北京大学

学士学位论文

题目 付费搜索广告在航空机票领域的最优关键词选取

姓　　名　虞高然

学　　号　00601037

院　　系　数学科学学院

专　　业　概率统计

指导教师　姚远

2010年7月

Key Words Selection Strategy in Paid Search Advertising for Online Airplane Ticket Booking

Gaoran Yu

Supervisor: Yuan Yao

School of Mathematical Sciences, Peking University

June, 2010

Submitted in total fulfilment of the requirements for the degree of Bachelor in Statistics

版权声明

摘　　要

付费搜索广告（搜索引擎竞价排名）作为一种新型的广告形式引起了诸多企业的关注。本文试图通过对付费搜索广告在航空机票预订领域为使用它的企业带来利润的大小进行分析，从而确立合理的广告投资方法，为企业在该领域购买广告关键词提供有价值的参考。

关键词： 付费搜索广告、航空机票、关键词、回归、LARS

Abstract

As a new kind of advertisement，Paid Searching Advertising is getting more and more popular among companies and enterprises. This essay is aimed at providing valuable suggestions for those who plans to invest in this kind of advertisement in the field of Online Airplane Ticket Booking. Our research is based on the data collected from the profits that Paid Searching Advertising has brought to its users in this special field.

Keywords: Paid Searching Advertising, Online Airplane Ticket Booking, Keywords Selections、Linear Regression、Least Angle Regression

目　　录

摘要 i

Abstract ii

目录 iii

第一章　引言 1

第二章　初期数据处理 3

第三章　简单的回归分析和变量选择： 7

3.1　关于地理因素的回归分析 7

3.2　LARS变量选择 14

第四章　主成分分析 19

4.1　变量分类 19

4.2　主成分分析 20

4.3　主成分对于利润做简单线性回归 31

第五章　交叉变量的生成和变量选择 36

5.1　交叉变量的生成 36

5.2　LARS变量选择结果 36

第六章　存在问题及改进 38

第七章　总结 40

第八章　附录 41

8.1　含交叉变量的LARS变量选择过程： 42

8.2　R程序代码： 45

致谢 48

第一章　引言

所谓的“搜索引擎竞价排名”，即“付费搜索广告”，是指在某些搜索引擎（如百度）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。相较于固定付费广告，“搜索引擎竞价排名”具有如下优势：

1 排名可靠，用科学的几大参数来决定排名位置，

2 自主投放地域，选择自己的潜在客户群所在地进行投放，精准锁定目标客户群，获取更多效益。

3 搜索推广不仅仅以文字形式进行展现，同时百度也最新推出百度网盟推广，联合30万家大型网站，以图片、flash动画的形式进行推广展现，更具吸引力，每天过亿的展现。

4 消费透明，每天的数据统计报告详细记载每一笔开支，精准的百度统计工具完善的统计客户的进出入状态，有针对性的分析当前市场的竞争优势。

5 保障客户的隐私，搜索推广的一种受保护的推广模式，百度是绝不公开客户的推广信息和数据，妥善保管客户的推广计划。

得益于其诸多无可比拟的优势，“搜索引擎竞价排名”已逐渐成为各企业青睐的对象。然而，其另一特点在于：付费搜索广告需要企业为每一次用户点击付费。所以对企业来说，其广告投入相比固定付费广告会更为高昂。因此，购买合适的关键词，从而使企业获得较好宣传效果的同时，尽可能的降低成本，成为一个亟待解决的问题。

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，所以销售类的网站便可能使这些访客转化为消费者。对于航空公司以及其各代理商，最关心的问题就是如何购买特定的关键词以吸引更多的旅客购买自己的机票，从而提升公司自身的利润。然而，根据中国民航局（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。并且，由于中国文化博大精深，国民的表达习惯各不相同，由地名、介词、修饰性短语构成的关键词数量将会非常庞大。若以本文采用

的数据为凭，如果企业希望覆盖中国航空网络的各种可能出现的关键词搜索，至少需要购买5000000个关键词。无论企业的资金多么雄厚，这对于任何一个企业来讲，都是一笔巨额的开支。因此，本文试图通过对30315个先前购买关键词的企业盈利状况的分析调查，得出关键词中各个部分（如地名、修饰性短语、介词）对于最后企业获利的影响，从而为企业选取一种更为理性的关键词投资方式。

第二章　初期数据处理

本文的原始数据由北京大学光华管理学院的王汉生教授提供。是一个30315 × 2的矩阵。其第一列表示航空机票搜索的关键词，第二列表示该关键词搜索为广告厂商提供的利润。下面提供原始数据的前十行样例：

关键词	利润
乌鲁木齐-阿克苏-机票	14.12
乌鲁木齐阿克苏飞机票价	9.06
乌鲁木齐到阿克苏-机票	-1.18
乌鲁木齐到阿克苏打折机票	-0.48
乌鲁木齐到阿克苏机票	31.94
乌鲁木齐-阿勒泰-机票	-1.14
乌鲁木齐-阿勒泰-特价机票	-0.49
乌鲁木齐阿勒泰订机票	9.58
乌鲁木齐阿勒泰飞机票	-0.49

虽然每行数据仅仅是二维向量，但事实上，由于第一列包含的文字信息十分庞大，需要经过特殊处理，才能化为适合处理的数据。

根据人工对前10000个样本的观察和统计，我们可将关键词分为如下三个部分：

出发地点、抵达地点、连接词

其中出发点和抵达地点均由地名构成，共有172种不同的地名，如“北京”,“乌鲁木齐”等，而连接词共有25种，我们将其分为7类，按中国公民的语言习惯，同一类别任意两类不会出现在同一关键词中。

第一类：”-”,”到”,”去”,”飞”,”至”

第二类：”飞机票”,”机票”,”飞机”

第三类：”便宜”,”打折”,”折扣”,”优惠”,”特惠”,”特价”,”低价”

第四类：”价格”,”价”

第五类：”往返”,”来回”

第六类：”的”

第七类：”查询”,”预定”,”预订”,”定”,”订”

对于字符串型自变量，通常采用化为拟变量的方式，即如果关键词中包含该特定字符串，则令其对应的自变量值为1，否则值为0。如：

$$f(x)=\begin{cases}1 & \text{当“到”在关键词中}\\ 0 & \text{其他情形。}\end{cases}$$

对于每个关键字，我们对其做如下处理：

一、将原始数据矩阵赋予矩阵B

二、对矩阵B进行处理，提取其出发地点和抵达地点

三、对具有相同出发地点的关键词所对应的利润进行算术平均，并对关键词的个数进行统计

四、对25个不同的连接词设立25个拟变量，并对根据每个关键词所含的连接词确立25个拟变量的数值

需要说明的是：

在对关键词提取出发地点及抵达地点的处理中，由于涉及的地名过多，长度也不统一，所以提取的难度较大，于是我们采取一种较为方便的方法：将关键词中的所有上述25个连接词全部剔除，这样关键词中仅涉及地理信息。将地理信息的前两个字作为出发地点，最后两个字作为抵达地点。事实上，我们并不关心每个地名提取是否准确，只要在数据处理上不引起混淆即可。如“爱尔兰”作为出发地被命名为“爱尔”,而作为抵达地点时被命名为“尔兰”。由于所有地名中，任取两个，其开头的两个字符互不相同，而结尾的两个字符也互不相同。所以如此处理并不会引起混淆，使得两个不同的地名被视为相同。

此外，在建立拟变量的过程中，由于某些连接词是“包含于”其他连接词中的，如“飞”包含于“飞机”,“飞机”包含于“飞机票”。而事实上，用户在键入关键词时，其分别代表不同的语言习惯和表达涵义。所以，在确立拟变量的数值时，不仅仅是一个字符串匹配的过程。还需要判断该用户键入这个字符串是否真的要表达相应的意思。例如：我们需要检验关键词中是否有“飞”，这里“飞”表示“到”

或“至”的意思。但如果关键词中“飞”仅以“飞机票”的形式出现，我们便不该认为“飞”对应的拟变量为1. 为了防止类似的情况发生，我们将关键词以特定的形式排列：”-”,”到”,”去”,”至”,”飞机票”,”机票”,”飞机”,”飞”,”便宜”,”打折”,”折扣”,”优惠”,”特 惠”,”特 价”,”低 价”,”价 格”, ”价”,”往 返”,”来 回”,”的”,”查 询”,”预 定”,”预订”,”定”,”订”。其中可能包含于其他连接词的连接词会置于包含它的连接词的后面。算法采用标准的字符串匹配，但是对于每个连接词，在匹配之后将它从关键词中剔除，这样便能避免上述问题。

具体的R语言代码请见附录，并将得到的矩阵几位B6

为了方便之后的运算，将所得矩阵分为两部分：
矩阵B为30315 × 8的矩阵，其前6行列举如下：

关键词	利润	出发地点	抵达地点
乌鲁木齐阿克苏	14.12	乌鲁	克苏
乌鲁木齐阿克苏	9.06	乌鲁	克苏
乌鲁木齐阿克苏	-1.18	乌鲁	克苏
乌鲁木齐阿克苏	-0.48	乌鲁	克苏
乌鲁木齐阿克苏	31.94	乌鲁	克苏
乌鲁木齐阿勒泰	-1.14	乌鲁	克苏

