目录

[数据处理 1](#_Toc62144881)

[碳减排 1](#_Toc62144882)

[签到表 2](#_Toc62144883)

[领取金币 3](#_Toc62144884)

[驾驶评分 3](#_Toc62144885)

[APP用户信息 3](#_Toc62144886)

[相关性： 4](#_Toc62144887)

[正态性、方差齐性 4](#_Toc62144888)

[正态性 4](#_Toc62144889)

[方差齐性、方差分析等等 5](#_Toc62144890)

[相关关系： 9](#_Toc62144891)

[大概的逐步回归一下 11](#_Toc62144892)

[线性回归 11](#_Toc62144893)

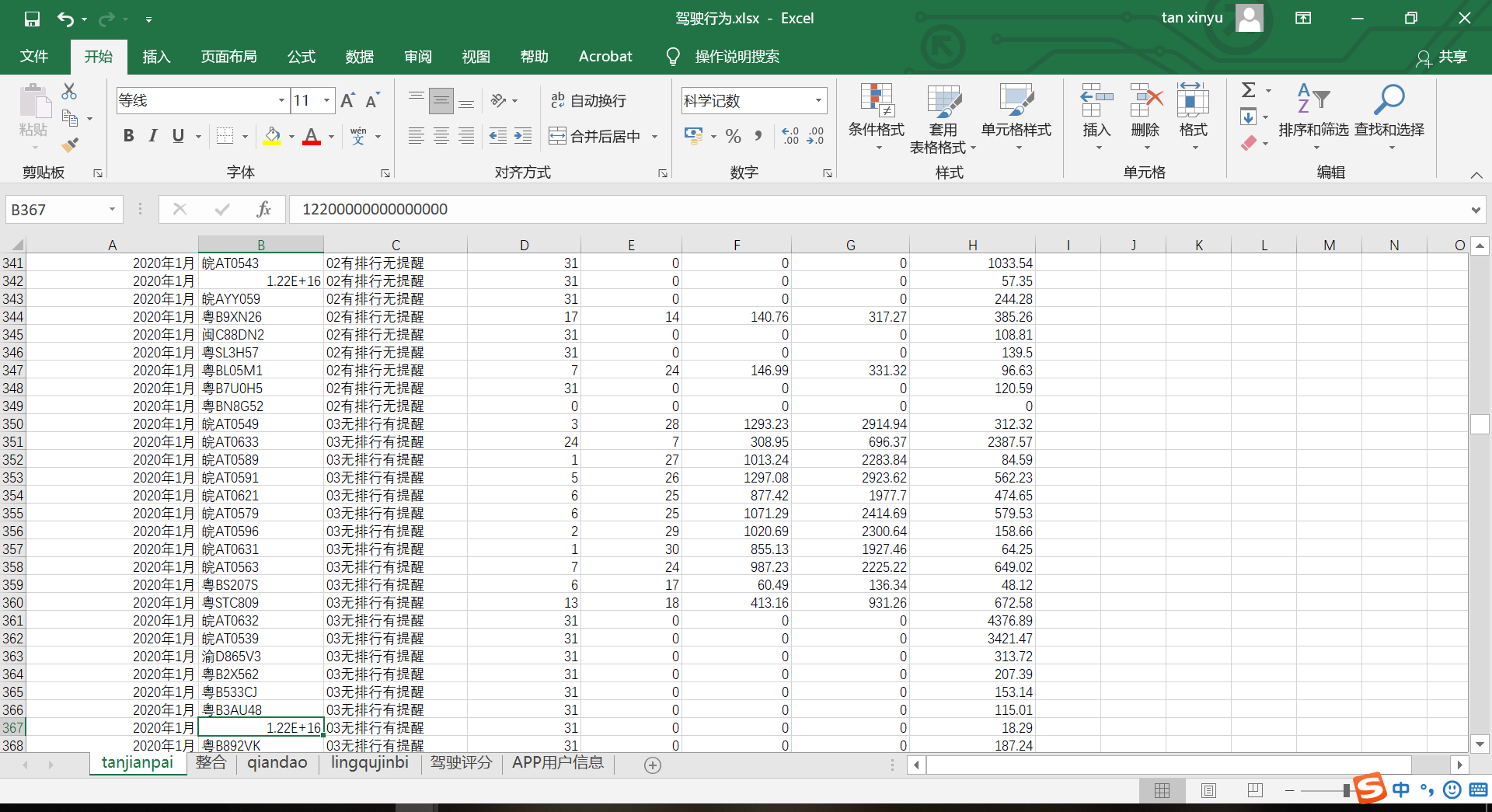
[别的回归试试 12](#_Toc62144894)

