a



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**】

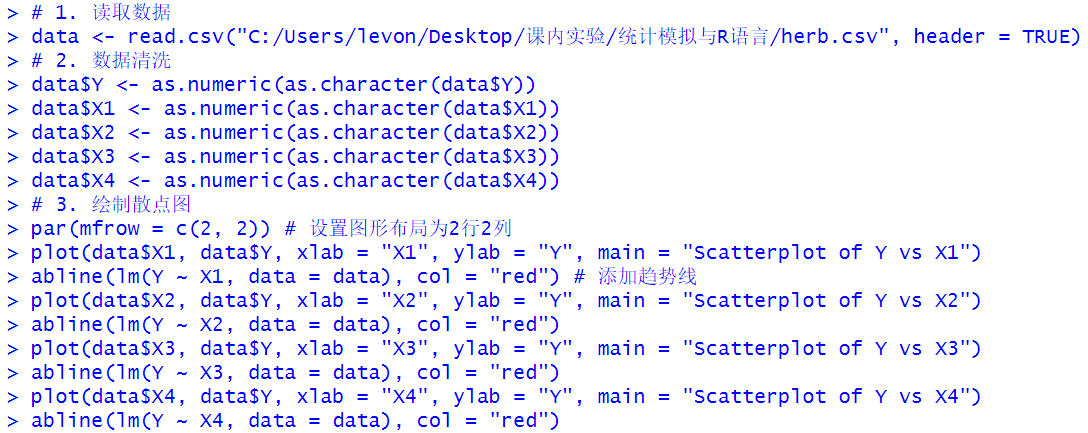
1. 你将有三天的时间完成这个大作业，5月25日上午8点下发，5月27日晚上11：59分之前提交。
2. 将你的答案以pdf文件先提交到http://www.xzc.cn/HFSanONFd1，文件名以“班级-姓名-学号”形式，纸质版上交请看群通知后面再交，纸质版需与pdf电子版保持一致，否则成绩无效。
3. 每题的答案列在题目下方，工整排版。所提交的答案要有必要的统计步骤、公式、假设、检验、R语言分析截图、结果分析与结论等。所有R语言输出的结果必须有相应的文字分析;不必要的R语言结果输出应该被避免。
4. 你可以利用你的笔记、任何书籍或线上资源，但内容必须由你自己完成。不能与其他任何人(无论是否与课程有关)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。

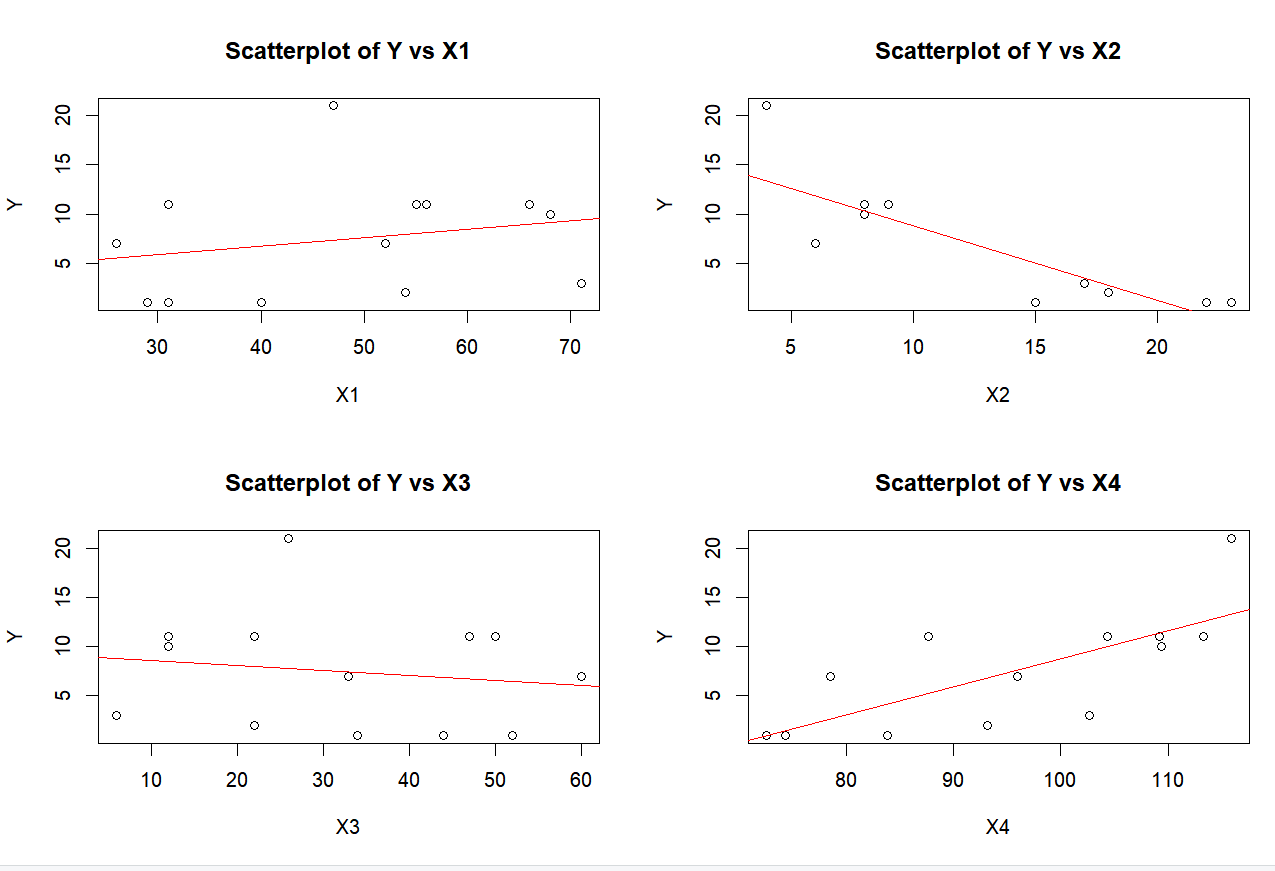
【**分数设置**】

|  |  |  |  |
| --- | --- | --- | --- |
| 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有关，现测得13组数据，见herb.csv文件，其中为草药提取物的抑菌效果，为提取物中的四种成分的含量。
2. **试用探索性数据分析方法，分别探究抑菌效果与每种成分含量之间是否存在线性关系。**



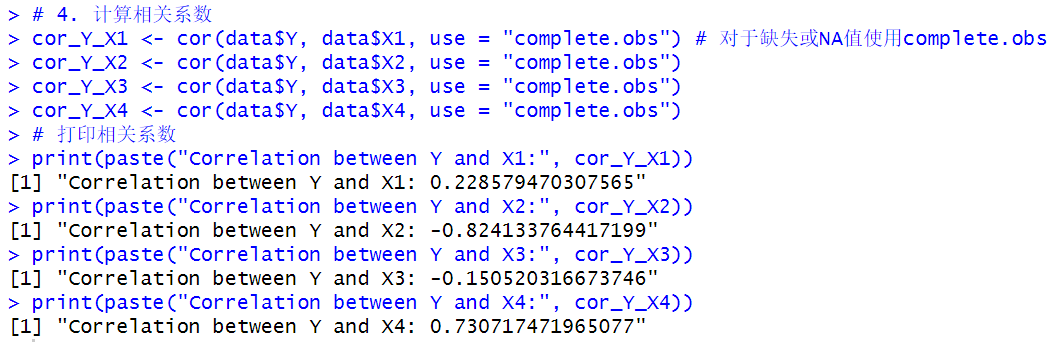


**·散点图分析：**

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。



**·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。

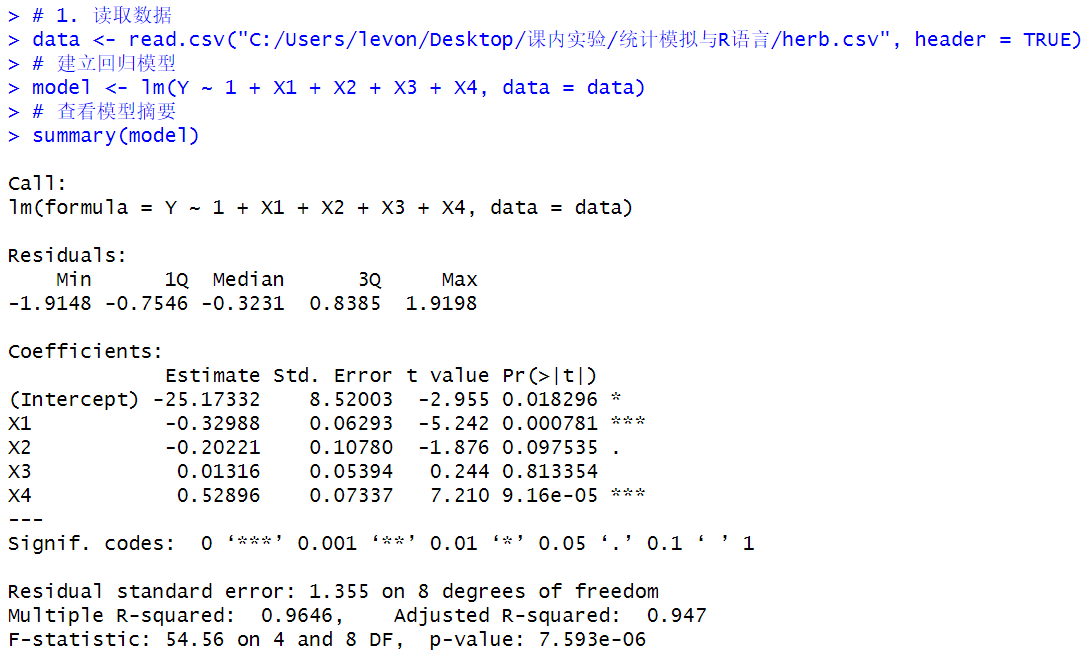
**·结论**

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。

1. **建立抑菌效果与四种成分含量的回归方程，并对回归方程进行显著性检验与回归诊断。**



与四种成分含量的回归方程为：

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):自变量的系数等于零，即该自变量对因变量Y没有影响.

备择假设(H1):自变量的系数不等于零，即该自变量对因变量Y有显著影响.

1. 统计量的计算

对于模型中每个自变量，计算其系数的t统计量:

=

其中,是估计的回归系数,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.