出发地利润	目的地利润	出发地次数	目的地次数
11.74461	10.692	490	5
11.74461	10.692	490	5
11.74461	10.692	490	5
11.74461	10.692	490	5
11.74461	10.692	490	5
11.74461	0.820	490	5

矩阵B6为30315 × 30的矩阵，其前5列举如下：

常系数1	出发点平均利润	目的地平均利润	出发点重复次数	目的地重复次数
1	11.74461	10.692	490	5
1	11.74461	10.692	490	5
1	11.74461	10.692	490	5
1	11.74461	10.692	490	5
1	11.74461	10.692	490	5
1	11.74461	0.820	490	5

对于B6的后25列，为各种连接词在关键词中的出现状况。这里仅列其中的前9列：

-	到	去	至	飞机票	机票	飞机	飞	便宜
1	0	0	0	0	1	0	0	0
0	0	0	0	1	0	0	0	0
1	1	0	0	0	1	0	0	0
0	1	0	0	0	1	0	0	0
0	1	0	0	0	1	0	0	0
1	0	0	0	0	1	0	0	0

第三章　简单的回归分析和变量选择：

3.1　关于地理因素的回归分析

由于自变量关键词中包含巨大的文字信息，包括172个地名和25个连接词，如果将其统一视为拟变量处理，其总共有2^{197}种不同的取值。但是，其取值的不连续性使得这个问题更接近于一个分类问题，虽然使用SVM分类方法可以将处于197维空间的30315个点分类。但对于每个自变量来讲，我们很难直观地发现其对利润的影响。在核函数的选取方面，也不容易找到理想的核函数从而对其本征展开式进行线性组合以获得理想的超曲面。这里我们更倾向于用一种简单直观的方法来选取合适的自变量从而对利润进行回归，这个自变量可视为在某区间内连续。很显然地，每个不同的地名对于利润都有不同的影响，由于缺乏每个地域必要的信息（如人口、GDP、航空客流量、消费偏好等），所以很难准确地估计地名对利润的影响。这里采取一种简单的做法：将含有相同出发地的关键词对应的利润的算术平均视为该出发地对于利润的影响值，对目的地也做此处理。这个做法并不是非常准确，由于每个关键词的搜索频率互不相同，从直观上来讲，搜索频率更高的关键词应该赋予更多的权值。关于如何改进，将在后文进行讨论。

首先我们将出发地点的利润平均和抵达地点的利润平均作为两个自变量，原始的利润作为因变量来进行回归，假设出发地的利润平均为X，抵达地的利润平均为y，则利润profit = f(x,y);
首先我们尝试了线性模型：即porfit = ax + by,在R中输入语句：

```
lm.sol<-lm(B[,2]~B[,5]+B[,6])
summary(lm.sol)
```

得到如下结果：

Call:
lm(formula = B[, 2] B[, 5] + B[, 6])

Residuals:

Min	1Q	Median	3Q	Max
-88.600	-16.683	-8.868	1.818	3035.175

Coefficients:

	Estimate Std.	Error t	value	$Pr(>\|t\|)$
(Intercept)	-18.53829	0.99429	-18.64	<2e-16 ***
B[, 5]	1.22615	0.05331	23.00	<2e-16 ***
B[, 6]	1.12183	0.03742	29.98	<2e-16 ***

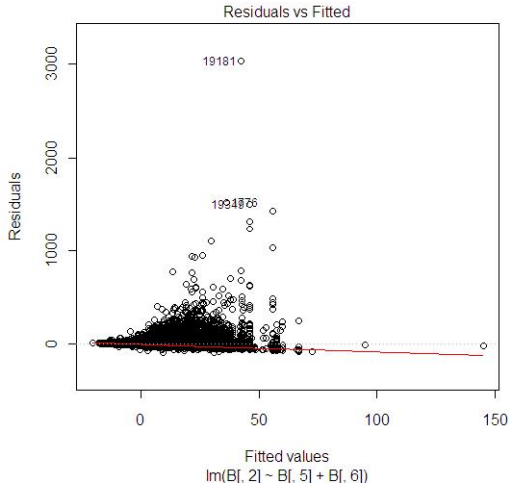
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘’ 1

Residual standard error: 49.08 on 30312 degrees of freedom
Multiple R-squared: 0.03984, Adjusted R-squared: 0.03977
F-statistic: 628.8 on 2 and 30312 DF, p-value: < 2.2e-16

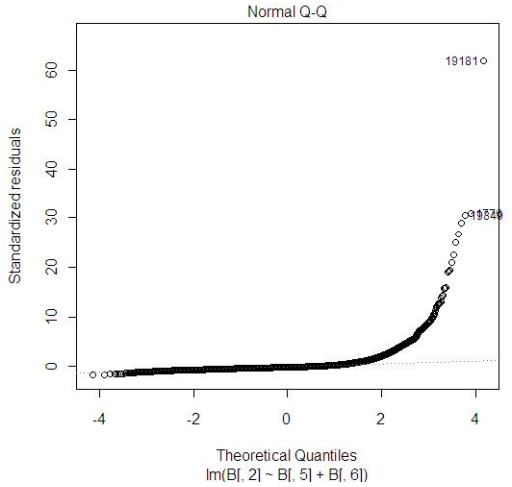
从统计学的角度来讲，这个结果已经相当令人满意了，其自变量的显著性检验的p值均在10^{-16}量级，而F统计量对应的p值小于2.2×10^{-16}。利润和出发地点利润平均、抵达地点利润平均有较为显著地线性关系。

从实际角度来看，出发地点的利润平均和抵达地点的利润平均前面的系数十分相近，分别为1.22615和1.12183，说明出发地点和目的地点对利润的影响比较相近，这从直观上来讲非常符合实际状况。

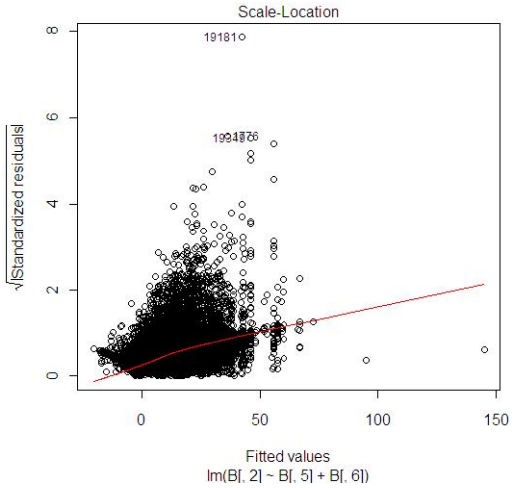
下面是利用R语言绘出的残差图、标准化QQ图等，从而做更进一步的分析和改进



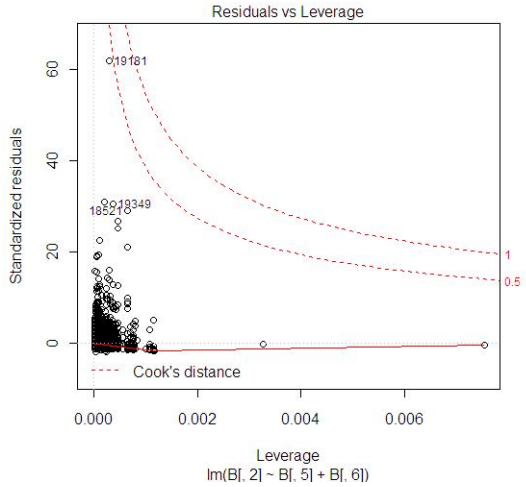
caption 图1 Residuals vs Fitted



caption 图2 Normal QQ



caption 图3 Scale-Location



caption 图4 Residuals vs Leverage

从图1我们不难得出线性回归对较低的利润拟合较好，但对于高额利润率，仍存在较大的误差。从QQ图中我们也能看出明显的重尾，说明利润较高的部分仍然拟合地不够准确。从图4的Cook距离可以看出在30315个数据中，尽管直观上有利润高达3000多的不合常理的情况，但从统计学的角度来讲，没有异常值的出现。对于前面提到的问题，这里我们提出三种改进方案：

1.采用多项式回归，增加二次项甚至三次项以保证在利润增长后回归曲线的增长率会随之增加。

2.采用指数函数进行非线性回归，指数函数能够对多数利润模型做出很好的解释。

3.采用分段多项式和样条，进行分段回归。

首先尝试一下多项式回归：

先假定多项式最高此项为2，在R中键入如下命令：

```
C<-matrix(0,nrow = 30315, ncol = 3)
C<-B[,c(2,5,6)]
lm.sol<-lm(C[,1]~1+C[,2]+I(C[,2]^2)+C[,3]+I(C[,3]^2)
+I(C[,3]*C[,2])
summary(lm.sol)
```

我们得到：

Call:

$lm(formula = C[,] 1 + C[,2] + I(C[,2]^2) + C[,3] + I(C[,3]^2) + I(C[,3] \times C[,2])$

