北京大学

本科生毕业论文

题目：网上机票预订的付费搜索广告的关键字选择问题

姓　　名：许　欣[1]

学　　号：00601100

院　　系：数学科学学院

专　　业：统计学

指导教师：姚　远　教授

二〇一〇年六月

[1] 作者邮箱：1990xuxin@sina.com

摘要

付费搜索广告作为现代企业新兴的广告方式，其最重要的组成部分是竞价排名的关键字的选择问题. 当可供选择的关键字数量过于庞大时，企业很难选择使利益最大化的最恰当的关键字. 本文简单介绍了付费搜索广告的主要特点；对一组实际的关于网上机票预订的付费搜索广告数据进行分析，从作为关键字的字符串中提取变量，引入代理变量以降低巨量关键字引起的高维数；通过非参数方差分析的分类模型，主要由哑变量组成的线性回归模型以及逻辑斯谛回归模型等方法对关键字与利润之间的关系进行探索以获得最恰当的关键字.

关键词：付费搜索广告；关键字选择；代理变量；方差分析；逻辑斯谛回归

目录

§1 引言

§1.1 付费搜索广告

付费搜索也叫竞价排名(PPC,即 Pay Per Click). 随着网络技术的发展与 Google 等搜索引擎的市场化全球化，付费搜索广告逐渐成为企业普遍的广告手段，美国付费搜索广告 2007 年收益超过 10 亿美元，据预测在 2012 年或将达到 154 亿美元[1].

付费搜索不同于其他传统广告形式,企业不再向广告公司支付固定的广告费用. 而是由企业对他们感兴趣的关键字竞价投标,当消费者使用搜索引擎搜索这些关键字时,将在搜索结果中出现该企业的推广信息. 关于搜索引擎如何实现付费搜索广告以及决定企业推广信息出现的位置，文献[2]进行了理论与实际应用上的详细讨论.

付费搜索广告具有其他广告形式无法比拟的优点. 其一,关键字的存在决定了付费搜索广告具有指向性，不同于一般广告形式面对所有消费者，而是通过搜索引擎的“筛选”直接指向对自己的商品和服务感兴趣的消费者，其效率会得到很大提高. 其二，付费搜索广告的基本特点是按点击量多少支付广告费用，即使点击量很少造成广告效果差，对企业也不会造成较大损失. 其三，消费者对付费搜索广告的反应可通过其进入企业推广信息页面后的行为来观察,企业将得到很多有用的反馈信息，这些信息可以用于企业的广告效果评估，广告策略调整甚至市场战略等更一般的决策.关于搜索引擎与付费搜索广告特点的详细讨论，可以参见文献[3][4].

§1.2 关键字选取

要做好付费搜索广告，最重要的地方就是选择恰当的关键字参与竞价投标. 好的关键字能增加消费者点击量，提高消费者对企业推广信息的兴趣，从而提高企业利润，达到付费搜索广告的最佳效果. 但是当可供选择的关键字数量过多时,选取最恰当的关键字这一工作变得十分困难. 参与竞价投标前企业可以通过

[1] 转引自文献[1]

调查目标人群，了解排名算法与竞价后台等手段选择适合自己的关键字．获得反馈信息后，应该对数据进行分析，不断调整关键字选择策略．

下面以本文中网上机票预订的关键字数据[1]为例来说明关键字数量庞大的问题．考虑关键字“北京到桂林往返特价机票”，这一关键字主要由以下部分组成：出发地部分“北京”，连接词部分“到”，目的地部分“桂林”，是否往返部分“往返”和价格相关部分“特价”．其他关键字条目基本由这几部分组成．在这组数据中，出发地(包括城市，省份，地区和国家)有 142 个，目的地有 258 个；连接词有 6 种，分别为“-”，“到”，“飞”，“去”，“至”和无连接词[2]；关于价格的关键字有 10 种，分别为“打折”，“低价”，“价格”，“价钱”，“便宜”，“票价”，“特价”，“优惠”，“折扣”和无关键字．因此可能的关键字组合数为 $142 \times 258 \times 6 \times 2 \times 10 \approx 4,000,000$ 个．在如此多关键字中选取恰当者十分困难，高维数对于数据分析是不利的，注意到关键字的数量之大主要是由出发地与目的地造成的，因此考虑使用与出发地及目的地相关的代理变量取代它们，达到降低维数的目的．