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

5.结论分析

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

如果不拒绝零假设，则得出结论:没有足够的统计学证据表明自变量对因变量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：

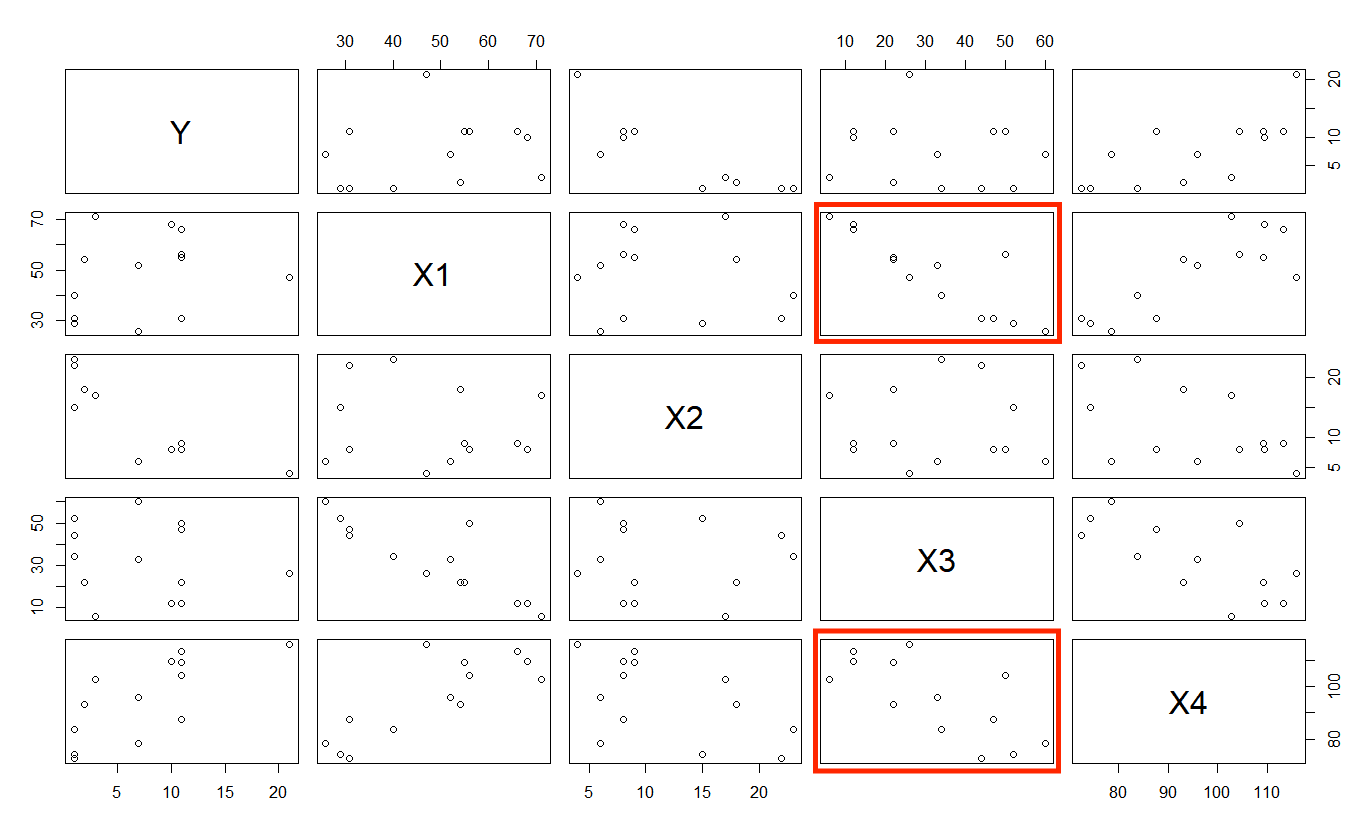
模型整体显著（p值远小于0.05）。

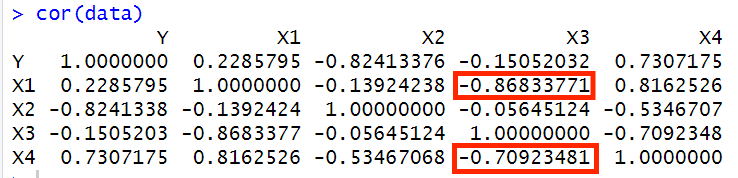
X1和X4对抑菌效果Y有显著的线性影响。

X2和X3在当前数据下对Y的影响不显著。

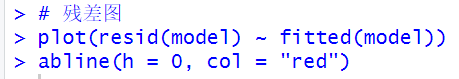
**·回归诊断：**

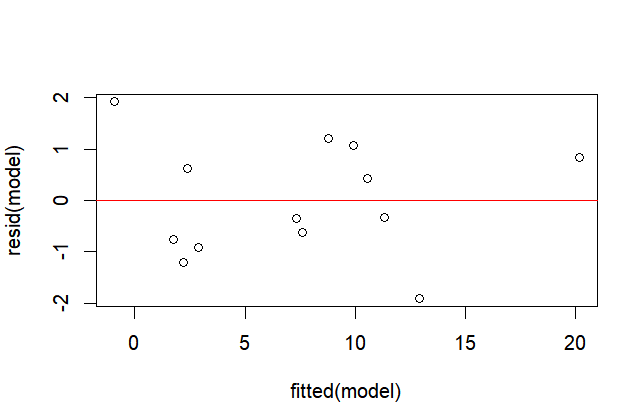
IMG_256





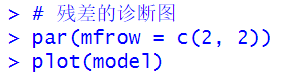
从散点图矩阵和相关系数矩阵中可以初步观察到该模型具有多重共线性。

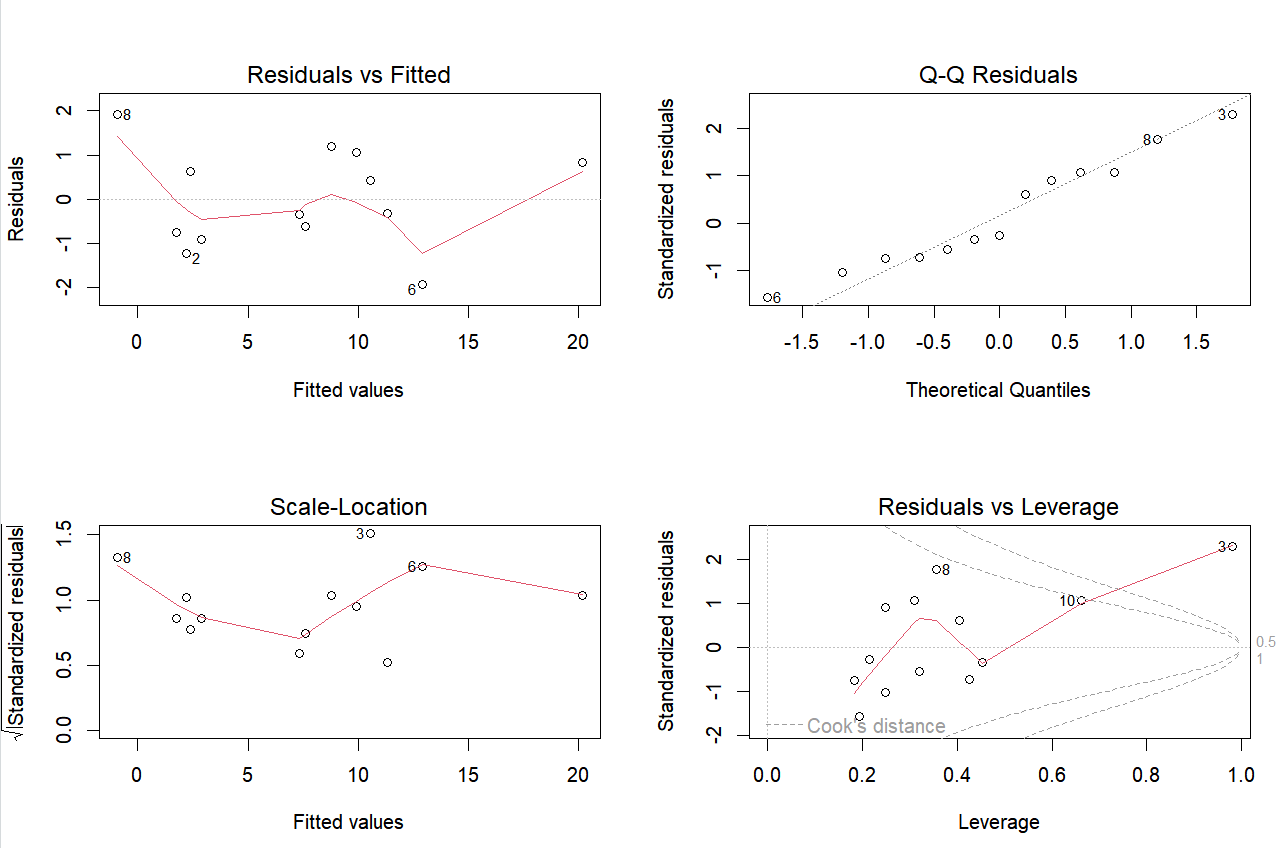




·分析与结论：

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。





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）

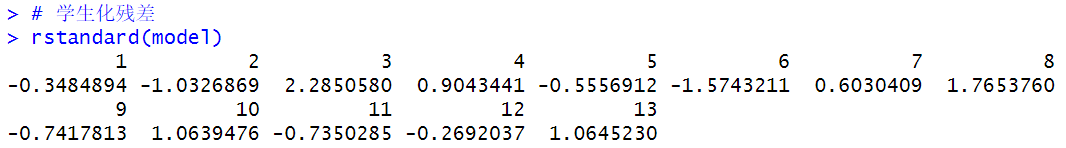
Q-Q图（Quantile-QuantilePlot）用于评估残差的正态性。图中的点近似落在一条直线上，表明残差服从正态分布。