Residuals:

Min	1Q	Median	3Q	Max
-89.107	-16.917	-8.944	1.975	3034.835

Coefficients:

	Estimate	Std.Error	t value	$Pr(>\|t\|)$	
(Intercept)	8.4339382	2.2169274	3.804	0.000142	***
C[, 2]	-0.6522197	0.2081290	-3.134	0.001728	**
$I(C[,2]^2)$	-0.0058526	0.0053612	-1.092	0.274988	
C[, 3]	-0.6813604	0.1286874	-5.295	1.2e-07	***
$I(C[,3]^2)$	-0.0005025	0.0023895	-0.210	0.833439	
I(C[, 3]×C[, 2])	0.1366380	0.0067798	20.154	<2e-16	***

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

Residual standard error: 49.07 on 30310 degrees of freedom

Multiple R-squared: 0.04031, Adjusted R-squared: 0.04019

F-statistic: 318.3 on 4 and 30310 DF, p-value: <2.2e-16

这里面C[，2]表示出发地利润平均，而C[，3]表示目的地利润平均。由于出发

地利润平均和目的地利润平均的显著性很差，所以剔除这两项后进行回归。

在R中输入：

```
lm.sol<-lm(C[,1]~1+C[,2]+C[,3]+I(C[,3]*C[,2]))
summary(lm.sol)
```

得到：

Call:

$lm(formula = C[,] 1 + C[,2] + C[,3] + I(C[,2] \times C[,3]))$

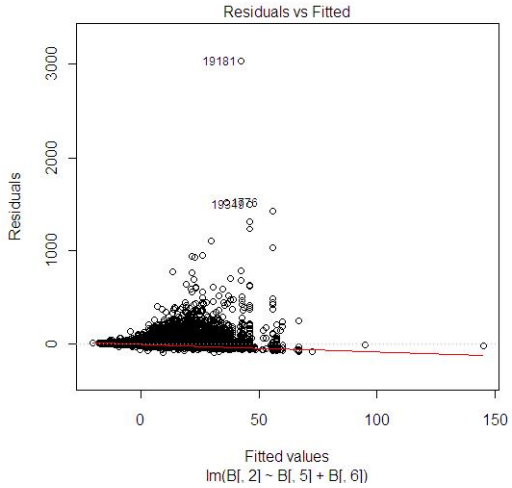
Residuals:

Min	1Q	Median	3Q	Max
-155.346	-15.472	-8.328	1.347	3018.793

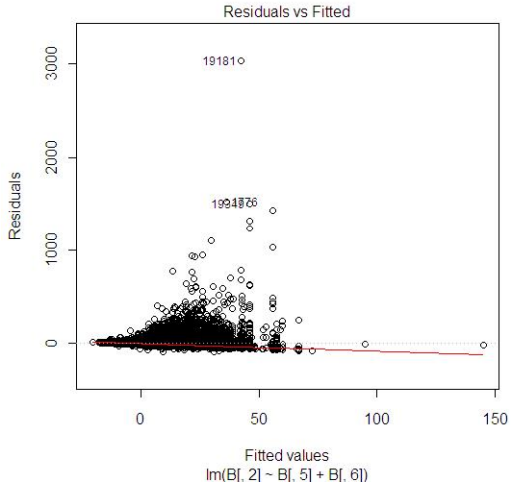
Coefficients:

	Estimate	Std.Error	t value	$Pr(>\|t\|)$	
(Intercept)	9.963607	1.705340	5.843	5.19e-09	***
C[, 2]	-0.842595	0.113959	-7.394	1.46e-13	***
C[, 3]	-0.717532	0.097120	-7.388	1.53e-13	***
$I(C[,3] \times C[,2])$	0.137662	0.006715	20.500	<2e-16	***

由于在原来线性回归的基础上添加了二次交叉项，多项式回归的残差平方和自然会小于先前的模型。但通过对残差图和QQ图的观察（图5和图6）我们并没有发现这个模型有很大的改进。事实上，考虑到该模型的复杂程度和实际意义，（一般来讲，旅客的出发地和目的地是相互独立的，所以交叉项的实际意义远没有在回归方程中来的重要）其直观性和简洁性不如之前的线性模型，与实际的联系也不是十分紧密。因此多项式模型的尝试并不成功。



caption 图5 Residuals vs Fitted



caption 图6 Normal QQ

如果采用指数函数：（由于大多数利润分析的情况下取指数函数能对利润做到很好的拟合，所以在这里我们还是有必要讨论一下这个问题）

profit=Ae^{Bx+Cy}的模型，注意到由于e^{Bx+Cy}恒正，所以profit的回归值只可能是恒为正或是恒为负的，显然对于长期盈利的公司，这种模型可能合适，但对于“付费搜索广告”这种高风险高回报，并且收益率普遍低于零的行业来讲，这显然不是个可取的模型。

另外一个可能的模型是：profit=$Ae^{Bx}+Ce^{Dy}$。为了不使拟合值恒为正或是恒为负，系数A，C必然异号。但就实际情况而言，A，C中任意一个为负代表着总利润会随出发地点平均利润（或是抵达地点平均利润）的上升而下降。公司购买会带来

高利润的关键词，其最终利润反而会因此而下降，这显然也是有悖于常理的。所以这种模型也不做过多讨论。

3.2 LARS变量选择

在加上代表25个连接词的拟变量和2个分别表示出发地和目的地在关键词中出现次数的自变量，线性回归中的自变量数目已经很多了。所以我们采用LARS算法进行变量选择。

首先简要介绍一下LARS算法的原理：

对于模型（1.1），假定

$$\sum_{i=1}^{n} y_i = 0,\ \sum_{i=1}^{n} x_{ij} = 0,\ \sum_{i=1}^{n} x_{ij}^2 = 1,\ j = 1, 2, \ldots, p.$$

开始置估计 $\hat{\mu}$ 为零，找到与 Y 相关性最大的变量，不妨设为 x_i，加入模型，$\hat{\mu}$ 沿着相关方向前进，直到有其他变量与残差的相关性与 x_i 相同，设为 x_j。这时，残差在 x_i, x_j 所张成平面上的投影与这两个变量的夹角相同。$\hat{\mu}$ 再沿着此投影方向前进，直到有第三个变量与现在的残差的相关性和前两个变量相同。$\hat{\mu}$ 前进的方向变为残差在这三个变量所张平面的投影方向。继续如此下去，变量会按照重要性一一加入模型。举个例子，当 $X = (x_1, x_2)$ 时，如图7。

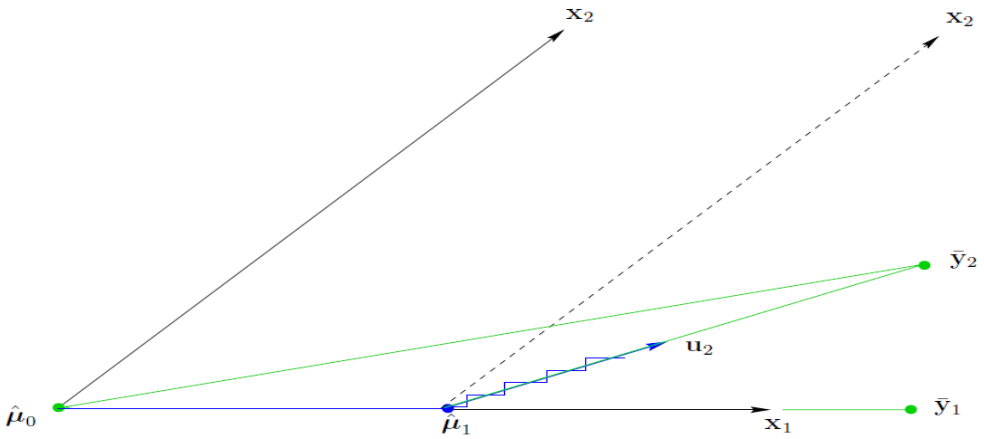


图 3.1: 图7：LARS:$\bar{y_2}$是y在x_1, x_2所张平面的投影，从$\hat{\mu}_0$出发，前进到$\hat{\mu}_1$，再沿着u_2前进。

设 $\mathcal{A}$ 为{1,2,...,p}的子集,定义矩阵

$$X_{\mathcal{A}} = (\cdots s_j x_j \cdots)_{j\in\mathcal{A}}, \tag{2.11}$$

其中 $s_j = \pm 1$. 令

$$\mathcal{G}_{\mathcal{A}} = X'_{\mathcal{A}} X_{\mathcal{A}},\ A_{\mathcal{A}} = (1'_{\mathcal{A}} \mathcal{G}_{\mathcal{A}}^{-1} 1_{\mathcal{A}})^{-\frac{1}{2}}, \tag{2.12}$$

则等角向量为：

$$u_{\mathcal{A}} = X_{\mathcal{A}} w_{\mathcal{A}},\ w_{\mathcal{A}} = A_{\mathcal{A}} \mathcal{G}_{\mathcal{A}}^{-1} 1_{\mathcal{A}} \tag{2.13}$$