本文将从作为字符串的关键字中提取代理变量与哑变量，通过多种统计方法对这组数据进行分析与探索，寻求关键字与企业利润之间的关系，找到最恰当的关键字．本文的第二部分是数据中变量的提取与简单处理，第三部分将数据作为一个基于关键字的分类问题进行方差分析，第四部分对数据建立线性回归模型并讨论相关的问题，第五部分建立逻辑斯谛(Logistic)回归模型，第六部分将对全文进行总结归纳．

[1] 数据来自文献[1]
[2] 有 12 例中连接词为“往”，因数量过少本文除第三部分外不予考虑

§2 数据与变量

数据共 N=30315 项，两列——关键字和利润，关键字以字符串的形式表现，利润为实数，数字越大表示利润越高，没有缺失数据.

§2.1 因变量的调整

因变量 Y 是与企业利润正相关的数字，其具体意义并不明确，表 2-1 是对 Y 的简要分析结果.

表 2-1 因变量 Y 的简单统计量

统计量	最小值	最大值	中位数	均值	标准差
值	-79.20	3078.00	-0.58	13.75	50.09

进一步的分析表明，Y 的绝大部分项都集中在 0 附近，对 Y 的正态性检验其 p 值几乎为零，因此有必要对 Y 进行处理. 引入反映关键字是否盈利的变量 Y0，即当 Y>0 时 Y0 为 1，否则为 0，Y0 将在第五部分逻辑斯谛回归模型中使用.另一方面，将 Y 做简单的截断处理，令

$$Y1=\begin{cases}100, \text{当} Y>100\\ Y, \text{当} -10\le Y\le 100\\ -10, \text{当} Y<-10\end{cases}, \tag{2-1}$$

以下主要以 Y1 为因变量进行分析.

§2.2 哑变量与代理变量的提取

哑变量 X1 到 X5 分别对应连接词“-”，“到”，“飞”，“去”，“至”，均为零则表示无连接词；X6 对应关键字“往返”；X7 到 X15 分别对应与价格相关的关键字“打折”，“低价”，“价格”，“价钱”，“便宜”，“票价”，“特价”，“优惠”，“折扣”，均为零则表示无关键字. 表 2-2 给出了各类数据的总量，可以看出数据在各类别中的分配很不均匀.

表 2-2 数据的分类情况

类别	数据量	百分比	类别	数据量	百分比	类别	数据量	百分比
-	5935	19.58%	往返	194	0.64%	打折	5554	18.32%
到	10615	35.02%	无	30121	99.36%	低价	64	0.21%
飞	3559	11.74%				价格	657	2.17%
去	1525	5.03%				价钱	20	0.07%
至	4272	14.09%				便宜	188	0.62%
无	4409	14.54%				票价	476	1.57%
						特价	7451	24.58%
						优惠	55	0.18%
						折扣	1579	5.21%
						无	14271	47.08%

对于出发地与目的地，文献[1]引入的代理变量是城市人口与 GDP，本文考虑对出发地与目的地各引入两个代理变量：出发地在所有数据中作为出发地出现的次数 X16，出发地作为出发地出现的数据的利润均值 X17，目的地在所有数据中作为目的地出现的次数 X18，目的地作为目的地出现的数据的利润均值 X19. 表 2-3 给出代理变量 X16 到 X19 各自的简单统计量.

表 2-3 四个代理变量的简单统计量

变量	最小值	最大值	中位数	均值	标准差
X16	1	3601	848	1208.531	1090.610
X17	-7.96	27.84	10.06	10.54535	3.041576
X18	1	1777	660	657.3912	471.1136
X19	-10	100	11.38	10.54542	4.342167

