付费搜索广告中的关键字选取
——网上机票预定业务中关键字的盈利性分析

00601014 祝垚

摘要

付费搜索广告（搜索引擎竞价排名）逐渐成为最受企业青睐的网络广告手段之一。中小型企业希望能够通过少量搜索广告吸引顾客，而大型企业则致力于减少在此类广告中的投入，提高回报率。本文通过统计方法分析了某航空公司在使用付费搜索广告过程中积累下来的数据，得到一些可以指导航空公司进行广告竞价的规律和方法。

关键词：付费搜索广告，主成分分析，分段线性回归，lasso 变量选择

目 录

1. 背景介绍

与固定付费广告相比，付费搜索广告（搜索引擎竞价排名）凭借其自主，透明，自由的优点逐渐成为最受企业青睐的网络广告手段之一。在竞价过程中企业购买特定的关键词使得网络用户能在搜索结果的显要位置看到该企业的广告。与固定付费广告不同的是，付费搜索广告需要企业为每一次用户点击付费，而在企业对于数量众多的关键字竞标之后，搜索引擎会以一定的算法决定哪些关键字会在搜索的结果中出现。

由于通过付费搜索广告点击进入网站的客户通常都具有很明确的意向，所以销售类的网站很容易使这些访客转化为消费者。对于空运行业中的企业来说，将拥有网上订购机票意向的访客转变为顾客的可能性也就更大，从而为公司来带相当的利润。当问题是这家公司需要购买相当大数量的关键字，以保证能够将大部分潜在顾客引入自己的网站，虽然这些公司通常规模比价大，资金比较雄厚，但是关键字全集的数量更是大的惊人，所以如何选择关键字就成为节省成本提高利润率的关键。以关键字“北京至上海打折机票预定”为例，它包括出发地目的地两个城市，并且有连词“至”，名词“机票”，修饰词“打折”和动词“预定”。据中国民航局（China Aviation Administration of China）的数据显示,截止至 08 年初我国就拥有 152 个大型民用机场，那么两个城市的有序组合约有 22500 个。保守的说关键字中的句式组合有 200 个（经过分析其实际数量远远高于 200），那么一家公司如果希望覆盖中国空运网络，就至少要竞标 450 万个关键字，这显然是不可能的。事实上我们将要分析的数据集也只有 3 万个关键字（范围为国内和国际航线），但是作为一般数据，其数量已经足以进行一些统计分析，找到投标这些关键字盈利情况的一点规律。

2．数据集概况与预处理

数据集为一个 30315×2 的矩阵，其第一列为关键字，第二列为公司为对应关键字评定的利润率指数，下面是前 6 行（head）的数据样例：

	关键词	利润
1	乌鲁木齐-阿克苏-机票	14.12
2	乌鲁木齐阿克苏飞机票价	9.06
3	乌鲁木齐到阿克苏-机票	-1.18
4	乌鲁木齐到阿克苏打折机票	-0.48
5	乌鲁木齐到阿克苏机票	31.94
6	乌鲁木齐-阿勒泰-机票	-1.14

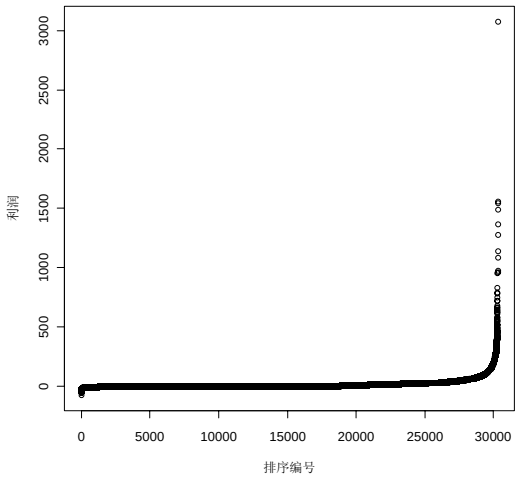
利润的分位表：

0%	5%	10%	15%	20%
-79.2	-5.3	-3.436	-2.56	-1.98
25%	30%	35%	40%	45%
-1.56	-1.28	-1.1	-0.84	-0.66
50%	55%	60%	65%	70%
-0.58	-0.53	-0.44	5.8	8.44
75%	80%	85%	90%	95%
11.53	18.14	26.68	39.192	71.69

虽然原始数据的每一个观测只是 2 维变量，但是关键字维度中包含的信息量很大，可以提取出非常多的变量，如出发地，连接词，目的地等等。对于利润做分位数表，发现其相对集中，虽然最小值和最大值分别为-79.2 和 3077.54，但有 60%的数据集中在[-5.2，5.8]之中。利润的平均值 13.75，说明有相当一部分关键字的竞标都取得了不错的收益，使得总收益平均值达到比较高的数值。但是经过查询发现 30315 个数据中有 11561 个为正数，其余 18754(61.86%)个数据为负数。对照分位数表可以发现数据在从负半轴接近原点是非常缓慢，在[-5.2，0]区间中聚集了超过 56%的观测，而更夸张的是在[-1.56，0]的区间内分布着超过 35%的数据。根据这样的情况我们再根据百分位表可以将观测根据利润分为以下几类：

编号	组别	盈利系数范围	样本个数
1	非正常负盈利	[-79.2,-26.3）	21
2	高负盈利	[-26.3，-10.3958)	约 3000
3	一般负盈利	[-10.3958，0）	约 15700
4	一般正盈利	[0，9.6286)	约 2800 个
5	高正盈利	[9.6296,22.07736)	约 3700 个
6	非正常正盈利	[22.07736，3077.54)	约 5100 个

排序后的利润图表：



根据更加详细的千分位表可以找出利润分布的大致突变点，以各段的上升速度决定其分类。其中(1)和(6)两类上升速度都非常快，而由于(1)中的样本个数过少，我们可以认为(1)中的观测均为异常值，(6)中包括一些异常值，但是观测量非常多，大多数是正常的观测，所以在剔除异常值之后，找出其中的特征仍然是很有价值的工作。高负盈利的部分(2)数据量较少，根据付费搜索广告的运作机制我们可以猜想这一部分关键字在搜索时出现的频率相对较高，导致企业向搜索引擎缴纳的费用较高，但是相对而言通过这些关键字转变为顾客的访问者又相对较少，导致盈利与成本的比值下降，从而得出较高的负盈利率。具有较低负盈利率的组(3)是包含观测量最大的组别，我们猜测这一部分关键字在搜索时出现次数较低，而访客到顾客的转化率稍低，于是出现轻微的负盈利现象。一般正盈利(4)的区间与(3)的基本对称，其形成原因大概是通过这些链接进入到航空公司网站的人转化为消费者的概率略高于平均值，较少的点击率使得这些竞标为公司带来利润。(5)则是一些收益率较高的关键字，它们转化访客的能力明显高于在搜索中出现的可能性，体现出平均值左右的收益率，但是组(5)中的数据个数并不是很多，相比组(4)的密度没有太大的提升。(6)是收益率最高的一组，跨度也是最大的一组，但是由于大多数超高收益率的观测可能是异常值，所以真正的跨度会大大的缩小。仍然这是正盈利组中数据最集中的组别，这些关键字可能是搜索的热门，但是由于访客的目的很明确，通过这些广告订购机票的人也相对最多，这导致相对高的成本和更高的收益，使得收益率至少为平均值的 1.5 倍左右。

总体来说，对于利润的分布情况我们可以看出大多数关键字处于亏损状态，但是盈利的关键字带来的巨大利润弥补了这些亏损并且将利润均值提高到一个比较高的水平，甚至超过了 75%分位点：负盈利的关键字大多成本较低，略有收入，所以使大部分变量都集中在绝对值较小的负半轴区域内，极少数的关键字拥有较大的点击量但是缺乏吸引消费者的能力，处于绝对值较大的负半轴段上。反观正盈利的关键字，它们在接近原点区域的密度远远小于负盈利的部分，而且在很大的范围内相对均匀的分散，所以在回归的过程中可能出现比较可信的结果。为了了解正负盈利的两类观测拥有的特点，可以用主成分方法初步对数据进行探索。

虽然原始数据的每一个观测只是 2 维变量，但是关键字维度中包含的信息量很大，可以提取出非常多的变量，如出发地，连接词等等。在不考虑到地理因素和语言因素的交互影响的情况下，我们可以把关键字分为两个部分，第一个部分是一个二元有序的地理位置向量，

第二个部分是句法因素向量。其中句法因素可以被一个多维 0-1 向量表示，此向量中每一个维度均代表一个词汇，如“优惠”,“打折”等等，每一个关键字的对应向量根据关键字中是否包含此词汇决定取值，包括则取 1，反之为 0.。这样我们就把句法因素完全的表示成了数学语言，但是因为在回归过程中，如果自变量都只有两个取值，那么模型的运算结果就很不自由。假设有 N 个自变量，那么自变量向量的所有取值也只有 2^N 个，然而最终因变量利润取值范围为实数，这使得回归或分类非常不准确，所以通常情况下会把二分类的变量当作哑变量(dummy variable)来处理。

现在我们需要从地理因素中提取出一些可以取实数值的变量作为回归的因变量，那么我们应该给每一个出发地或者目的地附一个权值来体现出地理方面的信息。但是数据中只有利润是取实数值的变量，简单易行的办法是将相同出发地点的关键字的利润取平均值，赋给出发地，关键字中包括的另外一条数字信息是这个出发地点在所有关键字中作为起点的出现次数，这样对于每一个观测，我们能够提取出一个 4 维地理特征向量(出发地平均利润,目的地平均利润,出发地出现次数,目的地出现次数)，然后用这 4 个取实值的变量就可以对因变量利润做回归分析了。

提取地理变量后的数据样例：

关键字	利润	出发地	目的地	出发地平均利润	目的地平均利润	出发地出现次数	目的地出现次数
青岛到杭州打折飞机票	-1.28	青岛	杭州	19.94048	14.36427	901	728
青岛到杭州打折机票	50.64	青岛	杭州	19.94048	14.36427	901	728
青岛到杭州的飞机票	-8.64	青岛	杭州	19.94048	14.36427	901	728
青岛到杭州的机票	68.5	青岛	杭州	19.94048	14.36427	901	728
青岛到杭州飞机票	-7.7	青岛	杭州	19.94048	14.36427	901	728
青岛到杭州机票	219.7	青岛	杭州	19.94048	14.36427	901	728

对于语言因素变量的具体处理需要先提取所有的词汇，共有如下几类：

1. 介于两个地理位置之间的连接词：“到”,“去”,“飞”,“至”,“-”
2. 主要名词：“机票”,“飞机票”
3. 主要修饰词：“便宜”,“特价”,“折扣”,“打折”,“低价”,“优惠”,“特惠”
4. 补充名词：“价格”,“价”
5. 路线修饰词：“往返”,“来回”
6. 动词：“查询”,“预定”,“预订”,“定”,“订”
7. 助词：“的”

以上组别可以进行组合，加上出发地和目的地可以形成最终的关键字，为了探究 24 个词汇的作用可以将其作为虚拟变量进行回归，之后可以加入这些词汇的交互影响进行回归，得到词汇组合对于利润率的影响。

3. 利润与地理因素变量的主成分分析与回归

首先对数据进行标准化，然后对于 2-6 组分别做主成分分析，由于组 6 包含一些极大的数据，我们又将这些过大的数据从组 6 中分离出来形成组 7。

高负盈利：

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Standard deviation	1.4595911	1.0858290	0.9660147	0.54458465	0	0	0	0
Proportion of Variance	0.4693365	0.2597435	0.2055840	0.06533601	0	0	0	0
Cumulative Proportion	0.4693365	0.7290800	0.9346640	1.00000000	1	1	1	1

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
关键字								1.000
利润							1.000	
出发地						1.000		
目的地					1.000			
出发地平均利润	0.285	0.861	-0.422					
目的地平均利润	-0.669	0.354	0.247	-0.605				
出发地重复数量	0.335	0.313	0.872	0.169				
目的地重复数量	-0.600	0.189		0.777				

一般负盈利：

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Standard deviation	1.4464187	0.9637858	0.8840008	0.40269940	0	0	0	0
Proportion of Variance	0.5276974	0.2342922	0.1971070	0.04090335	0	0	0	0
Cumulative Proportion	0.5276974	0.7619896	0.9590967	1.00000000	1	1	1	1

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
关键字								1.000
利润							1.000	
出发地						1.000		
目的地					1.000			
出发地平均利润	0.323	0.846	0.423					
目的地平均利润	-0.595	0.338	-0.203	-0.700				
出发地重复数量	0.404	0.280	-0.870					
目的地重复数量	-0.615	0.302	-0.152	0.712				

一般正盈利：

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Standard deviation	1.4133849	0.9661415	0.8881492	0.43333155	0	0	0	0
Proportion of Variance	0.5112141	0.2388710	0.2018616	0.04805323	0	0	0	0
Cumulative Proportion	0.5112141	0.7500851	0.9519468	1.00000000	1	1	1	1

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
关键字								1.000
利润							1.000	
出发地						1.000		
目的地					1.000			
出发地平均利润	0.237	0.944	-0.230					
目的地平均利润	-0.612	0.236	0.295	-0.694				
出发地重复数量	0.398	0.120	0.906					
目的地重复数量	-0.641	0.198	0.195	0.716				

高正盈利：

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Standard deviation	1.377948	0.9312530	0.9048278	0.42304462	0	0	0	0
Proportion of Variance	0.504494	0.2304231	0.2175316	0.04755135	0	0	0	0
Cumulative Proportion	0.504494	0.7349170	0.9524486	1.00000000	1	1	1	1

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
关键字								1.000
利润							1.000	
出发地						1.000		
目的地					1.000			
出发地平均利润	0.224	0.954	-0.199					
目的地平均利润	-0.614	0.219	0.317	-0.689				
出发地重复数量	0.404		0.906					
目的地重复数量	-0.640	0.182	0.199	0.719				

最高正盈利：

Importance of components:

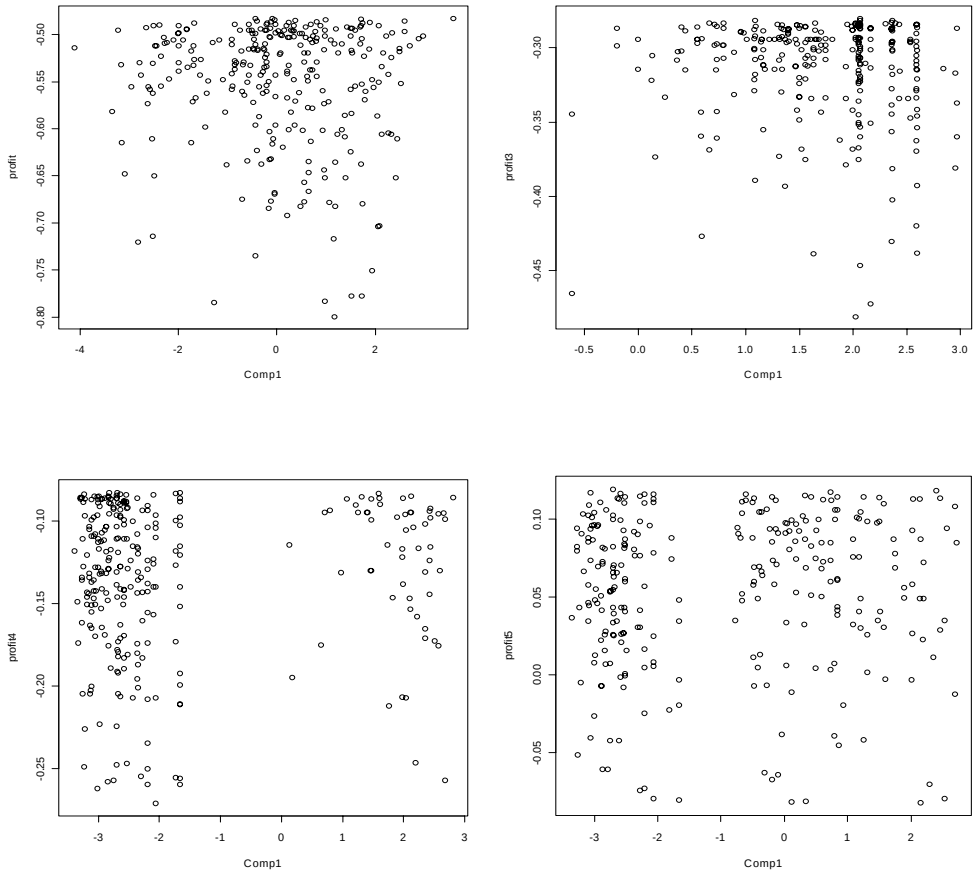
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Standard deviation	1.3865975	0.9338802	0.8992060	0.50822709	0	0	0	0
Proportion of Variance	0.4978836	0.2258444	0.2093849	0.06688713	0	0	0	0
Cumulative Proportion	0.4978836	0.7237279	0.9331129	1.00000000	1	1	1	1

