钻石属性对钻石价格的影响权重分析

# 摘要

钻石的价格受许多因素所影响，如切工，颜色，纯净度等。但是各个因素对钻石价格的影响所占权重却不同。本研究的目的在于探究该权重的详细信息。

首先从R中自带数据集diamonds 中的部分变量作为度量钻石价格的预测变量，随后筛出台宽比和亭深比为最佳的钻石，并剔除克拉数在0.3-2.0以外的数据组。使用判断矩阵得到切工，颜色，纯净度的量化参数。

为了消除量纲，将数据标准化。为了使绘出图更清晰，运用R语言，从筛选出来的6689个数据中随机抽取1000个，利用R进行探索性数据分析，再对克拉数进行简单回归分析，知克拉数carat对价格price影响较为明显。随后加入其余5个变量，得到多元线性回归模型。建立模型后，求得残差以及标准化残差、杠杆值等数据，并进行观测度量的求取。由观测及分析结果知，得到模型较符合实际钻石属性对钻石价格的影响权重。

# 正文

## 问题描述

钻石作为一种著名的奢侈品，一直以高昂的价格闻名。决定一颗钻石价格的因素有许多，克拉数、切工、净度、颜色等。本研究的目的在于确认各个因素对钻石价格的影响权重，利用多元线性回归方法确定构成钻石价格函数的各预测变量的回归系数。

## 变量选择

经过对钻石属性的了解，选取了最主要的几个因素作为预测变量：克拉数carat、切工cut、颜色color、净度clarity、亭深比depth、台宽比table。钻石价格为响应变量。

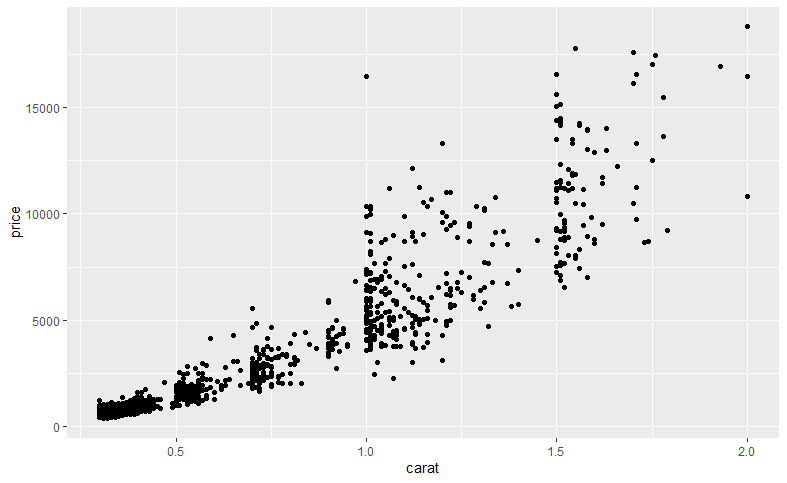
## 数据收集整理

数据集采用R语言自带数据集diamonds。

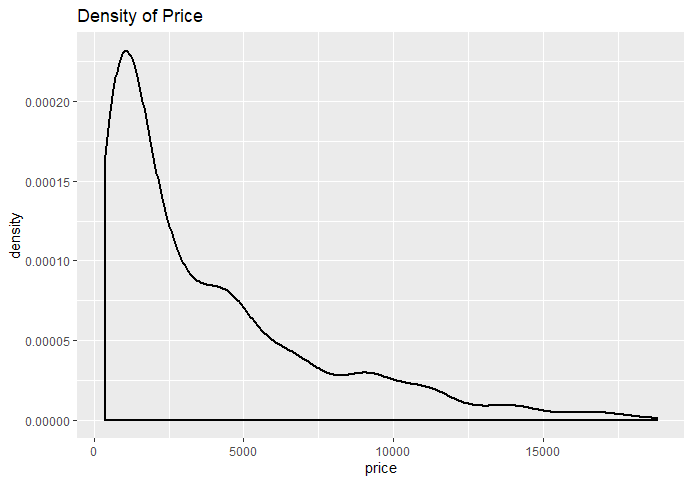
由于钻石的克拉数、亭深比和台宽比过大或过小会导致钻石价格出现较大波动，根据常见的钻石分级，取table∈[57,60], depth∈[59.3,61], carat∈[0.3,2.0]。

