統計諮詢：Mid-term Project 2

國立成功大學統計學系暨數據科學研究所

陳溫茹（R26091040)、廖傑恩（RE6094028)、戴庭筠（R26091032）

2021-05-30

1 問題敘述

本研究想要探討的研究問題為:

1. 不同種類的作物的產量沒有差異。
2. 氮肥與作物產量有正向線性關聯：氮肥濃度越高，作物產量越多。
3. 氮肥與作物產量的關聯在不同種類的作物上沒有差異。

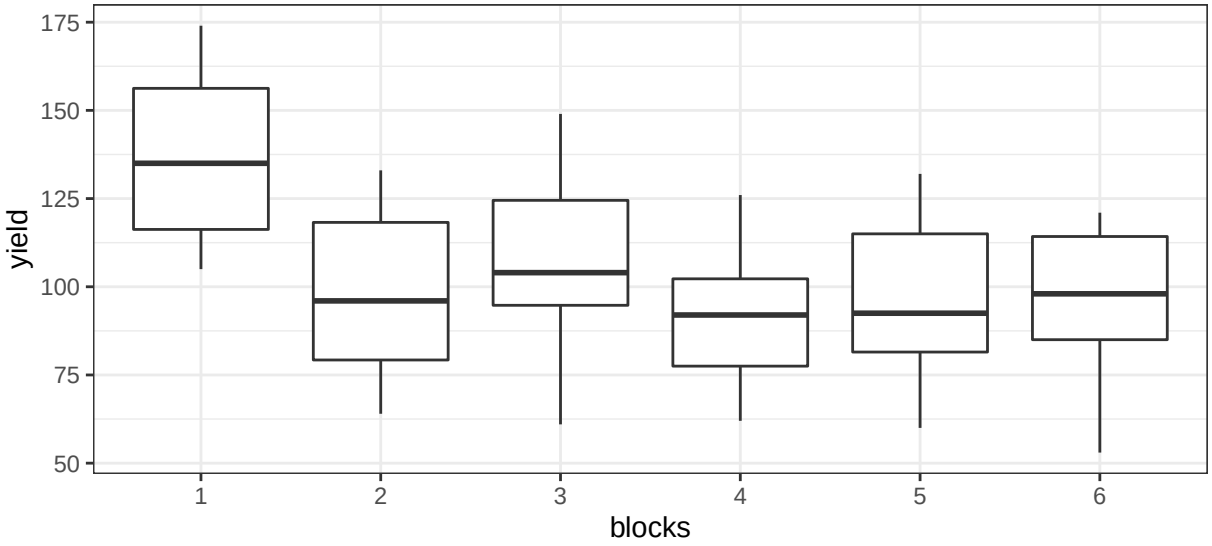
2 資料集敘述

關於作物實驗的資料集 `oats` 是由 Yates（1937）收集的，資料有 72 列、6 欄，每一列為一個實驗單位的資料。實驗採取裂區設計（split plot designs），實驗的田地被分為 72 個實驗單位：原始田地被分為 6 個 blocks；每個 block 被隨機分為 3 塊 plot；每個 plot 被隨機分為 4 塊 subplot，也就是實驗單位。資料欄位說明如下:

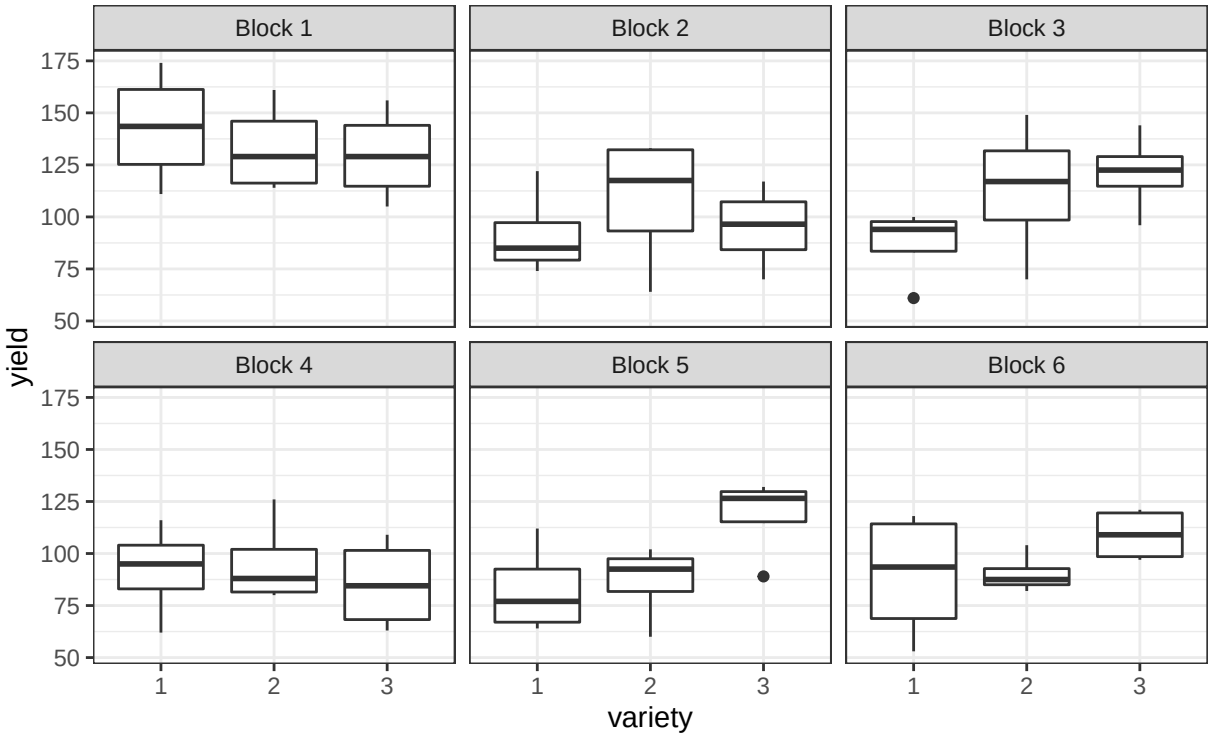
- `yield`: 作物產量，依變項。
- `blocks`: 田地的第一層分裂單位。原始田地被分為 6 個 blocks。
- `plots`: 田地的第二層分裂單位。每一個 block 被隨機分為 3 塊 plots。
- `variety`: 作物種類，獨變項之一，有 3 種作物。在每一個 block 中，3 種作物被隨機分配到 3 塊 plot 中。
- `subplots`: 田地的第三層分裂單位，也就是實驗單位。每一個 plot 被隨機分為 4 塊 subplots。
- `nitrogen`: 單一種氮肥的濃度，獨變項之二，有 0、0.2、0.4 與 0.6 這 4 種濃度。在每一個 plot 中，4 種氮肥濃度被隨機分配到 4 塊 subplot 中。

3 資料探索

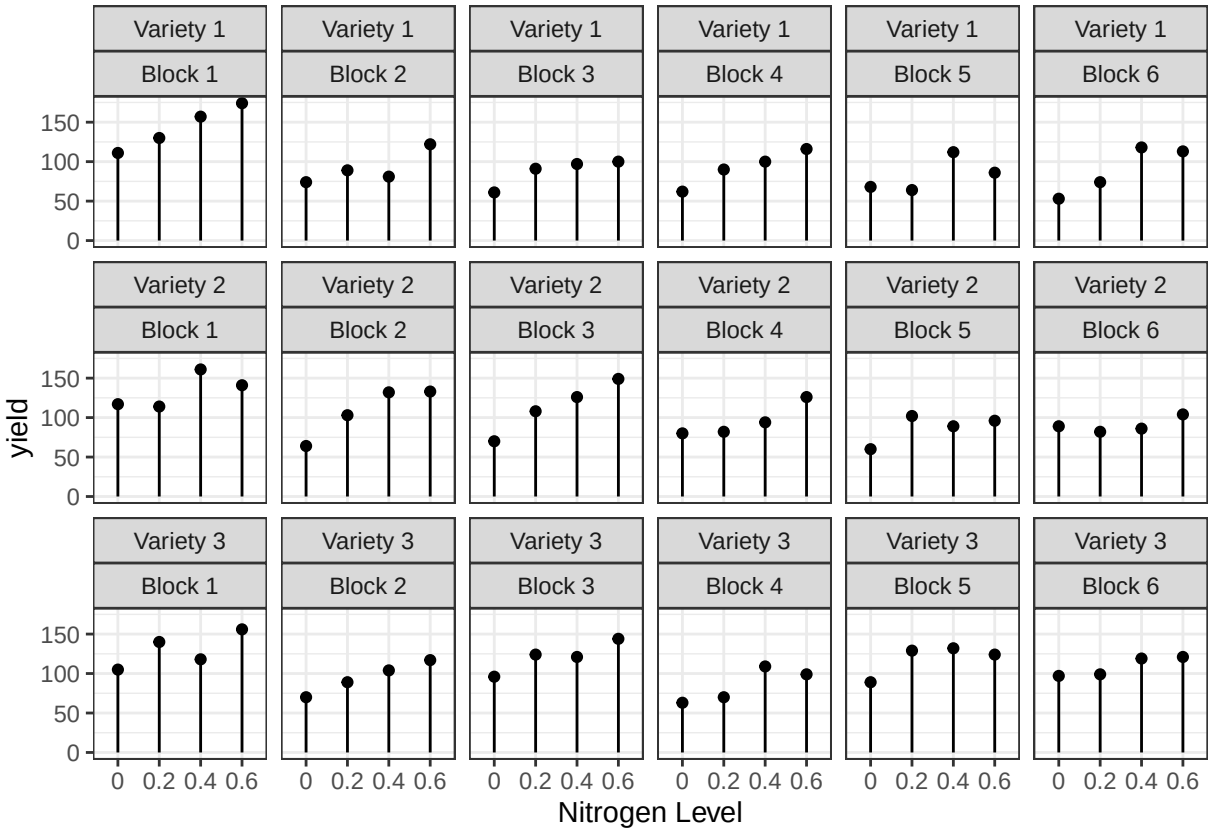
下圖是 6 個 block 的作物產量盒鬚圖。可以看出不同 block 間的產量似乎有所差異，因此在之後的分析中，我們應固定住 block 對產量的效果。