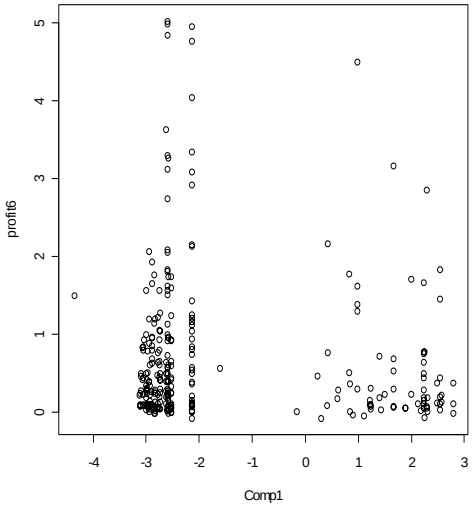
Loadings:

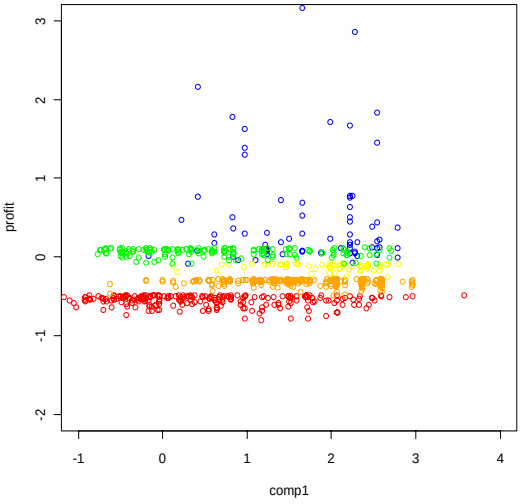
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
关键字								1.000
利润							1.000	
出发地						1.000		
目的地					1.000			
出发地平均利润	0.125	0.991						
目的地平均利润	-0.623	0.116	0.389	-0.669				
出发地重复数量	0.413		0.900	0.137				
目的地重复数量	-0.652		0.189	0.730				

通过上面几组的主成分分析，我们可以看到关于地理因素的主成分基本模式是固定的，第一主成分均为 4 项的线性组合，值得注意的是与出发地有关的两项均具备正系数，而与目的地相关的两项均有负系数，而且关于目的地的系数绝对值较大，也就是说目的地对于第一主成分的贡献更大。第一个主成分大致占到总信息量的一半左右，反映的信息主要是出发地目的地之间的差异，并且对于目的地的情况更加敏感。另外，将第一主成分中的四项系数相加，可以发现收益率绝对值越高的组别其系数和的绝对值也越高，可以理解为收益率离原点偏差越大，其起始点与终点的差异越能决定最终收益率的大小。

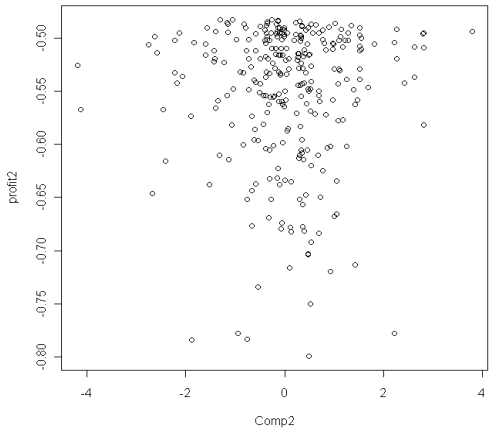
第二主成分的构成为至少包括两项的正系数线性组合，其中出发地平均利润的系数大概在[0.8，1]之中，而其他三个变量的系数都相对较小。随着组别收益率的提高，这些系数呈现比较明显的下降。在第三主成分中则是出发地重复数量的系数绝对值最低为 0.8，远远高于其他三项。

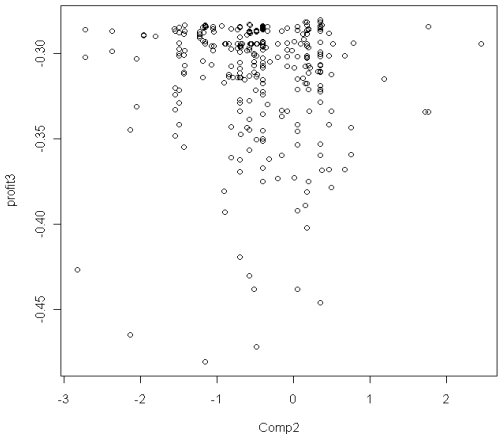


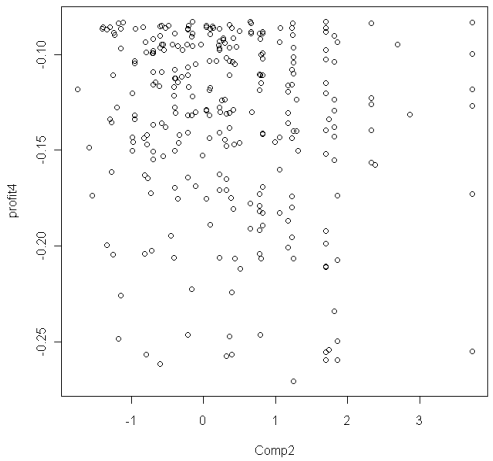


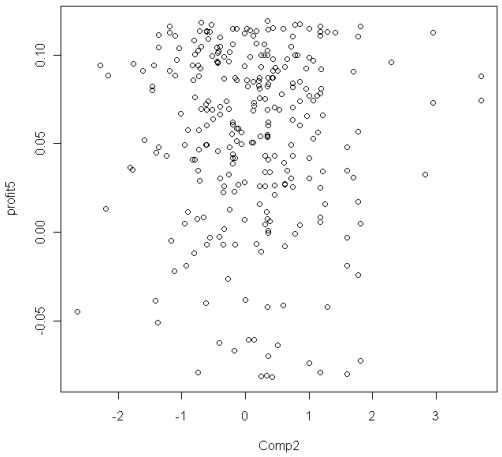


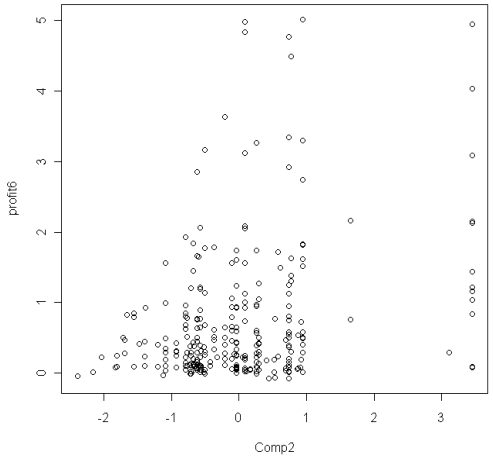
通过主成分一与利润的二维图，我们可以看出负盈利组的主成分比较倾向于集中分布，而正盈利组有分为左右两部分的趋势。最后一组图将各组的主成分一进行比较，可以认为此主成分与利润没有线性关系。

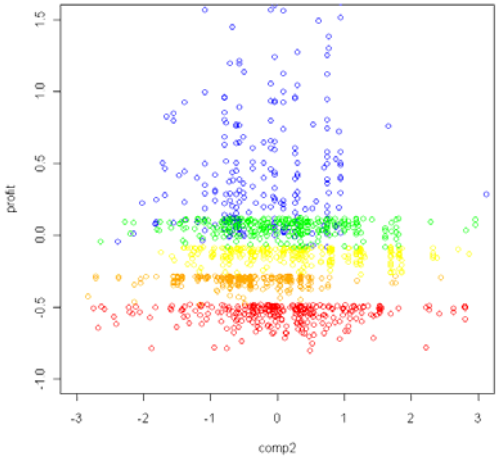




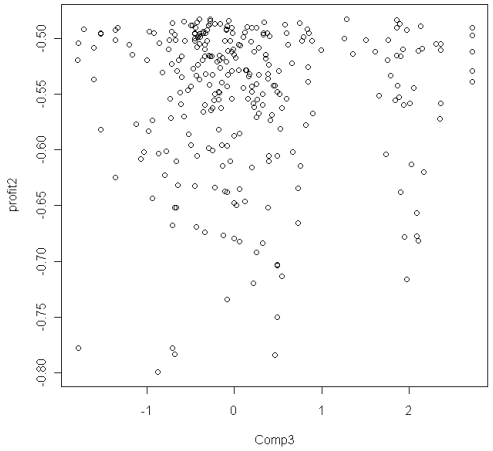


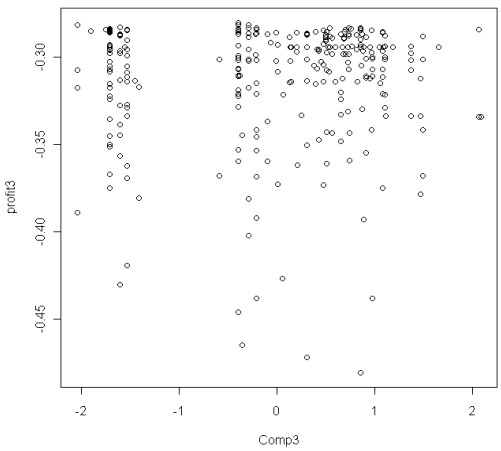


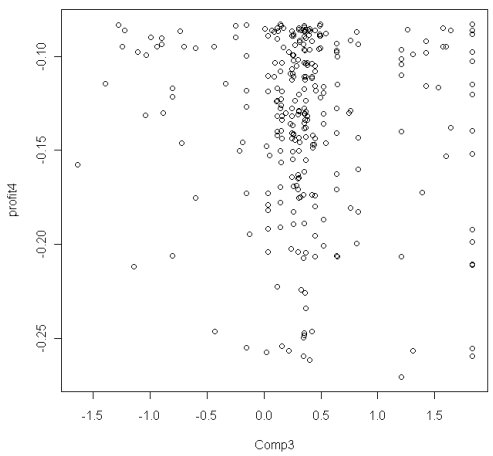


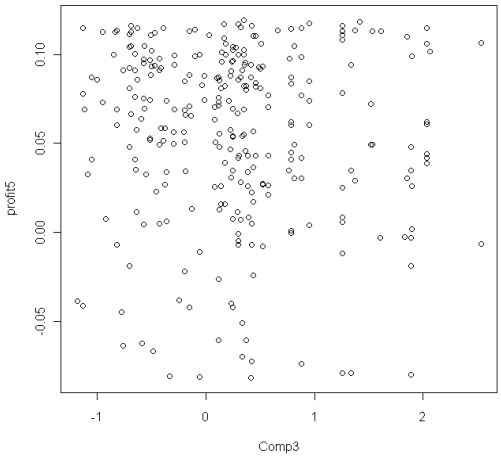


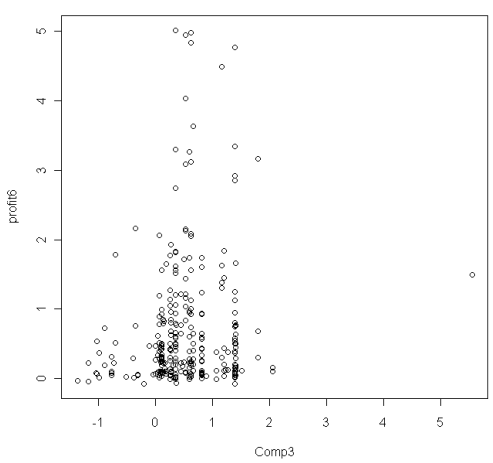
组 2，3，6 的第二主成分分布较集中，而 4，5 两组的分布较分散，通过比较各组的主成分分布，同样可以看到这些主成分对于利润没有较好的线性关系。

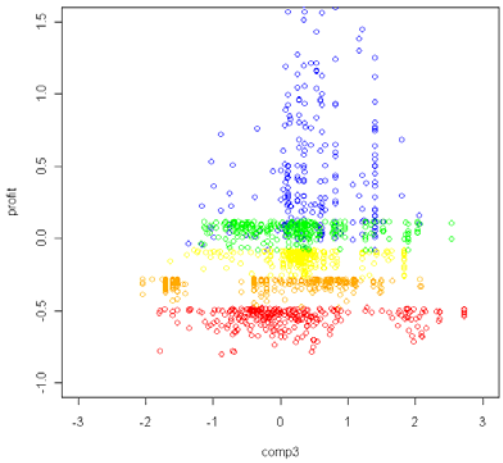








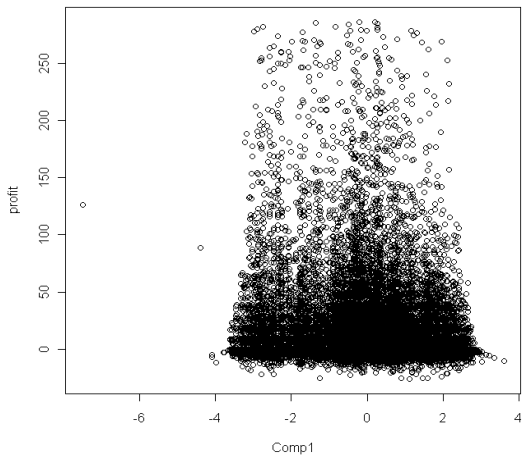


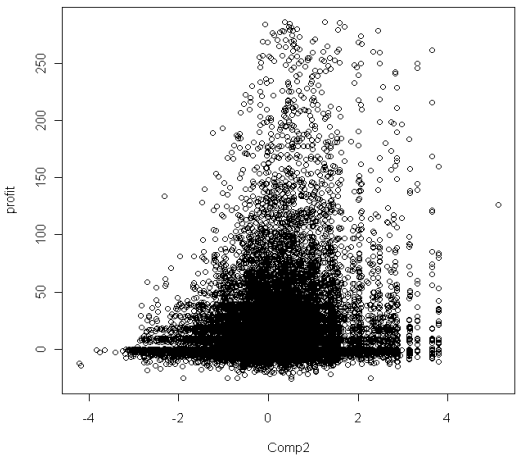


第三主成分对于利润也没有可视的回归关系，这样分别用主成分对利润进行回归可能会得到比较差的结果。其中的原因一是以上的主成分是基于我们从关键字和利润中提取的 4 个变量，其中包含了部分关于利润的信息，但是同时也有更多关于地理方面的信息，所以其中的主成分可能与信息无关。原因二是每组主成分的系数都有差异，可能使各组数据向原点靠拢，破坏了应有的关系。为了消除原因二的干扰，做出数据整体的主成分分析（组 2 到 6）进行比较

Importance of components:

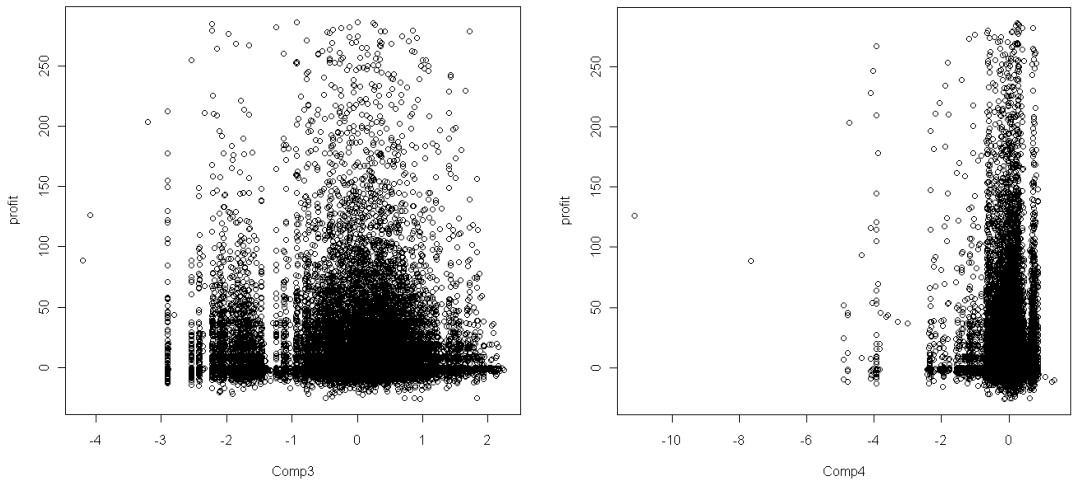
	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	1.4298161	0.9762306	0.8928436	0.43852063
Proportion of Variance	0.5127767	0.2390412	0.1999488	0.04823341
Cumulative Proportion	0.5127767	0.7518178	0.9517666	1.00000000





Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4
出发地平均利润	0.252	0.880	0.403	
目的地平均利润	-0.615	0.297	-0.238	-0.691
出发地重复数量	0.395	0.284	-0.871	
目的地重复数量	-0.635	0.239	-0.151	0.719