根据上述数据筛选原则，初步得到6689个数据。

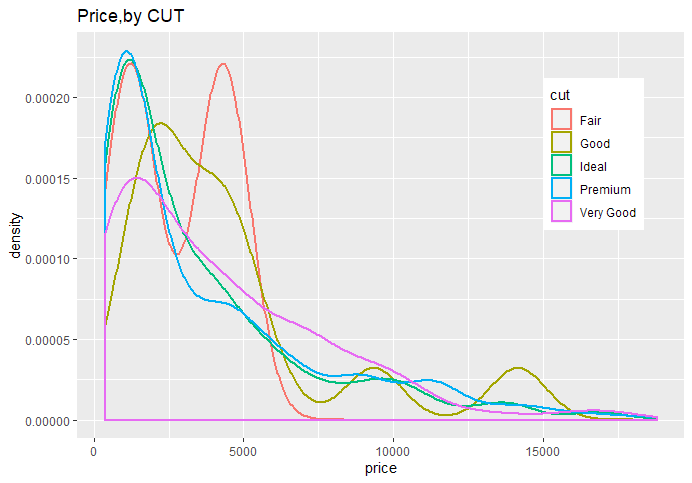
抽取一千个数据进行分析



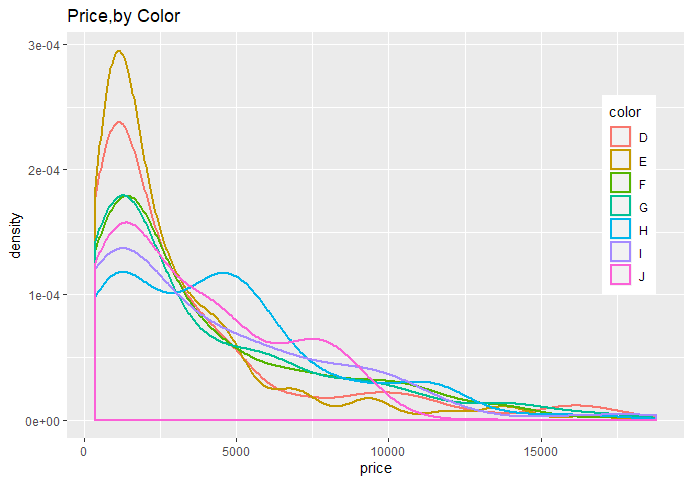
随着钻石克拉数增加，钻石的价格呈上升趋势，且在某些重量上相对集中。钻石的克拉数越高价格的离散程度越大，而克拉数较低时价格的离散程度较小。初步分析的结果是：钻石的克拉数越高，其价格越高；钻石的克拉数越高其价格受其他因素影响越大。



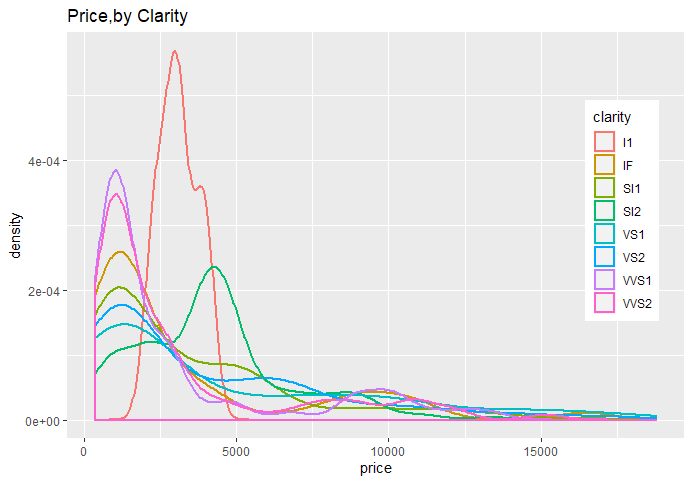
钻石的价格呈正偏峰分布，多数的值都集中在价格较低部分。



将价格按切工分组，各种切工的钻石分布情况大多集中于价格较低区域。



将钻石价格按颜色分组，颜色好(D)的钻石的价格集中在价格较低的区域，而颜色差(J)的钻石价格的分布更平均。



将价格按净度分组，发现和第二幅图相似，净度较好(IF)的钻石多集中在低价格区域，而净度较差(l1)的钻石的价格分布更平均一些。

## 建立判断矩阵

对于颜色，切工，纯净度三个非参数变量，采用构造判断矩阵的方法，分别对这三个变量进行量化。

**纯净度的判断矩阵**：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **l1** | **lF** | **VVS1** | **VVS2** | **VS1** | **VS2** | **Sl1** | **Sl2** |
| **l1** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| **lF** | **1/2** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| **VVS1** | **1/3** | **1/2** | **1** | **2** | **3** | **4** | **5** | **6** |
| **VVS2** | **1/4** | **1/3** | **1/2** | **1** | **2** | **3** | **4** | **5** |
| **VS1** | **1/5** | **1/4** | **1/3** | **1/2** | **1** | **2** | **3** | **4** |
| **VS2** | **1/6** | **1/5** | **1/4** | **1/3** | **1/2** | **1** | **2** | **3** |
| **Sl1** | **1/7** | **1/6** | **1/5** | **1/4** | **1/3** | **1/2** | **1** | **2** |
| **Sl2** | **1/8** | **1/7** | **1/6** | **1/5** | **1/4** | **1/3** | **1/2** | **1** |

