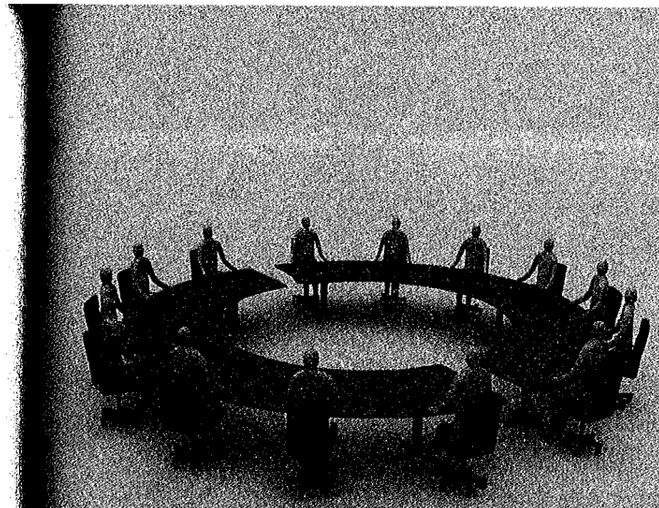
按照 MCR 做预测：

```
> a$Y.hat=predict(probit1,a)
> table(a$Y,a$Y.hat)

     1   2   3   4   5
1  28   0  72  21   0
2  49   0 145  69   0
3  34   0 303 164   0
4   7   0 188 234   0
5   0   0  24 113   0
>
```

第 5 章

泊松回归

——以付费搜索广告为例

5.1 背景介绍

企业生产一种产品或者服务，除了关注该产品或者服务的生产过程以外，同样重要的是关注相关信息在目标客户中的传递过程。一个企业的产品服务再好，如果客户不知道，那么一切都等于零。因此，现代企业营销实践格外看重各种媒体信息的传播作用。企业希望通过最有效的方式将相关信息准确传递给目标客户。但是怎样才能达此目的呢?

最传统的信息传播方式莫过于口碑（word of mouth），即通过现有客户的口口相传将企业的产品服务信息传递给新客户。过去的研究表明，口碑对产品服务信息的传递起着重要的作用。例如，新产品上市，口碑的作用将直接影响销量，进而影响产品的生命周期。直至今日，相关营销研究仍然认为口碑传播对企业产品服务的成功极其重要。通过口碑传播所获得的客户，同通过其他手段（如促销）获得的客户相比，具有更高的忠诚度，能够为企业带来更多的利润。

但口碑传播也有它的劣势，那就是不容易控制。企业很难通过资源的投入过多地改变口碑传播的过程。因此，除了口碑传播以外，企业还需要其他手段帮助传播产品服务信息，最常用的手段之一就是广告，通常的广告媒介包括电视、电台、报纸、杂志、灯箱、路牌、门户网站等。总而言之，对于一个高度商业化的社会而言，信息的有效传播能够带来可观的商业价值。因此，但凡人们有可能留心注意采集信息的地方，都可以通过合理巧妙的形式展示广告。过去人们忽视的很多死角（如电梯、卫生间、出租车前排坐椅椅背）如今都被充分地利用了起来，而且常常有意想不到的效果。很多企业，尤其是新兴企业，在其诞生之初同行业领先者相比就有着先天的劣势。因此，广告成为此类企业的生命线。广告的有效投放直接决定了企业的收入、利润，甚至存活。但是，常见的广告方式效果如何呢？有一句行业俗语："广告费的80%都被浪费了，而且不知道浪费在哪。"这从一个侧面反映出一个现象，即对某些产品传统媒体的广告效果不尽如人意，甚至广告效果在逐年递减。其中一个重要的原因就是这些广告都是被动形式的，也就是说，此类广告（如电视、电台、报纸）的展现无法做到根据客户的不同而不同。因此，有可能给想买衣服的消费者看洋酒的广告，而给想喝酒的顾客展示电

器广告。所以，此类广告的绝大多数投向了无效客户，这就难怪它们的效果不尽如人意。

那么有没有什么广告是允许消费者主动表达购买意愿的呢？有。最常见的可能就是付费搜索广告（paid search advertising），它是当下备受关注的搜索引擎营销（search engine marketing）的核心之一。有研究机构预测，美国付费搜索广告的营业额将从2007年的大约80亿美元成长为2012年的154亿美元（*US Online Advertising Forecast, 2007 to 2012*，jupiterresearch.com，2007）。届时，付费搜索广告占整个互联网广告的份额将远远超过50%。那么，到底什么是付费搜索广告？举一个简单的例子。例如春节期间某人需要购买一张回老家重庆的机票，需要找一个好的机票代理，但是又不知道到底哪家好，怎么办呢？他会打开一个常用的搜索引擎（如谷歌），然后在它的提示下输入汉字"重庆　机票"（请注意"重庆"和"机票"之间的空格），如图5—1所示。

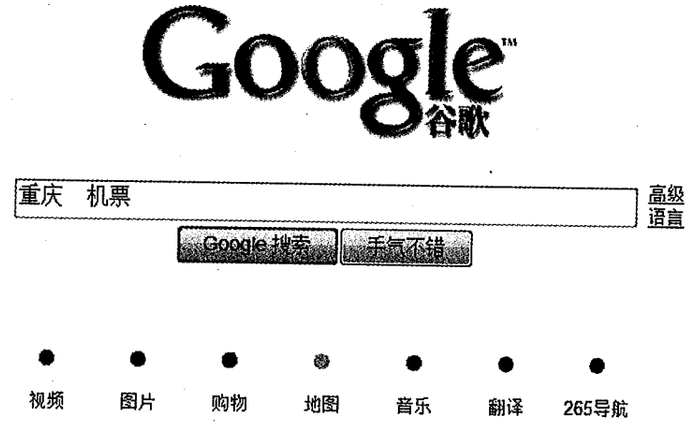


图5—1　谷歌页面

谷歌会展示它的搜索查询结果如图5—2所示。可以看到两种搜索结果。一种叫做自然搜索结果（organic search result）。这部分结果的排列顺序是按照搜索引擎的复杂算法，根据展示页面同搜索关键词的相关度得出的。如果一个企业的主页能够在自然搜索结果中被排在非常前面，这是一件很值得开心的事情。因为用户可以很容易看到企业的主页链接，并有可能因此点击浏览企业的主页。因为，用户已经通过关键词"重庆　机票"表达了购买机

票的意愿，因此绝对是企业的目标客户，很有可能在企业的网站上购买机票。更重要的是，作为自然搜索结果，谷歌还不会向企业收取任何费用！你看，这是多好的免费广告！但是，自然搜索也有它的缺点。因为它太好了，所以竞争对手（其他的机票代理网站）都会拼命琢磨搜索引擎的潜在算法，通过各种手段（如页面优化、增加点击量等）努力把它们的主页挤进比较靠前的位置。因此，对于任何一个有商业价值的关键词（例如："重庆　机票"），没有任何企业能够保证自己的主页永远被自然搜索结果排在前面，这是自然搜索的最大缺点。

