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。

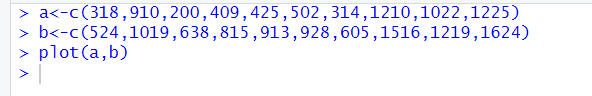
【实验内容与实现】

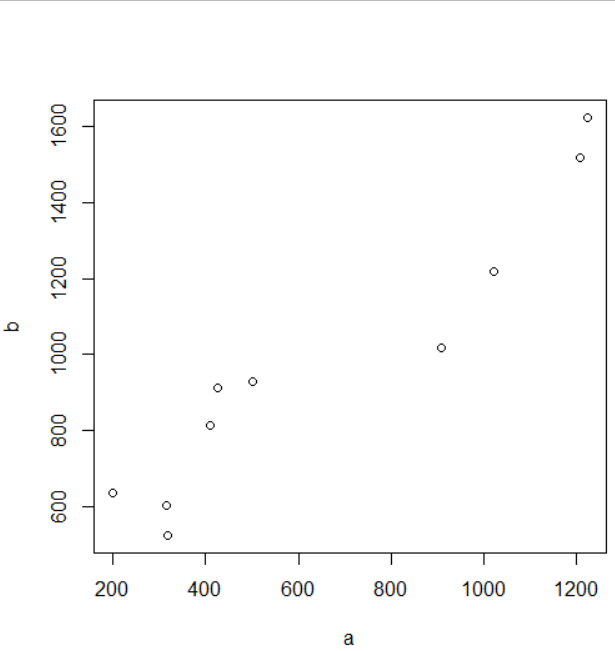
1. 有10个同类企业的生产性固定资产价值(X)和工业总产值(Y )资料如下:



问题如下：

1. 探索性分析X与Y之间的关系

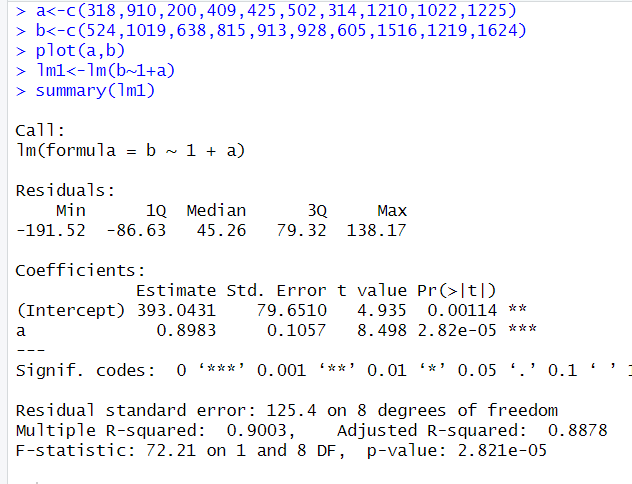




对数据进行探索性分析，从散点图中可以看出a,b之间存在着一定的线性关系。

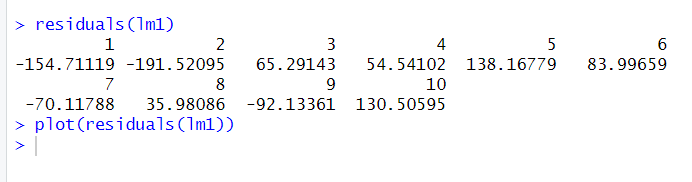
1. 建立X与Y之间的线性回归方程，并对其回归方程进行显著性检验和回归诊断。

对X,Y建立回归关系R实现如下：

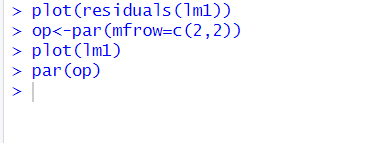


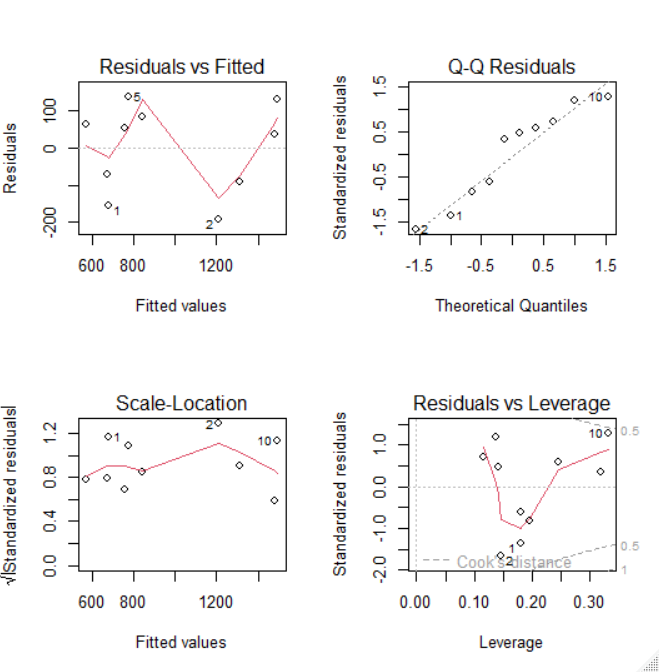
由图中的数据可以得出，P-value为2.821\*10-5远小于0.05故a，b之间的线性关系较为显著。