容易推导，$u_{\mathcal{A}}$ 满足：

$$X'_{\mathcal{A}} u_{\mathcal{A}} = A_{\mathcal{A}} 1_{\mathcal{A}},\ \|u_{\mathcal{A}}\|^2 = 1.$$

由于R语言中提供LARS的程序包，在R中键入如下命令行：

```
local({pkg <- select.list(sort(.packages(all.available = TRUE)))
+ if(nchar(pkg)) library(pkg, character.only=TRUE)})
lars(B6,B[,2],type = "lar")
```

得到：

Call:

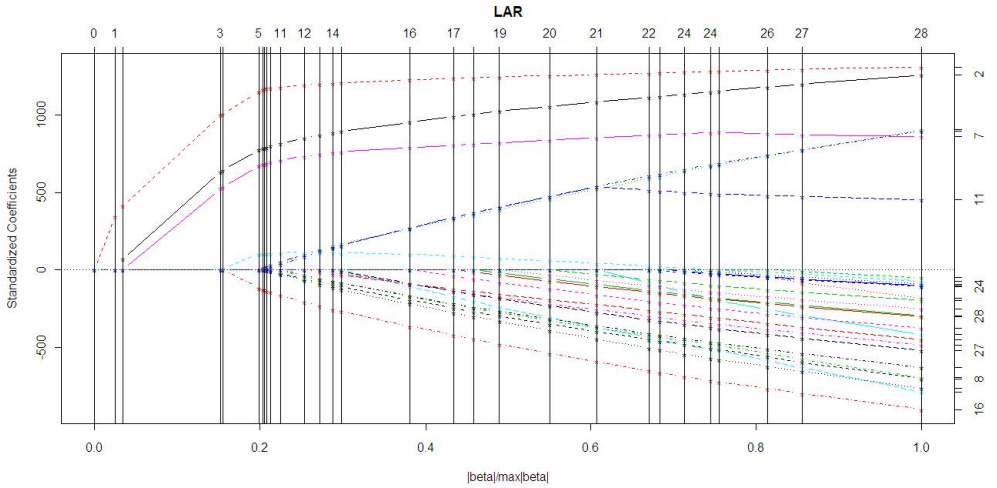
lars(x = B6, y = B[, 2], type = ”lar”)

R-squared: 0.089

Sequence of LAR moves:

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

	Df	Rss	Cp
0	1	76044721	2938.956
1	2	75248209	2592.667
2	3	74993527	2483.302
3	4	72685966	1476.279
4	5	72658437	1466.241
5	6	72166062	1252.941
6	7	72165853	1254.850
7	8	72123128	1238.168
8	9	72109573	1234.240
9	10	72081070	1223.777
10	11	72041262	1208.370
11	12	71934193	1163.552
12	13	71691825	1059.572
13	14	71547106	998.291
14	15	71431177	949.599
15	16	71357869	919.544
16	17	70823419	687.846
17	18	70533340	563.003
18	19	70417612	514.399
19	20	70271079	452.325
20	21	70019609	344.365
21	22	69822673	260.251
22	23	69640942	182.786
23	24	69610042	171.274
24	25	69541954	143.502
25	26	69462564	110.787
26	27	69375537	74.733
27	28	69329942	56.795
28	29	69261802	29.000



captionLARS

从LARS的变量选择结果我们不难看出，目的地的利润平均和出发地的利润平均(var3,var2)分别在第一个和第二个被选入，且系数也是所有自变量里面最大的。这与前面的结论相吻合。而出发地出现的次数与目的地出现的次数(var4,var5)分别在第六和第十位选入，并且系数也相对较小，说明出发地和目的地地名的搜索热门程度对利润的影响相对较小。此外，连接词“到”(var7)作为第三个被选入的变量，具有较大的正回归系数，所以我们认为含有“到”这个连接词对利润的提高石油明显帮助的。而且，对于已经固定的航线，连接词的选择更是尤为重要，在此推荐企业能够选择“到”作为购买的关键词之一。还有一个个比较特殊的变量“特价”(var19),作为第四个被选入的变量,“特价”在被选入后随着选入的连接词的个数的增加，其系数逐渐向负轴偏移，因此我们认为“特价”在连接词较少、关键词句式较为简单的情况下不失为一种明智的投资对象，在其他情况下，由于它的回归系数为负，我们并不推荐选用此变量。与“特价”类似的还有“折扣”(var11)，作为第五个选入的变量，我们对于该关键词有与“特价”相同的评价。最后提一下比较常见的几个连接词:“机票”、“去”、“打折”，他们均对利润有积极的影响。由于三条曲线过于集中，很难分辨，根据它们的纵截距我们看出三者显著性均很优秀，且回归系数较大，对利润的贡献十分突出。

根据表格中CP值的分布，我们不难看出在所有变量均选入的时候CP值最小，所以可以认为30个变量都比较显著。如果要分辨某特定连接词或是地名变量，只需查证它的变量序号根据B6的构成原理（”出发地利润平均”、”目的的利润平均”、”出发地次数”、”目的地次数”、”-”,”到”,”去”,”至”,”飞机票”,”机票”,”飞

机”,”飞”,”便宜”,”打折”,”折扣”, ”优惠”,”特惠”,”特价”,”低价”,”价格”,”价”,”往返”,”来回”,”的”,”查询”,”预定”,”预订”,”定”,”订”）查证它对应的变量序号然后回到Lars变量选择结果中找到该变量是在第n次被选入。从而在图中找到相应的曲线（左数第n个竖线与横轴交点为起点的曲线）。曲线在右边纵轴上的截距即代表该变量的回归系数的大小，也就是它对利润的影响。

第四章　主成分分析

4.1　变量分类

首先我们对利润做一下分位数表：

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

虽然利润的整体跨度较大，从-79.2到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(62%) 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

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。

4.2 主成分分析

下面确定一下参与主成分分析的自变量。根据上文的回归结果我们不难发现，

对利润影响最主要的部分还是地理因素。对于地理因素而言，出发地点和目的地点的利润平均固然十分重要，应作为自变量。此外出发地点和目的地点在关键词中的出现次数，一定程度上反应了该地点的搜索频率。从而我们选用出发地平均利润,目的地平均利润,出发地出现次数,目的地出现次数这4个变量进行分析。

在剔除异常值（利润率>287.0952或<-26.3的样本）后，我们得到新的矩阵：Bnew。

在R中键入如下命令：

```
B[(B[,2]>=(-26.3))&(B[,2]<(287.0952)),]->Bnew
Bp6.pr<-princomp(B6[,2:5]) summary(Bp6.pr,loadings=TRUE)
```

得到如下结果：

高负盈利：

Importance of components:

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

Loadings:

关键字	Comp.1	Comp.2	Comp.3	Comp.4
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
Standard deviation	1.4464187	0.9637858	0.8840008	0.40269940
Proportion of Variance	0.5276974	0.2342922	0.1971070	0.04090335
Cumulative Proportion	0.5276974	0.7619896	0.9590967	1.00000000

Loadings:

关键字	Comp.1	Comp.2	Comp.3	Comp.4
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
Standard deviation	1.4133849	0.9661415	0.8881492	0.43333155
Proportion of Variance	0.5112141	0.2388710	0.2018616	0.04805323
Cumulative Proportion	0.5112141	0.7500851	0.9519468	1.00000000

Loadings:

关键字	Comp.1	Comp.2	Comp.3	Comp.4
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
Standard deviation	1.377948	0.9312530	0.9048278	0.42304462
Proportion of Variance	0.504494	0.2304231	0.2175316	0.04755135
Cumulative Proportion	0.504494	0.7349170	0.9524486	1.00000000

Loadings:

关键字	Comp.1	Comp.2	Comp.3	Comp.4
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

超高正盈利：

Importance of components:

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

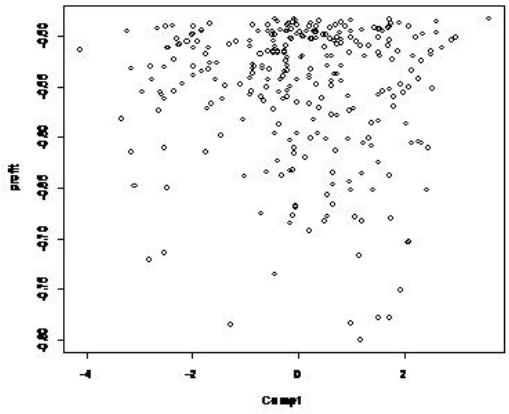
Loadings:

关键字	Comp.1	Comp.2	Comp.3	Comp.4
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