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。

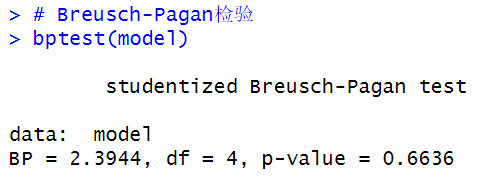


·计算结果分析：

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。



·检验结果分析：

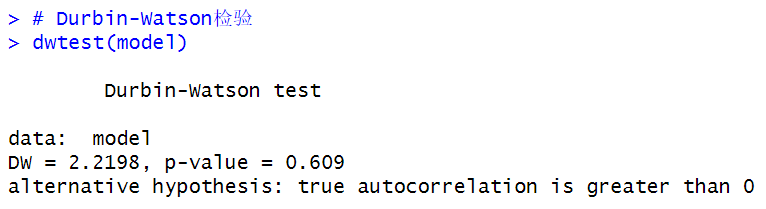
BP值：检验的统计量为2.3944。这是基于残差平方和与自变量之间关系的度量。

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。



·检验结果分析：

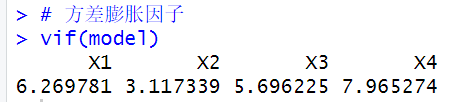
DW值：检验的Durbin-Watson统计量为2.2198。这个值表明残差之间可能存在轻微的正自相关，因为DW值大于2（但接近2）。

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。

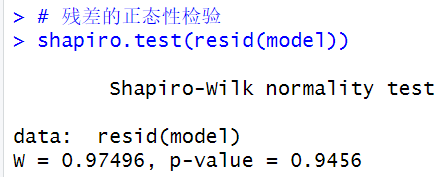
·结论：

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。



·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。



·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）。

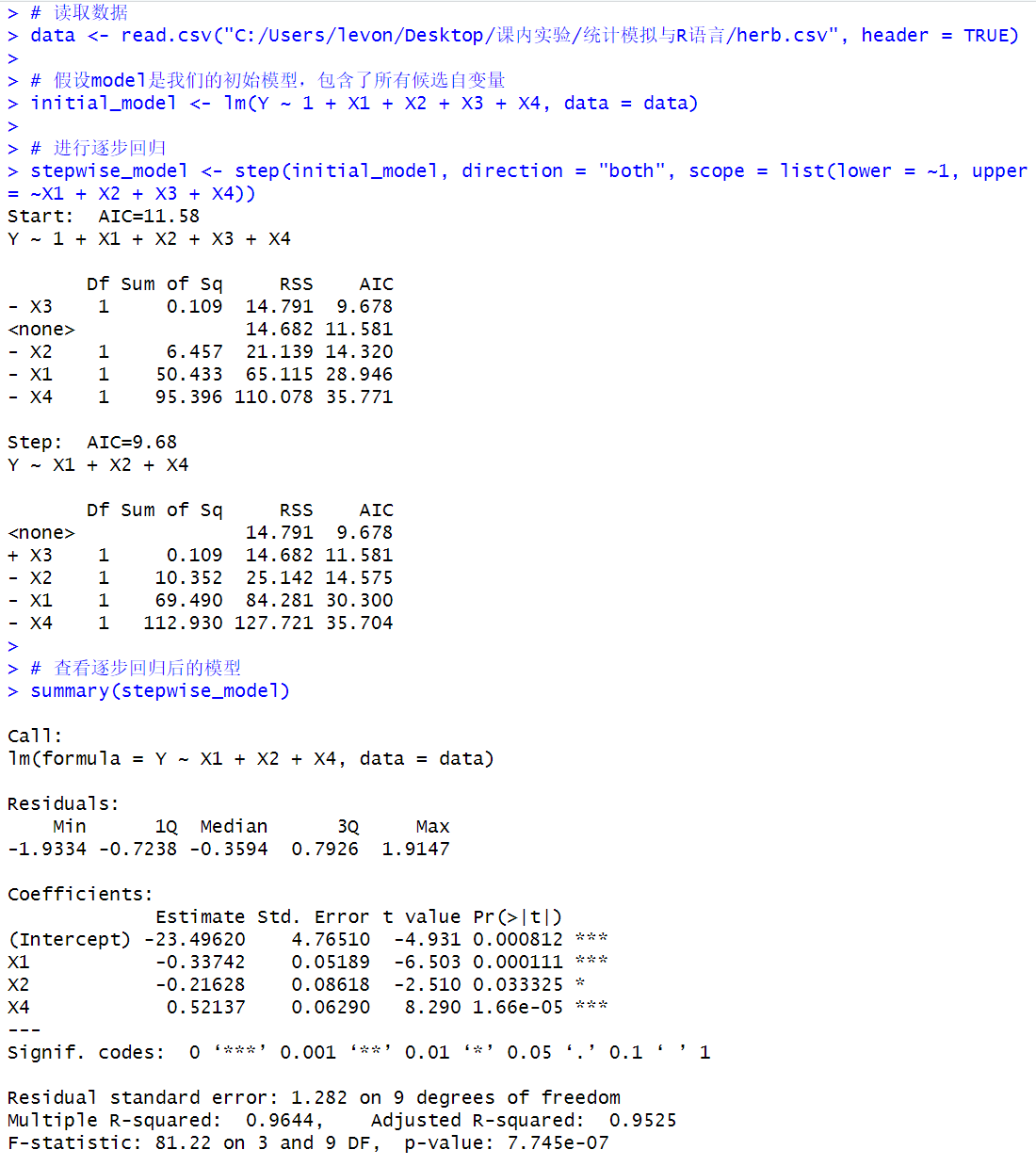
**·结论：**

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。

1. **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**。



·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。

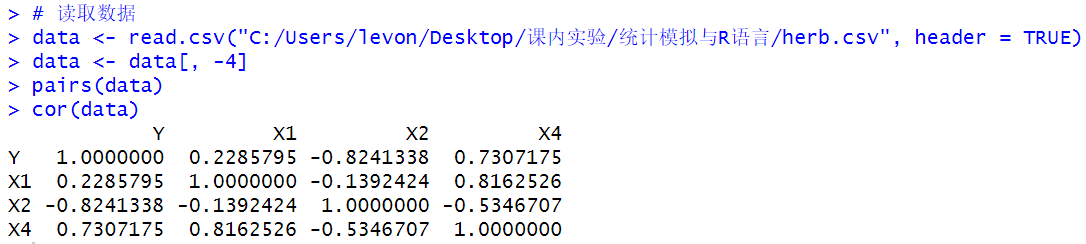
逐步删除：

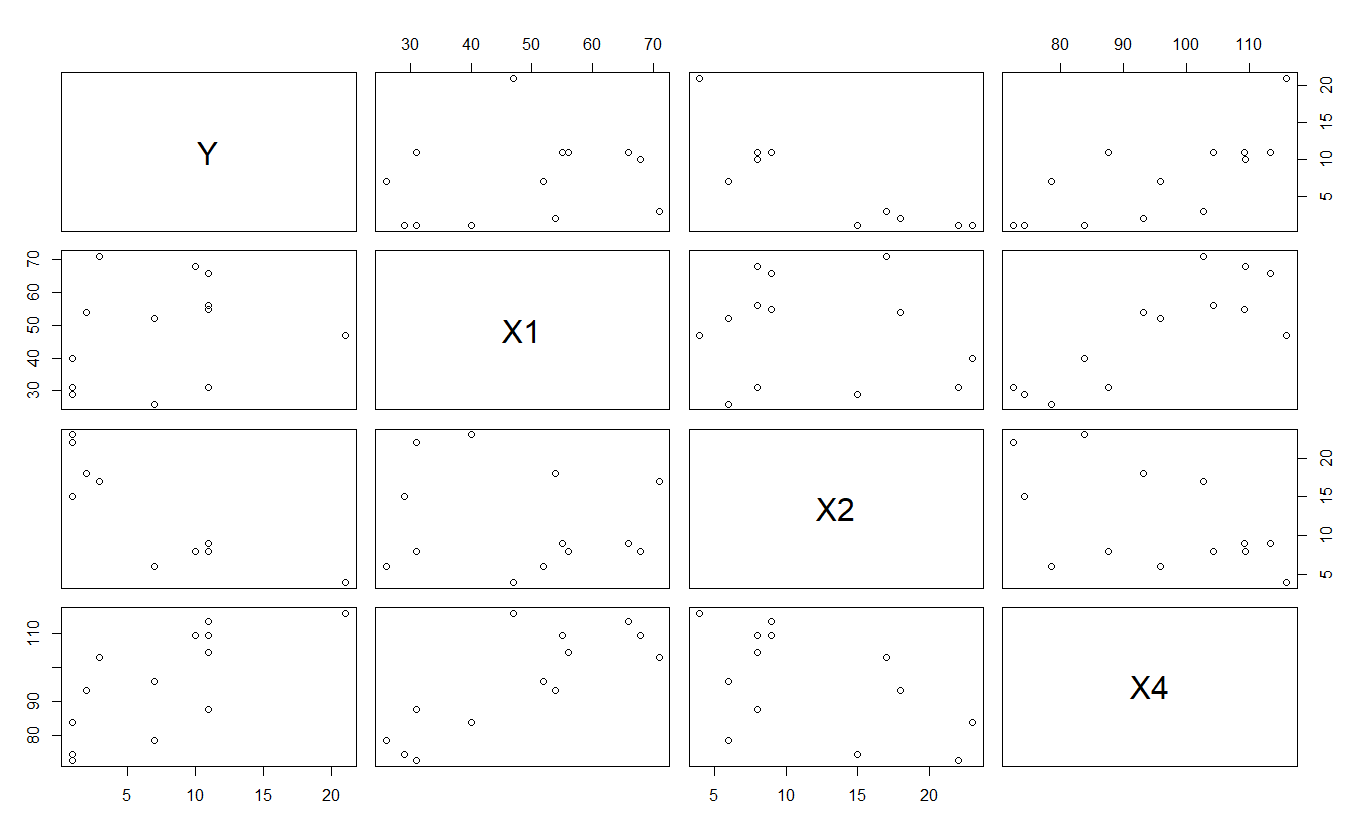
第一步，模型删除了X3，AIC从11.581降低到9.678，残差平方和（RSS）从14.682增加到14.791，自由度（Df）减少了1。

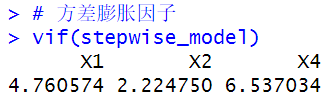
第二步，模型考虑重新添加X3或删除X2或X1或X4。根据AIC值，删除X2会使得AIC从9.678增加到14.320，因此X2被保留。删除X1会使得AIC从9.678增加到28.946，所以X1也被保留。删除X4会使得AIC从9.678增加到35.771，因此X4也被保留。模型没有进行任何改变，因为添加X3会使AIC从9.678增加到11.581。最终模型：经过逐步回归后，最终模型只包含

X1,X2,X4这三个自变量。

·多重共线性检验：







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):自变量的系数等于零，即该自变量对因变量Y没有影响.

