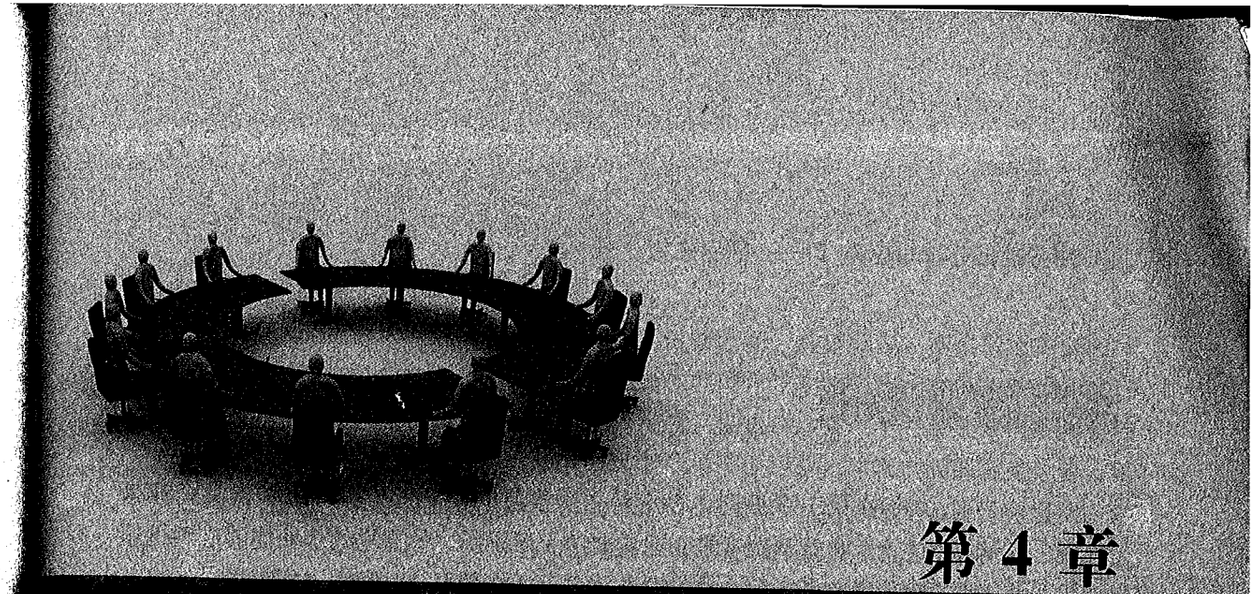


第4章

定序回归

——以消费者偏好度研究为例

4.1 背景介绍

我们在第 1 章提到，客户是企业重要的无形资产。对于零售类企业来说，这就意味着消费者是企业重要的无形资产。什么样的企业能够准确迅速地了解其消费者，什么样的企业就有可能在最短的时间内占领市场。

在实际工作中，了解消费者的方法五花八门，最基本的莫过于一线工作人员的直接接触，以及经验积累。例如，企业的客服人员往往能够第一时间了解顾客的抱怨和不满。当然，除了被动地接受顾客的抱怨投诉以外，客服人员还可以通过主动联系的方式了解顾客。这样做的好处是能够比较深入地了解消费者，但时间、精力耗费太大，难以大规模调查。而且被访问的顾客能否代表其他的大多数消费者是一个很难回答的问题。因此，管理者往往还要借助于其他手段去更加主动地了解顾客。从统计学的角度讲，就是要分析各种不同的消费者数据，以获得更加准确的判断。

可供分析的典型数据有两种。一种是实际交易数据（transaction data）。对于绝大多数现代化超市而言，此类数据非常容易获得。此类数据真实准确地记录了消费者的购买行为，非常准确地反映了消费者在复杂环境下的选择和偏好，这是此类数据最大的优点。但这也正是此类数据最大的缺点，因为真实的购买消费环境是极其复杂的。因此，如何控制去除各种干扰因素的影响非常困难。所以，人们还会关注试验数据。顾名思义，试验数据就是在一个实验环境下获得的数据。例如，有一新一旧两种不同的产品，厂商希望对比消费者对这两种产品的不同态度。在一个试验环境下，可以首先招募一批志愿消费者，然后将他们随机分组，分别尝试新产品或者旧产品，最后做出评价。此类试验最大的好处就是可以作随机分组。因此，如果有其他因素（如消费者性别）同时影响消费者的偏好，通过随机分组的方式，该影响因素会自动被平均掉（至少在样本量足够大的情况下）。试验数据所发现的相关关系有可能是因果关系。但试验数据也有其致命伤，即试验环境一般来说是高度简化的，距离真实情境相去甚远，我们有足够的理由相信消费者行为在真实情境和试验环境下是非常不一样的。最后，甚至谁也无法保证招募的试验者能够有效地代表真实的消费者。这么看来，没有什么方法可以看作完美的，但它们都是有用的。

