武漢理工大學

统计模拟与 R 语言

大作业

姓 名： 代 文 博

学 号： 0122210880110

专 业： 人 工 智 能

班 级： 2201

得 分：

《统计模拟与 R 语言》大作业

2023-2024 学年第 2 学期

【考试说明】

1. 你将有三天的时间完成这个大作业，5 月 25 日上午 8 点下发，5 月 27 日晚上 11：59 分之前提交。
2. 将你的答案以 pdf 文件先提交到 http://www.xzc.cn/HFSanONFd1，文件名以“班级-姓名-学号”形式，纸质版上交请看群通知后面再交，纸质版需与 pdf 电子版保持一致，否则成绩无效。
3. 每题的答案列在题目下方，工整排版。所提交的答案要有必要的统计步骤、公式、假设、检验、R 语言分析截图、结果分析与结论等。所有 R 语言输出的结果必须有相应的文字分析;不必要的 R 语言结果输出应该被避免。
4. 你可以利用你的笔记、任何书籍或线上资源，但内容必须由你自己完成。不能与其他任何人(无论是否与课程有关)合作或讨论这个大作业。
5. 一定要在你的答案中写清楚解题思路，除此之外还应该包括 R 语言代码和推理，所有的结果都需要解释。
6. 在截止日期后提交，成绩将按每小时 5%的比例扣分。

【分数设置】

大题	小题	分值	成绩
1	（1）	10	
	（2）	15	
	（3）	20	
2	（1）	10	
	（2）	30	
3	（1）	5	
	（2）	5	
	（3）	5	

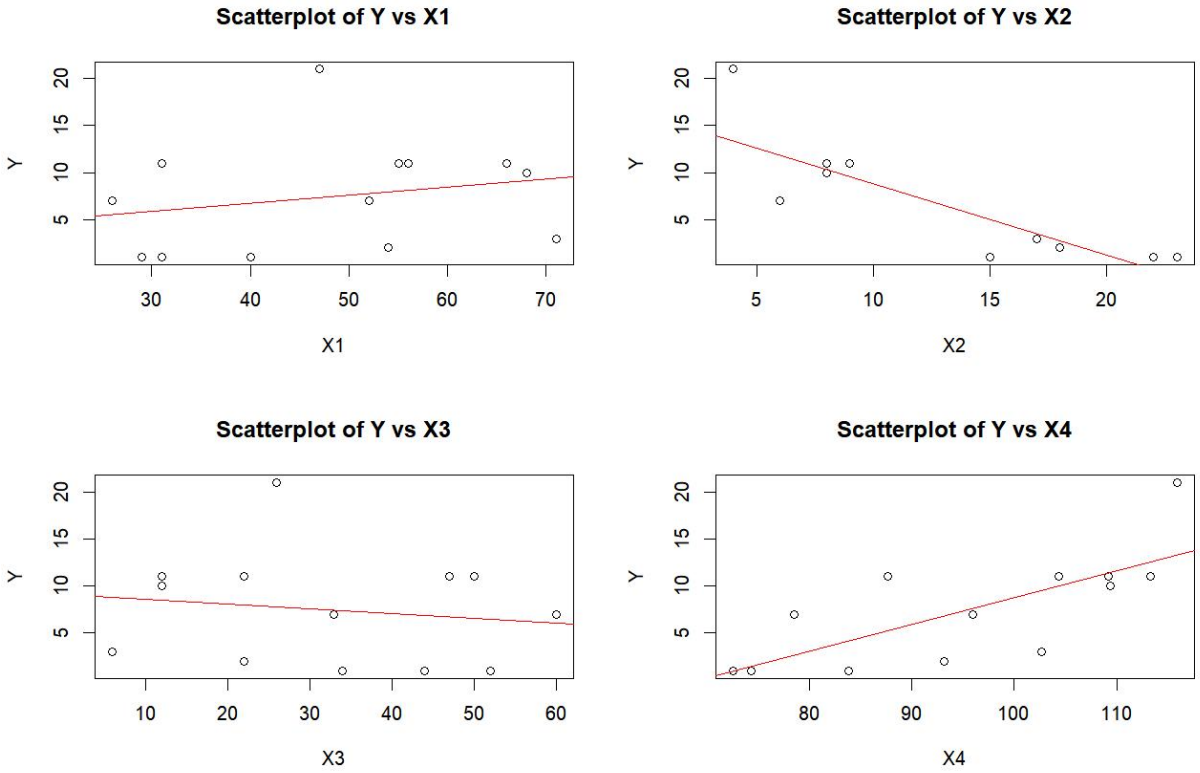
【考试题目】

1. 某校化生学院学生的实验课题为探究某种草药提取物在抑制细菌生长方面的效果是否与提取物中的四种成分X_1, X_2, X_3, X_4有关，现测得 13 组数据，见 herb.csv 文件，其中Y为草药提取物的抑菌效果，X_1, X_2, X_3, X_4为提取物中的四种成分的含量。

（1） 试用探索性数据分析方法，分别探究抑菌效果Y与每种成分含量之间是否

存在线性关系。

```
> # 1. 读取数据
> data <- read.csv("C:/Users/levon/Desktop/课内实验/统计模拟与R语言/herb.csv", header = TRUE)
> # 2. 数据清洗
> data$Y <- as.numeric(as.character(data$Y))
> data$X1 <- as.numeric(as.character(data$X1))
> data$X2 <- as.numeric(as.character(data$X2))
> data$X3 <- as.numeric(as.character(data$X3))
> data$X4 <- as.numeric(as.character(data$X4))
> # 3. 绘制散点图
> par(mfrow = c(2, 2)) # 设置图形布局为2行2列
> plot(data$X1, data$Y, xlab = "X1", ylab = "Y", main = "Scatterplot of Y vs X1")
> abline(lm(Y ~ X1, data = data), col = "red") # 添加趋势线
> plot(data$X2, data$Y, xlab = "X2", ylab = "Y", main = "Scatterplot of Y vs X2")
> abline(lm(Y ~ X2, data = data), col = "red")
> plot(data$X3, data$Y, xlab = "X3", ylab = "Y", main = "Scatterplot of Y vs X3")
> abline(lm(Y ~ X3, data = data), col = "red")
> plot(data$X4, data$Y, xlab = "X4", ylab = "Y", main = "Scatterplot of Y vs X4")
> abline(lm(Y ~ X4, data = data), col = "red")
```



• 散点图分析：

YvsX4：散点图显示数据点大致沿着一条直线排列，这可能表明 Y 和 X4 之间存在正的线性关系。

YvsX2：数据点呈现出从左上到右下的带状分布，这通常意味着负的线性关系。

YvsX1 和 YvsX3：数据点分布较为随机，没有明显的线性模式，这可能表明 Y 和 X3 或 X4 之间没有强烈的线性关系。

```
> # 4. 计算相关系数
> cor_Y_X1 <- cor(data$Y, data$X1, use = "complete.obs") # 对于缺失或NA值使用complete.obs
> cor_Y_X2 <- cor(data$Y, data$X2, use = "complete.obs")
> cor_Y_X3 <- cor(data$Y, data$X3, use = "complete.obs")
> cor_Y_X4 <- cor(data$Y, data$X4, use = "complete.obs")
> # 打印相关系数
> print(paste("Correlation between Y and X1:", cor_Y_X1))
[1] "Correlation between Y and X1: 0.228579470307565"
> print(paste("Correlation between Y and X2:", cor_Y_X2))
[1] "Correlation between Y and X2: -0.824133764417199"
> print(paste("Correlation between Y and X3:", cor_Y_X3))
[1] "Correlation between Y and X3: -0.150520316673746"
> print(paste("Correlation between Y and X4:", cor_Y_X4))
[1] "Correlation between Y and X4: 0.730717471965077"
```

• **分析：**

Y 和 X1 的相关系数约为 0.23，表明它们之间存在微弱的正线性关系。

Y 和 X2 的相关系数约为-0.82，表明它们之间存在较强的负线性关系。

Y 和 X3 的相关系数约为-0.15，表明它们之间存在非常弱的负线性关系，几乎可以认为没有线性关系。

Y 和 X4 的相关系数约为 0.73，表明它们之间存在中等强度的正线性关系。

• **结论**

Y 和 X2 之间存在较强的负线性关系，这表明随着 X2 含量的增加，抑菌效果（Y）可能会降低。

Y 和 X4 之间存在中等强度的正线性关系，这表明随着 X4 含量的增加，抑菌效果可能会提高。

Y 和 X1 以及 Y 和 X3 之间的关系较弱，可能没有明显的线性关系，或者需要更多的数据来确定它们之间的关系。

（2） 建立抑菌效果Y与四种成分含量X_1, X_2, X_3, X_4的回归方程，并对回归方程进行显著性检验与回归诊断。

```
> # 1. 读取数据
> data <- read.csv("C:/Users/levon/Desktop/课内实验/统计模拟与R语言/herb.csv", header = TRUE)
> # 建立回归模型
> model <- lm(Y ~ 1 + X1 + X2 + X3 + X4, data = data)
> # 查看模型摘要
> summary(model)

Call:
lm(formula = Y ~ 1 + X1 + X2 + X3 + X4, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9148 -0.7546 -0.3231  0.8385  1.9198

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -25.17332    8.52003  -2.955 0.018296 *
X1           -0.32988    0.06293  -5.242 0.000781 ***
X2           -0.20221    0.10780  -1.876 0.097535 .
X3            0.01316    0.05394   0.244 0.813354
X4            0.52896    0.07337   7.210 9.16e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.355 on 8 degrees of freedom
Multiple R-squared:  0.9646,    Adjusted R-squared:  0.947
F-statistic: 54.56 on 4 and 8 DF,  p-value: 7.593e-06
```

Y与四种成分含量X_1, X_2, X_3, X_4的回归方程为：

Y = -25.17332 - 0.32988X1 - 0.20221X2 + 0.01316X3 + 0.52896X4

对于**回归方程显著性检验**的详细步骤：

1.假设：

零假设（H0）：回归方程中自变量系数全为零（即自变量对因变量无显著影响）。

备择假设（H1）：回归方程中自变量系数不全为零（即自变量对因变量存在显著性影响）。

2.统计量：显著性检验通常使用 F 检验进行。F 检验的统计量为 F 值，计算公式为：

$$F=(SSR/k)/(SSE/(n-k-1))$$

其中，SSR 为回归平方和，SSE 为残差平方和，k 为自变量的个数，n 为样本容量。

3.确定拒绝域：在进行 F 检验时，我们需要根据显著性水平 α 和自由度 k 和 n-k-1 来确定 F 分布上的临界值。根据临界值，我们可以确定拒绝域。如果计算得到的 F 值落在拒绝域内，则拒绝原假设。

4.决策：根据计算得到的 F 值和拒绝域的临界值，我们做出决策。如果计算得到的 F 值落在拒绝域内，则拒绝原假设；否则接受原假设。

5. 结论分析：根据决策的结果，我们可以得出结论。如果拒绝了零假设，我们可以认为对应的自变量系数在回归模型中是显著的，即自变量对因变量有显著影响；如果接受了零假设，则说明自变量对因变量的影响不显著。

- 对于**回归系数显著性检验**的详细步骤：

1. 假设

零假设(H0):自变量X_j的系数β_j等于零，即该自变量对因变量 Y 没有影响.