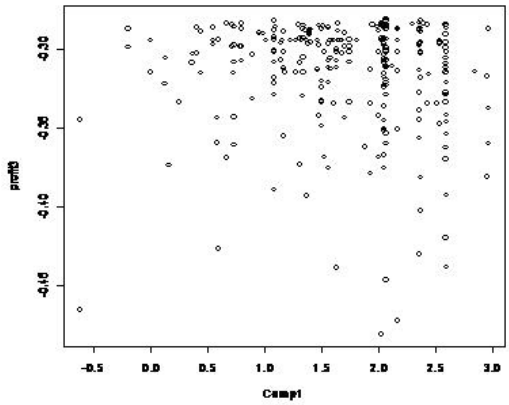
比较上面几组的主成分分析，我们可以看到关于地理因素的主成分模式是基本一致的，不随盈利的不同组别而产生明显变化，第一主成分均为4项的线性组合，值得注意的是与出发地有关的两项均具备正系数，而与目的地相关的两项均有负系数，而且关于目的地的系数绝对值较大。第一个主成分大致占到总信息量的一半左右，反映的信息主要是出发地目的地之间的差异，并且对于目的地的情况更加敏感。另外，将第一主成分中的四项系数相加，可以发现收益率绝对值越高的组别其系数和的绝对值也越高，可以理解为收益率离原点偏差越大，其起始点与终点的差异越能决定最终收益率的大小。

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。

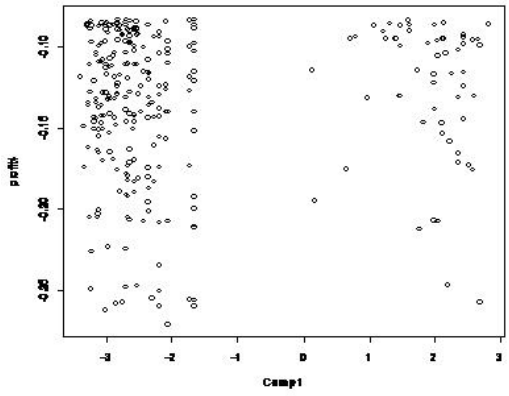
下面我们给出各个主成分非别与上述的5类利润的二维关系图。



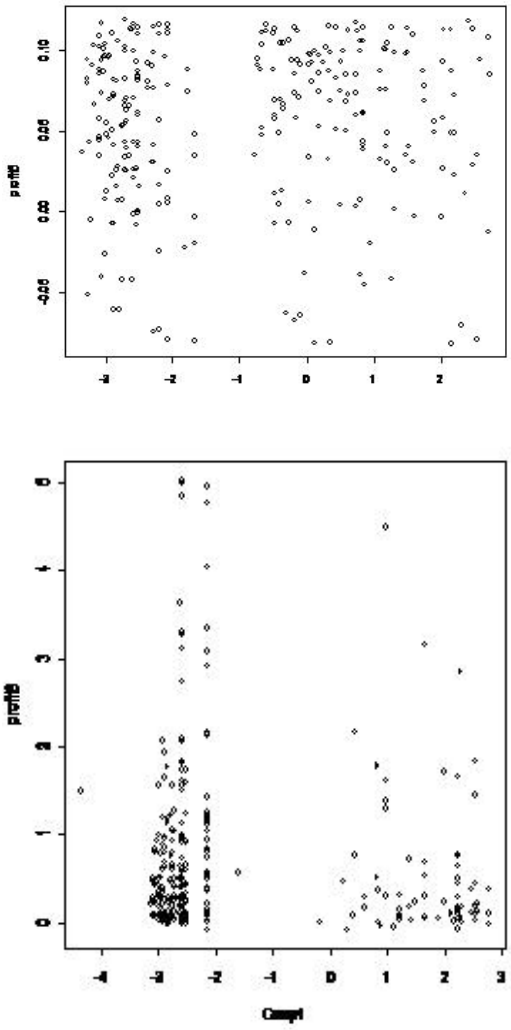
caption 图8



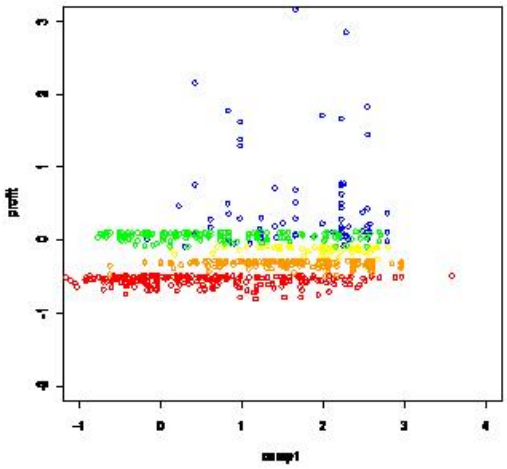
caption 图9



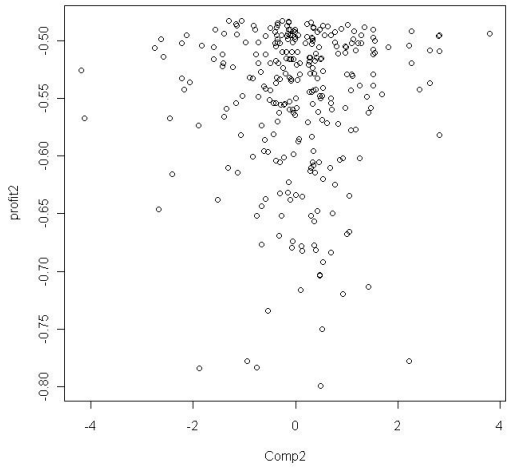
caption 图10



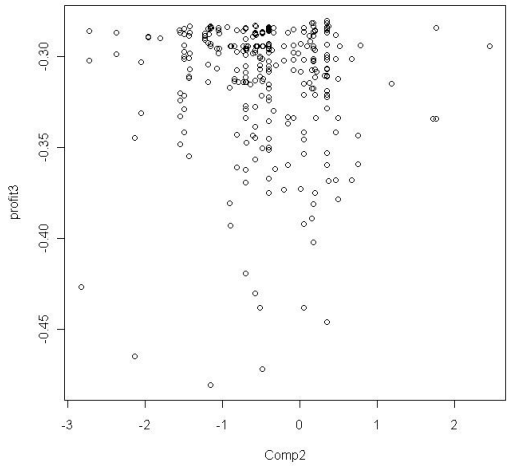
caption 图12



caption 图13

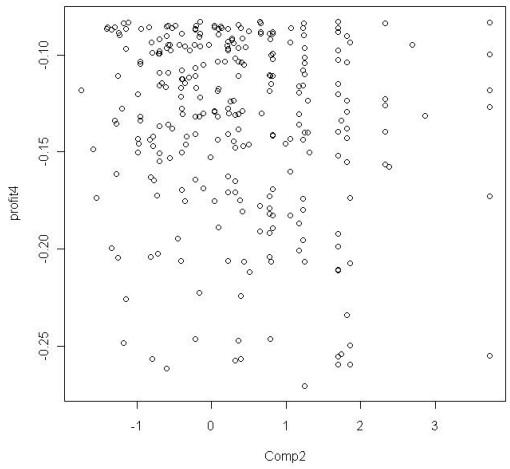


caption 图14

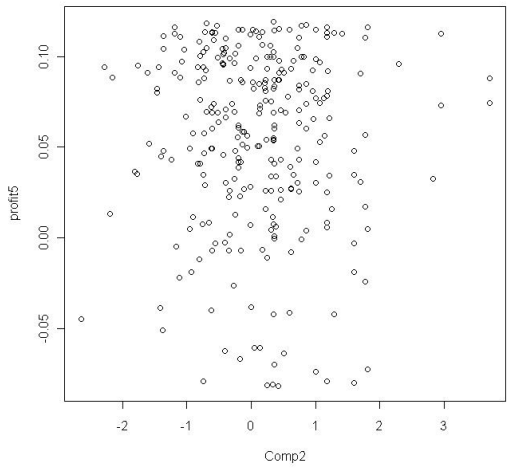


caption 图15

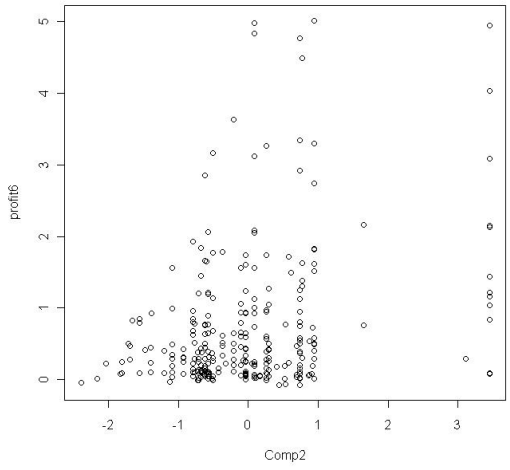
通过Comp1 VS Profit（图9-图12），我们可以看出负盈利组的主成分比较倾向于集中分布，而正盈利组有分为左右两部分的趋势。图13将各组的主成分一进行比较，可以认为第一主成分与利润没有线性关系。