# 数据处理

由于碳减排是按月分的，所以先就按月对数据进行计算和合并

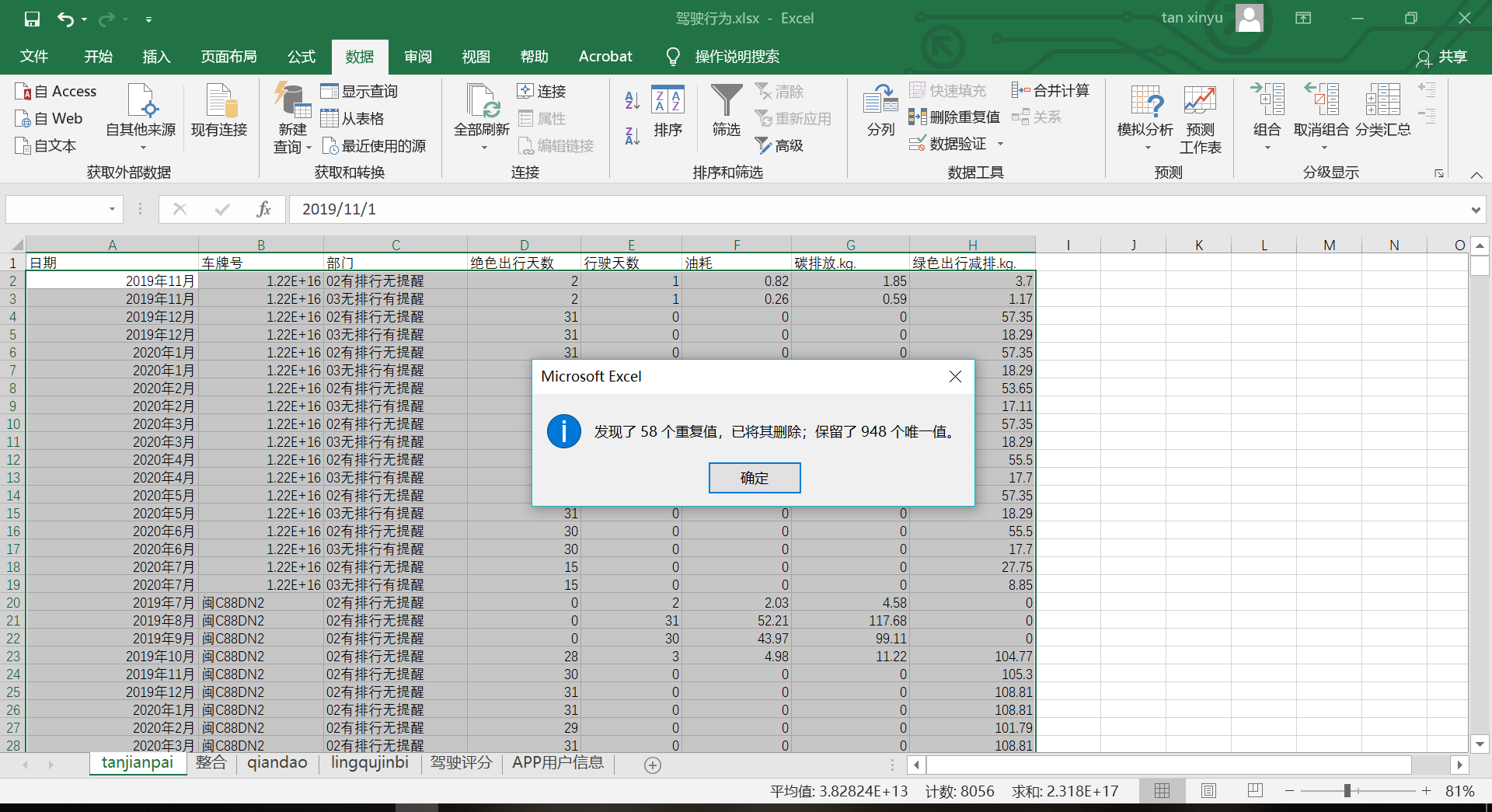
## 碳减排

几乎每个月都会有同一个车牌号缺失的数据，这个我没删掉，因为如果光看APP功能的影响（有排行无提醒等）和碳减排关系的话应该不具体到某辆车：



78辆车有具体信息，那个乱码的好像是两个车牌号（在02,03组）

操作去重



## 签到表

有60辆车签过到，编号51473（粤TEST01）、51483（粤BT2220）、51291（粤B5P05E）是2016年就用的老人

他们2018、 2017、2016的数据还挺多，删了。由于碳减排数据最早到2019/06，所以此前的删了。

感觉这个签到也许能反映使用积极性，这里先不考虑2016-2018年使用APP对那三个2019、2020年驾驶行为的影响。

## 领取金币

这也能反映使用积极性。签到的一批人就是领金币的一批人，因为好像就是靠签到攒的金币。

但是考察时间段内（2019.06-2020.09）领金币跟签到的排名不完全对应。估计有的人在2019.6之前签到的多一直没领，后来一看这么多钱，想起来了，一股脑全领了所以记录上他就很高。