备择假设(H1):自变量X_j的系数β_j不等于零，即该自变量对因变量 Y 有显著影响.

2. 统计量的计算

对于模型中每个自变量X_j，计算其系数$\widehat{\beta}_j$的 t 统计量:

$$t_j=\frac{\widehat{\beta}_j}{SE(\widehat{\beta}_j)}$$

其中,$\widehat{\beta}_j$是估计的回归系数,$SE(\widehat{\beta}_j)$是该估计系数的标准误。

3.确定拒绝域

选择一个显著性水平 α 常用的有 0.05、0.01），这代表了犯第一类错误（错误地拒绝一个真实的零假设）的概率上限。根据所选的显著性水平和自由度（df=n-p-1，其中 n 是样本大小,p 是模型中自变量的数量加 1），从 t 分布表中确定临界值。

4.决策

如果计算出的 t 统计量$|t_j|$大于临界值，或者计算出的 p 值小于显著性水平 α，则拒绝零假设H_0.

如果计算出的 t 统计量$|t_j|$小于或等于临界值，或者计算出的 p 值大于显著性水平 α，则不能拒绝零假设H_0.

5.结论分析

如果拒绝零假设H_0，则得出结论:有统计学证据表明自变量X_j对因变量 Y 有显著影响，其系数显著不同于零.

如果不拒绝零假设H_0，则得出结论:没有足够的统计学证据表明自变量X_j对因变量 Y 有显著影响，其系数不显著不同于零。

·模型摘要信息：

Residualstandarderror（残差标准误差）:1.355，表示观测值与模型预测值之间差异的标准差。

MultipleR-squared（多重 R 平方）:0.9646，表示模型解释的变异占总变异的比例，值越接近 1 越好。

AdjustedR-squared（调整后的 R 平方）:0.947，考虑了模型中变量的数量，对多重 R 平方进行了调整。

F-statistic（F 统计量）:54.56，用于整体模型的显著性检验。

p-value（p 值）:7.593e-06，表示模型整体的显著性水平，值越小表示模型整体越显著。

·显著性检验分析：

截距项(Intercept):估计系数为-25.173,标准误差为 8.520,t 值为-2.955,p 值为 0.018296。截距项在 0.05 的显著性水平下是显著的,因为 p 值小于 0.05。

X1:估计系数为-0.3299,标准误差为 0.0629,t 值为-5.242,p 值为 0.000781。X1 在 0.01 的显著性水平下对 Y 有显著影响，因为 p 值远小于 0.01。

X2:估计系数为-0.2022,标准误差为 0.1078,t 值为-1.876,p 值为 0.097535。X2 的 p 值为 0.097535，大于 0.05，因此在 0.05 的显著性水平下不显著。

X3:估计系数为 0.0132,标准误差为 0.0539,t 值为 0.244,p 值为 0.813354。X3 的 p 值远大于 0.05，因此在 0.05 的显著性水平下不显著。

X4:估计系数为 0.5290,标准误差为 0.0734,t 值为 7.210,p 值为 9.16e-05。X4 在 0.01 的显著性水平下对 Y 有显著影响，因为 p 值远小于 0.01。

·结论：

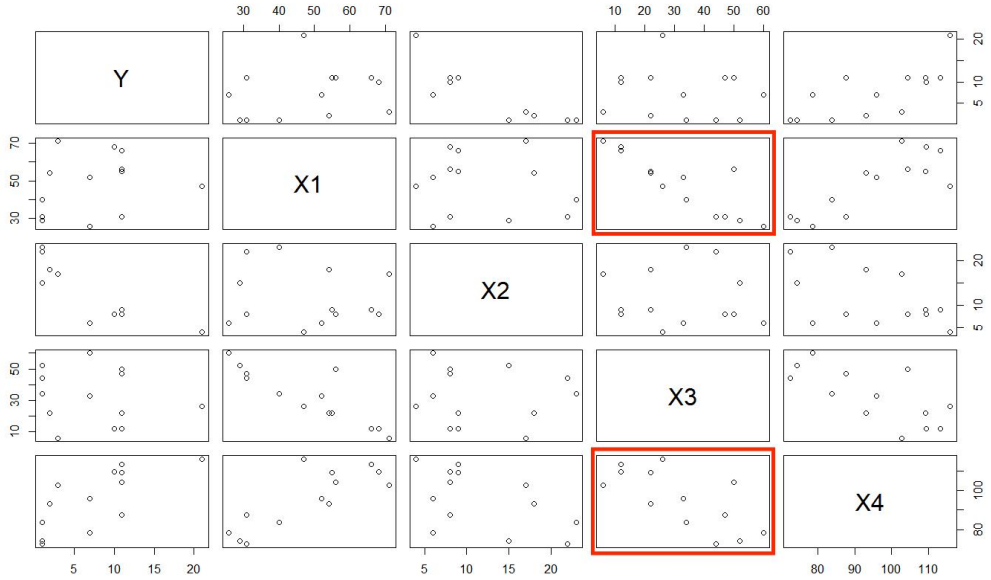
模型整体显著（p 值远小于 0.05）。

X1 和 X4 对抑菌效果 Y 有显著的线性影响。

X2 和 X3 在当前数据下对 Y 的影响不显著。

- **回归诊断：**

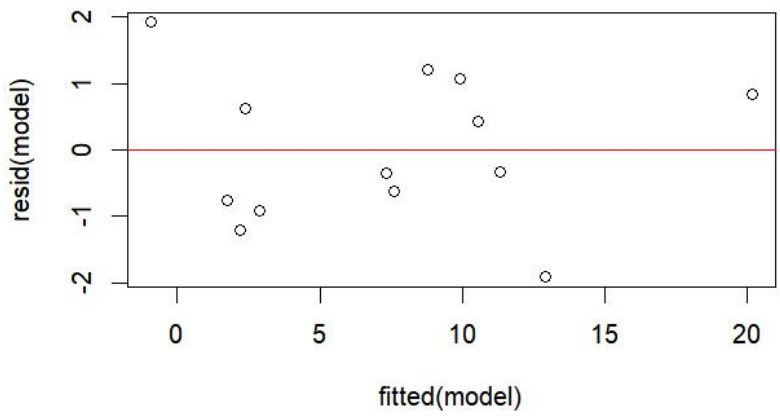
```
> pairs(data)
```




```
> cor(data)
             Y          X1          X2          X3          X4
Y    1.0000000  0.2285795 -0.82413376 -0.15052032  0.7307175
X1   0.2285795  1.0000000 -0.13924238 -0.86833771  0.8162526
X2  -0.8241338 -0.1392424  1.00000000 -0.05645124 -0.5346707
X3  -0.1505203 -0.8683377 -0.05645124  1.00000000 -0.7092348
X4   0.7307175  0.8162526 -0.53467068 -0.70923481  1.0000000
```

从散点图矩阵和相关系数矩阵中可以初步观察到该模型具有多重共线性。

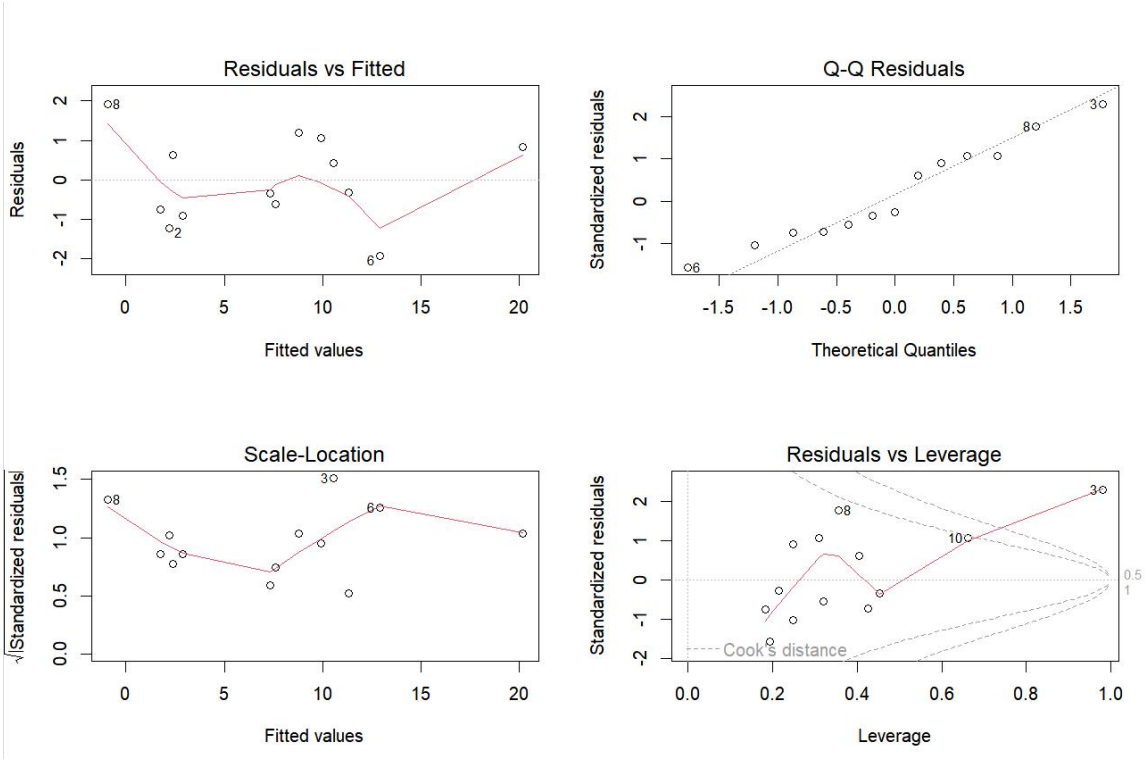
```
> # 残差图
> plot(resid(model) ~ fitted(model))
> abline(h = 0, col = "red")
```



• 分析与结论：

残差图上的点的散步并没有呈现出一定的趋势（随横坐标的增大而增大或减小），故初步判断回归模型不存在异方差。

```
> # 残差的诊断图
> par(mfrow = c(2, 2))
> plot(model)
```



1.ResidualsvsFitted（残差 vs 拟合值）

这张图展示了残差（实际观测值与模型预测值之差）随着模型拟合值的变化情况。此图表明残差的分散程度随着拟合值的变化基本保持恒定，可以初步确定该模型具有方差齐性。

2.Q-QResiduals（残差 Q-Q 图）

Q-Q 图（Quantile-QuantilePlot）用于评估残差的正态性。图中的点近似落在一条直线上，表明残差服从正态分布。

3.Scale-Location（尺度-位置图）

这张图用于评估残差的同方差性（方差齐性），在图中我们看到残差的尺度（标准差）与拟合值的位置（水平）无关，这表明模型具有方差齐性。

4.ResidualsvsLeverage（残差 vs 杠杆值）

这张图展示了残差与杠杆值（Leverage）之间的关系。杠杆值衡量的是每个观测值对模型拟合的影响程度。理想情况下，我们希望看到残差随机分布，没有与杠杆值相关的模式。而图中第三个点，第八个点和第八个点远离其他点，并且具有较高的杠杆值，这可能表明他们是强影响点或离群点，这些点可能对模型拟