下圖是的 3 種作物的產量盒鬚圖，並以 6 個 block 分層來看。可以看到在大部分的 block 中，不同的作物產量的中位數差異不大。



下圖是的 4 種氮肥濃度的產量棒棒糖圖（lollipop plot），並以 6 個 block 各 3 個 plots（共 18 個 plots）分層來看。在大部分的 plot 中，似乎都可以看到「氮肥濃度與作物產量呈現正相關」的趨勢。



4 資料分析

實驗採取裂區設計（split plot designs），實驗的田地被分為 72 個實驗單位：原始田地被分為 6 個 blocks；每個 block 被隨機分為 3 塊 plot；每個 plot 被隨機分為 4 塊 subplot，也就是實驗單位。因應這樣的實驗設計，我們將採取 Between-Within 混合設計的多因子變異數分析（analysis of variance, ANOVA）。

4.1 變數與模型定義

令 Y_{ijk} 為 block i 中接受因子 A 為水準 j 的固定處置（fixed treatment）以及因子 B 為水準 k 的固定處置的實驗單位的作物產量（依變項），其中因子 A 為作物種類，有 3 水準（i.e., $j = 1, 2, 3$），因子 B 為氮肥濃度，有 4 水準（i.e., $k = 1, 2, 3, 4$）。模型定義如下：

$$Y_{ijk} = \mu + \rho_i + \alpha_j + \eta_{ij} + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk} = \mu_{ijk} + \epsilon_{ijk}$$

其中 $i = 1, ..., 6$， $j = 1, 2, 3$， $k = 1, ..., 4$； μ 為全體總平均； ρ_i 為 block i 的效果； α_j 為作物 j 的效果； η_{ij} 為 block i 中第 j 個 plot 的效果，且 $\eta_{ij} \sim N(0, \sigma_\eta^2)$； β_k 為氮肥濃度 k 的效果； $(\alpha\beta)_{jk}$ 為作物 j 與氮肥濃度 k 的交互作用； ϵ_{ijk} 為 block i 中第 j 個 plot 中第 k 個 subplot 的效果，且 $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$。

4.2 檢查前提假設是否滿足

ANOVA 有若干個前提假設（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）

此分析獨變項為作物種類與氮肥濃度，前者為含有 3 個水準（level）的類別變數，後者原始雖為連續變數，因為我們只取 4 種濃度，因此可當作含有 4 個水準的類別變數；依變項為作物產量，為連續變項。符合。

2. 各組樣本依變項獨立

此分析中，不同種類的作物產量互不影響彼此，不同氮肥濃度下的作物產量也互不影響彼此，符合前提假設。

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

- 針對依變項進行常態檢定

有學者宣稱，在常態分布下，以 Bartlett 檢定變異數同質性會有較高的統計檢定力（power）（Lim & Loh，1996）。不過也有學者則認為以 Levene 檢定才有較高的統計檢定力。因此我們兩種檢定方式都進行。在進行變異數同質性檢定前，我們先以 Shapiro-Wilk 常態檢定法對依變項（Y）進行常態檢定，檢定的假說如下，顯著水準設定為 0.05。

$$H_0 : Y \sim ND \;\; v.s. \;\; H_1 : Y \; does \; not \sim ND$$

檢定結果：檢定統計量為 0.9838，其 p 值為 0.4807，不小於顯著水準，因此我們不拒絕 H_0，也就是說我們沒有足夠的證據證明母體分配不服從常態分佈。

- 針對依變項進行變異數同質檢定

我們以 Bartlett 與 Levene 兩種檢定方法來檢驗依變項變異數同質是否成立，也就是檢驗各作物種類與氮肥濃度之組合下，作物產量變異數是否相同。研究假說如下：

$$\begin{cases} H_0 : \sigma^2_{(jk)} = \sigma^2_{(jk)'}, \; \forall \; i = 1,2,3, \; k = 1,2,3,4, \; (jk) \neq (jk)' \\ H_1 : Not \; H_0 \end{cases}$$