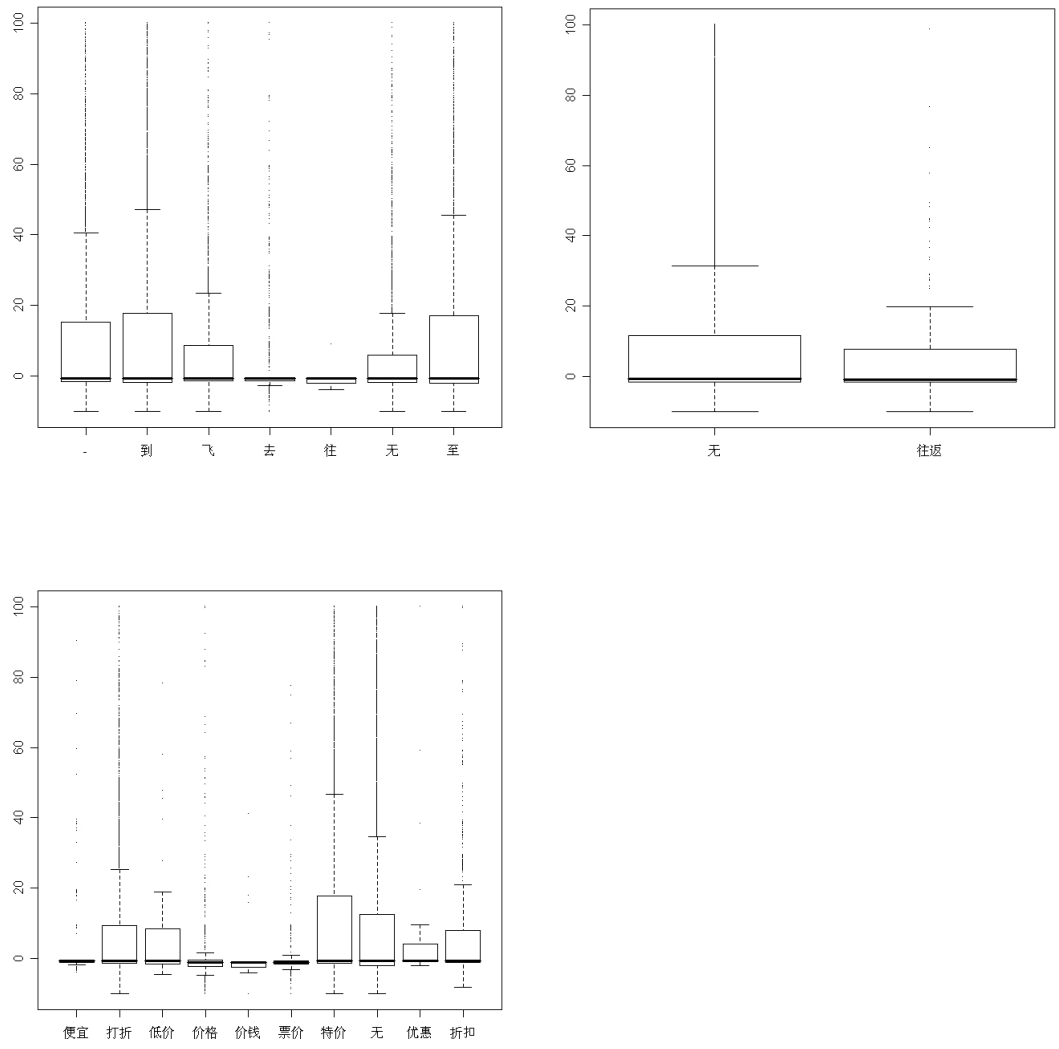
之后对这四个变量进行极差正规化变换，使其处于零一之间，与哑变量量级相同.

此外再引入反映关键字长度的变量 L.

§3 方差分析

相对于出发地与目的地，航空公司将更为看重连接词等关键字对利润的影响. 这是因为对特定的出发地与目的地，航空公司如果有售两地之间的机票，只能考虑是否投入付费搜索广告而无法改变作为出发地和目的地的关键字；但对于其他关键字，如两地之间的连接词，相互之间完全可以替代，因此航空公司可以自由充分地在几种不同类别的关键字中选择恰当者，对利润的提高会比出发地与目的地关键字的选取效果更好. 这一部分将只考虑哑变量 X1~X5, X6, X7~X15，将数据作为分类问题进行单因素方差分析，通过考察各类别之间因变量 Y1 均值的差异来推断恰当的关键字. 首先做出 Y1 关于各类别的盒型图如下.

图 3-1 Y1 关于各分类变量的盒型图



从图 3-1 来看，三种分类方法的各类别之间 Y1 的水平可能存在差异，因此有必要进行进一步的方差分析. 本文第二部分指出，变量 Y 不服从正态分布，经过简单处理后得到的因变量 Y1 仍不满足正态分布假定(事实上 Y 的分布情况十分坏以至于几乎无法通过有意义的变换或处理使其满足正态分布假定)，因此本文使用 Kruskal-Wallis 检验的非参数方差分析方法. Kruskal-Wallis 检验不要求数据来自正态分布总体，也不要求各类的方差相等，虽然在同等条件下一般比参数检验的功效低，但其适用性更加广泛. 文献[5]对 Kruskal-Wallis 检验的性质和特点进行了详细的讨论.