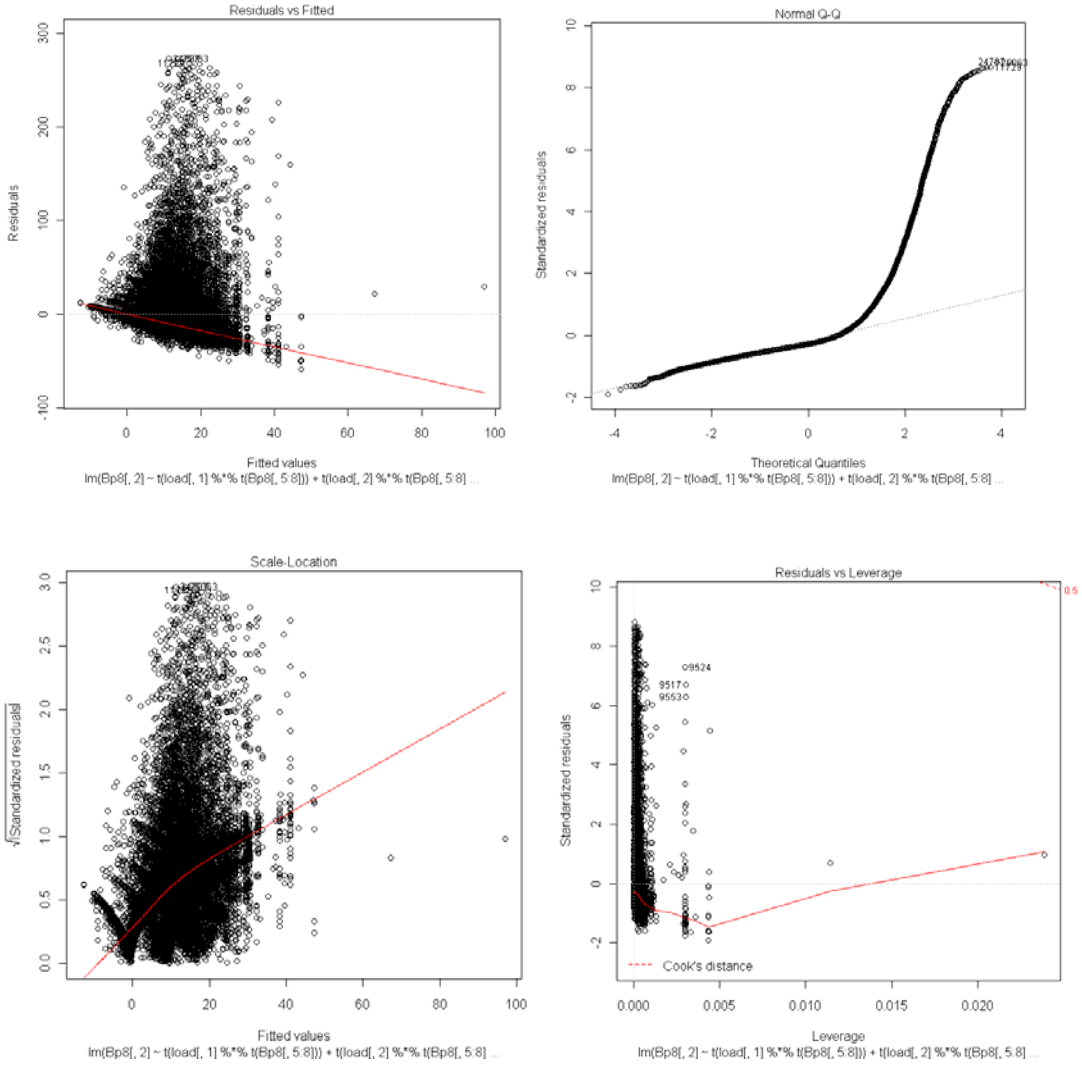


整体的主成分分析与分组后的大致相同，其与利润也没有明显的线性关系，下面用主成分对于利润做简单线性回归

```
     Residuals:
    Min      1Q  Median      3Q     Max
-58.892 -14.168  -8.410   1.496 272.959

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  11.8017     0.1781  66.271  < 2e-16 ***
Comp1                        -1.9631     0.1245 -15.762  < 2e-16 ***
Comp2                         5.8617     0.1824  32.134  < 2e-16 ***
Comp3                        -1.6196     0.1995  -8.120 4.83e-16 ***
Comp4                        -3.0356     0.4061  -7.475 7.93e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.93 on 30166 degrees of freedom
Multiple R-squared: 0.04444,    Adjusted R-squared: 0.04431
F-statistic: 350.7 on 4 and 30166 DF,  p-value: < 2.2e-16
```

回归结果模型的 P 值非常小，可以认为模型中的各变量在真实模型中都很重要。每个变量的 P 值很小，说明这些变量都显著的影响着因变量。相对于系数的绝对值而言，对于系数估计的标准误也较小，说明系数的估计准确。四个回归系数之中只有第二主成分的系数大于 0，而且绝对值明显高出其他三个主成分，前面提到第二主成分是四个变量的正线性组合，说明了出发地和目的地的平均水平和受关注程度都与最终的利润正相关。常数项为 11.8017 小于利润的均值，说明 4 个主成分的综合作用还是正面的。主成分 4 在负系数项里系数绝对值最大，其结构是目的地重复数量-目的地平均利润，而系数大小相似。这反应了目的地的情况，因为数据进行过标准化，所以可以理解为其出现的频繁程度和利润均值在所有城市中的地位相比较，如果平均利润高或出现频率低则会降低住成分 4，升高利润，反之则降低利

润。第一主成分如前文所述，描述出发地与目的地之间的差异，结果回归系数小于 0，可以理解为两地之间差距越大对于最终利润率越不利。



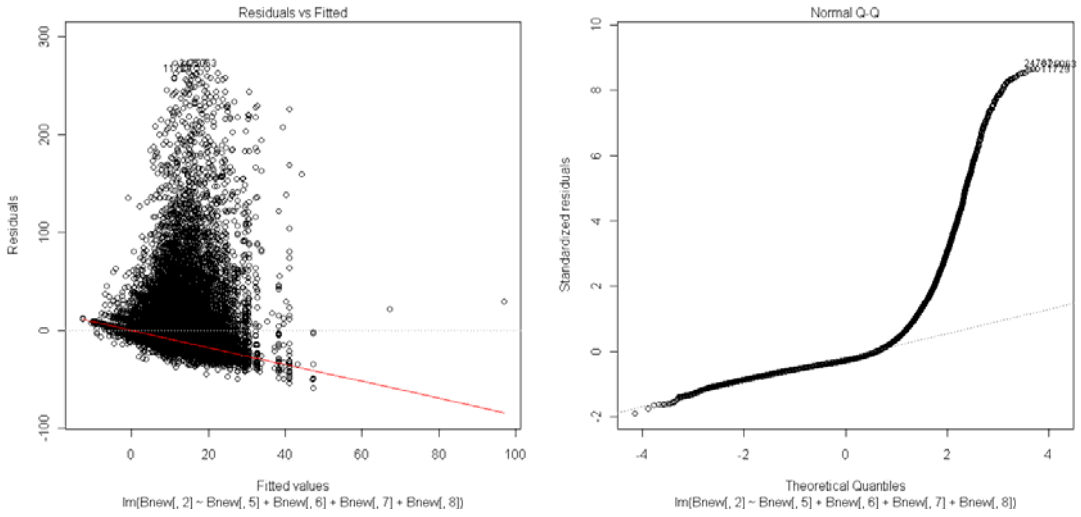
但是注意到主成分的线性回归模型中的 R-squared 统计量非常的小，这个统计量是回归平方和与总方差的比值，代表模型对于真是情况的拟合程度。主成分线性回归模型的 R-squared 统计量只有 0.04 左右，并没有很好的拟合数据，但是其线性模型和回归系数均显著。从回归的 normal Q-Q 图中我们可以看到模型对于数据左半部份拟合较好，但是右半部分明显有偏差，normal Q-Q 中曲线的斜率表示假设服从正态分布，其标准差的值，也就是数据的分散程度。所以我们将对于利润较大的一部分数据做模型上的改进，由于图中右半部分接近于直线，所以我们考虑做分段的回归。

为了便于比较我们先做出 4 个地理因素直接回归利润的结果（已经去除异常值）

```
Residuals:
    Min      1Q  Median      3Q     Max
-58.892 -14.168  -8.410   1.496 272.959

Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.151e+01  6.744e-01 -17.070   <2e-16 ***
Bnew[, 5]    7.418e-01  3.412e-02  21.737   <2e-16 ***
Bnew[, 6]    7.139e-01  3.937e-02  18.131   <2e-16 ***
Bnew[, 7]    1.911e-03  1.741e-04  10.975   <2e-16 ***
Bnew[, 8]    1.510e-03  6.562e-04   2.301   0.0214 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.93 on 30166 degrees of freedom
Multiple R-squared: 0.04444,    Adjusted R-squared: 0.04431
F-statistic: 350.7 on 4 and 30166 DF,  p-value: < 2.2e-16
```

其中可以发现目的地重复数量这一变量比其他各项的显著程度差，此结果大致与主成分回归相同，关于模型和变量的 P 值很小，均可以认为显著，但是对于利润的回归较差。关于残差的散点图和 Q-Q 图也都类似。

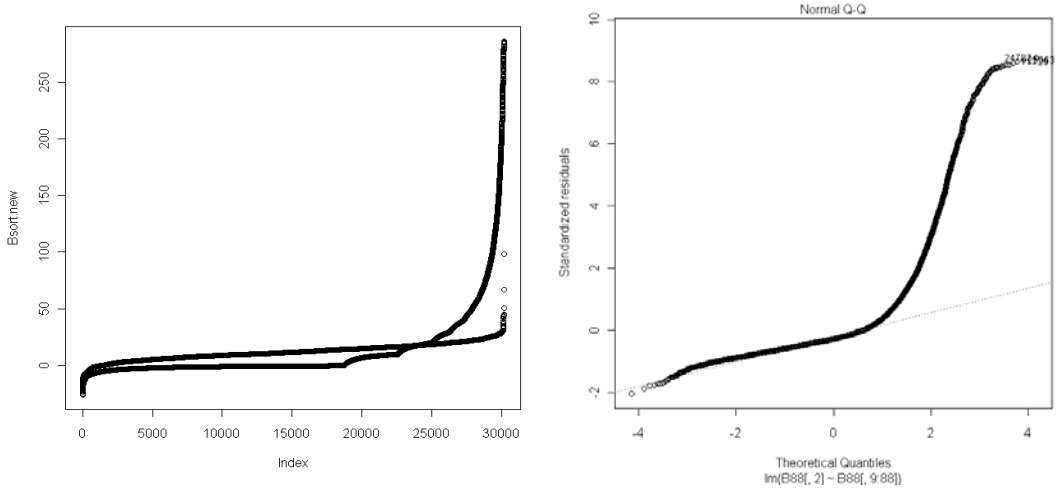


下面做分段回归，由于不能明确分界点，只能先假定一些分位数作为分界点再加以选择。先将 4 个地理变量的每个 10 分位点作为一个分割点进行分段回归，直接回归得到如下结果：

```
Residuals:
    Min      1Q  Median      3Q     Max
-61.884 -14.593  -8.125   1.839 269.989

Coefficients: (8 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.851e+01  1.884e+00  -9.828  < 2e-16 ***
B88[, 9:88]26 -1.109e+02  4.172e+01  -2.659  0.00784 **
B88[, 9:88]41  6.031e-01  3.001e-01   2.010  0.04445 *
B88[, 9:88]51  1.103e+00  3.575e-01   3.084  0.00204 **
B88[, 9:88]61  2.794e-02  1.465e-02   1.908  0.05640 .
```

```
B88[, 9:88]66  1.300e-01  5.065e-02   2.567  0.01025 *

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.85 on 30098 degrees of freedom
Multiple R-squared: 0.05137,    Adjusted R-squared: 0.0491
F-statistic: 22.64 on 72 and 30098 DF,  p-value: < 2.2e-16
```

上面列出的带有显著性的变量，即常数项，出发地重复次数大于 50%分位数时的常数加成，出发地平均利润，目的地平均利润，出发地出现次数的线性组合，还有出发地出现次数大于 50%分位数之后的附加项。与原始回归相比，其模型的 P 值都很小，说明模型显著，残差的标准误没有明显变化。F 统计量明显的减少，原因是变量增多导致拟合的误差减小。R-squared 统计量的大小由 0.44 提高到了 0.51，虽然提高比较明显，但是还没有达到模型充分拟合的标准



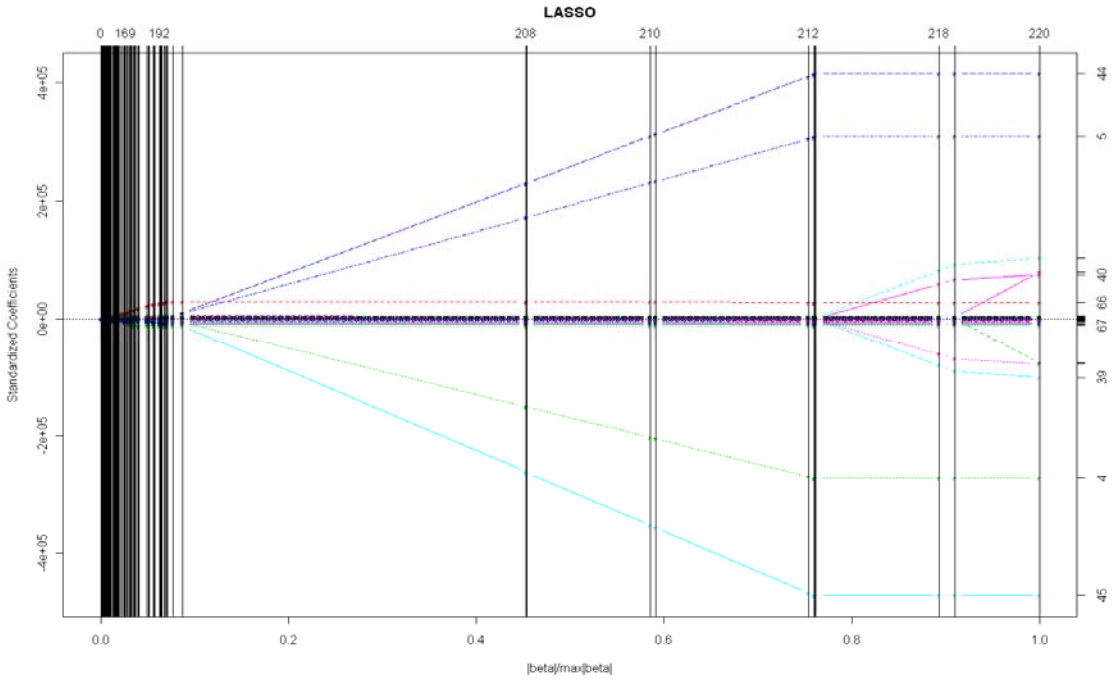
从上面的 Q-Q 图中可以看到新的拟合仍然没有解决在均值右侧数据分散的问题，而从左边的真实值与预测值的对照图中我们可以看出回归的主要问题集中在对于较大的数据的预测上。（左图中较平的一条为预测值排序后的散点图，尾端较高的一条为真实值排序后的散点图）由于利润率在超过 50 之后明显上升加速，可以尝试在某一个分位数后加入指数函数的成分进行回归。

但是通过对于所有 80 个变量的回归系数估计，我们也可以看到随着 4 个变量的不断增大，对于回归的影响

```
Coefficients: (8 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.851e+01  1.884e+00  -9.828  < 2e-16 ***
B88[, 9:88]1         NA         NA      NA       NA
B88[, 9:88]2   1.013e+00  6.548e+00   0.155  0.87702
B88[, 9:88]3  -1.438e+01  5.702e+01  -0.252  0.80084
B88[, 9:88]4  -3.649e+03  9.296e+03  -0.393  0.69465
```

```
B88[, 9:88]5    3.662e+03  9.300e+03   0.394  0.69373
B88[, 9:88]6   -3.731e+01  4.059e+01  -0.919  0.35811
B88[, 9:88]7   -1.588e+01  7.168e+01  -0.221  0.82473
B88[, 9:88]8    5.740e+01  9.162e+01   0.626  0.53103
B88[, 9:88]9   -1.465e+00  6.643e+01  -0.022  0.98240
B88[, 9:88]10 -1.900e+00  3.640e+00  -0.522  0.60163
B88[, 9:88]41   6.031e-01  3.001e-01   2.010  0.04445 *
B88[, 9:88]42 -2.628e-02  7.419e-01  -0.035  0.97174
B88[, 9:88]43   1.197e+00  5.103e+00   0.235  0.81455
B88[, 9:88]44   3.108e+02  7.914e+02   0.393  0.69455
B88[, 9:88]45 -3.119e+02  7.917e+02  -0.394  0.69363
B88[, 9:88]46   2.651e+00  2.990e+00   0.887  0.37518
B88[, 9:88]47   8.708e-01  5.031e+00   0.173  0.86258
B88[, 9:88]48 -3.605e+00  5.825e+00  -0.619  0.53604
B88[, 9:88]49 -1.597e-02  3.976e+00  -0.004  0.99680
B88[, 9:88]50          NA         NA      NA       NA
B88[, 9:88]80   4.130e-03  6.378e-03   0.648  0.51725
```

