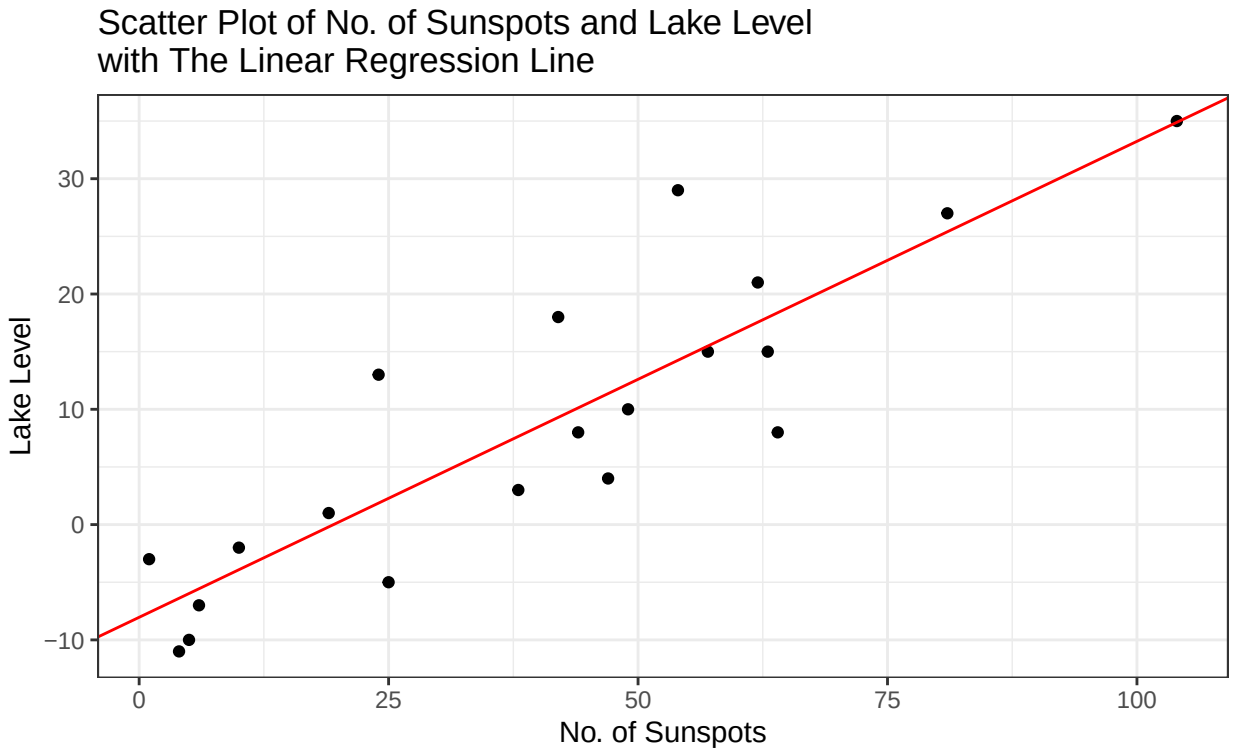
我們建立了以太陽黑子個數（X）來預測湖水水位（Y）的線性迴歸模型：

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

其中 $\hat{\beta}_0$ = -8.0418，$\hat{\beta}_1$ = 0.4128，其模型解釋力可由經由自由度校正的決定係數 R^2_{adj} 描述，為 0.7602，顯示我們建立的迴歸模型可以解釋依變項（i.e., 湖水水位）大部分的變異。R^2_{adj} 公式如下：

$$R^2 = 1 - \frac{SS_E/(n-K)}{SS_T/(n-1)}$$

將迴歸線繪製於兩變項的散佈圖上：


Scatter Plot of No. of Sunspots and Lake Level
with The Linear Regression Line
Lake Level
30
20
10
0
−10
0
25
50
75
100
No. of Sunspots