由于这部分和签到功能几乎一致，但是这个表他没有记录时间，所以不予合并。

## 驾驶评分

* 删除所有行驶里程=0的行，则“插拔”列全为“未拔出”，这一列删除。
* 行驶里程不知道有没有用，加进去吧。列“总行驶里程”
* 油耗到底可以干嘛，和碳减排会不会有关系？——*一般来讲汽车****启动****时，发动机缸内是冷的，为了使启动顺利，会适当的加浓油气混合气，这种情况下会造成CO和PM的排放的上升。（氧气少了，燃料多了，不完全燃烧的多了，CO和碳粒就多了）汽车****加速****时，为了加速的平稳，会加浓混合气，CO和PM的排放会上升。在****匀速或减速****行驶时，为了降低油耗，提高燃烧效率，会增大空燃比，造成NOX的排放上升。还有点火角对排放、油耗的影响...等等等等。*这都是驾驶行为，应该是有关系的吧。
* 整合精确到月，增加列“参与评分天数”
* 算“每月均分”
* 平均速度（还是平均）、行驶时长、急加速、急减速、急转弯、疲劳驾驶、夜间驾驶时长（求和）。我发现一个月的行驶里程相加 不等于（最后一天总里程-上个月最后一天总里程），我怀疑设备成精了？！目前“月总里程”使用每月行驶里程相加。

上述表格合并后，此时数据为948条月度数据，其中无“驾驶评分”相关数据记录的有426条，完整数据522条。缺失近50%。

## APP用户信息

有43名用户的信息（车牌、年龄、性别、收入、历史出险、驾龄、是否已婚）

车牌都是“**皖xxxx**”，缺AT0626和0540，以及其他省份车主的信息

全部已婚，这列删去

历史出险次数没办法按照时间标进去，所以删掉了

其他列也有些许缺失值

最后增加了一些对数列~~，所有缺失值标为NA~~

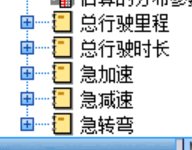
数据的具体情况见output.doc

# 相关性：

## 正态性、方差齐性

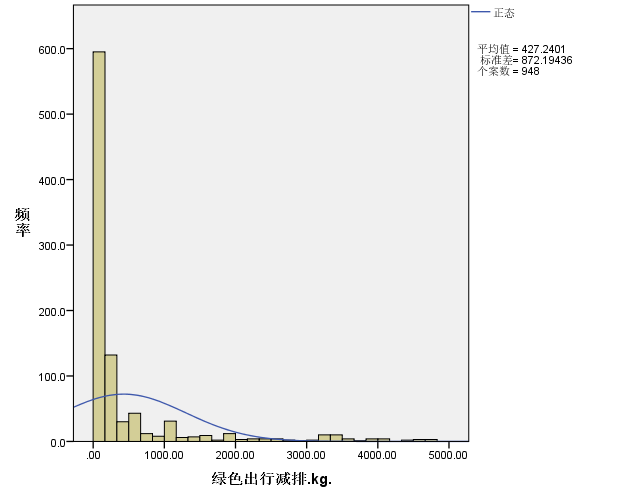
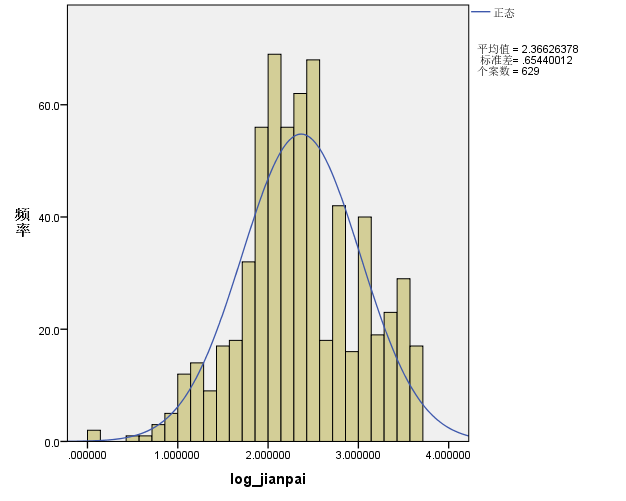
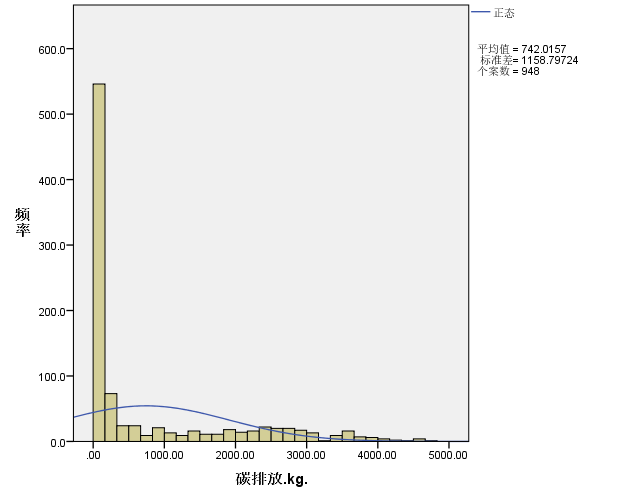
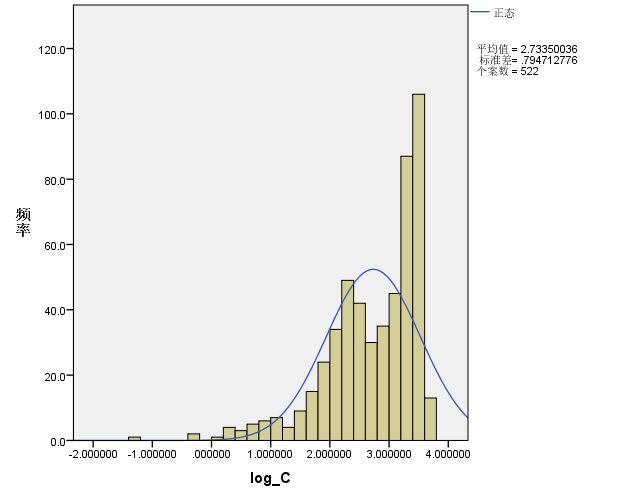
### 正态性