其中项 1 至 10 和 41 至 50 对应变量出发地平均利润，11 至 20 和 51 至 60 对应目的地平均利润，21 至 30 和 61 至 70 对应出发地重复次数，31 至 40 和 71 至 80 对应目的地重复次数。超过一定分位数的加成并不像想象中一样，而是正负相间隔，高分位数的加成总和也上下波动。下面用 LARS 变量选择方法筛选变量，比较如果变量较少时的回归情况。



变量选择的过程如下所示：

```
Var  52 42 53 43 44 25 -43 4 54 15 -44 -54 51 28  7 14 23 22  8 13
Step  1  2  3  4  5  6   7 8  9 10  11  12 13 14 15 16 17 18 19 20

Var    2 41 -53 34 33 37 17 12 -42 18 36  5  3 -7 -52 40 19 73 10 -36
Step  21 22  23 24 25 26 27 28  29 30 31 32 33 34  35 36 37 38 39  40

Var  32 -19 50 -3 70 -10 77 -37 -5 27 69  6 39 20 46 71 24 -6 56 35 ……
Step 41  42 43 44 45  46 47  48 49 50 51 52 53 54 55 56 57 58 59 60 ……
```

（全表详见附录）

Step表示进行选择的步数，var表示有所变化的变量，如果符号为正则表示被选入模型当中，反之则表示从模型中剔除。模型在220步左右达到稳定。下面是每步的残差平方和和Cp统计量。

```
Call: lars(x = B88[, 9:88], y = B88[, 2], type = "lasso")
    Df      Rss       Cp
0    1 30205077 1558.748    1    2 29876444 1215.548    2    3 29554626  879.506
3    4 29308522  622.996    4    5 29287987  603.425    5    6 29254883  570.653
6    7 29204547  519.779    7    6 29173792  485.474    8    7 29134239  445.927
9    8 29125877  439.144   10    9 29122324  437.412   11    8 29103879  416.037
12   7 29033470  340.079   13    8 29030994  339.478   14    9 29014913  324.586
15  10 28991923  302.437   16   11 28975248  286.921   17   12 28972771  286.319
18  13 28924449  237.561   19   14 28887473  200.722   20   15 28875387  190.026
21  16 28869682  186.033   22   17 28862153  180.125   23   16 28856597  172.289
24  17 28806283  121.439   25   18 28804768  121.848   26   19 28795106  113.698
27  20 28788343  108.595   28   21 28772626   94.085   29   20 28746001   64.117
30  21 28739574   59.367   31   22 28738554   60.295   32   23 28735010   58.573
33  24 28732767   58.217   34   23 28729061   52.323   35   22 28712596   33.029
36  23 28709852   32.147   37   24 28705609   29.690   38   25 28704588   30.617
39  26 28702913   30.857   40   25 28700961   26.807   41   26 28700323   28.137
42  25 28697184   22.840   43   26 28696400   24.016   44   25 28694785   20.320
45  26 28694123   21.625   46   25 28693070   18.518   47   26 28691967   19.360
48  25 28691110   16.459   49   24 28690595   13.919   50   25 28690282   15.590
60  33 28685737   26.816   70   37 28679108   27.853   80   43 28676261   36.861
90  43 28673451   33.910  100   45 28672129   36.522  110   47 28670952   39.285
120 49 28669338   41.590  130   51 28666903   43.033  140   51 28665874   41.952
150 55 28662881   46.807  160   59 28660910   52.737  170   61 28658625   54.337
180 67 28656380   63.978  190   69 28654036   65.517  200   71 28653691   69.154
201 72 28653691   71.154  202   73 28653691   73.154  203   72 28653691   71.154
204 73 28653688   73.151  205   72 28653688   71.151  206   73 28653684   73.147
```

207	72	28653684	71.147	208	73	28653573	73.031	209	72	28653573	71.031
210	73	28653554	73.010	211	72	28653553	71.009	212	73	28653544	73.000
213	72	28653544	71.000	214	73	28653544	73.000	215	72	28653544	71.000
216	73	28653544	73.000	217	72	28653544	71.000	218	73	28653544	73.000
219	72	28653544	71.000	220	73	28653544	73.000				

（全表详见附录）

一般来说 Cp 统计量在小于变量数时认为模型是比较符合实际的，当筛选进行到 41 步之后，Cp 统计量一直保持不超过变量数目，可以认为 40 步之后的模型均何以接受。残差项的结果依旧比较差。注意到最初被选进模型的几项都在后面被除去，这可能是因为后面的变量更好的包括了这些变量中的信息。而最终的回归中有些项的系数明显高于其他项，其中第 4，5，26，44，45 的系数为绝对值最大的 5 个系数，他们分别表示超过 30%和 40%分位数的出发地平均利润的一次项和常数项，以及出发地重复数在超过 50%分位数之后的附加常数项，在数据逐渐增大的过程中自变量对于因变量的影响主要取决于编号较小的项，所以我们可以认为出发地平均利润在超过 30%分位数后会给利润带来额外的正影响，伴随有一个较小的常数惩罚。同样的，出发地重复次数在超过 50%分位数之后会带来一些负面影响，在从某出发地网上预订机票的顾客数量一定的情况下，过多的关键字必然会导致浪费，其总体利润均值会下降，购买其中某些关键字可能出现负盈利的情况。

为了进一步体现地理变量数值增加过程中对于利润的加速作用，我们又加入了指数因素，但是回归结果没有明显改进（部分系数估计详见附录）。我们认为直线的分段回归可以描述曲线的大致走向。

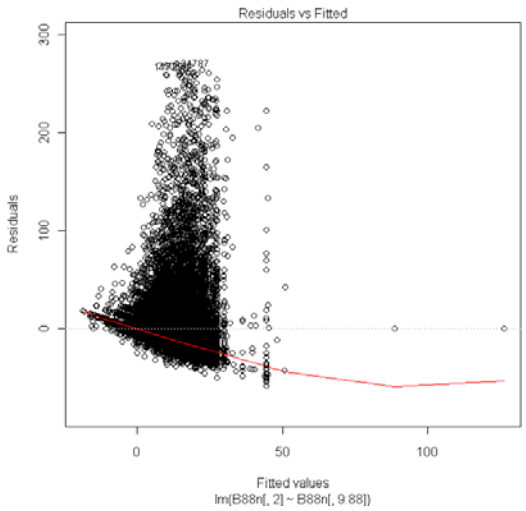
观察到出发地和目的地平均利润过于集中，其分界点分别为：-9.765000 ，7.668175，10.821253，11.583718，11.750615，13.324775，14.333917，15.995046 ，17.979273，19.733913，30.553709 和-11.500000，4.632778，7.650370，9.551698，11.065882，12.856607，14.364272，16.027289，20.398889，24.132805，126.250000 其中有些分界点之间的间隔过小，所以考虑重新规定分界点，可能能够增加对于较大数据的预测的准确性，尝试使用数值上面的等间距方法去分界点，重新规定后 4 项地理变量的分界点分别为：(-10,-6,-2,2,6,10,14,18,22,26,30)，(-12,0,12,24,36,48,60,72,84,96,108) ， (0,300,600,900,1200,1500,1800,2100,2400,2700,3000) ，(0,200,400,600,800,1000,1200,1400,1600,1800,2000)。重新回归的结果如下：

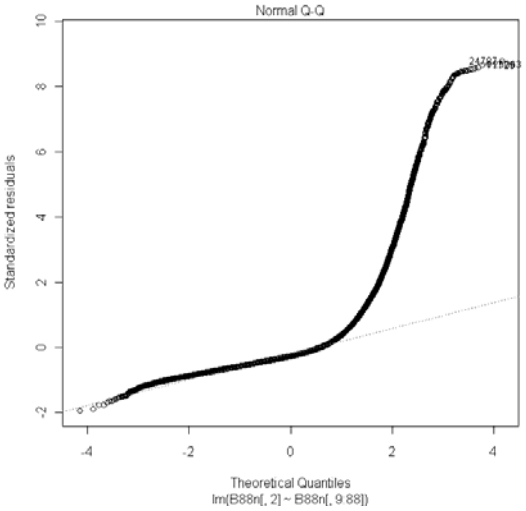
```
Residual standard error: 30.85 on 30115 degrees of freedom
Multiple R-squared: 0.0509,      Adjusted R-squared: 0.04916
F-statistic: 29.36 on 55 and 30115 DF,  p-value: < 2.2e-16
```

（具体模型估计详见附录）

依然没有对于R-squared统计量的明显改进，可能是地理提取出的地理变量并不能很好的包含整个与利润率相关的地理信息，如果需要能够接近真实值的回归，应该引入或者获取更多的变量。

重新分段后的回归图像：


Residuals vs Fitted
Residuals
Fitted values
lm(B88n[, 2] ~ B88n[, 9:88])

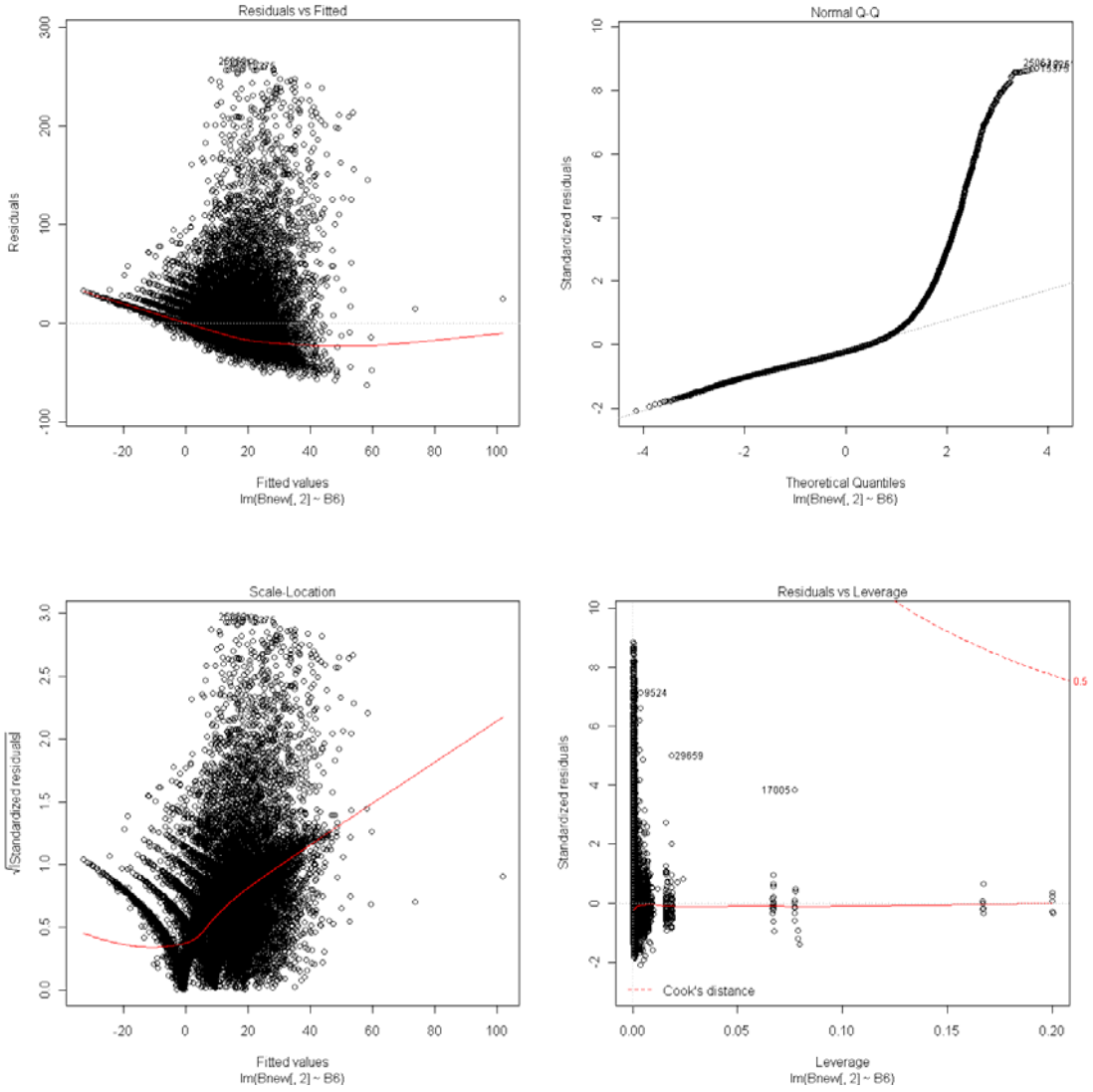

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(B88n[, 2] ~ B88n[, 9:88])

4. 对于关键字的选择

如前文所述对于关键字作用的回归以地理变量的回归作为基础，加入虚拟变量的影响。由于没有在上一步骤中得到更好的模型，只能用基本的线性模型进行回归。对于每一个在关键字出现过的词汇，我们都赋予一个虚拟变量进行回归，这样因变量共有 4 个取实值的变量和 24 个虚拟变量。先做出简单的线性回归结果：

```
Residuals:
    Min      1Q  Median      3Q     Max
-62.364 -14.957  -6.605   4.313 265.467

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.850e+01  2.792e+00  -6.624 3.55e-11 ***
B61                 NA         NA      NA       NA
B62          8.564e-01  3.345e-02  25.600  < 2e-16 ***
B63          7.061e-01  3.829e-02  18.442  < 2e-16 ***
B64          3.695e-03  1.804e-04  20.490  < 2e-16 ***
B65          6.487e-03  6.540e-04   9.920  < 2e-16 ***
B66         -5.773e+00  4.400e-01 -13.122  < 2e-16 ***
B67          6.840e+00  4.433e-01  15.431  < 2e-16 ***
B68         -1.380e+01  8.704e-01 -15.854  < 2e-16 ***
B69         -6.175e-01  6.012e-01  -1.027  0.30436
B610        -8.890e+00  6.362e-01 -13.972  < 2e-16 ***
B611        -2.003e+01  2.229e+00  -8.982  < 2e-16 ***
B612        -2.821e-01  4.477e-01  -0.630  0.52853
B613        -7.395e+00  4.974e-01 -14.866  < 2e-16 ***
B614        -1.653e+01  8.283e-01 -19.953  < 2e-16 ***
B615        -2.308e+01  3.791e+00  -6.090 1.14e-09 ***
B616        -1.959e+01  4.068e+00  -4.816 1.47e-06 ***
B617        -3.497e+01  1.342e+01  -2.606  0.00917 **
B618         6.811e+00  2.716e+00   2.508  0.01216 *
B619        -4.340e+00  2.720e+00  -1.596  0.11060
B620        -1.581e+01  2.193e+00  -7.211 5.68e-13 ***
B621        -2.161e+01  8.334e+00  -2.593  0.00952 **
B622        -1.547e+01  1.556e+00  -9.946  < 2e-16 ***
B623        -1.874e+01  1.239e+00 -15.123  < 2e-16 ***
B624        -6.570e-01  7.558e-01  -0.869  0.38470
B625        -1.712e+01  1.450e+00 -11.808  < 2e-16 ***
B626        -1.872e+01  1.573e+00 -11.896  < 2e-16 ***
```

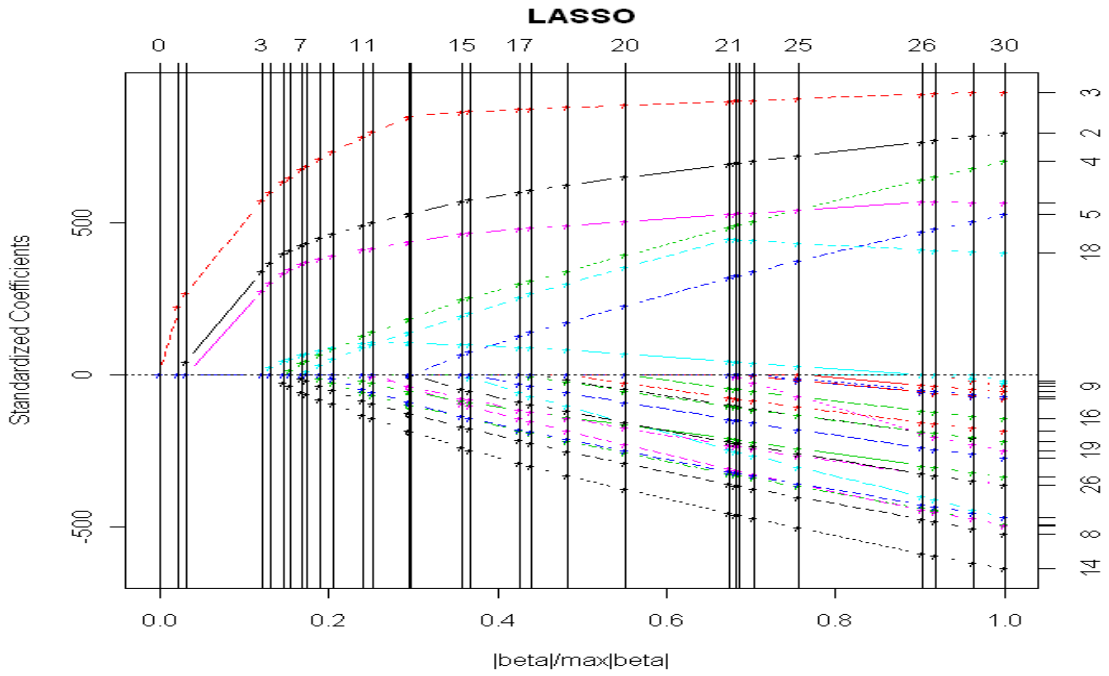
```
B627          -2.220e+01   1.224e+01   -1.813   0.06977 .
B628          -1.793e+01   2.486e+00   -7.214  5.57e-13 ***
B629          -1.882e+01   7.752e+00   -2.428   0.01519 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.97 on 30142 degrees of freedom
Multiple R-squared: 0.1035,      Adjusted R-squared: 0.1027
F-statistic: 124.3 on 28 and 30142 DF,  p-value: < 2.2e-16
```