将判断矩阵进行按列归一化处理：

**0.3679 0.4355 0.4027 0.3545 0.3109 0.2748 0.2456 0.2222**

**0.1840 0.2177 0.2685 0.2659 0.2487 0.2290 0.2105 0.1944**

**0.1226 0.1089 0.1342 0.1773 0.1865 0.1832 0.1754 0.1667**

**0.0920 0.0726 0.0671 0.0886 0.1244 0.1374 0.1404 0.1389**

**0.0736 0.0544 0.0447 0.0443 0.0622 0.0916 0.1053 0.1111**

**0.0613 0.0435 0.0336 0.0295 0.0311 0.0458 0.0702 0.0833**

**0.0526 0.0363 0.0268 0.0222 0.0207 0.0229 0.0351 0.0556**

**0.0460 0.0311 0.0224 0.0177 0.0155 0.0153 0.0175 0.0278**

取对角线上的数，得到

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **l1** | **lF** | **VVS1** | **VVS2** | **VS1** | **VS2** | **Sl1** | **Sl2** |
| **0.3679** | **0.2177** | **0.1342** | **0.0886** | **0.0622** | **0.0458** | **0.0351** | **0.0278** |

**颜色的判断矩阵**：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | D | E | F | G | H | I | J |
| D | 1 | 3 | 5 | 7 | 9 | 11 | 13 |
| E | 1/3 | 1 | 3 | 5 | 7 | 9 | 11 |
| F | 1/5 | 1/3 | 1 | 3 | 5 | 7 | 9 |
| G | 1/7 | 1/5 | 1/3 | 1 | 3 | 5 | 7 |
| H | 1/9 | 1/7 | 1/5 | 1/3 | 1 | 3 | 5 |
| I | 1/11 | 1/9 | 1/7 | 1/5 | 1/3 | 1 | 3 |
| J | 1/13 | 1/11 | 1/9 | 1/7 | 1/5 | 1/3 | 1 |

将判断矩阵进行归一化处理：

0.5115 0.6150 0.5109 0.4198 0.3525 0.3028 0.2653

0.1705 0.2050 0.3065 0.2998 0.2742 0.2477 0.2245

0.1023 0.0683 0.1022 0.1799 0.1958 0.1927 0.1837

0.0731 0.0410 0.0341 0.0600 0.1175 0.1376 0.1429

0.0568 0.0293 0.0204 0.0200 0.0392 0.0826 0.1020

0.0465 0.0228 0.0146 0.0120 0.0131 0.0275 0.0612

0.0393 0.0186 0.0114 0.0086 0.0078 0.0092 0.0204

取对角线上数得：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| D | E | F | G | H | I | J |
| 0.5115 | 0.2050 | 0.1022 | 0.0600 | 0.0392 | 0.0275 | 0.0204 |

**切工的判断矩阵：**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Ideal** | **Premium** | **Very Good** | **Good** | **Fair** |
| **Ideal** | **1** | **3** | **5** | **7** | **9** |
| **Premium** | **1/3** | **1** | **3** | **5** | **7** |
| **Very Good** | **1/5** | **1/3** | **1** | **3** | **5** |
| **Good** | **1/7** | **1/5** | **1/3** | **1** | **3** |
| **Fair** | **1/9** | **1/7** | **1/5** | **1/3** | **1** |

**得到归一化矩阵：**

**0.5595 0.6415 0.5245 0.4286 0.3600**

**0.1865 0.2138 0.3147 0.3061 0.2800**

**0.1119 0.0713 0.1049 0.1837 0.2000**

**0.0799 0.0428 0.0350 0.0612 0.1200**

**0.0622 0.0305 0.0210 0.0204 0.0400**

取对角线数得到：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Ideal** | **Premium** | **Very Good** | **Good** | **Fair** |
| **0.5595** | **0.2138** | **0.1049** | **0.0612** | **0.0400** |