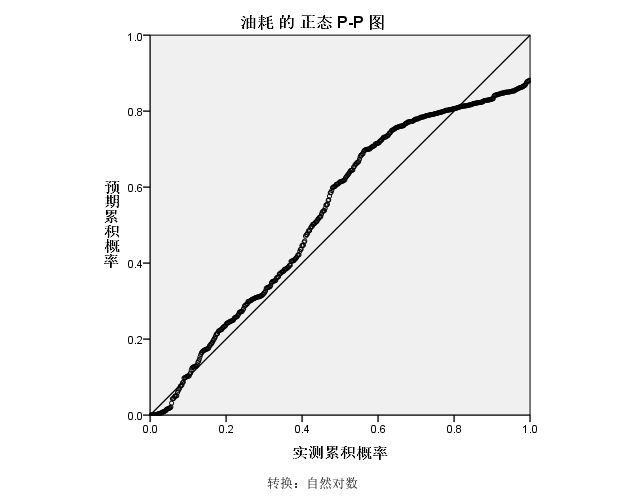
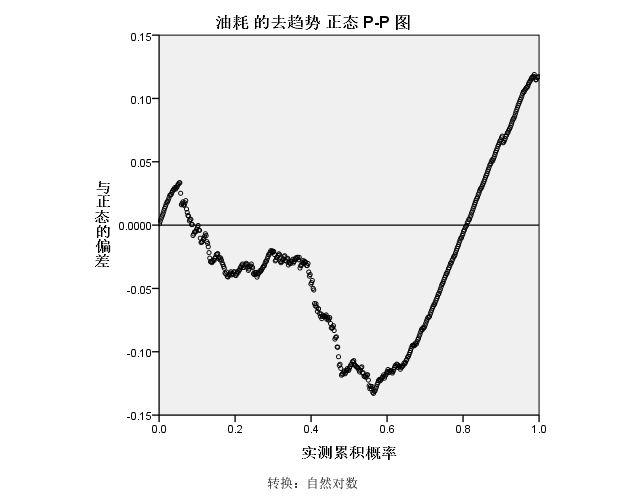
变量变换后基本满足正态性。

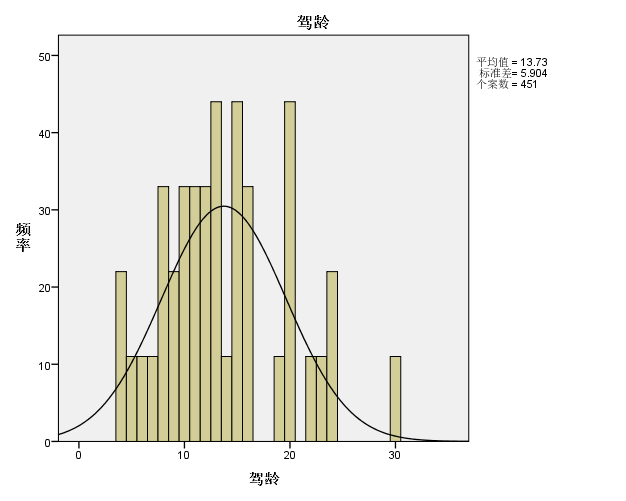
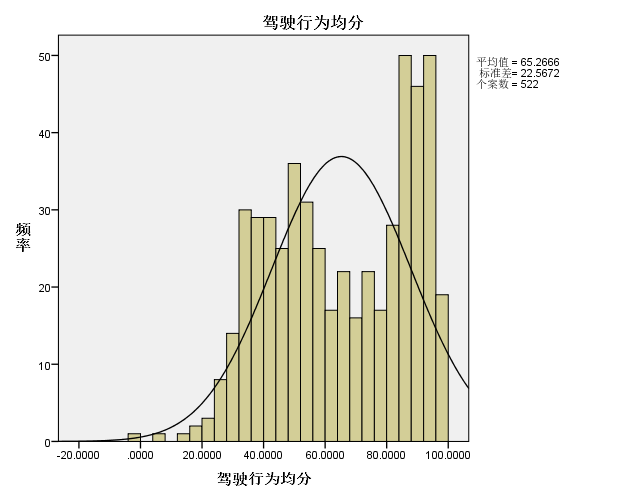
左边这些和碳减排、碳排放、油耗、夜间驾驶时长一起取对数，