合有较大影响，需要进一步调查。并且第三个点的 cook 距离大于 1，表明该观测点可能是一个强影响点，对模型的拟合有较大的影响，需要重点关注。

```
> # 学生化残差
> rstandard(model)
         1          2          3          4          5          6          7          8
-0.3484894 -1.0326869  2.2850580  0.9043441 -0.5556912 -1.5743211  0.6030409  1.7653760
         9         10         11         12         13
-0.7417813  1.0639476 -0.7350285 -0.2692037  1.0645230
```

• 计算结果分析：

残差大小：学生化残差的绝对值大于 2 或 3 通常表示该观测点可能是一个异常值或强影响点。在此数据集中，第 3、5、6 和 8 个观测值的学生化残差的绝对值较大，表明这些点可能对模型拟合有较大影响。

• 结论：

根据提供的学生化残差，第 3、5、6 和 8 个观测点可能需要进一步的调查，以确定它们是否为异常值或强影响点。如果这些观测点确实对模型有较大影响，可能需要考虑将它们从模型中移除，或者寻找这些点影响模型的原因。

```
> # Breusch-Pagan检验
> bptest(model)

        studentized Breusch-Pagan test

data:  model
BP = 2.3944, df = 4, p-value = 0.6636
```

• 检验结果分析：

BP 值：检验的统计量为 2.3944。这是基于残差平方和与自变量之间关系的度量。

自由度（df）：检验的自由度为 4。自由度等于模型中自变量的数量减 1。

p 值：检验的 p 值为 0.6636。这是一个非常高的 p 值，远大于常用的显著性水平（如 0.05 或 0.01）。

• 结论：

由于 p 值远大于 0.05，我们没有足够的证据拒绝原假设，即我们没有证据表明模型存在异方差性。换句话说，根据 Breusch-Pagan 检验的结果，我们可以认为模型的误差项具有方差齐性（Homoscedasticity），即误差项的方差不随自变量的变化而变化。

```
> # Durbin-Watson检验
> dwtest(model)

        Durbin-Watson test

data:  model
DW = 2.2198, p-value = 0.609
alternative hypothesis: true autocorrelation is greater than 0
```

• 检验结果分析：

DW 值：检验的 Durbin-Watson 统计量为 2.2198。这个值表明残差之间可能存在轻微的正自相关，因为 DW 值大于 2（但接近 2）。

p 值：检验的 p 值为 0.609，这是一个非常高的 p 值，远大于常用的显著性水平（如 0.05 或 0.01）。

备择假设：这里的备择假设是存在大于 0 的自相关，即残差之间存在正相关。

• 结论：

由于 p 值远大于 0.05，我们没有足够的证据拒绝原假设，即我们没有证据表明模型残差存在显著的一阶自相关。换句话说，根据 Durbin-Watson 检验的结果，我们可以认为模型的残差不存在显著的自相关性。

```
> # 方差膨胀因子
> vif(model)
      X1       X2       X3       X4
6.269781 3.117339 5.696225 7.965274
```

• 检验结果分析与结论：

X1、X3 和 X4 的 VIF 值均大于 5，表明这些变量可能与其他变量存在较高的共线性，这可能会影响模型系数的稳定性和解释性。X2 的 VIF 值为 3.117339，虽然低于 5，但仍接近通常的警戒线，可能也需要关注。

```
> # 残差的正态性检验
> shapiro.test(resid(model))

        Shapiro-Wilk normality test

data:  resid(model)
W = 0.97496, p-value = 0.9456
```

• 检验结果分析：

W 值：Shapiro-Wilk 检验的 W 值为 0.97496。W 值越接近 1，表示残差越接近正态分布。

p 值：检验的 p 值为 0.9456，这是一个非常高的 p 值，远大于常用的显著性水平（如 0.05 或 0.01）。

• 结论：

由于 p 值远大于 0.05，我们没有足够的证据拒绝原假设，即我们没有证据表明模型残差显著偏离正态分布。换句话说，根据 Shapiro-Wilk 正态性检验的结果，我们可以认为模型的残差满足正态分布的假设。

（3） 如果有上述检验存在问题，做相应处理后找到最优回归模型，对最优回归模型进行显著性检验和残差诊断，并解释模型的含义。

• 分析

在第二问中模型进行了显著性检验和回归诊断之后，发现最大的问题在于该模型具有多重共线性，因而考虑使用**逐步回归**。

```
> # 读取数据
> data <- read.csv("C:/Users/levon/Desktop/课内实验/统计模拟与R语言/herb.csv", header = TRUE)
>
> # 假设model是我们的初始模型，包含了所有候选自变量
> initial_model <- lm(Y ~ 1 + X1 + X2 + X3 + X4, data = data)
>
> # 进行逐步回归
> stepwise_model <- step(initial_model, direction = "both", scope = list(lower = ~1, upper
= ~X1 + X2 + X3 + X4))
Start:  AIC=11.58
Y ~ 1 + X1 + X2 + X3 + X4

       Df Sum of Sq     RSS    AIC
- X3    1     0.109  14.791  9.678
<none>               14.682 11.581
- X2    1     6.457  21.139 14.320
- X1    1    50.433  65.115 28.946
- X4    1    95.396 110.078 35.771

Step:  AIC=9.68
Y ~ X1 + X2 + X4

       Df Sum of Sq     RSS    AIC
<none>               14.791  9.678
+ X3    1     0.109  14.682 11.581
- X2    1    10.352  25.142 14.575
- X1    1    69.490  84.281 30.300
- X4    1   112.930 127.721 35.704
>
```

• 逐步回归过程分析：

初始模型：开始时，模型包含了所有自变量 X1,X2,X3,X4。

模型评估：模型使用赤池信息准则（AkaikeInformationCriterion,AIC）来评估。AIC 是一个衡量模型拟合优度的指标，同时对模型复杂度进行惩罚。较低的 AIC 值通常表示更好的模型。

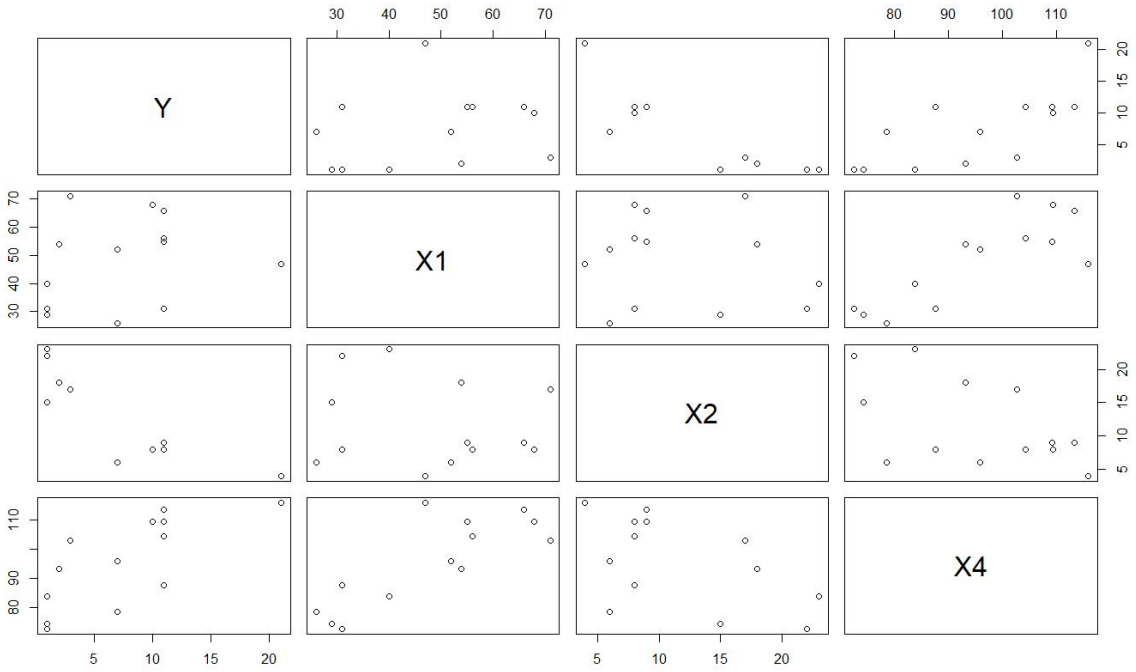
逐步删除：

第一步，模型删除了 X3，AIC 从 11.581 降低到 9.678，残差平方和（RSS）从 14.682 增加到 14.791，自由度（Df）减少了 1。

第二步，模型考虑重新添加 X3 或删除 X2 或 X1 或 X4。根据 AIC 值，删除 X2 会使得 AIC 从 9.678 增加到 14.320,因此 X2 被保留。删除 X1 会使得 AIC 从 9.678 增加到 28.946，所以 X1 也被保留。删除 X4 会使得 AIC 从 9.678 增加到 35.771，因此 X4 也被保留。模型没有进行任何改变，因为添加 X3 会使 AIC 从 9.678 增加到 11.581。最终模型：经过逐步回归后，最终模型只包含
X1,X2,X4 这三个自变量。

• 多重共线性检验：

```
> # 读取数据
> data <- read.csv("C:/Users/levon/Desktop/课内实验/统计模拟与R语言/herb.csv", header = TRUE)
> data <- data[, -4]
> pairs(data)
> cor(data)
           Y         X1         X2         X4
Y   1.0000000  0.2285795 -0.8241338  0.7307175
X1  0.2285795  1.0000000 -0.1392424  0.8162526
X2 -0.8241338 -0.1392424  1.0000000 -0.5346707
X4  0.7307175  0.8162526 -0.5346707  1.0000000
```




```
> # 方差膨胀因子
> vif(stepwise_model)
      X1       X2       X4
4.760574 2.224750 6.537034
```

X1 的 VIF 值为 4.76，这通常被认为是可接受的，因为 VIF 值小于 5。X2 的 VIF 值为 2.22，这也表明 X2 与其他自变量的共线性较弱。X4 的 VIF 值为 6.54，这表明 X4 与其他自变量存在一定程度的多重共线性，但是相较于逐步回归之前的模型，多重共线性的问题已经得到了很大程度上的改善，并且，再次观察散点图矩阵和相关系数矩阵，并没有发现明显的多重共线性，因此，基本可以认为此问题得到了解决，**该模型即为最优回归模型**。

对于**最优回归模型显著性检验**的详细步骤：

1.假设：在回归分析中，显著性检验是用来确定回归方程中自变量的系数是否显著不为零，即自变量对因变量的影响是否存在显著性。显著性检验的假设如下：零假设（H0）：回归方程中自变量系数全为零（即自变量对因变量无显著影响）。备择假设（H1）：回归方程中自变量系数不全为零（即自变量对因变量存在显著性影响）。

2.统计量：显著性检验通常使用 F 检验进行。F 检验的统计量为 F 值，计算公式为：
F=(SSR/k)/(SSE/(n-k-1))
其中，SSR 为回归平方和，SSE 为残差平方和，k 为自变量的个数，n 为样本容量。

3.确定拒绝域：在进行 F 检验时，我们需要根据显著性水平 α 和自由度 k 和 n-k-1

来确定 F 分布上的临界值。根据临界值，我们可以确定拒绝域。如果计算得到的 F 值落在拒绝域内，则拒绝原假设。

4. 决策：根据计算得到的 F 值和拒绝域的临界值，我们做出决策。如果计算得到的 F 值落在拒绝域内，则拒绝原假设；否则接受原假设。

5. 结论分析：根据决策的结果，我们可以得出结论。如果拒绝了零假设，我们可以认为对应的自变量系数在回归模型中是显著的，即自变量对因变量有显著影响；如果接受了零假设，则说明自变量对因变量的影响不显著。在结论分析中，还可以评价回归方程的拟合程度和解释能力。

• 对于**回归系数显著性检验**的详细步骤：

1. 假设：

零假设(H0):自变量X_j的系数β_j等于零，即该自变量对因变量 Y 没有影响.

