实验四 R语言综合实验

【实验目的】

1. 掌握使用R语言进行一元线性回归方程和多元线性回归方程分析。
2. 掌握使用R语言进行回归方程和回归参数的假设检验。
3. 掌握使用R语言进行回归诊断（检验：异常值、非线性、残差、多重共线性等）。
4. 掌握使用R语言进行单因素方差和双因素方差分析。

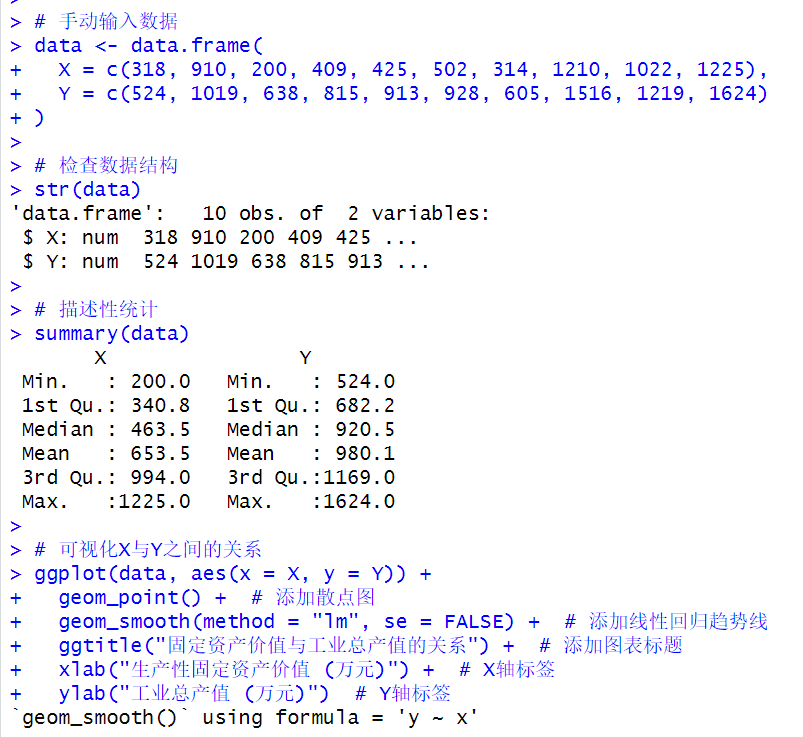
【实验内容与实现】

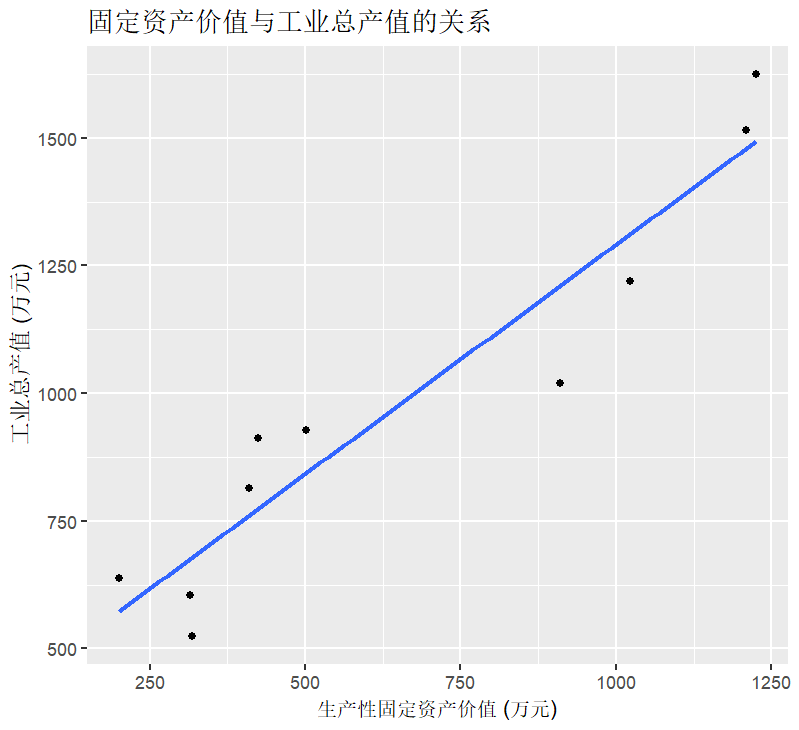
1. 有10个同类企业的生产性固定资产价值(X)和工业总产值(Y )资料如下:

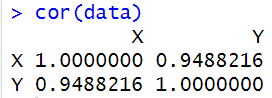


问题如下：

1. 探索性分析X与Y之间的关系







通过散点图可以观察到固定资产价值和工业总产值之间可能存在正相关关系，即固定资产价值较高的企业往往有更高的工业总产值。这种关系在视觉上表现为点的分布倾向于沿直线上升。同时，X和Y的相关系数高达0.9488216，可以推测他们具有很强的线性相关性。

1. 建立X与Y之间的线性回归方程，并对其回归方程进行显著性检验和回归诊断。

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：

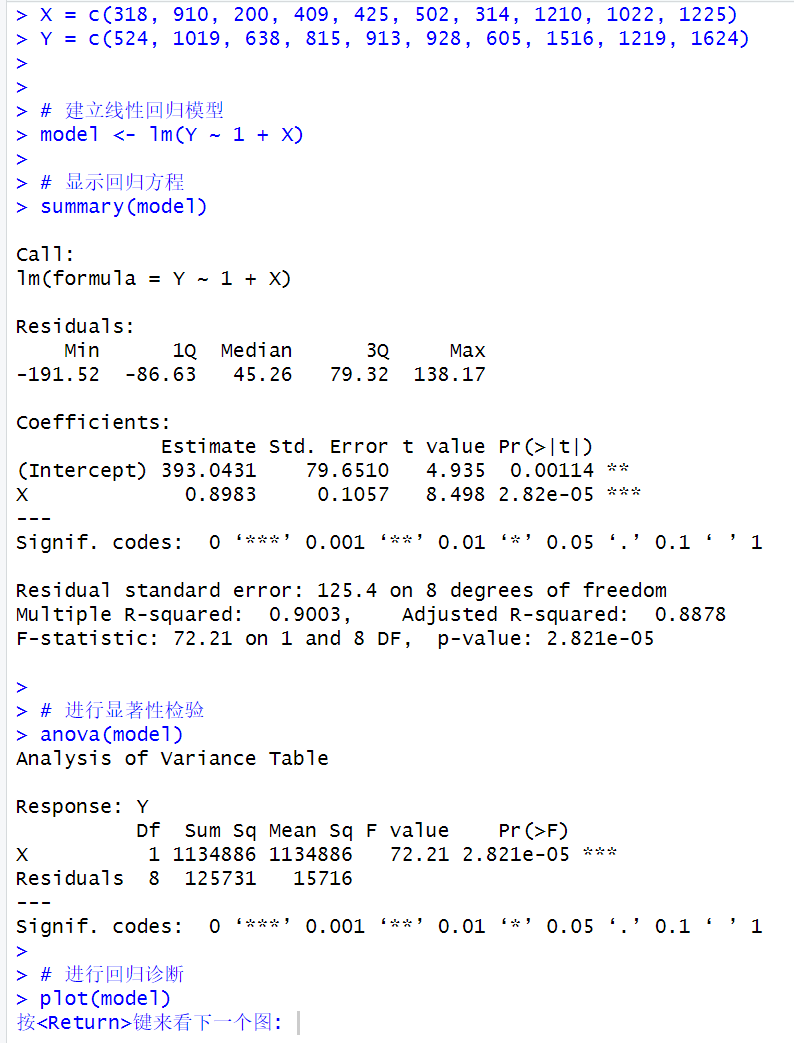
F=(SSR/k)/(SSE/(n-k-1))

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。

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。

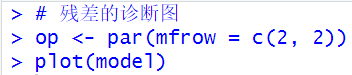
5.结论分析：根据决策的结果，我们可以得出结论。如果拒绝了零假设，我们可以认为对应的自变量系数在回归模型中是显著的，即自变量对因变量有显著影响；如果接受了零假设，则说明自变量对因变量的影响不显著。

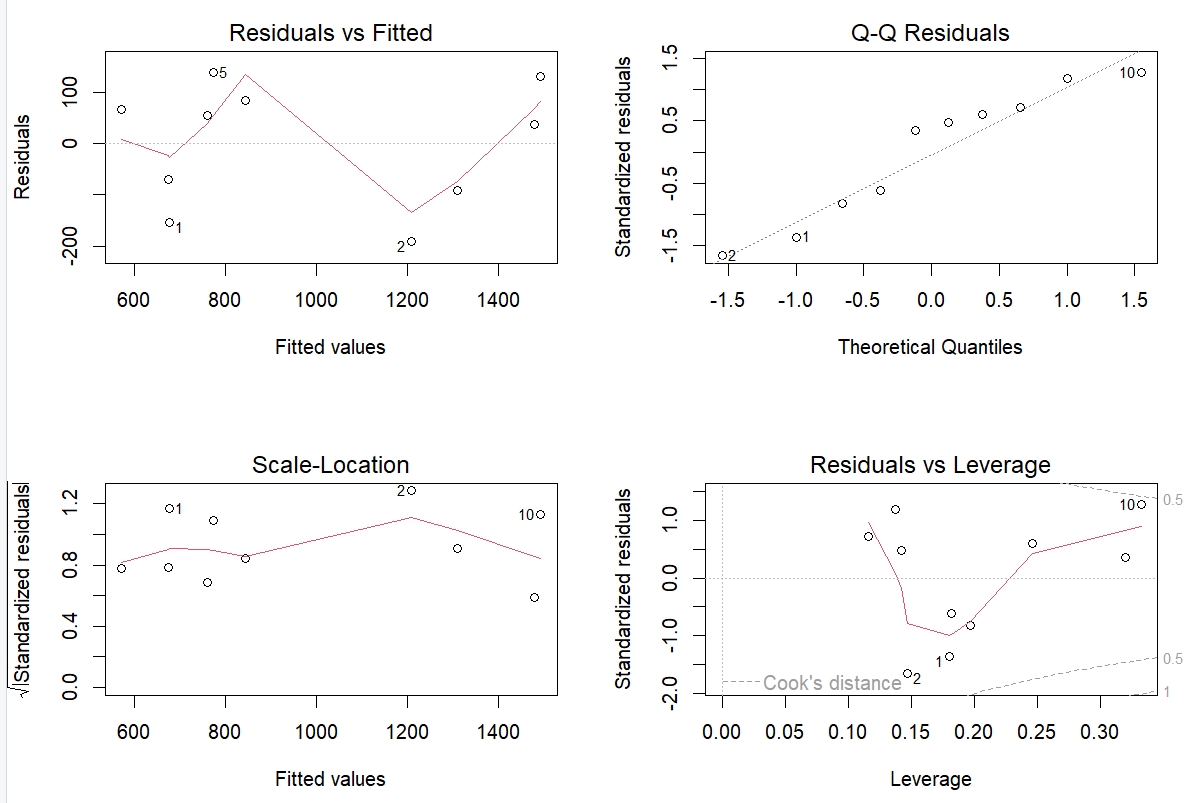


·分析与结论：

1. value = 2.821e-05 << 0.05，故拒绝原假设，至少有一个自变量的系数不等于零，即模型中的自变量对因变量有显著影响，模型的解释能力较强。

·回归诊断





1.Residuals vs Fitted（残差vs拟合值）

由图可知，模型的残差随机分布在拟合值的周围，没有显示出明显的模式或趋势，这表明模型已经适当地捕捉了数据中的线性关系，并且满足同方差性假设。

2.Q-Q Residuals（残差Q-Q图）

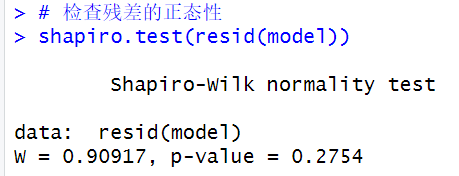
图中的数据点围绕着一条直线排列，这表明残差近似地遵循正态分布。这是线性回归模型的一个重要假设。图片中没有明显显示任何异常的数据点，这表明模型可能没有受到异常值的显著影响。

3.Scale-Location（尺度-位置图）

残差围绕着红线随机分布，没有明显的模式或趋势。残差的散布在不同的拟合值范围内看起来相对均匀，没有出现随着拟合值增加而散布增加或减少的情况，这表明同方差性假设可能得到了满足。图中没有明显的非线性形状或趋势。

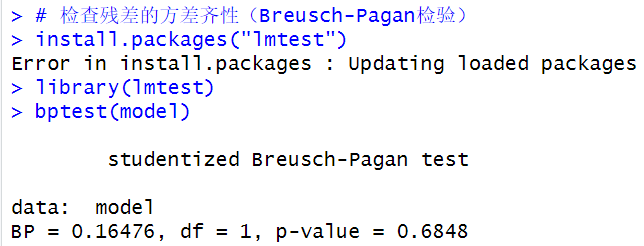
4.Residuals vs Leverage（残差vs杠杆值）

根据图片内容，似乎没有明显的异常点，即没有点同时在高杠杆值和高Cook距离区域，这表明模型可能没有受到个别观测值的显著影响，即没有异常点。



·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0.2754，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0.90917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。



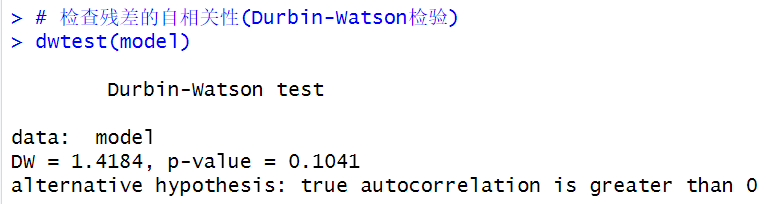
·检验结果分析及结论：

BP值：检验统计量BP值为1.6476，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1，这通常意味着在模型中有1个预测变量（不包括常数项）。

p-value：p值为0.6848，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。



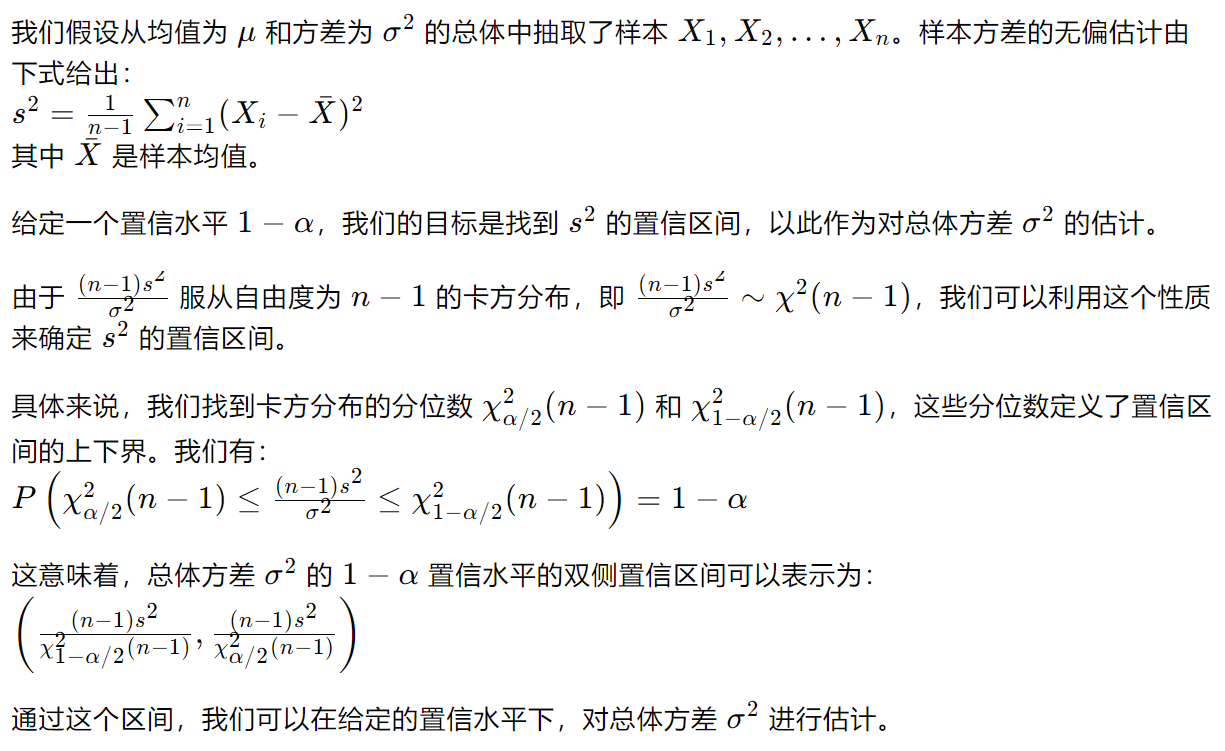
·检验结果分析及结论：

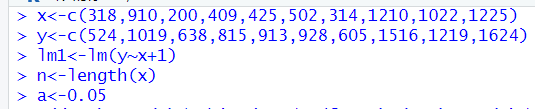
DW值：Durbin-Watson统计量的值为1.4184，表明残差之间没有显著的自相关性。

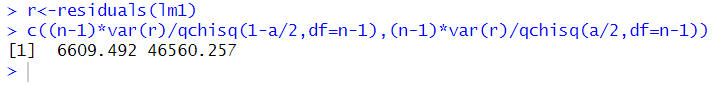
p-value：p值为0.1041，这个值表示在零假设（残差之间不存在自相关性）成立的情况下，观察到的统计量（或更极端）的概率。p值大于常用的显著性水平（如0.05或0.01），这意味着没有足够的证据拒绝零假设。

根据DW检验的结果，DW值为1.4184，且p值为0.1041，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。