加入语言成分的回归 R-squared 统计量有明显的提高，从 0.05 左右提高到了 0.1 左右，总体模型的 P 值很低，说明模型显著，大多数变量的 P 值检验也都显著，比只有地理变量的模型拟合有很大改进。在关键词汇中，不显著的有：“至”，“特价”，“机票”“飞机票”“的”“定”“订”，得到的回归系数 4 项地理因素都为正值符合常理，关键字的回归系数基本都为负值，只有“到”和“机票”两个词汇具有大于 0 的系数。观察 plot 图表我们可以看到残差明显的减小，但是重尾的问题依旧没有解决，其根本原因可能是提取出的地理变量不能包

含与利润率相关的所有信息。预测值和学生化标准差的图标有明显的 V 字形，其原因是出发地重复次数和目的地重复次数两个变量在很多样本中都取相同的值，比如最大的出发地重复数量有 3600 个之多那么这些观测的散点很可能聚集在一起或者形成某种图案。由 COOK 距离的判定可以得知异常值问题并不是非常严重。下面用 LASSO 方法再观察变量选择的过程：

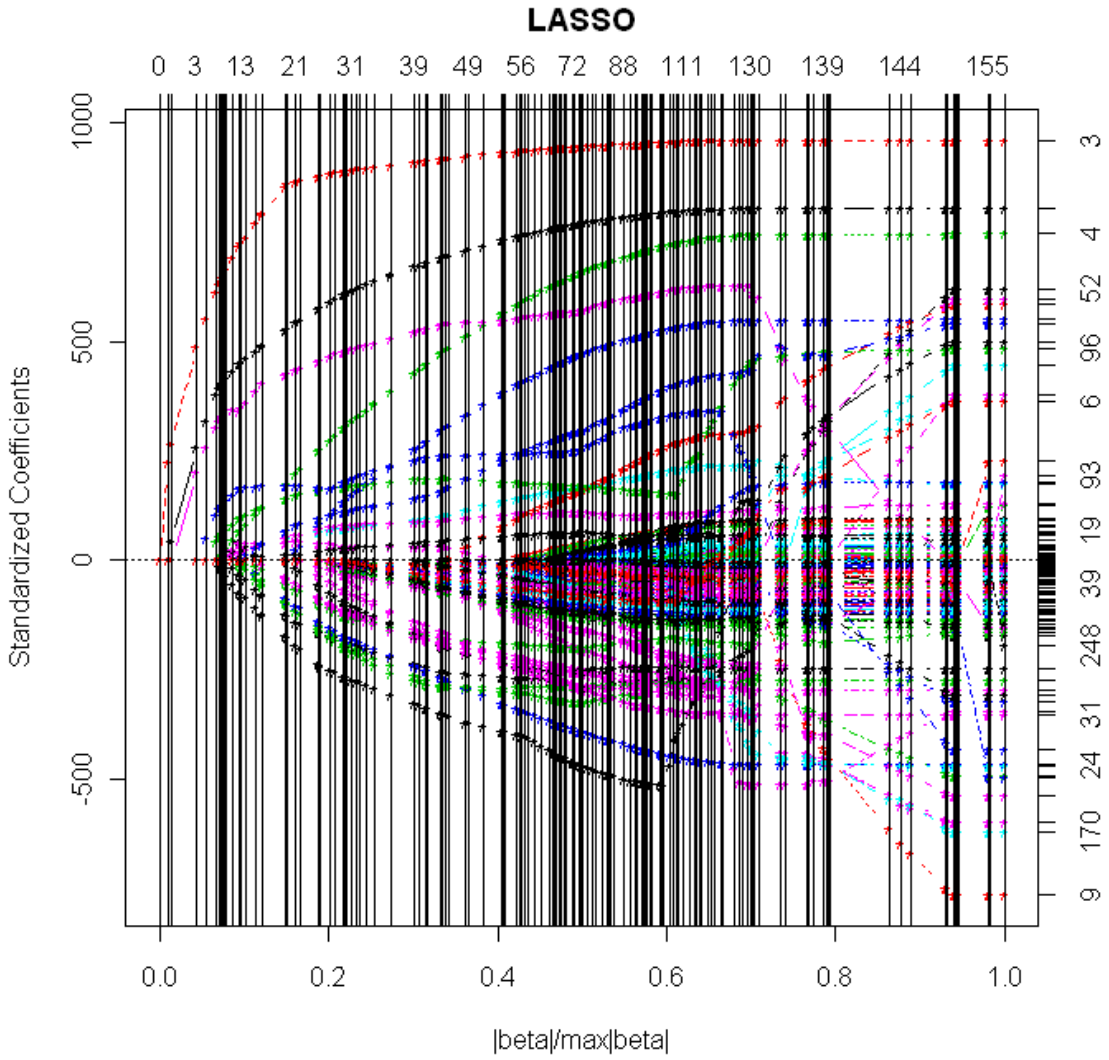
```
Call:
lars(x = B6, y = Bnew[, 2], type = "lasso")
R-squared: 0.103
Sequence of LASSO moves:

Var  3 2 7 12 14 4 8 18 10 23 22 25 13 26  5  6 11 28 20 15 16 19 17 21 29 27 -12  9 24 12
Step 1 2 3  4  5 6 7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26  27 28 29 30
```



对于地理因素的四个变量来讲，只有目的地重复次数是在第 15 位被选入模型的，其系数也是最小的一个表示这是价值相对小的地理变量。地理连接词中“到”具有较大的正回归系数，说明带有此关键词的搜索给公司带来利润的可能性较大，“至”在第 3 为被选入模型，但是当所有的连接词都被选入模型之后被剔除，最后的回归结果是此变量不显著。由于其余的连接词“-”，“飞”，“去”都对利润存在负影响，所以认为“至”是排名第二的连接词变量，次于“到”但优于其他三个。表示价格下降的 7 个变量的最终回归结果均非正，但是关键词“特价”在第 4 位被选入模型，其回归系数一直大于 0，当所有的同类词汇都被选入模型中之后被剔除，说明“特价”对于利润有正面的作用。其余项中有 5 项对于利润的不利影响比较明显，其从大到小的排序为：“特惠”“低价”“便宜”“优惠”“折扣”。两个主题词汇“机票”和“飞机票”的回顾系数都不大，并且显著程度都不高。“机票”的系数估计为正，而“飞机票”的系数估计为负，说明虽然对利润率影响不大，但是“机票”略好于“飞机票”。对于“往返”和“来回”这一对词汇的系数预测均为负值，而且其 P 值较小，对于模型均

显著。这说明在搜索的过程中较少的人会使用“往返”和“来回”，或是以其为搜索关键字的访客较少购买机票。由于目前网上预订机票的相关网站都可以提供往返订购功能，所以在搜索的过程中使用这个词汇可能是不必要的，因而导致购买此关键字的利润较低。“来回”比往返在措辞上更加随意，可能在搜索时更少见，所以其显著程度稍低。关键词“的”在模型接近稳定是才被选入模型，且最终结果为不显著，说明对于利润没有影响。

考虑到可能这些词中有一些固定搭配，我们还应该再模型中加入一些交叉变量考察词汇或地理因素变量之间的交互关系。例如：如果在链接词的位置使用了“飞”，那么根据中文的语言习惯，在后面的主题词上一般应该使用“机票”而不是“飞机票”。方便起见，我们加入所有 28 个变量的交叉项，所谓的同位词的交叉项的取值应该为 0。



（最终回归结果见附录）

通过对于回顾结果的分析，发现加入交叉项之后显著的单个词汇有“价格”,“查询”,“预定”,“预订”，它们的回归系数均小于零，可以认为带有这些词汇的关键字利润被一定程度的降低。在词汇组合中，变量显著并且对于利润有正面作用的只有两个（至，特价）和（飞，折扣），其中前者更加显著并且系数绝对值更大。其他对于利润有负作用的关键词组合有：（——，打折）（——，特价），（到，价），（到，查询）。显著性稍差的几个有（特惠，机票），（去，特价），（到，折扣）。

5. 关于全文的一点思考：

从开始分析数据的时候就有两种大概的思路，一种是分类一种是回归。由于利润的分布有明显的特点，而如果利润小于零，一般的企业不会关注其具体的值，所以可以选择分类的方法。但是由于分类过程最终的取值也是离散的，因变量和自变量的取值都过于不自由，可能导致模型无法做调整或者改进。另外，对于大于零的变量，其具体的取值可以帮助企业在竞标此关键字时定价，有更深远的意义，所以本文的最终方法确定为回归。

但正如以上回归模型所显示的，本文并没有构造出能够很好的关于利润率的模型，这主要有两个原因，第一个是我们从原始数据中提取实质变量的过程。我个人认为存在问题最大的可能是取出发地和目的地平均利润的这一环节。本文采用的这两个变量不能很好的描述出发地和目的地的情况。另外一个是收益率本身是由企业计算得到的，其中用到的数据和公式我们无从得知。很有可能公司在计算得到此收益率的时候使用了很多维数据，比如点击次数，购买次数等等。根据付费搜索广告的运行机制，这些变量包含了许多有价值的信息，如果我们只适用利润率这一变量，那么我们的全部信息只是真是情况的一下部分，另外企业使用的公式可能也不合理，在评定利润率的时候产生系统偏差。所以就此看来，我们只能够拟合出总平方和十分之一的回归平方和也是可以理解的，我们最好的预期是模型包含了地理和语言两个因素，但是只包含这两个因素。

在文章的后半部分对于语言因素做回顾的过程中，我曾经考虑过使用迭代的方法：第一次我们对于地理因素的估计使用的是简单的去平均值方法，但是当我们从这样的地理因素出发，得到语言因素之后，我们可以用得语言因素得到的系数作用在原始收益率上，即从原始数据利润率中减去语言因素影响，这样我们就可以得到新的一些地理因素，然后我们可以在从新的地理因素出发归纳出一些新的地理因素变量，然后第二次对于语言因素做回归。进行这样的迭代若干次之后，我们可能能够得到一组稳定的估计，使用地理因素和语言因素之一估计另外一个，得到的恰是另外一个因素最终的结果，这种结果可能更加贴近实际。但是这种方法存在一个很严重的问题，就是如果我们在第一次用语言因素的回归结果反向估计地理因素是产生了与初始的简单地理因素方向相同，但是幅度更大的偏差，那么迭代的步数越多，偏差就可能被扩大的越多，最后无法稳定或者产生一个很不符合实际的结果，而我们由于掌握的信息太少，可能无法判断得到结果的可靠性。

6.结论

关键字与利润的二维数据不能包含大部分与利润率相关的信息，为了更加精准的预测和分析利润的影响因素，需要从外部获得更多的数据。

就能够提取出的变量信息来看，出发地与目的地的情况很大程度上决定了最终的利润。目的地与出发地之间的指标相差越多越不利于利润的升高，而目的地本身的利润均值和出现的频繁程度在所有城市中的排名相比也能在一定程度上决定最终利润，出现程度偏高对于利润有负影响，在竞标可以降低出价或适当减少此类关键字。提取出的地理变量对于利润的影响不是简单线性的，当出发地平均利润超过总体的 30%分位数时会对利润产生额外的提升，而出发地的重复次数超过总体的 50%分位数是则会给利润带来更大的负面影响，所以当上述两种情况出现时，对于竞标时出价的调整力度应该更大。

从关键词中的语言因素考虑，“到”和“至”的效果在地理连接词中排在前两名，其余几个的都有很大可能减低原有利润率。“特价”是修饰机票最好的词汇，“打折”紧随其后排名第二，对于提高利润的作用明显好于其他词语如“折扣”“便宜”等等。大多数人会使用“机票”的说法，而相对较少的人会选择“飞机票”。带有“往返”和“来回”两词的关键字可以不购买，它们往往会降低收益率。另外助词“的”的作用非常不明显，所以在竞标时应该成对的选择，无论是包括“的”，其价值都在同一水平。

7. 附录

1. 分段回归时的 LASSO 变量选择步骤，逐步残差和 Cp 统计量

```
Call:
lars(x = B84, y = B88[, 2], type = "lasso")
R-squared: 0.051
Sequence of LASSO moves:

Var  56 46 57 47 48 29 -47 8 58 19 -48 -58  2 -55 32 11 18 27 26 12 17  6  1 45 -57 38 37 41 21 16 -46 22
40  9  7 -11 -56 44 23 77 14 -40 36 -23 54 -7 74 -14 81 -41 -9
Step  1  2  3  4  5  6  7 8  9 10  11  12 13  13 14 15 16 17 18 19 20 21 22 22  23 24 25 26 27 28  29 30
31 32 33  34  35 36 37 38 39  40 41  42 43 44 45  46 47  48 49

Var  31 73 10 43 24 50  4 -75 28 -10 60 39 51 76 -77 40 -4 47 13 71 84 -31 20 -74 -17 57 46 49 82 33 30  3
-65 68 -28 -33 -49 72 -30 10  7 33 -10 70 14 10 -33 30 -70 56
Step 50 51 52 53 54 55 56  56 57  58 59 60 61 62  63 64 65 66 67 68 69  70 71  72  73 74 75 76 77 78 79 80
80 81  82  83  84 85  86 87 88 89  90 91 92 93  94 95  96 97

Var  28 -18 -50 -16  34 -47 -34  11  34 -13  17  53  16 -17 -51  78  18 -27 -34  67  17  64  33  41 -19 -73
-39  9  39  59  66  27 -24 -67  19 -46 -59  62  48 -33 -22
Step 98  99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123
124 125 126 127 128 129 130 131 132 133 134 135 136 137 138

Var  46  61  70  -8 -46  33  47  73  23 -33  67  74 -43  24  46 -27  4  63  31  22  58 -63  50 -24  43
80 -74  33  79  8 -48 -23 -17  27 -73  83  77  34  69  17  24
Step 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163
164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179

Var  49  -9 -43  59  23  52 -34  42  73 -82  51  63  48 -73  82  34 -24 -14  9 -48  43  14  48  -8  8
-34  73  -1 -33  1  74 -54  13 -43  24 -54  43 -24  24 -44  44
Step 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204
205 206 207 208 209 209 210 211 212 213 214 215 216 217 218 219

Var  -33  34
Step 220 221

Call: lars(x = B88[, 9:88], y = B88[, 2], type = "lasso")
    Df      Rss       Cp
0    1 30205077 1558.748
```

1	2	29876444	1215.548
2	3	29554626	879.506
3	4	29308522	622.996
4	5	29287987	603.425
5	6	29254883	570.653
6	7	29204547	519.779
7	6	29173792	485.474
8	7	29134239	445.927
9	8	29125877	439.144
10	9	29122324	437.412
11	8	29103879	416.037
12	7	29033470	340.079
13	8	29030994	339.478
14	9	29014913	324.586
15	10	28991923	302.437
16	11	28975248	286.921
17	12	28972771	286.319
18	13	28924449	237.561
19	14	28887473	200.722
20	15	28875387	190.026
21	16	28869682	186.033
22	17	28862153	180.125
23	16	28856597	172.289
24	17	28806283	121.439
25	18	28804768	121.848
26	19	28795106	113.698
27	20	28788343	108.595
28	21	28772626	94.085
29	20	28746001	64.117
30	21	28739574	59.367
31	22	28738554	60.295
32	23	28735010	58.573
33	24	28732767	58.217
34	23	28729061	52.323
35	22	28712596	33.029
36	23	28709852	32.147
37	24	28705609	29.690
38	25	28704588	30.617
39	26	28702913	30.857
40	25	28700961	26.807
41	26	28700323	28.137
42	25	28697184	22.840
43	26	28696400	24.016
44	25	28694785	20.320