本章将要演示的数据是标准的试验数据。具体地说，本案例所关注的是一个新产品定位的问题。生产厂商在具体开发一款新产品之前，希望知道自己的目标客户都关心产品的哪些特质。显然，这是一个对企业来说极其重要的问题。在竞争日益激烈的市场上，为了保持技术或者产品的领先优势，每年企业都会投入很大的资源用于新产品研发。但遗憾的是，有研究表明，绝大多数新产品都不成功。在投放市场后，无法迅速产生足够的销量，因此无法带来足够的利润支撑企业的正常运转，更无法弥补企业前期所做的研发投入。这样的产品，往往在上市很短一段时间后即匆匆退出。这对企业显然是一个巨大的损失。导致新产品失败的原因众多，有可能是产品的销售渠道没有铺开，不够畅通；有可能是企业的广告投放不到位；有可能是促销手段不合适；还有可能是销售团队出了问题。在众多原因中，产品定位的失败往往是一个重要的原因。简单地说，就是企业不了解自己的目标客户到底需要什么，生产了一款消费者不需要的产品。为了避免或者极小化该风险，很多企业在新产品立项的初期就会做消费者调查研究，以获得对消费者偏好（preference）的准确判断，进而指导新产品的研发。

那么如何做消费者调查研究呢？最简单的方法就是问卷调查。例如，将企业正在考虑的不同产品给消费者尝试（如果还没有生产出来，至少描述给消费者），然后征求消费者的宝贵意见，了解他们喜欢哪一个产品或者功能多一些。本章以手机为例，一个手机制造商想推出一款面向商务人士的手机，需要了解针对此类消费人群，手机的什么功能是最重要的。也许人们会问，为什么不把所有的功能都集中在一款高级手机上呢？这样做纵然能使手机的功能无比强大，但缺点也很明显：第一，因为该手机集中了很多功能，造价必定昂贵，消费者是否愿意为那些他们并不认可的功能买单呢？答案并不乐观。第二，为了实现很多功能，手机就必须在体积和重量上有所牺牲。如果牺牲过大，也会影响消费者的偏好程度。因此，实际的产品设计必须是一个权衡取舍的过程。在这个过程中，需要通过调查研究的方式了解消费者最看重什么、第二看中什么、最不看重什么、第二不看重什么，等等。

当消费者面对一个产品，态度如何评判呢？如果我们希望用第 1 章的线性回归解决问题，就需要该消费者为该产品的偏好程度打出一个连续的得分，最好还是正态分布的。这当然不可能，这样做会让被调查者迅速失去兴趣，失去配合的耐心。因此，消费者的评判过程一定要简单！所以，实际工作中，人们只会问个大概，如：您对该产品：很喜欢、一般喜欢、无所谓、

不喜欢、极其不喜欢。这就简单多了。为了方便起见，人们常常将其编号，如：5＝很喜欢，4＝一般喜欢，3＝无所谓，2＝不喜欢，1＝极其不喜欢。请注意，这样的编号只有代码的意义，没有任何数值意义。例如，我们不能说：3（无所谓）+2（不喜欢）=5（很喜欢）。因此，该数据从本质上讲是一个定性数据。但是，同一般的定性数据（如颜色）不同的是，本数据有顺序（order）特征，如5（很喜欢）要优于4（一般喜欢），4（一般喜欢）要优于3（无所谓）等。统计学上称此类数据为定序（ordinal）数据。如果此类数据恰好是我们感兴趣的因变量，那么前面所讲的各种模型都不再适用。线性回归要求因变量是定量指标，逻辑回归要求因变量的取值只有两个可能，我们需要一种特殊的、特别为定序数据设计的回归方法，这就是下面将要讲述的定序回归。