我們同樣令顯著水準為 0.05。檢定結果如下表。兩種檢定方法的檢定統計量之 p 值都不小於顯著水準，因此我們在兩檢定中都不拒絕 H_0，意味著我們無法證明至少有一組母體變異數與其他組不同，也就是說我們沒有足夠的證據證明變異數同質性不存在，通過此前提假設。

檢定方法	檢定統計量	p 值
Bartlett	K^2=9.182	0.6051
Levene	F =0.6716	0.6716

4. 殘差（residuals）服從常態分配：待配適完模型後診斷。

以上步驟顯示，在我們的資料中，ANOVA 的前提假設均滿足（殘差常態假設待檢驗），因此我們可以進行 ANOVA。

4.3 研究假說

我們欲以 Between-Within 混合的 two-way ANOVA 檢驗 3 個研究問題，其與對應的虛無假設與對立假設陳列如下：

1. 不同種類的作物的產量是否有顯著差異：$H_0 : \alpha_j = 0 \quad \forall j = 1, 2, 3 \quad v.s. \quad H_1 : Not\ H_0$。
2. 不同氮肥濃度下，作物產量是否有顯著差異：$H_0 : \beta_k = 0 \ \forall k = 1, 2, 3, 4 \quad v.s. \quad H_1 : Not\ H_0$。
3. 氮肥與作物產量的關聯在不同種類的作物上是否有差異，亦即兩者之間有無交互作用：$H_0 : (\alpha\beta)_{jk} = 0 \ \ \forall j = 1, 2, 3, \ \forall k = 1, 2, 3, 4 \quad v.s. \quad H_1 : Not\ H_0$

4.4 分析結果

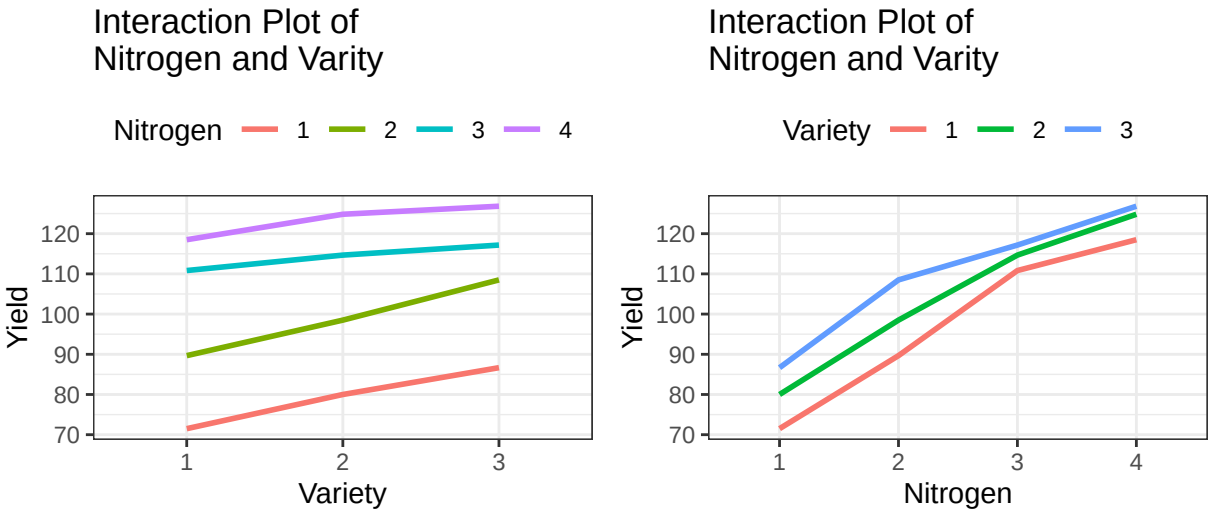
4.4.1 主要結果

變異來源	自由度	平方和	均方	F 值	p 值
殘差：blocks	5	15875	3175		
作物種類（V）	2	1786	893	1.49	0.27
殘差：blocks:plots	10	6013	601		
氮肥濃度（N）	3	20020	6673	37.69	0
V 與 N 交互作用	6	322	54	0.3	0.93
殘差：blocks:plots:subplots	45	7969	177		