1. 设为(b)部分所构建的线性回归模型的误差方差。计算的95%置信区间。







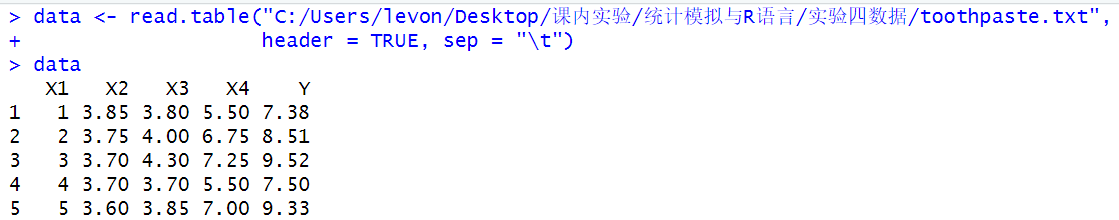
因此，的95%置信区间为[6609.492, 46560.257]。

1. 某大型牙膏制造商为了更好地拓展商品市场，有效地管理库存，公司董事会要求销售部门根据市场调查，找出公司生产的牙膏销售量与销售价格、广告投入等之间的关系，从而预测出在不同价格和广告费用下销售量。为此，销售部的研究人员收集了过去30个销售周期（每个销售周期为4周）公司生产的牙膏的销售周期（X1）、销售价格（X2）、其他厂家生产同类牙膏的市场平均销售价格（X3）、投入的广告费用（X4）、销售量（Y），数据见附件“toothpaste.txt”。试根据这些数据建立一个线性回归函数模型，分析牙膏销售量与其他因素的关系，为制定价格策略和广告投入策略提供数量依据。

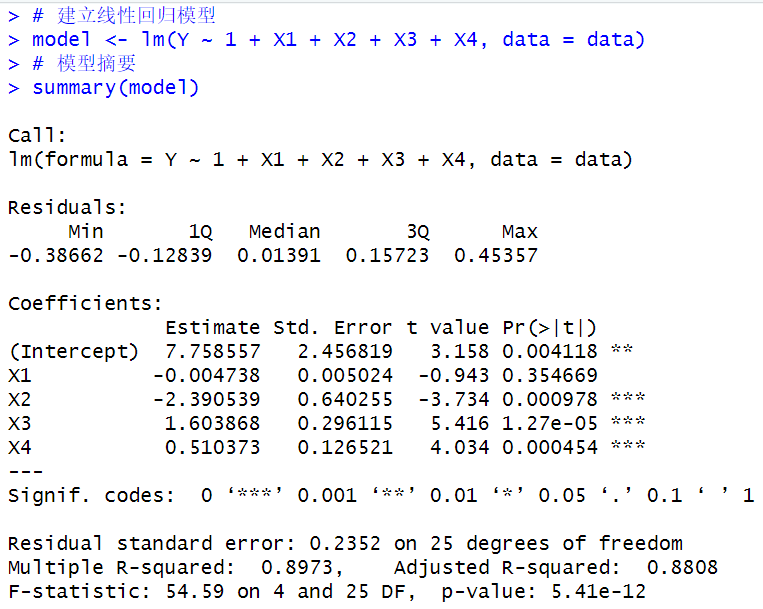
问题如下：

1. 建立牙膏销量与其他因素的回归模型；

首先读入数据：



建立多元线性回归模型：



·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。

·系数:

截距: 截距为7.758557，标准误为2.456819，t值为3.158，对应的p值为0.004118，表示模型的截距在统计上是显著的。

X1: 系数为-0.004738，标准误为0.005024，t值为-0.943，p值为0.354669，表示X1对Y的影响在统计上不显著。

X2: 系数为-2.390539，标准误为0.640255，t值为-3.734，p值为0.000978，表示X2对Y有显著的负影响。

X3: 系数为1.603868，标准误为0.296115，t值为5.416，p值为1.27e-05，表示X3对Y有显著的正影响。

X4: 系数为0.510373，标准误为0.126521，t值为4.034，p值为0.000454，表示X4对Y有显著的正影响。

·模型拟合度:

残差标准误: 为0.2352，表示观测值与模型预测值之间的标准偏差。

自由度: 为25（df = n - p - 1，其中n为样本量，p为自变量数量加1）。

R平方: 为0.8973，表明模型解释了因变量变异的89.73%。

调整R平方: 为0.8808，调整R平方考虑了模型中变量的数量，通常用于多元回归分析。

F统计量: 为54.59，对应的p值为5.41e-12，表示模型整体显著。

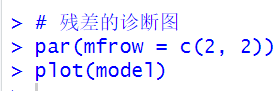
显著性代码: 提供了不同显著性水平的代码，如\*\*\*表示p < 0.001。

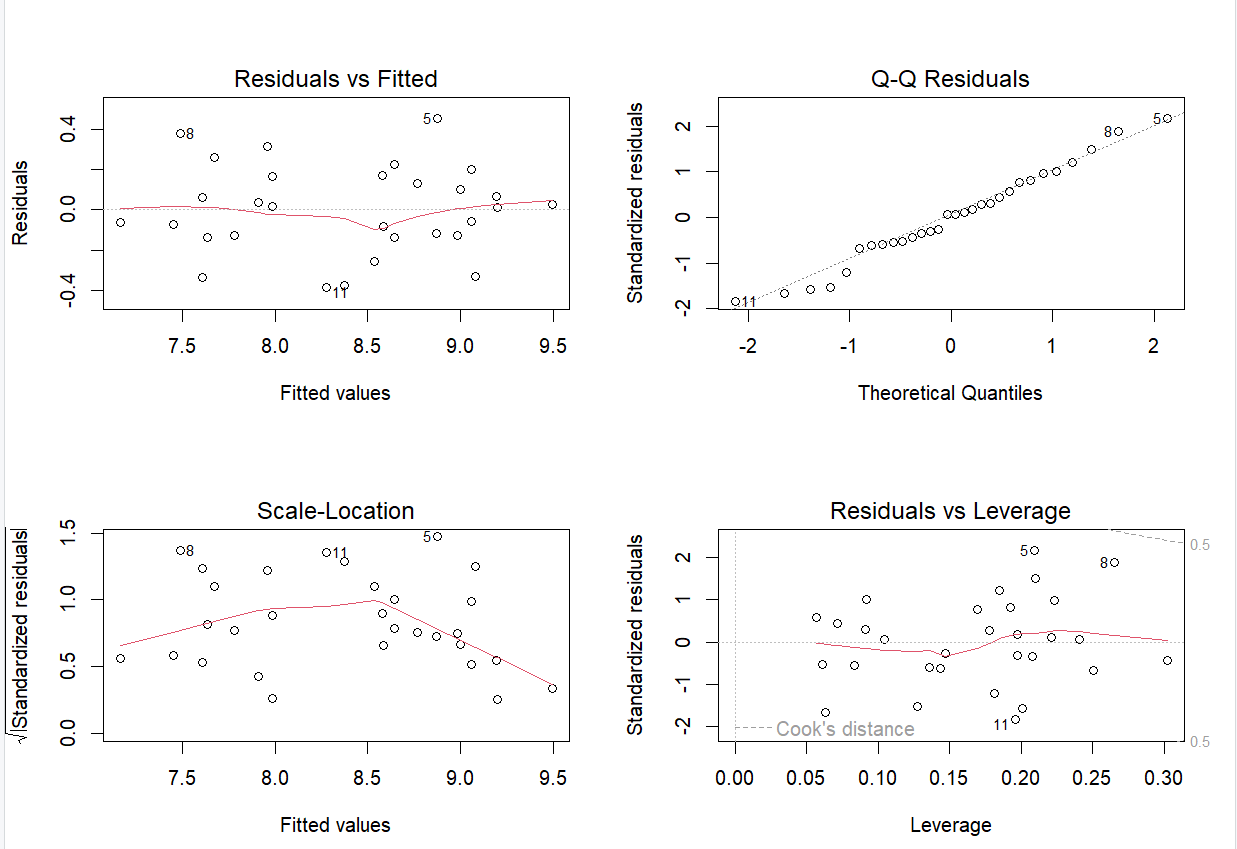
·结论:

这个线性回归模型整体上是显著的，其中X2和X4对牙膏销售量有显著的正影响，而X3也有显著的正影响。X1的影响不显著。模型拟合度较高，可以解释大部分销售量的变化。

1. 对得到的线性模型做回归诊断，分析是否存在异常样本，是否有多重共线性问题存在；如果有需要删除的异常样本和多重共线性，做相应处理后找到最优回归模型。

·回归诊断：





1.Residuals vs Fitted（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-Q Residuals（残差Q-Q图）

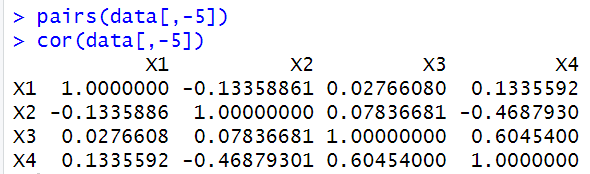
Q-Q图（Quantile - Quantile Plot）用于评估残差的正态性。图中的点近似落在一条直线上，表明残差服从正态分布。

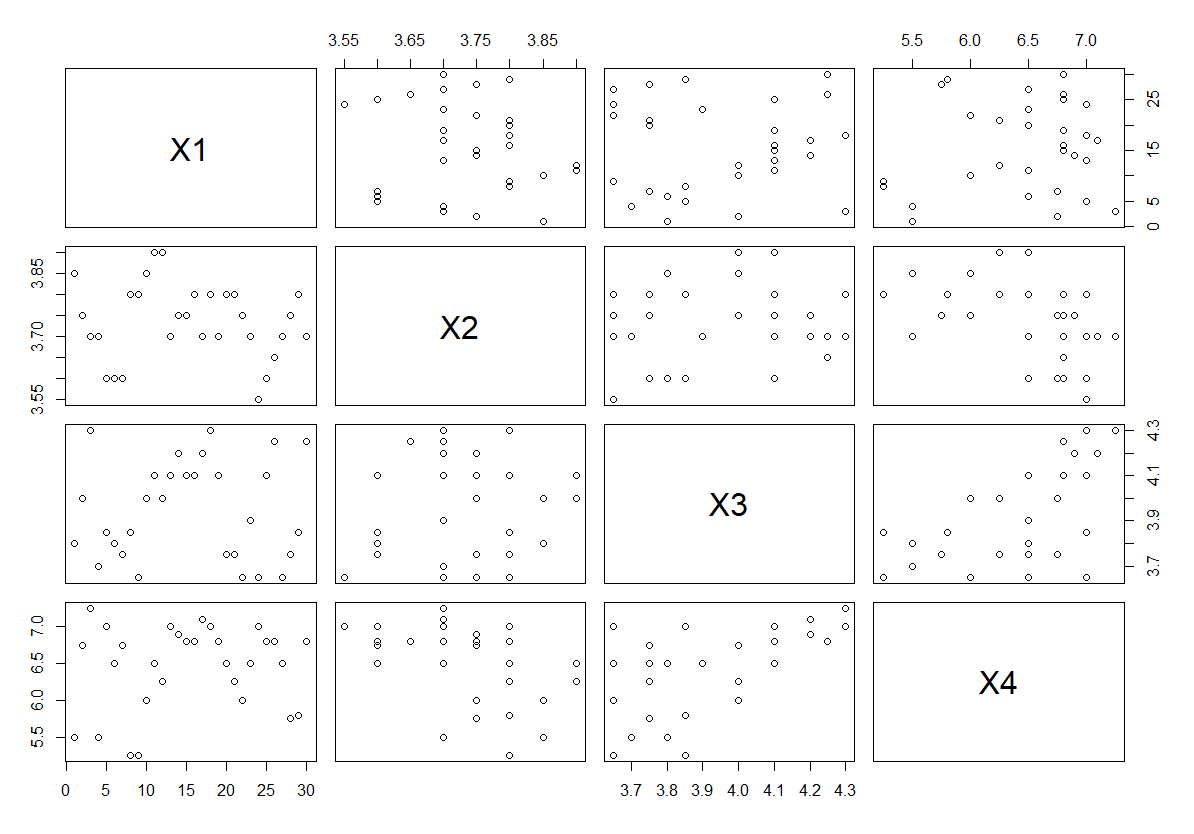
3.Scale-Location（尺度-位置图）

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。

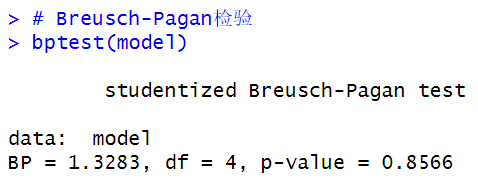
4.Residuals vs Leverage（残差vs杠杆值）

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距离小于0.5，表明不存在明显的异常点。





从相关系数矩阵和散点图矩阵中可以看出X3和X4具有较为明显的线性关系，因此推断此样本的数据间存在多重共线性。



·检验结果分析：

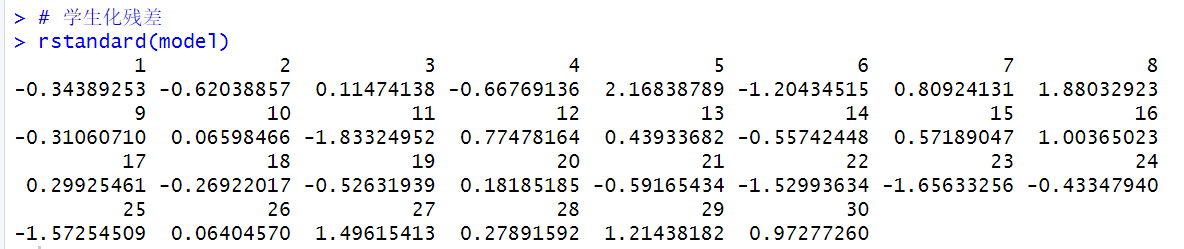
BP值：检验的统计量为1.3283。这是基于残差平方和与自变量之间关系的度量。

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0.8566。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）。

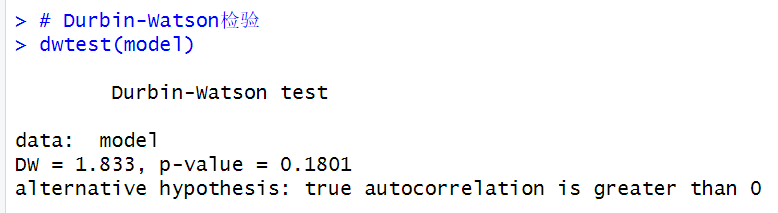
·结论：

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。



·计算结果分析及结论：

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，虽然第5个观测值的学生化残差的绝对值略大于2，但是在杠杆图中第五个点的Cook距离是小于0.5的，因而不认为这个点是异常点。



·检验结果分析：

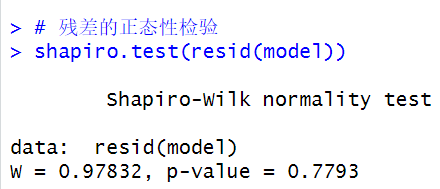
DW值：检验的Durbin-Watson统计量为1.833。这个值表明残差之间可能存在轻微的正自相关，因为DW值小于2（但接近2）。

p值：检验的p值为0.1801，这是一个较高的p值，大于显著性水平0.05。

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。



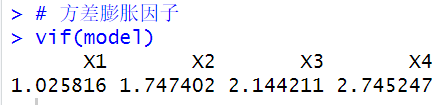
·检验结果分析：

W值：Shapiro-Wilk检验的W值为0.97832。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0.7793，这是一个非常高的p值，远大于显著性水平0.05。

·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。



方差膨胀因子是一个衡量自变量之间线性关系强度的指标。如果一个变量可以通过其他变量的线性组合来预测，那么这些变量之间存在多重共线性。VIF值大于1表示存在共线性，VIF值越高，共线性越严重。

·分析：

X1: VIF值为1.025816，接近1，表示X1与其他自变量之间的共线性较弱。

X2: VIF值为1.747402，略高于1，可能表明X2与其他变量存在一定程度的共线性，但通常不认为是严重的问题。

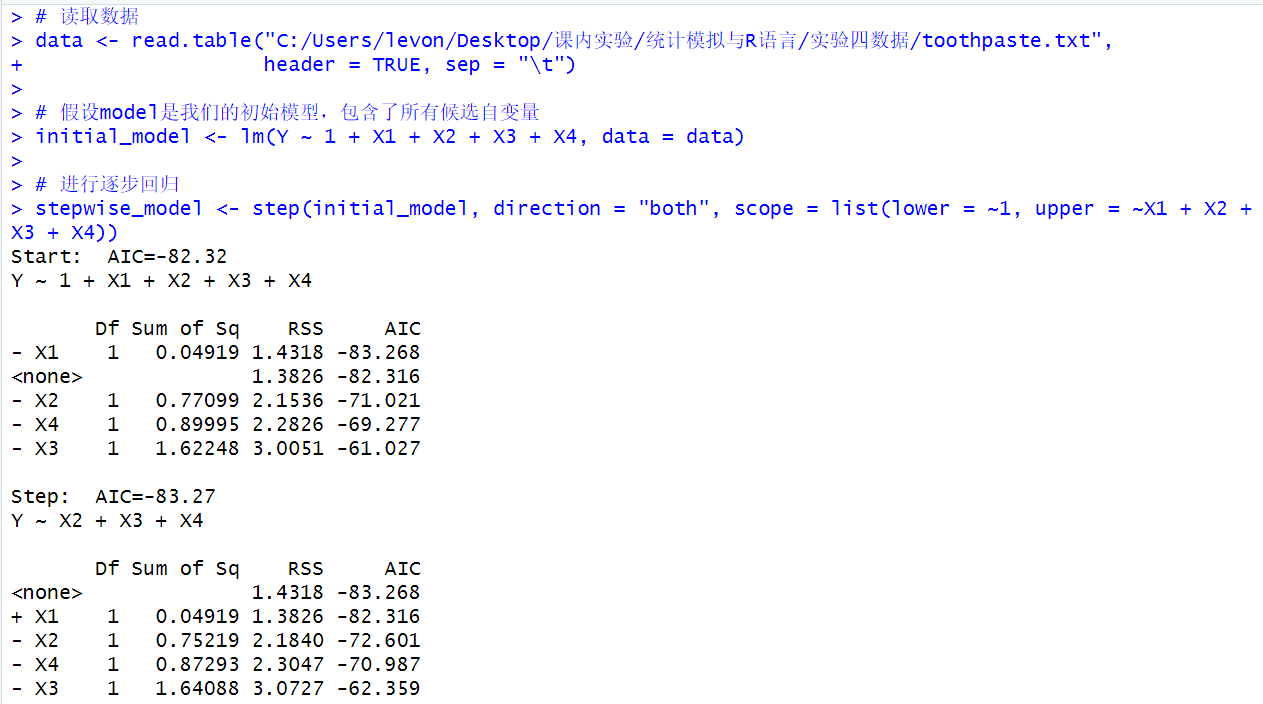
X3: VIF值为2.144211，高于2，表明X3与其他变量之间存在中等程度的共线性。