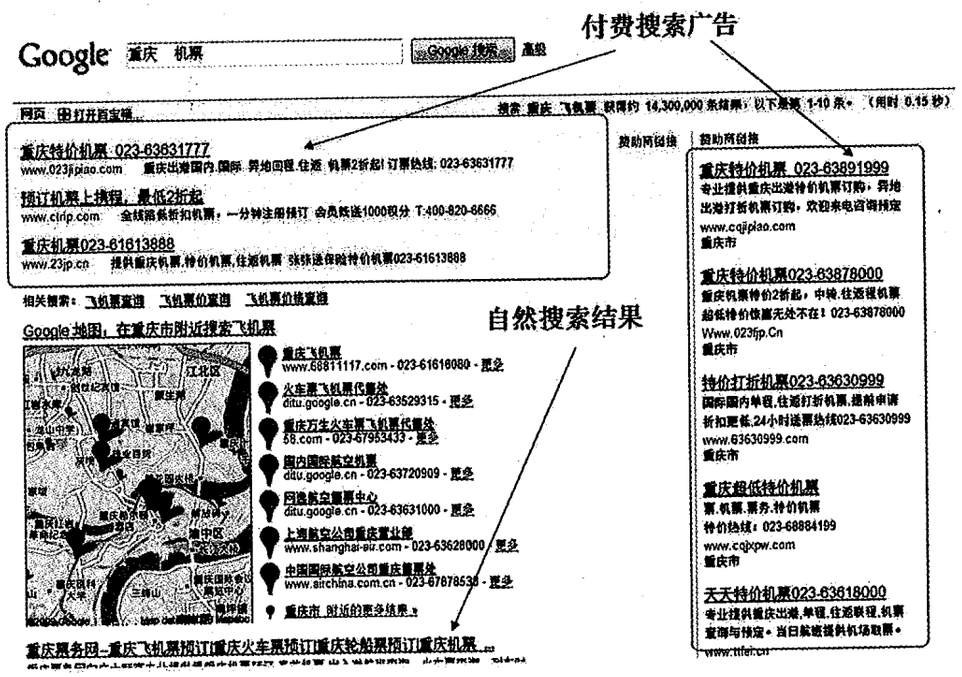


图 5—2　谷歌搜索结果页面

但是，如果企业愿意支付一定的费用，情况就不一样了。例如，企业告诉谷歌，我愿意为每一个"重庆　机票"点击支付 1 元人民币，那么谷歌会根据企业的出价，再结合企业主页的质量好坏，产生一个综合得分，按照该综合得分同其他竞拍该关键词的机票代理对比排序，决定是否展现该企业的主页。如果决定展现，还将同时决定位置。如果企业的最终得分最高，主页

将有幸被展现在最靠前的位置。这就是第二种搜索结果，即付费搜索结果。付费搜索广告的缺点是明显的，那就是得花钱。对于某些竞争激烈的行业（如教育培训），每个点击花费 100 多元人民币是司空见惯的事情。但是，付费搜索的好处是稳定可靠。原则上讲，只要能出足够高的价钱，企业的主页一定能够得到展示。

如今付费搜索广告已经被越来越多的广告客户所接受。为什么？前面已经提到它的一个巨大优点，那就是更加准确地瞄准目标客户。除此以外，它还有几个极其重要的特点。第一，付费搜索广告价格便宜。没有几百万元的现金，想做电视广告是很困难的；没有几十万元，想在体面的平面媒体上做广告也是不容易的。而在搜索引擎上做广告要多少钱呢？以机票代理为例，一般来说每个点击不超过 1 元钱。但是假设该点击能够带来一个订单，那么所产生的利润大概是 10～20 元，如果该订单是国际机票，利润将更加可观。即使是昂贵的教育培训类广告（如 MBA 培训），最多也就 100 多元一个点击，但是如果能够产生一个订单，那么利润是极其丰厚的。第二，付费搜索广告的效果是可以追踪的。前面提到过一句业内俗语："广告费的 80%都被浪费了，而且不知道浪费在哪。"这句话的另外一个含义就是面对众多的广告投放方式，到底哪一种对企业自身有效，很难评价。但如果企业所运营的是 B2C 业务，如网上电子商城，技术上，付费搜索广告可以追踪到底是哪一个关键词产生了订单，因此产生了广告效果、效果多大。同传统广告方式的糊涂账相比，这是一个了不起的优势。第三，付费广告的门槛低。例如，对谷歌而言，任何人只要有一张信用卡，理论上都可以开通其 AdWords 账户，然后竞拍感兴趣的关键词。相比较，如果想要做央视的标王，没有几亿元的现金和强大的团队是不可能的。这一点使得付费搜索广告尤其受到众多中小企业的追捧。

虽然付费搜索广告有这么多的优点，但要想把它的优点发挥到极致却是一件很不容易的事情。例如，以机票为例，能够表达重庆机票的关键词太多了！例如，"重庆机票"（请注意"重庆"和"机票"之间没有空格）和"重庆　机票"是两个不同的关键词，但是可以表达同样的购买意向。这两个词在互联网上所产生的搜索量是不一样的，能够带给企业的利润也是不一样的。两个不同的关键词面对的是两个并不完全相同的竞拍人群，应该如何出价？除此以外，还有"重庆飞机票"、"重庆　电子客票"、"重庆便宜机票"等。如果把所有的排列组合做完，总共会有多少类似的关键词呢？上亿个！

而一般来说一个普通的广告商至多能够维护不超过1万个关键词，除非是大型客户。那么，如何从上亿个可能的关键词中找出最出色的1万个是付费搜索广告研究的一个核心问题。它要求我们研究不同关键词的效果（如点击量）和它们的特征（如长度、展现量、排名等）之间的回归关系。这样一个关系，对于指导人们的搜索引擎营销意义极大。

由此可见，在这个问题中，因变量是点击量。它的取值为非负的整数，这是一种新的因变量数据类型。首先，它不同于第1章线性回归中的连续型因变量。点击量可能是0个、1个、2个等，但不可能是1.5个。因此，线性回归不能处理此类因变量。此外，它还不同于第3章中的0—1变量，因为点击量的取值范围更加丰富，不仅仅是0和1。显然，点击量也不同于第4章中的定序变量，点击量具有数值意义，因为1个点击加2个点击等于3个点击。所有这些说明，前面讨论的回归模型都不再适用，我们需要一种全新的、面向计数数据（count data）因变量的回归模型，这就是本章将要介绍的泊松回归（Poisson regression）。

5.2 数据介绍

本数据来源于国内某培训公司。该公司的主要业务是帮助各种各样的客户应对各类考试，类似的企业在国内极多。本案例数据提供者目前主要关心各种家教和MBA联考培训业务，它设计了几千个关键词，如“暑期家教”、“MBA培训”等，我们从中随机抽取了200个关键词某天的数据作为演示案例。该数据包含下面的变量信息。

1. 关键词长度（X_1）

该指标刻画了关键词的长度。例如“MBA”包含三个字符“M”，“B”，还有“A”，因此其长度为3。考虑关键词长度很重要。一般来说，关键词越短，搜索的人越多，因此所产生的展现（display，impression）就会越多。但是，更多的展现能否带来更多的点击（click）却不一定。相反，稍微长一点的关键词（如“MBA培训”）往往代表着更加清晰的搜索意图、购买意愿，因此，有可能相关点击量反而更高。

2. 展现量（X_2）

当消费者向搜索引擎输入一个关键词，搜索引擎便会依据一定规则把相关网站展现出来。对于一个给定的网站，每一次相关搜索都有可能使其被搜索引擎展现。而在一定时间段内（如一天以内），该网站被展现的次数就是展现量。由于搜索引擎对展现并不收费，因此展现本身对广告商而言是一个免费的午餐，而对于搜索引擎而言是一种投入。如果一个网站的展现量太高，但是没有产生足够点击，那么搜索引擎就仅仅忙于展现，为该网站做免费广告，无法产生足够收入。那么，搜索引擎会认为该网站的页面质量太差，进而影响该网站在同等或者类似出价情况下的竞价排名。