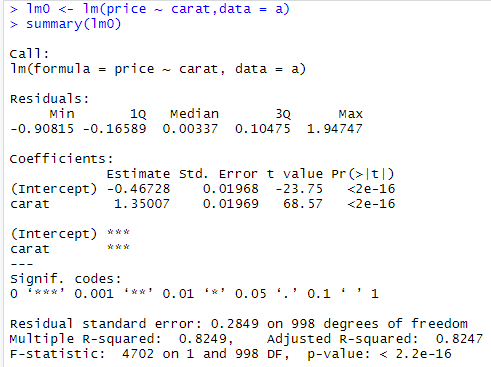
## 模型设定

1. 所有钻石的形状都是较为规则的，即长宽高不会过大或过小。
2. 钻石只受表中变量所影响，其他因素对钻石的影响极小，忽略不计。
3. 市场不存在恶意抬升或打压钻石价格的情况。
4. 响应变量钻石价格与预测变量关系的模型关于回归参数是线性的。
5. 误差项符合正态性假定。
6. 预测变量是非随机的，且取值没有测量误差。

## 模型拟合

1. 本多元线性回归模型运用软件R进行数据分析以及模型建立。
2. 数据共有6个预测变量，分别为克拉数carat，切工cut，颜色color，clarity净度，depth庭深比，width台宽比。响应变量为价格price。

3. 运用R语言，从筛选出来的6689个数据中随机抽取1000个，随后对数据进行标准化处理。通过二元图矩阵的分析得知，克拉数carat对价格price的影响最为显著，因此优先对carat进行简单回归分析，得到如下结果：

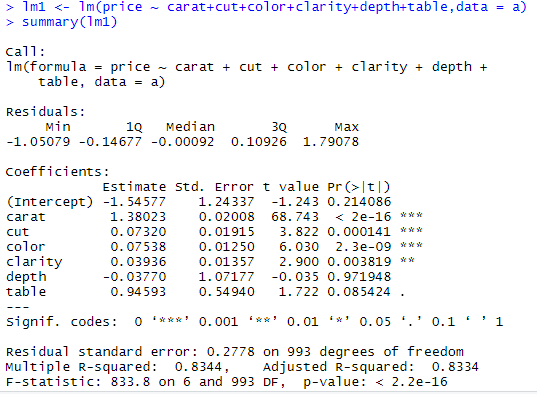


此时得到简单回归模型：

Y = -0.46728 + 1.35007 \* X1 + ε

此时的R方为0.8249，可知克拉数carat对价格price影响较为明显。

1. 随后加入其余5个变量：得到多元线性回归模型：如下：



得到多元线性回归模型：

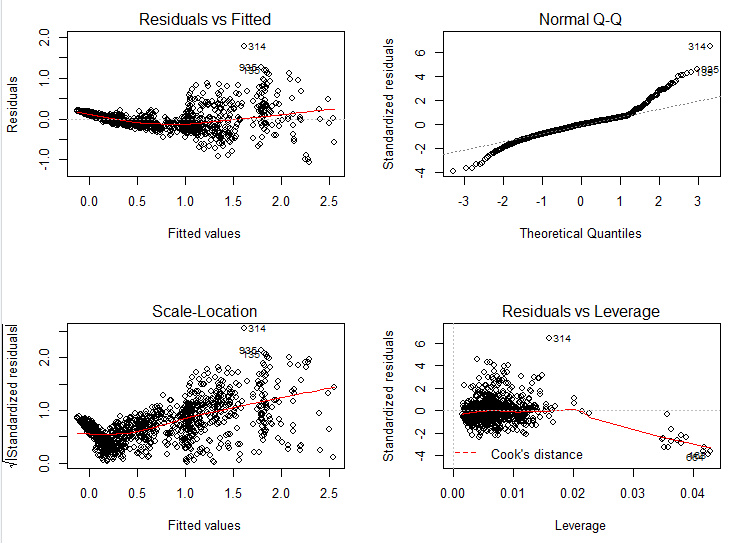
Y = -1.54577 + 1.38023 \* X1 + 0.0732 \* X2 + 0.07538 \* X3 + 0.3936 \* X4