表 3-1 给出了三种分类方式下 Kruskal-Wallis 检验的结果，包括各关键字所在类别 Y1 的均值，近似的卡方统计量以及检验的 p 值(Prob > Chi-Square)，三次检验的自由度分别为 6,1 和 9. 其中关于关键字中是否包含“往返”字样的分类变量 X6 的 Kruskal-Wallis 检验因为只分两类，与独立两组的 Wilcoxon 秩和检验结果是相同的.

表 3-1 Kruskal-Wallis 检验结果

类别	均值	类别	均值	类别	均值
-	11.42555	往返	5.696598	打折	9.161858
到	14.08686	无	10.57696	低价	5.695156
飞	6.217398			价格	2.55898
去	3.859436			价钱	3.0235
往	-0.4991667			便宜	5.130798
至	12.29069			票价	0.9258824
无	4.966525			特价	13.86106
				优惠	5.064182
				折扣	5.294243
				无	10.74769
Chi-Square	212.6091	Chi-Square	4.6726	Chi-Square	384.1336
P-value	<2.2e-16	P-value	0.03065	P-value	<2.2e-16

表 3-1 第一部分是关于连接词分类的结果，检验的 p 值几乎为 0，可见含有不同连接词的关键字之间，其利润水平有极为显著的差异；从均值上看，“到”，“至”和“-”这三个连接词比无连接词的情况利润水平高，“飞”可能是恰当的，而“去”和“往”则低于无连接词的情况下的利润水平.

第二部分是关于是否含有“往返”字样分类的结果，检验的 p 值为 0.03065，考虑到 Kruskal-Wallis 检验作为非参数检验的低功效性，可以认为是否含有“往返”字样其利润水平有显著差异；从均值上看，含有“往返”字样的关键字其利润水平比不含“往返”字样的关键字低.

第三部分是关于价格相关关键字分类的结果，检验的 p 值同样几乎为 0，可见含有价格相关信息的关键字，其利润水平有极为显著的差异；从均值上看，只有含有“特价”字样的关键字利润水平高于不含价格相关信息的关键字，含有其他价格相关信息的关键字利润水平都会更低.

§4 线性回归模型

§4.1 简单线性回归

表 4-1 简单线性回归的参数估计及其他结果

Variable	Estimate	Std. Error	T value	P value
Intercept	-24.69249	1.64402	-15.020	< 2e-16
X1(-)	10.22042	0.50484	20.245	< 2e-16
X2(到)	13.89681	0.46634	29.799	< 2e-16
X3(飞)	2.35647	0.53564	4.399	1.09e-05
X4(去)	-1.60729	0.69187	-2.323	0.02018
X5(至)	8.05392	0.53073	15.175	< 2e-16
X6(往返)	-3.48460	1.65903	-2.100	0.03570
X7(打折)	0.04997	0.44979	0.111	0.91154
X8(低价)	-11.32345	2.86821	-3.948	7.90e-05
X9(价格)	-5.71599	0.94861	-6.026	1.70e-09
X10(价钱)	-18.90251	5.08233	-3.719	0.00020
X11(便宜)	-7.68106	1.68809	-4.550	5.38e-06
X12(票价)	-8.69145	1.08995	-7.974	1.59e-15
X13(特价)	6.16713	0.42625	14.468	< 2e-16
X14(优惠)	-9.27855	3.07574	-3.017	0.00256
X15(折扣)	-6.75214	0.65173	-10.360	< 2e-16
X16(出发地次数)	8.40185	0.49062	17.125	< 2e-16
X17(出发地均值)	45.39023	1.63324	27.791	< 2e-16
X18(目的地次数)	8.28975	0.76311	10.863	< 2e-16
X19(目的地均值)	116.26899	4.94809	23.498	< 2e-16
L(关键字长度)	-2.87018	0.16220	-17.695	< 2e-16
Multiple R-Squared	0.1236			
Adjusted R-squared	0.1231			
F-statistic	213.7			
p-value	< 2.2e-16			

表 4-1 给出了 Y1 对哑变量 X1 到 X15，代理变量 X16 到 X19，以及 L 等所有变量简单线性回归的结果. 除了代表“打折”字样的变量 X7 外，所有估计系数 t 检验的 p 值均很小. 具体来说有以下结果.