4.2 数据介绍

手机市场竞争非常激烈，为了确立自己在市场中的相对优势地位，开发新功能是手机制造商常用的手段。因为与开发新产品相比，在现有产品的基础上增加新功能所需要的投入少，且承担的风险较小，因此不失为一种既快速又有效的方式。当企业决定为现有产品增加新功能时，往往面对众多选择，例如，下一个升级换代产品是增加拍照功能还是增加收音机或者MP3播放器？对这个问题的正确回答依赖于目标客户。不同的目标客户对不同新功能的评价喜好是不一样的，而这种喜好将直接决定消费者的支付意愿，对产品的实际销售价格产生影响。有趣的是，在价格形成过程中，企业为实现新功能所耗费的成本不见得能起多大作用。因此，研究消费者对新功能的偏好就有着特别重要的意义。我们尤其感兴趣的是，当消费者面对不同功能取舍的时候（如有拍照功能但是没有触摸屏），哪些功能更为消费者看重。

我们的数据来源于问卷调查，被调查对象是北京大学光华管理学院的MBA学生和高级经理培训班学员。对于每一个被调查者，我们为其呈现几款具体的手机描述，然后要求被调查者对其偏好程度打分（5度李克特量表）。在实际调查中，有可能一个被调查者评价了若干款手机，因此会产生多条观测。从统计学上讲，来自同一个体的观测肯定是相关的，因此需要特殊的回归模型（如面板数据模型，panel data model）来处理。为了简便起

见，本案例假设所有样本都是独立的。严格地讲，这是一个错误的假设，相应产生的结论是有可能有偏（biased）的。

在为被调查者描述产品时，我们考虑了以下几个产品属性。首先是品牌，这里考虑了两个知名的国际品牌（诺基亚、摩托罗拉）、一个亚洲地区的知名品牌（三星），以及一个本土知名品牌（波导）。考虑的手机功能包括：拍照功能、收看电视功能、触摸屏、电话本多条记录、MP3功能和游戏数目。详细情况如表4—1所示。

表4—1　　变量描述

变量类型	变量含义	变量名	变量水平
因变量	对该产品的偏好程度	Y	1＝根本不喜欢；2＝比较不喜欢；3＝一般喜欢；4＝比较喜欢；5＝非常喜欢
自变量	手机品牌	X_1	诺基亚、摩托罗拉、三星和波导
	能否拍照	X_2	共两种（1＝能，0＝不能）
	能否收看电视	X_3	共两种（1＝能，0＝不能）
	有无触摸屏	X_4	共两种（1＝有，0＝无）
	电话本能否多条记录	X_5	共两种（1＝能，0＝不能）
	有无MP3	X_6	共两种（1＝有，0＝无）
	游戏数目	X_7	数值型

然后将这七个要素按不同方式组成12个产品组合，如表4—2所示。

表4—2　　产品设计方案

品牌	能否拍照	能否收看电视	触摸屏	电话本能否多条记录	MP3	游戏数目
诺基亚	不能	不能	无	能	有	3
	能	不能	有	不能	有	5
	不能	能	有	不能	无	7
波导	能	能	无	能	无	3
	不能	不能	无	不能	有	5
	能	不能	有	能	有	7
摩托罗拉	不能	能	有	能	无	3
	能	能	无	不能	无	5
	不能	不能	无	能	无	7
三星	能	不能	有	不能	无	3
	不能	能	有	不能	有	5
	能	能	无	能	有	7

对于表4—2中的每一款产品邀请多个被调查者互相独立地评估他们对

该产品的偏好程度。由于被调查人群是北京大学光华管理学院的 MBA 学生和高级经理培训班的学员，因此可以假设目标人群是高端商务人士。

值得一提的是，本案例的解释变量所涉及的数据类型比较丰富。其中 X_7 游戏数目是一个数值型的定量变量，而 $X_2 \sim X_6$ 都涉及某个功能的有无，因此全部都是取值 0—1 的哑变量。相对最复杂的是 X_1 手机品牌。显然，这是一个定性的因素，有 4 个不同的水平，即诺基亚、摩托罗拉、三星、波导。