3. 平均点击价格（X_3）

这是指在一定时间以内（通常是一天内）所发生的所有点击的平均价格，人们也常常称其为单位点击成本（cost per click，CPC）。不同关键词的长度不一样，表述方式不一样，都会造成不同的搜索量、不同的竞争程度。受到人们追捧的热点词的平均点击价格往往都很贵。因此，在控制排名的前提下，关键词的平均点击价格在一个侧面反映该词的竞争激烈程度。

4. 平均排名（X_4）

这是指在一定时间内（通常是一天内）所发生的所有点击的平均排名情况。就某一次具体展现而言，一个特定网站的排名是一个整数，如第一名、第二名、第三名等。但是，对于一定时间内发生的所有点击而言，其平均排名更像是一个连续变量。毋庸置疑，排名是决定广告效果的一个重要因素，人们一般相信，排名越靠前，越能够引起搜索者的注意，进而越能够产生较大的点击量。但是，也有研究表明也许排名第二、第三是更好的选择。

以上讨论的是付费搜索广告研究中涉及的几个最常见的变量，从它们能够衍生出来各种常见的指标。例如，通过点击量除以展现量可以获得点击率（conversion rate），通过点击量乘以单位点击成本再除以展现量可以获得千次展现成本。当然，实际上能够影响付费搜索广告效果的因素非常多。谁能够把握理解更多的因素，谁就有可能更加准确地预测各个关键词的广告效果，就有可能在竞争激烈的市场中取得优势。

5.3 描述分析

同前面几章一样，我们假设本案例涉及的数据存放在目录“D:\商务数据分析与应用\案例数据”下的 CSV 文件“第 5 章.csv”中。简单展示如下：

	A	B	C	D	E
1	关键词长度	展现量	平均点击价格	平均排名	点击量
2	3	761	146.37	2.28	11
3	5	8	105.72	1	1
4	5	2	0	3.5	0
5	5	2	0	1.5	0
6	5	1	0	3	0
7	7	1	0	1	0
8	5	20	0	2.63	0
9	7	8	0	3.88	0
10	5	5	43.07	1.6	2

其中比较值得注意的是最后一列，这是因变量，取值为非负整数。然后 SAS 读入：

```
data A0;
    infile "D:\商务数据分析与应用\案例数据\第5章.csv"
    firstobs=2 delimiter=",";
    input Y X1-X4;
run;
```

原始数据“第 5 章.csv”被读入 SAS 环境，并存放在数据集 A0 中。在 SAS 的资源管理器下可以找到该数据，然后展示如下：

	Y	X1	X2	X3	X4	id
1	11	3	761	146.37	2.28	1
2	1	5	8	105.72	1	2
3	0	5	2	0	3.5	3
4	0	5	2	0	1.5	4
5	0	5	1	0	3	5
6	0	7	1	0	1	6
7	0	5	20	0	2.63	7
8	0	7	8	0	3.88	8
9	2	5	5	43.07	1.6	9
10	1	5	49	48.1	1.83	10

模仿第 1 章对所有变量做描述分析如下：

```
data A0; set A0; id=_n_; run;
proc transpose data=A0 out=A1; by id; var Y X1-X4; run;
proc sort data=A1; by _name_; run;
proc univariate data=A1 noprint;
    by _name_; var col1;
    output out=A2 n=n mean=mean std=std min=min median=median max=max;
run;
data A2; set A2; format mean 5.3 std 5.3 min 5.3 median 5.3 max 5.3; run;
```

结果存放在 SAS 数据集 A2 中。在 SAS 的资源管理器中，展现如下：

	以前的变量名	COL1 的非缺失值数	COL1 的均值	COL1 的标准差	COL1 的最大值	COL1 的中位数	COL1 的最小值
1	X1	200	6.705	2.248	13.00	7.000	2.000
2	X2	200	40.18	200.7	2242	3.000	1.000
3	X3	200	6.834	22.50	148.4	0.000	0.000
4	X4	200	4.540	4.068	26.00	3.000	1.000
5	Y	200	0.385	1.472	14.00	0.000	0.000

从上表可以获得下列主要结论：

- 首先从最后一行可以看到在 $n=200$ 个关键词中，平均每个关键词能够产生的点击为 0.385，这等价于平均 1/0.385≈2.6 个关键词产生一个点击。整个账户产生了 200×0.385＝77 个点击。根据我们的有限经验，这样一个点击率是相当不错的。如果能够保持该点击率，并把关键词的个数从 200 个拓展到 2 000 个，那么就能够获得超过 700 个点击。当然，如何从 200 个关键词成长为2 000 个，是一个非常不容易的过程，纯粹依靠工程师或者业务员的经验判断是远远不够的，这时候特别需要统计学的帮助。
- 从第二行可以看出这 200 个关键词的平均长度为 6.7，标准差为 2.2。这说明该账户的大多数关键词为长尾词。同长度比较短的短尾词相比，长尾词能够产生的展现量、点击量都比较小，因为搜索者都比较懒惰，不愿意在搜索引擎上敲入太多汉字。但是，长尾词所产生的点击率（click through rate）或者转化率（conversion rate）都不错。该结果也很合理，因为如果一个消费者输入的汉字越多，他表达的购买意向越强烈，他越知道自己在找什么，因此越容易点击相应网站。
- 从上表第三行可以看出，平均展现量为 40.18。但是，最大值为 2 242，这说明不同关键词的展现量差别很大。这也是一个很合理的现象。常常很多长度较短的热点词备受关注，搜索量很大，点击量也很大。例如，

“机票”、“手机”、“家教”、“MBA”等词就能够产生很多展现。少数的热点词，往往能够产生整个账户超过一半以上的展现。如果企业的目的是产生足够的展现，为企业的品牌服务（因此不在乎订单量），那么这些词是必须关注的。再如一个B2C企业，最关心的是订单和销售，那么这些词不一定能够产生大量订单，甚至有可能以极快的速度消耗企业的广告预算，是不折不扣的毒药。因此，认真仔细的统计分析非常必要。

- 从第四行可以看到，这200个词的平均竞价为6.8，但是最高能够达到146.4。这说明不同关键词的竞价差别特别大。这立刻产生一个问题，为什么？尤其是那些极其昂贵的关键词，那么高的价格是否必要？此外，那些非常便宜的关键词是否又太便宜了？能否通过调高它们的价格，获得更好的位置，然后产生更高的销售利润？由此可见，合理竞价是成功搜索引擎营销的另一个关键。

- 最后，从第五行的结果可以看到，这200个词的平均排名为4.54，在前5名以内。这说明，该企业的竞价策略比较激进，努力争取了很多的好位置。在实际工作中，很多广告商发现排名第一太贵，排名太靠后没效果。因此，一个最简单的傻瓜原则就是咬准第2到第5或者第2到第10之间的某个位置。但是，显然这不可能是一个最优的策略，那么最优的又会是什么呢？

5.4 统计模型

前面介绍了数据变量，并做了描述分析。下面我们详细讨论如何构造一个关于定序因变量的回归模型，为此定义解释变量向量 $X=(1, X_1, X_2, X_3, X_4)'$，相应的回归系数为 $\beta=(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)'$，其中 β_0 是截距项。再定义线性组合：

$$X'\beta=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4$$

如何探讨因变量 Y 和 $X'\beta$ 的关系呢？同0—1逻辑回归一样，直接定义 $Y=X'\beta+\epsilon$ 是不合适的，因为等号的右边是一个取值任意的量，而等号的左边是一个离散的定量的指标，该指标值为非负整数。