备择假设(H1):自变量的系数不等于零，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，计算其系数的t统计量:

=

其中,是估计的回归系数,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.

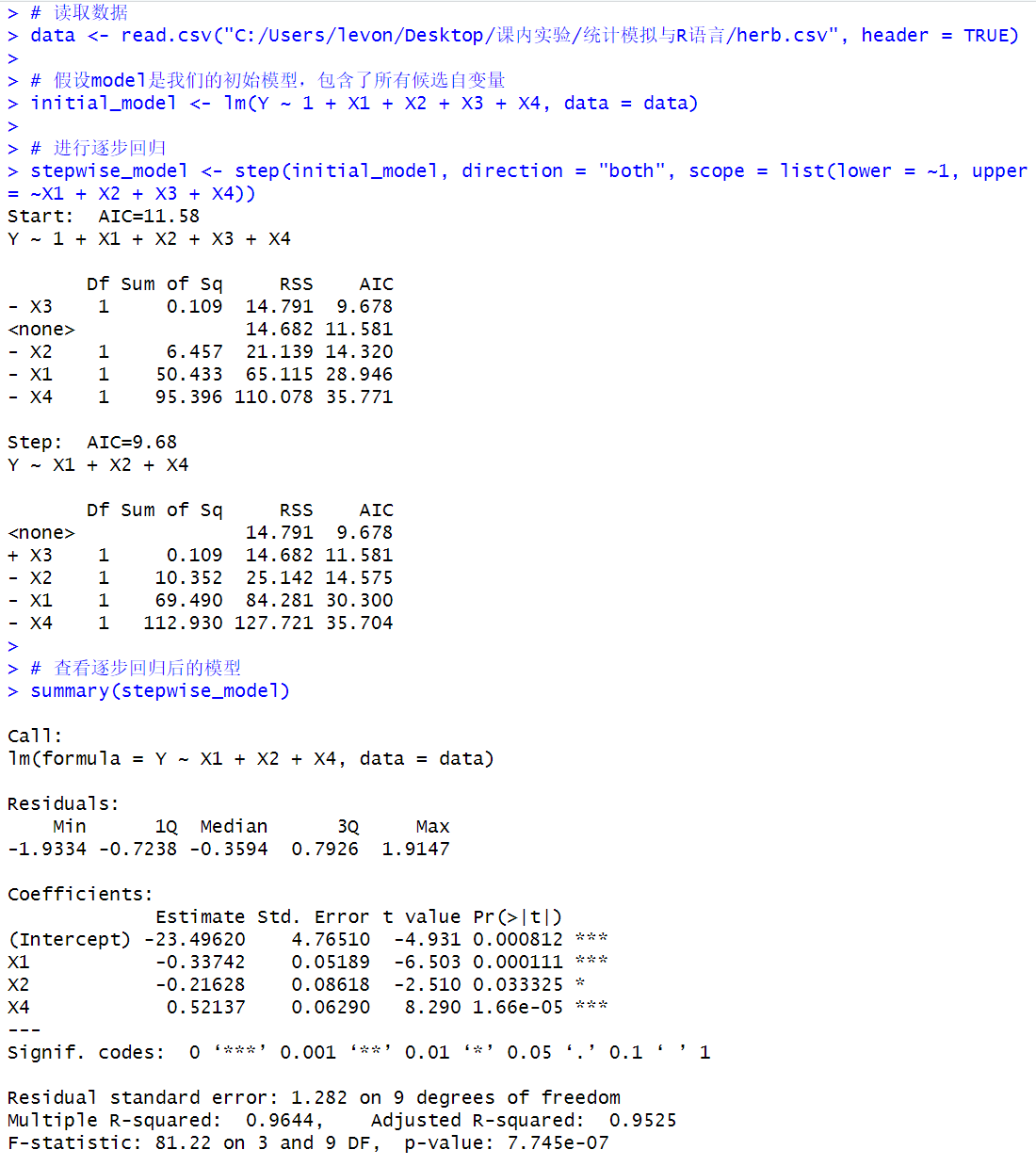
如果计算出的t统计量小于或等于临界值，或者计算出的p值大于显著性水平α，则不能拒绝零假设.

5.结论分析

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

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

·**显著性检验**：



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。

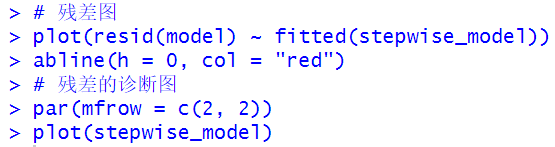
X2：-0.21628，t值为-2.510，p值为0.033325，表示X2对Y有显著的负影响。

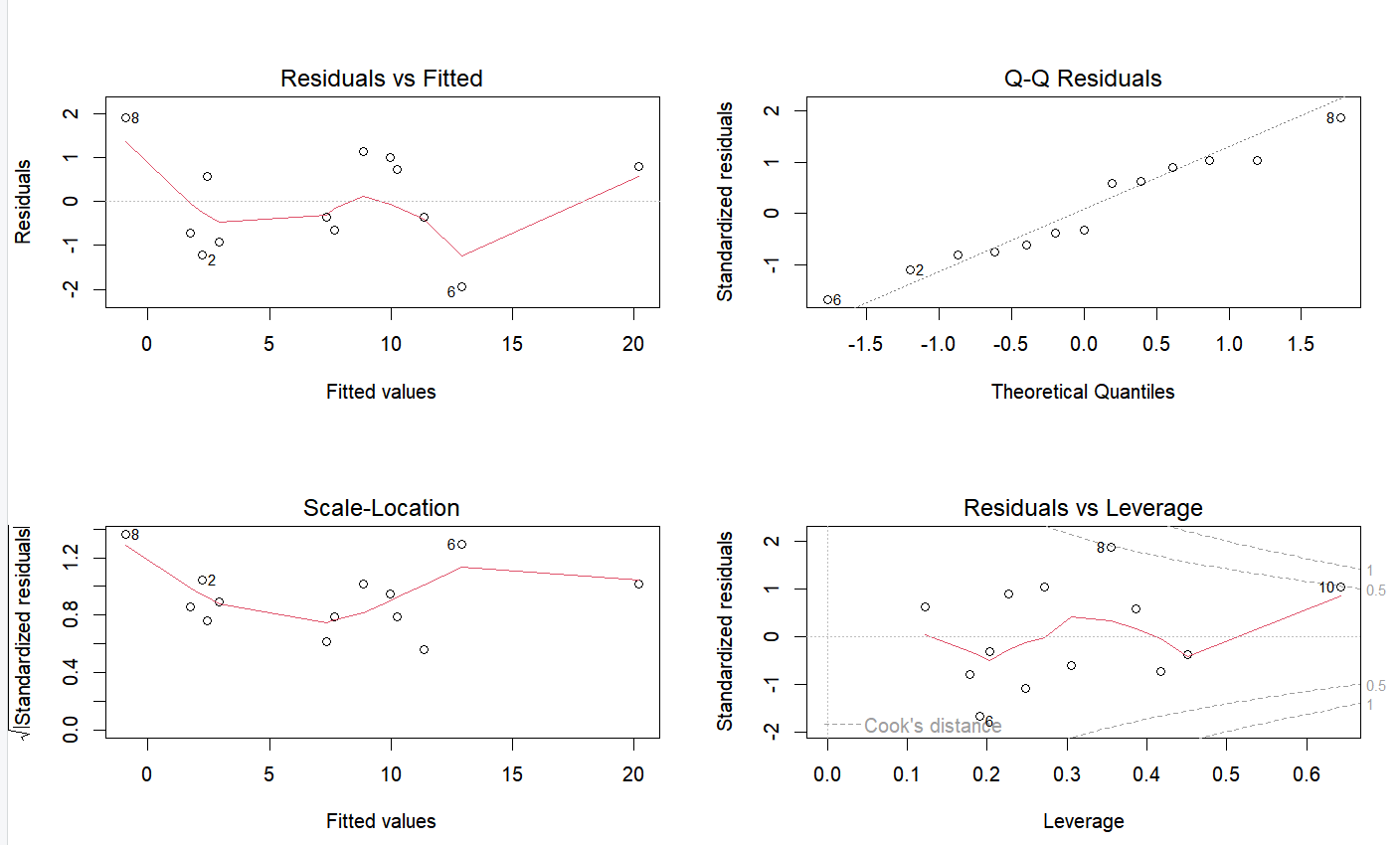
X4：0.52137，t值为8.290，p值为1.66e-05，表示X4对Y有显著的正影响。

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**：





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）

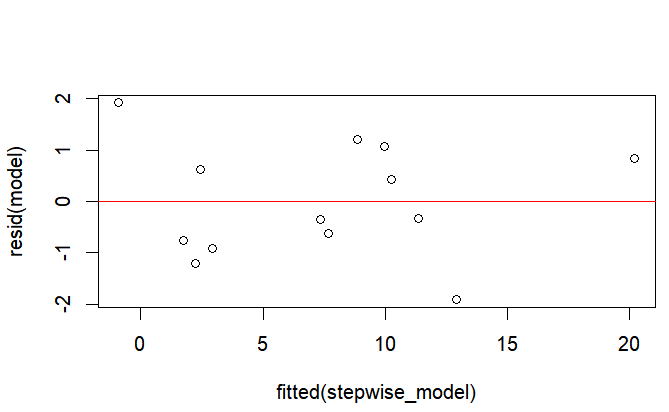
Q-Q图（Quantile-QuantilePlot）用于评估残差的正态性。图中的点近似落在一条直线上，表明残差服从正态分布。