X4: VIF值为2.745247，也高于2，表明X4与其他变量之间存在中等程度的共线性。

·结论:

根据VIF值，X3和X4显示出中等程度的共线性，而X1和X2的共线性较弱。通常，VIF值大于5或10时，才会被认为是严重共线性，需要采取措施来解决。在这种情况下，X3和X4的共线性可能需要进一步的调查。

通过上面的残差诊断图和各种检验方法，可以判断出数据中不存在异常点，但是存在多重共线性，因此使用逐步回归来解决这一问题：



·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。

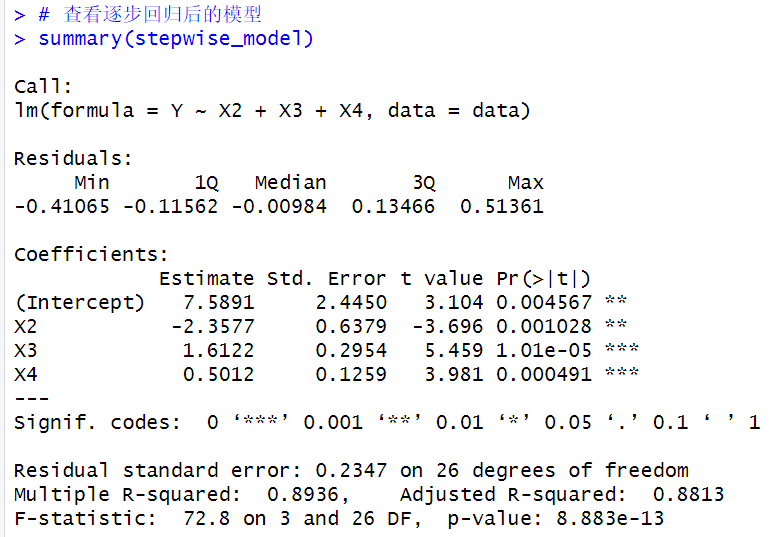
模型评估：模型使用赤池信息准则（AkaikeInformationCriterion,AIC）来评估。AIC是一个衡量模型拟合优度的指标，同时对模型复杂度进行惩罚。较低的AIC值通常表示更好的模型。

逐步删除：

第一步，模型删除了X1，AIC从-82.316降低到-83.268，残差平方和（RSS）从1.3826增加到1.4318，自由度（Df）减少了1。

第二步，模型考虑重新添加X1或删除X2或X3或X4。根据AIC值，删除X2会使得AIC从-83.268增加到-72.601，因此X2被保留。删除X4会使得AIC从-83.268增加到-70.987，所以X4也被保留。删除X3会使得AIC从-83.268增加到-62.359，因此X3也被保留。模型没有进行任何改变，因为添加X3会使AIC从-83.268增加到-82.316。最终模型：经过逐步回归后，最终模型只包含

X2,X3,X4这三个自变量。



·系数估计:

截距: 截距项的估计是 7.5891，标准误为 2.4450，t值为 3.104，p值为 0.004567，表示模型的截距在统计上是显著的。

X2: 系数为 -2.3577，标准误为 0.6379，t值为 -3.696，p值为 0.001028，表示 X2 对 Y 有显著的负影响。

X3: 系数为 1.6122，标准误为 0.2954，t值为 5.459，p值为 1.01e-05，表示 X3 对 Y 有显著的正影响。

X4: 系数为 0.5012，标准误为 0.1259，t值为 3.981，p值为 0.000491，表示 X4 对 Y 有显著的正影响。

·模型拟合度:

残差标准误差: 为 0.2347，表示观测值与模型预测值之间的标准偏差。

自由度: 为 26（df = n - (k + 1)，其中 n 为样本量，k 为自变量数量）。

多重R平方: 为 0.8936，表明模型解释了因变量变异的 89.36%。

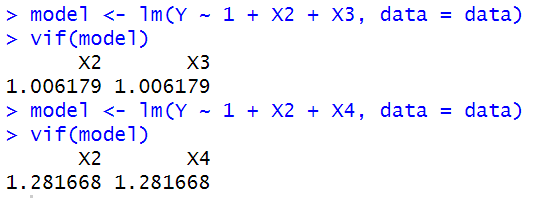
调整R平方: 为 0.8813，调整R平方考虑了模型中变量的数量，通常用于多元回归分析。

F统计量: 为 72.8，对应的 p值为 8.883e-13，表示模型整体显著。

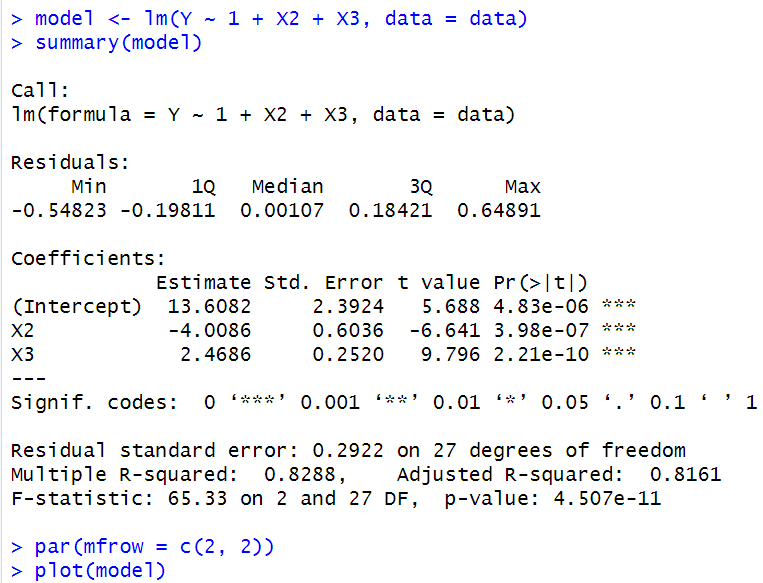
·结论：

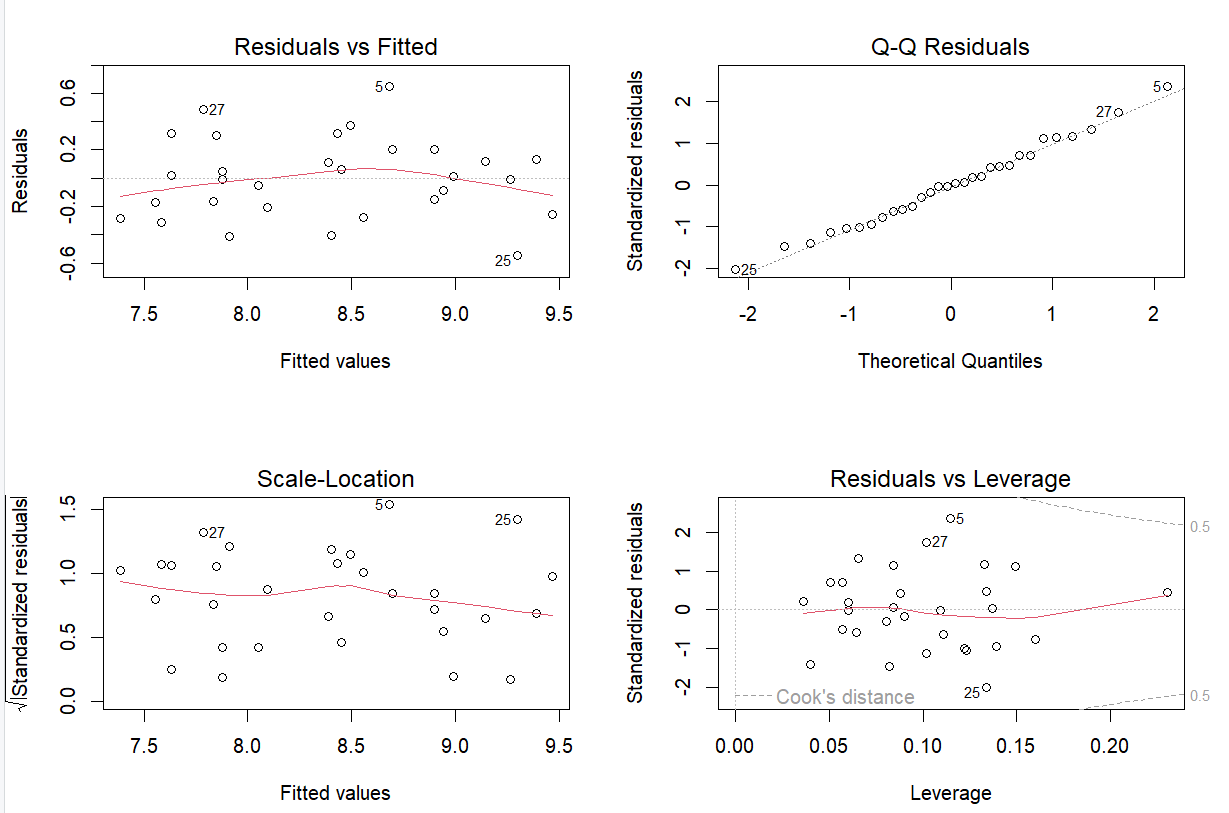
逐步回归后的模型显示 X2、X3 和 X4 与因变量 Y 有显著的线性关系。模型拟合度较高，调整R平方值为 0.8813，表明模型能够解释大部分因变量的变异。所有自变量的系数都在统计上显著，其中 X2 与 Y 负相关，而 X3 和 X4 与 Y 正相关。

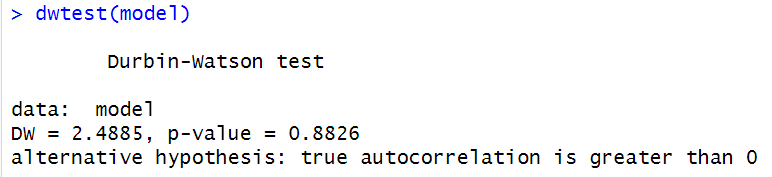
然而，X3与X4仍然存在于模型当中，他们的多重共线性问题并没有得到解决，考虑直接去除其中的某一个变量，并对比方差膨胀因子：



可以看到，去除X4之后的方差膨胀因子更小，同时也解决了多重共线性问题，对该模型进行显著性检验和回归诊断：

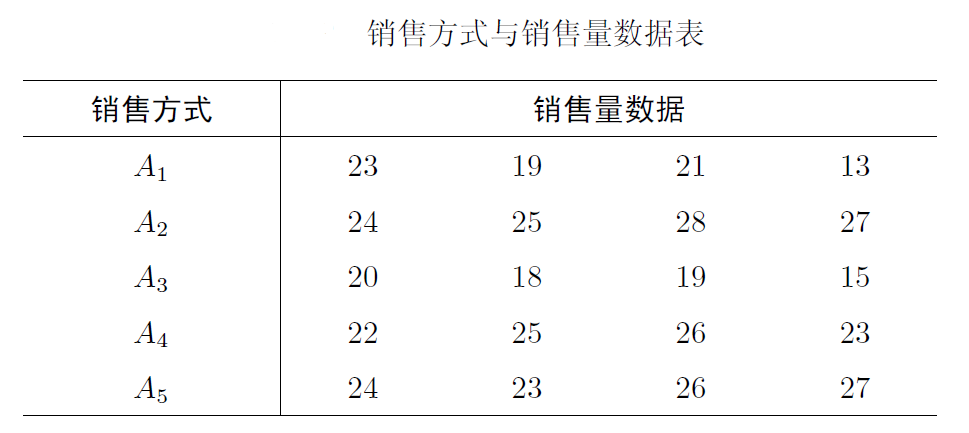






可以发现，模型是极为显著的，并且没有异常点，同时残差也满足独立性，等方差性和正态性（详细的分析过程同上），因此认为该模型（Y ~ 1 + X2 + X3）为最优模型。并且此模型整体拟合度很高，可以作为预测牙膏销售量的有效工具。

1. 某商店以各自的销售方式卖出新型手表，连续四天手表的销售量如表所示，试考察销售方式之间是否有显著差异。（提示：方差分析的步骤包括1.正态性检验，2.方差齐性检验，3.方差分析，4.多重比较）



**1.正态性检验**

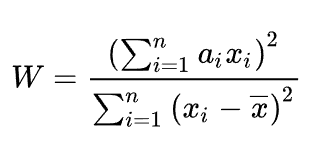
·假设

零假设(H0):数据来自一个正态分布的总体。这意味着样本数据的分布特性与正态分布一致。

备择假设(H1):数据不是来自正态分布的总体。这表明样本数据的分布特性与正态分布不一致。

·构造统计量

使用Shapiro-Wilk检验来定量评估数据的正态性。Shapiro-Wilk检验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。

·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。

·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：

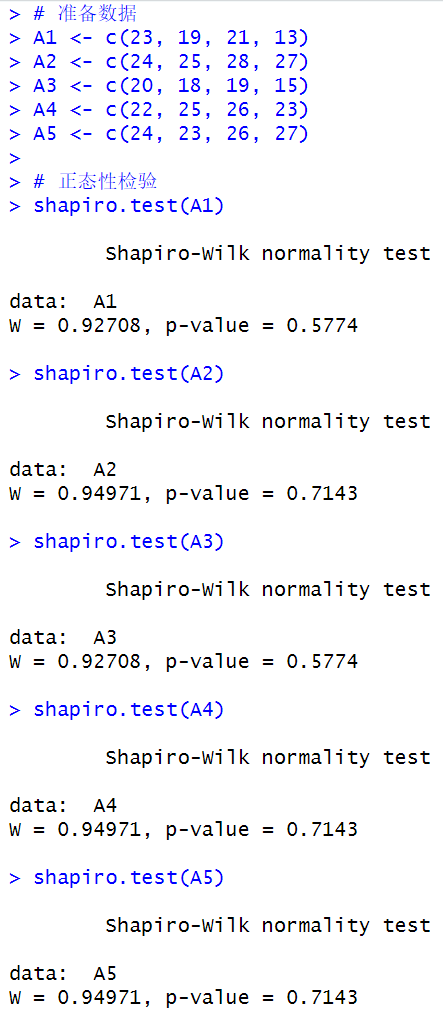
如果p≤α，则拒绝零假设（H0），接受备择假设（H1），认为数据不来自正态分布。

如果p>α，则不能拒绝零假设（H0），认为数据来自正态分布。

·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。



以上数据的P-value均大于0.05，因此数据满足正态性。

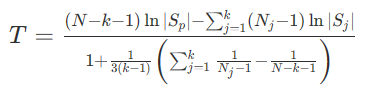
**2.方差齐性检验**

·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。

·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。

·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。

·决策：

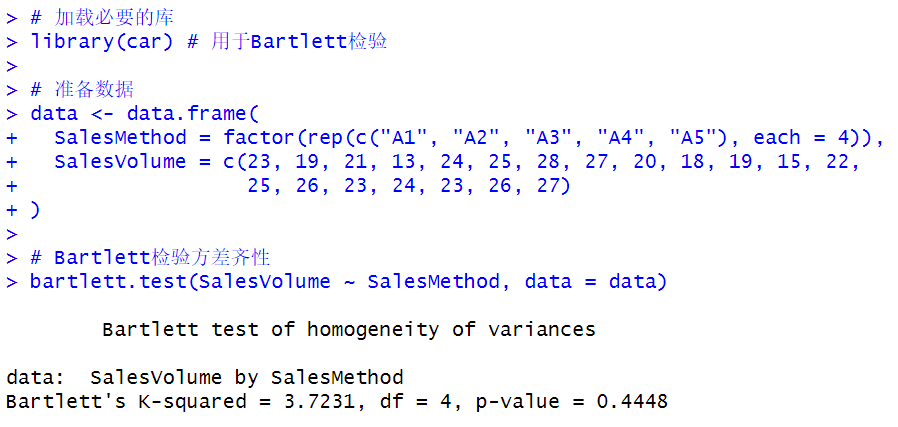
若统计量T落在拒绝域内，则拒绝零假设。

若统计量T不在拒绝域内，则接受零假设。

·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（0.4448）大于显著性水平（0.05），不能拒绝零假设，即认为各总体方差相等，满足方差齐性，可以继续进行ANOVA分析。

**3.方差分析**

·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。

·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：

·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。

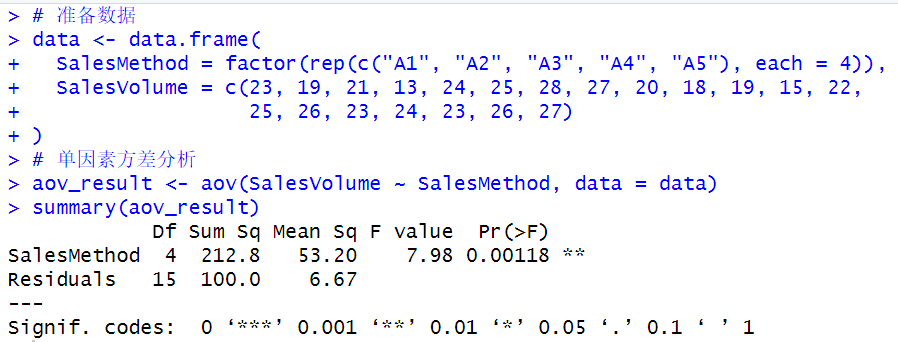
·决策：

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。

·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。



·结果分析：

销售方式的F值为7.98，P值为0.00118。由于P值远小于0.05，我们可以拒绝原假设（即所有销售方式的平均销售量相同），认为至少有两种销售方式之间存在显著差异。