要解决该矛盾，首先需要找到一种可以描述非负整数的概率分布，在此基础上通过适当修改，产生一个合理的回归模型。有哪些概率分布能够产生

非负整数呢？首先可以想到的大概是二项式分布（binomial distribution）。二项式分布的概率模型说的是随机投掷 m 次硬币，总共能够获得正面的概率。所以，二项式分布能够产生非负整数。但是，二项式分布有一个缺陷，那就是它的取值不能超过一个事先设定的最大值 m，而这个最大值 m 没有任何实际意义。例如，如果我们想用二项式分布描述点击量，m 应该是多少呢？从理论上讲，一个关键词的点击量的取值可以任意大，只要有一个人乐得用足够快的速度不停点击。所以，我们需要一种概率模型，它不仅仅能够产生非负整数，而且没有上界。你会发现满足这样条件的概率分布极其有限，而其中最常见的一种就是泊松分布（Poisson distribution）。该分布的概率函数为：

$$P(Y=k)=\frac{\lambda^k}{k!}\exp(-\lambda), k\geqslant 0$$

乍一看，泊松分布的概率函数很奇怪。但是，该分布在概率论中有着极其重要的作用，对很多更加复杂的概率模型而言，这是其基础建筑。简单地说，泊松分布是一种有着很多优良性质的概率分布，同指数分布（exponential distribution）这种连续分布有着密切联系。它还有一个特征，就是均值方差相等，即：$E(Y)=\mathrm{var}(Y)=\lambda$。

下面再考虑，如何能够建立一种 $X'\beta$ 和 Y 之间的回归模型呢？首先注意到，泊松分布是一个单参数模型。也就是说，它的概率分布完全由一个单一的参数 λ 决定。因此，如果因变量 X 能够影响 Y 的行为，那么它必须通过影响 λ 来实现。所以，只要能够建立一种 λ 和 $X'\beta$ 之间的函数关系，那么就可以获得一个关于计数数据的回归模型。也许我们可以首先尝试假设：$\lambda=X'\beta$。这似乎是一个不错的假设，因为等号的左右两边都是连续的。但还存在一个问题：等号左边 λ 是一个正数，而其右边 $X'\beta$ 有可能是负数。因此，我们需要做进一步的修改如下：

$$\log(\lambda)=X'\beta$$

即

$$\lambda(X'\beta)=\exp(X'\beta)$$

这同前面的泊松概率函数一起就构成了人们常用的泊松回归模型。

同其他回归模型类似，泊松回归也关心回归系数 β。对于一个给定的解释变量 X_j，$\beta_j=0$ 意味着在给定其他解释变量的前提下，该指标对于解释条

件均值$\lambda(X'\beta)$没有任何帮助。因此，对于解释定序变量Y的随机行为也没有任何帮助。但是，如果$\beta_j>0$，那么在给定其他解释变量不变的前提下，指标X_j的上升会带来条件均值$\lambda(X'\beta)$的上升。这等价于说Y的取值更有可能变大。从某个角度看来，这似乎是一种“正”相关。相反，如果$\beta_j<0$，那么在给定其他解释变量不变的前提下，指标X_j的上升会带来条件均值$\lambda(X'\beta)$的下降。也就是说，因变量Y取值更有可能变小，这似乎是一种“负”相关。

同逻辑回归以及定序回归类似，泊松回归的参数可以通过极大似然估计获得。具体地说，我们用(Y_i, X_i)代表来自第i个个体的数据，其中Y_i是因变量，$X_i=(X_{i1}, X_{i2}, \cdots, X_{ip})'$是相应的解释变量。它们的联合似然函数为：

$$\prod_{i=1}^{n} P(Y_i \mid X_i) = \prod_{i=1}^{n} \frac{\lambda(X_i'\beta)^{Y_i}}{k!}\exp\{-\lambda(X_i'\beta)\}$$

对它做对数变换后，得到对数似然函数为：

$$\mathcal{L}(\beta) = \sum_{i=1}^{n}\log\{P(Y_i \mid X_i)\} = C + \sum_{i=1}^{n}[Y_i\log\{\lambda(X_i'\beta)\} - \lambda(X_i'\beta)]$$

其中C是一个和回归系数β无关的常数。可以通过极大化该对数似然函数获得极大似然估计，即$\hat{\beta}=\text{argmax}_\beta\mathcal{L}(\beta)$。标准的统计学理论告诉我们，该估计量是渐进无偏的、相合一致的，而且是极限正态的。因此，可以对每个系数的估计误差有所判断，进而计算相应的p-值，再做统计学推断，即假设检验$H_0:\beta_j=0$，$H_1:\beta_j\neq0$。同逻辑回归以及定序回归一样，泊松回归没有“残差”这个概念，因此无法定义残差平方和。但是，可以定义离差为$DEV=-2\mathcal{L}(\hat{\beta})$。然后，也可以检验全局检验$H_0:\tilde{\beta}=0$，$H_1:\beta\neq0$，其中$\tilde{\beta}=(\beta_1, \beta_2, \cdots, \beta_p)'$。当某一个解释变量为多水平定性因素时，该因素的显著性水平也可以模仿第4章案例计算。由于整个过程类似，为节省篇幅，这里不再赘述。

5.5 预测评估

对泊松回归而言，预测在实际中有着非常重要的应用。例如，前面提到一般工程师或者业务员能够根据他们的宝贵经验生成大概200个左右的关键

词，我们称之为种子词。但是，如何把这200个关键词成长为2万个呢？从词典里找出上亿个词都不是难事，关键是如何找到最有用的2万个，这就需要对关键词的效果做出预测。

具体情况如下。(Y_i, X_i)（$i=1, \cdots, n$）代表历史数据。在假设(Y_i^*, X_i^*)（$i=1, \cdots, m$）是未来数据。对于未来数据而言，解释变量X_i^*是已知的，但是因变量Y_i^*却是未知的。就本案例而言，Y_i是某关键词所产生的点击量，X_i是它各种预测指标，例如排名、长度、展现量等。那么，X_i^*可以是另外一个关键词的相关指标，而且已知。但是，未知的是Y_i^*，这是该词上线后所产生的流量。如何预测呢？首先通过分析历史数据建立泊松回归模型，获得极大似然估计$\hat{\beta}$。然后，将此估计应用于未来数据X_i^*，对其因变量Y_i^*的均值估计如下：

$$E(Y_i^*=1|X_i^*)\approx\lambda(X_i^{*\prime}\hat{\beta})=\exp(X_i^{*\prime}\hat{\beta})$$

值得注意的是，虽然因变量Y_i^*取值整数，但这并不妨碍它的均值是一个带小数点的正实数。因此，我们可以直接采用$\lambda(X_i^{*\prime}\hat{\beta})$对$Y_i^*$做预测。

接下来，再讨论如何对预测精度做出合理判断。当然，最简单的做法莫过于直接计算平均的绝对预测误差，即

$$\text{绝对预测误差}=m^{-1}\sum_{i=1}^{m}|Y_i^*-\lambda(X_i^{*\prime}\hat{\beta})|$$

当然，是否一定要用绝对值，用平方误差（就像最小二乘估计一样）可以吗？当然可以。但是，不管是绝对误差，还是均方误差，都有一个致命缺陷，那就是忽略了计数数据的异质性。简单地说，少数关键词的点击量很大，而多数很少。因此，同样一个单位的预测误差，对那些高点击量的关键词，可以忽略，但是对于那些低点击量的词，就很大了。因此，一个更加合理的标准应该相对地看待预测误差。一个可能的标准是：

$$\text{相对预测误差}=m^{-1}\sum_{i=1}^{m}\frac{|Y_i^*-\lambda(X_i^{*\prime}\hat{\beta})|}{Y_i^*}$$