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。

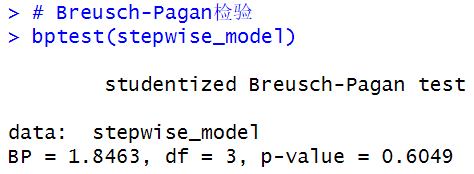
4.ResidualsvsLeverage（残差vs杠杆值）

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。



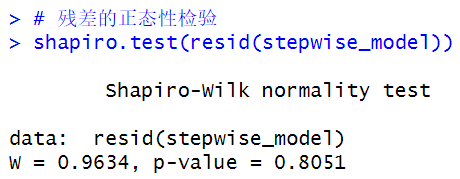
·检验结果分析及结论：

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）。

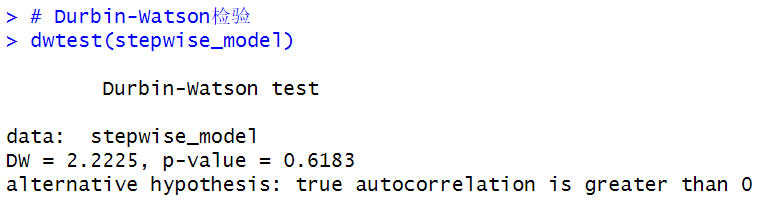
p-value：p值为0.6049，这个值表示在假设方差齐性的情况下，观察到的统计量（或更极端）的概率。

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。



·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。

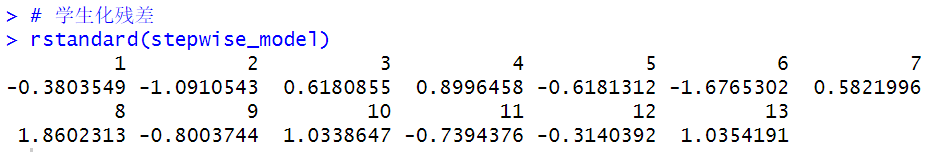


·检验结果分析及结论：

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。



·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。

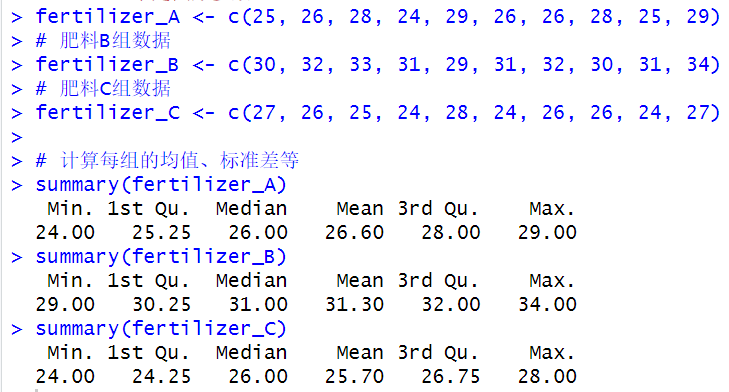
1. 某研究小组想要比较三种不同肥料在提高植物生长方面的效果是否存在显著差异。他们在相同的条件下，将一片土地分为三组，并分别施加了肥料A、肥料B和肥料C。随后，他们测量了每组植物的平均生长高度（单位：cm），得到了以下数据：

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

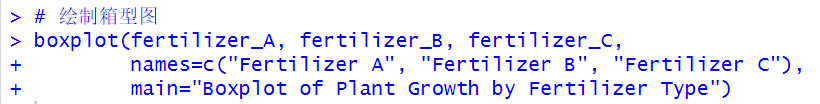
肥料C组：27,26,25,24,28,24,26,26,24,27

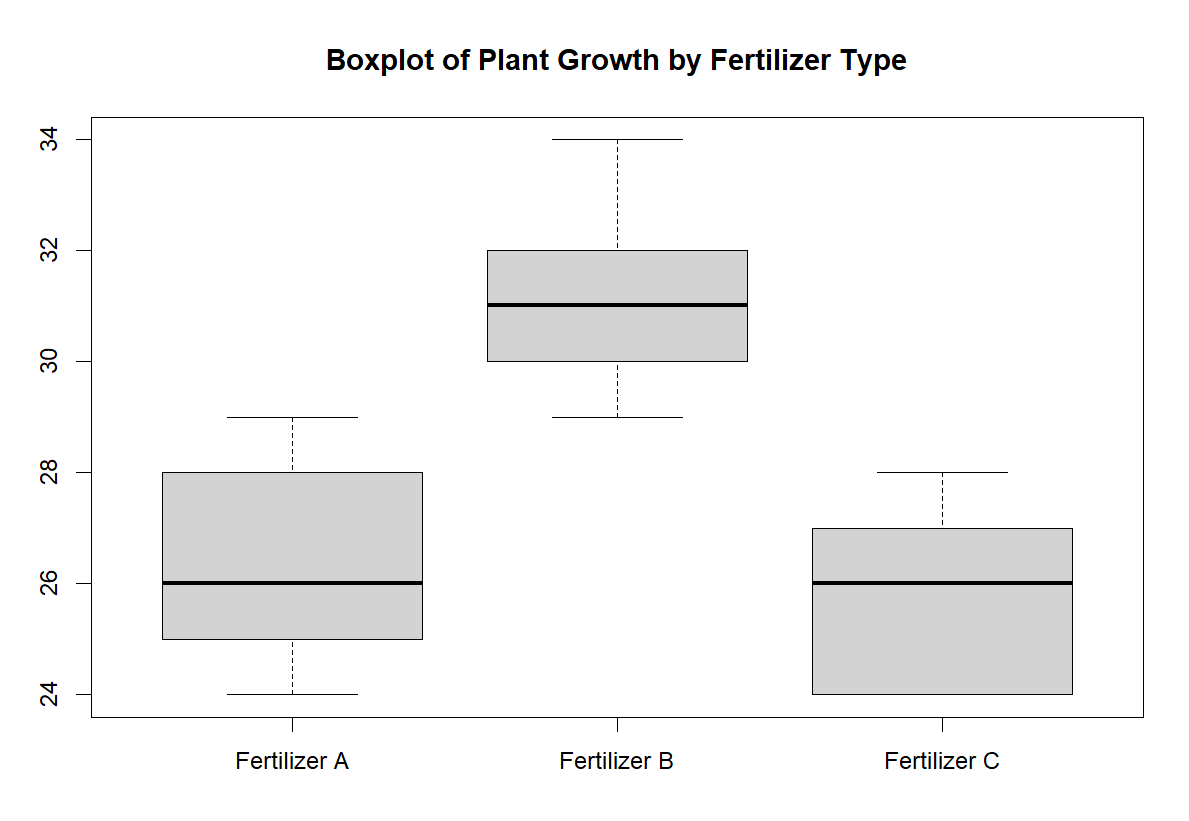
1. **试用探索性数据分析方法判断不同肥料的效果是否存在显著差异。**



·显著性差异探究：

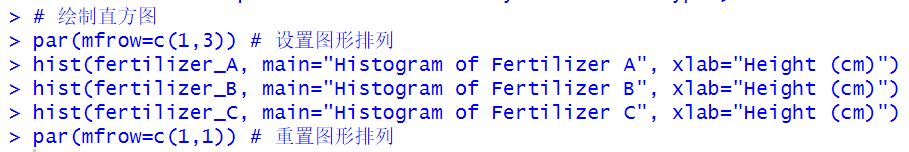
从描述性统计来看，肥料B组的平均生长高度明显高于肥料A和C组，可能与其他两组存在显著差异，而肥料A和C组的平均生长高度相近，存在显著差异的可能性不大。