备择假设(H1):自变量X_j的系数β_j不等于零，即该自变量对因变量 Y 有显著影响.

2. 构造统计量：

对于模型中每个自变量X_j，计算其系数$\widehat{\beta}_j$的 t 统计量:

$$t_j=\frac{\widehat{\beta}_j}{SE(\widehat{\beta}_j)}$$

其中,$\widehat{\beta}_j$是估计的回归系数,$SE(\widehat{\beta}_j)$是该估计系数的标准误。

3. 确定拒绝域

选择一个显著性水平 α 常用的有 0.05、0.01），这代表了犯第一类错误（错误地拒绝一个真实的零假设）的概率上限.

根据所选的显著性水平和自由度（df=n-p-1，其中 n 是样本大小,p 是模型中自变量的数量加 1），从 t 分布表中确定临界值。

4. 决策

如果计算出的 t 统计量$|t_j|$大于临界值，或者计算出的 p 值小于显著性水平 α，则拒绝零假设H_0.

如果计算出的 t 统计量$|t_j|$小于或等于临界值，或者计算出的 p 值大于显著性水平 α，则不能拒绝零假设H_0.

5. 结论分析

如果拒绝零假设H_0，则得出结论:有统计学证据表明自变量X_j对因变量 Y 有显著影响，其系数显著不同于零.

如果不拒绝零假设H_0，则得出结论:没有足够的统计学证据表明自变量X_j对

因变量 Y 有显著影响，其系数不显著不同于零。

• 显著性检验：

```
> # 查看逐步回归后的模型
> summary(stepwise_model)

Call:
lm(formula = Y ~ X1 + X2 + X4, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9334 -0.7238 -0.3594  0.7926  1.9147

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.49620    4.76510  -4.931 0.000812 ***
X1           -0.33742    0.05189  -6.503 0.000111 ***
X2           -0.21628    0.08618  -2.510 0.033325 *
X4            0.52137    0.06290   8.290 1.66e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.282 on 9 degrees of freedom
Multiple R-squared:  0.9644,    Adjusted R-squared:  0.9525
F-statistic: 81.22 on 3 and 9 DF,  p-value: 7.745e-07
```

截距：-23.49620

X1：-0.33742，t 值为-6.503，p 值为 0.000111，表示 X1 对 Y 有显著的负影响。

X2：-0.21628，t 值为-2.510，p 值为 0.033325，表示 X2 对 Y 有显著的负影响。

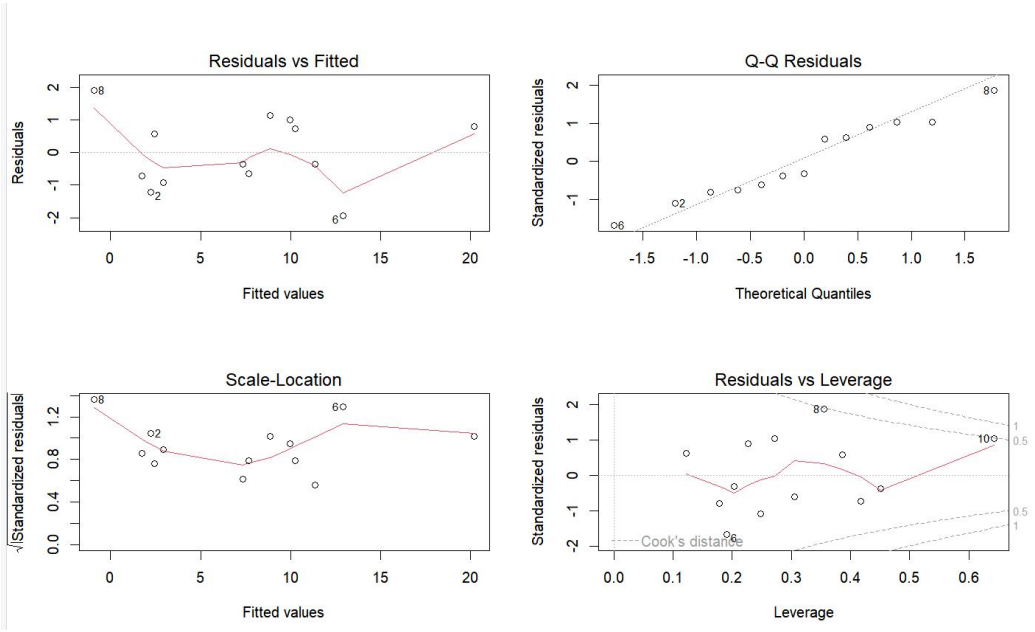
X4：0.52137，t 值为 8.290，p 值为 1.66e-05，表示 X4 对 Y 有显著的正影响。

结论：

逐步回归分析完成后，模型仅包括 X1、X2 和 X4，因为这些变量对因变量 Y 有显著的预测能力。模型的 R 平方值为 0.9644，调整后的 R 平方值为 0.9525，表明模型解释了大部分的变异性。**F 统计量的 p 值为 7.745e-07，表明模型整体非常显著**。

• 残差诊断：

```
> # 残差图
> plot(resid(model) ~ fitted(stepwise_model))
> abline(h = 0, col = "red")
> # 残差的诊断图
> par(mfrow = c(2, 2))
> plot(stepwise_model)
```



1.ResidualsvsFitted（残差 vs 拟合值）

这张图展示了残差（实际观测值与模型预测值之差）随着模型拟合值的变化情况。此图表明残差的分散程度随着拟合值的变化基本保持恒定，可以初步确定该模型具有方差齐性。

2.Q-QResiduals（残差 Q-Q 图）

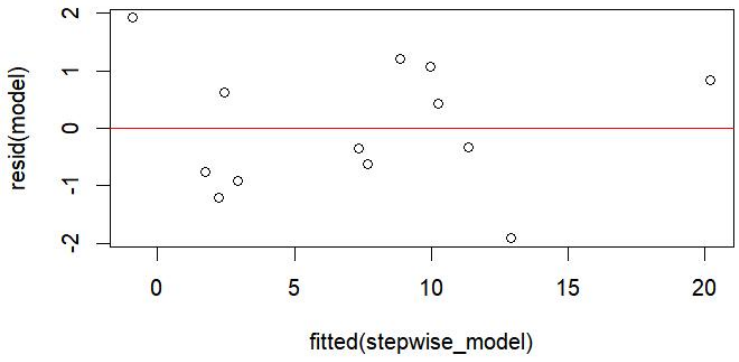
Q-Q 图（Quantile-QuantilePlot）用于评估残差的正态性。图中的点近似落在一条直线上，表明残差服从正态分布。

3.Scale-Location（尺度-位置图）

这张图用于评估残差的同方差性（方差齐性）。在图中我们看到残差的尺度（标准差）与拟合值的位置（水平）无关。这表明模型具有方差齐性。

4.ResidualsvsLeverage（残差 vs 杠杆值）

这张图展示了残差与杠杆值（Leverage）之间的关系。杠杆值衡量的是每个观测值对模型拟合的影响程度。理想情况下，我们希望看到残差随机分布，没有与杠杆值相关的模式。图中所有点的 cook 距离都小于 0.5，表明不存在异常值或强影响点。



·分析与结论：
残差图上的点的散步并没有呈现出一定的趋势（随横坐标的增大而增大或减小），故初步判断回归模型不存在异方差。

```
> # Breusch-Pagan检验
> bptest(stepwise_model)

        studentized Breusch-Pagan test

data:  stepwise_model
BP = 1.8463, df = 3, p-value = 0.6049
```

·检验结果分析及结论：

BP 值：检验统计量 BP 值为 1.8463，这个值是用于判断残差是否具有恒定方差。

自由度：自由度为 3，这通常意味着在模型中有 3 个预测变量（不包括常数项）。

p-value：p 值为 0.6049，这个值表示在假设方差齐性的情况下，观察到的统计量（或更极端）的概率。

由于 p 值大于常用的显著性水平（如 0.05），我们没有足够的证据拒绝方差齐性的零假设。这意味着在 5%的显著性水平下，没有显著的证据表明残差的方差随着自变量的变化而变化，因此可以认为模型满足方差齐性的假设。

```
> # 残差的正态性检验
> shapiro.test(resid(stepwise_model))

        Shapiro-Wilk normality test

data:  resid(stepwise_model)
W = 0.9634, p-value = 0.8051
```

·检验结果分析及结论：

Shapiro-Wilk 检验的零假设（H0）是数据来自正态分布，备择假设（H1）是数据不来自正态分布。如果 p 值大于显著性水平（通常设为 0.05 或 0.01），则不能拒绝零假设，意味着没有足够的证据表明数据不服从正态分布。在本例中，p 值为 0.8051，远大于 0.05，因此我们不能拒绝零假设，即没有足够的证据表明残差不服从正态分布。W 统计量 0.9634 是一个接近 1 的值，它表示数据的正态性是可接受的，因为 W 值越接近 1，数据越接近正态分布。

```
> # Durbin-Watson检验
> dwtest(stepwise_model)

        Durbin-Watson test

data:  stepwise_model
DW = 2.2225, p-value = 0.6183
alternative hypothesis: true autocorrelation is greater than 0
```

• 检验结果分析及结论：

DW 值：Durbin-Watson 统计量的值为 2.2225，表明残差之间没有显著的自相关性。

p-value：p 值为 0.6183，这个值表示在零假设（残差之间不存在自相关性）成立的情况下，观察到的统计量（或更极端）的概率。p 值大于常用的显著性水平（如 0.05 或 0.01），这意味着没有足够的证据拒绝零假设。

根据 DW 检验的结果，DW 值为 2.2225，位于 1.5 到 2.5 的可接受范围内，且 p 值为 0.6183，远大于 0.05。这表明在 5%的显著性水平下，没有足够的证据拒绝残差之间不存在自相关性的零假设。因此，可以认为这个回归模型的残差之间没有显著的自相关性，模型的自相关性假设得到了满足。

```
> # 学生化残差
> rstandard(stepwise_model)
         1          2          3          4          5          6          7
-0.3803549 -1.0910543  0.6180855  0.8996458 -0.6181312 -1.6765302  0.5821996
         8          9         10         11         12         13
 1.8602313 -0.8003744  1.0338647 -0.7394376 -0.3140392  1.0354191
```