驾龄、绿色出行天数、行驶天数等等有一点点偏，但是也看得过去。

下面是部分的展示：

### 方差齐性、方差分析等等

按1234四个不同的组分，看方差齐性，

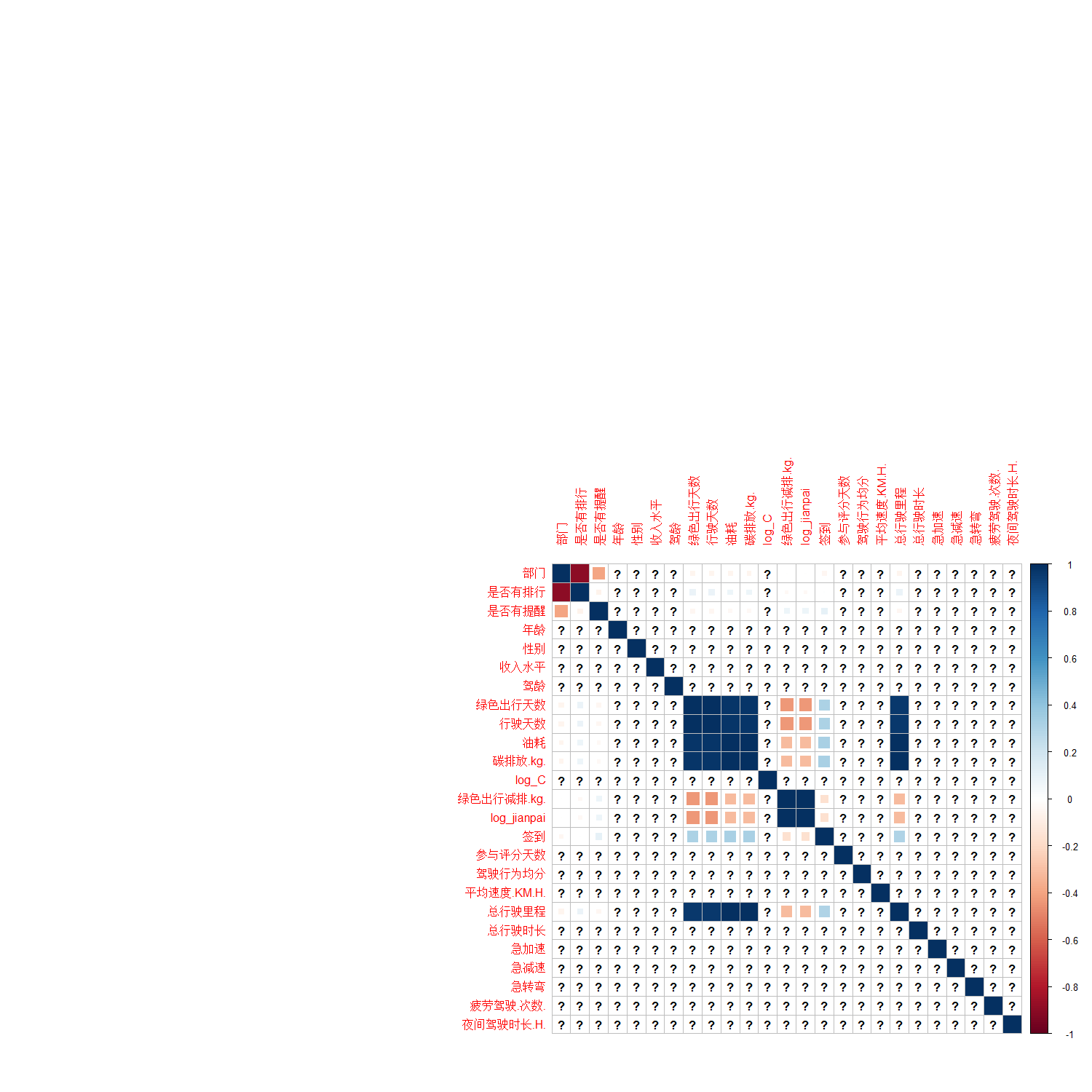
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **方差齐性检验** | | | | |
|  | 莱文统计 | 自由度 1 | 自由度 2 | 显著性 |
| 年龄 | 10.834 | 3 | 436 | .000 |
| 驾龄 | 6.467 | 3 | 447 | .000 |
| 绿色出行天数 | 1.035 | 3 | 944 | .376 |
| 行驶天数 | 2.003 | 3 | 944 | .112 |
| 油耗 | 4.236 | 3 | 944 | .006 |
| log\_g | .912 | 3 | 518 | .435 |
| 碳排放.kg. | 4.236 | 3 | 944 | .006 |
| log\_C | .905 | 3 | 518 | .439 |
| 绿色出行减排.kg. | 9.257 | 3 | 944 | .000 |
| log\_jianpai | 1.790 | 3 | 625 | .148 |
| 参与评分天数 | .086 | 3 | 519 | .968 |
| 驾驶行为均分 | 1.966 | 3 | 518 | .118 |
| 平均速度.KM.H. | .918 | 3 | 518 | .432 |
| 总行驶里程 | 1.789 | 3 | 944 | .148 |
| log\_m | 4.921 | 3 | 518 | .002 |
| 总行驶时长 | 2.884 | 3 | 518 | .035 |
| log\_t | .712 | 3 | 518 | .545 |
| 急加速 | 9.539 | 3 | 518 | .000 |
| log\_a1 | 3.178 | 3 | 488 | .024 |
| 急减速 | 3.851 | 3 | 518 | .010 |
| log\_a2 | 6.637 | 3 | 489 | .000 |
| 急转弯 | 44.775 | 3 | 518 | .000 |
| log\_a3 | 32.086 | 3 | 449 | .000 |
| 疲劳驾驶.次数. | 7.935 | 3 | 518 | .000 |
| 夜间驾驶时长.H. | 14.816 | 3 | 518 | .000 |
| log\_ni | 3.498 | 3 | 488 | .015 |