——连接词中，“到”，“-”与“至”对利润有显著的正面影响，“飞”的存在也优于无连接词的情况，而连接词“去”将使利润下降.

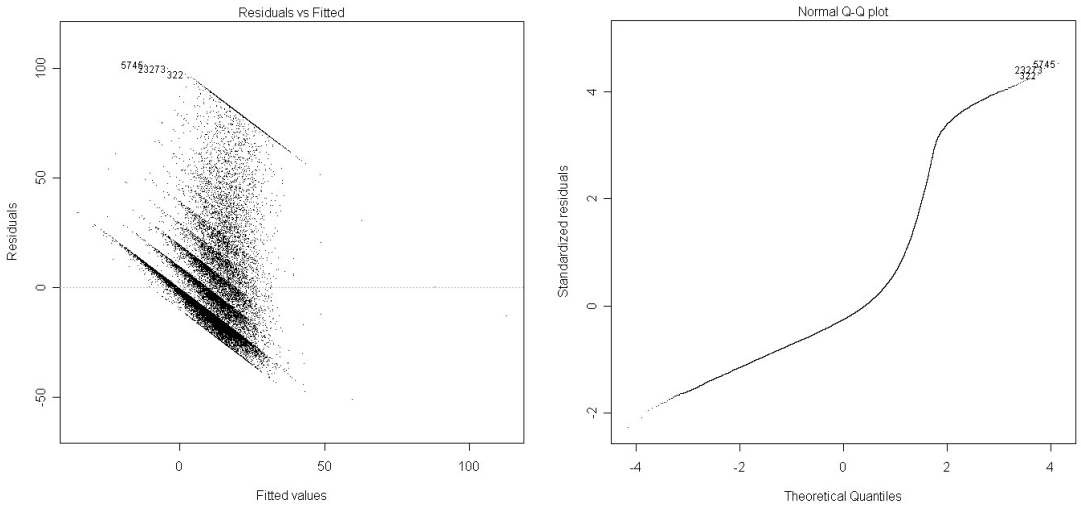
——“往返”对利润产生负影响. 但因为含有“往返”字样的关键字数量很少，这种影响可能不是显著的.

——与价格相关的关键字中，“特价”对利润有显著的正面影响，“打折”对利润水平无影响，其他字样均产生负影响.

——出发地与目的地的出现次数(可视作其热门程度)对利润有显著正面影响. 此外，与预期相同，出发地与目的地均值与利润表现出强烈的正相关.

——关键字越长，其利润水平越低.

图 4-1 残差诊断图



观察图 4-1 发现残差与拟合值有很强的负线性关系，并且表现为很多平行的“层”，这是因为参与模型的自变量中大部分是只取零一的哑变量，造成拟合值只能在离散的不同水平的层次之间跳动. 此外，残差的正态 QQ 图表现出比较严重的轻尾症状. 这些结果表明简单的线性回归模型可能不够恰当，需要尝试进一步的调整和改进.

§4.2 其他模型

考虑从全模型中去除几个变量. 表 2-2 表明部分类别数据量很少，对应的哑变量对整体数据的影响相应的会很小，现将占数据比例小于 1%的哑变量 X6, X8, X10, X11, X14 (分别对应关键字“往返”，“低价”，“价钱”，“便宜”，“优惠”) 从模型中去除. 此外出发地与目的地利润平均值，即变量 X17 与 X19 与 Y1 相关性很强，可能会对其他变量在模型中的作用造成很大影响，因此也从模型中去除. 下面考虑 Y1 对剩余 13 个自变量的回归.

表 4-2 新模型的参数估计及其他结果

Variable	Estimate	Std. Error	T value	P value
Intercept	18.3236	1.1905	15.392	< 2e-16
X1(-)	10.3084	0.5131	20.089	< 2e-16
X2(到)	13.6798	0.4755	28.771	< 2e-16
X3(飞)	2.4203	0.5465	4.429	9.51e-06
X4(去)	-0.8861	0.7060	-1.255	0.209436
X5(至)	9.3418	0.5397	17.310	< 2e-16
X7(打折)	1.5328	0.4489	3.415	0.000639
X9(价格)	-4.0561	0.9649	-4.204	2.63e-05
X12(票价)	-7.7109	1.1117	-6.936	4.12e-12
X13(特价)	7.4897	0.4251	17.620	< 2e-16
X15(折扣)	-5.0217	0.6587	-7.624	2.53e-14
X16(出发地次数)	10.4781	0.4801	21.826	< 2e-16
X18(目的地次数)	18.4065	0.5422	33.950	< 2e-16
L(关键字长度)	-3.2520	0.1621	-20.064	< 2e-16
Multiple R-Squared	0.08362			
Adjusted R-squared	0.08323			
F-statistic	212.7			
p-value	< 2.2e-16			