回归诊断：



残差值分布相对均匀。





从图1看出残差均匀分布在直线两侧，未发现明显异常。

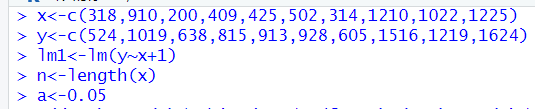
从图2看出数据分布在45度对角线周围，基本服从正态分布。

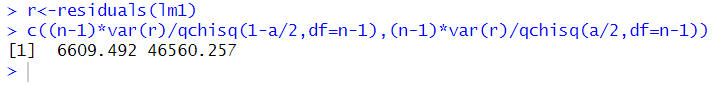
从图3可以看出，方差基本相同，具有方差齐性。

从图4可以看出，不存在异常点对回归结果造成严重影响。

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。

由题可知：





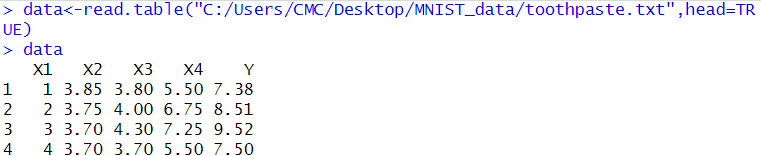
最终求得的置信区间为6609.492到46560.257。

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：

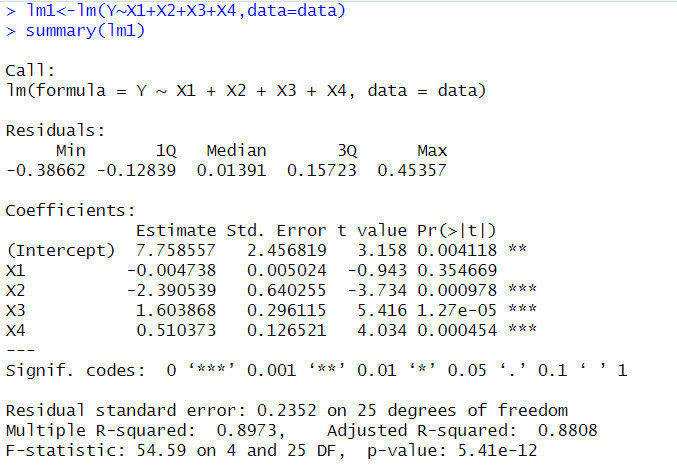
1. 建立牙膏销量与其他因素的回归模型；

首先读入相关数据：



接下来对数据进行线性回归。

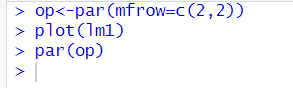
回归得到的相关结果如下：

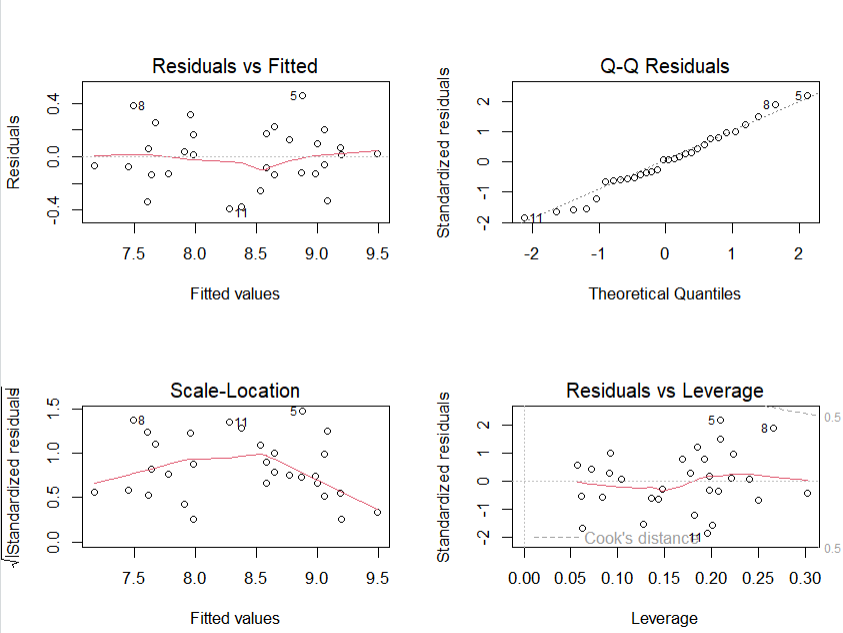


从图中可以看出，除X2的P-value值大于0.05，与Y之间的线性关系不显著，其他变量的P-value值均小于0.05表明与Y之间的线性相关关系比较显著。

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。

进行线性回归诊断如下：





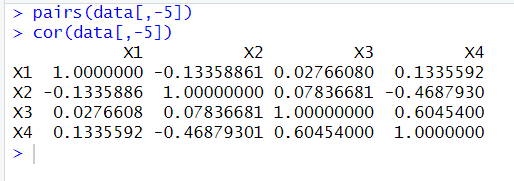