4.3 描述分析

同前面几章一样，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 4 章.csv”中。简单展示如下：

	A	B	C	D	E	F	G	H
1	Y	X1	X2	X3	X4	X5	X6	X7
2	3	诺基亚	0	0	0	1	1	3
3	4	诺基亚	1	0	1	0	1	5
4	4	诺基亚	0	1	1	0	0	7
5	4	波导	1	1	0	1	0	3
6	3	波导	0	0	0	0	1	5
7	4	波导	1	0	1	1	1	7
8	5	摩托罗拉	0	1	1	1	0	3
9	4	摩托罗拉	1	1	0	0	0	5
10	3	摩托罗拉	0	0	0	1	0	7

其中比较值得注意的是第二列 Y，这是因变量，取值 1～5。然后 SAS 读入：

```
data A0;
    infile "D:\商务数据分析与应用\案例数据\第4章.csv"
    firstobs=2 delimiter=",";
    input Y X1$ X2-X7;
run;
```

原始数据“第 4 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	X5	X6	X7
1	3	诺基亚	0	0	0	1	1	3
2	4	诺基亚	1	0	1	0	1	5
3	4	诺基亚	0	1	1	0	0	7
4	4	波导	1	1	0	1	0	3
5	3	波导	0	0	0	0	1	5
6	4	波导	1	0	1	1	1	7
7	5	摩托罗拉	0	1	1	1	0	3
8	4	摩托罗拉	1	1	0	0	0	5
9	3	摩托罗拉	0	0	0	1	0	7
10	3	三星	1	0	1	0	0	3

本案例涉及的解释变量比较复杂，数据类型比较丰富，因此，需要根据相应的数据特点采用不同的描述分析方法。首先考虑第一个解释变量 X_1 手机品牌。该因素是一个定性变量，具有 4 个不同的水平，适合做列联表分析，可以通过下面的 SAS 程序实现。

```
proc freq data=A0; table Y*X1; run;
```

SAS 会自动对消费者偏好 Y 以及手机品牌 X_1 做列联表如下：

Y　X1

频数 百分比 行百分比 列百分比	波导	摩托罗拉	诺基亚	三星	合计
1	37 2.55 30.58 10.51	24 1.65 19.83 6.59	34 2.34 28.10 9.19	26 1.79 21.49 7.12	121 8.34
2	80 5.51 30.42 22.73	64 4.41 24.33 17.58	53 3.65 20.15 14.32	66 4.55 25.10 18.08	263 18.13
3	98 6.75 19.56 27.84	138 9.51 27.54 37.91	132 9.10 26.35 35.68	133 9.17 26.55 36.44	501 34.53
4	109 7.51 25.41 30.97	108 7.44 25.17 29.67	116 7.99 27.04 31.35	96 6.62 22.38 26.30	429 29.57
5	28 1.93 20.44 7.95	30 2.07 21.90 8.24	35 2.41 25.55 9.46	44 3.03 32.12 12.05	137 9.44
合计	352 24.26	364 25.09	370 25.50	365 25.16	1451 100.00

从最后一行第二列可以看出，样本中一共有352人测试了波导手机，占整个样本 $n=1\,451$ 的24.26%。然后从同一列的第二行可以看出，在352个波导样本中，有37个人给出了“1=根本不喜欢”的答案。这部分样本（37个）占整个波导样本（352个）的10.51%。该比率是4个不同品牌中最高的，明显地高于摩托罗拉以及三星。同样的规律也出现在“2=比较不喜欢”这一行中。根据最后一列的第三行，我们知道在整个样本中一共有263个样本给出了“2=比较不喜欢”这个答案，其中有80个来自波导，是所有品牌中最多的，占263个样本的30.42%。所有分析都似乎表明波导在被测样本中是一个比较弱势的品牌。这和我们的目标人群是高端商务人群有关，如果面对另外一类消费人群，结论可能完全不一样。

接下来，分析 $X_2 \sim X_6$。它们都是取值0—1的哑变量，用以标明某个功能的有无。对这部分变量，可以模仿第1章的办法，对其做描述分析如下：

```
proc sort data=A0; by Y; run;
proc univariate data=A0 noprint;
    by Y;
    var X2-X6;
    output out=A1 mean=m2-m6;
run;
data A1; set A1; format m2-m6 5.2; run;
```

具体的分析结果A1如下表所示：

	Y	X2的均值	X3的均值	X4的均值	X5的均值	X6的均值
1	1	0.43	0.44	0.47	0.21	0.54
2	2	0.38	0.43	0.43	0.39	0.50
3	3	0.48	0.49	0.54	0.42	0.45
4	4	0.58	0.53	0.55	0.62	0.52
5	5	0.58	0.64	0.38	0.85	0.60