表 4-2 给出的各变量回归系数与完全模型相差不大，没有得到有意义的新结论. 变量 X4 的作用不再显著，相对的，变量 X7 的作用变得显著起来. 此外截距由负值变为正值. 由于去除了变量 X17 与 X19，大部分变量的回归系数相对原模型都增大了，这一事实证明了变量 X17 与 X19 对其他变量的影响.

关于关键字与利润数据的其他线性模型(LARS 方法等)可以参见文献[1][6].

§5 逻辑斯谛回归模型

§5.1 逻辑斯谛回归[1]

当因变量不是连续型变量而是只取几个离散值的分类变量，甚至是只取零一的分类变量(例如实验成功与否，产品达标与否等)时，正态线性模型是不合适的，因为零一响应无法与正态误差对应起来. 在这种情况下可以采用逻辑斯谛(Logistic)回归模型.

逻辑斯谛模型关心作为因变量的分类变量 Y 取 1 的概率 p 对自变量 X 的依赖关系，通过对 p 进行 logit 变换，即定义

$$\mathrm{logit}(p)=\ln\frac{p}{1-p}, \tag{5-1}$$

变换(5-1)对 p 严格单调上升，并且将零一之间的变量 p 转化为在整个实数范围内变化的变量. 从而可以用 logit 尺度拟合一个线性模型如下

$$\mathrm{logit}(p)=b_0+\sum_{i=1}^{k}b_i x_i\ . \tag{5-2}$$

结合(5-1)(5-2)两式得到模型

$$p=\frac{e^{\,b_0+\sum_{i=1}^{k}b_i x_i}}{1+e^{\,b_0+\sum_{i=1}^{k}b_i x_i}} \tag{5-3}$$

(5-3)式表示在原始概率尺度下将得到 S 型的拟合曲线.

逻辑斯谛回归模型作为广义线性模型的一个特例，具有以下几条性质：首先，因变量是由在已知次数的试验中的成功次数构成的相互独立的二项计数组成的；其次，只以线性的形式依赖于自变量；第三，logit 变换将自变量的线性形式与二项计数的期望值联系起来. 一般的参数估计算法可以参见文献[7].

[1] 本部分内容主要来自文献[7]

§5.2 计算结果

相对利润的具体数额，航空公司可能更关心关键字是否能带来利润. 因此下面将利用本文第二部分中引入的分类变量 Y0 对各自变量进行逻辑斯谛回归，分析何种关键字更可能为航空公司带来利润.

表 5-1 逻辑斯谛回归的参数估计

Variable	Estimate	Std. Error	z value	P value
Intercept	-3.47123	0.16686	-20.803	< 2e-16
X1(-)	0.95622	0.05029	19.014	< 2e-16
X2(到)	1.16415	0.04735	24.587	< 2e-16
X3(飞)	0.27493	0.05368	5.122	3.02e-07
X4(去)	-0.24363	0.07243	-3.364	0.000769
X5(至)	0.75191	0.05231	14.374	< 2e-16
X6(往返)	-0.30219	0.17202	-1.757	0.078963
X7(打折)	0.13292	0.04394	3.025	0.002484
X8(低价)	-1.09517	0.29502	-3.712	0.000205
X9(价格)	-0.84370	0.10898	-7.742	9.79e-15
X10(价钱)	-1.89062	0.56931	-3.321	0.000897
X11(便宜)	-0.92220	0.18230	-5.059	4.22e-07
X12(票价)	-1.61671	0.14967	-10.802	< 2e-16
X13(特价)	0.62339	0.04171	14.947	< 2e-16
X14(优惠)	-0.91077	0.32066	-2.840	0.004507
X15(折扣)	-0.53208	0.06582	-8.084	6.28e-16
X16(出发地次数)	0.83738	0.04770	17.557	< 2e-16
X17(出发地均值)	3.94273	0.16514	23.875	< 2e-16
X18(目的地次数)	0.68008	0.07602	8.946	< 2e-16
X19(目的地均值)	10.58558	0.52003	20.356	< 2e-16
L(关键字长度)	-0.28279	0.01658	-17.060	< 2e-16

对照表 5-1 与表 4-1，可以发现逻辑斯谛回归的各自变量系数的估计和标准差比线性回归对应的值低一个量级，而其相对大小没有多大变化. 因此将得到与第四部分中类似的结论.