从图1看出残差均匀分布在直线两侧，未发现明显异常。

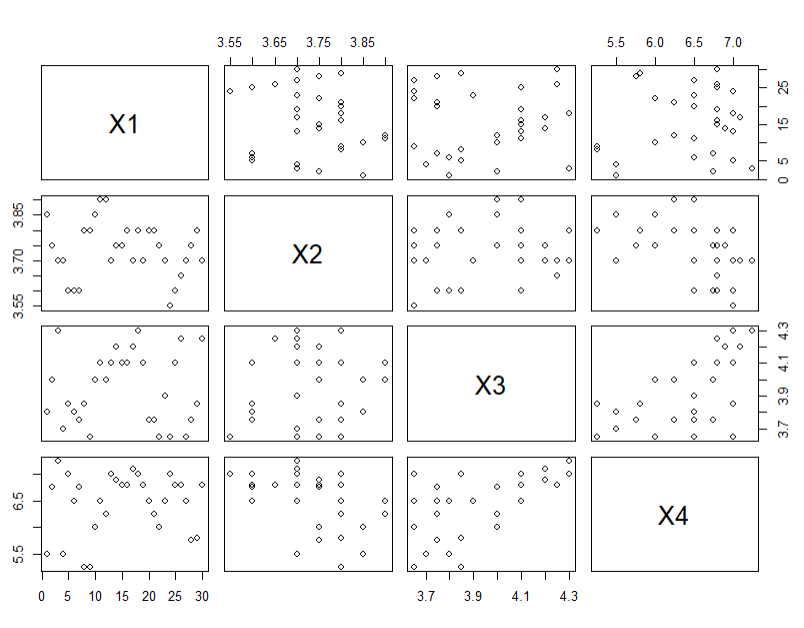
从图2看出数据分布在45度对角线周围，基本服从正态分布。

从图3可以看出，方差基本相同，具有方差齐性。

从图4可以看出，不存在异常点对回归结果造成严重影响。

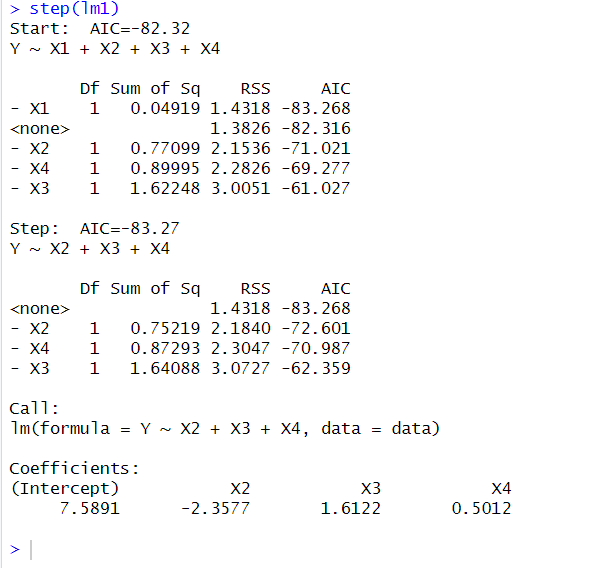
接着对样本进行分析是否存在多重共线性：





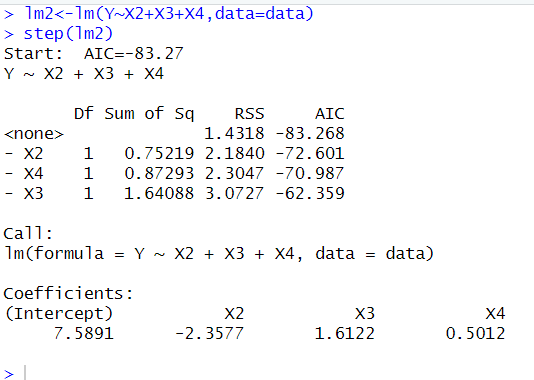
从图中可以看出，X3和X4之间存在着近似的线性关系，因此认为该样本数据间存在多重共线性。

采用逐步回归的方法，探索出最合适的线性回归模型：

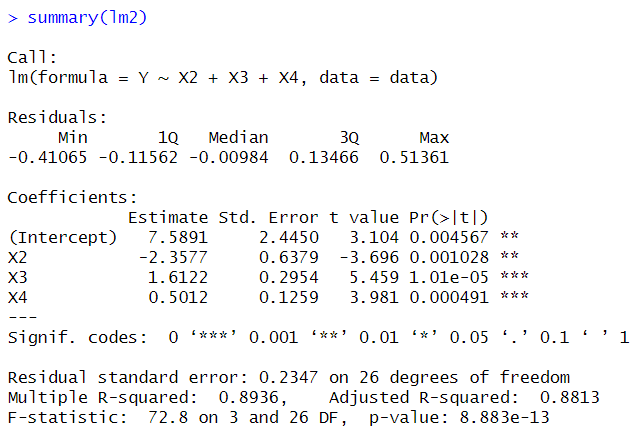


由此可知，当X1去除时线性模型的AIC值更小，模型更优。

因此删去X1属性，再次对数据进行回归操作。

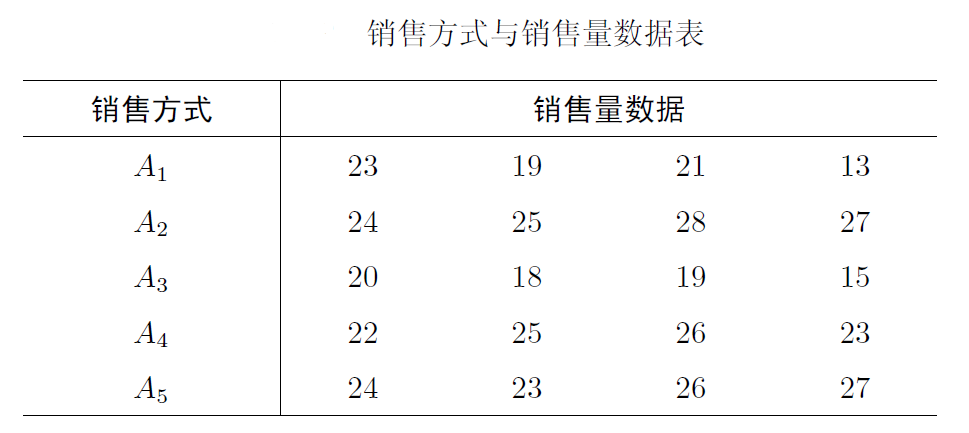


再次回归后得到了最优结果

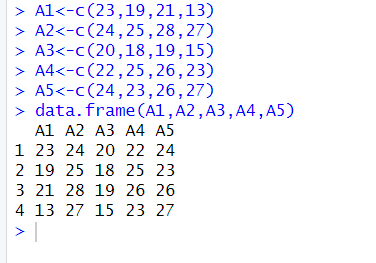


可以看出X2,X3,X4与Y之间的相关性都比较显著。

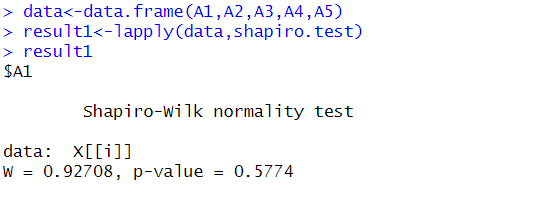
1. 某商店以各自的销售方式卖出新型手表，连续四天手表的销售量如表所示，试考察销售方式之间是否有显著差异。（提示：方差分析的步骤包括1.正态性检验，2.方差齐性检验，3.方差分析，4.多重比较）

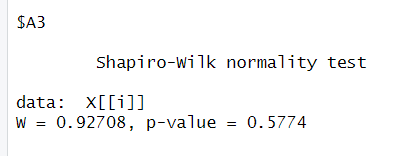
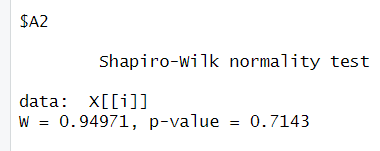