如何解读上表？最简单的方式就是对每一个变量对比其首个和最后一个数据。例如，从第三列的第二行我们知道在回答 $Y=1$（根本不喜欢）的样本中，具备拍照功能的样本只占到43%；但是，同样的比例在回答 $Y=5$（非常喜欢）的样本中，占到58%。这说明更好的满意度（即 $Y=5$）和更高的拥有拍照功能比例之间有一定的相关性。同样的规律也出现在 X_3（能

否收看电视）、X_4（有无触摸屏）、X_5（电话本能否多条记录），还有 X_6（有无MP3）中。其中，尤其以 X_5（电话本能否多条记录）表现最为明显。其最后一个和首个数据之差高达85%－21%=64%。这非常清楚地告诉我们，高端商务人群十分看中电话本的多条记录功能，而该功能的实现是极其容易的。

最后对 X_7（游戏数目）做描述分析。由于该变量具有数值意义，因此均值、方差等传统描述统计量均适用。分析如下：

```
proc univariate data=A0 noprint;
    by Y;
    var X7;
    output out=A2 mean=mean std=std min=min median=median max=max;
run;
data A2; set A2; format mean 5.3 std 5.3; run;
```

	Y	X7的均值	X7的标准差	X7的最大值	X7的中位数	X7的最小值
1	1	5.066	1.340	7	5	3
2	2	5.061	1.512	7	5	3
3	3	4.904	1.582	7	5	3
4	4	5.037	1.724	7	5	3
5	5	5.000	1.940	7	5	3

从上表可以大概看出，不同的偏好分组（即不同的 Y 取值）所对应的游戏数目 X_7 的均值是非常类似的。这似乎提示我们，游戏多少对商务人士不重要。

4.4 统计模型

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 $X=(1, X_{11}, X_{12}, X_{13}, X_2, \cdots, X_7)'$，相应的回归系数为 $\beta=(\beta_0, \beta_{11}, \beta_{12}, \beta_{13}, \beta_2, \cdots, \beta_7)'$，其中 β_0 是截距项。再定义：

$$X'\beta=\beta_0+\beta_{11}X_{11}+\beta_{12}X_{12}+\beta_{13}X_{13}+\beta_2X_2+\cdots+\beta_7X_7$$

应该如何探讨因变量 Y 和 $X'\beta$ 的关系呢？同0—1逻辑回归一样，直接定义 $Y=X'\beta+\varepsilon$ 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性的指标。如何调和这个矛盾呢?

前辈统计学家想到了一个巧妙而又智慧的办法。他们分析到，因变量 Y 是消费者表达出来的明确产品偏好，其实，在消费者做出明确表达之前，内心深处思考酝酿着一个更加细致入微的产品偏好，该偏好没有被表达出来，是潜在的（latent）。该变量甚至是连续的，因为面对极其接近的产品的时候，人们对它们的偏好是很接近的，因此才会表现出难以取舍、左右为难。假设用 Z 来表达该潜在偏好，可以想象当 Z 的取值特别高时，明确表达的产品偏好 Y 就会高；相反，当 Z 的取值特别低时，明确表达的产品偏好 Y 就会低。数学上，可以假设：

$$Y=\begin{cases}1, Z<c_1\\ 2, c_1\leqslant Z<c_2\\ 3, c_2\leqslant Z<c_3\\ 4, c_3\leqslant Z<c_4\\ 5, c_4\leqslant Z\end{cases}$$

其中 $c_1\sim c_4$ 是 4 个未知的阈值，需要基于数据估计。第一次接触该潜在变量模型的读者可能会觉得有点不习惯。但想一下我们平时的考试成绩评定，就会觉得该模型非常合理。例如，Z 就是一个人的期末成绩，是连续的 0～100 分。如果 $Z<60$，那么定义 $Y=1$ 不及格；如果 $60\leqslant Z\leqslant 90$，定义 $Y=2$ 良好；最后如果 $Z>90$，定义 $Y=3$ 优秀。