45 26 28694123 21.625
46 25 28693070 18.518
47 26 28691967 19.360
48 25 28691110 16.459
49 24 28690595 13.919
50 25 28690282 15.590
51 26 28689885 17.172
52 27 28689157 18.408
53 28 28688673 19.899
54 29 28687878 21.064
55 30 28687626 22.800
56 31 28687415 24.578
57 32 28687003 26.146
58 31 28686983 24.125
59 32 28686769 25.900
60 33 28685737 26.816
61 34 28685156 28.205
62 35 28685067 30.112
63 34 28684914 27.951
64 35 28683496 28.462
65 34 28682951 25.890
66 35 28680799 25.628
67 36 28680743 27.570
68 37 28679844 28.625
69 38 28679804 30.584
70 37 28679108 27.853
71 38 28679002 29.741
72 37 28678959 27.696
73 36 28678676 25.399
74 37 28678608 27.328
75 38 28678403 29.112
76 39 28678187 30.885
77 40 28677904 32.588
78 41 28676808 33.437
79 42 28676536 35.151
80 43 28676261 36.861
81 44 28675771 38.347
82 43 28675323 35.877
83 42 28675196 33.743
84 41 28675010 31.548
85 42 28674610 33.128
86 41 28674329 30.833
87 42 28673905 32.388
88 43 28673873 34.354

89	44	28673771	36.247
90	43	28673451	33.910
91	44	28673448	35.907
92	45	28673127	37.570
93	46	28673112	39.554
94	45	28673011	37.449
95	46	28672992	39.429
96	45	28672928	37.361
97	46	28672915	39.347
98	47	28672610	41.026
99	46	28672257	38.656
100	45	28672129	36.522
101	44	28672113	34.505
102	45	28672074	36.464
103	44	28672055	34.444
104	43	28671820	32.197
105	44	28671682	34.053
106	45	28671470	35.830
107	44	28670974	33.308
108	45	28670966	35.300
109	46	28670965	37.299
110	47	28670952	39.285
111	46	28670900	37.230
112	45	28670699	35.019
113	46	28670554	36.867
114	47	28670455	38.764
115	46	28670338	36.641
116	45	28670161	34.455
117	46	28670155	36.449
118	47	28669684	37.953
119	48	28669633	39.900
120	49	28669338	41.590
121	50	28669100	43.340
122	49	28668921	41.152
123	48	28668814	39.040
124	47	28668362	36.565
125	48	28668240	38.437
126	49	28668192	40.387
127	50	28668156	42.348
128	51	28667404	43.559
129	52	28667233	45.379
130	51	28666903	43.033
131	50	28666808	40.932
132	51	28666753	42.875

133 50 28666746 40.867
134 49 28666573 38.686
135 50 28666544 40.655
136 51 28666380 42.482
137 50 28666289 40.387
138 49 28666247 38.343
139 50 28666112 40.202
140 51 28665874 41.952
141 52 28665683 43.751
142 51 28665232 41.277
143 50 28664916 38.945
144 51 28664849 40.875
145 52 28663964 41.945
146 53 28663717 43.686
147 54 28663516 45.474
148 53 28663453 43.409
149 54 28663294 45.242
150 55 28662881 46.807
151 54 28662859 44.784
152 55 28662674 46.590
153 56 28662334 48.232
154 55 28662303 46.200
155 56 28662278 48.174
156 57 28661968 49.848
157 58 28661887 51.763
158 59 28661415 53.267
159 60 28661055 54.890
160 59 28660910 52.737
161 60 28660374 54.174
162 59 28660321 52.119
163 60 28660289 54.085
164 61 28659843 55.617
165 60 28659685 53.451
166 61 28659612 55.374
167 62 28659200 56.941
168 63 28658917 58.644
169 62 28658747 56.466
170 61 28658625 54.337
171 60 28658360 52.059
172 61 28658023 53.704
173 60 28657598 51.258
174 61 28657422 53.073
175 62 28657186 54.825
176 63 28656956 56.584

177 64 28656744 58.361
178 65 28656526 60.132
179 66 28656493 62.098
180 67 28656380 63.978
181 66 28656318 61.913
182 65 28656046 59.628
183 66 28655826 61.397
184 67 28655665 63.227
185 68 28654622 64.132
186 67 28654442 61.943
187 68 28654423 63.923
188 69 28654412 65.912
189 68 28654103 63.587
190 69 28654036 65.517
191 70 28654009 67.488
192 71 28653796 69.265
193 70 28653782 67.249
194 71 28653758 69.225
195 72 28653749 71.215
196 71 28653707 69.171
197 70 28653700 67.163
198 71 28653699 69.162
199 70 28653698 67.162
200 71 28653691 69.154
201 72 28653691 71.154
202 73 28653691 73.154
203 72 28653691 71.154
204 73 28653688 73.151
205 72 28653688 71.151
206 73 28653684 73.147
207 72 28653684 71.147
208 73 28653573 73.031
209 72 28653573 71.031
210 73 28653554 73.010
211 72 28653553 71.009
212 73 28653544 73.000
213 72 28653544 71.000
214 73 28653544 73.000
215 72 28653544 71.000
216 73 28653544 73.000
217 72 28653544 71.000
218 73 28653544 73.000
219 72 28653544 71.000
220 73 28653544 73.000

2. 重新分段之后的回归

```
Call:
lm(formula = B88n[, 2] ~ B88n[, 9:88])

Residuals:
    Min      1Q  Median      3Q     Max
-57.894 -14.563  -8.136   1.868 270.717

Coefficients: (25 not defined because of singularities)
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -3.731e+01  1.073e+02  -0.348 0.728178
B88n[, 9:88]1          NA         NA      NA       NA
B88n[, 9:88]2   7.996e+00  1.445e+02   0.055 0.955862
B88n[, 9:88]3   1.110e+01  9.678e+01   0.115 0.908729
B88n[, 9:88]4   1.991e-01  3.229e+00   0.062 0.950843
B88n[, 9:88]5  -1.177e+00  5.878e+00  -0.200 0.841288
B88n[, 9:88]6   5.019e+00  6.507e+00   0.771 0.440493
B88n[, 9:88]7  -2.898e+01  9.151e+00  -3.166 0.001546 **
B88n[, 9:88]8  -6.201e+01  9.487e+01  -0.654 0.513337
B88n[, 9:88]9   2.114e+02  1.607e+02   1.315 0.188484
B88n[, 9:88]10  4.135e+01  4.509e+01   0.917 0.359082
B88n[, 9:88]11         NA         NA      NA       NA
B88n[, 9:88]12 -4.952e-01  2.867e+00  -0.173 0.862869
B88n[, 9:88]13  5.342e+00  3.347e+00   1.596 0.110484
B88n[, 9:88]14  3.932e+01  2.987e+01   1.316 0.188150
B88n[, 9:88]15  1.112e+01  9.772e+01   0.114 0.909364
B88n[, 9:88]16  1.506e+02  2.497e+02   0.603 0.546486
B88n[, 9:88]17  1.565e+02  1.404e+02   1.114 0.265098
B88n[, 9:88]18         NA         NA      NA       NA
B88n[, 9:88]19         NA         NA      NA       NA
B88n[, 9:88]20  1.491e+02  1.627e+02   0.916 0.359633
B88n[, 9:88]21         NA         NA      NA       NA
B88n[, 9:88]22  1.582e+00  3.428e+00   0.462 0.644386
B88n[, 9:88]23  7.571e+00  6.660e+00   1.137 0.255657
B88n[, 9:88]24  1.312e+00  2.794e+01   0.047 0.962541
B88n[, 9:88]25  4.315e+00  8.375e+00   0.515 0.606350
B88n[, 9:88]26  2.091e+00  1.375e+01   0.152 0.879090
B88n[, 9:88]27 -1.817e+00  2.383e+01  -0.076 0.939223
B88n[, 9:88]28         NA         NA      NA       NA
B88n[, 9:88]29         NA         NA      NA       NA
B88n[, 9:88]30  5.063e+00  2.748e+01   0.184 0.853848
B88n[, 9:88]31         NA         NA      NA       NA
```

```
B88n[, 9:88]32 -3.817e+00  5.157e+00  -0.740 0.459230
B88n[, 9:88]33  5.440e+00  6.070e+00   0.896 0.370173
B88n[, 9:88]34 -1.130e+01  9.823e+00  -1.150 0.250101
B88n[, 9:88]35  5.377e+00  1.359e+01   0.396 0.692292
B88n[, 9:88]36 -5.922e-01  2.108e+00  -0.281 0.778782
B88n[, 9:88]37 -1.035e+00  3.690e+00  -0.280 0.779113
B88n[, 9:88]38 -3.333e-01  7.594e+00  -0.044 0.964991
B88n[, 9:88]39 -1.232e+01  1.087e+01  -1.134 0.256979
B88n[, 9:88]40          NA         NA      NA       NA
B88n[, 9:88]41 -8.000e-01  1.182e+01  -0.068 0.946060
B88n[, 9:88]42 -3.114e+00  3.837e+01  -0.081 0.935327
B88n[, 9:88]43  3.952e+00  3.656e+01   0.108 0.913919
B88n[, 9:88]44  4.621e-01  2.079e+00   0.222 0.824093
B88n[, 9:88]45  2.798e-01  8.979e-01   0.312 0.755373
B88n[, 9:88]46 -4.543e-01  7.202e-01  -0.631 0.528235
B88n[, 9:88]47  1.911e+00  6.366e-01   3.002 0.002684 **
B88n[, 9:88]48  2.747e+00  4.801e+00   0.572 0.567294
B88n[, 9:88]49 -9.904e+00  7.498e+00  -1.321 0.186562
B88n[, 9:88]50          NA         NA      NA       NA
B88n[, 9:88]51  1.030e+00  1.483e+00   0.695 0.487316
B88n[, 9:88]52 -3.931e-02  1.487e+00  -0.026 0.978917
B88n[, 9:88]53 -4.187e-01  2.491e-01  -1.681 0.092795 .
B88n[, 9:88]54 -1.320e+00  1.125e+00  -1.174 0.240554
B88n[, 9:88]55  4.160e-01  2.261e+00   0.184 0.854036
B88n[, 9:88]56 -2.759e+00  4.603e+00  -0.600 0.548824
B88n[, 9:88]57          NA         NA      NA       NA
B88n[, 9:88]58          NA         NA      NA       NA
B88n[, 9:88]59          NA         NA      NA       NA
B88n[, 9:88]60          NA         NA      NA       NA
B88n[, 9:88]61  1.993e-02  5.587e-03   3.567 0.000362 ***
B88n[, 9:88]62 -1.106e-02  9.341e-03  -1.184 0.236506
B88n[, 9:88]63 -1.110e-02  1.042e-02  -1.065 0.286876
B88n[, 9:88]64 -4.790e-04  3.012e-02  -0.016 0.987313
B88n[, 9:88]65          NA         NA      NA       NA
B88n[, 9:88]66          NA         NA      NA       NA
B88n[, 9:88]67          NA         NA      NA       NA
B88n[, 9:88]68          NA         NA      NA       NA
B88n[, 9:88]69          NA         NA      NA       NA
B88n[, 9:88]70          NA         NA      NA       NA
B88n[, 9:88]71 -5.744e-04  7.889e-03  -0.073 0.941961
B88n[, 9:88]72  1.760e-02  1.831e-02   0.961 0.336625
B88n[, 9:88]73 -1.667e-02  1.782e-02  -0.935 0.349548
B88n[, 9:88]74  1.636e-02  1.444e-02   1.133 0.257224
B88n[, 9:88]75 -8.316e-03  1.671e-02  -0.498 0.618813
```

```
B88n[, 9:88]76          NA          NA       NA       NA
B88n[, 9:88]77          NA          NA       NA       NA
B88n[, 9:88]78          NA          NA       NA       NA
B88n[, 9:88]79          NA          NA       NA       NA
B88n[, 9:88]80          NA          NA       NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.85 on 30115 degrees of freedom
Multiple R-squared: 0.0509,     Adjusted R-squared: 0.04916
F-statistic: 29.36 on 55 and 30115 DF,  p-value: < 2.2e-16
```

3. 加入指数项后的部分估计系数

```
B128[, 5:128]89           NA          NA       NA       NA
B128[, 5:128]90    2.678e+03   6.127e+03    0.437    0.662
B128[, 5:128]91   -3.551e+03   4.400e+03   -0.807    0.420
B128[, 5:128]92    2.004e+03   2.176e+03    0.921    0.357
B128[, 5:128]93   -1.667e+02   7.608e+02   -0.219    0.827
B128[, 5:128]94           NA          NA       NA       NA
B128[, 5:128]95    1.218e+00   2.227e+01    0.055    0.956
B128[, 5:128]96    3.064e+01   1.143e+02    0.268    0.789
B128[, 5:128]97    1.280e+02   3.177e+02    0.403    0.687
B128[, 5:128]98   -1.920e+02   5.797e+02   -0.331    0.741
B128[, 5:128]99   -4.577e+01   6.347e+02   -0.072    0.943
B128[, 5:128]100   1.220e+02   4.589e+02    0.266    0.790
B128[, 5:128]101 -9.419e+01   2.525e+02   -0.373    0.709
B128[, 5:128]102   7.746e+01   2.312e+02    0.335    0.738
B128[, 5:128]103 -2.742e+01   2.547e+01   -1.076    0.282

 Residual standard error: 30.86 on 30081 degrees of freedom
Multiple R-squared: 0.05156,    Adjusted R-squared: 0.04875
F-statistic: 18.37 on 89 and 30081 DF,  p-value: < 2.2e-16
```

4. 加入关键字交叉项后的 LASSO 变量选择步骤

```
Call:
lars(x = B9, y = Bnew[, 2], type = "lasso")
R-squared: 0.118
Sequence of LASSO moves:

Var  3 2 7 146 96 -257 14 -162 4 37 133 66  8 52 10 18 219 23 40 22 171 -195 -52 25 26  5 59 148 11 128 32
48 31 105 28 223 98 259 20 49 15 -177 158 50 68 13 -159 77 108
Step 1 2 3   4  5    5  6    6 7  8   8  9 10 11 12 13  13 14 15 16  17   17  18 19 20 21 22  23 24  24 25
```

26 27 28 29 30 31 31 32 33 34 34 35 36 37 38 38 39 40

Var 35 -132 78 45 73 16 180 53 87 106 17 -191 -192 -214 41 29 27 -66 61 118 116 226 69 55 30 82 -232 94 -253 91 247 154 -236 183 -204 88 21 222 97 -258 101 39 -134 79
Step 41 41 42 43 44 45 45 46 47 48 49 49 49 49 50 51 52 53 54 55 56 57 58 59 60 61 61 62 62 63 63 64 64 65 65 66 67 67 68 68 69 70 70 71

Var -229 99 -260 90 -246 92 -252 153 -261 95 115 33 67 83 233 63 86 38 -127 81 -230 46 111 42 66 172 200 71 120 100 70 184 -209 93 248 47 75 80 -231 103 84 -234 110 196
Step 71 72 72 73 73 74 74 75 75 76 77 78 79 80 80 81 82 83 83 84 84 85 86 87 88 89 89 90 91 92 93 94 94 95 95 96 97 98 98 99 100 100 101 102

Var -237 123 34 -125 170 -199 119 85 228 56 24 62 147 -99 109 54 51 107 99 89 245 64 60 -259 121 259 36 -126 12 136 9 -8 -98 -66 169 194 52 -68 19 -225
Step 102 103 104 104 105 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 119 120 121 122 123 123 124 124 125 125 126 127 128 129 130 130 131 132 133 133

Var 66 -146 6 146 8 -13 68 -46 46 13 -14 14 -93 -7 7 144 221 185 -205 186 -210 -180 -16 16 180 98 -259 93 -259
Step 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 151 152 152 153 154 155 155 156 157 158 158

5. 词汇关键字交互影响的回归结果：

Residuals:

Min	1Q	Median	3Q	Max
-109.731	-18.067	-6.072	6.428	3009.016

Coefficients: (165 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-4.056e+01	2.415e+01	-1.679	0.093144	.
B101	NA	NA	NA	NA	
B102	1.376e+00	5.319e-02	25.862	< 2e-16	***
B103	1.010e+00	6.066e-02	16.643	< 2e-16	***
B104	4.899e-03	2.941e-04	16.656	< 2e-16	***
B105	1.137e-02	1.038e-03	10.951	< 2e-16	***
B106	-1.036e+00	5.312e+00	-0.195	0.845437	
B107	3.671e+00	2.862e+01	0.128	0.897941	
B108	-1.083e+01	6.122e+00	-1.769	0.076872	.
B109	-3.214e+01	4.473e+01	-0.718	0.472466	
B1010	-1.437e+01	2.283e+01	-0.629	0.529076	
B1011	1.655e+01	2.416e+01	0.685	0.493165	
B1012	NA	NA	NA	NA	
B1013	2.190e+00	2.419e+01	0.091	0.927844	