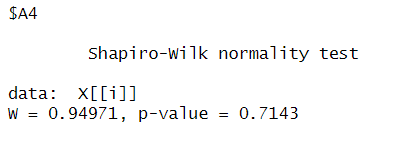
首先输入相关数据：

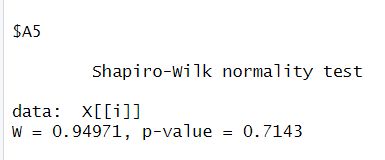


下面进行正态性检验：



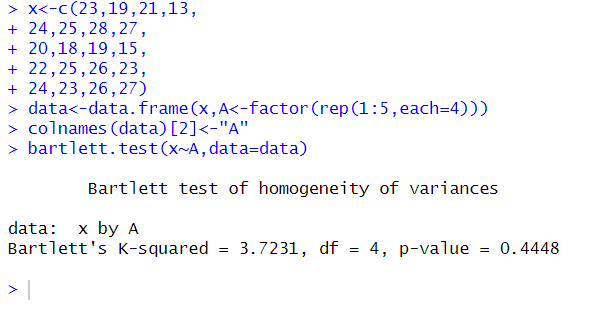






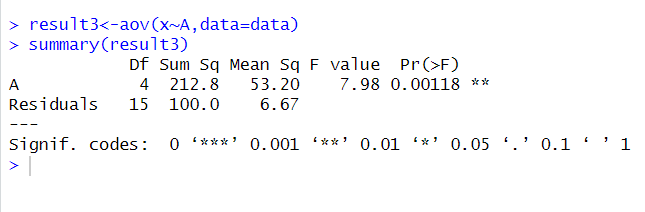
由以上数据P-value均大于0.05，因此数据满足正态性。

下面进行方差齐性检验：



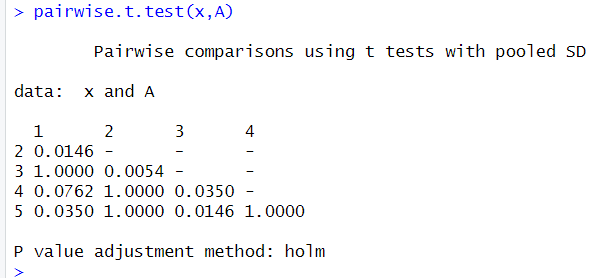
由此结果可知P-value大于0.05，因此数据满足方差齐性。

下面进行方差分析：



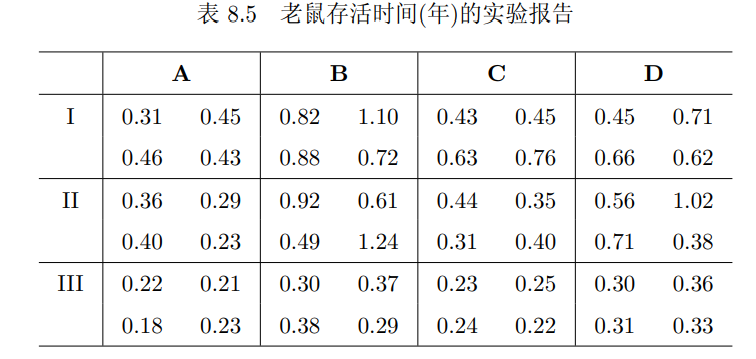
由P-value小于0.05，因此认为差异存在显著性，故销售方式之间存在着显著性差异。

下面进行多重性比较：

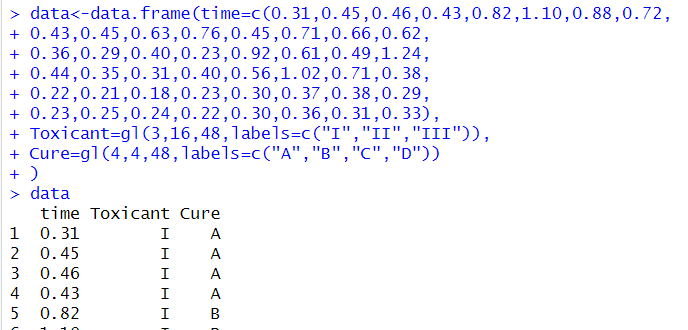


由此可以看出1和2，1和5，2和3，3和4，3和5之间的P值小于0.05存在显著性差异。

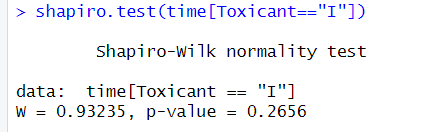
1. 有一个关于检验毒品强弱的试验, 给48只老鼠注射I、II、III三种毒药(因素A), 同时有A、B、C、D 4种治疗方案(因素B), 这样的试验在每一种因素组合下都重复四次测试老鼠的存活时间, 数据如下表所示. 试分析毒药和治疗方案以及它们的交互作用对老鼠存活时间有无显著影响.