- 0.0377 \* X5 + 0.94593 \* X6 + ε

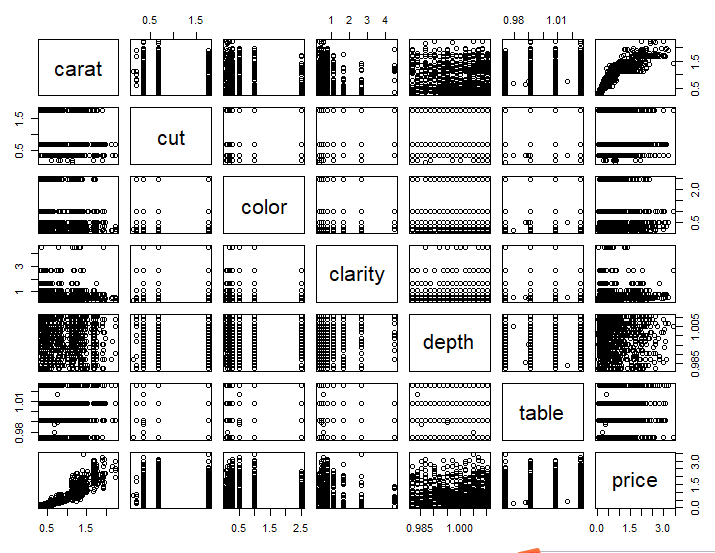
此时的R方为0.8344，相比简单回归有所提升，说明剩余的五个变量有所影响。

由变量的显著性检验结果可知carat，color两个变量影响最为显著，而table和width两个变量较不显著，对响应变量的影响占比较小。

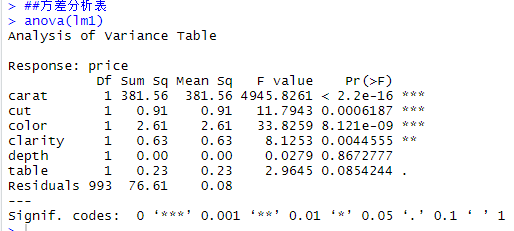
1. 关于多元线性回归的分析图



1. 图矩阵：



1. 方差分析表anova



可知残差平方和为76.61，回归平方和为385.94。

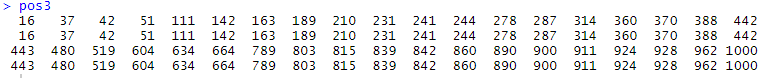
## 模型评价

1.求残差以及标准化残差，杠杆值等数据：

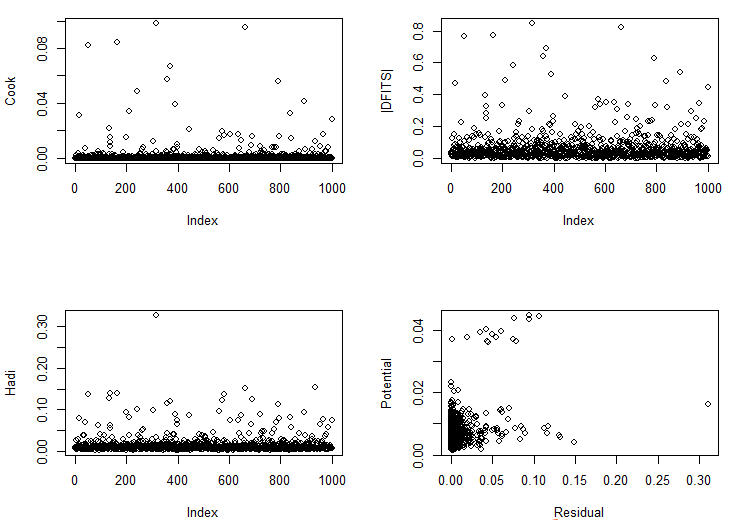
标准差>3的异常点：

6

高杠杆点：杠杆值大于 2\*(p+1)/n：



1. 进行观测度量的求取：



（图片内依次为Cook距离，Welsch and Kuh 度量，Hadi度量，位势-残差图）

3. 由图像可知目前的模型对于数据的拟合效果较好，有部分数据为强影响点或者高杠杆点，并且存在一定的异方差问题，但是不应程式化的删除或降低权重，因此暂时不作处理。在接下来的模型改进中，将进一步分析。