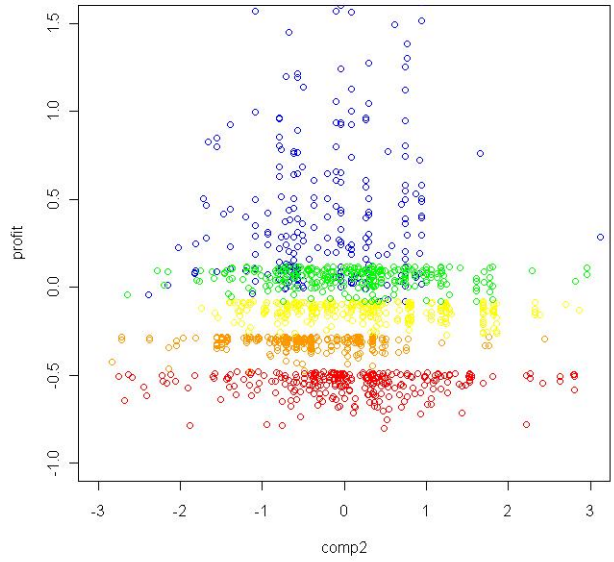
caption 图16



caption 图17



caption 图18

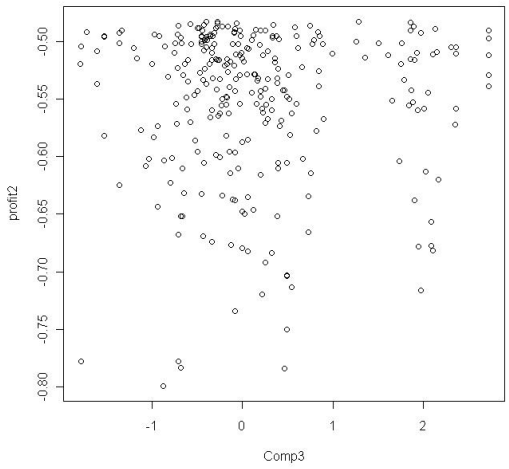


caption 图19

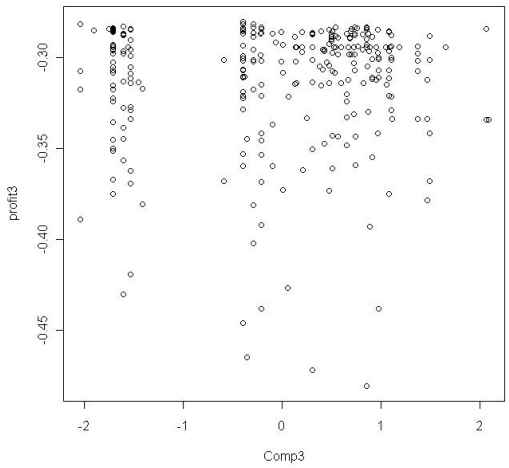
通过Comp2 VS Profit（图14-图18），我们发现在组2、3、6中点的分布比较分散，而在4、5中比较集中。说明对于一般正盈利和高正盈利两组样本，第二主成分对其影响较为有限。此外通过对图14-图19的观察我们发现第二主成分与利润之间没有明显的线性关系。

从图25可以看出，第三主成分对于利润也没有明显的函数关系，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。

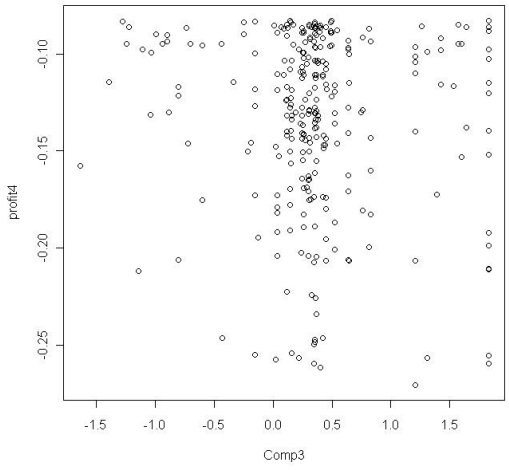
Importance of components:



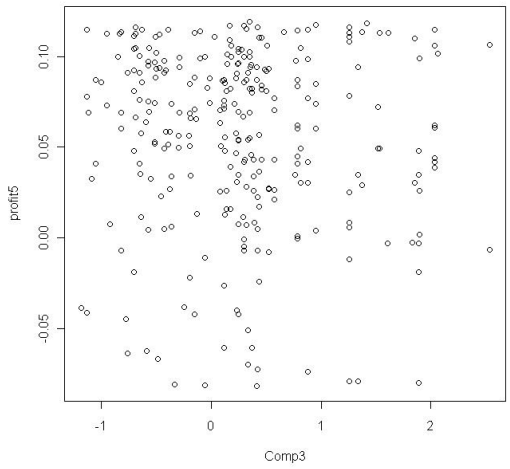
caption 图20



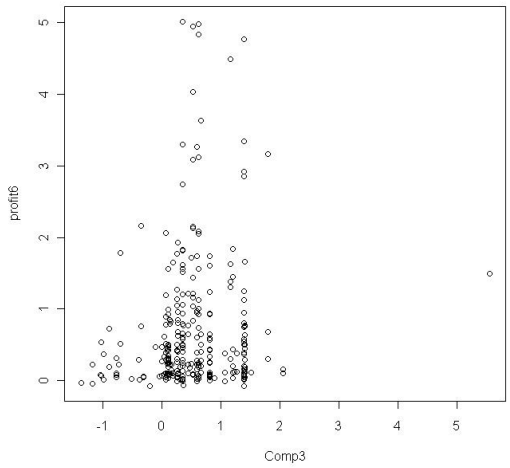
caption 图21



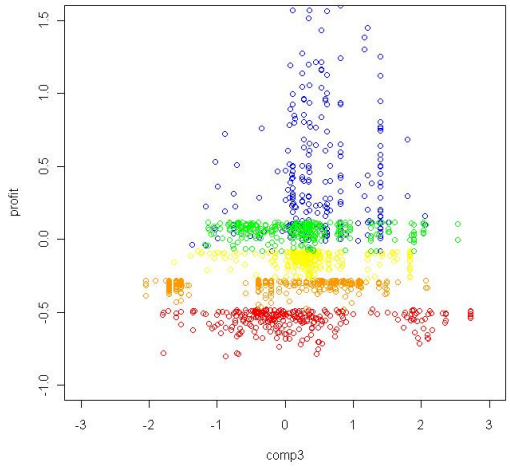
caption 图22



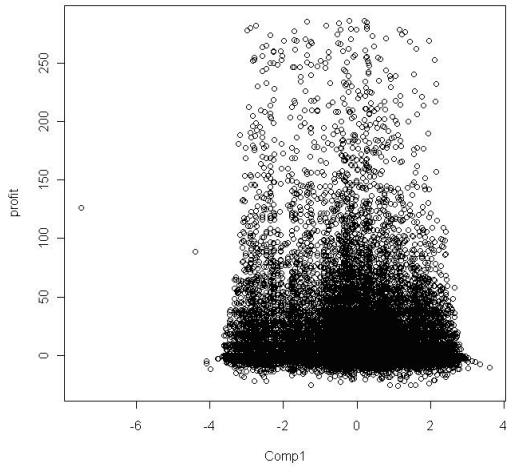
caption 图23



caption 图24



caption 图25



caption 图26

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	1.429816	1 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

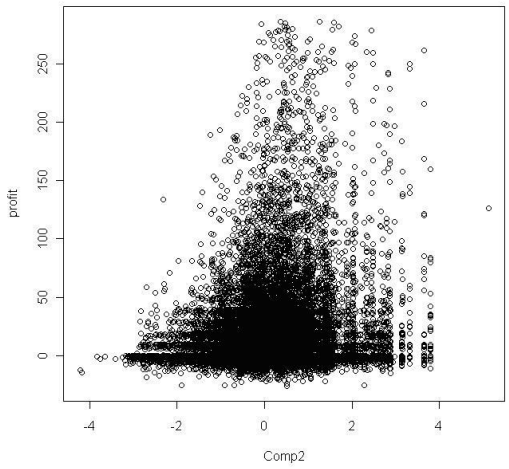
Loadings:

关键字	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

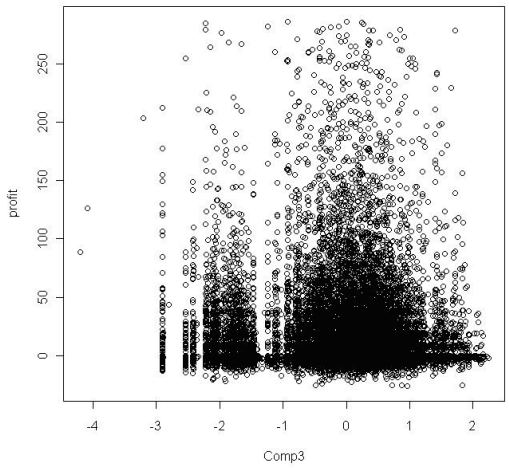
整体的主成分分析与分组后的大致相同，其与利润也没有明显的线性关系。

4.3 主成分对于利润做简单线性回归

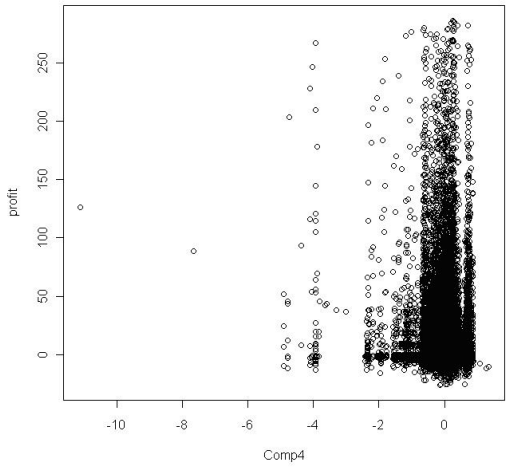
Residuals:



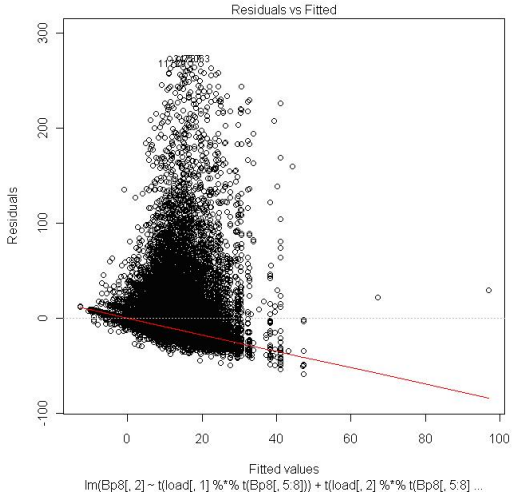
caption 图27



caption 图28



caption 图29



caption 图30

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	***

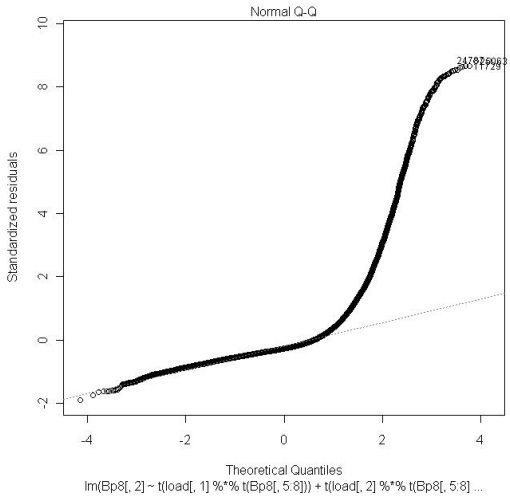
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30.93 on 30166 degrees of freedom

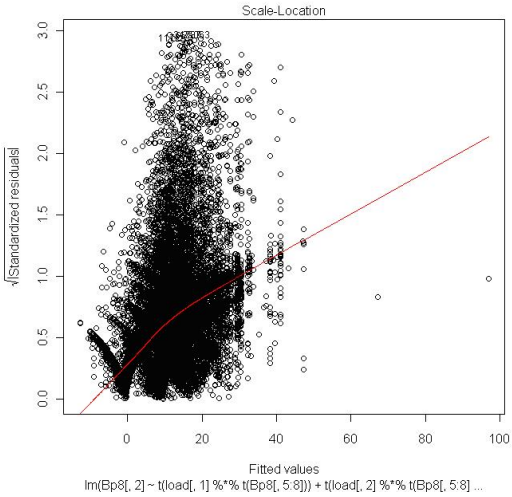
Multiple R-squared: 0.04444, Adjusted R-squared: 0.04431

F-statistic: 350.7 on 4 and 30166 DF, p-value: <2.2e-16

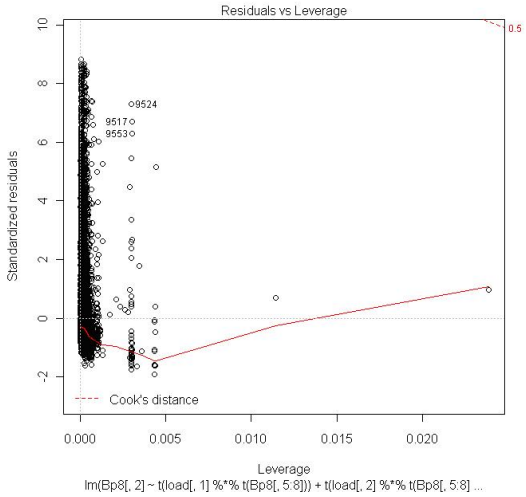
回归结果模型的P值非常小，可以认为模型中的各变量在真实模型中都很重要。每个变量的P值很小，说明这些变量都显著的影响着因变量。相对于系数的绝对值而言，对于系数估计的标准误也较小，说明系数的估计准确。四个回归系数之中只有第二主成分的系数大于0，而且绝对值明显高出其他三个主成分，前面提到第二主成分是四个变量的正线性组合，说明了出发地和目的地的平均水平和



caption 图31



caption 图32



caption 图33

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。主成分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想比较，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。总体来讲，主成分分析并没有很好的解决问题，还需通过其他途径。

第五章　交叉变量的生成和变量选择

5.1　交叉变量的生成

由于在这个问题中共有25个不同的连接词，所以如果考虑2次交叉项就有C_{25}^{2} = 300项，而如果考虑3次交叉项则项数会高达C_{25}^{3} = 项考虑到程序运算的时间代价，我们将其分为7类，按中国公民的语言习惯，同一类别任意两类不会出现在同一关键词中。

第一类：”-”,”到”,”去”,”飞”,”至”

第二类：”飞机票”,”机票”,”飞机”

第三类：”便宜”,”打折”,”折扣”,”优惠”,”特惠”,”特价”,”低价”

第四类：”价格”,”价”

第五类：”往返”,”来回”

第六类：”的”

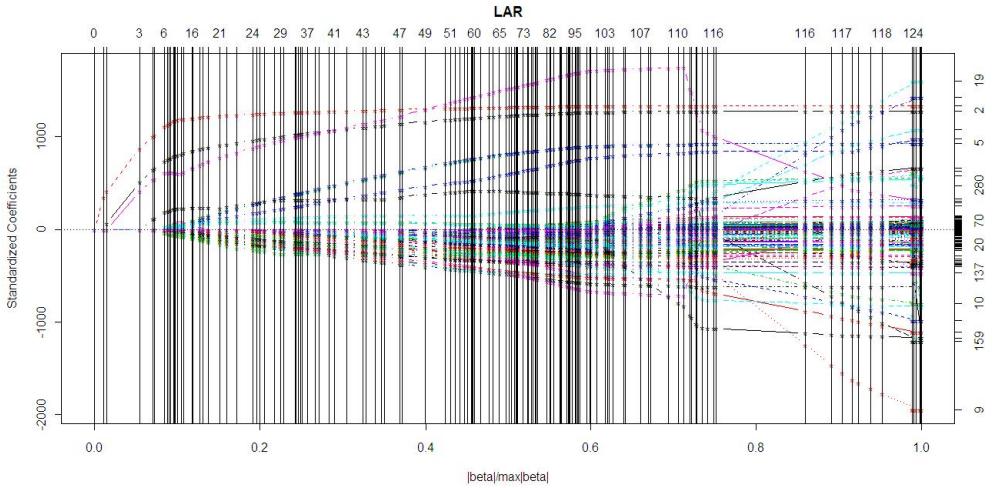
第七类：”查询”,”预定”,”预订”,”定”,”订”

则要生成2次交叉项只需在这六类中选取2类并在两类中分别选取1个连接词即可。按照此算法，只需考虑254个二次交叉项。可大幅减少程序运行所需的时间，为以后处理更多数据或是生成更高维的交叉项提供了便利。具体R语言代码请见附录.

5.2　LARS变量选择结果

利用生成的交叉项矩阵，我们可以做更进一步的LARS变量选择。

利用与前文类似的方法，可以观察到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。在由连接词形成的交叉变量中，变量显著并且对于利润有正面影响的只有两个（至，特价）和（飞，折扣），其中前者更回



caption LARS2

归系数绝对值更大。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)。如果要查询一般的连接词形成的交叉变量对利润的影响，请利用前文介绍的方法。唯一的变化是，在查询变量对应的地地理因素或是连接词、连接词组合时，Var1-Var30分别为常数项、4个地理因素和24个连接词，其排列顺序与前文一致。后254个变量：Var31-Var284,请在R语言运行前两个程序后键入：head(B8)，用变量对应的序列号减去30来查询其对应的连接词组合。

LARS变量选择的具体步骤表格请见附录。

第六章　存在问题及改进

通过之前叙述的步骤我们不难发现上述的问题处理方法还存在诸多问题，最主要的集中于地理因素。由于各地条件各不相同，我们很难整体上描述地理因素对于利润的影响，所以需要先区别不同地理位置由各地自身因素（如地理位置、GDP、人口、旅游资源等）而对利润产生的影响。这样才好消除地域差别，从而在整体上描述地理因素对利润的作用。在确立不同地域对利润的影响的过程中，我们采用了以包含该地的所有关键词对应的利润平均作为该地的影响力。注意到有些地名其实出现次数很少，有的甚至只有5次，所以这种算法在一定程度上无法准确衡量不同地域对利润的影响力。另外，虽然中国人的语言习惯不同，但可从整体上把握他们的用词规律，也就是说，可以掌握每种关键词被搜索的频率大小，在取利润平均时，不应该做算术平均，而是加权平均，对于出现频率较高的关键词赋予较高的权值，这样可以更为准确地描述地理因素的影响。