4.多重比较

·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：

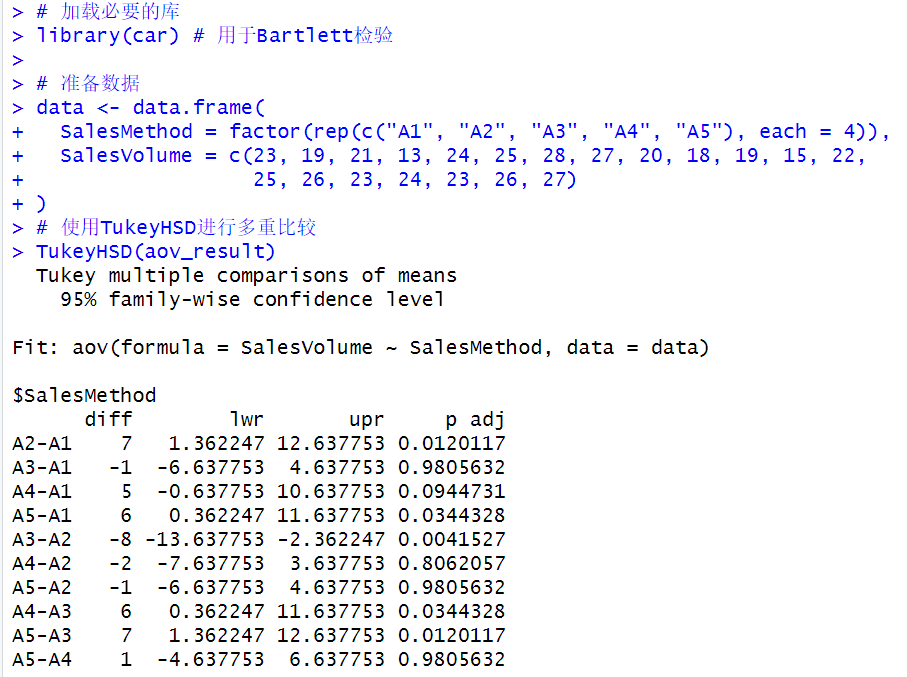
拒绝域由Tukey检验的临界值确定，该临界值基于给定的显著性水平（通常为0.05）和自由度。在R中，TukeyHSD()函数根据组间比较的数量和样本大小自动确定拒绝域。

·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（本题为0.05），则认为组间差异是统计学上显著的。



·分析与结论

A2-A1: A2的销售量平均比A1高7个单位，差异显著，置信区间不包含0，P值小于0.05。

A3-A1: A3与A1之间的销售量差异不显著，置信区间包含0，P值远大于0.05。

A4-A1: A4的销售量平均比A1高5个单位，差异接近显著，但P值略高于0.05。

A5-A1:A5的销售量平均比A1高6个单位，差异显著，P值小于0.05。

A3-A2: A3的销售量平均比A2低8个单位，差异显著，P值小于0.01。

A4-A2:A4与A2之间的销售量差异不显著，P值大于0.05。

A5-A2: A5与A2之间的销售量差异不显著，P值远大于0.05。

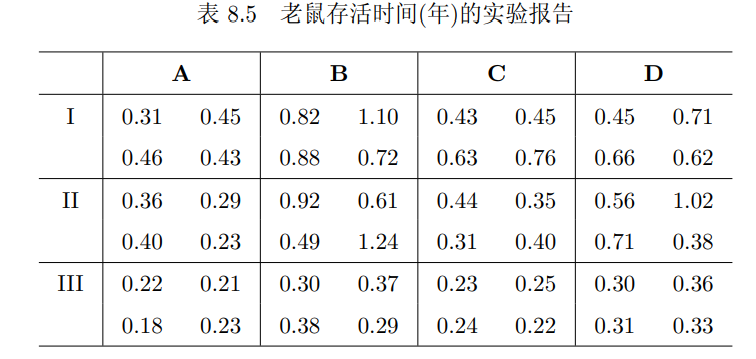
A4-A3:A4的销售量平均比A3高6个单位，差异显著，P值小于0.05。

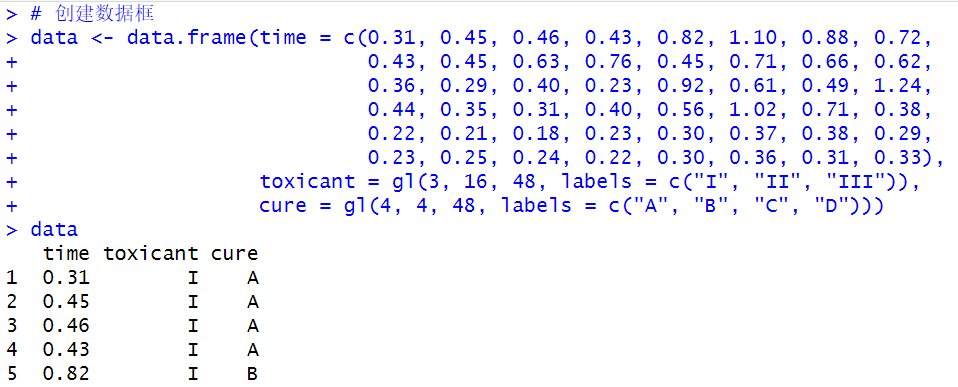
A5-A3:A5的销售量平均比A3高7个单位，差异显著，P值小于0.05。

A5-A4:A5与A4之间的销售量差异不显著，置信区间包含0，P值大于0.05。

**综上，A2-A1，A5-A1，A3-A2，A4-A3，A5-A3这五组销售方式的组内存在显著差异，另外五组的差异不显著。**

1. 有一个关于检验毒品强弱的试验, 给48只老鼠注射I、II、III三种毒药(因素A), 同时有A、B、C、D 4种治疗方案(因素B), 这样的试验在每一种因素组合下都重复四次测试老鼠的存活时间, 数据如下表所示. 试分析毒药和治疗方案以及它们的交互作用对老鼠存活时间有无显著影响.





**步骤一（正态性检验）：**

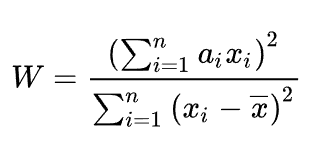
1.假设

零假设(H0):数据来自一个正态分布的总体。这意味着样本数据的分布特性与正态分布一致。

备择假设(H1):数据不是来自正态分布的总体。这表明样本数据的分布特性与正态分布不一致。

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（例如，0.05或0.01）。

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：

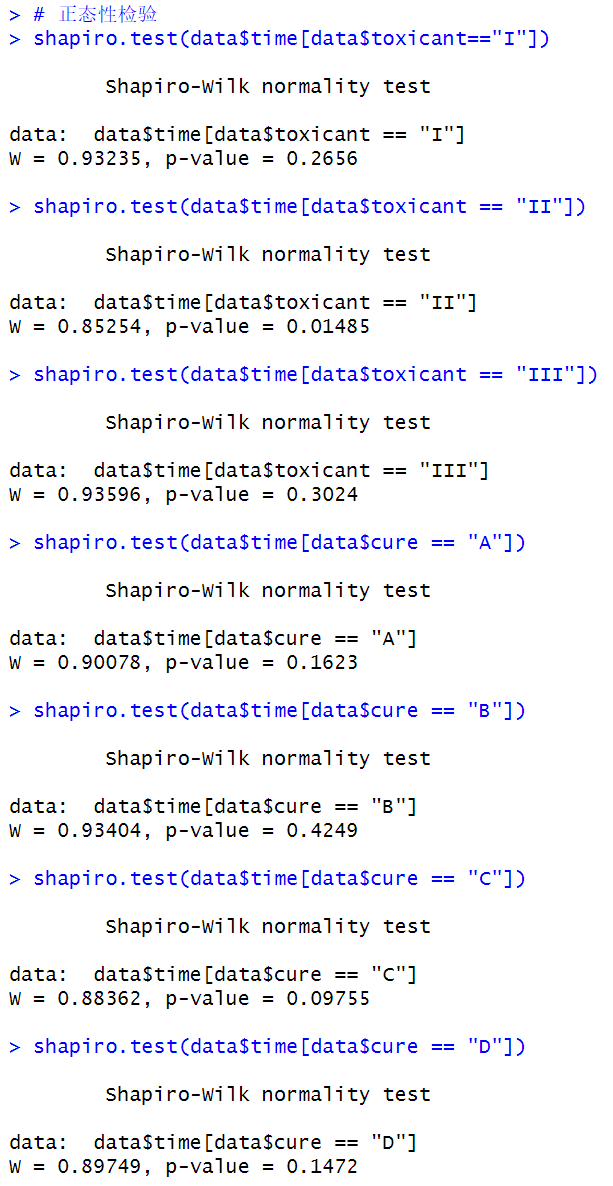
如果p≤α，则拒绝零假设（H0），接受备择假设（H1），认为数据不来自正态分布。

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。



**·结果分析：**

对于毒药I (toxicant == "I")，p值为0.2656，大于0.05，表明正态性假设无法被拒绝。

对于毒药II (toxicant == "II")，p值为0.01485，小于0.05，表明正态性假设被拒绝。

对于毒药III (toxicant == "III")，p值为0.3024，大于0.05，正态性假设无法被拒绝。

对于治疗方案A (cure == "A")，p值为0.1623，大于0.05，正态性假设无法被拒绝。

对于治疗方案B (cure == "B")，p值为0.4249，大于0.05，正态性假设无法被拒绝。

对于治疗方案C (cure == "C")，p值为0.09755，大于0.05，正态性假设无法被拒绝。

对于治疗方案D (cure == "D")，p值为0.1472，大于0.05，正态性假设无法被拒绝。

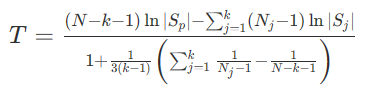
**步骤二（方差齐性检验）：**

·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。

·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。

·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。

·决策：

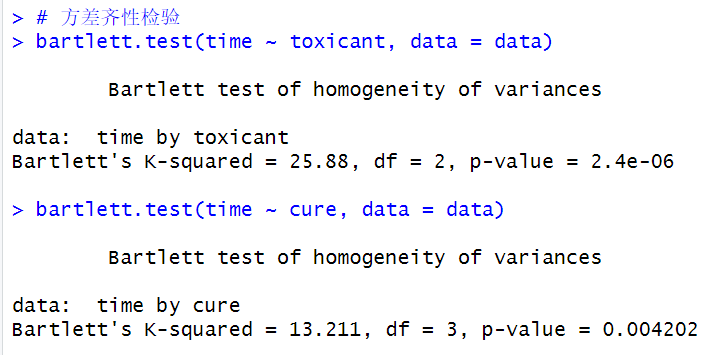
若统计量T落在拒绝域内，则拒绝零假设。

若统计量T不在拒绝域内，则接受零假设。

·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。



**·结果分析：**

对于毒药类型，p值为2.4e-06，远小于0.05，表明不同毒药类型下存活时间的方差不相等。

对于治疗方案，p值为0.004202，小于0.05，表明不同治疗方案下存活时间的方差不相等。

**步骤三（方差分析）：**

·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。

·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：

·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。

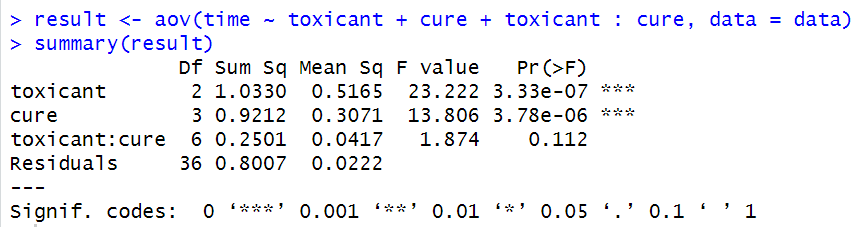
·决策：

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。

·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。



**·结果分析：**

毒药类型(toxicant)：F值为23.222，p值为3.33e-07，远小于0.05，表明毒药类型对存活时间有显著影响。

治疗方案(cure)：F值为13.806，p值为3.78e-06，远小于0.05，表明治疗方案对存活时间有显著影响。

毒药类型与治疗方案的交互作用(toxicant:cure)：F值为1.874，p值为0.112，大于0.05，表明交互作用对存活时间没有显著影响。

综上所述，毒药类型和治疗方案对老鼠的存活时间有显著影响，而它们的交互作用对存活时间没有显著影响。

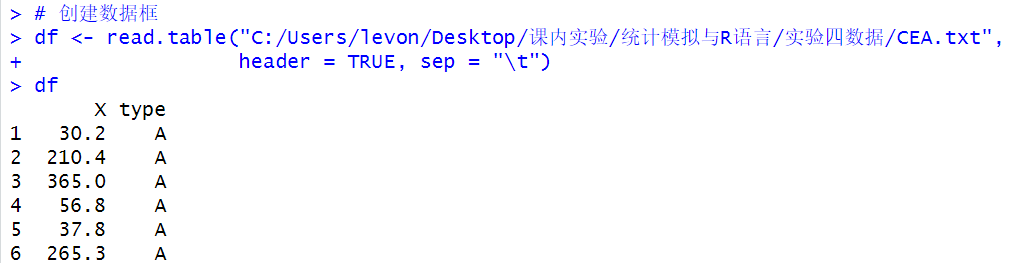
然而，需要注意的是：由于方差分析（ANOVA）的假设包括数据的正态性和方差齐性，当这些假设不满足时，ANOVA的结果可能会受到影响。

但是，虽然Bartlett检验表明方差不齐，但这并不意味着ANOVA的结果完全不可靠。Bartlett检验对小样本数据比较敏感，而且当样本量较大时，即使是小的方差差异也可能导致显著的p值。在实际研究中，如果样本量足够大，即使方差不齐，ANOVA的结果也可能是稳健的。而Shapiro-Wilk检验对小样本（n < 50）的数据比较敏感。当样本量增加时，非正态分布的数据也可能近似正态分布。此外，ANOVA对正态性假设的违反具有一定的鲁棒性，尤其是当样本量较大时。在这种情况下，如果样本量足够大，或者数据的偏态或峰态不是非常极端，ANOVA的结果仍然可以提供有用的信息。

【作业】（说明：请附上每个题的代码、作图、文字分析和描述。）

1. 在对比研究中观察正常人、萎缩性胃炎和胃癌三个不同群体(用TYPE=A, B和C表示), 记录的数据附件“CEA.txt”，试对该组数据作方差分析。

数据读入：



1. 检验三个群体中CEA含量的分布是否为正态分布, 方差是否相等？

**·正态性检验：**

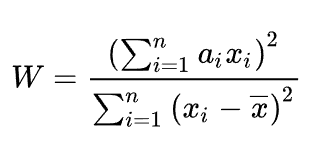
1.假设

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0.01）。

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：

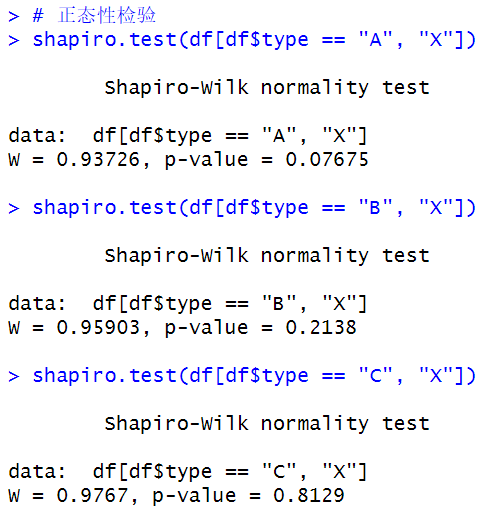
如果p≤α，则拒绝零假设（H0），接受备择假设（H1），认为数据不来自正态分布。

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。



·结果分析

对于TYPE A群体，W值为0.93726，p值为0.07675。

对于TYPE B群体，W值为0.95903，p值为0.2138。

对于TYPE C群体，W值为0.9767，p值为0.8129。

在显著性水平α=0.01的情况下，所有三个群体的p值都大于0.01，因此我们不能拒绝正态分布的假设。这意味着对于每个群体，CEA含量的数据可以被认为是近似正态分布的。

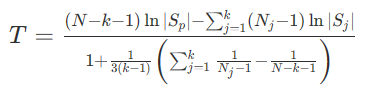
**·方差齐性检验：**

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：

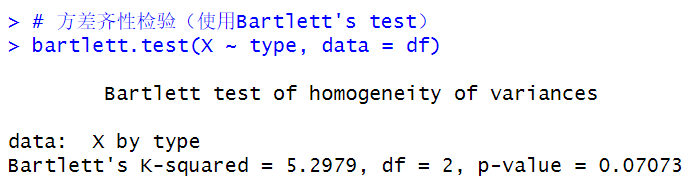
若统计量T落在拒绝域内，则拒绝零假设。

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。



·结果分析：

Bartlett's K-squared值为5.2979，自由度df=2，p值为0.07073。

在显著性水平α=0.01的情况下，p值（0.07073）大于0.01，因此我们不能拒绝方差齐性的假设。这意味着我们可以认为三个群体的CEA含量数据具有相等的方差。

1. 试用方差分析(ANOVA)过程比较这三个群体CEA含量有无显著差异？若有显著差异, 请指出哪些群体间CEA的平均含量有显著差异?