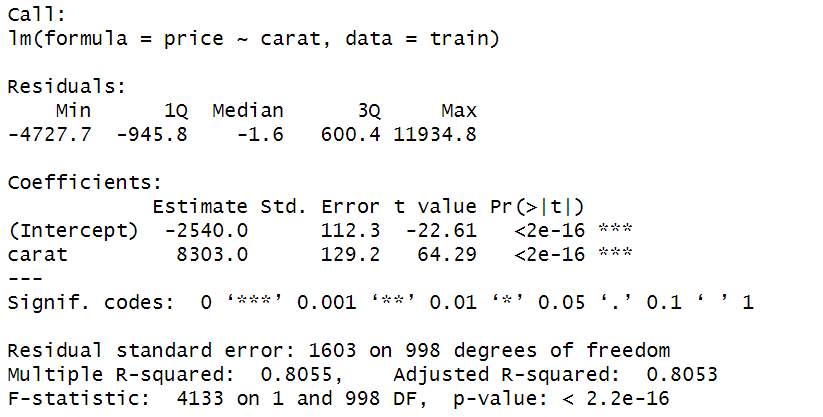
## 模型改进

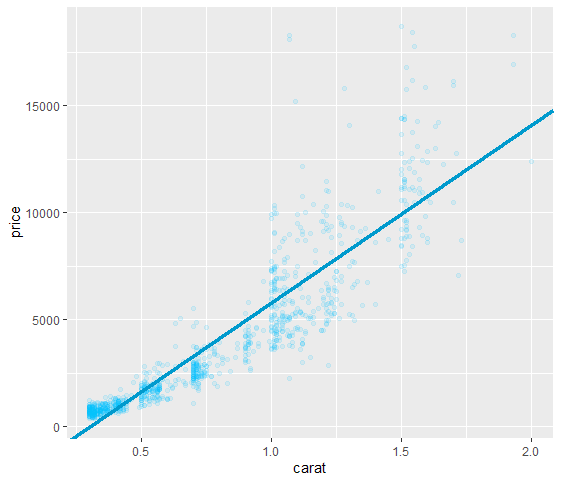
1. 抽取数据

在筛选的数据里随机抽取2000个数据，一半用于拟合，一半用于检验

2.

价格和克拉的简单线性回归

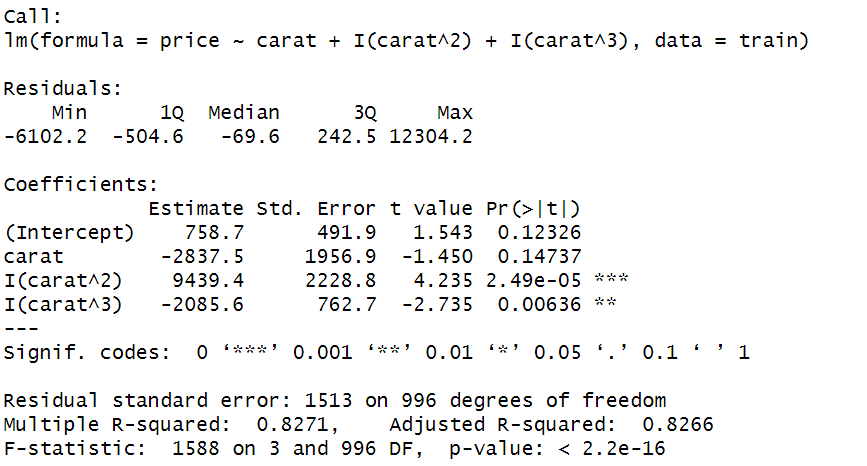


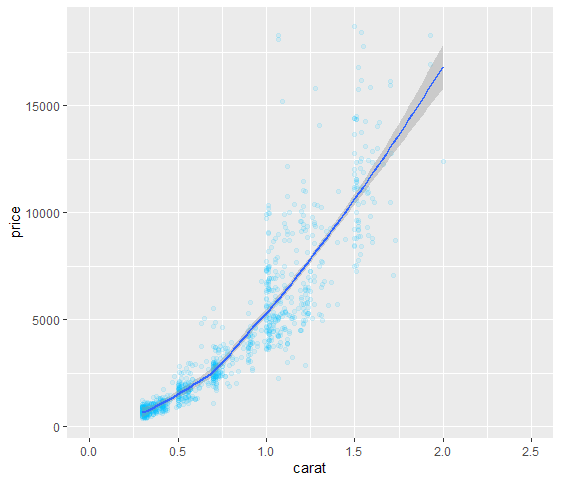


这时R方为0.8055，进一步分析。

3.

多项式回归

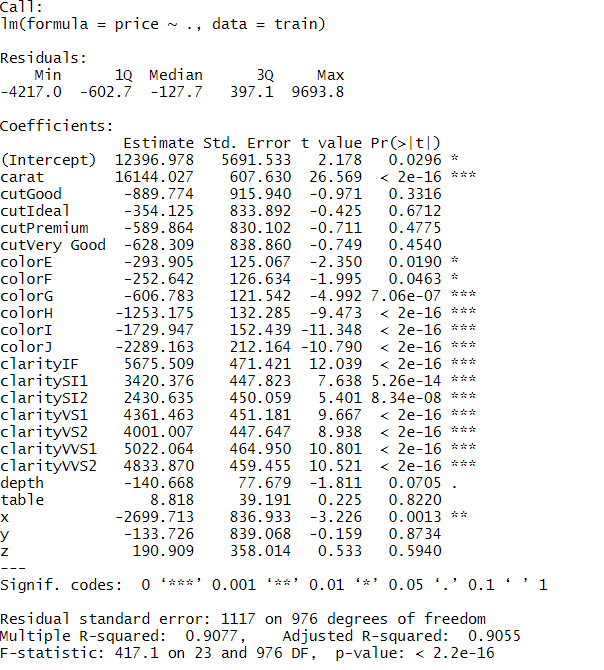




这时R方为0.8271，继续分析。

4.