单因素方差分析：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVA** | | | | | | |
|  | | 平方和 | 自由度 | 均方 | F | 显著性 |
| 年龄 | 组间 | 207.825 | 3 | 69.275 | 1.086 | .355 |
| 组内 | 27808.900 | 436 | 63.782 |  |  |
| 总计 | 28016.725 | 439 |  |  |  |
| 驾龄 | 组间 | 726.759 | 3 | 242.253 | 7.239 | .000 |
| 组内 | 14959.778 | 447 | 33.467 |  |  |
| 总计 | 15686.537 | 450 |  |  |  |
| 绿色出行天数 | 组间 | 2643.620 | 3 | 881.207 | 4.889 | .002 |
| 组内 | 170148.971 | 944 | 180.243 |  |  |
| 总计 | 172792.591 | 947 |  |  |  |
| 行驶天数 | 组间 | 488.284 | 3 | 162.761 | .966 | .408 |
| 组内 | 159051.527 | 944 | 168.487 |  |  |
| 总计 | 159539.810 | 947 |  |  |  |
| 油耗 | 组间 | 823461.656 | 3 | 274487.219 | 1.039 | .375 |
| 组内 | 249474404.900 | 944 | 264273.734 |  |  |
| 总计 | 250297866.500 | 947 |  |  |  |
| log\_g | 组间 | 2.821 | 3 | .940 | 1.491 | .216 |
| 组内 | 326.598 | 518 | .630 |  |  |
| 总计 | 329.419 | 521 |  |  |  |
| 碳排放.kg. | 组间 | 4183607.209 | 3 | 1394535.736 | 1.039 | .375 |
| 组内 | 1267458455.000 | 944 | 1342646.668 |  |  |
| 总计 | 1271642062.000 | 947 |  |  |  |
| log\_C | 组间 | 2.817 | 3 | .939 | 1.491 | .216 |
| 组内 | 326.230 | 518 | .630 |  |  |
| 总计 | 329.047 | 521 |  |  |  |
| 绿色出行减排.kg. | 组间 | 7148996.879 | 3 | 2382998.960 | 3.154 | .024 |
| 组内 | 713255679.000 | 944 | 755567.457 |  |  |
| 总计 | 720404675.900 | 947 |  |  |  |
| log\_jianpai | 组间 | 5.556 | 3 | 1.852 | 4.395 | .005 |
| 组内 | 263.378 | 625 | .421 |  |  |
| 总计 | 268.934 | 628 |  |  |  |
| 参与评分天数 | 组间 | 281.649 | 3 | 93.883 | .841 | .472 |
| 组内 | 57907.441 | 519 | 111.575 |  |  |
| 总计 | 58189.090 | 522 |  |  |  |
| 驾驶行为均分 | 组间 | 6574.161 | 3 | 2191.387 | 4.387 | .005 |
| 组内 | 258759.964 | 518 | 499.537 |  |  |
| 总计 | 265334.125 | 521 |  |  |  |
| 平均速度.KM.H. | 组间 | 668.820 | 3 | 222.940 | 4.326 | .005 |
| 组内 | 26694.789 | 518 | 51.534 |  |  |
| 总计 | 27363.609 | 521 |  |  |  |
| 总行驶里程 | 组间 | 10299327.970 | 3 | 3433109.325 | .219 | .883 |
| 组内 | 14805205010.000 | 944 | 15683479.880 |  |  |
| 总计 | 14815504340.000 | 947 |  |  |  |
| log\_m | 组间 | 5.941 | 3 | 1.980 | 1.847 | .138 |
| 组内 | 555.476 | 518 | 1.072 |  |  |
| 总计 | 561.417 | 521 |  |  |  |
| 总行驶时长 | 组间 | 174098.794 | 3 | 58032.931 | 2.023 | .110 |
| 组内 | 14862043.900 | 518 | 28691.204 |  |  |
| 总计 | 15036142.690 | 521 |  |  |  |
| log\_t | 组间 | 2.658 | 3 | .886 | 1.646 | .178 |
| 组内 | 278.910 | 518 | .538 |  |  |
| 总计 | 281.568 | 521 |  |  |  |
| 急加速 | 组间 | 67599342.840 | 3 | 22533114.280 | 3.885 | .009 |
| 组内 | 3004399829.000 | 518 | 5799999.671 |  |  |
| 总计 | 3071999172.000 | 521 |  |  |  |
| log\_a1 | 组间 | 5.357 | 3 | 1.786 | 2.258 | .081 |
| 组内 | 385.898 | 488 | .791 |  |  |
| 总计 | 391.255 | 491 |  |  |  |
| 急减速 | 组间 | 46282010.610 | 3 | 15427336.870 | 1.848 | .137 |
| 组内 | 4323343307.000 | 518 | 8346222.600 |  |  |
| 总计 | 4369625318.000 | 521 |  |  |  |
| log\_a2 | 组间 | 2.207 | 3 | .736 | 1.050 | .370 |
| 组内 | 342.494 | 489 | .700 |  |  |
| 总计 | 344.700 | 492 |  |  |  |
| 急转弯 | 组间 | 360550306.000 | 3 | 120183435.300 | 12.847 | .000 |
| 组内 | 4845742363.000 | 518 | 9354714.986 |  |  |
| 总计 | 5206292669.000 | 521 |  |  |  |
| log\_a3 | 组间 | 23.217 | 3 | 7.739 | 9.735 | .000 |
| 组内 | 356.928 | 449 | .795 |  |  |
| 总计 | 380.145 | 452 |  |  |  |
| 疲劳驾驶.次数. | 组间 | 5704.419 | 3 | 1901.473 | 4.977 | .002 |
| 组内 | 197916.012 | 518 | 382.077 |  |  |
| 总计 | 203620.431 | 521 |  |  |  |
| 夜间驾驶时长.H. | 组间 | 89176.927 | 3 | 29725.642 | 4.553 | .004 |
| 组内 | 3381632.067 | 518 | 6528.247 |  |  |
| 总计 | 3470808.994 | 521 |  |  |  |
| log\_ni | 组间 | 2.628 | 3 | .876 | 1.527 | .207 |
| 组内 | 280.011 | 488 | .574 |  |  |
| 总计 | 282.639 | 491 |  |  |  |