言归正传，因为潜变量 Z 是一个连续变量，因此完全可以对其假设一个普通的线性回归模型，即 $Z=X'\beta+\varepsilon$，其中假设随机扰动项服从一个均值为 0 而方差为 1 的标准正态分布。细心的读者也许会问，为什么要假设方差为 1？假如 ε 的方差不是 1，而是 σ，那么我们可以重新定义一个新的残差 $\tilde{\varepsilon}=\varepsilon/\sigma$，新的回归系数 $\tilde{\beta}=\beta/\sigma$，还有一个新的潜变量 $\tilde{Z}=Z/\sigma$，前面提到的回归模型仍然成立，但是回归系数从 β 变成了 $\tilde{\beta}$，即 $\tilde{Z}=X'\tilde{\beta}+\tilde{\varepsilon}$。

因此，如果不限制 ε 的方差为 1，回归系数 β 是不可识别的（nonidentifiable）。因此，限制 ε 的方差为 1 是必须的。

假设随机扰动项服从一个均值为 0、方差为 1 的标准正态分布，可以计算出 Y 的各个取值的条件概率。以 $Y=2$ 为例，结果如下：

$$\begin{aligned}P(Y=2|X)&=P(c_1\leqslant Z<c_2)=P(c_1-X'\beta\leqslant\varepsilon<c_2-X'\beta)\\&=\Phi(c_2-X'\beta)-\Phi(c_1-X'\beta)\end{aligned}$$

依此类推，可以得到其他所有关于 Y 的条件概率，结果如下：

$$P(Y=K|X)=p_k(X'\beta)=\begin{cases}\Phi(c_1-X'\beta), k=1\\ \Phi(c_2-X'\beta)-\Phi(c_1-X'\beta), k=2\\ \Phi(c_3-X'\beta)-\Phi(c_2-X'\beta), k=3\\ \Phi(c_4-X'\beta)-\Phi(c_3-X'\beta), k=4\\ 1-\Phi(c_4-X'\beta), k=5\end{cases}$$

其中 $\Phi(t)$ 代表一个标准正态分布（standard normal）的分布函数（distribution function），这就是人们常说的关于定序数据的 PROBIT 回归模型。

同普通线性回归以及 0—1 变量逻辑（LOGIT）回归类似，对于定序 PROBIT 回归而言，人们关心回归系数 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概率 $p_k(X'\beta)$ 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 $\beta_j>0$，在给定其他解释变量不变的前提下，指标 X_j 的上升会带来条件概率 $p_k(X'\beta)$ 的下降（请注意 β 前面的负号）。也就是说，因变量 Y 取值为偏小（即 $Y\leqslant k$）的可能性会变小，这等价于说 Y 的取值更有可能变大。从某个角度看来，好像是一种“正”相关。当然，如果 $\beta_j<0$，在给定其他解释变量不变的前提下，指标 X_j 的上升会带来条件概率 $p_k(X'\beta)$ 的上升（请注意 β 前的负号）。也就是说，因变量 Y 取值为偏小（即 $Y\leqslant k$）的可能性会变大，这等价于说 Y 的取值更有可能变小，这好像是一种“负”相关。

定序 PROBIT 回归模型应该如何估计呢？从其推导过程来看，PROBIT 回归似乎也是一个线性回归，只不过“因变量”不是 Y，而是潜变量 Z。因此如果人们能够观测到 Z，那就可以用 Z 作为因变量做一个普通的最小二乘估计，所有的问题都能够迎刃而解。但问题的难处就在于 Z 是一个未知变量（这也是称其为潜变量的原因）。因此，这种天真的最小二乘想法无法实施。那么应该如何解决该问题？同 0—1 逻辑回归一样，我们将考虑极大似然准则。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。那么，它们的联合似然函数为：

$$\prod_{i=1}^{n}P(Y_i\mid X_i)=\prod_{i=1}^{n}\prod_{k=1}^{5}\{p_k(X'_i\beta,c)\}^{I(Y_i=k)}$$

其中 $c=(c_1, c_2, c_3, c_4)'$。对它做对数变换后，得到对数似然函数（log-

likelihood function）为：

$$\mathcal{L}(\beta)=\sum_{i=1}^{n}\log\{P(Y_i\mid X_i)\}=\sum_{i=1}^{n}\sum_{k=1}^{5}I(Y_i=k)\log\{p_k(X_i'\beta,c)\}$$