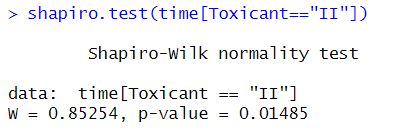


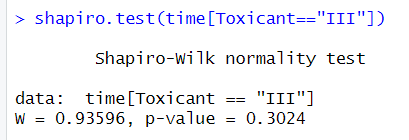
首先完成数据输入：

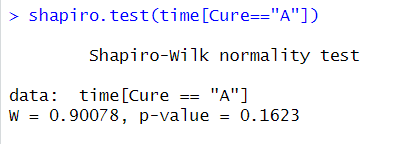


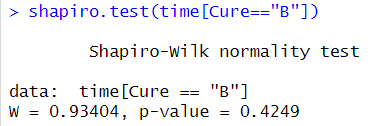
分别进行正态性检验和方差齐性检验：

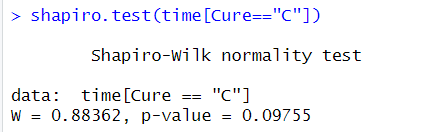


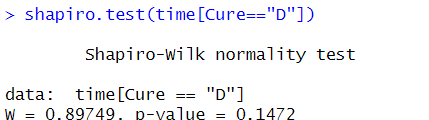






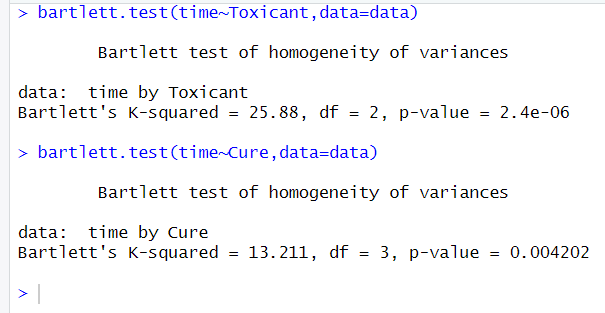






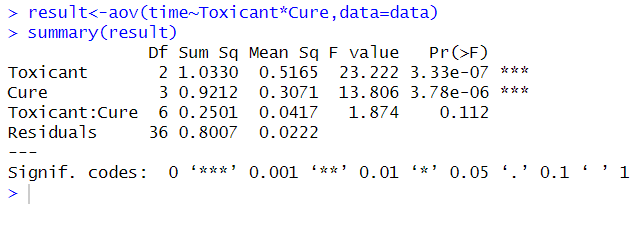
以上P-value存在小于0.05的情况，因此数据并不完全满足正态性。

方差齐性检验：



由此看出P-value均小于0.05，不满足方差齐性。

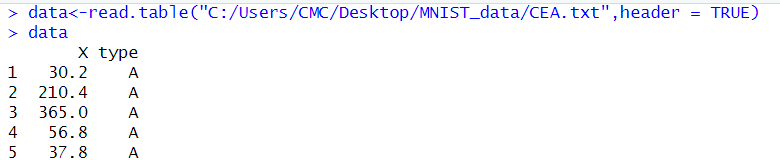
方差分析：



因此数据仅仅在Toxicant和Cure影响下显著，在Toxicant和Cure的共同影响下并不显著。由于不满足正态性和方差齐性，无法对方差分析结果进行保证。

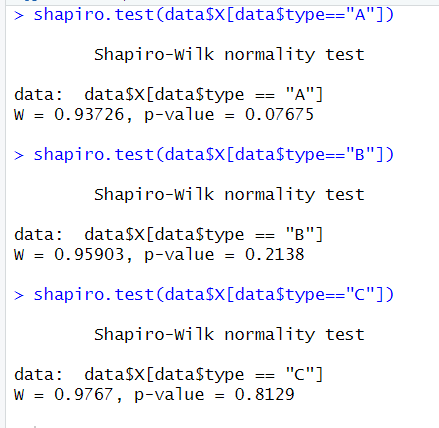
【作业】（说明：请附上每个题的代码、作图、文字分析和描述。）

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。



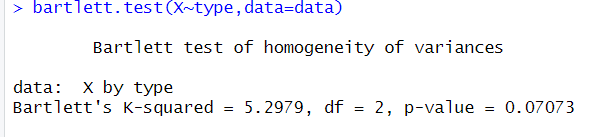
首先读入并输出数据。

进行正态性检验：



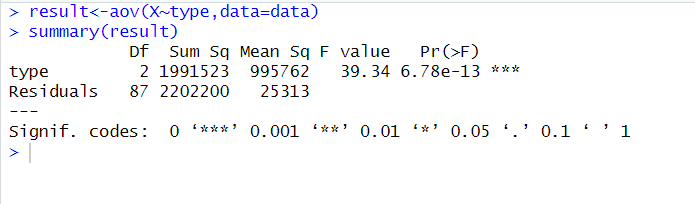
P-value均大于0.05因此满足正态性。

方差齐性检验：



由于P-value大于0.05因此通过方差齐性检验。

方差分析：



由题可知，P-value小于0.05因此CEA存在显著性差异。

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？

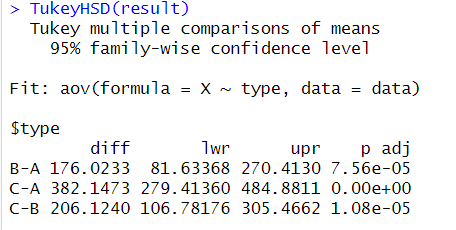
通过第一题的正态性检验表明，三个群体的CEA含量属于正态分布，方差相等。

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?