还有一个 多重比较.doc

## 相关关系：



相关系数矩阵

1. 删除所有log\_jianpai为缺失值的行，除日期和车牌（字符串类型）做相关关系矩阵，还是有很多缺失值

**这个相关系数矩阵怕是得弃了**

1. 既有分类变量又有连续变量，用信息增益（IG）看一下。

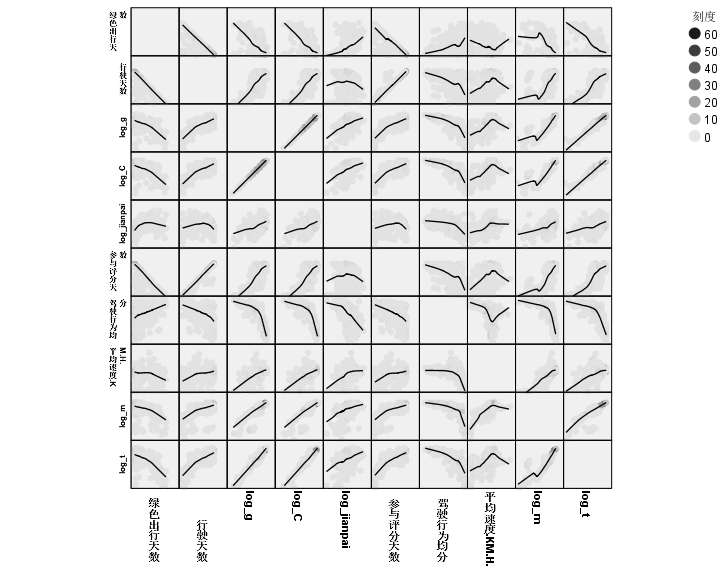
见g\_result.csv，大概长这样：



绿色出行碳减排不用看了，他就等于因变量。

有点怪，车牌号影响挺大

3、下面是几个满足正态分布、方差齐性的变量，做相关矩阵散点图：



## 大概的逐步回归一下

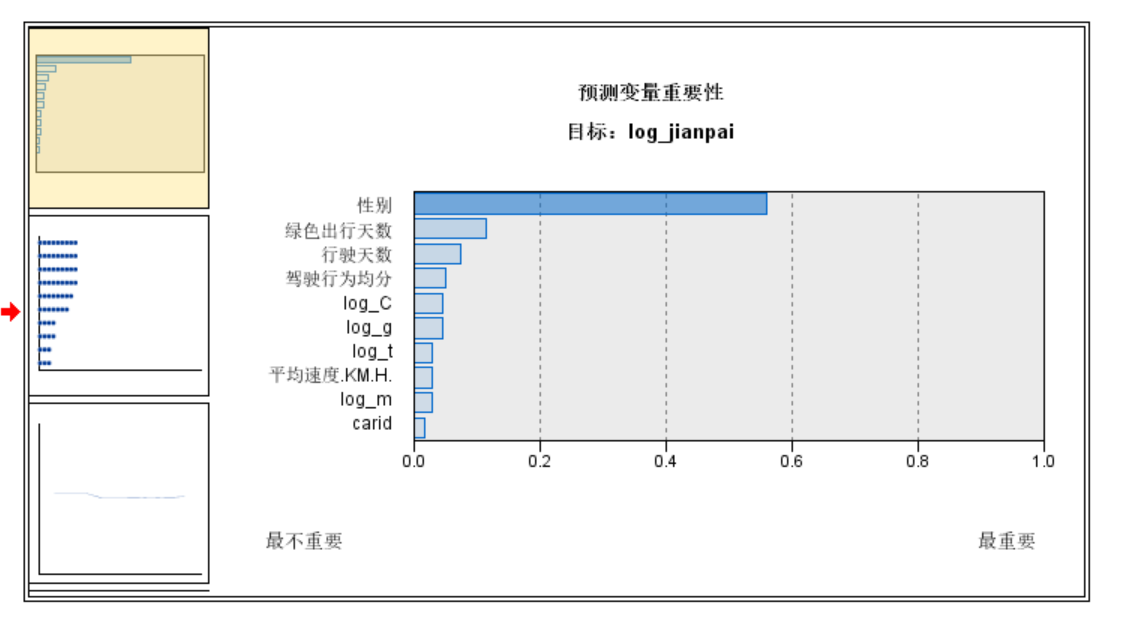
根据此前各属性的分布情况和相关关系，主要选carid、部门、是否有排行、是否有提醒、绿色出行天数、行驶天数、log\_g、log\_C、参与评分天数、驾驶行为均分、平均速度.KM.H.、log\_m、log\_t为备选特征