B1014	-2.080e+01	1.584e+01	-1.314	0.188955	
B1015	-8.028e-01	5.342e+01	-0.015	0.988010	
B1016	4.553e+00	4.073e+01	0.112	0.911003	
B1017	-2.977e+01	5.304e+01	-0.561	0.574647	
B1018	-5.905e+01	4.765e+01	-1.239	0.215302	
B1019	2.134e+01	2.980e+01	0.716	0.474009	
B1020	-2.086e+01	3.056e+01	-0.682	0.494974	
B1021	-3.007e+01	9.403e+00	-3.198	0.001386	**
B1022	-9.586e+00	2.439e+01	-0.393	0.694265	
B1023	-2.962e+01	1.507e+01	-1.966	0.049257	*
B1024	-4.588e+01	3.372e+01	-1.361	0.173639	
B1025	-3.233e+01	2.765e+01	-1.169	0.242392	
B1026	-1.907e+01	3.292e+00	-5.793	7.00e-09	***
B1027	-1.842e+01	5.232e+00	-3.521	0.000431	***
B1028	-1.772e+01	6.334e+00	-2.797	0.005165	**
B1029	-3.509e+01	3.369e+01	-1.041	0.297687	
B1030	-2.501e+01	1.440e+01	-1.737	0.082413	.
B1031	-5.813e+00	5.389e+00	-1.079	0.280707	
B1032	NA	NA	NA	NA	
B1033	NA	NA	NA	NA	
B1034	8.274e+00	2.863e+01	0.289	0.772555	
B1035	NA	NA	NA	NA	
B1036	1.760e+01	2.867e+01	0.614	0.539434	
B1037	-8.134e+00	6.225e+00	-1.307	0.191336	
B1038	NA	NA	NA	NA	
B1039	NA	NA	NA	NA	
B1040	2.662e+01	4.471e+01	0.595	0.551530	
B1041	NA	NA	NA	NA	
B1042	3.336e+01	4.474e+01	0.746	0.455836	
B1043	8.193e-01	2.280e+01	0.036	0.971334	
B1044	NA	NA	NA	NA	
B1045	1.096e+01	2.312e+01	0.474	0.635578	
B1046	-1.108e+01	2.270e+01	-0.488	0.625368	
B1047	-5.568e+00	1.912e+00	-2.912	0.003598	**
B1048	3.318e+00	3.814e+00	0.870	0.384380	
B1049	NA	NA	NA	NA	
B1050	NA	NA	NA	NA	
B1051	-9.302e+00	1.677e+00	-5.546	2.95e-08	***
B1052	8.602e+00	2.657e+01	0.324	0.746156	
B1053	-1.714e+01	9.323e+00	-1.839	0.065933	.
B1054	-2.602e+00	1.985e+00	-1.311	0.189873	
B1055	-7.857e+00	3.963e+00	-1.982	0.047452	*
B1056	-4.898e+00	4.855e+01	-0.101	0.919634	
B1057	1.525e+01	5.326e+01	0.286	0.774709	

B1058	-1.224e+00	1.772e+00	-0.691	0.489802	
B1059	-1.882e+01	2.658e+01	-0.708	0.478749	
B1060	7.816e+00	2.108e+01	0.371	0.710830	
B1061	2.163e+00	3.924e+00	0.551	0.581528	
B1062	1.726e+01	1.070e+01	1.613	0.106746	
B1063	NA	NA	NA	NA	
B1064	NA	NA	NA	NA	
B1065	-7.467e+00	3.534e+00	-2.113	0.034602	*
B1066	9.283e+00	3.948e+01	0.235	0.814115	
B1067	6.835e+00	1.212e+01	0.564	0.572891	
B1068	5.241e+00	2.753e+00	1.904	0.056980	.
B1069	5.243e+00	5.222e+00	1.004	0.315382	
B1070	2.028e+01	4.883e+01	0.415	0.677887	
B1071	NA	NA	NA	NA	
B1072	9.722e+00	2.479e+00	3.922	8.79e-05	***
B1073	-9.889e-01	2.968e+01	-0.033	0.973423	
B1074	1.223e+01	1.271e+01	0.962	0.335978	
B1075	2.505e+00	2.882e+00	0.869	0.384661	
B1076	1.120e+01	5.580e+00	2.008	0.044701	*
B1077	3.127e+01	5.512e+01	0.567	0.570477	
B1078	NA	NA	NA	NA	
B1079	-2.082e+00	2.390e+00	-0.871	0.383648	
B1080	-4.057e+00	4.275e+01	-0.095	0.924395	
B1081	-7.686e-01	1.156e+01	-0.066	0.947013	
B1082	9.596e+00	4.792e+01	0.200	0.841276	
B1083	-4.025e+00	4.809e+00	-0.837	0.402586	
B1084	-1.843e+01	6.613e+00	-2.787	0.005320	**
B1085	1.450e+01	1.413e+01	1.026	0.304751	
B1086	5.928e-01	4.812e+01	0.012	0.990171	
B1087	9.424e+00	6.176e+00	1.526	0.127079	
B1088	-5.132e+00	8.307e+00	-0.618	0.536726	
B1089	1.383e+01	8.311e+00	1.665	0.095985	.
B1090	1.260e+00	1.513e+01	0.083	0.933640	
B1091	-4.003e+00	4.828e+01	-0.083	0.933929	
B1092	NA	NA	NA	NA	
B1093	-9.815e+00	9.173e+00	-1.070	0.284662	
B1094	-7.507e+00	3.890e+01	-0.193	0.846988	
B1095	1.955e+01	1.711e+01	1.143	0.253160	
B1096	3.908e+01	5.837e+01	0.670	0.503139	
B1097	2.996e+00	1.096e+01	0.273	0.784527	
B1098	6.194e+01	4.353e+01	1.423	0.154704	
B1099	-4.506e+00	1.970e+01	-0.229	0.819093	
B10100	3.756e+01	5.836e+01	0.644	0.519831	
B10101	8.341e+00	5.500e+01	0.152	0.879460	

B10102	3.547e+01	2.758e+01	1.286	0.198400	
B10103	3.711e+01	2.782e+01	1.334	0.182297	
B10104	2.165e+01	2.762e+01	0.784	0.432997	
B10105	2.597e+01	2.796e+01	0.929	0.353098	
B10106	7.502e+00	1.257e+01	0.597	0.550545	
B10107	-4.022e+00	8.624e+00	-0.466	0.640953	
B10108	1.063e+01	1.656e+01	0.642	0.520814	
B10109	NA	NA	NA	NA	
B10110	NA	NA	NA	NA	
B10111	-1.535e+01	5.343e+00	-2.872	0.004078	**
B10112	-1.385e+01	6.167e+00	-2.245	0.024760	*
B10113	-1.905e+01	9.173e+00	-2.077	0.037854	*
B10114	NA	NA	NA	NA	
B10115	5.113e+00	3.104e+01	0.165	0.869168	
B10116	-1.048e+01	4.777e+01	-0.219	0.826310	
B10117	1.635e+01	3.412e+01	0.479	0.631763	
B10118	NA	NA	NA	NA	
B10119	NA	NA	NA	NA	
B10120	NA	NA	NA	NA	
B10121	7.067e+00	6.568e+00	1.076	0.281944	
B10122	-5.551e+00	7.817e+00	-0.710	0.477610	
B10123	-5.085e+00	1.114e+01	-0.457	0.647954	
B10124	NA	NA	NA	NA	
B10125	1.712e+01	4.977e+01	0.344	0.730859	
B10126	5.012e+00	1.147e+01	0.437	0.662082	
B10127	3.873e-01	1.175e+01	0.033	0.973714	
B10128	7.841e+00	1.574e+01	0.498	0.618324	
B10129	NA	NA	NA	NA	
B10130	NA	NA	NA	NA	
B10131	NA	NA	NA	NA	
B10132	NA	NA	NA	NA	
B10133	NA	NA	NA	NA	
B10134	NA	NA	NA	NA	
B10135	NA	NA	NA	NA	
B10136	-6.576e-01	1.458e+01	-0.045	0.964027	
B10137	-7.371e+00	5.341e+01	-0.138	0.890235	
B10138	NA	NA	NA	NA	
B10139	NA	NA	NA	NA	
B10140	NA	NA	NA	NA	
B10141	NA	NA	NA	NA	
B10142	NA	NA	NA	NA	
B10143	NA	NA	NA	NA	
B10144	NA	NA	NA	NA	
B10145	NA	NA	NA	NA	

B10146	NA	NA	NA	NA
B10147	NA	NA	NA	NA
B10148	NA	NA	NA	NA
B10149	NA	NA	NA	NA
B10150	NA	NA	NA	NA
B10151	-2.193e+01	5.345e+01	-0.410	0.681567
B10152	-2.578e+01	4.059e+01	-0.635	0.525365
B10153	-3.682e+00	2.333e+01	-0.158	0.874588
B10154	NA	NA	NA	NA
B10155	NA	NA	NA	NA
B10156	-2.571e+01	4.127e+01	-0.623	0.533294
B10157	NA	NA	NA	NA
B10158	NA	NA	NA	NA
B10159	-1.670e+01	2.979e+01	-0.561	0.575029
B10160	NA	NA	NA	NA
B10161	NA	NA	NA	NA
B10162	NA	NA	NA	NA
B10163	-4.230e+01	2.986e+01	-1.417	0.156532
B10164	-1.562e+00	1.695e+01	-0.092	0.926610
B10165	NA	NA	NA	NA
B10166	NA	NA	NA	NA
B10167	5.402e+00	9.135e+00	0.591	0.554312
B10168	-1.652e+01	2.474e+01	-0.668	0.504375
B10169	1.320e+01	1.441e+01	0.916	0.359797
B10170	NA	NA	NA	NA
B10171	-9.051e-01	2.778e+00	-0.326	0.744555
B10172	NA	NA	NA	NA
B10173	NA	NA	NA	NA
B10174	NA	NA	NA	NA
B10175	NA	NA	NA	NA
B10176	NA	NA	NA	NA
B10177	NA	NA	NA	NA
B10178	-1.574e+00	2.459e+01	-0.064	0.948948
B10179	NA	NA	NA	NA
B10180	NA	NA	NA	NA
B10181	NA	NA	NA	NA
B10182	NA	NA	NA	NA
B10183	NA	NA	NA	NA
B10184	NA	NA	NA	NA
B10185	NA	NA	NA	NA
B10186	NA	NA	NA	NA
B10187	NA	NA	NA	NA
B10188	NA	NA	NA	NA
B10189	NA	NA	NA	NA

B10190	NA	NA	NA	NA
B10191	NA	NA	NA	NA
B10192	NA	NA	NA	NA
B10193	NA	NA	NA	NA
B10194	NA	NA	NA	NA
B10195	NA	NA	NA	NA
B10196	NA	NA	NA	NA
B10197	NA	NA	NA	NA
B10198	NA	NA	NA	NA
B10199	NA	NA	NA	NA
B10200	NA	NA	NA	NA
B10201	NA	NA	NA	NA
B10202	NA	NA	NA	NA
B10203	NA	NA	NA	NA
B10204	NA	NA	NA	NA
B10205	NA	NA	NA	NA
B10206	NA	NA	NA	NA
B10207	NA	NA	NA	NA
B10208	NA	NA	NA	NA
B10209	NA	NA	NA	NA
B10210	NA	NA	NA	NA
B10211	NA	NA	NA	NA
B10212	NA	NA	NA	NA
B10213	NA	NA	NA	NA
B10214	NA	NA	NA	NA
B10215	NA	NA	NA	NA
B10216	NA	NA	NA	NA
B10217	NA	NA	NA	NA
B10218	NA	NA	NA	NA
B10219	NA	NA	NA	NA
B10220	NA	NA	NA	NA
B10221	NA	NA	NA	NA
B10222	NA	NA	NA	NA
B10223	NA	NA	NA	NA
B10224	NA	NA	NA	NA
B10225	NA	NA	NA	NA
B10226	NA	NA	NA	NA
B10227	NA	NA	NA	NA
B10228	NA	NA	NA	NA
B10229	NA	NA	NA	NA
B10230	NA	NA	NA	NA
B10231	NA	NA	NA	NA
B10232	NA	NA	NA	NA
B10233	NA	NA	NA	NA

B10234	NA	NA	NA	NA
B10235	NA	NA	NA	NA
B10236	NA	NA	NA	NA
B10237	NA	NA	NA	NA
B10238	NA	NA	NA	NA
B10239	NA	NA	NA	NA
B10240	NA	NA	NA	NA
B10241	NA	NA	NA	NA
B10242	NA	NA	NA	NA
B10243	NA	NA	NA	NA
B10244	NA	NA	NA	NA
B10245	NA	NA	NA	NA
B10246	NA	NA	NA	NA
B10247	NA	NA	NA	NA
B10248	NA	NA	NA	NA
B10249	NA	NA	NA	NA
B10250	NA	NA	NA	NA
B10251	NA	NA	NA	NA
B10252	NA	NA	NA	NA
B10253	NA	NA	NA	NA
B10254	NA	NA	NA	NA
B10255	NA	NA	NA	NA
B10256	NA	NA	NA	NA
B10257	NA	NA	NA	NA
B10258	NA	NA	NA	NA
B10259	NA	NA	NA	NA
B10260	NA	NA	NA	NA
B10261	NA	NA	NA	NA
B10262	NA	NA	NA	NA
B10263	NA	NA	NA	NA
B10264	NA	NA	NA	NA
B10265	NA	NA	NA	NA
B10266	NA	NA	NA	NA
B10267	NA	NA	NA	NA
B10268	NA	NA	NA	NA
B10269	NA	NA	NA	NA
B10270	NA	NA	NA	NA
B10271	NA	NA	NA	NA
B10272	NA	NA	NA	NA
B10273	NA	NA	NA	NA
B10274	NA	NA	NA	NA
B10275	NA	NA	NA	NA
B10276	NA	NA	NA	NA
B10277	NA	NA	NA	NA

```
B10278              NA          NA       NA        NA
B10279              NA          NA       NA        NA
B10280              NA          NA       NA        NA
B10281              NA          NA       NA        NA
B10282              NA          NA       NA        NA
B10283              NA          NA       NA        NA
B10284              NA          NA       NA        NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47.62 on 30195 degrees of freedom
Multiple R-squared: 0.09945,    Adjusted R-squared: 0.0959
F-statistic: 28.02 on 119 and 30195 DF,  p-value: < 2.2e-16
```