**·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。

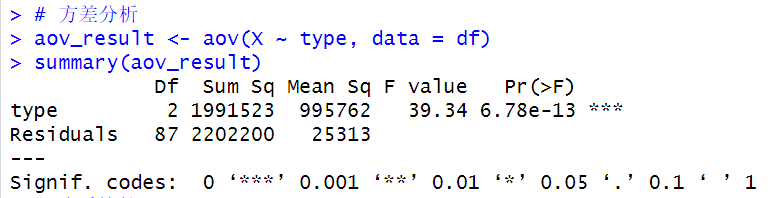
4.决策：

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。



结果分析：

自由度Df为2（因为有三个群体）。总平方和sum sq为1991523。平均平方和Mean Sq为995762。F值为39.34。

p值为6.78e-13，远小于显著性水平α=0.05。由于p值远小于0.05，我们拒绝零假设，即至少有两个群体间CEA含量的平均值存在显著差异。

**·多重比较：**

1.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。

2.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。

3.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（本题为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。

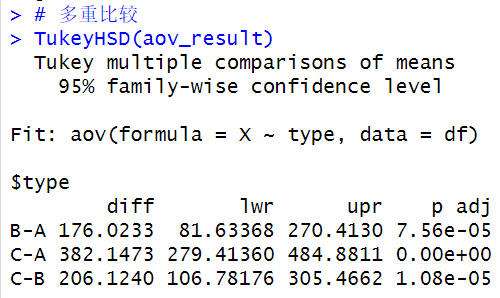
4.决策：

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。

5.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。



·结果分析：

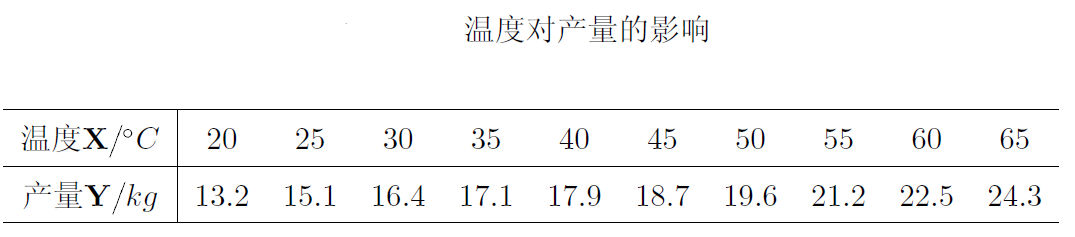
B与A相比，平均差异为176.0233，调整后的p值为7.56e-05，表示B群体的CEA含量显著高于A群体。

C与A相比，平均差异为382.1473，调整后的p值为0.00e+00（实质上是小于检测阈值，可以认为是0），表示C群体的CEA含量显著高于A群体。

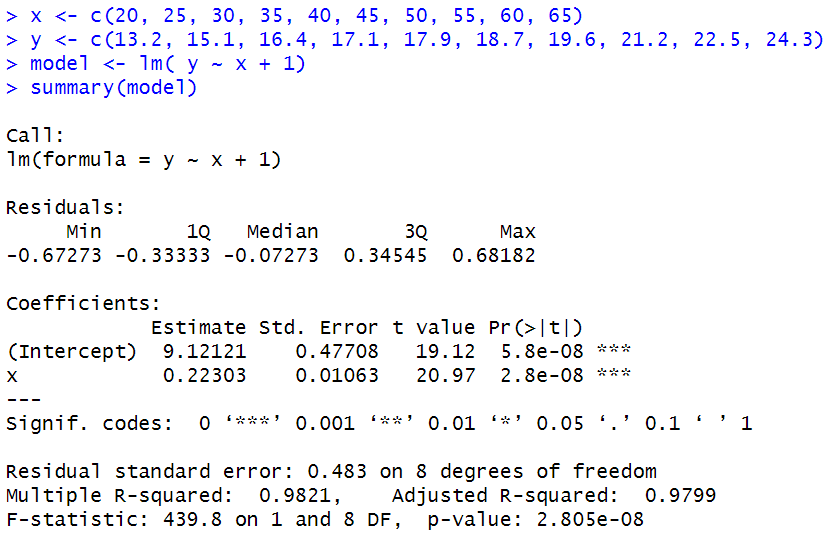
C与B相比，平均差异为206.1240，调整后的p值为1.08e-05，表示C群体的CEA含量显著高于B群体。

因此，三个群体两两之间的CEA含量均有显著差异。

1. 考察温度对产量的影响, 测得10组数据(见下表)



1. 试建立X与Y之间的回归方程；



X与Y之间的回归方程为: **Y=9.12121+0.22303X**

1. 对其回归方程进行显著性检验和回归诊断；

**·显著性检验**

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。

2.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

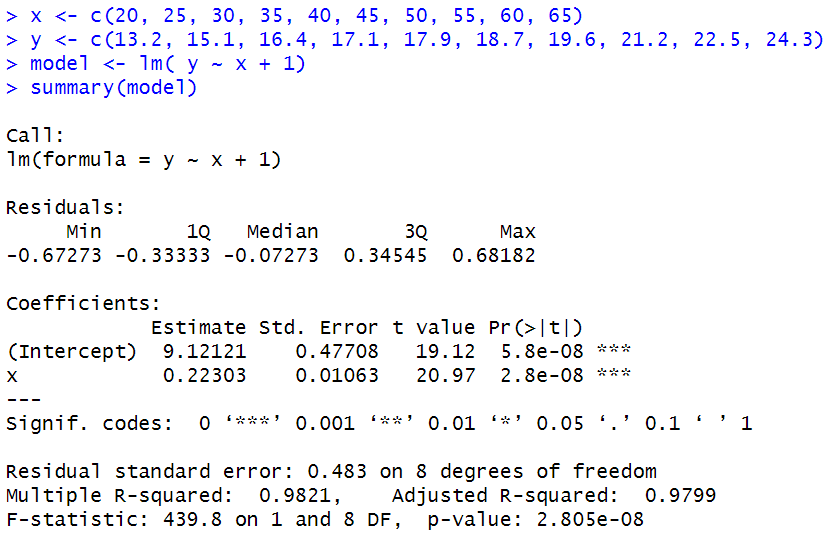
如果计算出的t统计量大于临界值，或者计算出的p值小于显著性水平α，则拒绝零假设.

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。



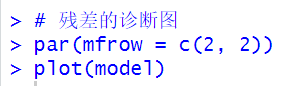
·参数分析与结论：

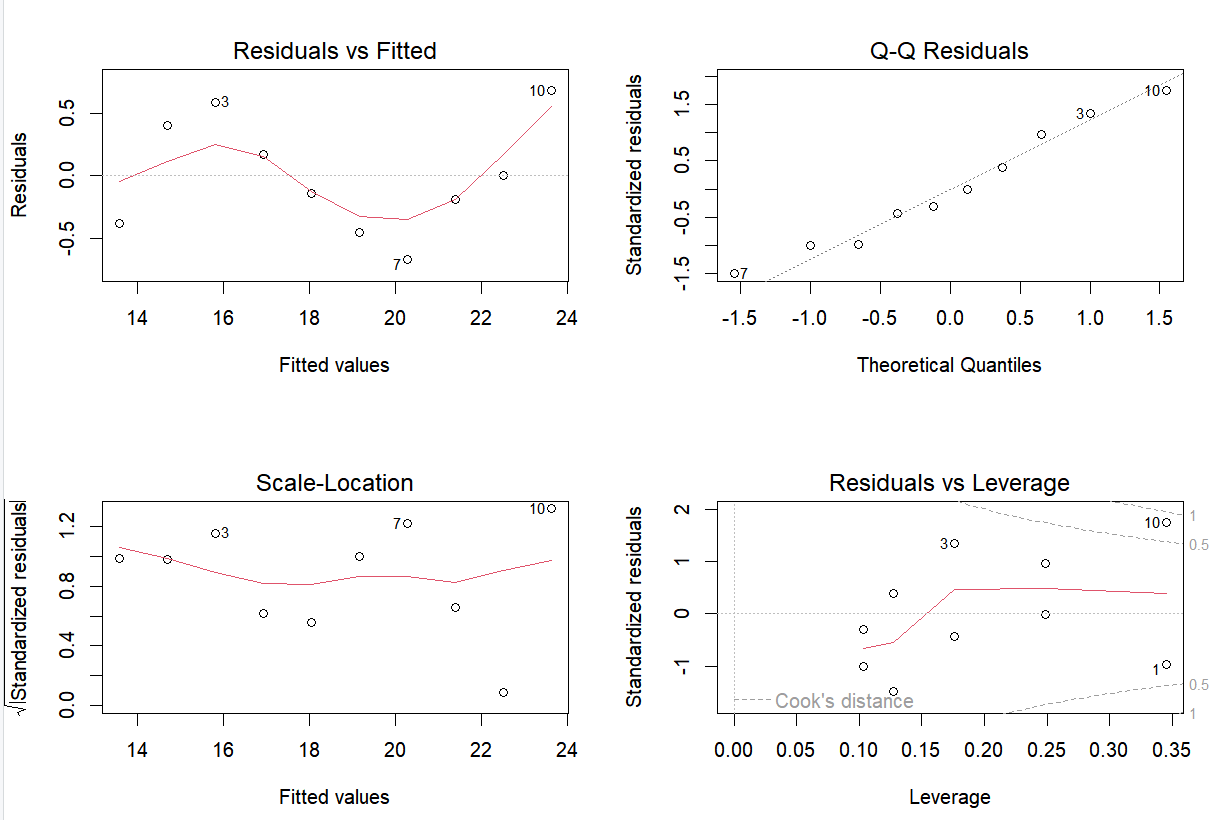
截距（Intercept）估计值为9.12121，标准误差为0.47708，t值为19.12，P值为5.8e-08。截距项在0.05的显著性水平下是显著的，因为p值小于0.05。

X估计值为0.22303，标准误差为0.01063，t值为20.97，P值为2.8e-08，X在0.05的显著性水平下对Y有显著影响，因为p值远小于0.05。

F统计量为439.8，对应的 p 值为 2.805e-08，说明F值落在拒绝域内，故而拒绝原假设（自变量对因变量无显著影响），因此模型整体的显著性非常高。

**·回归诊断**





1.Residuals vs Fitted（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-Q Residuals（残差Q-Q图）

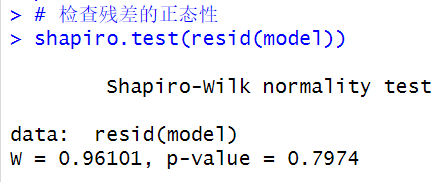
Q-Q图（Quantile-Quantile Plot）用于评估残差的正态性。图中的点近似落在一条直线上，表明残差服从正态分布。

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。

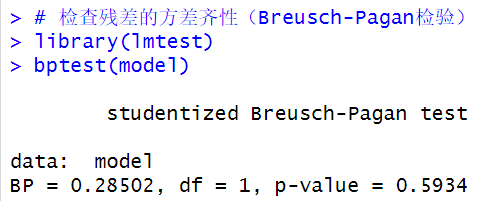
4.Residuals vs Leverage（残差vs杠杆值）

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。图中所第十个点的cook距离大于0.5，说明第十个点是异常点。



·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0.7974，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0.96101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。

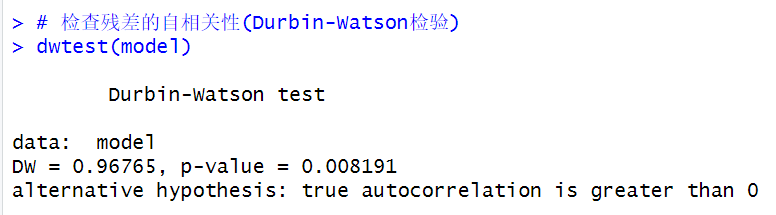


·检验结果分析及结论：

BP值：检验统计量BP值为0.28502，这个值是用于判断残差是否具有恒定方差。

p-value：p值为0.5934，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：

Durbin-Watson统计量 (Dw): 0.96765

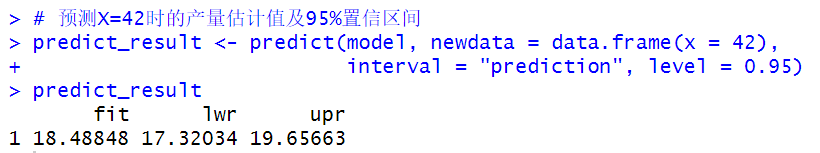
p值: 0.008191

备择假设: 真实的自相关性大于0。

检验结果表明残差的自相关性存在，因为 Dw 值接近1，但 p 值小于0.05，这表明我们有足够证据拒绝原假设，认为残差存在正自相关。

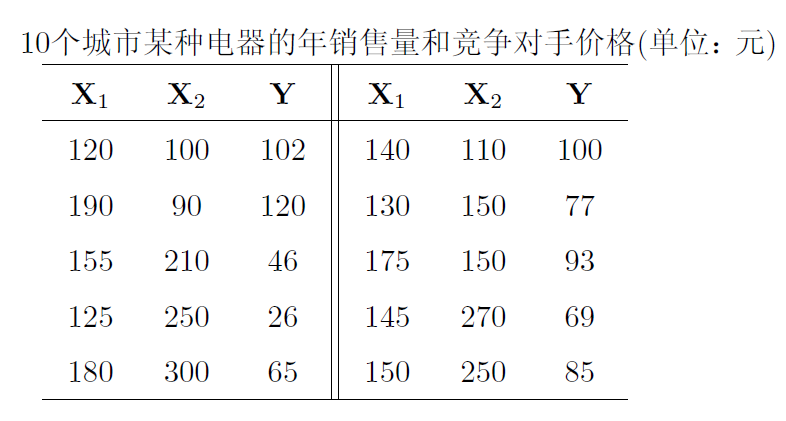
综上所述，通过回归诊断，发现该模型的残差存在正自相关性，并且第十个数据点是异常点。

1. 预测X=42℃时产量的估计值及预测值区间（置信度为95%）。



根据本题所建立的模型，X=42℃时产量的估计值为18.48848，0.95置信度的预测值区间[17.32034,19.65663]。

1. 某厂生产的一种电器的年销售量Y与竞争对手的价格X1及本厂的价格X2有关。下表是10个城市中记录的资料。



1. 建立Y 与X1及X2的回归关系, 并说明回归方程式在 的水平上是否显著？并解释回归系数的含义;

·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。

2.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

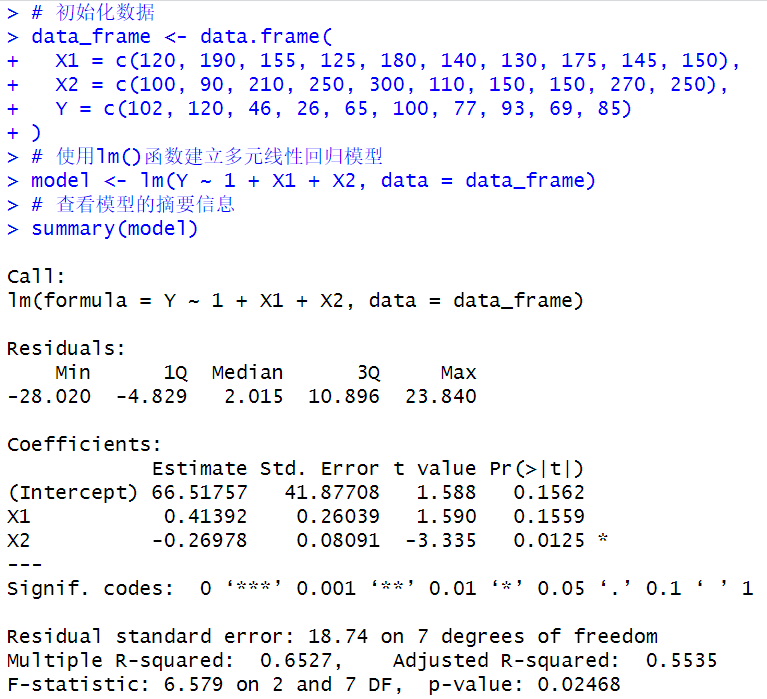
如果计算出的t统计量大于临界值，或者计算出的p值小于显著性水平α，则拒绝零假设.

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。



·回归方程

Y = 66.51757 + 0.41392X1 - 0.26978X2

·模型显著性

模型的F统计量是6.579，对应的p-value是0.02468。由于p-value小于常用的显著性水平0.05，因此我们可以认为整个回归模型在95%的置信水平上是显著的。

·回归系数的含义

截距(Intercept):

66.52，表示当X1和X2都为0时，预测的年销售量是66.52。截距的p-value是0.1562，大于0.05，这意味着截距对Y的影响在95%的置信水平上不显著。

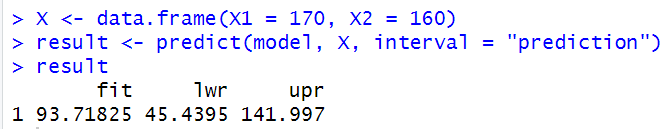
X1的系数(0.41):

表示竞争对手的价格每增加1单位，年销售量平均增加0.41单位。系数的p-value是0.1559，大于0.05，这意味着X1对Y的影响在95%的置信水平上不显著。

X2的系数(-0.27):

表示本厂的价格每增加1单位，年销售量平均减少0.27单位。系数的p-value是0.0125，小于0.05，这意味着X2对Y的影响在95%的置信水平上是显著的。

1. 已知某城市中本厂电器的售价X2为160元, 竞争对手售价X1为170元, 使用上述建立起来的回归模型预测该城市的年销售量;



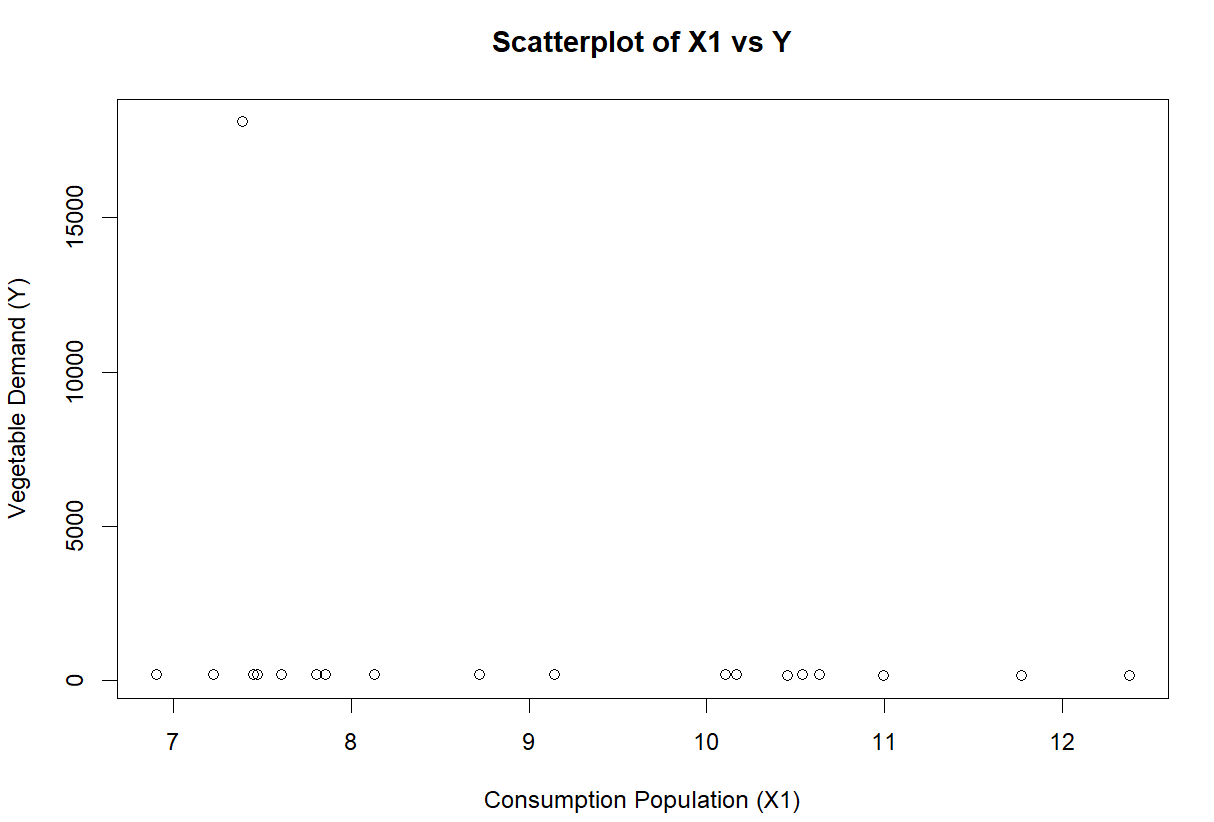
根据本题所建立起来的回归模型，预测出的年销售量的估计值为93.71825，在0.95置信水平下的预测值区间为[45.4395,141.997]。