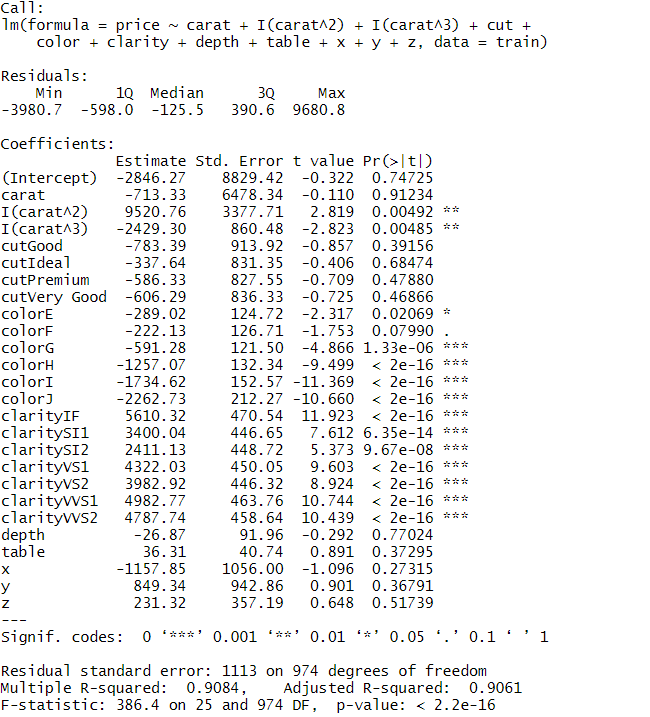
多元线性回归



这时R方为0.9077，继续分析。

5.

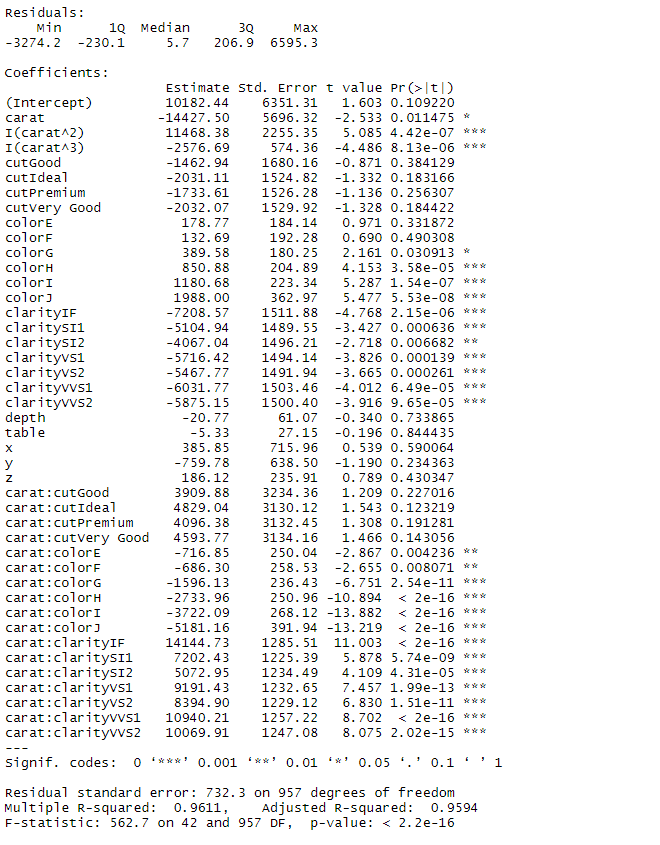
添加变量



这时R方为0.9084，继续分析。

6.

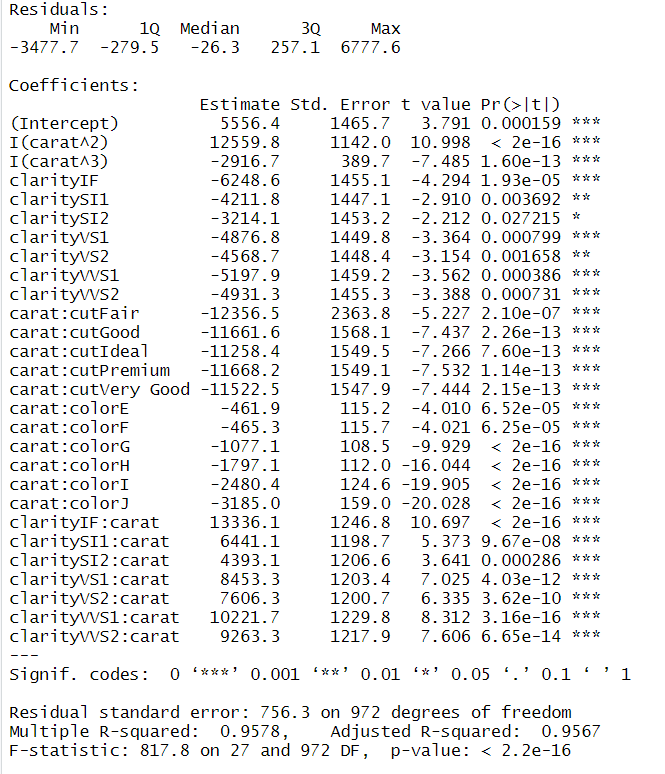
加入交互项



这时R方为0.9611，说明模型拟合效果较好，但是仍然可以简化模型。

7.