对于改进方案。这里我们要在此对总体的利益分布做一下分析：为什么有些企业的利润为负，而有些企业的利润很高。我们认为：亏损严重的企业是因为过高的点击频率和较低的购买次数。而轻微亏损的企业是因为较低的点击频率和更低的购买次数。同样，对于高利润的企业，首先需要很高的点击率和点击访客对其很高的消费偏好，也就是说，点击率高时利润高的必要条件。那么我们可以粗略地判断：是否每个关键词的点击率和利润之间存在一定的正相关关系。那么我们认为应该存在一个单调递增的函数f(x)，使得f(|profit|)可以代表该关键词的权值。

其次在地理因素影响的估计上，还存在另一个问题。由于我们采用的算法是直接对关键词对应的利润取算术平均，但并没有考虑到连接词对利润的影响。事实上，更为优秀得做法应该是将利润减去连接词的影响作用后再取加权平均，从而得到地理因素的影响。但是这里的难点在于我们队连接词对利润的影响也一无所知。

可能可行的改进方案：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取平均值方法，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收敛的估计，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。

另外一个问题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，比如点击频率，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会和我们对于关键词的评定方式存在差异。利润率不仅仅取决于关键词的选取，也在于公司内部的运营机制、合作伙伴及财务状况等一系列因素。而事实上，后面这些因素都已超出了我们关注的范围，而受公司本身因素的影响。所以就此看来，由于缺少必要信息，我们在进行回归分析的时候最好的情况下R-square也只有不到0.1，而多数情况下甚至低至0.04，其结果还远不能令人满意。

第七章　总结

根据前面的分析，我们可以为企业在选择航空机票方面的付费搜索广告的关键词上提供一些力所能及的建议。首先显而易见的是选择平均利润率高的地名。如“大连”、“武汉”、“青岛”、“厦门”、“沈阳”等作为出发地都十分理想，而“米兰”、“罗纳”、“泰国”、“海南”、“韩国”、“上海”等是目的地的上上之选。此外，“到”、“至”作为地理连接词是十分理想的，其对利润的正面影响远高于“飞”、“-”。而作为价格修饰词，“特价”、“折扣”优于“便宜”、“打折”。在名词方面，“机票”是可作为连接词的首选，它对于利润的提升具有较大的作用。值得注意的是“往返”、“来回”两个连接词对于利润有负面的影响，也就是购买这两个关键词往往会入不敷出，不建议企业购买。

第八章 附录

8.1 含交叉变量的LARS变量选择过程：

Var						3	2	7	159	102
Step						1	2	3	4	5
Var	280	16	201	34	72	47	4	8	10	5
Step	5	6	6	7	8	9	10	11	12	13
Var	151	184	11	-133	21	22	55	163	-192	
Step	14	14	15	15	16	17	18	19	19	
Var	26	-34	27	51	37	15	-198	53	111	14
Step	20	21	22	23	24	25	25	26	27	28
Var	-183	104	-282	43	-132	6	28	59	23	93
Step	28	29	29	30	30	31	32	33	34	35
Var	20	-249	168	-263	65	112	56	113	84	17
Step	36	36	37	37	38	39	40	41	42	43
Var	213	94	31	83	79	18	-158	-237	54	36
Step	43	44	45	46	47	48	48	48	49	50
Var	30	179	-274	122	105	-283	123	76	62	88
Step	51	52	52	53	54	54	55	56	57	58
Var	-255	24	-170	-277	29	9	89	-256	95	-268
Step	58	59	59	59	60	61	62	62	63	63
Var	40	-131	164	46	103	281	-9	153	98	-275
Step	64	64	65	66	67	67	68	69	70	70
Var	68	85	-252	87	-253	107	74	121	101	181
Step	71	72	72	73	73	74	75	76	77	78
Var	117	177	-260	106	108	99	-271	77	128	73
Step	79	80	80	81	82	83	83	84	85	86
Var	52	126	58	116	42	-145	75	96	-269	90
Step	87	88	89	90	91	91	92	93	93	94
Var	-257	69	100	276	125	66	81	67	48	82
Step	94	95	96	96	97	98	99	100	101	102
Var	80	60	152	61	171	-284	-181	91	136	97
Step	103	104	105	106	107	107	108	109	110	111

	Df	Rss	Cp
0	1	76044721	3246.325
1	2	75248209	2896.816
2	3	74993527	2786.422
3	4	73052844	1931.979
4	5	72550596	1712.331
5	7	72486657	1688.114
6	9	72159976	1547.947
7	10	72086834	1517.668
8	11	72030563	1494.835
9	12	71999823	1483.269
10	13	71912173	1446.589
11	14	71890872	1439.188
12	15	71878336	1435.656
13	16	71808158	1406.686
14	18	71729928	1376.162
15	18	71660338	1345.451
16	19	71464648	1261.091
17	20	71459017	1260.606
18	21	71301943	1193.288
19	21	71245857	1168.536
20	22	71137684	1122.799
21	23	70944287	1039.451
22	24	70827716	990.006
23	25	70673529	923.962
24	27	70446332	827.698
25	28	70392820	806.082
26	29	70343912	786.499
27	29	70284335	760.207
28	29	70152834	702.174
29	29	70083724	671.675
30	30	70042915	655.666
31	31	69907449	597.883
32	32	69903033	597.935
33	33	69875701	587.872

	Df	Rss	Cp
34	33	69873199	586.769
35	34	69845673	576.621
36	35	69829724	571.582
37	35	69768517	544.571
38	36	69692752	513.135
39	37	69653186	497.674
40	38	69554238	456.007
41	39	69503486	435.610
42	40	69389725	387.406
43	41	69255843	330.323
44	41	69232740	320.127
45	42	69159541	289.824
46	43	69120658	274.664
47	44	68999737	223.300
48	45	68986134	219.297
49	44	68863690	163.262
50	45	68774462	125.884
51	45	68753097	116.456
52	46	68734497	110.247
53	46	68734442	110.223
54	47	68719823	105.772
55	48	68703285	100.473
56	48	68693774	96.276
57	49	68680342	92.348
58	50	68678988	93.751
59	49	68675995	90.430
60	49	68671398	88.401
61	49	68655192	81.249
62	49	68633111	71.504
63	51	68633036	75.472
64	52	68621017	72.167
65	53	68606861	67.920
66	53	68593041	61.821

	Df	Rss	Cp
67	53	68587329	59.301
68	53	68583099	57.434
69	54	68577293	56.871
70	55	68574663	57.711
71	56	68572196	58.622
72	57	68571182	60.175
73	58	68561818	58.042
74	59	68555226	57.133
75	60	68554006	58.595
76	60	68548907	56.345
77	61	68547430	57.693
78	62	68543613	58.008
79	62	68542804	57.651
80	62	68541279	56.978
81	63	68529944	53.976
82	64	68526695	54.542

由于在第81行CP值最小，所以我们只截取其中的82行，其余数据省略

8.2 R程序代码：

这里忽略简单或是重复性的指令，只列出三段比较有价值的代码：

矩阵读入和地名提取：

```
rc<-read.csv("E:/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","至")
```

```
for(i in 1:28){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]))
```

拟变量的赋值：

```
B6<-matrix(0,nrow=30315,ncol=30) B6[,1]<-1 B6[,2]<-B[,5]
B6[,3]<-B[,6] for(i in 1:30315)B6[i,4]<-length(grep(B[i,3],B[,3]))
for(i in 1:30315)B6[i,5]<-length(grep(B[i,4],B[,4]))
C<-c("-","到","去","至","飞机票","机票","飞机","飞","便宜","打折","折
扣","优惠","特惠",
"特价","低价","价格","价","往返","来回","的","查询","预定","预订","定","订")
for(i in 1:25){ B6[c(grep(C[i],B[,1])),i+5]<-1
B[,1]<-gsub(C[i],"",B[,1]) }
B6<-B6[,c(1,2,3,4,5,6,7,8,9,13,11,12,10,14,15,
16,17,18,19,20,21,22,23,24,25,26,27,28,29,30)]
```

交叉变量的生成：

```
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]} } } } } }
```

参考文献

1.Efron，B.，Johnstone，I.，Hastie，T. and Tibshirani，R.(2004)，Least Angle Regression，*Ann. Statist.*，**32**，407-499.

2.Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning, Data Mining, Inference, and Prediction, *Stanford University*

3.[美] S.韦斯伯格（Weisberg, S.）；王静龙等译, 北京: 中国统计出版社，**1998.3**

4.高惠璇编著，应用多元统计分析，北京: 北京大学出版社，**2005.1**

致　　谢

指导老师：姚远教授

合作伙伴：祝垚

数据提供：王汉生教授

LATEX指导：李康、蒋龙龙、张子立

R语言指导：程晓行

感谢北京大学数学科学学院和父母对我的支持和帮助

谨以此文献给我伟大的父亲和亲爱的母亲!