这个标准就好懂多了。它说的是平均而言，泊松模型的预测误差$|Y_i^*-\lambda(X_i^{*\prime}\hat{\beta})|$相对于真实水平$Y_i^*$有多大。

上面定义的相对预测误差虽然很好懂，但是在实际中很难应用。因为很多长尾词的真实点击量$Y_i^*=0$。因此，把它放在分母是一个问题。当然，

$Y_i^*=0$ 并不说是说该词没有贡献，因为也许下一次它的点击量就不是 0 了。从统计学理论的角度讲，此类词的均值 $\lambda(X_i^{*\prime}\beta)$ 很低，但不是严格为 0。因此，更合理的相对预测误差应该定义如下：

$$\text{相对预测误差}=m^{-1}\sum_{i=1}^{m}\frac{|Y_i^*-\lambda(X_i^{*\prime}\beta)|}{\lambda(X_i^{*\prime}\beta)}$$

由于 β 是一个未知的参数，因此实际操作中必须用估计量替代。相应修正如下：

$$\text{相对预测误差}=m^{-1}\sum_{i=1}^{m}\frac{|Y_i^*-\lambda(X_i^{*\prime}\hat{\beta})|}{\lambda(X_i^{*\prime}\hat{\beta})}$$

这似乎是一个更合理的关于预测精度的评判标准。

但实际上该标准也有问题。对于本案例而言，大部分关键词是长尾词，它们的点击量是 0。因此，如果我们能够成功地预测它们相应的 $\lambda(X_i^{*\prime}\hat{\beta})$ 值很小，那么这就是一个不错的预测。但是，如果按照上面定义的相对预测误差，会有 $|Y_i^*-\lambda(X_i^{*\prime}\hat{\beta})|/\lambda(X_i^{*\prime}\hat{\beta})=1$，这是一个很大的数字。因此，对相对预测误差再做稍微修改，使得其分母不会特别小，如下：

$$\text{相对预测误差}=m^{-1}\sum_{i=1}^{m}\frac{|Y_i^*-\lambda(X_i^{*\prime}\hat{\beta})|}{1+\lambda(X_i^{*\prime}\hat{\beta})}$$

这样，前面提到的所有问题就都解决了。例如，该标准可以考虑关键词的异质性，也能兼顾长尾词的特征。当然，为什么分母要加 1 呢？加 2 可以吗？加 0.5 可以吗？当然都可以。但是，到底加多少才适合，这需要经验和时间。以上就是本案例最终采用的预测精度标准。

5.6 SAS 编程

到此为止，逻辑回归的基本理论就介绍完了。我们回到本章的搜索引擎案例，通过 SAS 程序具体分析如下：

```
proc genmod data=A0;
    model Y=X1-X4/dist=poisson;
    output out=A3 p=p;
run;
```


相应的分析结果如下所示。

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.2785	0.4085	0.4778	2.0792	9.79	0.0018
X1	1	-0.3866	0.0753	-0.5341	-0.2390	26.37	<.0001
X2	1	0.0007	0.0005	-0.0003	0.0017	1.64	0.2004
X3	1	0.0180	0.0027	0.0128	0.0232	46.06	<.0001
X4	1	-0.2011	0.0889	-0.3753	-0.0268	5.12	0.0237
Scale	0	1.0000	0.0000	1.0000	1.0000		

我们可以获得的主要结论如下：

● 首先注意到，第二个解释变量 X_2 展现量在 5%的水平下不显著。这似乎在说展现量的大小同点击量的大小无关。这显然不对，没有充足的展现量，哪里来的点击量？但是，上面的结果到底是什么意思呢？它说的是，在控制其他变量的前提下（尤其是平均点击价格 X_3 和平均排名 X_4），展现量 X_2 不再重要。那么它的信息被谁代替了呢？很有可能被平均点击价格 X_3 代替了，因为该价格是所有竞价参与者长期经验的反映，一个关键词的点击量越大，其平均价格就会越贵，当然这常常伴随着更高的展现量。所以，在给定平均点击价格 X_3 的前提下，展现量 X_2 不再重要。

● 显然，平均排名 X_4 很重要。它的极大似然估计量为 −0.201，是负的。这说明，排名越靠前的关键词，点击量越大。这同我们的经验常识相符合。

● 平均点击价格 X_3 是一个很重要的解释变量，对预测点击量的大小意义重大。它的极大似然估计量为 0.018，是正的。这说明，价格越昂贵的关键词，点击量越大。一般的常识认为，价格越高，排名越靠前，因此点击量越大。但是，该解释在这里不大合适。原因是，该结果已经控制了平均排名 X_4 的作用。它反映的是在排名还有其他因素（如长度）相同的情况下，越贵的关键词越能够产生更多的点击。这也是一个非常合理的结论。不同的关键词，即使排名相同，长度相同，产生的点击量也不相同。因此，人们会乐意为产生更多点击量的关键词支付更高的价格。久而久之，市场上就会反映出来，价格贵的词一般来说点击量更高。

● 最后，我们还发现关键词长度 X_1 也很重要。它的极大似然估计量为 −0.387，是负的。这说明，越长的关键词，点击量越小。这也是一个非常合理的结论。搜索者都是懒惰的，搜索引擎之所以存在，就是因为人们希望

通过最少的提问获得最准确的结果。因此，相比较而言，愿意输入很长关键词的搜索者人数，远远少于不愿意的人数。因此，关键词越长，搜索量越少，点击量越小。

下面再展示一下如何用 SAS 程序对预测精度予以评估。和前面几章一样，需要强调一下，这里没有区分内外样本，这是一个值得注意的缺陷。

```
data A3; set A3; RPE=abs(p-Y)/(1+p); run;
proc univariate data=A3; var RPE; run;
```

相应的 SAS 输出如下：

基本统计测度

位置		变异性	
均值	0.249256	标准偏差	0.47852
中位数	0.137405	方差	0.22898
众数	0.055493	极差	5.47045
		四分位极差	0.22723

NOTE: 显示的众数是 2 个众数的最小值（计数为 4）。

由此可见，最终的相对预测误差的均值是 24.9%，这是一个相当不错的数字。根据我们的有限实际经验，该精度能够满足绝大多数搜索引擎营销关键词研究的需要。

5.7 总结讨论

本章通过付费搜索广告的案例，对泊松回归模型的核心理论做了简要论述，对相应的 SAS 编程做了详细展示。从理论上讲，泊松回归不是处理计数数据的唯一回归模型，只要能够找到一种没有上界的、取值为非负整数的概率分布，就可以获得一个回归模型。例如，常见的计数回归模型，还有负二项（negative binomial）回归模型。除此以外，人们还常常发现计数数据中 0 的个数有可能远远超出模型所能解释的范围。这说明，该因变量取值是否为 0 是一个独立的过程，在取值有可能非 0 的情况下，是另外一个过程。对这一类数据，人们可以考虑带有零膨胀的泊松（zero inflated Poisson）回归模型。