• 检验结果分析及结论：

大多数残差都集中在-2 到+2 的范围内，这意味着模型拟合得较好。

• 对于模型含义的解释：

模型公式：Y~ X1+X2+X4

这个模型表示草药提取物的抑菌效果（Y）与四种成分中的三种成分含量（X1,X2,X4）有线性关系。

系数解释：

截距（Intercept）:-23.49620

截距项表示当所有自变量（X1,X2,X4）为 0 时，草药提取物的抑菌效果的预测值。这里，截距有统计学上的显著性（p 值为 0.000812），但需要注意，当所有自变量为 0 时的情况可能在实际应用中没有意义。

X1:-0.33742

这个系数表示成分 X1 含量每增加一个单位，草药提取物的抑菌效果预计会减少 0.33742 个单位。这个影响是显著的（p 值为 0.000111），表明 X1 的含量与抑菌效果有显著的负相关关系。

X2:-0.21628

X2 的系数表示成分 X2 含量每增加一个单位，抑菌效果预计会减少 0.21628 个单位。这个影响在统计上也是显著的（p 值为 0.033325），意味着 X2 的含量与抑菌效果有负相关关系。

X4:0.52137

X4 的系数表示成分 X4 含量每增加一个单位，抑菌效果预计会增加 0.52137 个单位。这个影响非常显著（p 值为 1.66e-05），表明 X4 的含量与抑菌效果有显著的正相关关系。

残差标准误差（Residualstandarderror）:1.282

这表示观测值与模型预测值之间的差异的标准差为 1.282，可以用来评估模型的预测精度。

R 平方（MultipleR-squared）:0.9644

R 平方值为 0.9644，表示模型能够解释 96.44%的因变量变异性，这是一个非常高的值，表明模型拟合度很好。

调整后的 R 平方（AdjustedR-squared）:0.9525

调整后的 R 平方考虑了模型中变量的数量，为 0.9525，仍然表明模型有很高的解释能力。

F 统计量（F-statistic）:81.22

F 统计量为 81.22，对应的 p 值为 7.745e-07，远小于 0.05，表明模型整体上是统计显著的。

综上所述，这个逐步回归模型表明草药提取物的抑菌效果与成分 X1、X2 和 X4 的含量有显著的线性关系，其中 X1 和 X2 的含量增加会降低抑菌效果，而 X4 的含量增加则会提高抑菌效果。模型整体拟合度很高，可以作为预测草药提取物抑菌效果的有效工具。

2. 某研究小组想要比较三种不同肥料在提高植物生长方面的效果是否存在显著差异。他们在相同的条件下，将一片土地分为三组，并分别施加了肥料 A、肥料 B 和肥料 C。随后，他们测量了每组植物的平均生长高度（单位：cm），得到了以下数据：

肥料 A 组：25,26,28,24,29,26,26,28,25,29
肥料 B 组：30,32,33,31,29,31,32,30,31,34
肥料 C 组：27,26,25,24,28,24,26,26,24,27

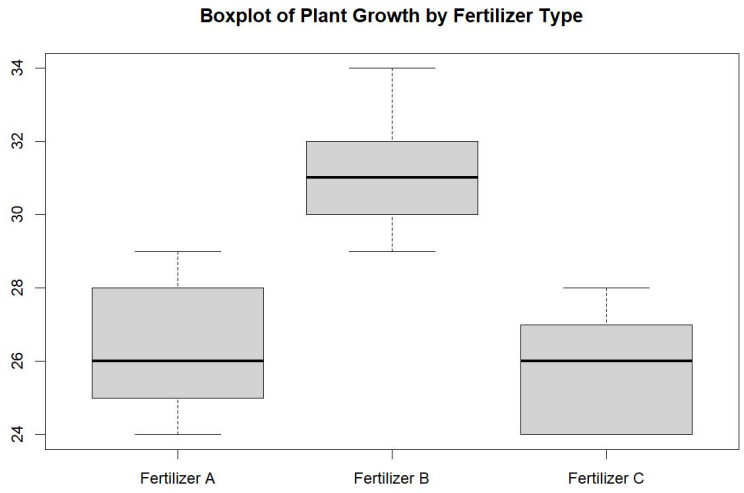
（1） 试用探索性数据分析方法判断不同肥料的效果是否存在显著差异。

```
> fertilizer_A <- c(25, 26, 28, 24, 29, 26, 26, 28, 25, 29)
> # 肥料B组数据
> fertilizer_B <- c(30, 32, 33, 31, 29, 31, 32, 30, 31, 34)
> # 肥料C组数据
> fertilizer_C <- c(27, 26, 25, 24, 28, 24, 26, 26, 24, 27)
>
> # 计算每组的均值、标准差等
> summary(fertilizer_A)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  24.00   25.25   26.00   26.60   28.00   29.00
> summary(fertilizer_B)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  29.00   30.25   31.00   31.30   32.00   34.00
> summary(fertilizer_C)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  24.00   24.25   26.00   25.70   26.75   28.00
```

• 显著性差异探究：

从描述性统计来看，肥料 B 组的平均生长高度明显高于肥料 A 和 C 组，可能与其他两组存在显著差异，而肥料 A 和 C 组的平均生长高度相近，存在显著差异的可能性不大。

```
> # 绘制箱型图
> boxplot(fertilizer_A, fertilizer_B, fertilizer_C,
+         names=c("Fertilizer A", "Fertilizer B", "Fertilizer C"),
+         main="Boxplot of Plant Growth by Fertilizer Type")
```

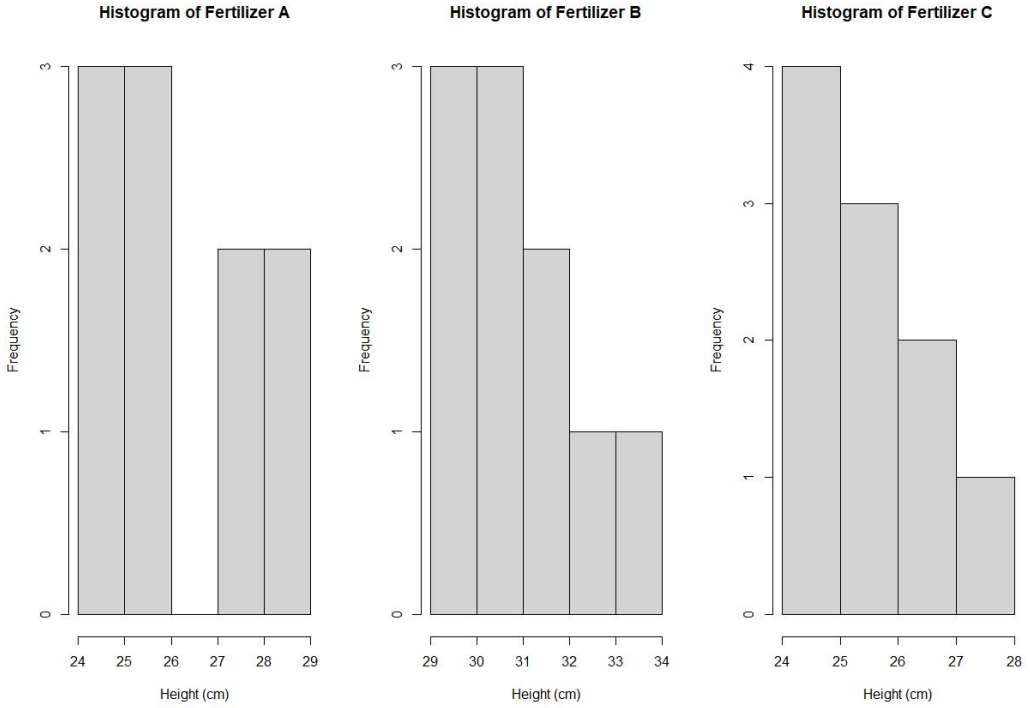


• 显著性差异探究：

肥料 B 组的中位数和平均值（如果离群值代表平均值）明显高于肥料 A 组和 C 组。

- 肥料 A 组和 C 组的中位数比较接近，但明显低于肥料 B 组，存在显著性差异的可能较小。

```
> # 绘制直方图
> par(mfrow=c(1,3)) # 设置图形排列
> hist(fertilizer_A, main="Histogram of Fertilizer A", xlab="Height (cm)")
> hist(fertilizer_B, main="Histogram of Fertilizer B", xlab="Height (cm)")
> hist(fertilizer_C, main="Histogram of Fertilizer C", xlab="Height (cm)")
> par(mfrow=c(1,1)) # 重置图形排列
```



• 分析：

肥料 A 的直方图显示了大部分植物生长高度集中在 24cm 到 29cm 之间，有一个较高的频率峰值在 26cm 左右。

肥料 B 的直方图显示了植物生长高度的分布较为分散，且整体上移，集中在 29cm 到 34cm 之间，有一个较高的频率峰值在 31cm 左右。

肥料 C 的直方图显示了植物生长高度集中在 24cm 到 28cm 之间，频率分布有递减的趋势。

• 显著性差异探究：

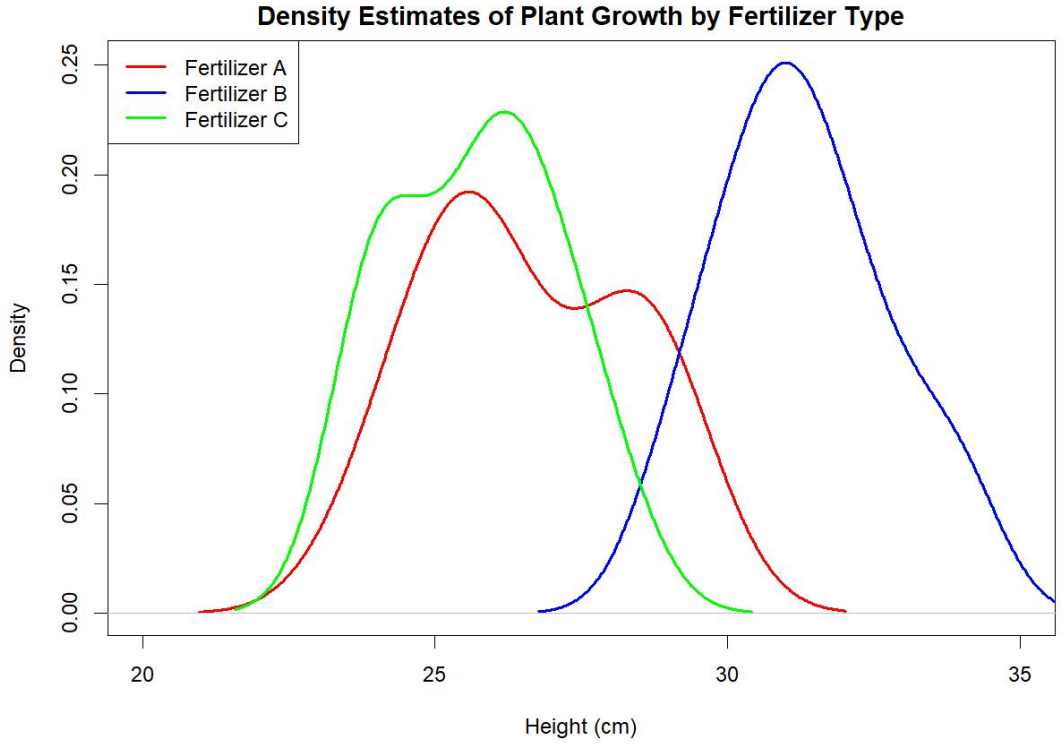
肥料 B 的植物生长高度普遍高于肥料 A 和肥料 C，这可能表明肥料 B 的效果与其他两种肥料有显著差异。