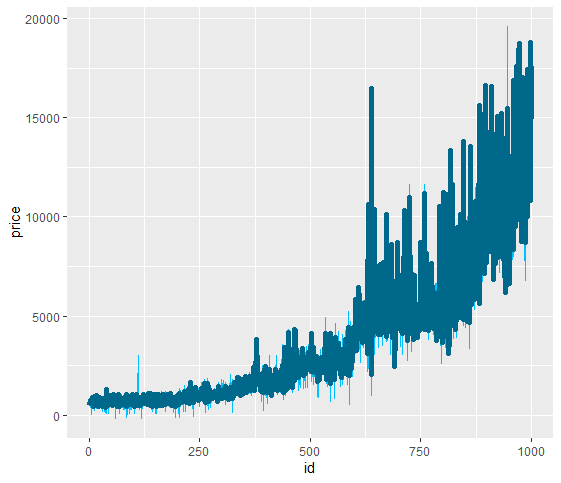
简化



这时R方为0.9578，符合预期。

8.

检测将另外1000个数据带入模型



可以看出，在价格较低时拟合的不错，但是价格越高，波动越大，越不准确。

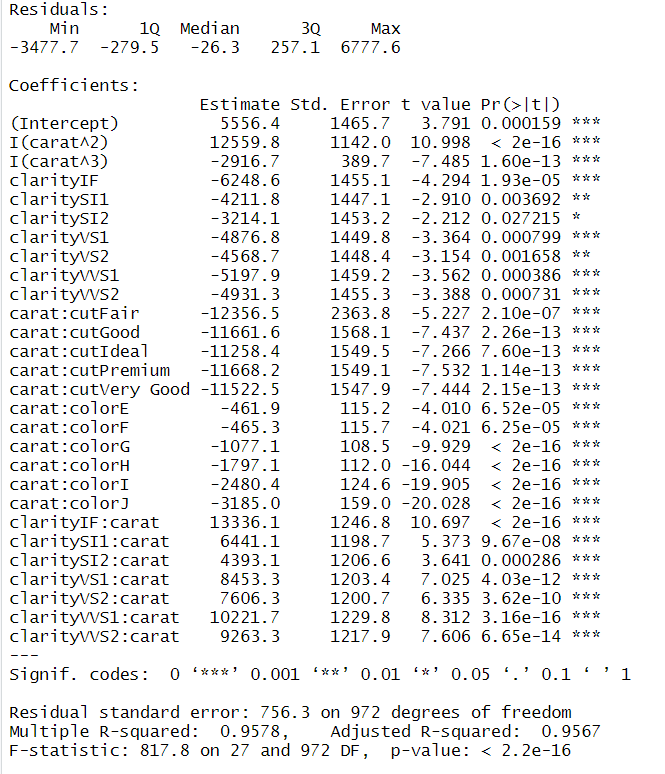
# 结论

由以上数据分析过程，得出的初步结论为：钻石的克拉数越高，其价格越高；钻石的克拉数越高其价格受其他因素影响越大。

进一步分析，初步建立的多元线性回归模型Y = -1.54577 + 1.38023 \* X1 + 0.0732 \* X2 + 0.07538 \* X3 + 0.3936 \* X4 - 0.0377 \* X5 + 0.94593 \* X6 + ε的R^2为0.8344，相比仅针对克拉数的简单回归，R^2有所提升，说明添加的cut, color, clarity, width, table 这五个变量对price确实有所影响。

由变量的显著性检验结果可知carat，color两个变量影响最为显著，而table和width两个变量较不显著，对响应变量的影响占比较小。

优化后的模型由clarity, carat^2, carat^3, carat:cut, carat:color, carat:clarity这些预测变量构成。回归系数如下：



至此，钻石的各个属性对钻石价格的影响权重已有了较清楚的结果，图中的回归系数极为该属性所占权重。

本模型的拟合效果在钻石价格较低时较为准确，但价格越高，波动就越大。这一点可以由越大的整块钻石越稀有，价格就越高，故其受各预测变量的影响波动大。考虑到消费者对于钻石存在不同的偏好，所以高价钻石的波动可以由此解释。

# 附录

本研究所涉及数据集为R语言自带数据集diamonds，分析中的R语言代码另附