通过方差分析的结果可以看出，三个群体的CEA含量存在显著性差异。

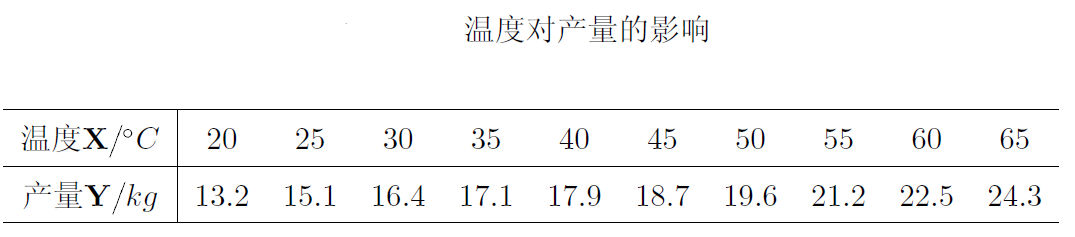
要指出哪些群体的CEA平均含量显著性差异显著需要进一步进行多重方差检验。

检验如下：

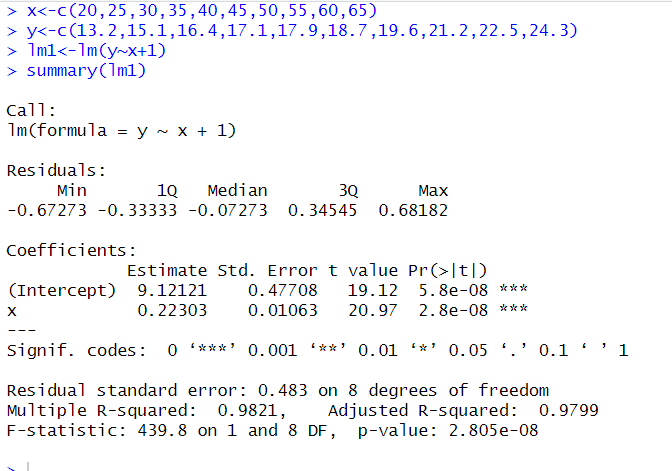


由图可知P-value远小于0.05，因此认为两两群体之间的CEA含量均有显著性差异。

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；

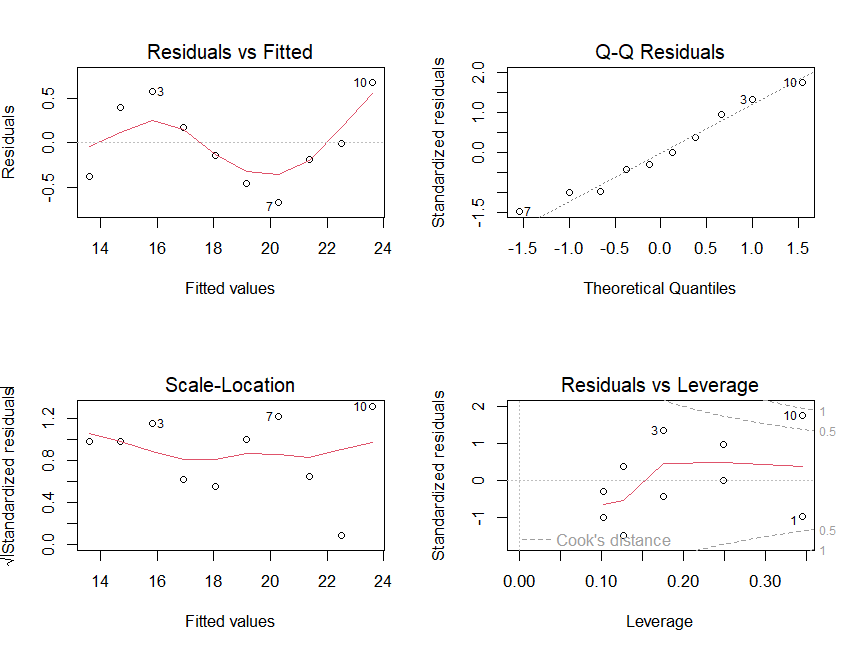


如图可知可建立的回归方程为Y=9.12121+0.22303X

1. 对其回归方程进行显著性检验和回归诊断；

由（1）中可知，P-value均远小于0.05，因此该回归模型存在显著性。

回归诊断：



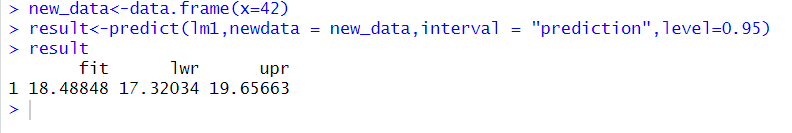
从图1看出残差均匀分布在直线两侧，未发现明显异常。

从图2看出数据分布在45度对角线周围，基本服从正态分布。

从图3可以看出，方差基本相同，具有方差齐性。

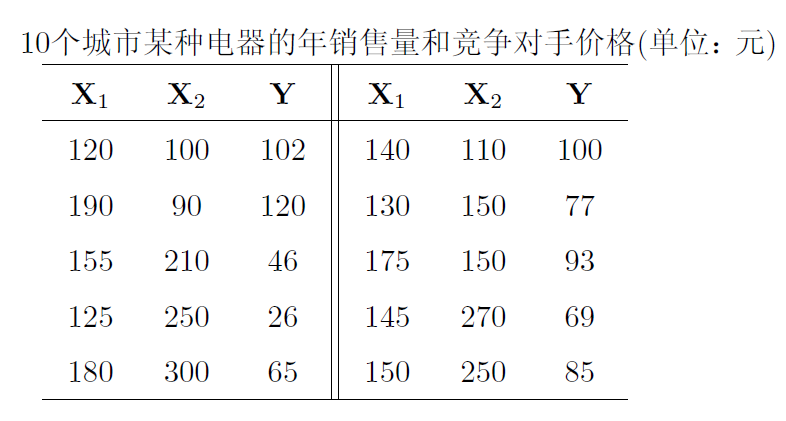
从图4可以看出，存在一个点超过0.5的线对回归影响较大属于异常值。

1. 预测X=42℃时产量的估计值及预测值区间（置信度为95%）。

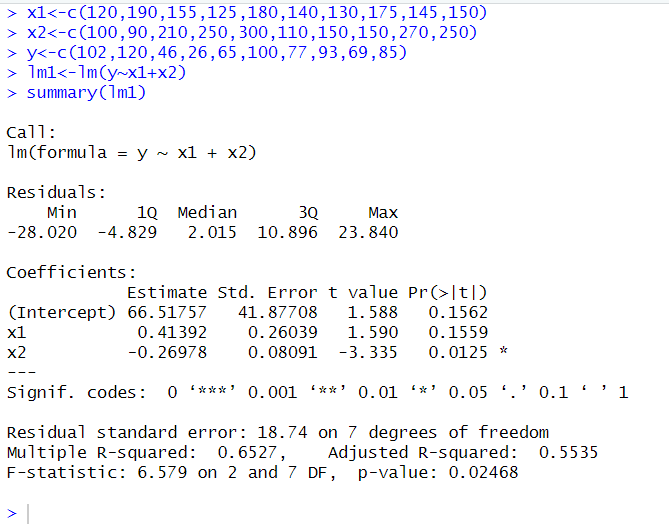


由题可知在（2）的模型下，预测值为18.48848，0.95置信区间为17.32034到19.65663

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。



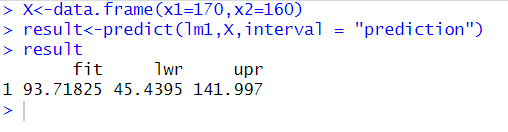
1. 建立Y 与X1及X2的回归关系, 并说明回归方程式在 的水平上是否显著？并解释回归系数的含义;



由此可知，x1的和截距的P-value大于0.05，因此x1和P-value和y之间没有显著的线性关系。X2P-value小于0.05因此存在显著的线性关系。

回归系数的含义：回归系数越大表明相关系越高，为正数表明是正相关，为负数表明是负相关。

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;