肥料 A 和肥料 C 的生长高度分布有一定的重叠，存在显著差异的可能不大。

```r
# 设置图形参数
par(mar = c(5, 5, 2, 1), # 设置边距：下，左，上，右
    mfrow = c(1, 1))    # 设置图形排列为1行1列

# 计算密度估计
density_A <- density(fertilizer_A)
density_B <- density(fertilizer_B)
density_C <- density(fertilizer_C)

# 绘制密度曲线
plot(density_A, main="Density Estimates of Plant Growth by Fertilizer Type",
     xlab="Height (cm)", ylab="Density", col="red", lwd=2, xlim=c(20, 35), ylim=
       c(0, max(density_A$y, density_B$y, density_C$y)))
lines(density_B, col="blue", lwd=2)
lines(density_C, col="green", lwd=2)
legend("topleft", legend=c("Fertilizer A", "Fertilizer B", "Fertilizer C"),
       col=c("red", "blue", "green"), lwd=2)
```



• 分析：

肥料 A 的密度曲线显示了植物生长高度主要集中在 25cm 到 29cm 之间，中心峰值大约在 26cm 左右。

肥料 B 的密度曲线显示了植物生长高度的分布较为分散，中心峰值较高，可能在 31cm 左右，且有一个较高的生长高度峰值在 34cm。

肥料 C 的密度曲线显示了植物生长高度的分布较为平均，中心峰值在 26cm 左右，但整体分布比肥料 A 和 B 要低。

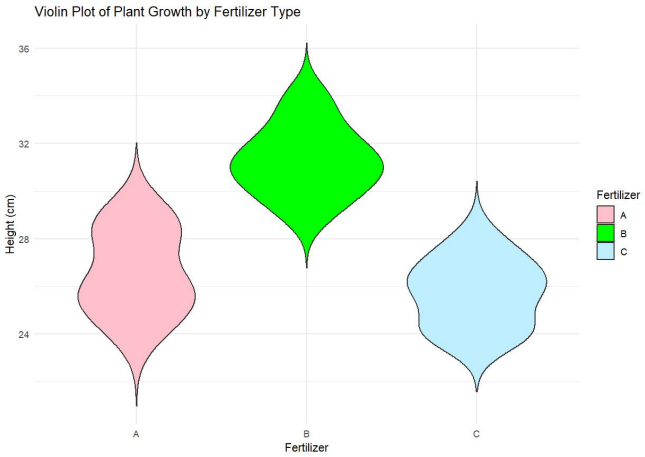
• 显著性差异探究：

肥料 B 的分布似乎有一个较高的峰值，并且整体上移，表明其可能与其他两组存在显著性差异。

肥料 A 和肥料 C 的分布较为接近，存在显著性差异的可能性不大。

```r
# 创建数据框
data <- data.frame(
  Height = c(fertilizer_A, fertilizer_B, fertilizer_C),
  Fertilizer = factor(rep(c("A", "B", "C"), each = length(fertilizer_A)))
)

# 创建小提琴图
ggplot(data, aes(x = Fertilizer, y = Height, fill = Fertilizer)) +
  geom_violin(trim = FALSE, scale = "area") +
  labs(x = "Fertilizer", y = "Height (cm)",
       title = "Violin Plot of Plant Growth by Fertilizer Type") +
  theme_minimal() +
  scale_fill_manual(values = c("A" = "pink", "B" = "green", "C" = "lightblue1"))
```



• 初步分析：

肥料 A 的小提琴图显示了植物生长高度的分布，其中心趋势（中位数）和密度分布的大致位置。

肥料 B 的小提琴图显示了一个更高的中心趋势，表明其可能促进了更高的植物生长。

肥料 C 的小提琴图显示了植物生长高度的分布，但看起来比肥料 A 和 B 的要低。

• 显著性差异探究：

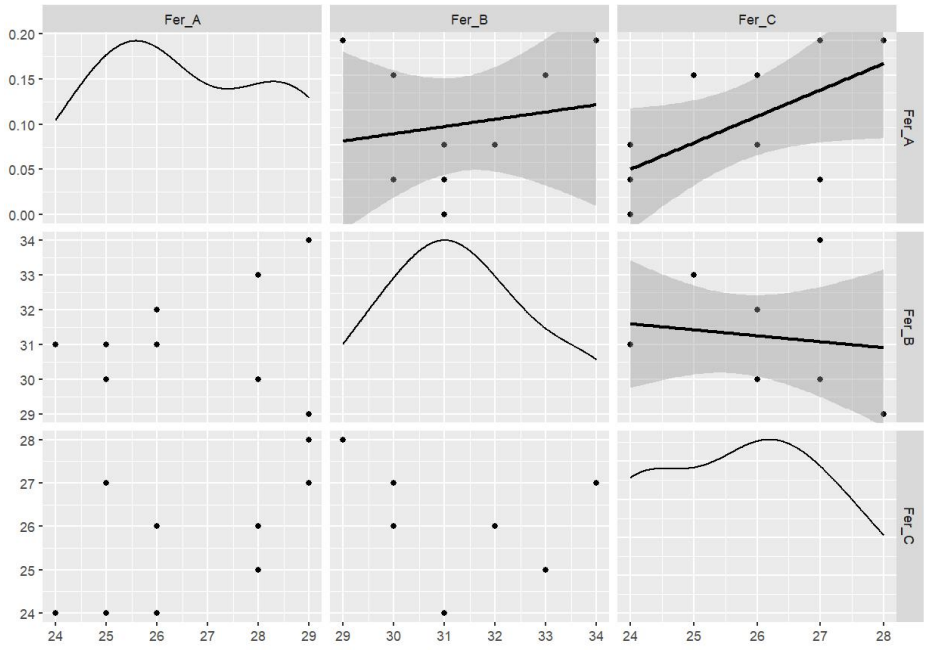
肥料 B 在促进植物生长方面效果最好，因为它的中心趋势和密度分布都位于较高的位置。

肥料 A 和肥料 C 的分布较为接近，但肥料 C 的中心趋势可能略低于肥料 A。

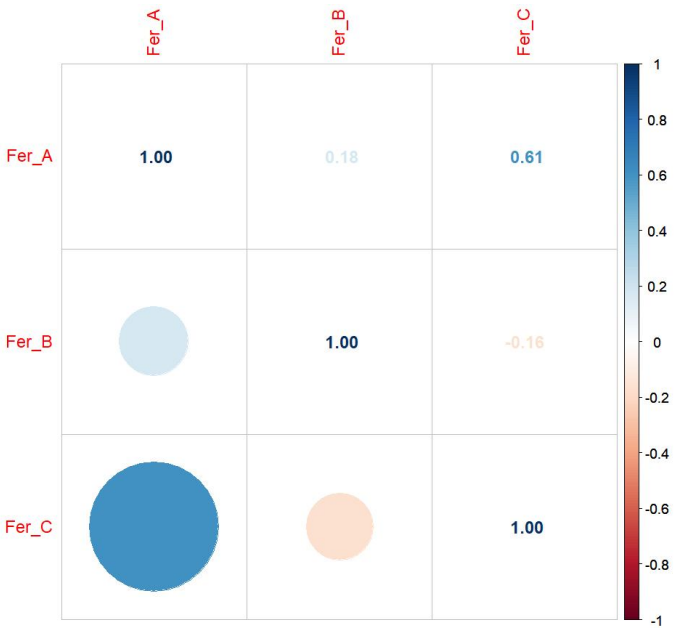
```r
# 创建数据框
data <- data.frame(
  Height = c(fertilizer_A, fertilizer_B, fertilizer_C),
  Fertilizer = factor(rep(c("A", "B", "C"), each = length(fertilizer_A)))
)

# 创建小提琴图加箱线图
ggplot(data, aes(x = Fertilizer, y = Height, fill = Fertilizer)) +
  geom_violin(trim = FALSE, color = "black", scale = "count") +  # 绘制小提琴图的轮廓
  geom_boxplot(width = 0.1, fill = NA, outlier.shape = NA) +  # 绘制箱线图，不显示异常点
  labs(x = "Fertilizer", y = "Height (cm)",
       title = "Violin and Boxplot of Plant Growth by Fertilizer Type") +
  theme_minimal() +
  scale_fill_manual(values = c("A" = "pink", "B" = "green", "C" = "lightblue1"))  # 设置颜色
```

```r
# 创建数据框
data <- data.frame(
  Fer_A = fertilizer_A,
  Fer_B = fertilizer_B,
  Fer_C = fertilizer_C
)
# 使用GGally包创建矩阵散点图
ggpairs(data,
        lower = list(continuous = "points", combo = "box"),
        upper = list(continuous = "smooth", combo = "box"))
```




```
> # 创建相关性热力图
> install.packages("corrplot")
Error in install.packages : Updating loaded packages
> library("corrplot")
> data.cor<-cor(data)
> corrplot(data.cor)
> corrplot(data.cor, method="number", type ="upper",tl.col= "n", tl.cex=0.8, tl.pos ="n",
+ tl.srt=45,add =T)
```



• 显著性差异探究：
相关性热力图可以看到，A 组和 C 组的相关系数达到了 0.61，而 B 组和 A 组，B 组和 C 组的相关系数的绝对值都小于 0.2,再结合矩阵散点图,可以初步判断出 B 组和 A 组，B 组和 C 组的相关性很弱，A 组和 C 组的相关性较强，因此可以推测 B 组会与其他两组有显著性差异。

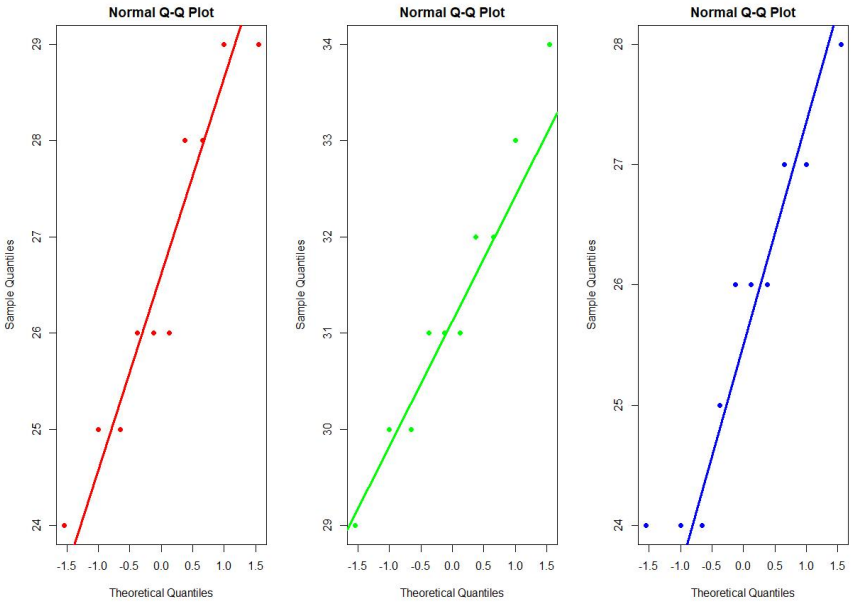
（2） 请用方差分析方法验证不同肥料的效果是否存在显著差异。（提示：做方差分析时需对其前提条件进行检验）。