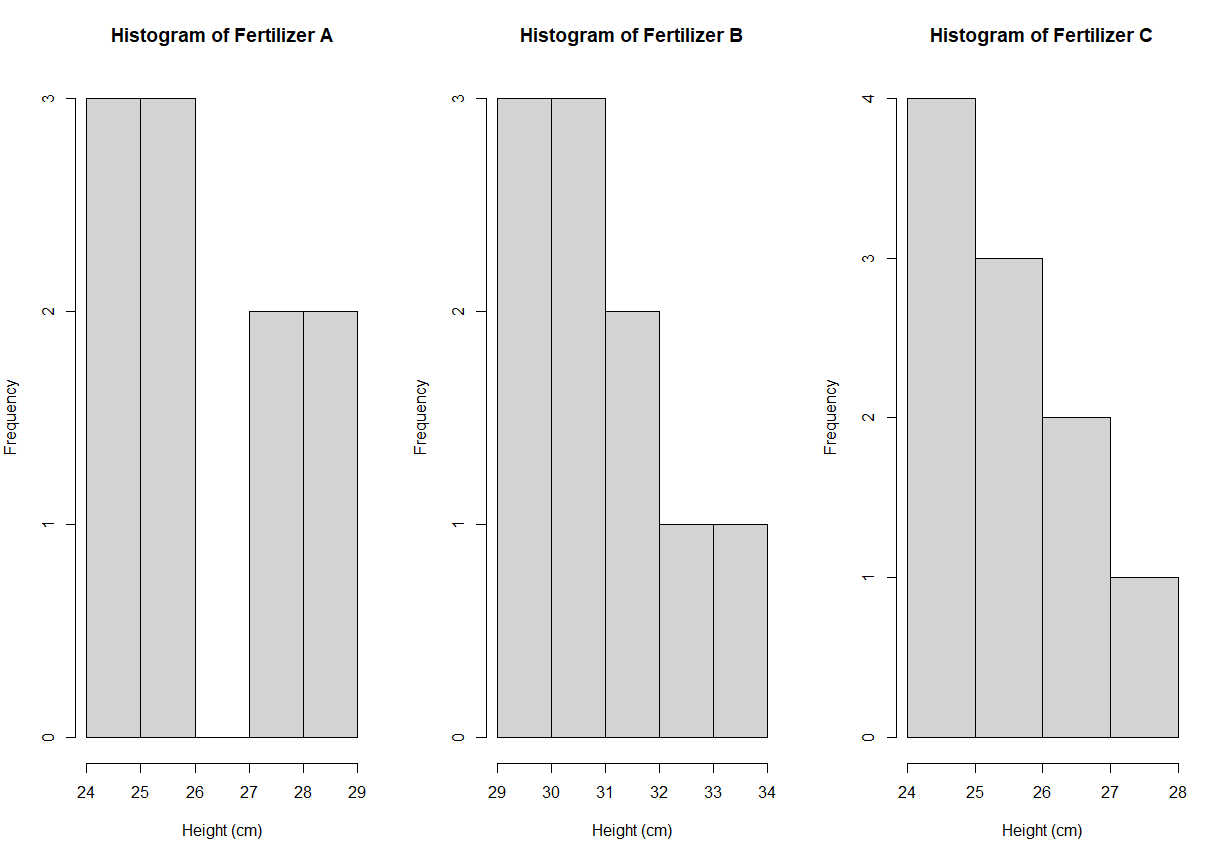




·显著性差异探究：

* 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。





·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。

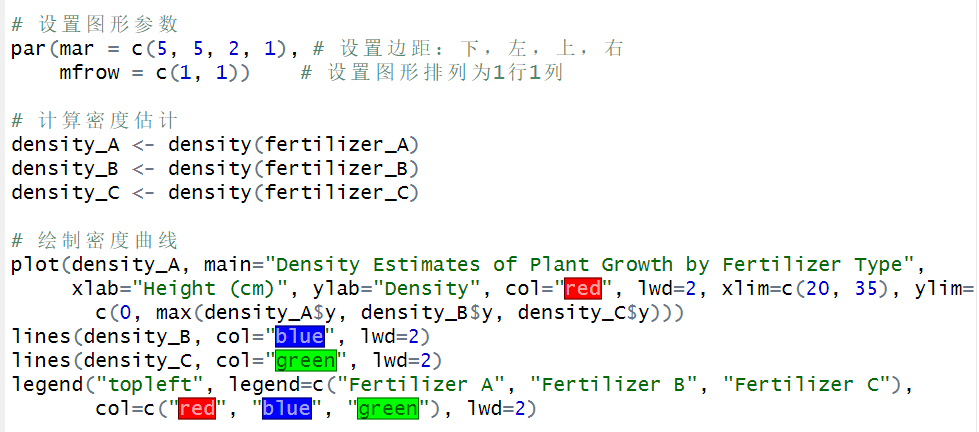
肥料B的直方图显示了植物生长高度的分布较为分散，且整体上移，集中在29cm到34cm之间，有一个较高的频率峰值在31cm左右。

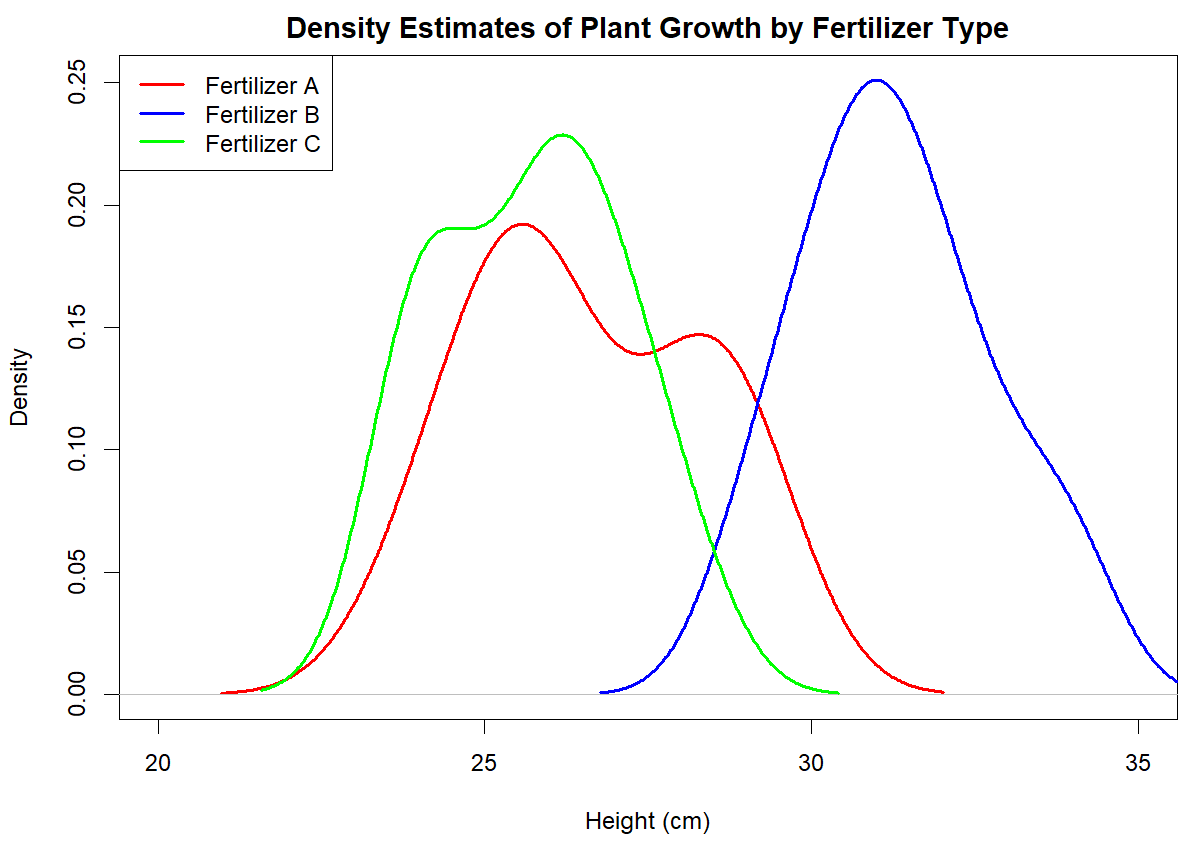
肥料C的直方图显示了植物生长高度集中在24cm到28cm之间，频率分布有递减的趋势。

·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。





·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。

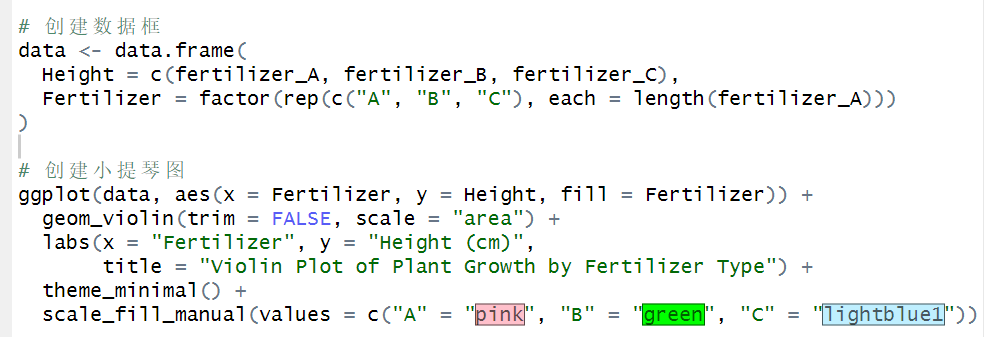
肥料B的密度曲线显示了植物生长高度的分布较为分散，中心峰值较高，可能在31cm左右，且有一个较高的生长高度峰值在34cm。

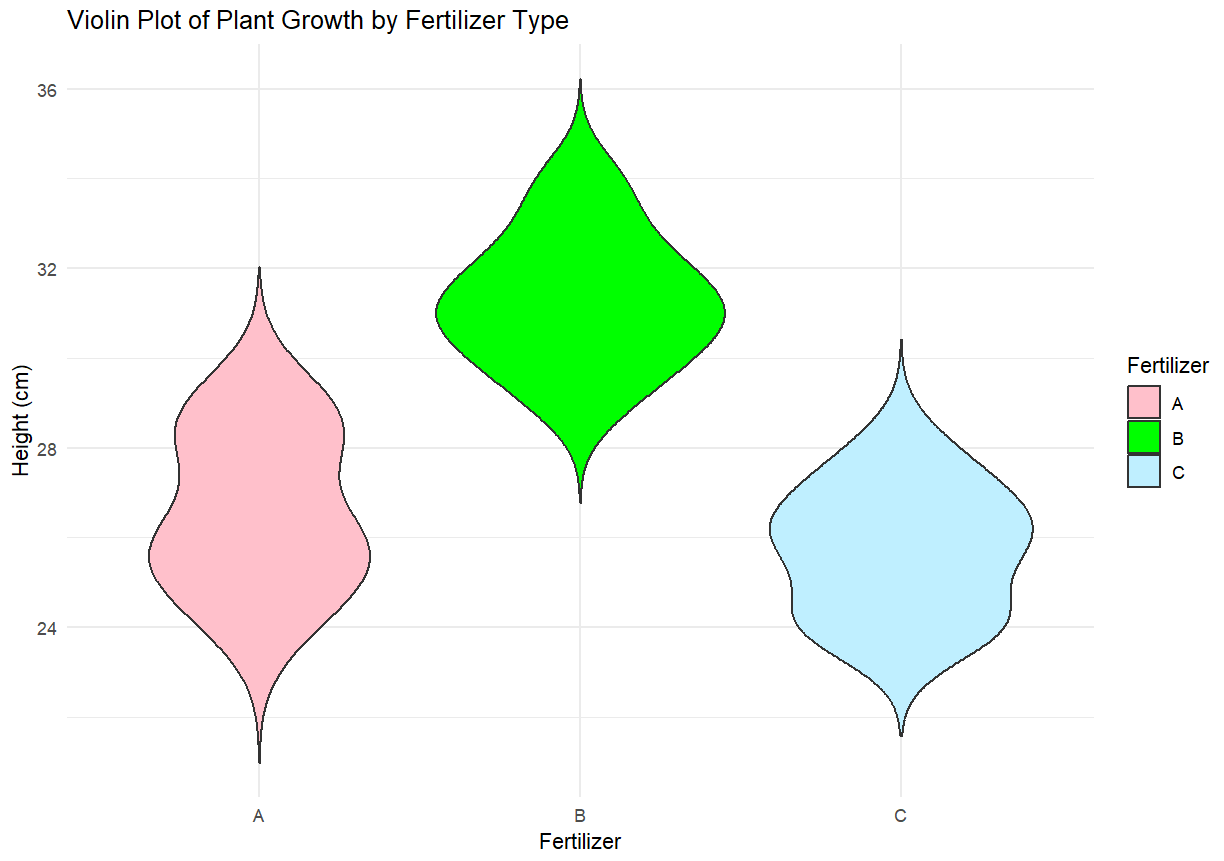
肥料C的密度曲线显示了植物生长高度的分布较为平均，中心峰值在26cm左右，但整体分布比肥料A和B要低。