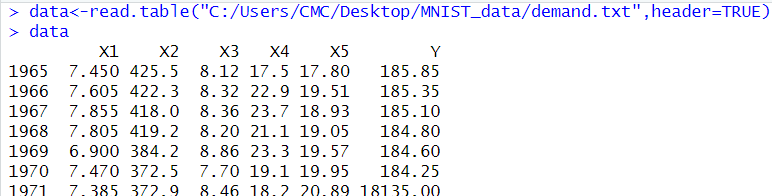


预测出的年销售量为93.71825，在0.95置信水平下的置信区间为45.4395到141.997。

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。

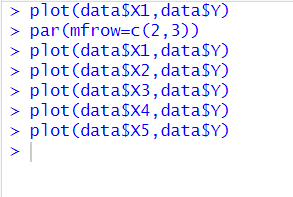
要求：

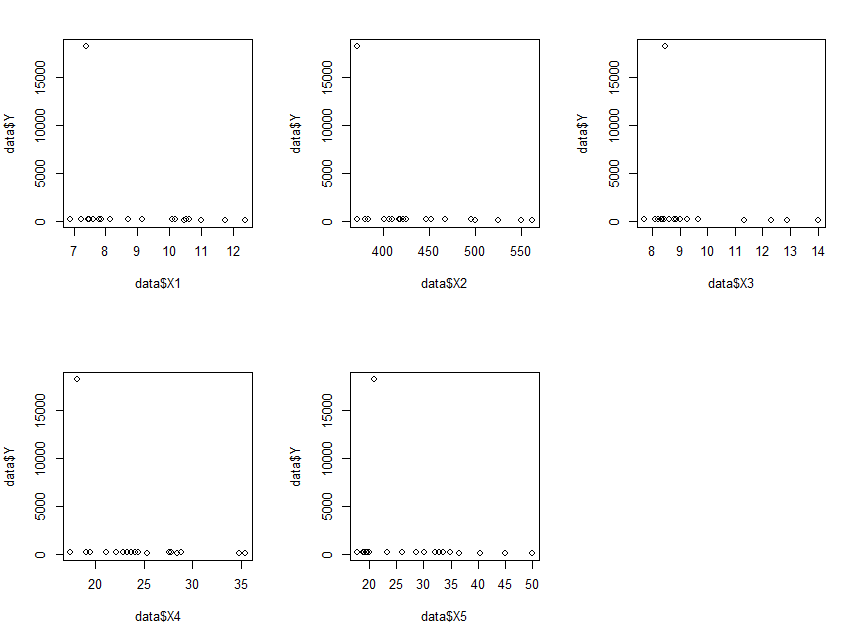
1. 建立数据集, 并画出各自变量与因变量的散点图；



读入数据并输出。

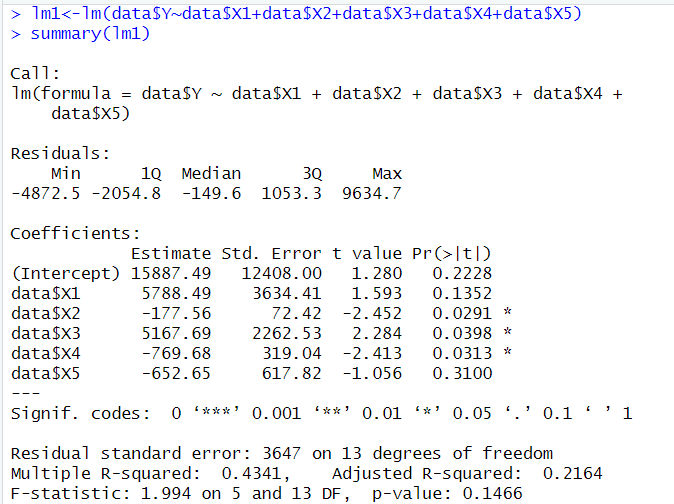
散点图如下：





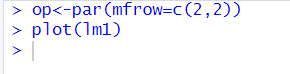
1. 进行线性回归分析, 对其回归方程进行显著性检验和回归诊断（提示：回归诊断包括异常值，多重共线性，残差的独立性、等方差性和正态性检验）。