我們的 ANOVA 模型通過診斷。而由以上的 ANOVA 結果表可做出以下結論：

1. 不同種類的作物產量並無差異，亦即 $\alpha_j = 0 \ ; \ \forall j = 1, 2, 3$。
2. 在施予不同氮肥濃度的情況下，作物產量有所差異，亦即在 $k = 1, 2, 3, 4$ 中，至少有一個 k 使得 $\beta_k \neq 0$。

3. 氮肥與作物產量的關聯在不同種類的作物上沒有差異，也就是說作物種類與氮肥濃度之間並無交互作用，亦即 $(\alpha\beta)_{jk} = 0 \ \ \forall j = 1, 2, 3, \ \forall k = 1, 2, 3, 4$。

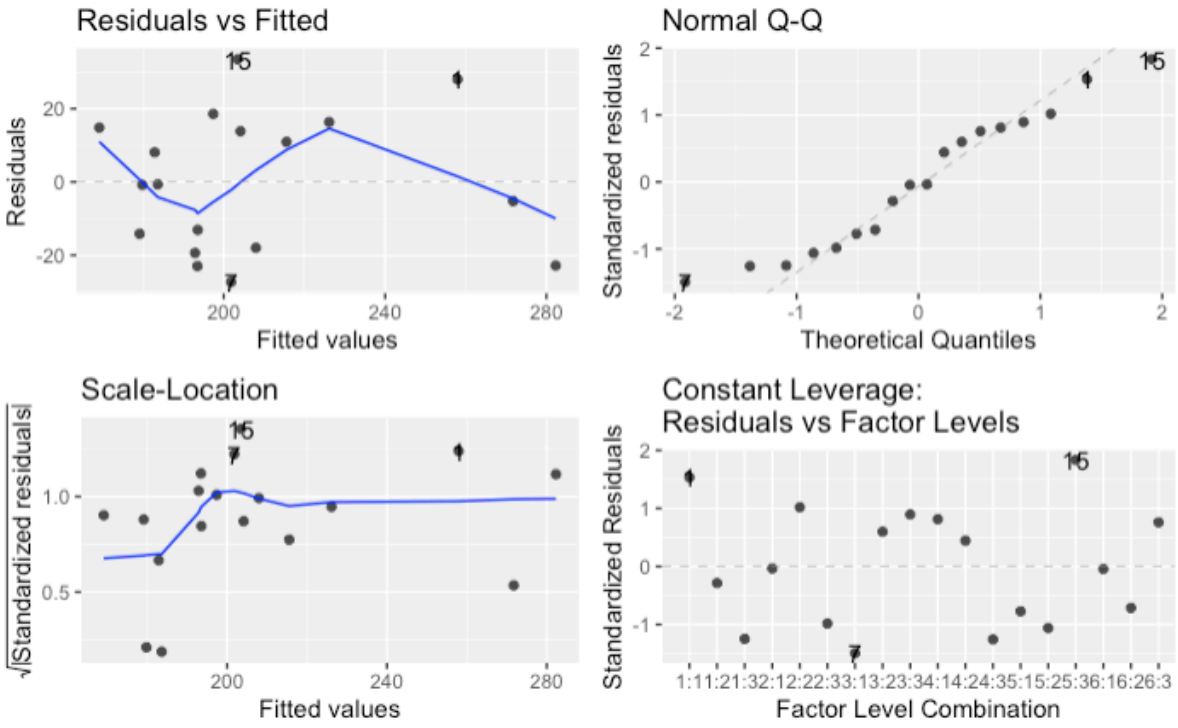


由以上兩張圖可再次確認，種子種類與氮肥之間並無交互作用，且對 3 種不同的種子而言，氮肥與產量的關係皆為正相關。

4.4.2 模型殘差診斷

因為模型假設中有兩個常態假設，我們需做兩次模型診斷，每次診斷都會做常態檢定與變異數同質性檢定，所有檢定的顯著水準均設定為 0.05。兩個常態假設為 $\eta_{ij} \sim N(0, \sigma_\eta^2)$ 與 $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$。

4.4.2.1 診斷 1：針對 η_{ij} 的診斷



由四張殘差圖的左上圖可以看出：殘差的期望值大致為零。左下圖顯示標準誤大約為定值，右上圖則顯示殘差大致服從常態分配。接著進行常態與同質變異數檢定來確認，顯著水準均設定為 0.05。

- 針對 η_{ij} 的常態檢定：以 Shapiro-Wilk 常態檢定法對殘差進行常態檢定，檢定假說：$H_0: Residuals \sim ND \quad v.s. \quad H_1: not\ H_0$。
- 針對 η_{ij} 的變異數同質性檢定：對殘差進行 NCV 變異數同質檢定（NCV test），檢定假設為：H_0: 殘差變異數具有同質性 v.s. H_1: 殘差變異數不具有同質性。