最后结合本章案例，作者希望对搜索引擎营销在中国的实践做一点粗浅的评论。在业界付费搜索广告的重要性已经毋庸置疑，但是，如何将它的效果充分发挥却不是一个简单的问题，这是一个复杂的统计优化（statistical optimization）问题。当企业花费丰厚的薪水招聘计算机工程师、广告咨询师的时候，都忽略了一点，那就是统计师的重要性。只有这三种专业精英的紧密配合，才能将付费搜索广告的效果发挥到极致。这样的理念，只有最优秀的、最有远见的引擎营销企业才可能拥有。

举一个具体的例子。作者同业界的合作伙伴，来自 CubeAD（博雅立方，www. cubead. com）的同仁一起，把宝贵的广告营销理念、先进的计算机网络技术，还有独特的统计学算法结合在一起，设计了 CubeSearch 平台（cubesearch. cubead. com）。在这个平台上，广告主可以根据自己的历史数据、过去表现，自动生成成千上万的关键词。然后对这些关键词自动跟踪，每日自动调整竞价策略。没有统计学的深刻介入，产生这样的智能系统是不可能的。那么效果如何呢？

图 5—3 为大家展示了一个账户的数据。系统从零时刻接手，经过

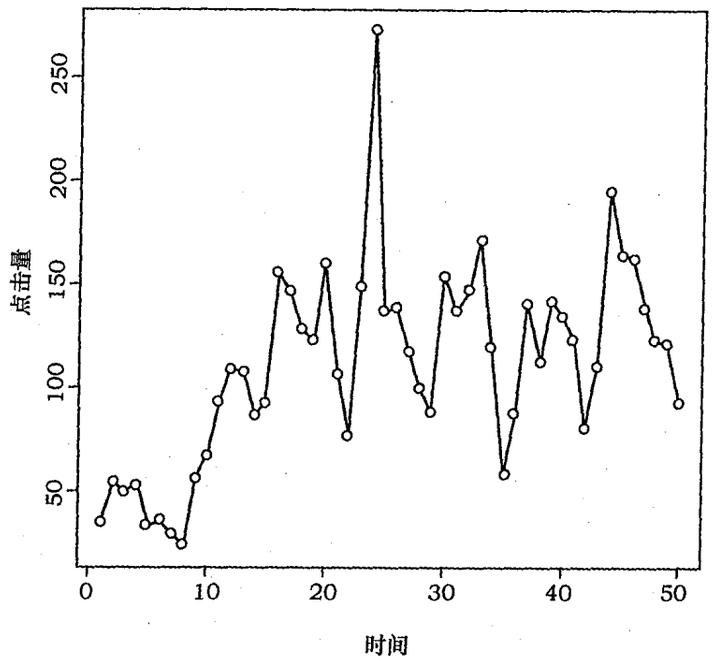


图 5—3　点击量时间分布图

大概一个星期的时间，其点击量开始大幅攀升。从最开始的不到 50 个一直成长到大概日均 150 个，增加了两倍有余。但是，单位点击价格如何呢？

从图 5—4 可以看到，单位点击价格一路下降。从最高的时候每个点击 40 元下降并最终稳定在 15 元左右，几乎下降了 2/3。因为广告效果非常理想，广告主决定多次增加广告预算，因此每日广告花费情况如图5—5所示。

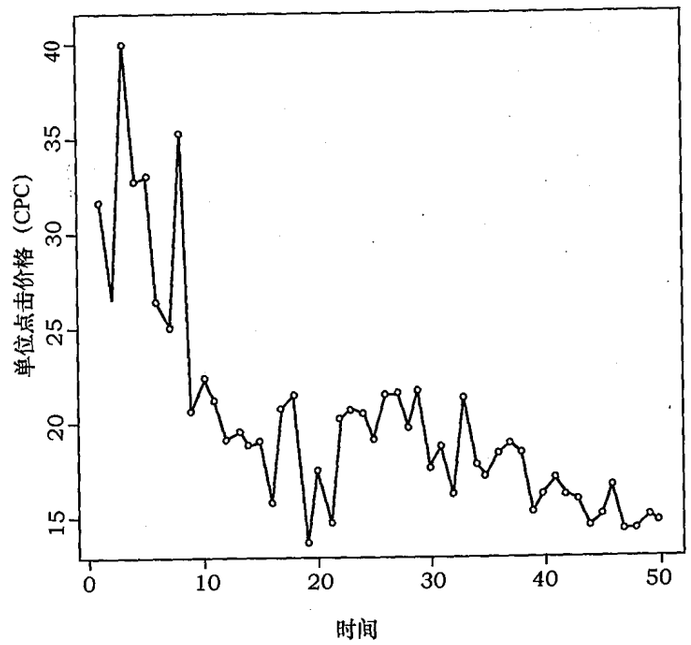


图 5—4　单位点击价格时间分布图

从图 5—5 可以看到，每日的广告花费从最开始的略多于 1 000 元不停增加，最后稳定在 2 000 元左右。广告预算增加了 1 倍。因此，这样的付费广告优化行为不仅为广告主带来了利益，还为搜索引擎吸引了更多的广告投入。当然，请不要忘记，消费者也因为更加准确的搜索展现节省了宝贵的时间。在这个游戏中，所有的人都得以获益。

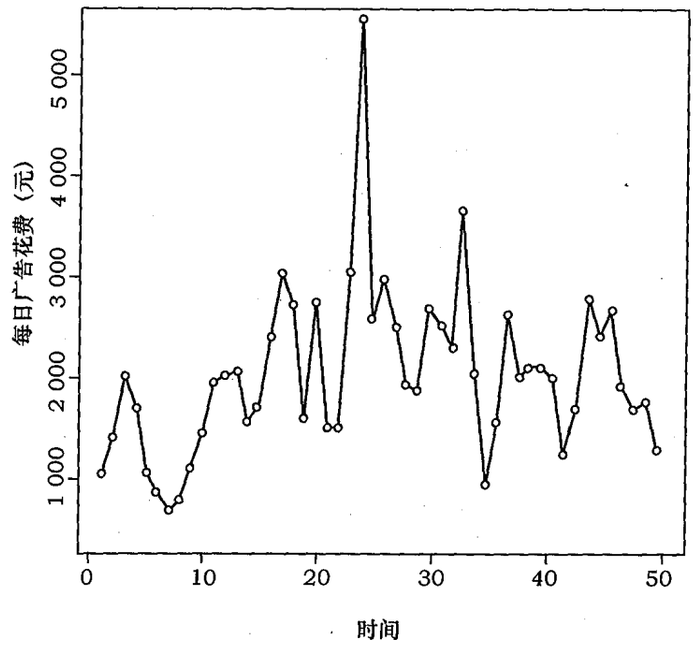


图 5—5　每日广告花费时间分布图

附录 5A 分析报告

付费搜索广告点击量研究

1. 研究目的

通过分析 200 个关键词的长度、展现量、排名等特征，理解各个关键词点击量的规律，为优化广告投放提供参考。

2. 背景介绍

现代企业格外关注信息在目标客户中的传递过程。因此，也格外看重各种媒体信息的传播作用。企业希望通过最有效的方式将信息准确传递给目标客户。最传统的信息传播依赖口碑，即通过现有客户的口口相传将企业的产品服务信息传递给新客户。过去的研究表明，口碑对产品服务信息的传递起着重要的作用。例如，新产品上市，口碑的作用将直接影响销量，进而影响产品的生命周期。直至今日，相关营销研究仍然认为口碑传播对企业产品服务的成功极其重要。通过口碑传播所获得的客户，同通过其他手段（如促销）获得的客户相比，具有更高的忠诚度，能够为企业带来更多的利润。