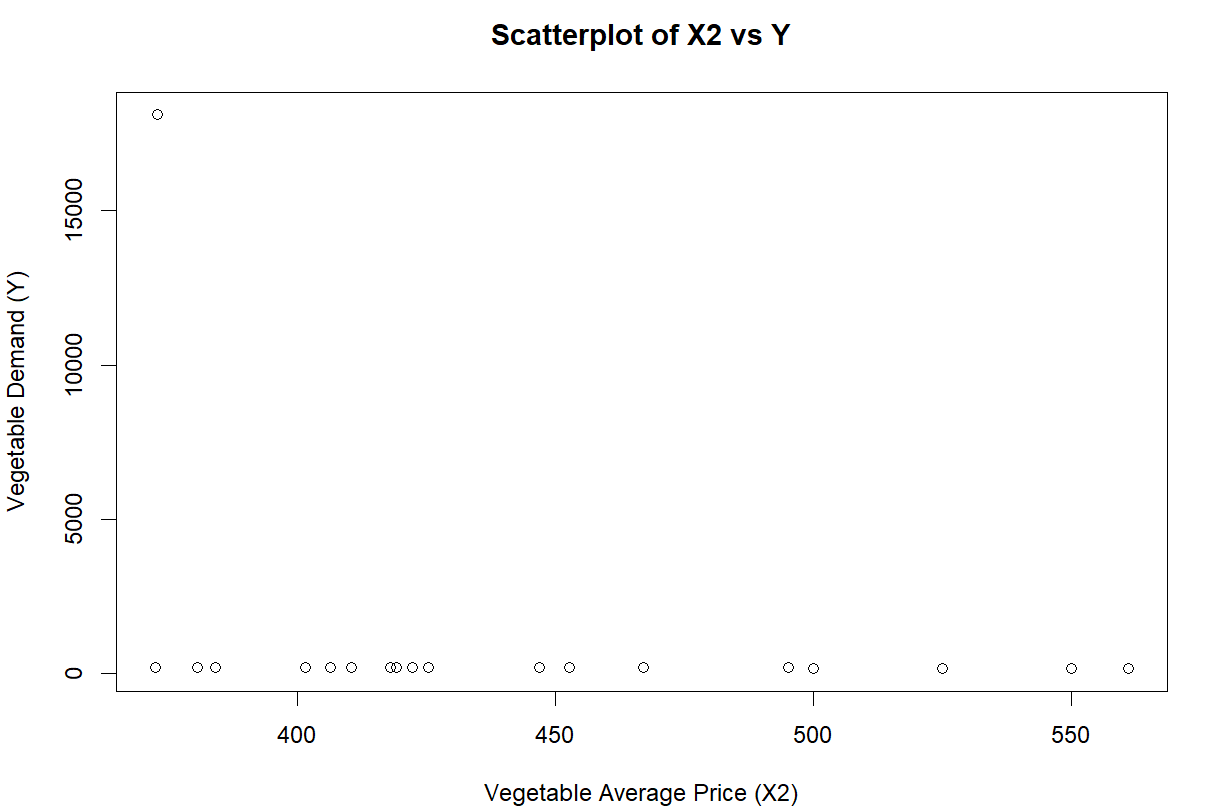
1. 某市消费人口X1(万人)、蔬菜年平均价格X2(分/kg)、瓜果年人均消费量X3(kg)、副食年人均消费量X4(kg)和粮食年人均消费量X5 (kg）是影响蔬菜需求量Y(万吨)的主要因素。调查数据如附件“demand.txt”。试用R语言对蔬菜需求量进行多元回归分析。

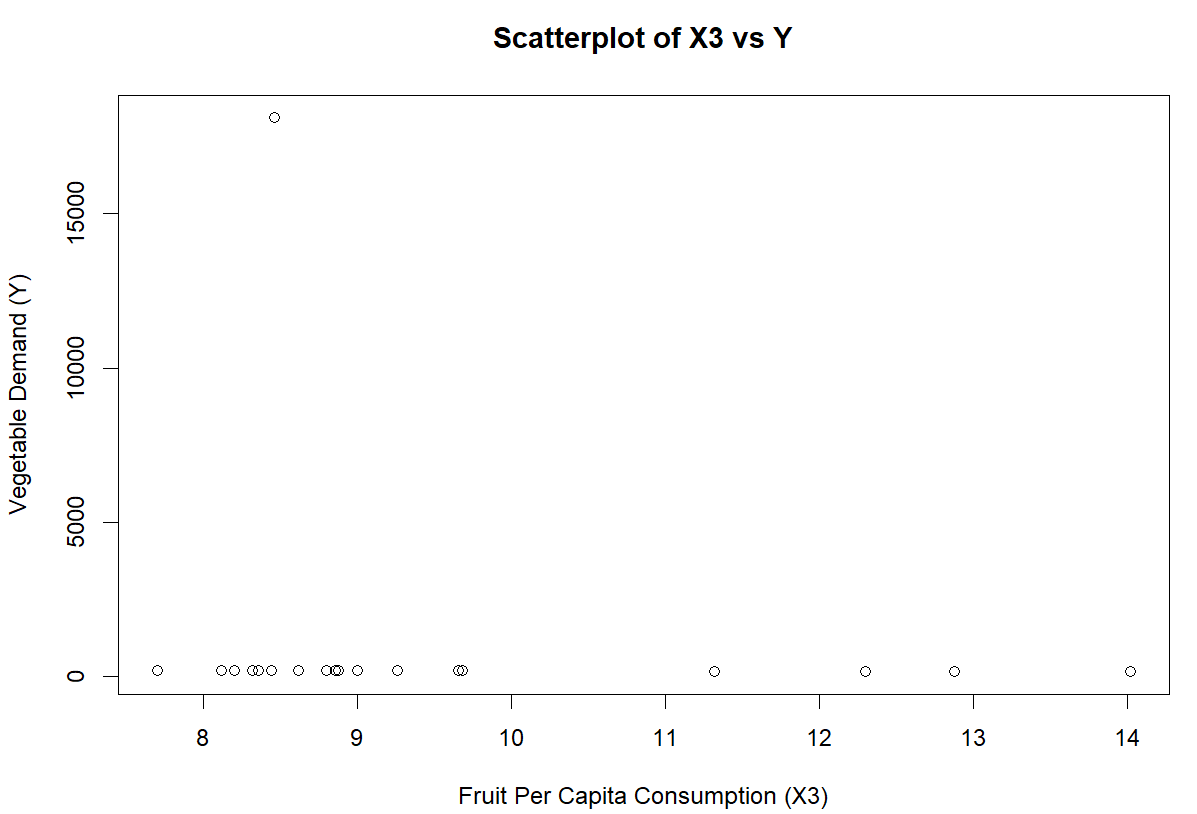
要求：

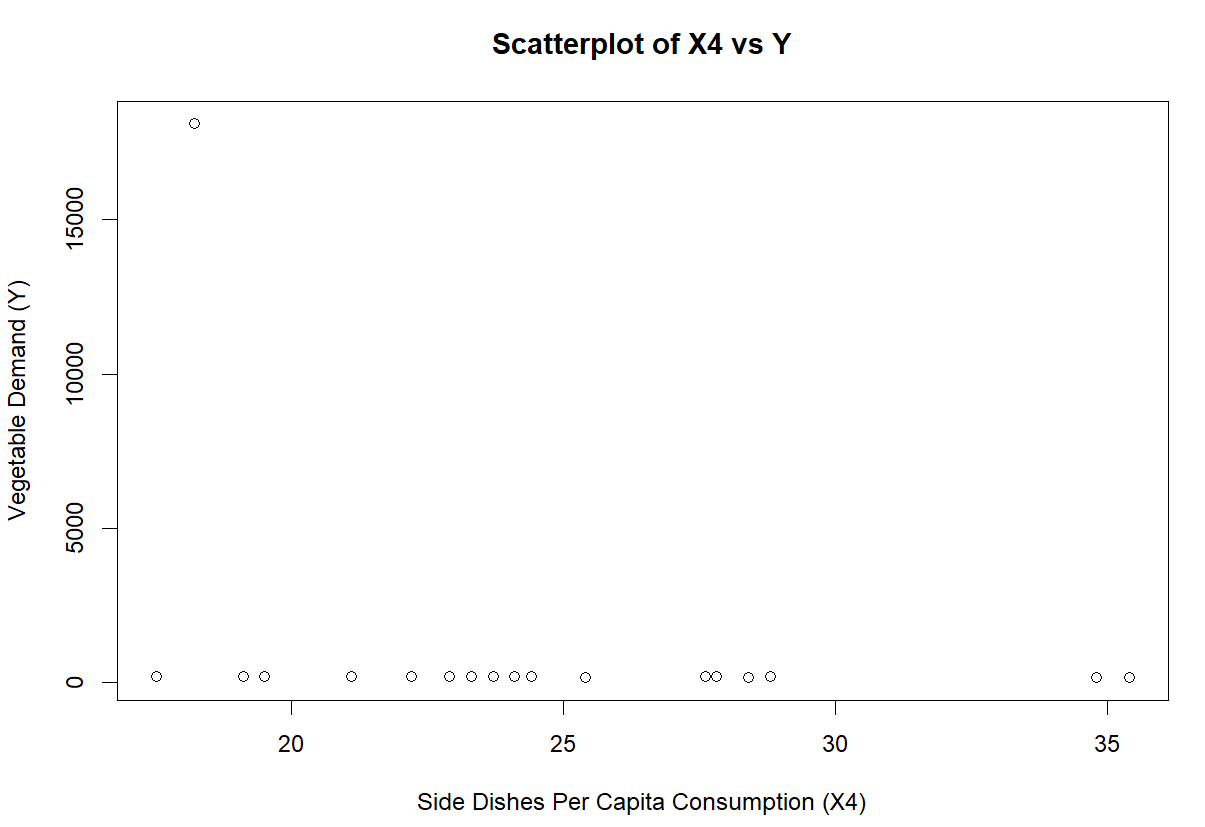
1. 建立数据集, 并画出各自变量与因变量的散点图；

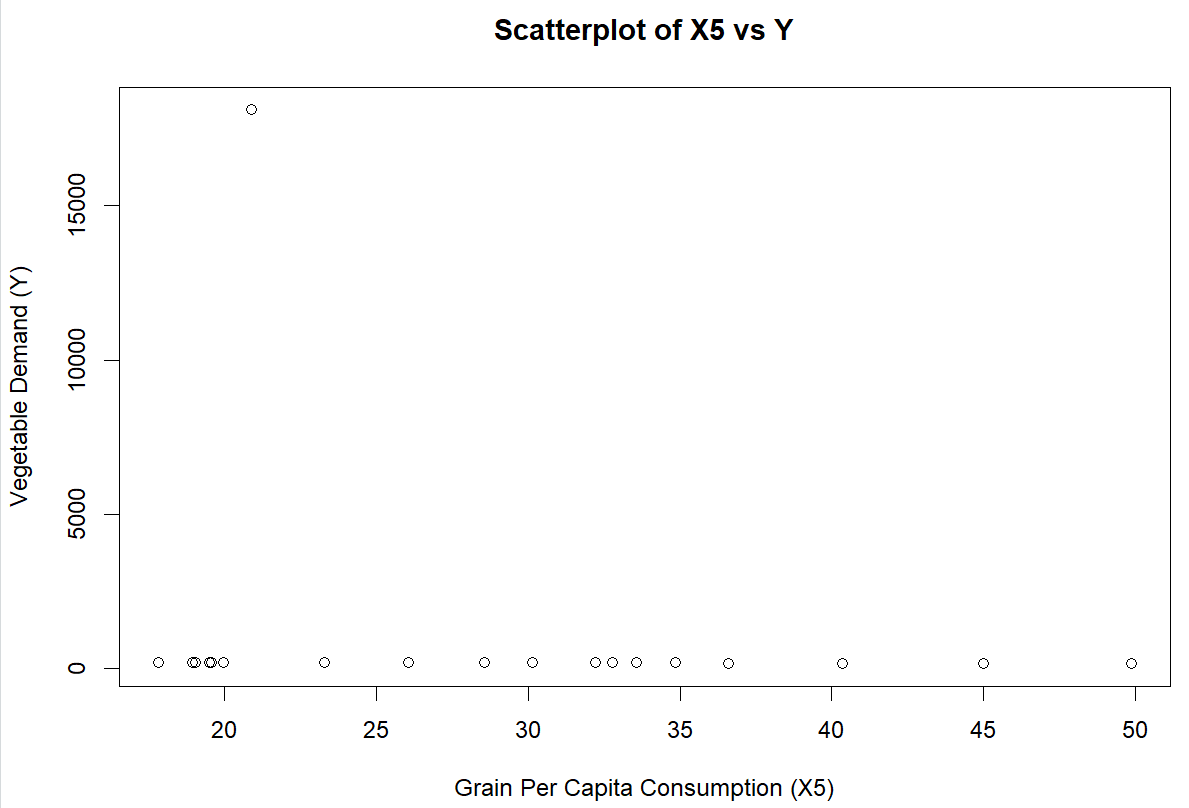












1. 进行线性回归分析, 对其回归方程进行显著性检验和回归诊断（提示：回归诊断包括异常值，多重共线性，残差的独立性、等方差性和正态性检验）。

·显著性检验

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。

2.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

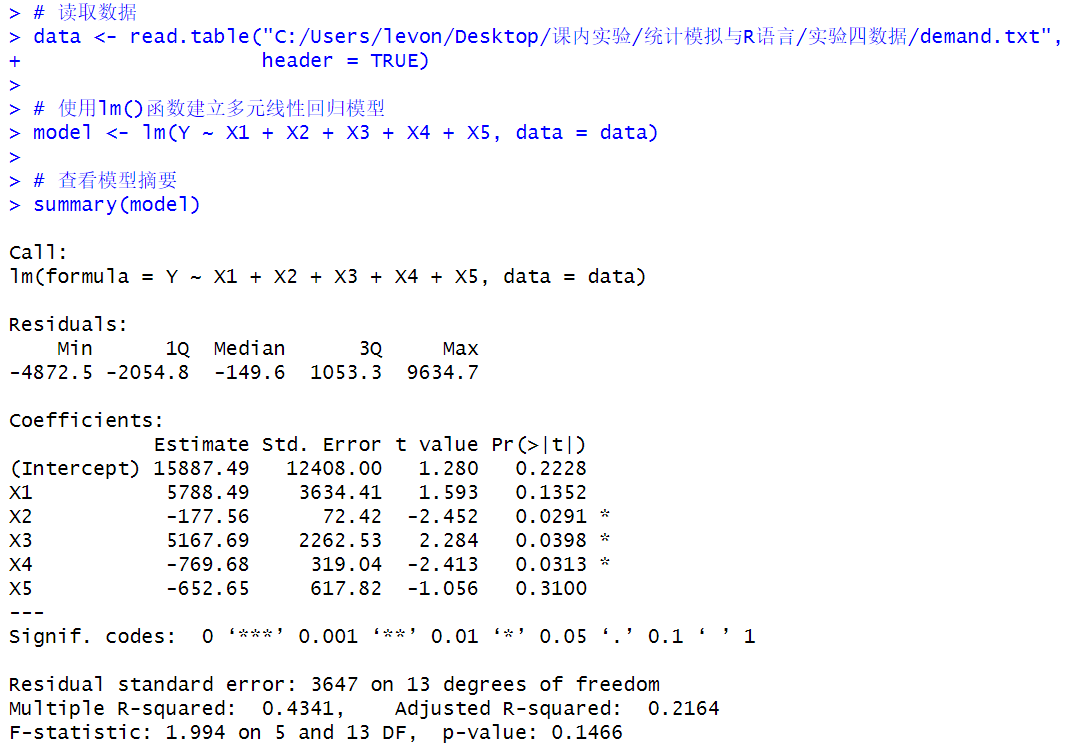
如果计算出的t统计量大于临界值，或者计算出的p值小于显著性水平α，则拒绝零假设.