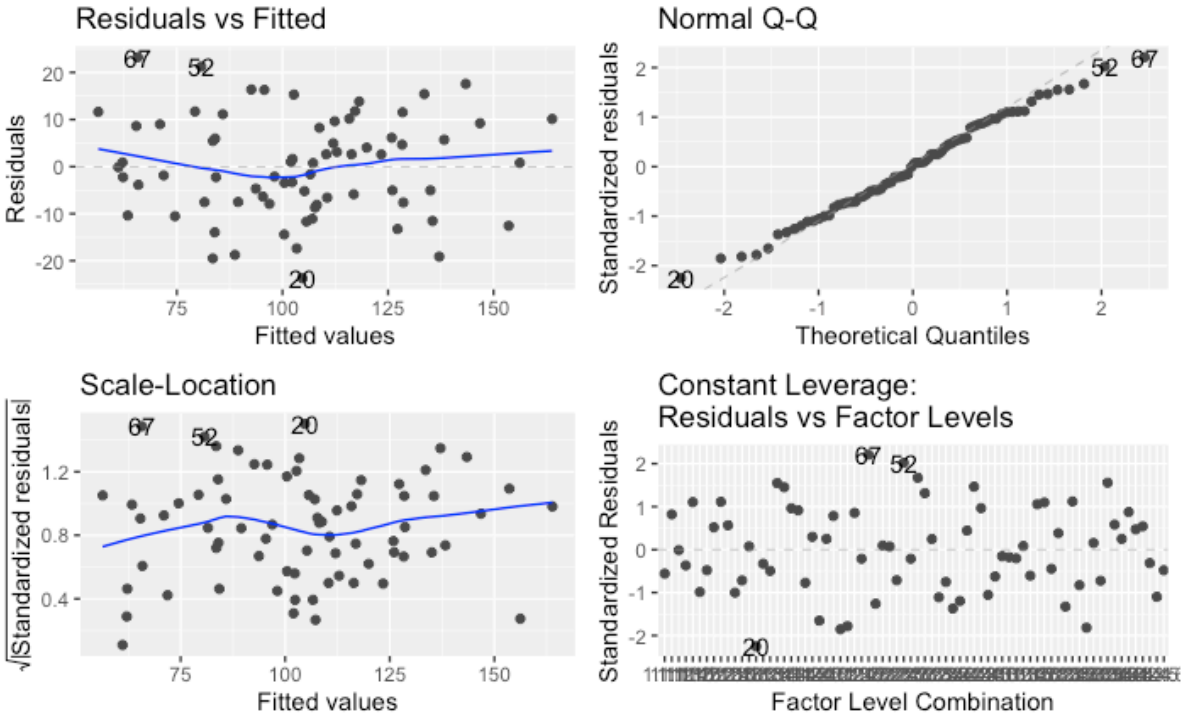
下表為檢定結果：

檢定方法	檢定統計量	p 值
Shapiro-Wilk	W=0.9446	0.3471
NCV	C=0.401	0.5266

在 Shapiro-Wilk 中，p 值大於顯著水準，不拒絕虛無假設，表示我們有足夠證據支持殘差服從常態分配。在 NCV 中，p 值大於顯著水準，不拒絕虛無假設，表示 η_{ij} 服從常態分配，且具有同質變異數。

4.4.2.2 診斷 2：針對 ϵ_{ijk} 的診斷

- ϵ_{ijk} 的殘差圖



由四張殘差圖的左上圖可以看出：殘差的期望值大致為零。左下圖顯示標準誤大約為定值，右上圖則顯示殘差大致服從常態分配。接著進行常態與同質變異數檢定來確認，顯著水準均設定為 0.05。

- 針對 ϵ_{ijk} 的常態檢定：以 Shapiro-Wilk 常態檢定法對殘差進行顯著水準為 0.05 的常態檢定，檢定的假說：$H_0: \ Residuals \sim ND \quad v.s. \quad H_1 : not \ H_0$。
- 針對 ϵ_{ijk} 的變異數同質性檢定：對殘差進行顯著水準為 0.05 的 NCV 變異數同質檢定（NCV test），其檢定假設為：H_0: 殘差變異數具有同質性 v.s. H_1: 殘差變異數不具有同質性。

檢定結果如下表：

檢定方法	統計量	p 值
Shapiro-Wilk	W=0.9899	0.8365
NCV	C=0.0844	0.5266

在 Shapiro-Wilk 中，因 p 值大於顯著水準，不拒絕虛無假設，表示我們有足夠證據支持殘差服從常態分配。在 NCV 中，因 p 值大於 0.05，不拒絕虛無假設，表示殘差擁有同質變異數。