但口碑传播也有它的劣势，那就是不容易控制。企业很难通过资源的投入过多地改变口碑传播的过程。因此，除了口碑传播以外，企业还需要其他手段帮助传播产品服务信息，最常用的手段之一就是广告。常见的广告媒介有电视、电台、报纸、杂志、灯箱、路牌、门户网站等。但是，这些传统广告都是被动形式的。也就是说，此类广告的展现无法做到根据客户的不同而不同。因此，有可能给想买衣服的消费者看洋酒的广告，而给想喝酒的顾客展示电器广告。所以，此类广告的绝大多数展现投向了无效客户，这就难怪它们的效果差强人意。

作为一个有效的补充，越来越多的企业考虑在搜索引擎上做广告。同传统的被动型广告相比，搜索引擎上的付费搜索广告是主动型广告。只有在消费者表达了购买某特定产品意愿的前提下（如“北京飞上海机票”），相关网页才会展示。因此，同传统广告形式相比，付费搜索广告更加有效，更加便

宜，而且由于信息技术的进步，还可以追踪比较其广告效果。有研究机构预测，美国付费搜索广告的营业额将从 2007 年的大约 80 亿美元成长为 2012 年的 154 亿美元（*US Online Advertising Forecast*，*2007 to 2012*，jupiterresearch. com，2007）。届时，付费搜索广告占整个互联网广告的份额将远远超过 50%。

3. 指标设计

本数据来源于国内某培训公司，该公司的主要业务是帮助各种各样的客户应对各类考试，目前主要关心各种家教和 MBA 联考培训业务，它设计了几千个关键词，如“暑期家教”、“MBA 培训”等。我们从中随机抽取了 200 个关键词某天的数据作为我们的演示案例。该数据包含下面的变量信息。

- 关键词长度（X_1）

该指标刻画了关键词的长度。例如“MBA”包含三个字符“M”，“B”，还有“A”，因此其长度为 3。考虑关键词长度很重要。一般来说，关键词越短，搜索的人越多，因此所产生的展现（display，impression）就越多。但是，更多的展现能否带来更多的点击（click）却不一定。相反，稍微长一点的关键词（如“MBA 培训”）往往代表着更加清晰的搜索意图、购买意愿，因此，有可能相关点击量反而更高。

- 展现量（X_2）

当消费者向搜索引擎输入一个关键词，搜索引擎便会依据一定规则把相关网站展现出来。对于一个给定的网站，每一次相关搜索都有可能使其被搜索引擎展现。而在一定时间段内（如一天以内），该网站被展现的次数就是展现量。由于搜索引擎对展现并不收费，因此展现本身对广告商而言是一个免费的午餐，而对于搜索引擎而言是一种投入。如果一个网站的展现量太高，但是没有产生足够点击，那么搜索引擎就仅仅忙于展现，为该网站做免费广告，无法产生足够收入。那么，搜索引擎会认为该网站的页面质量太差，进而影响该网站在同等或者类似出价情况下的竞价排名。

- 平均点击价格（X_3）

这是指在一定时间以内（通常是一天内）所发生的所有点击的平均价格，人们也常常称其为单位点击成本（cost per click，CPC）。不同关键词的长度不一样，表述方式不一样，都会造成不同的搜索量、不同的竞争程

度。受到人们追捧的热点词的平均点击价格往往都很贵。因此，在控制排名的前提下，关键词的平均点击价格在一个侧面反映该词的竞争激烈程度。

● 平均排名（X_4）

这是指在一定时间内（通常是一天内）所发生的所有点击的平均排名情况。就某一次具体展现而言，一个特定网站的排名是一个整数，如第一名、第二名、第三名等。但是，对于一定时间内发生的所有点击而言，其平均排名更像是一个连续变量。毋庸置疑，排名是决定广告效果的一个重要因素，人们一般相信，排名越靠前，越能够引起搜索者的注意，进而越能够产生较大的点击量。但是，也有研究表明也许排名第二、第三是更好的选择。

以上讨论的是付费搜索广告研究中涉及的几个最常见的变量，从它们能够衍生出来各种常见的指标。例如，通过点击量除以展现量可以获得点击率（conversion rate），通过点击量乘以单位点击成本再除以展现量可以获得千次展现成本。当然，实际上能够影响付费搜索广告效果的因素非常多。

4. 描述分析

我们首先计算因变量和各个解释变量的均值、标准差等描述统计量，如表 5—1 所示。

表 5—1　　各个变量的描述分析

变量名称	样本量	均值	标准差	最大值	中位数	最小值
关键词长度	200	6.705	2.248	13.00	7.000	2.000
展现量	200	40.18	200.7	2 242	3.000	1.000
平均点击价格	200	6.834	22.50	146.4	0.000	0.000
平均排名	200	4.540	4.068	26.00	3.000	1.000
点击量	200	0.385	1.472	14.00	0.000	0.000

从表 5—1 可以获得的主要结论如下：

● 首先从最后一行可以看到在 $n=200$ 个关键词中，平均每个关键词能够产生的点击为 0.385，这等价于平均 $1/0.385\approx2.6$ 个关键词产生一个点击。整个账户产生了 $200\times0.385=77$ 个点击。根据我们的有限经验，这样一个点击率是相当不错的。如果能够保持该点击率，并把关键词的个数从 200 个拓展到 2 000 个，那么就能够获得超过 700 个点击。当然，从 200 个关键词成长为 2 000 个是一个非常不容易的过程，仅仅依靠工程师或者业务员的经验判断是远远不够的。

● 从第二行可以看出这 200 个关键词的平均长度为 6.705，标准差为 2.248。这说明该账户的大多数关键词为长尾词。同长度比较短的短尾词相比，长尾词能够产生的展现量、点击量都比较小，因为搜索者都比较懒惰，不愿意在搜索引擎上敲入太多汉字。但是，长尾词所产生的点击率（click through rate）或者转化率（conversion rate）都不错。该结果也很合理，因为如果一个消费者输入的汉字越多，他表达的购买意向越强烈，他越知道自己在找什么，因此越容易点击相应网站。

● 从第三行可以看出，平均展现量为 40.18。但是，最大值为 2 242，这说明不同关键词的展现量差别很大。这也是一个很合理的现象。常常很多长度较短的热点词备受关注，搜索量很大，点击量也很大。少数的热点词（如“MBA”），往往能够产生整个账户超过一半以上的展现。

● 从第四行可以看到，这 200 个词的平均竞价为 6.834，但是最高能够达到 146.4。这说明不同关键词的竞价差别特别大。这立刻产生一个问题，为什么？尤其是那些极其昂贵的关键词，那么高的价格是否必要？此外，那些非常便宜的关键词是否太便宜了？能否通过调高它们的价格，获得更好的位置，然后产生更高的销售利润？由此可见，合理竞价是未来工作的一个重点。

● 最后，从第五行的结果可以看到，这 200 个词的平均排名为 4.54，在前 5 名以内。这说明，该企业的竞价策略比较激进，努力争取了很多的好位置。

5. 模型分析

在描述分析的基础上，我们通过泊松定序回归模型对各个因素同点击量大小之间的关系做模型分析，结果如表 5—2 所示。

表 5—2　　各个因素卡方检验结果

变量名称	参数估计	标准误差	卡方统计量	p-值
截距项	1.278 5	0.408 5	9.79	0.001 8
关键词长度	−0.386 6	0.075 3	26.37	<0.000 1
展现量	0.000 7	0.000 5	1.64	0.200 4
平均点击价格	0.018 0	0.002 7	46.06	<0.000 1
平均排名	−0.201 1	0.088 9	5.12	0.023 7