## 线性回归

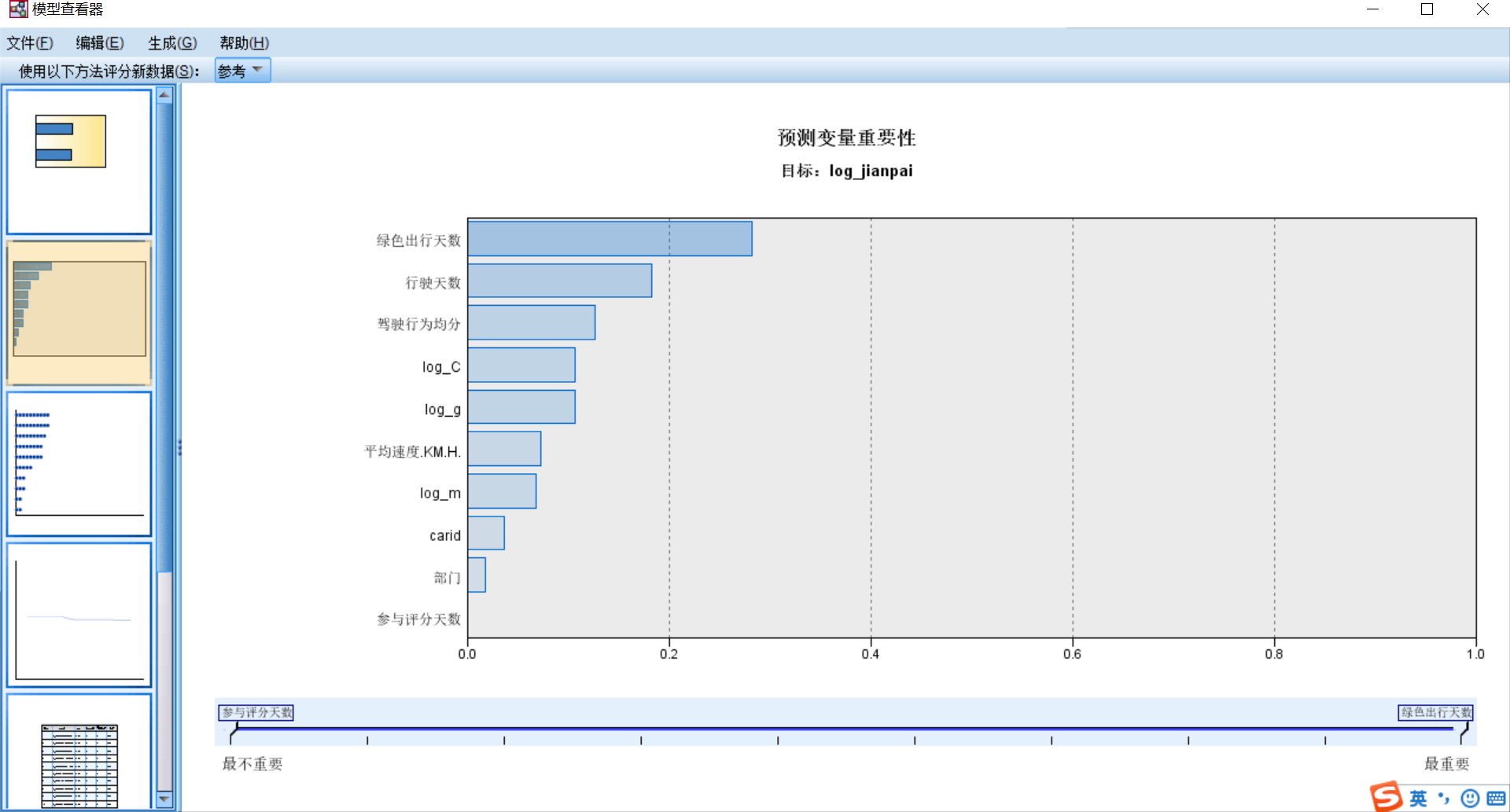
根据AIC，使用最佳子集特征选择；95%置信度，增强模型准确度方法

最后大概这样：

1. 性别作为名义变量放在特征子集里，准确度53%左右：



1. 之前性别缺的太多了，把性别移除特征子集，准确度52%左右



## 别的回归试试

晚点继续