——连接词中，“到”，“-”与“至”的获利概率最高，“飞”的获利概率较小，而连接词“去”的获利概率将低于无连接词的情况.

——“往返”对获利概率有负影响，但这种影响不显著.

——与价格相关的关键字中，“特价”的获利概率最高，“打折”与无价格信息时水平相同，其他字样均产生负影响.

——出发地与目的地的出现次数对获利概率有显著正面影响，其中出发地的影响更大.

——关键字越长，获利概率越小.

§6 结论

通过对数据进行方差分析，线性回归与逻辑斯谛回归等探索，基本确定了航空公司关于网上机票预订的付费搜索广告的最优关键字.

出发地与目的地的热门程度对关键字的利润有很大影响，航空公司应当对重点城市投入更多广告资金. 连接词的存在同样影响到关键字的利润，以“到”为连接词的关键字为航空公司带来最多利润，其次为“至”和“-”，连接词“飞”也是有利的，连接词“去”则会带来相反的效果,这可能与消费者对简单表达方式的偏好有关. 带有“往返”字样的关键字会使利润下降. 与价格相关的所有关键字中，只有“特价”这一信息是有利的，相对的，“折扣”、“便宜”等信息将使利润下降. 此外，关键字越长越不利，这与大多数消费者的搜索习惯是一致的.

本文关于这组数据的处理尚有可改进之处. 第三部分的方差分析并未进行多因素方差分析，可能忽视了分类变量之间的交互作用. 第四部分线性模型的性质不够好，未尝试变量选择等其他方法. 关于分类变量化的利润 Y0，还可以对其做 Probit 回归等广义线性回归.

致谢

在本文的写作过程中，指导老师姚远老师提供了大量的帮助与建议，在此我表示衷心感谢. 北京大学光华管理学院的王汉生老师提供了本文所使用的数据，参与同一课题的祝垚同学提供了宝贵的建议，此外我的父母在论文写作期间给我极大关心，在此对他们表示感谢.

参考文献

1. Minghua Jiang, Xuefeng Li, Chi-Ling Tsai and Hansheng Wang. (2010), Profitable Keyword Selection: Paid Search Advertising for Online Airplane Ticket Booking, manuscript.
2. Rutz, O. J. and Bucklin, R. E. (2007), A model for individual keyword performance in paid search advertising, *Working Paper*.
3. Chen, Y. and He, C. (2006), Paid placement: advertising and search on the Internet, *Working Paper*, University of Colorado, Boulder, CO.
4. Goldfarb, A. and Tucker, C. (2007), Search engine advertising: pricing ads to context, *Working Paper*.
5. Mahoney, M. and Magel, R. (1996), Estimation of the Power of the Kruskal-Wallis Test, *Biometrical Journal* Vol.38, 613–630.
6. Yao Zhu. (2010), Keyword Selection in Paid Search Advertising: Profitable Combination in Online Airplane Ticket Booking, manuscript.
7. Weisberg, S. (1998), Applied Linear Regression. Trans. Jinglong Wang, Xiaoyun Liang and Baohui Li. 北京，中国统计出版社.

附录：部分 R 程序

```
> attach(se)
##绘制盒型图
> boxplot(Y1~G1,pch=".")
> boxplot(Y1~X6,pch=".",names=c("无","往返"))
> boxplot(Y1~G2,pch=".")
##Kruskal-Wallis检验
> kruskal.test(Y1~G1)
> kruskal.test(Y1~X6)
> kruskal.test(Y1~G2)

##线性回归模型
> mod1<-lm(Y1~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10+X11+X12+X13+X14
   +X15+X16+X17+X18+X19+L)
> summary(mod1)
##残差诊断
> plot(mod1,pch=".")
##改进
> mod2<-lm(Y1 ~ X1+X2+X3+X4+X5+X7+X9+X12+X13+X15+X16+X18+L)
> summary(mod2)

##逻辑斯谛回归
> mod3<-glm(Y0~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10+X11+X12+X13
   +X14+X15+X16+X17+X18+X19+L,family=binomial)
> summary(mod3)
```