6. R 程序的片段

```
rc<-read.csv("e:/statistics/SE.csv")
A<- matrix(0,nrow=30315,ncol=6)
B<-cbind(rc,A)
B[,3]<-substr(B[1:30315,1],1,2)
C<-c("机票价格","机票价","的","飞机票","航空","飞机","飞","打折","优惠","特惠","钱
","折扣","机票","票价","低价","查询","往返","特价","便宜","-","预定","预订","价格
","订","定","折扣","来回")
for(i in 1:27){B[,7]<-gsub(C[i],"",B[,7])}
B[,4]<-substr(B[1:30315,7],(nchar(B[1:30315,7])-1),nchar(B[1:30315,7]))
 for(i in 1:30315){B[i,5]<-mean(B[grep(B[i,3],B[1:30315,3]),2])}
 for(i in 1:30315){B[i,6]<-mean(B[grep(B[i,4],B[1:30315,4]),2])}
for(i in 1:30315)B[i,7]<-length(grep(B[i,3],B[,3]))
for(i in 1:30315)B[i,8]<-length(grep(B[i,4],B[,4]))
lm.sol<-lm(B[,2]~B[,5]+B[,6])
summary(lm.sol)
plot(resid(lm.sol))
B[,7]<-as.numeric(B[,7])
colnames(B)<-c("关键字","利润","出发地","目的地","出发地平均利润","目的地平均利润
","出发地重复数量","目的地重复数量")
plot(B[,5],B[,2])
Bft<-matrix(0,ncol=1,nrow=30315)
B9<-cbind(B,Bft)
B9[,9]<-(B9[,5]*B[,6])
lm.sol2<-lm(B9[,2]~B9[,5]+B9[,6]+B9[,9])
summery(lm.sol2)
A<- matrix(0,nrow=30315,ncol=2)
B11<-cbind(B9,A)
B11[c(grep(" ",B[,1])),10]<-1
```

```
B11[c(grep("-",B[,1])),11]<-1
lm.sol3<-lm(B11[,2]~B11[,5]+B11[,6]+B11[,9]+B11[,10]+B11[,11])
summary(lm.sol3)
B9.res<-resid(lm.sol2)
B9.fit<-predict(lm.sol2)
plot(B9.res~B9.fit)
local({pkg <- select.list(sort(.packages(all.available = TRUE)))
+ if(nchar(pkg)) library(pkg, character.only=TRUE)})
B4<-cbind(B11[,5],B11[,6],B11[,7],B11[,8],B11[,9],B11[,10],B11[,11])
lars(B4,B11[,2],type="lasso")
B6<-matrix(0,nrow=30315,ncol=29)
B6[,1]<-1
B6[,2]<-B[,5]
B6[,3]<-B[,6]
B6[,4]<-B[,7]
B6[,5]<-B[,8]
B6[c(grep("-",B[,1])),6]<-1
B6[c(grep("到",B[,1])),7]<-1
B6[c(grep("去",B[,1])),8]<-1
B6[c(grep("至",B[,1])),9]<-1
B6[c(grep("飞机票",B[,1])),19]<-1
B6[c(grep("机票",B[,1])),18]<-1
B6[c(grep("飞机票",B[,1])),18]<-0
B6[c(grep("飞",B[,1])),10]<-1
B6[c(grep("飞机",B[,1])),10]<-0
B6[c(grep("便宜",B[,1])),11]<-1
B6[c(grep("打折",B[,1])),13]<-1
B6[c(grep("折扣",B[,1])),14]<-1
B6[c(grep("优惠",B[,1])),16]<-1
B6[c(grep("特惠",B[,1])),17]<-1
B6[c(grep("特价",B[,1])),12]<-1
B6[c(grep("低价",B[,1])),15]<-1
B6[c(grep("价格",B[,1])),23]<-1
B6[c(grep("价",B[,1])),22]<-1
B6[c(grep("特价",B[,1])),22]<-0
B6[c(grep("低价",B[,1])),22]<-0
B6[c(grep("价格",B[,1])),22]<-0
B6[c(grep("往返",B[,1])),20]<-1
B6[c(grep("来回",B[,1])),21]<-1
B6[c(grep("的",B[,1])),24]<-1
B6[c(grep("查询",B[,1])),25]<-1
B6[c(grep("预定",B[,1])),26]<-1
B6[c(grep("定",B[,1])),27]<-1
B6[c(grep("预定",B[,1])),27]<-0
```

```
B6[c(grep("预订",B[,1])),28]<-1
B6[c(grep("订",B[,1])),29]<-1
B6[c(grep("预订",B[,1])),29]<-0
lars(B6,B11[,2],type="lasso")
lm.sol4<-lm(B[,2]~B6[,2]+B6[,3]+B6[,7]+B6[,12]+B6[,14]
+B6[,8]+B6[,4]+B6[,13]+B6[,5])
summary(lm.sol4)
Bsort<-matrix(ncol=2,nrow=30315)
Bsort[,2]<-B[,2]
Bsort[,2]<-sort(Bsort[,2])
Bsort[,1]<-1:30315
plot(Bsort)
quantile(B[,2],probs=seq(0,1,0.01))
Bp<-B
Bp[,1]<-0
Bp[,2]<-0
Bp[,3]<-0
Bp[,4]<-0
for(i in 1:4){Bp[,i]<-as.numeric(Bp[,i])}
Bp<-scale(Bp)
Bp[(B[,2]>=(-26.3))&(B[,2]<(-10.3958)),]->Bp2
Bp[(B[,2]<(-26.3)),]->Bp1
Bp[(B[,2]>=(-10.3958))&(B[,2]<(0)),]->Bp3
Bp[(B[,2]>=(0))&(B[,2]<(9.6286)),]->Bp4
Bp[(B[,2]>=(9.6286))&(B[,2]<(20.07736)),]->Bp5
Bp[(B[,2]>=(9.6286))&(B[,2]<(287.0952)),]->Bp6
Bp[(B[,2]>=(287.0952))&(B[,2]<(4000)),]->Bp7
Bp6[,1:4]<-0
Bp6.pr<-princomp(Bp6)
summary(Bp6.pr,loadings=TRUE)
Bp[,2]<-B[,2]
scale(Bp[,2])->Bp[,2]
load2<-loadings(Bp2.pr)
Bp[(B[,2]>=(-26.3))&(B[,2]<(-10.3958)),2]->Bp2[,2]
plot(load2[5:8,1]%*%t(Bp2[1:282,5:8]),Bp2[1:282,2],xlab="Comp1",ylab="profit2")
load3<-loadings(Bp3.pr)
Bp[(B[,2]>=(-10.3958))&(B[,2]<(0)),2]->Bp3[,2]
plot(load3[5:8,1]%*%t(Bp3[1:282,5:8]),Bp3[1:282,2],xlab="Comp1",ylab="profit3")
load4<-loadings(Bp4.pr)
Bp[(B[,2]>=(0))&(B[,2]<(9.6286)),2]->Bp4[,2]
plot(load4[5:8,1]%*%t(Bp4[1:282,5:8]),Bp4[1:282,2],xlab="Comp1",ylab="profit4")
load5<-loadings(Bp5.pr)
Bp[(B[,2]>=(9.6286))&(B[,2]<(20.07736)),2]->Bp5[,2]
plot(load5[5:8,1]%*%t(Bp5[1:282,5:8]),Bp5[1:282,2],xlab="Comp1",ylab="profit5")
```

```
load6<-loadings(Bp6.pr)
Bp[(B[,2]>=(9.6286))&(B[,2]<(287.0952)),2]->Bp6[,2]
plot(load6[5:8,1]%*%t(Bp6[1:282,5:8]),Bp6[1:282,2],xlab="Comp1",ylab="profit6")
plot(c(-1,3),c(-1,1.5),col="white",xlab="comp1",ylab="profit")
points(load6[5:8,1]%*%t(Bp6[1:282,5:8]),Bp6[1:282,2],col="blue")
points(load2[5:8,1]%*%t(Bp2[1:282,5:8]),Bp2[1:282,2],col="red")
points(load3[5:8,1]%*%t(Bp3[1:282,5:8]),Bp3[1:282,2],col="orange")
points(load4[5:8,1]%*%t(Bp4[1:282,5:8]),Bp4[1:282,2],col="yellow")
points(load5[5:8,1]%*%t(Bp5[1:282,5:8]),Bp5[1:282,2],col="green")
Bp[,2]<-B[,2]
scale(Bp[,2])->Bp[,2]
load2<-loadings(Bp2.pr)
Bp[(B[,2]>=(-26.3))&(B[,2]<(-10.3958)),2]->Bp2[,2]
plot(load2[5:8,2]%*%t(Bp2[1:282,5:8]),Bp2[1:282,2],xlab="Comp2",ylab="profit2")

load3<-loadings(Bp3.pr)
Bp[(B[,2]>=(-10.3958))&(B[,2]<(0)),2]->Bp3[,2]
plot(load3[5:8,2]%*%t(Bp3[1:282,5:8]),Bp3[1:282,2],xlab="Comp2",ylab="profit3")

load4<-loadings(Bp4.pr)
Bp[(B[,2]>=(0))&(B[,2]<(9.6286)),2]->Bp4[,2]
plot(load4[5:8,2]%*%t(Bp4[1:282,5:8]),Bp4[1:282,2],xlab="Comp2",ylab="profit4")

load5<-loadings(Bp5.pr)
Bp[(B[,2]>=(9.6286))&(B[,2]<(20.07736)),2]->Bp5[,2]
plot(load5[5:8,2]%*%t(Bp5[1:282,5:8]),Bp5[1:282,2],xlab="Comp2",ylab="profit5")

load6<-loadings(Bp6.pr)
Bp[(B[,2]>=(9.6286))&(B[,2]<(287.0952)),2]->Bp6[,2]
plot(load6[5:8,2]%*%t(Bp6[1:282,5:8]),Bp6[1:282,2],xlab="Comp2",ylab="profit6")
```

所有主成分2对于利润的图

```
plot(c(-3,3),c(-1,1.5),col="white",xlab="comp2",ylab="profit")
points(load6[5:8,2]%*%t(Bp6[1:282,5:8]),Bp6[1:282,2],col="blue")
points(load2[5:8,2]%*%t(Bp2[1:282,5:8]),Bp2[1:282,2],col="red")
points(load3[5:8,2]%*%t(Bp3[1:282,5:8]),Bp3[1:282,2],col="orange")
points(load4[5:8,2]%*%t(Bp4[1:282,5:8]),Bp4[1:282,2],col="yellow")
points(load5[5:8,2]%*%t(Bp5[1:282,5:8]),Bp5[1:282,2],col="green")
Bp[,2]<-B[,2]
scale(Bp[,2])->Bp[,2]
load2<-loadings(Bp2.pr)
Bp[(B[,2]>=(-26.3))&(B[,2]<(-10.3958)),2]->Bp2[,2]
plot(load2[5:8,3]%*%t(Bp2[1:282,5:8]),Bp2[1:282,2],xlab="Comp3",ylab="profit2")
```

```
load3<-loadings(Bp3.pr)
Bp[(B[,2]>=(-10.3958))&(B[,2]<(0)),2]->Bp3[,2]
plot(load3[5:8,3]%*%t(Bp3[1:282,5:8]),Bp3[1:282,2],xlab="Comp3",ylab="profit3")
load4<-loadings(Bp4.pr)
Bp[(B[,2]>=(0))&(B[,2]<(9.6286)),2]->Bp4[,2]
plot(load4[5:8,3]%*%t(Bp4[1:282,5:8]),Bp4[1:282,2],xlab="Comp3",ylab="profit4")
load5<-loadings(Bp5.pr)
Bp[(B[,2]>=(9.6286))&(B[,2]<(20.07736)),2]->Bp5[,2]
plot(load5[5:8,3]%*%t(Bp5[1:282,5:8]),Bp5[1:282,2],xlab="Comp3",ylab="profit5")
load6<-loadings(Bp6.pr)
Bp[(B[,2]>=(9.6286))&(B[,2]<(287.0952)),2]->Bp6[,2]
plot(load6[5:8,3]%*%t(Bp6[1:282,5:8]),Bp6[1:282,2],xlab="Comp3",ylab="profit6")
plot(c(-3,3),c(-1,1.5),col="white",xlab="comp3",ylab="profit")
points(load6[5:8,3]%*%t(Bp6[1:282,5:8]),Bp6[1:282,2],col="blue")
points(load2[5:8,3]%*%t(Bp2[1:282,5:8]),Bp2[1:282,2],col="red")
points(load3[5:8,3]%*%t(Bp3[1:282,5:8]),Bp3[1:282,2],col="orange")
points(load4[5:8,3]%*%t(Bp4[1:282,5:8]),Bp4[1:282,2],col="yellow")
points(load5[5:8,3]%*%t(Bp5[1:282,5:8]),Bp5[1:282,2],col="green")
Bp[,2]<-B[,2]
Bp[(B[,2]>=(-26.3))&(B[,2]<(287.0952)),]->Bp8
Bp8.pr<-princomp(Bp8[,5:8])
summary(Bp8.pr,loadings=TRUE)
load<-loadings(Bp8.pr)
plot(load[,1]%*%t(Bp8[,5:8]),Bp8[,2],xlab="Comp1",ylab="profit")
plot(load[,2]%*%t(Bp8[,5:8]),Bp8[,2],xlab="Comp2",ylab="profit")
plot(load[,3]%*%t(Bp8[,5:8]),Bp8[,2],xlab="Comp3",ylab="profit")
plot(load[,4]%*%t(Bp8[,5:8]),Bp8[,2],xlab="Comp4",ylab="profit")
lm.comp<-lm(Bp8[,2]~t(load[,1]%*%t(Bp8[,5:8]))+t(load[,2]%*%t(Bp8[,5:8]))+t(load[,3
]%*%t(Bp8[,5:8]))+t(load[,4]%*%t(Bp8[,5:8])))
summary(lm.comp)
B[(B[,2]>=(-26.3))&(B[,2]<(287.0952)),]->Bnew
lm5<-lm(Bnew[,2]~Bnew[,5]+Bnew[,6]+Bnew[,7]+Bnew[,8])
summary(lm5)
quantile(Bnew[,5],probs=seq(0,1,0.1))->a
a1<-(1:11)
for(i in 1:11){as.numeric(a[i])->a1[i]}
quantile(Bnew[,6],probs=seq(0,1,0.1))->b
b1<-(1:11)
for(i in 1:11){as.numeric(b[i])->b1[i]}
quantile(Bnew[,7],probs=seq(0,1,0.1))->c
c1<-(1:11)
for(i in 1:11){as.numeric(c[i])->c1[i]}
quantile(Bnew[,8],probs=seq(0,1,0.1))->d
d1<-(1:11)
```

```
for(i in 1:11){as.numeric(d[i])->d1[i]}
B88<-matrix(0,ncol=88,nrow=30171)
for(i in 1:10)
{B88[(Bnew[,5]>=a1[i]),(8+i)]<-1}
for(i in 1:10)
{B88[(Bnew[,6]>=b1[i]),(18+i)]<-1}
for(i in 1:10)
{B88[(Bnew[,7]>=c1[i]),(28+i)]<-1}
for(i in 1:10)
{B88(Bnew[,8]>=d1[i]),(38+i)]<-1}
B88[,2]<-Bnew[,2]
B88[,5]<-Bnew[,5]
B88[,6]<-Bnew[,6]
B88[,7]<-Bnew[,7]
B88[,8]<-Bnew[,8]
for(i in 1:10)
{for(j in 1:30171){B88[j,48+i]<-B88[j,8+i]*B88[j,5]}}
for(i in 1:10)
{for(j in 1:30171){B88[j,58+i]<-B88[j,18+i]*B88[j,6]}}
for(i in 1:10)
{for(j in 1:30171){B88[j,68+i]<-B88[j,28+i]*B88[j,7]}}
for(i in 1:10)
{for(j in 1:30171){B88[j,78+i]<-B88[j,38+i]*B88[j,8]}}
lars(B88[,5:88],B88[,2],type="lasso")->lars4
summary(lars4)
plot(lars4)
lm7<-lm(B88[,2]~B88[,5:88])
summary(lm7)
plot(lm7)
lars(B88[,9:88],B88[,2],type="lasso")->lars4
summary(lars4)
plot(lars4)
B128<-matrix(0,ncol=138,nrow=30171)
B128[,1:88]<-B88[,1:88]
for(i in 1:40){B128[,88+i]<-exp(B128[,48+i]/10)}
lm8<-lm(B128[,2]~B128[,5:128])
a2<-c(-10,-6,-2,2,6,10,14,18,22,26,30)
b2<-c(-12,0,12,24,36,48,60,72,84,96,108)
c2<-c(0,300,600,900,1200,1500,1800,2100,2400,2700,3000)
d2<-c(0,200,400,600,800,1000,1200,1400,1600,1800,2000)
B88n<-matrix(0,ncol=88,nrow=30171)
for(i in 1:10)
{B88n[(Bnew[,5]>=a2[i]),(8+i)]<-1}
for(i in 1:10)
```

```
{B88n[(Bnew[,6]>=b2[i]),(18+i)]<-1}
for(i in 1:10)
{B88n[(Bnew[,7]>=c2[i]),(28+i)]<-1}
for(i in 1:10)
{B88n[(Bnew[,8]>=d2[i]),(38+i)]<-1}
B88n[,2]<-Bnew[,2]
B88n[,5]<-Bnew[,5]
B88n[,6]<-Bnew[,6]
B88n[,7]<-Bnew[,7]
B88n[,8]<-Bnew[,8]
for(i in 1:10)
{for(j in 1:30171){B88n[j,48+i]<-B88n[j,8+i]*B88n[j,5]}}
for(i in 1:10)
{for(j in 1:30171){B88n[j,58+i]<-B88n[j,18+i]*B88n[j,6]}}
for(i in 1:10)
{for(j in 1:30171){B88n[j,68+i]<-B88n[j,28+i]*B88n[j,7]}}
for(i in 1:10)
{for(j in 1:30171){B88n[j,78+i]<-B88n[j,38+i]*B88n[j,8]}}
lm(B88n[,2]~B88n[,9:88])
lm9<-lm(B88n[,2]~B88n[,9:88])
summary(lm9)
lm10<-lm(Bnew[,2]~B6)
summary(lm10)
Lst<-list(c("-","到","去","至","飞"),c("飞机票","机票"),c("便宜","打折","折扣","优惠","特惠","
特价","低价"),c("价格","价"),c("往返","来回"),c("的"),c("查询","预定","预订","定","订"))
B7<-matrix(0,nrow=30171,ncol=254)
B8<-matrix(0,nrow=2,ncol=254)
l<-c(5,2,7,2,2,1,5)
k = 0
for (i in 1:6){
for (j in 2:7){
if (any(i<j)) {
for (s in 1:9){
for (t in 1:7){
if (any(s<=l[i])){
if (any(t<=l[j])){
k = k + 1
B8[1,k]<-Lst[[i]][s]
  B8[2,k]<-Lst[[j]][t]
m <-sum(l[1:i])-l[i]+s
n <-sum(l[i:j])-l[j]+t
for(o in 1:30171)B7[o,k]<-B6[o,m+5]*B6[o,n+5]} }   }}}}}
```

7. 参考文献

【1】统计学习基础——数据挖掘，推理与预测/（美）黑斯蒂（Hastie，T.）等著；范明等译，北京：电子工业出版社，2004.1

【2】应用线性回归/（美）S.韦斯伯格（Weisberg, S.）；王静龙等译，
北京：中国统计出版社，1998.3.

【3】应用多元统计分析/高惠璇编著. 北京：北京大学出版社，2005.1

【4】Profitable Key Word Selection： Paid Search Advertising For Online Airplane Ticket Booking/Minghua.Jiang, Xuefeng Li, Chih-Ling Tsai, Hansheng Wang

【5】Statistics with R/Vincent Zoonekynd, 2005.8

特别感谢：国家，姚远老师，王汉生老师，虞高然同学