·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。





·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。

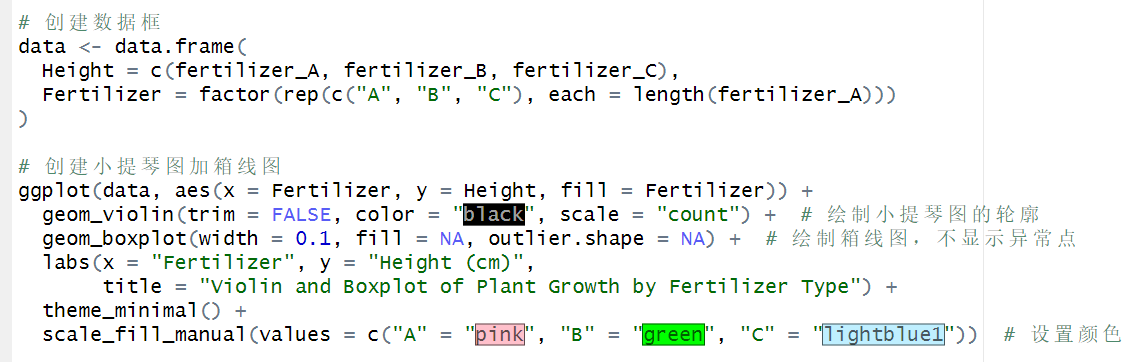
肥料B的小提琴图显示了一个更高的中心趋势，表明其可能促进了更高的植物生长。

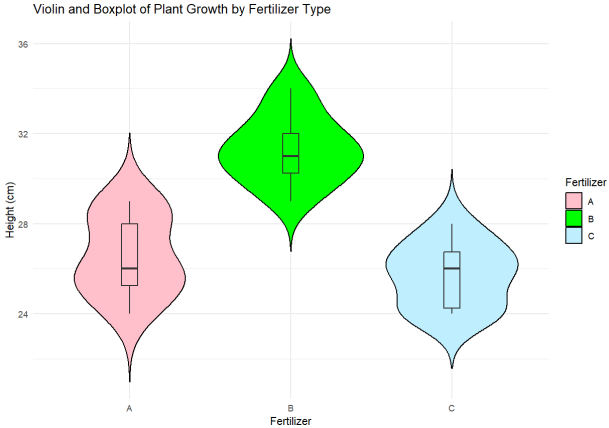
肥料C的小提琴图显示了植物生长高度的分布，但看起来比肥料A和B的要低。

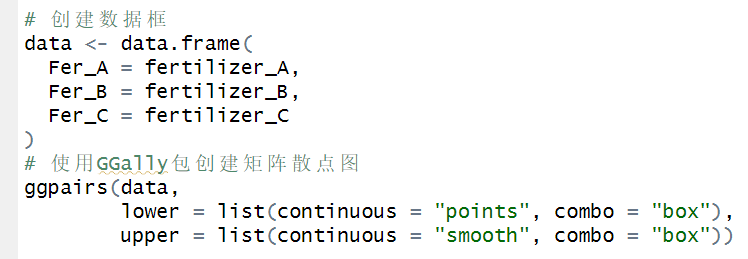
·显著性差异探究：

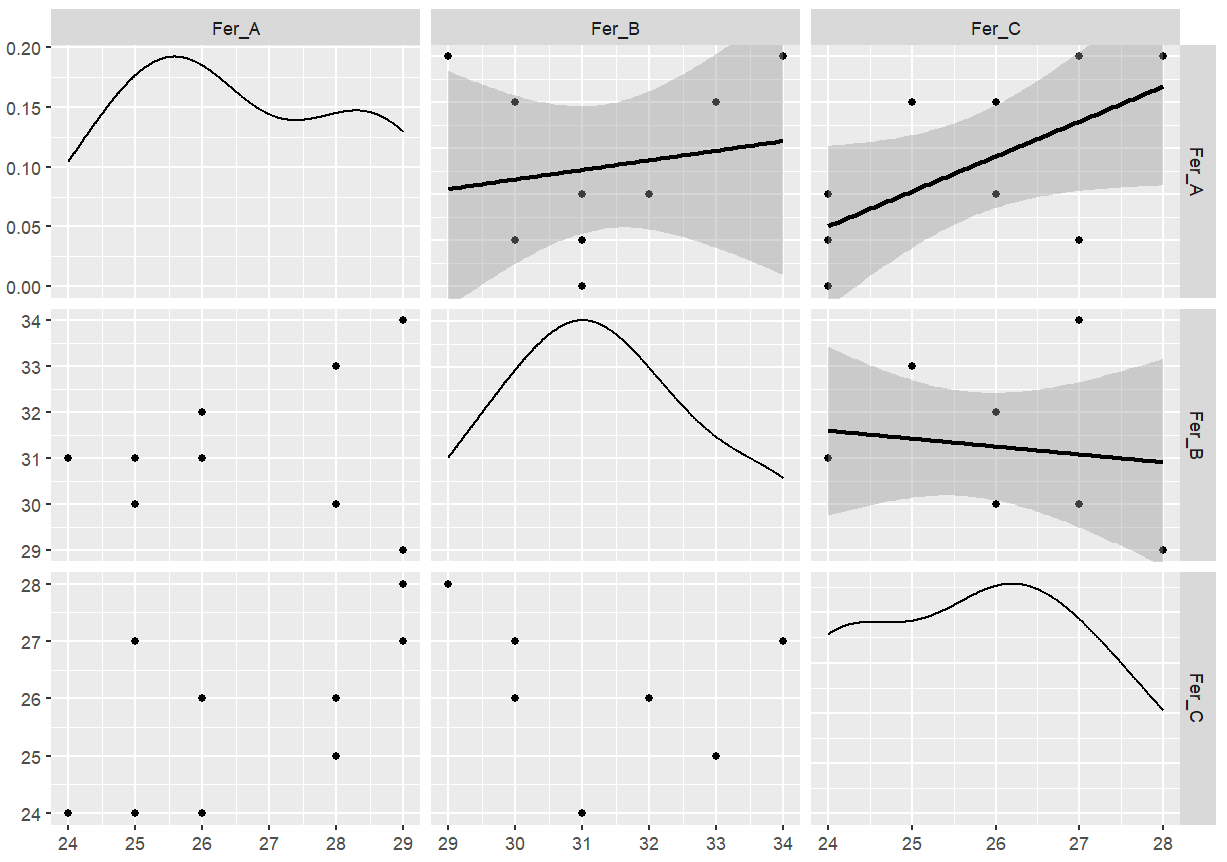
肥料B在促进植物生长方面效果最好，因为它的中心趋势和密度分布都位于较高的位置。

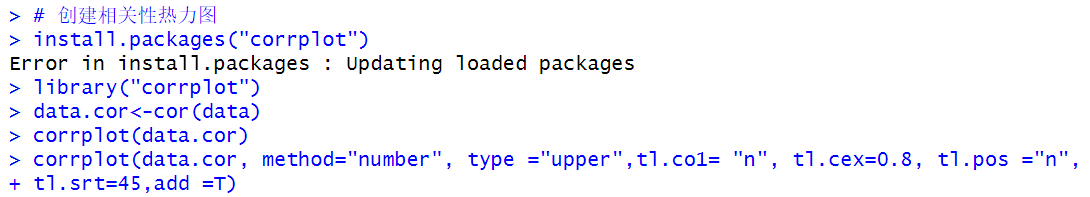
肥料A和肥料C的分布较为接近，但肥料C的中心趋势可能略低于肥料A。

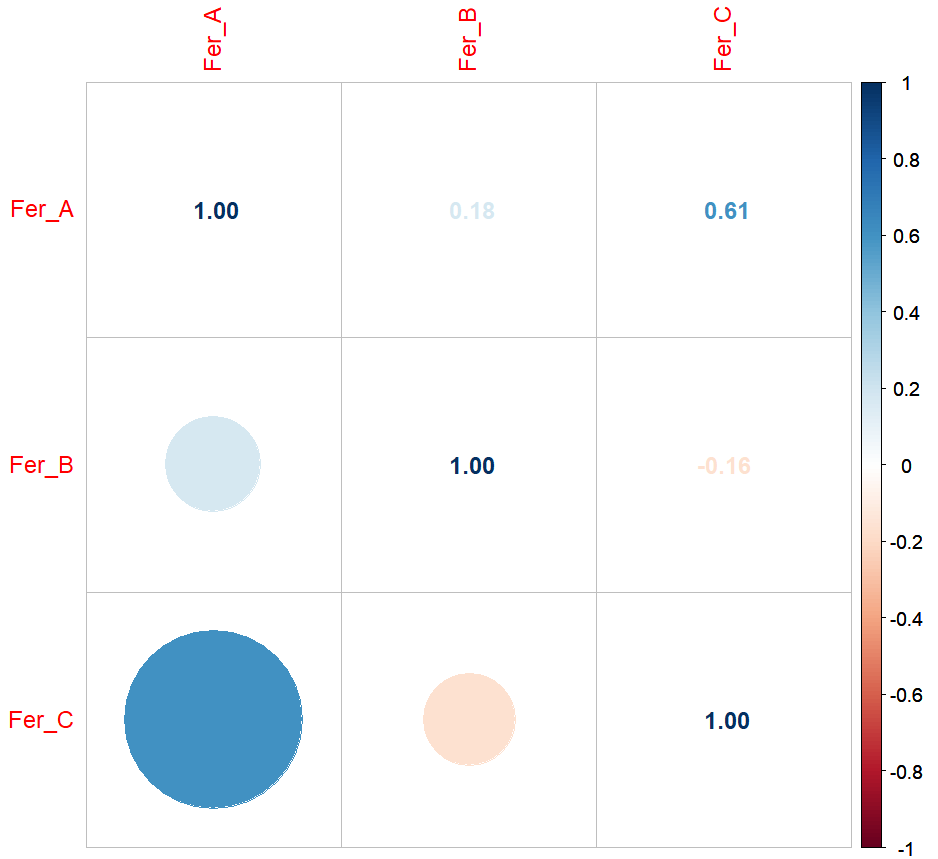










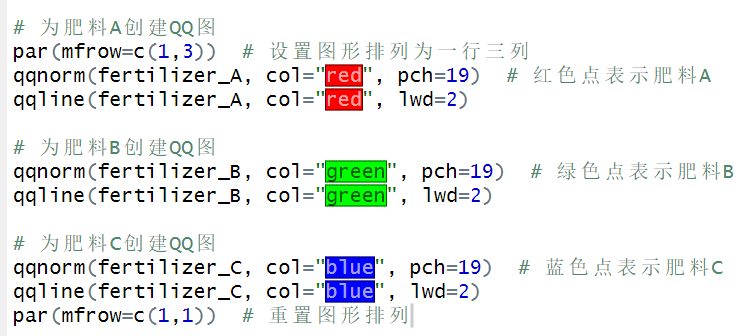


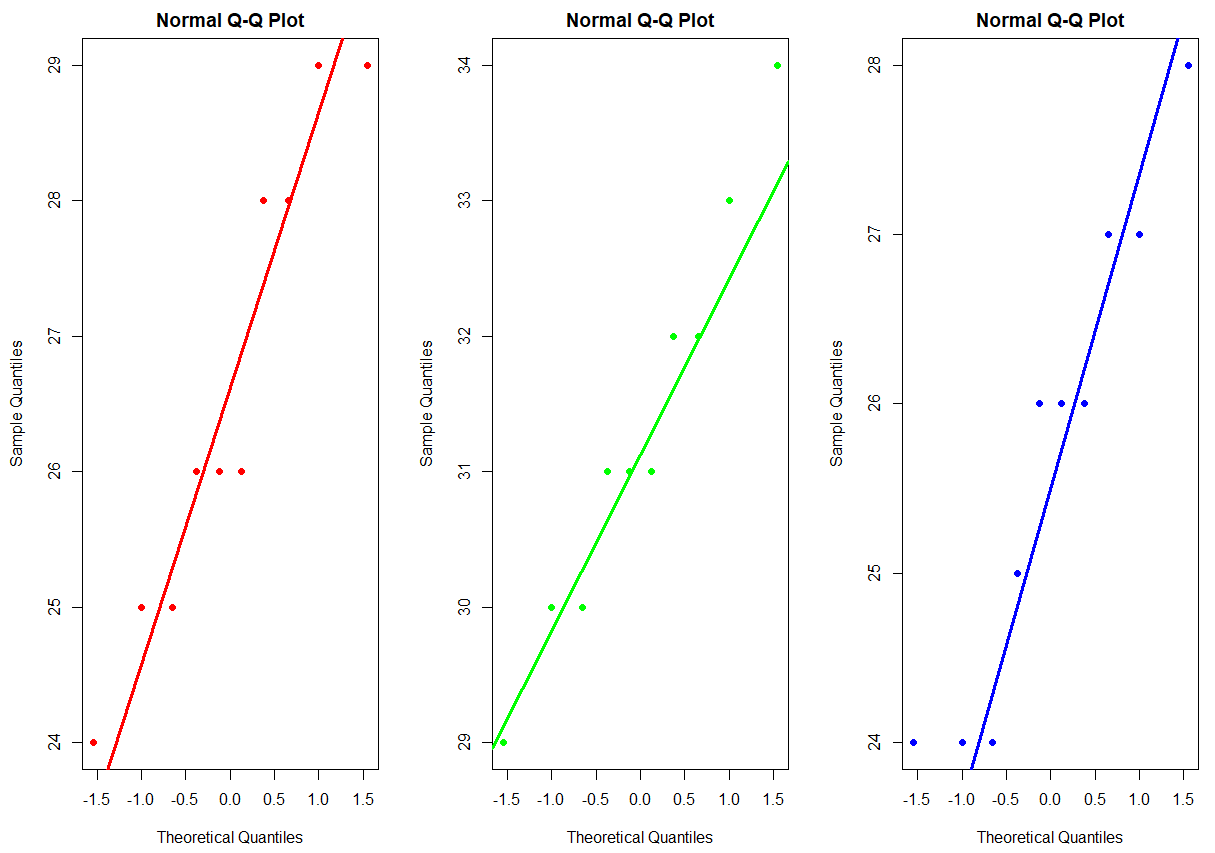
·显著性差异探究：

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。

1. **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）：**





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:

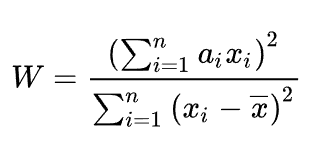
1.假设

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。



其中，

是排好序的样本值，

是样本均值，

是与样本大小相关的常数。

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：

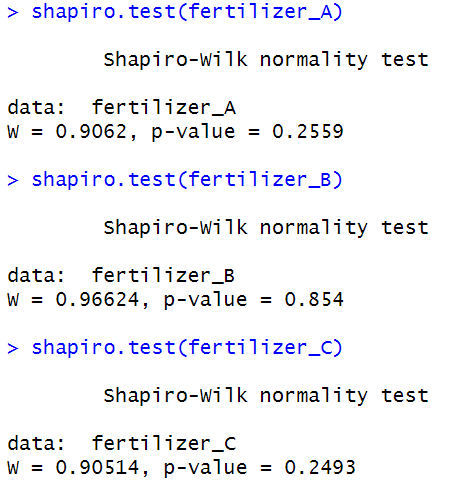
如果p≤α，则拒绝零假设（H0），接受备择假设（H1），认为数据不来自正态分布。

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。



**·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。

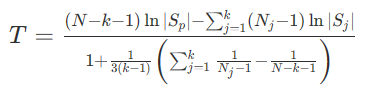
**步骤二（方差齐性检验）：**

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：



其中，N是总样本数，k是总体个数，是第个总体的样本量，是总体协方差矩阵，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：