从表 5—2 我们可以获得的主要结论如下：

● 首先注意到，第二个解释变量 X_2 展现量在5%的水平下不显著。这似乎在说展现量的大小同点击量的大小无关。这显然不对，没有充足的展现量，哪里来的点击量？但是，上面的结果到底是什么意思呢？它说的是，在控制其他变量的前提下（尤其是平均点击价格 X_3 和平均排名 X_4），展现量 X_2 不再重要。那么它的信息被谁代替了呢？很有可能被平均点击价格 X_3 代替了，因为该价格是所有竞价参与者长期经验的反映，一个关键词的点击量越大，其平均价格就会越贵，当然这常常伴随着更高的展现量。所以，在给定平均点击价格 X_3 的前提下，展现量 X_2 不再重要。

● 显然，平均排名 X_4 很重要。它的极大似然估计量为 -0.201，是负的。这说明，排名越靠前的关键词，点击量越大。这同我们的经验常识相符合。

● 平均点击价格 X_3 是一个很重要的解释变量，对预测点击量的大小意义重大。它的极大似然估计量为 0.018，是正的。这说明，价格越昂贵的关键词，点击量越大。一般的常识认为，价格越高，排名越靠前，因此点击量越大。但是，该解释在这里不大合适。原因是，该结果已经控制了平均排名 X_4 的作用。它反映的是在排名还有其他因素（如长度）相同的情况下，越贵的关键词越能够产生更多的点击。这也是一个非常合理的结论。不同的关键词，即使排名相同，长度相同，产生的点击量也不相同。因此，人们会乐意为产生更多点击量的关键词支付更高的价格。久而久之，市场上就会反映出来，价格贵的词一般来说点击量更高。

● 最后，我们还发现关键词长度 X_1 也很重要。它的极大似然估计量为 -0.387，是负的。这说明，越长的关键词，点击量越小。这也是一个非常合理的结论。搜索者都是懒惰的，搜索引擎之所以存在，就是因为人们希望通过最少的提问获得最准确的结果。因此，相比较而言，愿意输入很长关键词的搜索者人数，远远少于不愿意的人数。因此，关键词越长，搜索量越少，点击量越小。

6. 预测评估

接下来，我们考虑用建立的泊松回归模型做预测。假设 (Y_i, X_i)（$i=1, \cdots, n$）代表历史数据，而 (Y_i^*, X_i^*)（$i=1, \cdots, m$）是未来数据。通过分析历史数据建立泊松回归模型，获得极大似然估计 $\hat{\beta}$。然后，将此估计应用于未来数据 X_i^*，对其因变量 Y_i^* 的均值估计如下：

$$E(Y_i^*=1|X_i^*)\approx\lambda(X_i^{*\prime}\hat{\beta})=\exp(X_i^{*\prime}\hat{\beta})$$

计算相应的相对预测误差如下：

$$\text{相对预测误差}=m^{-1}\sum_{i=1}^{m}\frac{|Y_i^*-\lambda(X_i^{*\prime}\hat{\beta})|}{1+\lambda(X_i^{*\prime}\hat{\beta})}$$

计算表明，该相对误差大约为 24.9%。

7. 总结讨论

本研究分析了 200 个随机选取的关键词，研究了它们点击量大小同关键词长度、展现量、平均点击价格，以及平均排名之间的关系。关键词长度以及平均排名同点击量显著地负相关，而平均点击价格同点击量显著地正相关。我们也对所建立模型的预测精度予以评估，效果良好。但是值得注意的是本研究没有对内外样本予以区分，值得未来改进。

附录 5B　课后习题

超市消费者到访频数研究

1. 研究目的

研究某超市消费者下一个月的访问频数。希望理解该频数同消费者历史消费行为（主要是指历史访问频数和历史消费金额）的关系，为客户关系管理提供参考。

2. 数据介绍

某年度随机抽取的 3 995 个超市消费者（会员）。因变量是该消费者下一个月的访问频数。解释变量有：当月的访问频数和消费金额（元），以及上一个月的访问频数和消费金额。该数据存放在目录“D:\商务数据分析与应用\课后练习”下 CSV 文件“课后练习 5. csv”中。

3. 作业要求

- 问题理解：请参阅相关客户关系管理的教材，理解为什么人们关心消费者的新进性（recency）、访问频率（frequency）和货币价值（monetary value）。看这些营销学的智慧是否在数据上有所体现。
- 做完整的泊松定序回归分析，包括参数估计、假设检验和预测评估。
- 将分析结果汇总成如附录 5A 所示的简短研究报告。

附录 5C　R 程序演示

首先通过下面程序读入数据：

```
> a=read.csv("D:/商务数据分析与应用/案例数据/第5章.csv",header=T)
> names(a)=c("Y","X1","X2","X3","X4")
> a[c(1:5),]
  Y X1  X2     X3   X4
1 11  3 761 146.37 2.28
2  1  5   8 105.72 1.00
3  0  5   2   0.00 3.50
4  0  5   2   0.00 1.50
5  0  5   1   0.00 3.00
```

全变量描述分析：

```
> N=sapply(a,length)
> MU=sapply(a,mean)
> SD=sapply(a,sd)
> MIN=sapply(a,min)
> MED=sapply(a,median)
> MAX=sapply(a,max)
> result=cbind(N,MU,SD,MIN,MED,MAX)
> result
     N       MU          SD MIN MED     MAX
Y  200  0.38500    1.472282   0   0   14.00
X1 200  6.70500    2.247942   2   7   13.00
X2 200 40.17500 200.657673   1   3 2242.00
X3 200  6.83445   22.502247   0   0  146.37
X4 200  4.53970    4.068050   1   3   26.00
```

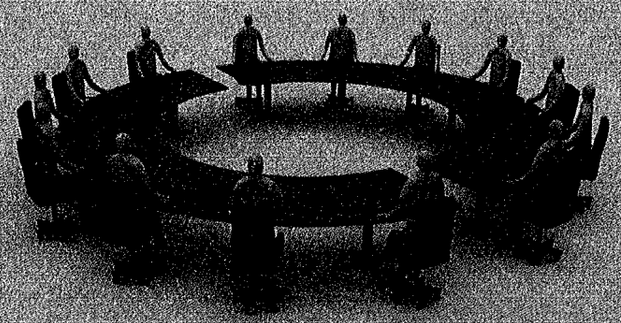
泊松回归：

```
> pos1=glm(Y~X1+X2+X3+X4,family=poisson(),data=a)
> summary(pos1)

Call:
glm(formula = Y ~ X1 + X2 + X3 + X4, family = poisson(), data = a)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0367  -0.6888  -0.4143  -0.2060   6.3874
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.2785068  0.4085366   3.129  0.00175 **
X1          -0.3865588  0.0752822  -5.135 2.82e-07 ***
X2           0.0006558  0.0005122   1.280  0.20043
X3           0.0179875  0.0026502   6.787 1.14e-11 ***
X4          -0.2010931  0.0889027  -2.262  0.02370 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 358.74  on 199  degrees of freedom
Residual deviance: 173.95  on 195  degrees of freedom
AIC: 259.75

Number of Fisher Scoring iterations: 6
```

预测与评估：

```
> pred=predict(pos1,a)
> lam=exp(pred)
> RME=abs(a$Y-lam)/(1+lam)
> summary(RME)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0006325 0.0463800 0.1374000 0.2493000 0.2730000 5.4710000
```

第6章

生存数据回归

——以员工离职管理为例