步骤一（正态性检验）：

```
# 为肥料A创建QQ图
par(mfrow=c(1,3))  # 设置图形排列为一行三列
qqnorm(fertilizer_A, col="red", pch=19)  # 红色点表示肥料A
qqline(fertilizer_A, col="red", lwd=2)

# 为肥料B创建QQ图
qqnorm(fertilizer_B, col="green", pch=19)  # 绿色点表示肥料B
qqline(fertilizer_B, col="green", lwd=2)

# 为肥料C创建QQ图
qqnorm(fertilizer_C, col="blue", pch=19)  # 蓝色点表示肥料C
qqline(fertilizer_C, col="blue", lwd=2)
par(mfrow=c(1,1))  # 重置图形排列
```



分析：
三组数据的数据点紧密地围绕着直线排列，这表明三组数据都有着良好的正态性。

Shapiro-Wilk 检验:
1. 假设
零假设(H0):数据来自一个正态分布的总体。
备择假设(H1):数据不是来自正态分布的总体。
2. 构造统计量

使用 Shapiro-Wilk 测试来定量评估数据的正态性。Shapiro-Wilk 测试的统计量通常表示为 W，它是根据样本数据计算得出的，用于衡量样本分布与正态分布的接近程度。

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$$

其中，

x_i是排好序的样本值，

$\bar{x}$是样本均值，

a_i是与样本大小相关的常数。

n 是样本大小。

3.确定拒绝域

选择一个显著性水平 α （本题为 0.05）。

根据显著性水平和样本大小，查找 Shapiro-Wilk 测试的临界值。这个临界值是从统计表中得到的，或者是通过统计软件计算得出的。

如果统计量 W 的值小于或等于临界值，或者对应的 p 值小于或等于显著性水平 α，则落在拒绝域内。

4.决策

计算 Shapiro-Wilk 测试的统计量 W 和 p 值。将 p 值与显著性水平 α 进行比较：

如果 p≤α，则拒绝零假设（H0），接受备择假设（H1），认为数据不来自正态分布。

如果 p>α，则不能拒绝零假设（H0），认为数据来自正态分布。

5.结论分析

如果拒绝了零假设，说明样本数据显著偏离正态分布，可能需要考虑数据转换或采用非参数方法进行后续分析。

如果没有拒绝零假设，认为样本数据近似正态分布，可以继续进行假设检验或其他需要正态分布假设的统计分析。

```
> shapiro.test(fertilizer_A)

        Shapiro-Wilk normality test

data:  fertilizer_A
W = 0.9062, p-value = 0.2559

> shapiro.test(fertilizer_B)

        Shapiro-Wilk normality test

data:  fertilizer_B
W = 0.96624, p-value = 0.854

> shapiro.test(fertilizer_C)

        Shapiro-Wilk normality test

data:  fertilizer_C
W = 0.90514, p-value = 0.2493
```

• **分析：**

对于所有三种肥料组，p 值都远大于 0.05，这表明：

我们没有足够的证据拒绝肥料 A 组数据来自正态分布的零假设。

我们没有足够的证据拒绝肥料 B 组数据来自正态分布的零假设。

我们没有足够的证据拒绝肥料 C 组数据来自正态分布的零假设。

因此，基于 Shapiro-Wilk 测试的结果，我们可以认为三组数据都近似正态分布，可以继续进行 ANOVA 分析。

步骤二（方差齐性检验）：

1. 假设：

零假设：各总体方差相等。

备择假设：至少有一对总体方差不相等。

2. 构造统计量（Bartlett-Box 检验）：Bartlett-Box 检验统计量的公式为：

$$T = \frac{(N-k-1)\ln|S_p| - \sum_{j-1}^{k}(N_j-1)\ln|S_j|}{1+\frac{1}{3(k-1)}\left(\sum_{j-1}^{k}\frac{1}{N_j-1} - \frac{1}{N-k-1}\right)}$$

其中，N 是总样本数，k 是总体个数，N_j是第 j 个总体的样本量，S_P是总体协方差矩阵，S_j是第 j 组的样本协方差矩阵。

3. 确定拒绝域：

在显著水平 α 下，查找临界值，若统计量 T 落在拒绝域内，则拒绝零假设。

4. 决策：

若统计量 T 落在拒绝域内，则拒绝零假设。

若统计量 T 不在拒绝域内，则接受零假设。

5. 结论分析：

如果拒绝了零假设，则表示各总体方差不相等。

如果接受了零假设，则表示各总体方差相等。

```
> fertilizer_A <- c(25, 26, 28, 24, 29, 26, 26, 28, 25, 29)
> fertilizer_B <- c(30, 32, 33, 31, 29, 31, 32, 30, 31, 34)
> fertilizer_C <- c(27, 26, 25, 24, 28, 24, 26, 26, 24, 27)
>
> # 进行Bartlett检验
> bartlett_test_result <- bartlett.test(list(fertilizer_A, fertilizer_B, fertilizer_C))
>
> # 输出测试结果
> print(bartlett_test_result)

        Bartlett test of homogeneity of variances

data:  list(fertilizer_A, fertilizer_B, fertilizer_C)
Bartlett's K-squared = 0.49105, df = 2, p-value = 0.7823

>
> # 决策
> alpha <- 0.05
> if(bartlett_test_result$p.value <= alpha) {
+   cat("拒绝零假设 - 方差不齐\n")
+ } else {
+   cat("不能拒绝零假设 - 方差齐\n")
+ }
不能拒绝零假设 - 方差齐
```

• 分析

由于 p-value（0.7823）大于显著性水平（0.05），脚本输出“不能拒绝零假设－方差齐”，这意味着没有足够的证据表明三个肥料组的方差不相等。并且，由于没有拒绝零假设，可以继续进行 ANOVA 分析。

步骤三（方差分析）：

1. 假设

零假设（H0）：所有处理组（或水平）的总体均值相等，即没有显著差异。

备择假设（H1）：至少有一个处理组的总体均值与其他组不同。

2. 构造统计量：

组间均方（MSB）：$MSB = SSB/(k-1)$，其中 SSB 为组间平方和，k 为组数；

组内均方（MSW）：$MSW = SSW/(n-k)$，其中 SSW 为组内平方和，n 为总体样本个数，k 为组数；

F 统计量：$F = MSB / MSW$

3. 确定拒绝域：

设显著性水平为α，自由度分别为 k-1 和 n-k 的 F 分布上 α 分位点为 Fα。若 F>Fα，则拒绝原假设。

4. 决策：

如果统计量 F 在拒绝域内，拒绝原假设，认为各组之间存在显著差异；如果统计量 F 在拒绝域外，接受原假设，认为各组之间不存在显著差异。

5. 结论分析：

当拒绝原假设时，表示各组之间存在显著差异，可进行进一步的事后比较分析以确定具体哪些组之间有显著差异；

当接受原假设时，表示各组之间不存在显著差异，可以进行其他统计方法或分析。

```
> # 创建数据框
> data <- data.frame(
+   PlantHeight = c(25, 26, 28, 24, 29, 26, 26, 28, 25, 29,
+                   30, 32, 33, 31, 29, 31, 32, 30, 31, 34,
+                   27, 26, 25, 24, 28, 24, 26, 26, 24, 27),
+   Fertilizer = factor(c(rep("A", 10), rep("B", 10), rep("C", 10)))
+ )
>
> # 进行ANOVA
> aov_model <- aov(PlantHeight ~ Fertilizer, data = data)
> summary(aov_model)
            Df Sum Sq Mean Sq F value   Pr(>F)
Fertilizer   2  180.9   90.43   36.66 2.02e-08 ***
Residuals   27   66.6    2.47
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

• ANOVA 结果分析：

Df（自由度）：组间自由度为 2，因为有三个组。

sumSq（组间平方和）：组间总的平方和为 180.9。

MeanSq（组间均方）：组间均方为 90.43，这是组间平方和除以组间自由度得到的。

Fvalue（F 值）：组间均方与组内均方的比值为 36.66，这是一个衡量组间变异与组内变异的指标。

Pr(>F)（p 值）：对应的 p 值为 2.02e-08，远小于 0.05 的显著性水平。输出中的显著性代码"***"表示 p 值小于 0.001，这是非常显著的结果。

·结论：

由于 ANOVA 的 p 值远小于 0.05，我们可以拒绝零假设（三种肥料对植物生长高度的影响没有差异），这表明至少有两种肥料在提高植物生长高度方面存在显著差异。此外，由于 ANOVA 结果显示显著差异，下一步可以进行多重比较测试，来确定具体哪些组之间存在显著差异。

步骤四（多重比较）：

方法一（多重 t 检验）：

·假设：

零假设(H0)：对于每一对比较，两个组的均值之间没有差异。

备择假设(H1)：对于至少一对比较，两个组的均值之间存在差异。

·构造统计量：

使用 t 统计量来量化每对组之间均值差异的显著性。t 统计量的计算公式为：

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

其中，$\bar{x}_1$和$\bar{x}_2$是两组的样本均值，s^2是合并的方差估计，n_1和n_2是两组的样本大小。

·确定拒绝域：

拒绝域由 t 分布的临界值确定，该临界值基于给定的显著性水平（本题为 0.05）和自由度（通常是两个组的样本大小之和减 2）。在 R 中，pairwise.t.test 函数会自动计算这些临界值。

·决策：

如果计算出的 t 统计量的绝对值大于临界值，则拒绝零假设，认为比较的两个组之间存在显著差异。如果 t 统计量的绝对值小于或等于临界值，则不能拒绝零假设，即没有足够证据表明两组之间存在显著差异。

·结果分析：

分析 pairwise.t.test 函数的输出，查看每对比较的均值差异、标准误差、

t 统计量、P 值和置信区间。如果 P 值小于显著性水平（通常为 0.05），则认为两组间的差异是统计学上显著的。

```
> print(pairwise_t_test_result)