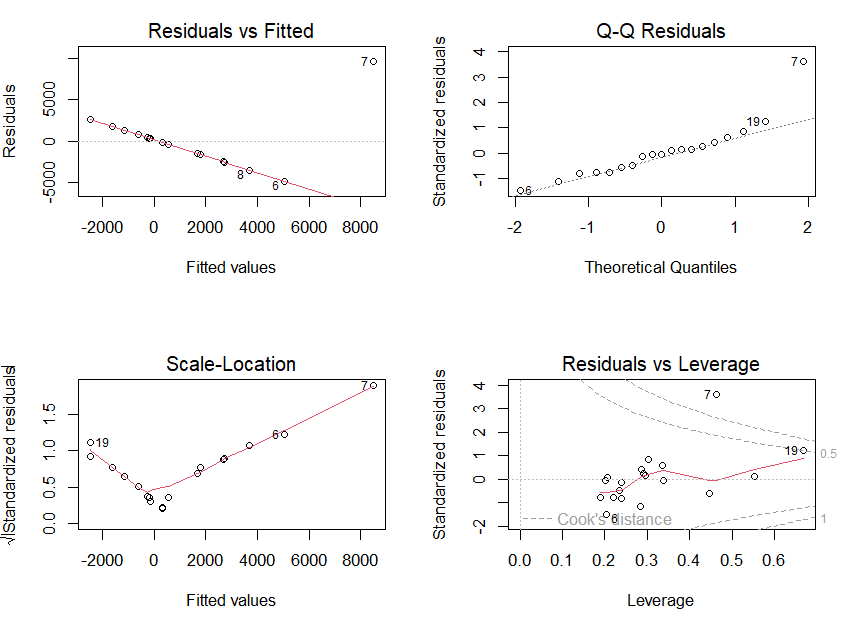
线性回归分析如下：



显著性检验表明X2、X3、X4与Y之间存在较为显著的线性关系。截距和X5与Y之间的线性关系不显著。

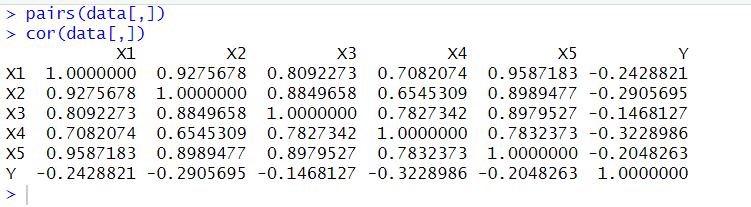
回归诊断：

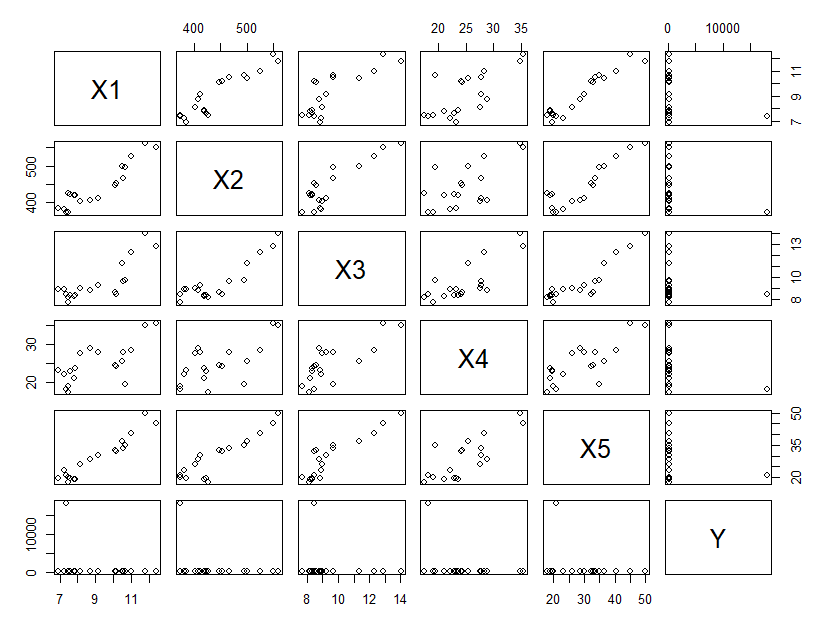




从图1可以看出残差并不是均匀分布在直线两侧。QQ图上表明数据存在异常值造成的严重偏态，图3的方差分布存在异常，方差齐不满足，图4也可以看出存在异常点严重影响了最终的回归效果。

多重共线性检验：

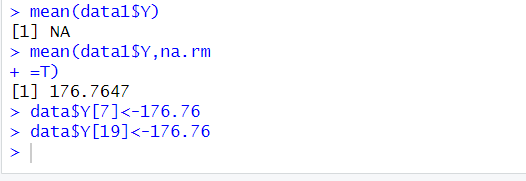




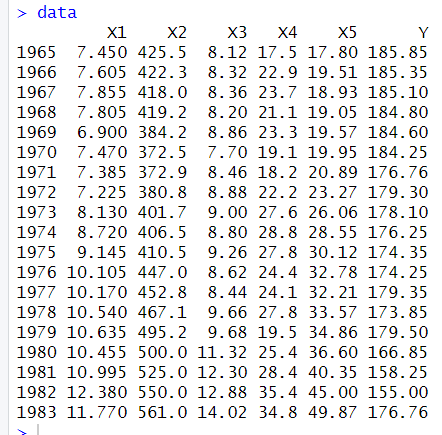
从这些结果中可以知道数据之间存在着多重共线性。

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）。

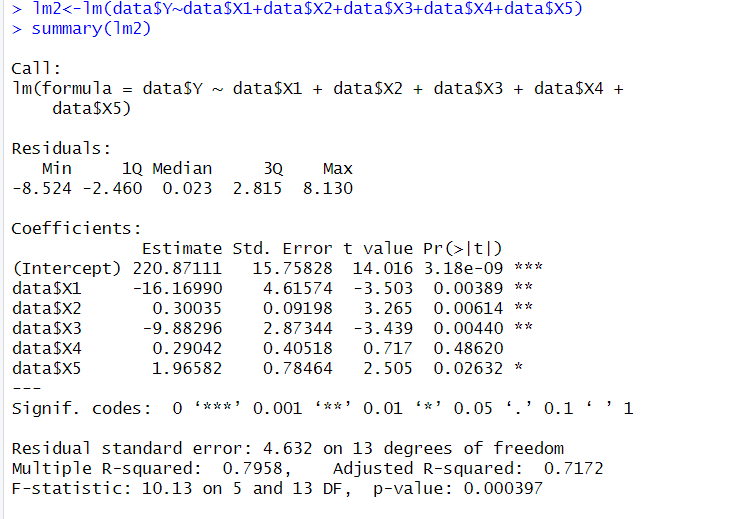
去除异常值：



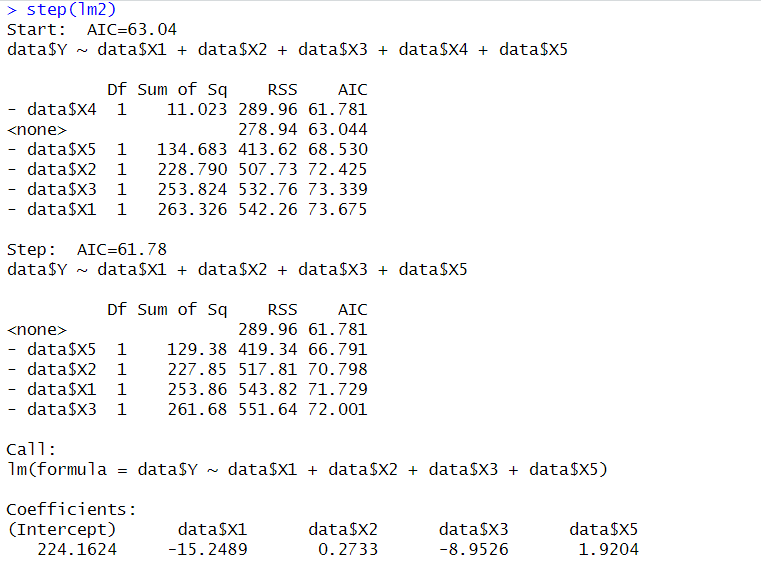
用均值来代替后data变为：



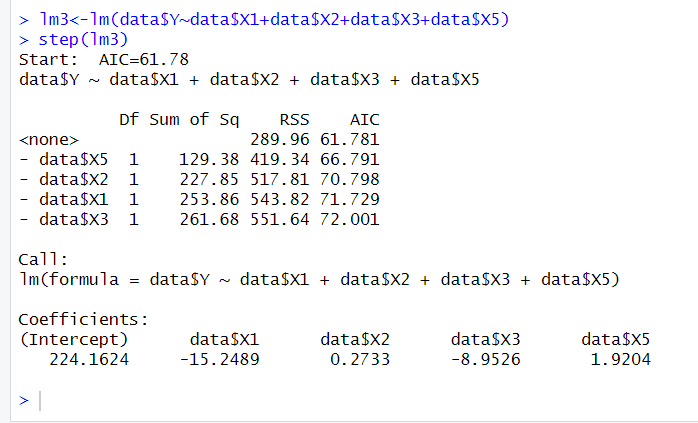
现在重新进行回归分析：



我们可以看出显著性有了明显的改善，但是仍然存在X4与Y之间的显著性不高，因此进行分步回归：



由此可以看出在去除X4时AIC值变小，模型更优。



重新构建回归模型，发现此时模型即为最优。

故最优的回归模型为Y=-15.2489X1+0.2733X2-8.9536X3+1.9204X5+224.1624