所有檢定顯示我們的 ANOVA 模型通過診斷。

4.4.3 線性效果之檢驗

根據前面的分析，我們得知在不同氮肥濃度下，至少有一組平均產量與其他組別不同，且根據交互作用圖，我們發現氮肥與產量呈現正相關。在進行事後比較前，我們想進一步確認隨著氮肥濃度增加，產量是如何變化。所以我們在模型中氮肥的效益分成三個部分：線性、平方與立方，並將顯著水準設定為 0.05，分析結果如下表。

變異來源	自由度	平方和	均方	F 值	p 值
氮肥濃度	3	20021	6674	37.69	2.46e-12
氮肥濃度（線性）	1	19536	19536	110.32	1.09e-13
氮肥濃度（平方）	1	480	480	2.71	0.106
氮肥濃度（立方）	1	4	4	0.02	0.887
作物種類與氮肥濃度交互作用	6	322	54	0.30	0.932
殘差	45	7969	177		

根據資料計算結果，發現只有線性關係效果的檢定統計量的 p 值小於顯著水準，其餘 p 值皆大於顯著水準，表示氮肥與產量間僅有線性關係。加上交互作用圖，可說明氮肥與產量為線性正相關。我們接著進行事後比較來進一步了解氮肥濃度是否對平均產量有顯著差異。

4.4.4 事後比較

接著，我們透過事後比較來檢驗作物平均產量是否有隨著氮肥濃度的提升而有顯著增加。因為氮肥各組樣本大小皆相同，因此我們選用 Tukey 方法進行事後比較，並將整體的信心水準設在 95%。以下為檢定假說：

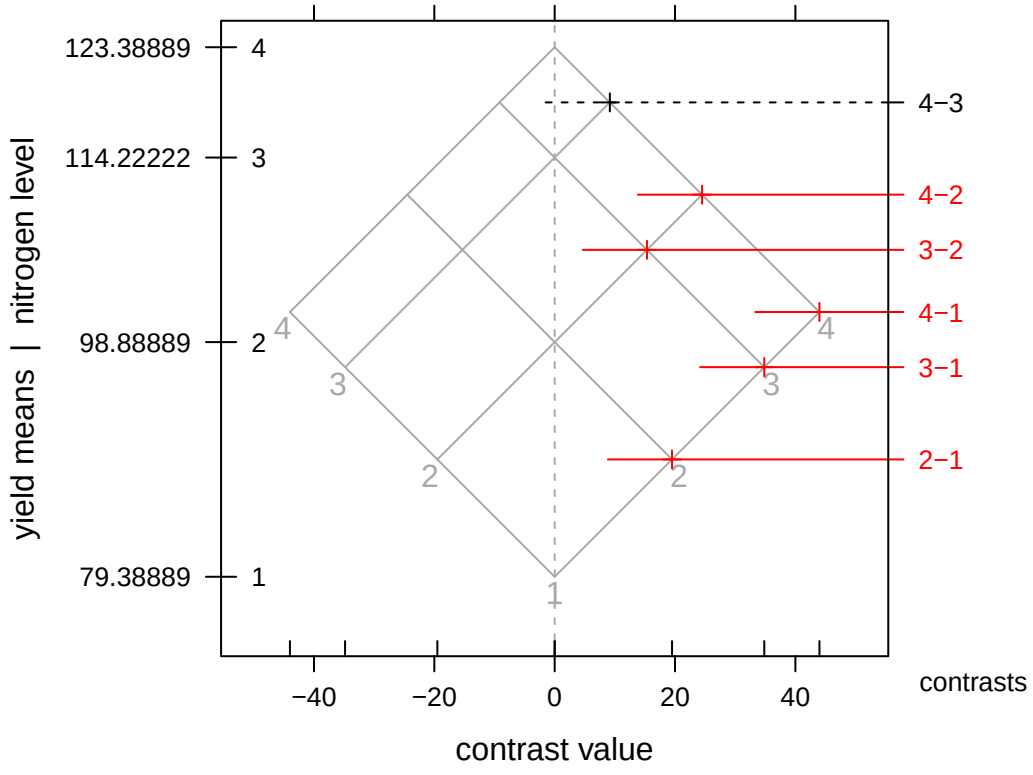
$$H_0 : \mu_{k_1} \leq \mu_{k_2} \quad v.s \quad H_1 : \mu_{k_1} > \mu_{k_2}$$

$$\forall k_1 > k_2, \quad (k_1, k_2) = (4,3), (4,2), (4,1), (3,2), (3,1), (2,1)$$