        Pairwise comparisons using t tests with pooled SD

data:  data$PlantHeight and data$Fertilizer

  A       B
B 1.0e-06 -
C 0.63    4.3e-08

P value adjustment method: bonferroni
```

• 分析：

AvsB：

肥料 A 与肥料 B 之间的比较显示，P 值为 0.000001，这是一个非常小的 P 值，远小于 0.05，表明肥料 A 和肥料 B 之间存在显著差异，肥料 B 的效果显著优于肥料 A。

AvsC：

肥料 A 与肥料 C 之间的比较显示，P 值为 0.63，这个 P 值远大于 0.05，表明肥料 A 和肥料 C 之间没有显著差异。

BvsC：

肥料 C 与肥料 B 之间的比较显示，P 值为 4.3e-08，这是一个非常小的 P 值，远小于 0.05，表明肥料 C 和肥料 B 之间存在显著差异，肥料 B 的效果显著优于肥料 C。

多重比较校正：

Bonferroni 校正。这种校正方法用于控制多重比较中的第一类错误（假阳性）的风险。然而，由于这里的 P 值非常小，即使不进行校正，这些显著性结果也很可能会保持不变。

• 结论：

应用 Bonferroni 校正后的成对 t 检验结果表明，肥料 B 在提高植物生长高度方面显著优于肥料 A 和肥料 C，而肥料 A 和肥料 C 之间没有显著差异。

方法二（同时置信区间 TUKEY 法）：

• 假设：

零假设(H0)：所有组间的均值相等，即没有显著差异。

备择假设(H1)：至少有两个组的均值不相等，即存在显著差异。

• 构造统计量：

Tukey 检验使用学生化秩次分布（studentsizedrangedistribution）来构造统计量，该统计量基于 ANOVA 模型的残差均方误差（MSE）和组间平均值差异。Tukey 检验的统计量通常是基于以下公式计算的：

$$Q = \frac{|\overline{Y}_{(i)} - \overline{Y}_{(j)}|}{SE}$$

其中：

$\overline{Y}_{(i)}$和$\overline{Y}_{(j)}$分别是第 i 组和第 j 组的样本均值。

SE（标准误差）是两组均值差异的标准误差，计算公式为：

$$SE = \sqrt{\frac{MSE}{n_i} + \frac{MSE}{nj}}$$

MSE（均方误差）是 ANOVA 模型中的残差均方误差。

n_i和n_j分别是第 i 组和第 j 组的样本大小。

• 确定拒绝域：

拒绝域由 Tukey 检验的临界值确定，该临界值基于给定的显著性水平（通常为 0.05）和自由度。在 R 中，TukeyHSD()函数根据组间比较的数量和样本大小自动确定拒绝域。

• 决策：

使用 TukeyHSD()函数得到的统计量结果，如果组间平均值差异的估计值大于由 Tukey 临界值确定的拒绝域，则拒绝零假设，认为这两个组之间存在显著差异。在 R 中，TukeyHSD()函数会输出一个列表，其中包含各组间比较的统计量、P 值等信息。

• 结果分析：

分析 TukeyHSD()函数的输出，查看每对比较的均值差异、标准误差、P 值和置信区间。

如果 P 值小于显著性水平（通常为 0.05），则认为组间差异是统计学上显著的。

```
> # 使用TukeyHSD进行多重比较
> tukey_result <- TukeyHSD(aov_model)
> print(tukey_result)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = PlantHeight ~ Fertilizer, data = data)

$Fertilizer
    diff       lwr        upr     p adj
B-A  4.7  2.958514  6.4414858 0.0000010
C-A -0.9 -2.641486  0.8414858 0.4175852
C-B -5.6 -7.341486 -3.8585142 0.0000000
```

• 分析：

B-A：肥料 B 与肥料 A 相比，平均生长高度的差异为 4.7 厘米。95%置信区间为[2.958514,6.4414858]厘米。调整后的 P 值为 0.0000010，远小于 0.05，表明肥料 B 和肥料 A 之间存在显著差异，肥料 B 的效果显著优于肥料 A。

C-A：肥料 C 与肥料 A 相比，平均生长高度的差异为-0.9 厘米。95%置信区间为[-2.641486,0.8414858]厘米。调整后的 P 值为 0.4175852，大于 0.05，表明肥料 C 和肥料 A 之间没有显著差异。

C-B：肥料 C 与肥料 B 相比，平均生长高度的差异为-5.6 厘米。95%置信区间为

[-7.341486,-3.8585142]厘米。调整后的 P 值为 0.0000000，远小于 0.05，表明肥料 C 和肥料 B 之间存在显著差异，肥料 B 的效果显著优于肥料 C。

• 结论：

根据 TukeyHSD 的多重比较结果，肥料 B 在提高植物生长高度方面显著优于肥料 A 和肥料 C，而肥料 A 和肥料 C 之间没有显著差异。这些结果可以帮助研究小组了解不同肥料对植物生长的具体影响，并为今后的肥料选择提供科学依据。

3. 在课上我们探讨了大数据时代统计学研究仍然需要抽样，请对以下情况进行分析，并给出一个例子进行说明：

（1） 如果通过抽样能够显著降低数据处理的复杂程度，且解决问题的效果没有太大的下降，抽样便是最优解；

• 分析：

a)成本效益：

处理全部数据通常需要巨大的计算资源和存储空间，而抽样可以显著降低这些成本。

b)时间效率：

全数据集的处理可能需要较长时间，而抽样可以在较短时间内得到近似的结果。

c)数据质量：

大数据集中可能包含许多噪声和异常值，抽样可以帮助筛选出更高质量的数据子集。

d)代表性：

合适的抽样方法可以确保样本的代表性，使得研究结果具有普遍性。

e)复杂性降低：

抽样可以减少数据的复杂性，简化模型，使得分析更加直观和易于理解。

• 例子：

假设我们想要研究一个大型在线零售商的客户满意度。该零售商拥有数百万的客户交易记录和反馈数据。如果我们尝试分析所有数据，首先需要处理和存储大量的信息，这将消耗大量的计算资源和时间。此外，数据中可能包含许多不相关的细节，如交易失败的记录，这些信息对于满意度分析并不重要。通过抽样，我们可以选择一个代表性的子集进行分析。例如，我们可以随机选择 10,000 个交易记录，这个子集应该能够代表整个客户群体。通过分析这个样本，我们可以估计整体客户满意度，同时减少处理时间、存储需求和计算成本。如果我们的抽样方法得当，比如使用分层抽样确保不同客户群体的代表性，那么从样本中得到的结果与全数据集的结果相比，差异可能不会太大。这样，我们就可以在较低的成本和时间内得到一个近似但足够准确的结果。

• 结论：

在大数据时代，抽样可以是一个有效的策略，特别是当抽样能够显著降低数据处理的复杂程度，同时保持解决问题的效果时。通过合理的抽样设计，我们可以在资源有限的情况下，高效地进行统计学研究。

（2） 若随着采样率的降低，解决问题的效果也快速下降，则应寻求大数据解决方案；

• 分析：

在大数据时代，尽管我们有能力处理和分析大规模数据集，但在某些情况下，抽样可能不足以提供足够的精确度或代表性，从而导致解决问题的效果快速下降。这通常发生在以下几种情况：

a)高变异性数据：

如果数据具有高度的变异性，即数据点之间的差异很大，那么较小的样本可能无法捕捉到这种变异性，导致分析结果不准确。

b)需要高精度：

在需要高精度预测或决策的场合，抽样可能无法提供足够的信息来确保结果的可靠性。

c)数据稀疏性：

在某些情况下，数据可能在某些特征或维度上非常稀疏，抽样可能无法捕捉到这些特征的全部信息。

d）动态变化：

如果数据集是动态变化的，随着时间的推移，抽样可能无法及时反映最新的数据趋势。

在这些情况下，寻求大数据解决方案，即分析整个数据集或尽可能多的数据，可能是更合适的选择。

• 例子：

假设我们正在研究一个社交媒体平台上用户的行为模式，以预测即将到来的选举结果。如果我们仅从平台上随机抽取一小部分用户的帖子和互动数据进行分析，可能会因为样本量太小而无法捕捉到所有重要的行为模式和趋势。特别是如果用户行为在不同的群体、地区或时间段内表现出显著的差异，抽样可能无法代表整个用户群体。随着采样率的降低，我们可能会发现预测的准确性迅速下降。这是因为样本可能无法充分代表整个用户群体，特别是如果存在高度的变异性或动态变化。在这种情况下，我们应该考虑使用大数据解决方案，即分析尽可能多的用户数据。通过分析整个数据集，我们可以更准确地捕捉到用户行为的细微差别，包括不同群体的特定行为模式，以及随时间变化的趋势。通过使用大数据，我们可以构建更复杂的模型，如深度学习网络，这些模型可以从大量的数据中学习并做出更准确的预测。此外，我们还可以实时更新模型，以反映最新的用户行为和趋势。

• 结论：

当抽样无法提供足够的信息或导致解决问题的效果快速下降时，转向大数据解决方案是合理的。这不仅可以提高分析的精度和可靠性，还可以帮助我们更好地理解和预测复杂的现象。然而，这也需要考虑到处理大数据所需的计算资源和隐私保护等问题。

（3） 如果某些问题的处理效果随着数据量上升有一定的提升，但当数据大到一定规模后再增加数据量带来的效果提升并不明显，这时，应选取一个有较大规模但并非总体的数据集来处理。

• 分析：

在统计学研究中，随着数据量的增加，我们通常能够获得更准确的估计和更可靠的结果。然而，这种提升并非无限制的。当数据量达到某个临界点后，进一步增加数据量所带来的效果提升可能会变得微不足道。这种现象通常被称为“边际递减收益”（diminishingreturns）。

在这种情况下，选择一个有较大规模但并非总体的数据集来处理，可以带来以下几个好处：

a)成本效益：

处理大规模数据集的成本（包括计算、存储和时间成本）通常高于处理较小数据集的成本。当数据量增加带来的边际效益降低时，继续增加数据量可能不再经济。

b)效率：

选择一个合适的数据规模可以提高处理效率，减少等待时间，使得研究结果能够更快地得到。

c)数据质量：

在某些情况下，过多的数据可能包含噪声和不相关的信息，这可能会降低分析的质量。选择一个有较大规模的数据集可以帮助我们集中关注更高质量的数据。

d)可管理性：

较小的数据集更易于管理和分析，可以使得研究过程更加清晰和可控。

• 例子：

假设我们正在研究一个在线教育平台上学生的学习行为，以优化课程设计和提高学习效果。起初，通过分析更多的学生数据，我们可以发现一些普遍的学习模式和行为趋势，从而对课程进行有效的调整。例如，我们可能发现学生在某个特定章节的学习时间更长，这可能意味着该章节需要更多的解释或不同的教学方法。然而，当我们分析了足够多的学生数据后，进一步增加数据量可能不再显著提高我们对学习行为的理解。例如，分析 10,000 名学生的数据可能已经足够揭示大部分的学习模式，而分析 100,000 名学生的数据可能只会带来微小的额外信息。在这种情况下，我们可以选择一个有较大规模的数据集，比如 10,000 名学生的数据，来进行分析。这样的数据集规模已经足够大，可以提供可靠的统计结果，同时避免了处理更大数据集所带来的成本和复杂性。通过选择一个合适的数据规模，我们不仅能够获得高质量的分析结果，还能够在合理的成本和时间内完成研究。

• 结论：

在大数据时代，选择合适的数据规模进行统计学研究是一个重要的决策。当数据量增加带来的效果提升变得不明显时，选择一个有较大规模但并非总体的数据集来处理，可以在保持研究质量的同时，提高成本效益和研究效率。这种方法可以帮助我们平衡数据量、研究成本和分析质量之间的关系，以达到最优的研究效果。