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。



·参数分析：

截距（Intercept）：15887.49，这是当所有自变量为0时，因变量Y的预测值。标准误差为12408.00，t值为1.280，p值为0.2228，这表明截距项在统计上不显著。

X1：5788.49，标准误差为3634.41，t值为1.593，p值为0.1352，这表明X1对Y的影响在统计上是不显著的。

X2：-177.56，标准误差为72.42，t值为-2.452，p值为0.0291，这表明X2对Y的影响在统计上是显著的，且为负相关。

X3：5167.69，标准误差为2262.53，t值为2.284，p值为0.0398，这表明X3对Y的影响在统计上是显著的，且为正相关。

X4：-769.68，标准误差为319.04，t值为-2.413，p值为0.0313，这表明X4对Y的影响在统计上是显著的，且为负相关。

X5：-652.65，标准误差为617.82，t值为-1.056，p值为0.3100，这表明X5对Y的影响在统计上不显著。

残差标准误差：3647，这是模型预测值与实际值之间差异的标准差。

多重R平方：0.4341，表示模型解释了43.41%的因变量变异。

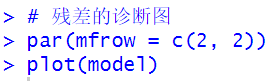
调整后的R平方：0.2164，考虑了自由度的影响后的R平方值，用于更准确地衡量模型的解释能力。

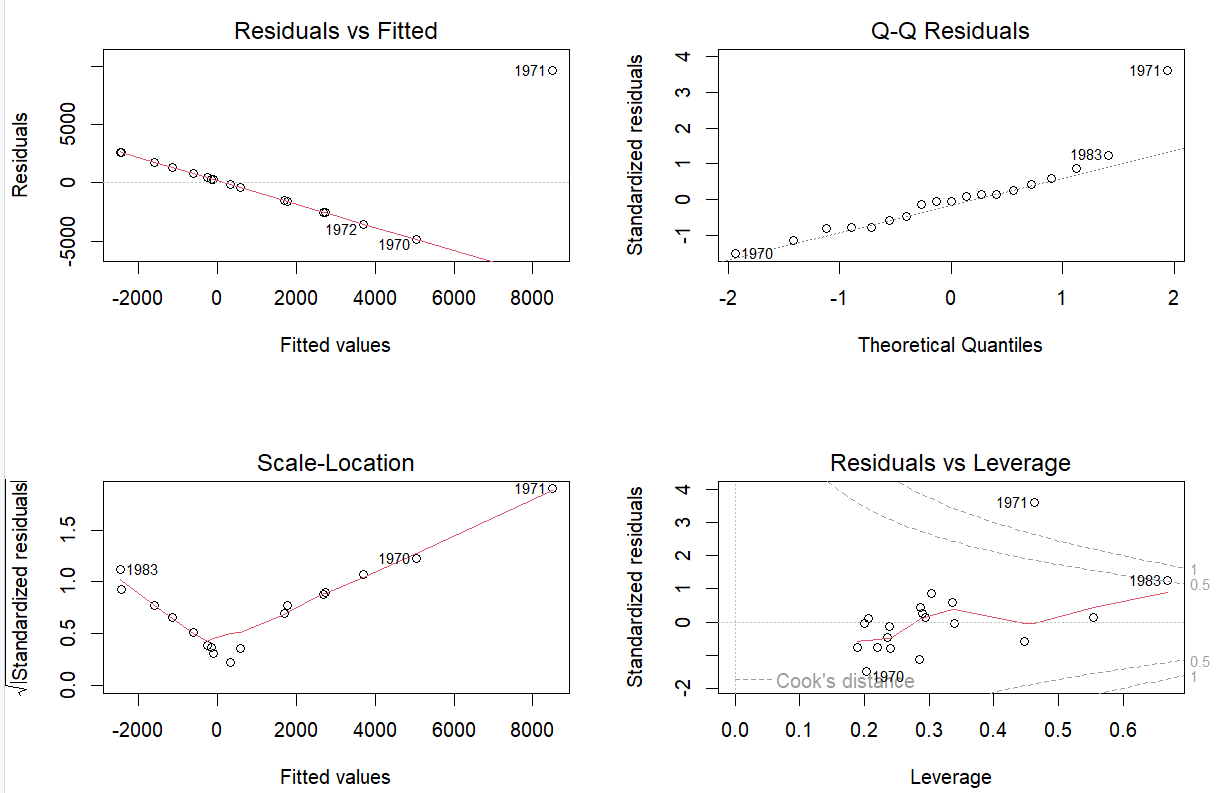
F统计量：1.994，用于检验模型中所有自变量的联合显著性，p值为0.1466，表明模型中所有自变量作为一个整体对因变量的解释能力不显著。

**·结论：**

在这个多元线性回归模型中，X2、X3和X4是显著的，它们对因变量Y有显著影响。X1和X5的系数不显著，可能对模型的贡献较小。模型的解释能力有限，调整后的R平方值较低，表明模型只解释了因变量变异的一小部分。模型的F统计量的P值较高，表明模型中所有自变量作为一个整体对因变量的解释能力不显著。

·回归诊断





1.Residuals vs Fitted（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。此图表明残差并没有均匀地分布在直线两侧，可以初步确定该模型不具有方差齐性。

2.Q-Q Residuals（残差Q-Q图）

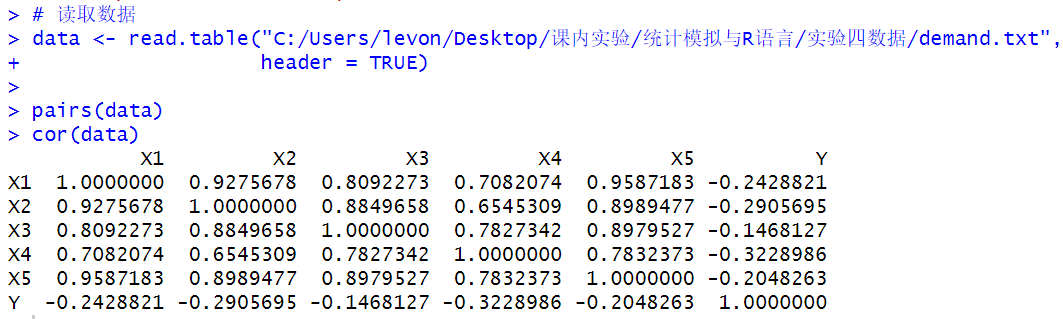
Q-Q图（Quantile - Quantile Plot）用于评估残差的正态性。本图中因为有一个异常点而产生了严重的偏态。

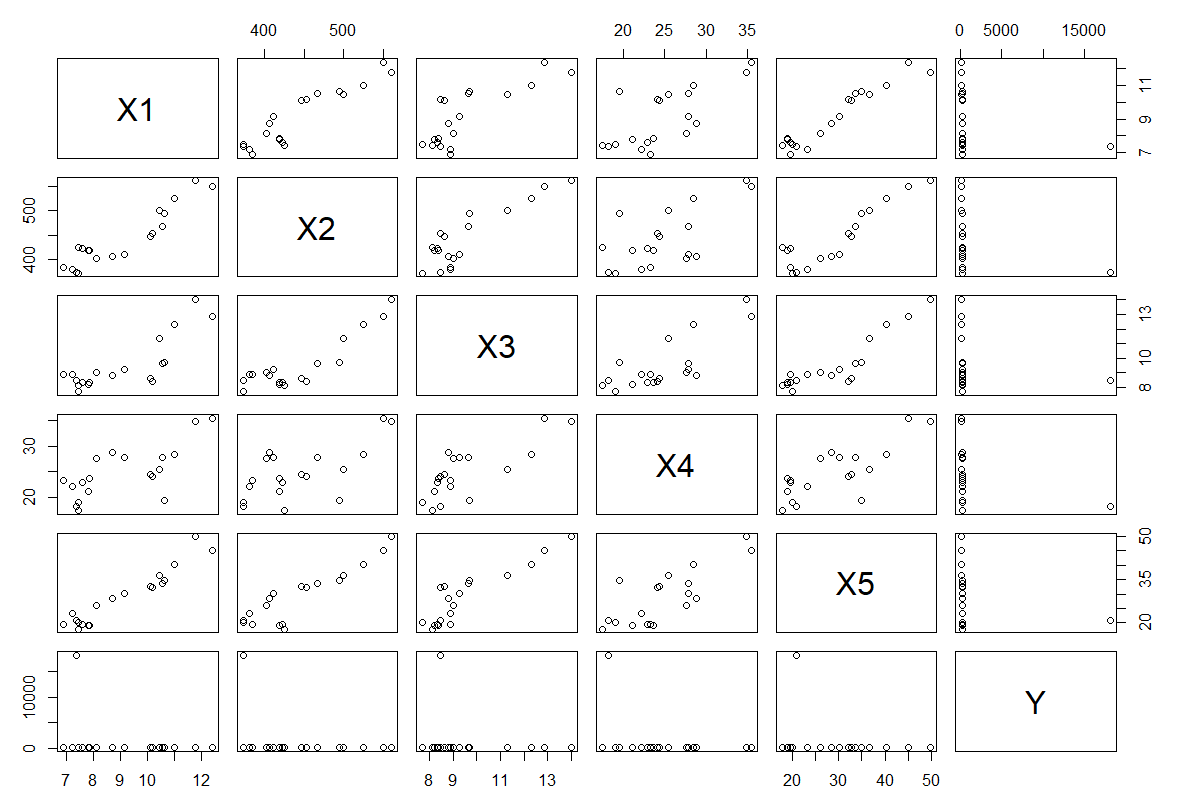
3.Scale-Location（尺度-位置图）

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）有很大关系，呈现出先下降后上升的曲线。这表明模型不具有方差齐性。

4.Residuals vs Leverage（残差vs杠杆值）

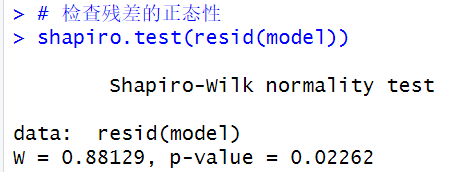
这张图展示了残差与杠杆值（Leverage）之间的关系。杠杆值衡量的是每个观测值对模型拟合的影响程度。理想情况下，我们希望看到残差随机分布，没有与杠杆值相关的模式。而图中1971年数据点的cook距离大于1，1983年数据点的cook距离为0.5，表明它们是异常点。





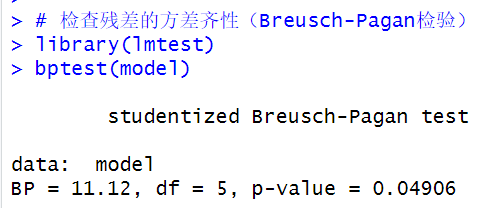
由散点图矩阵和相关系数矩阵可以看出，五个特征两两之间都存在很强的线性相关性，因此该模型具有多重共线性问题。

下面通过一些检验方法进一步诊断：



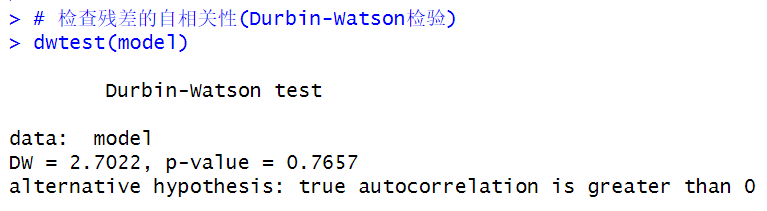
检验结果：Shapiro-Wilk检验的W值为0.88129，p值为0.02262。

分析：由于p值小于0.05，我们拒绝原假设（残差来自正态分布），这表明残差可能不服从正态分布。这可能影响模型的有效性和参数估计的可靠性。



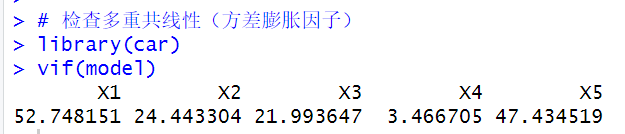
检验结果：Breusch-Pagan检验的BP值为11.12，自由度为5，p值为0.04906。

分析：p值接近0.05，表明我们不能强烈拒绝原假设（方差是齐的）。然而，接近显著性水平的结果提示我们可能存在异方差性，即误差项的方差可能不恒定。



检验结果：Durbin-Watson检验的Dw值为2.7022，p值为0.7657。

分析：Durbin-Watson检验的Dw值范围在0到4之间，其中2表示没有自相关性。Dw值接近2.7，表明残差之间可能不存在显著的自相关性。然而，p值大于0.05，我们不能拒绝原假设（不存在自相关性），但这个结果并不强烈支持这一点。

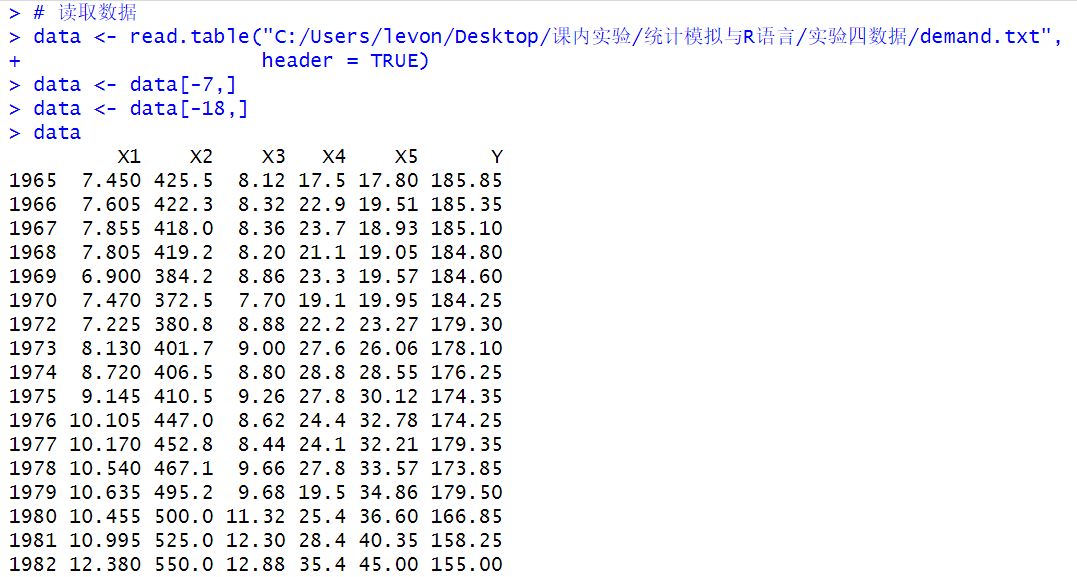


检验结果：VIF值分别为52.748151（X1）、24.443304（X2）、21.993647（X3）、3.466705（X4）、47.434519（X5）。

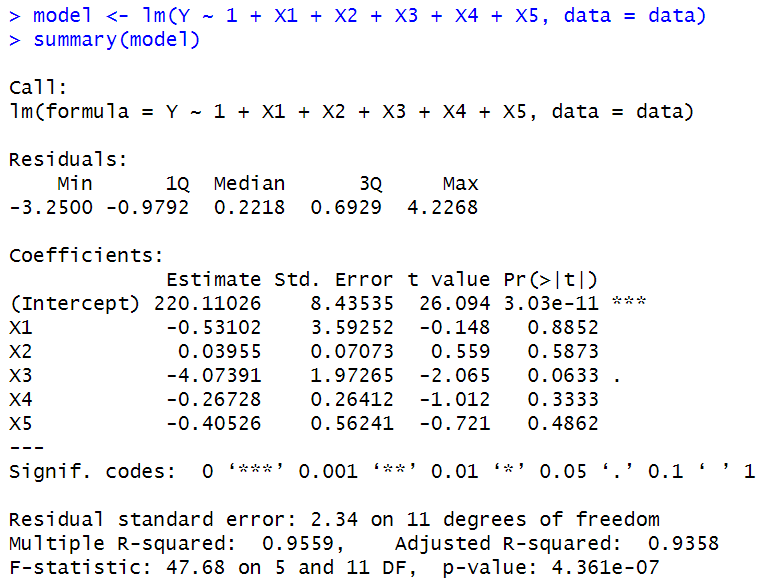
分析：通常，VIF值大于5或10被认为是多重共线性的一个警示信号。根据这个标准，X1、X2、X3和X5显示出较强的多重共线性，而X4的VIF值接近4，相对较低。多重共线性可能导致回归系数的估计不准确和不稳定。

1. 如果有回归诊断存在问题，做相应处理后找到最优回归模型（提示：可能包括异常值的去除和变量的删减）。

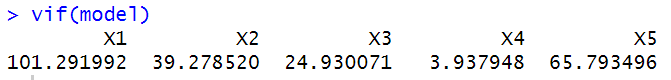
首先去除前面的分析中所发现的异常值，即1971年和1983年的数据。



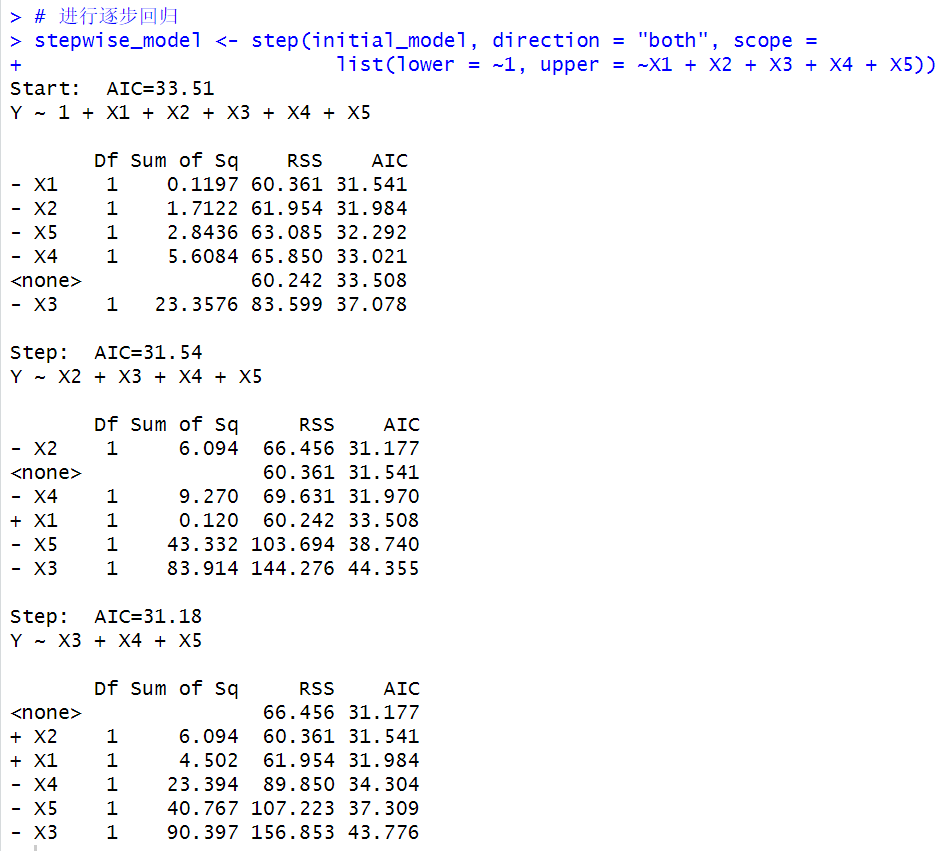
然后对模型进行显著性检验:



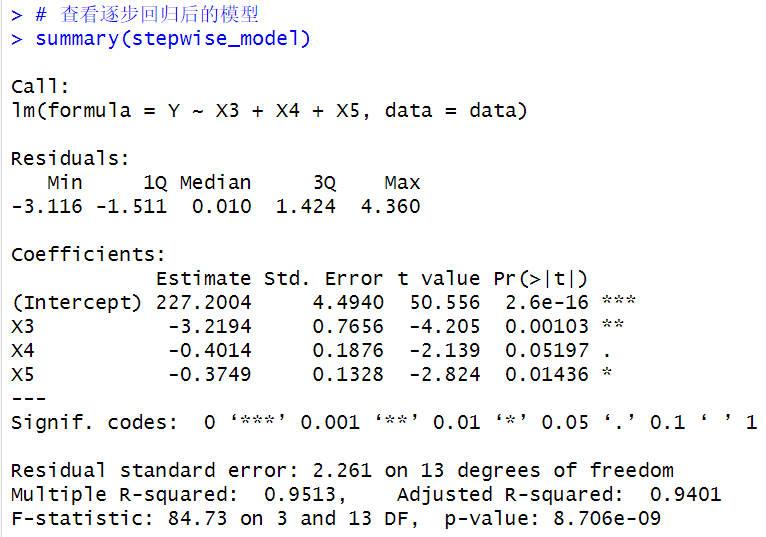
虽然模型整体变得显著了，但是绝大多数的自变量与Y之间的线性关系并不显著，然后再次查看方差膨胀因子：



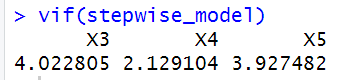
发现多重共线性的问题还是很严重，因此考虑逐步回归。



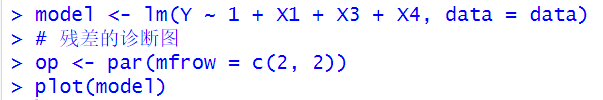
查看完成逐步回归后的模型以及其显著性检验的信息：

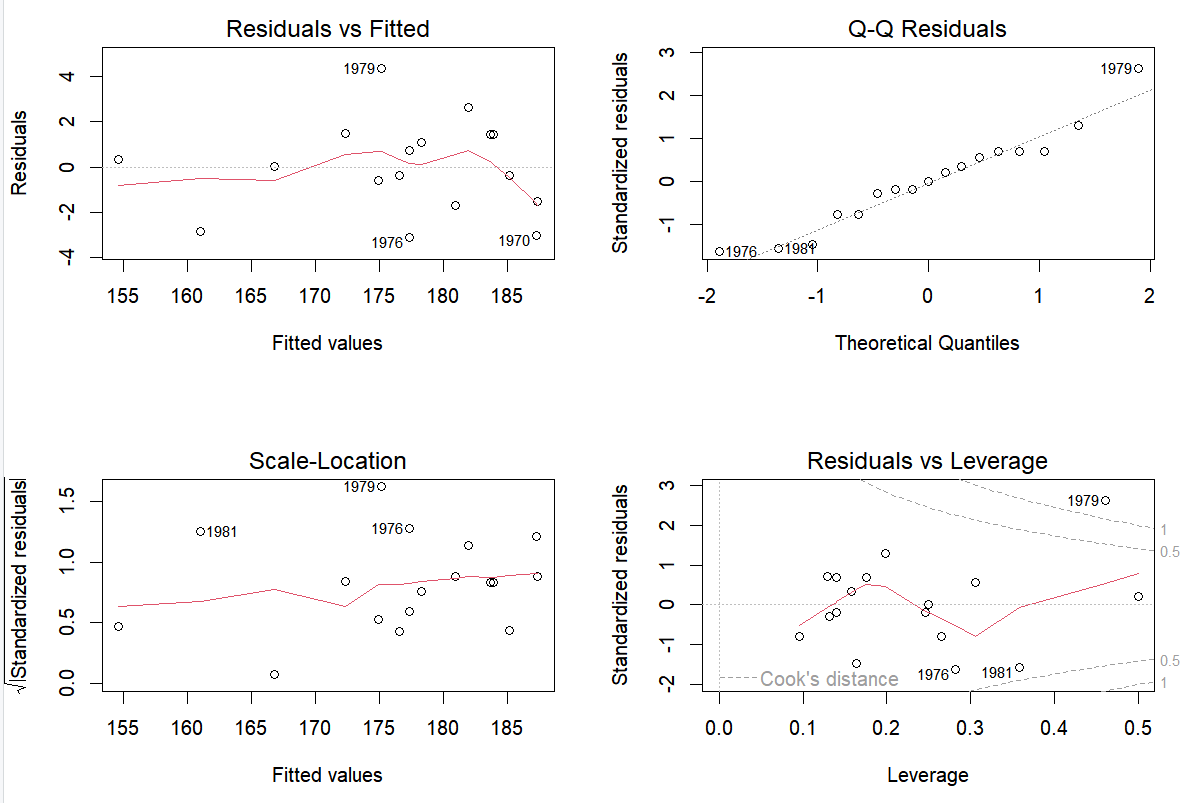


发现自变量和因变量的线性关系变得强烈了一些，并且模型整体也是显著的，再次查看方差膨胀因子：

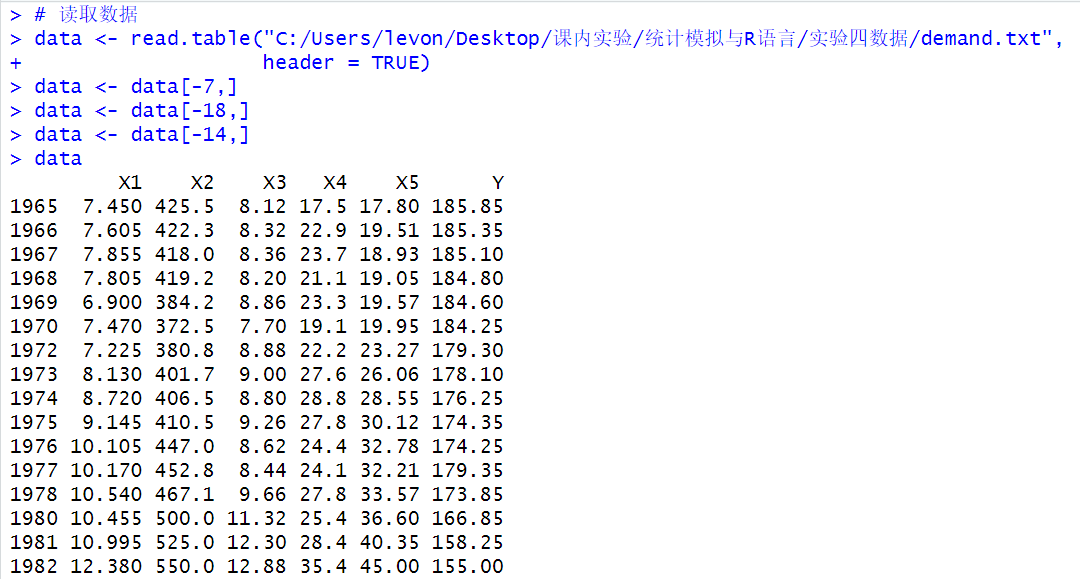


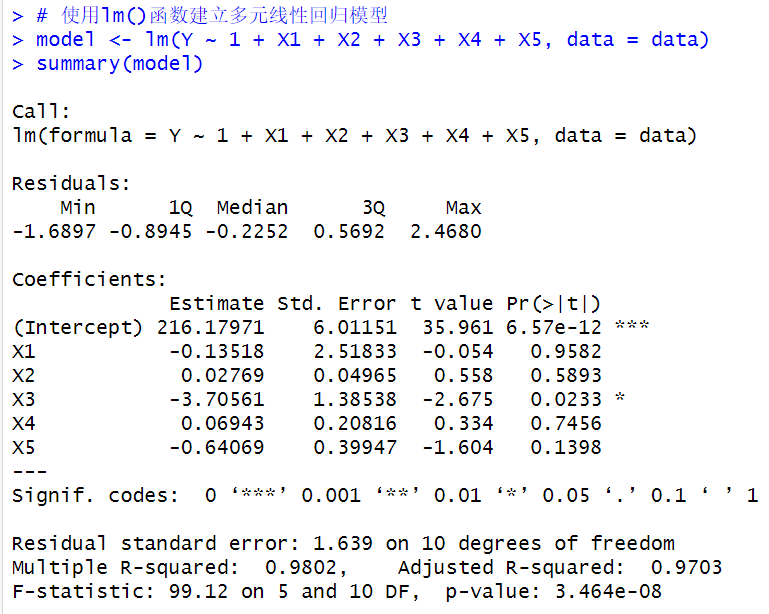
此时三个自变量的方差膨胀因子都小于5，可以认为多重共线性的问题得到了解决。下面再对该模型进行其他方面的诊断：

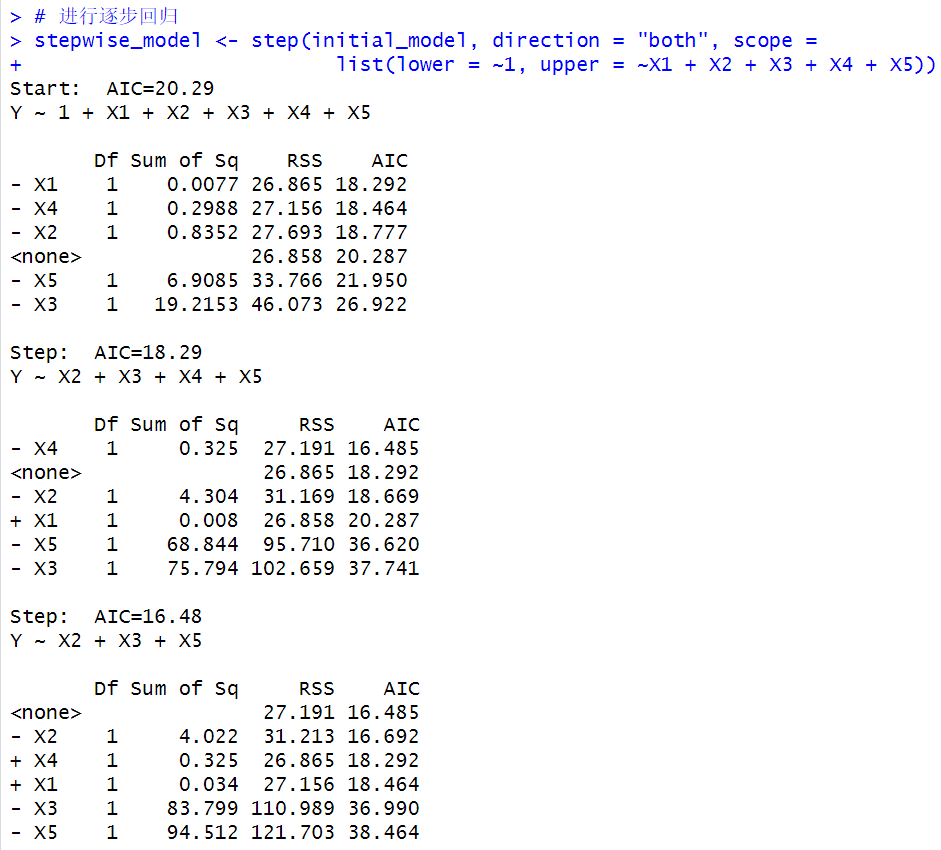


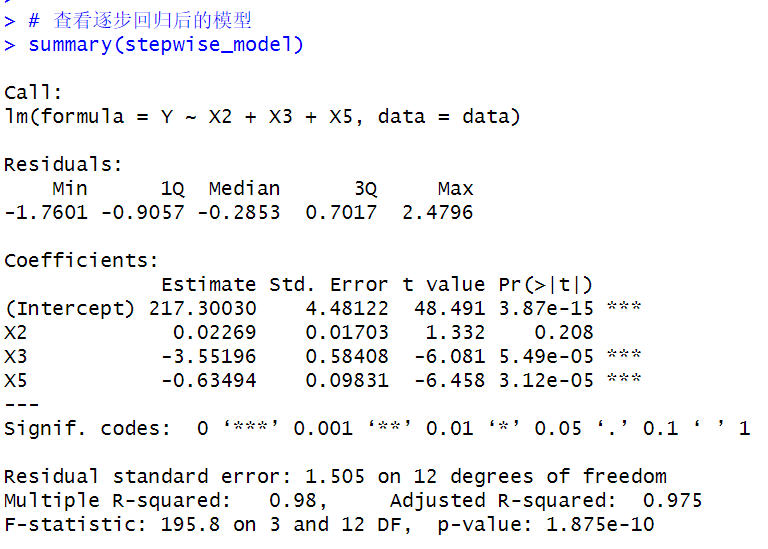


通过图一和图三可以初步判断残差具有方差齐性，通过图二可以初步判断残差分布具有正态性，但是图四中显示了新的异常点，并且cook距离大于1是很严重的异常点，因此考虑再次删除这个点，并重新进行逐步回归。

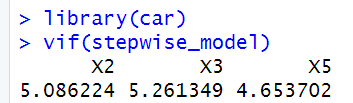




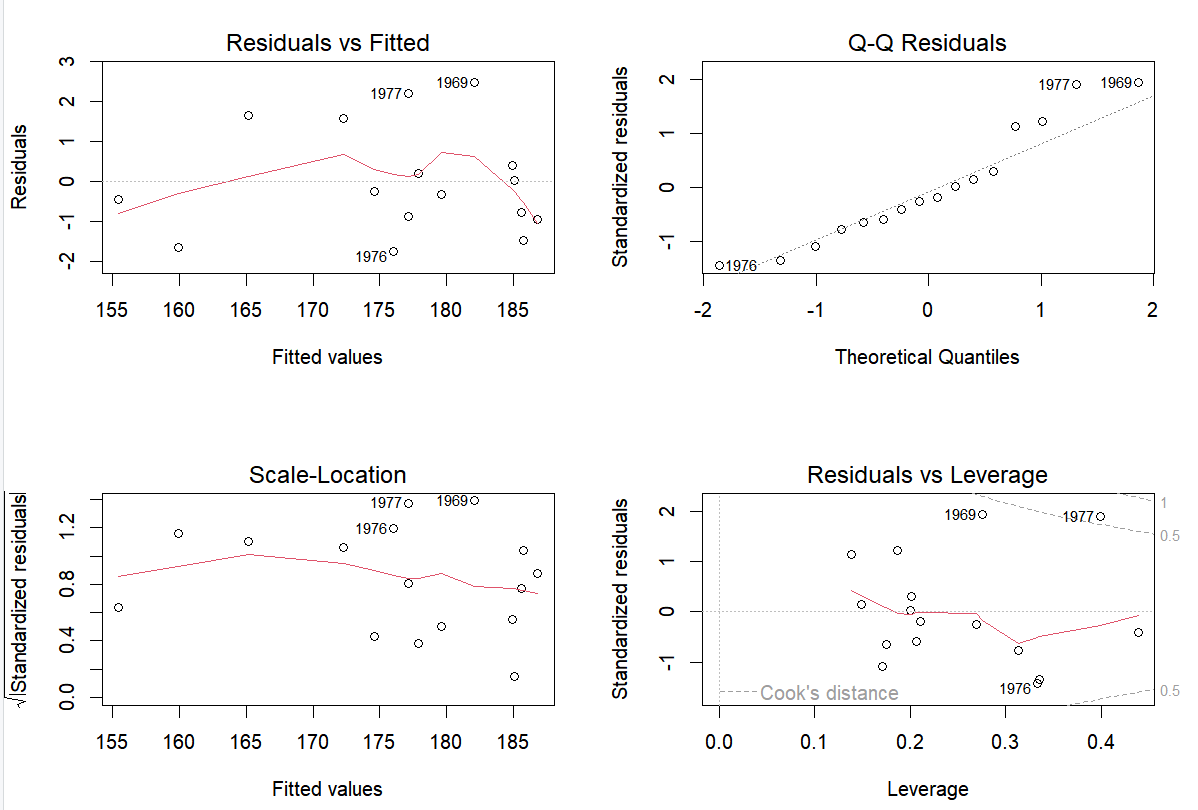




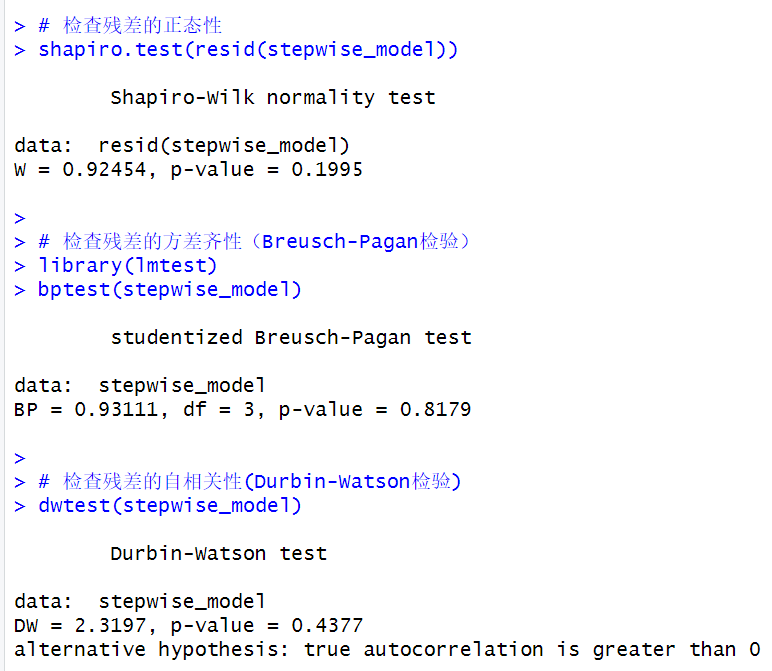
查看膨胀因子：



发现同样没有严重的多重共线性问题，因此进行进一步的残差诊断。



虽然1977年的数据的cook距离略大于0.5，不过相较于前面的两次建模中出现的cook距离大于1的异常点来说，已经可以不认为它是一个异常点了。再进行其它方面的检验：



发现残差满足正态性和方差齐性，并且不存在显著的自相关性。因此认为此模型为最优回归模型，即：

**Y** = 217.30030 + 0.022269 **X2** - 3.55196**X3** - 0.63494**X5**