其中 μ_1 為氮肥濃度為 0 的平均產量，μ_2 為氮肥濃度為 0.2 的平均產量，μ_3 為氮肥濃度為 0.4 的平均產量，μ_4 為氮肥濃度為 0.6 的平均產量。

下表為兩兩平均差異的 95% 信賴區間。4 為 μ_4，表示氮肥濃度為 0.6 的平均作物產量；3 為 μ_3，表示氮肥濃度為 0.4 的平均作物產量；2 為 μ_2，表示氮肥濃度為 0.2 的平均作物產量；1 為 μ_1，表示氮肥濃度為 0 的平均作物產量。

	組間差異平均估計值	CI 下界	CI 上界
4-3	9.1667	-1.5052	Inf
4-2	24.5000	13.8282	Inf
3-2	15.3333	4.6615	Inf
4-1	44.0000	33.3282	Inf
3-1	34.8333	24.1615	Inf
2-1	19.5000	8.8282	Inf



上圖為事後比較分析結果的視覺化，圖中的線為各組兩兩平均差異的 95% 信賴區

間，其中標注為紅色者表示其包含 0，也就是說在 95% 信心水準，兩組平均有顯著差異。

根據事後比較結果的圖表，我們發現，只有濃度 0.6 的氮肥與濃度 0.4 的氮肥平均產量差異的 95% 信賴區間（i.e., $[-1.5150, \infty)$）包含 0，表示我們無足夠證據支持氮肥濃度 0.6 下的平均產量顯著高於氮肥濃度 0.4 下的平均產量；而其餘兩兩氮肥濃度的配對組別的 95 信賴區間皆不包含 0，表示我們有足夠證據支持：除了濃度 0.6 與 0.4 的氮肥之外，基本上濃度高的氮肥平均作物產量大於濃度低的氮肥。

我們最終得到的比較結果：

$$\mu_4 > \mu_2 > \mu_1, \quad \mu_3 > \mu_2 > \mu_1$$

大致上來說，隨著氮肥濃度增加，平均產量也會提高。但當氮肥濃度由 0.4 提升為 0.6 時，平均產量並無顯著提高。

5 結論

1. 不同種類的作物的產量沒有差異。
2. 氮肥與作物產量有正向線性關聯：氮肥濃度越高，作物產量越多。四種氮肥濃度的關係為：$\mu_4 > \mu_2 > \mu_1, \quad \mu_3 > \mu_2 > \mu_1$。
3. 氮肥與作物產量的關聯在不同種類的作物上沒有差異。

6 附錄

6.1 線性效果的模型介紹

上述分析有提到在線性效果檢驗部分，我們將氮肥濃度的效益分為線性、平方與立方。於此我們簡單介紹其模型，原本為 ANOVA 模型，但在此我們轉用線性迴歸解釋。以下為模型公式：

$$Y_i = \beta_0 + \sum_{j=1}^{5} \beta_j x_{ji} + \sum_{j=6}^{7} \beta_j x_{ji} + \sum_{j=1}^{5}\sum_{k=6}^{7} \beta_{j,k} x_{ji} x_{ki} + \beta_8 x_{8i} + \beta_9 x_{8i}^2 + \beta_{10} x_{8i}^3 + \sum_{j=6}^{7} \beta_{j,8} x_{8i} + \beta_{j,9} x_{8i}^2 + \beta_{j,10} x_{8i}^3$$

其中 $\beta_1, \ldots, \beta_5$ 為第 2 到第 4 個集區的係數，以第 1 個集區作為基本（baseline）。β_6, β_7 為第 2 種、第 3 種作物的係數，β_8 為氮肥濃度的線性係數，β_9 為氮肥濃度的平方係數，$\beta_1 0$ 為氮肥濃度的立方係數。$\beta_{j,k}$ 為第 j 和 k 項的交互作用項。$\mathbf{X}$ 為設計矩陣（design matrix），x_{ij} 為第 i 個觀察值第 j 個變數的值。

$$x_{ij} = \begin{cases} 0, \ if \ observation_i \ not \ \in variable_j \\ 1, \ if \ observation_i \in variable_j \end{cases}$$

7 参考資料

1. Heiberger, R. M. & Burt Holland, B. H. (2015). Statistical Analysis and Data Display An Intermediate Course with Examples in R. Springer.

2. Rutherford, A. (2001). Introducing ANOVA and ANCOVA: a GLM approach. Sage.

3. Lim, T. S., & Loh, W. Y. (1996). A comparison of tests of equality of variances. Computational Statistics & Data Analysis, 22(3), 287-301.

4. Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association, 50(272), 1096-1121.

5. Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: the Bonferroni method. Bmj, 310(6973), 170.