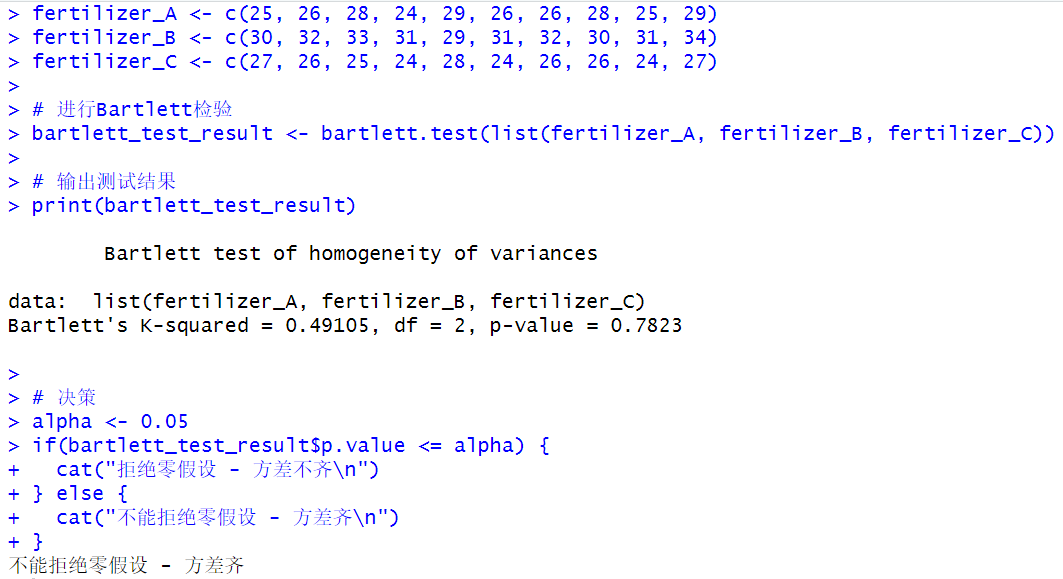
若统计量T落在拒绝域内，则拒绝零假设。

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。



**·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：

，其中SSB为组间平方和，k为组数；

组内均方（MSW）：，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：

3.确定拒绝域：

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

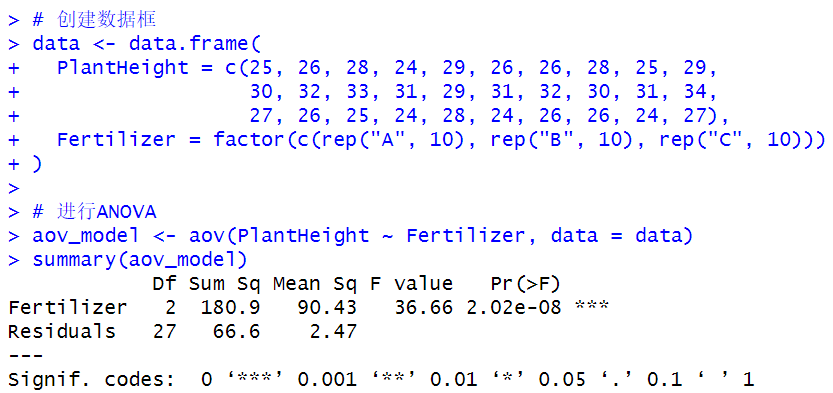
4.决策：

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。



**·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：

其中，和是两组的样本均值，是合并的方差估计，和是两组的样本大小。

·确定拒绝域：

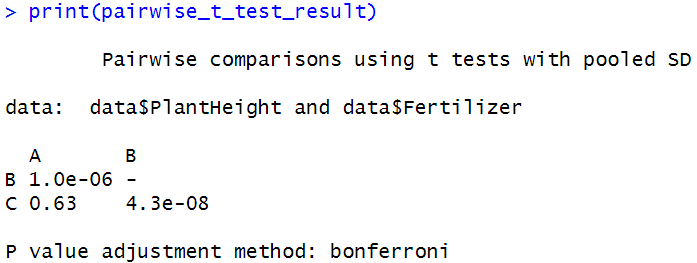
拒绝域由t分布的临界值确定，该临界值基于给定的显著性水平（本题为0.05）和自由度（通常是两个组的样本大小之和减2）。在R中，pairwise.t.test函数会自动计算这些临界值。

·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。



**·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：

其中：

和分别是第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 =

MSE（均方误差）是ANOVA模型中的残差均方误差。

和分别是第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。

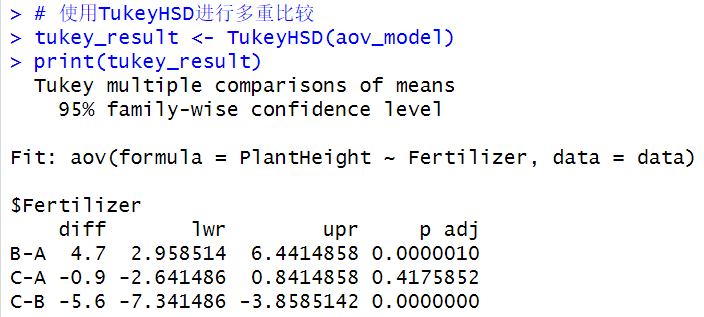
·决策：

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。



**·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。

1. 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：
2. **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。

1. **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。

1. **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。