然后通过极大化该对数似然函数获得极大似然估计（maximum likelihood estimator），即 $(\hat{\beta},\ \hat{c})=\text{argmax}_{\beta}\ \mathcal{L}(\beta,\ c)$，其中 $\hat{c}=(\hat{c}_1,\ \hat{c}_2,\ \hat{c}_3,\ \hat{c}_4)'$。标准的统计学理论告诉我们，该估计量是渐进无偏的（asymptotically unbiased）、相合一致的（consistent），而且是极限正态的（asymptotically normal）。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 $H_0:\beta_j=0$，$H_1:\beta_j\neq0$。

同逻辑回归一样，对定序 PROBIT 回归而言，没有“残差”这个概念，因此无法定义残差平方和。但可以定义离差为 $DEV=-2\,\mathcal{L}(\hat{\beta},\ \hat{c})$，然后做全局检验 $H_0:\tilde{\beta}=0$，$H_1:\tilde{\beta}\neq0$，其中 $\tilde{\beta}=(\beta_1,\ \beta_2,\ \cdots,\ \beta_p)'$。具体地说，我们考虑两个不同的互为竞争关系的模型。其中模型 A 允许 $\tilde{\beta}\neq0$，模型 B 假设 $\tilde{\beta}=0$。把它们各自相应的离差记为：DEV_A 和 DEV_B。由于模型 A 比模型 B 更加灵活，因此其相应的似然函数一定更大，即一定有 $DEV_B>DEV_A$。如果原假设 $\tilde{\beta}=0$ 正确，DEV_B-DEV_A 应该不会特别大。到底会多大呢？标准的似然比检验（likelihood ratio test）理论告诉我们，该差异在样本量足够大的情况下，应该服从一个自由度为 p 的卡方分布。对于本案例而言，一共有 7 个因素。其中，除了第一个因素手机品牌以外，每个因素都只消耗一个自由度，总共 6 个。但是，手机品牌有 4 个水平，对应 3 个哑变量，共消耗 3 个自由度。因此，对本案例而言自由度为 $p=9$。我们可以近似地计算出模型全局检验的 p-值，并以此作为依据，做相应的统计推断。

和方差分析一样，我们还可以单独检验多水平因素“手机品牌”的显著性。为此，对比下面两个模型：

模型 a：$X'\beta=\beta_0+\beta_{11}X_{11}+\beta_{12}X_{12}+\beta_{13}X_{13}+\beta_2X_2+\cdots+\beta_7X_7$

模型 b：$X'\beta=\beta_0+\beta_2X_2+\cdots+\beta_7X_7$

把它们的离差分别记作 DEV_a 和 DEV_b。由于模型 a 比模型 b 更加灵活，其相应的似然函数一定更大，即一定有 $DEV_b>DEV_a$。如果原假设 $\tilde{\beta}=0$正确，DEV_b-DEV_a 应该不会特别大。到底会多大呢？标准的似然比检验理论告诉我们，该差异在样本量足够大的情况下，应该服从一个自由度

为 $df=3$ 的卡方分布。因此，可以近似地计算出模型全局检验的 p-值，并以此为依据，对手机品牌这个因素的显著性有所判断。

4.5　预测评估

在实际工作中，预测评估定序变量是一个相当复杂的问题。同前一章讲的 0—1 变量相比，定序变量情形更加复杂。首先考虑预测问题。请注意 $(Y_i,\ X_i)$（$i=1,\ \cdots,\ n$）代表历史数据，再假设 $(Y_i^*,\ X_i^*)$（$i=1,\ \cdots,\ m$）是未来数据。对于未来数据而言，解释变量 X_i^* 是已知的，但因变量 Y_i^* 是未知的。就本案例而言，Y_i 是某消费者的未知偏好，而 X_i 是某一款手机的各种属性。如何预测呢？首先通过分析历史数据建立 PROBIT 定序回归模型，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^* 各种取值的概率估计如下：

$$P(Y_i^*=k\mid X_i^*)=p_k(X_i^{*\prime}\hat{\beta},\ \hat{c})$$

$$=\begin{cases}\Phi(\hat{c}_1-X_i^{*\prime}\hat{\beta}),k=1\\ \Phi(\hat{c}_2-X_i^{*\prime}\hat{\beta})-\Phi(\hat{c}_1-X_i^{*\prime}\hat{\beta}),k=2\\ \Phi(\hat{c}_3-X_i^{*\prime}\hat{\beta})-\Phi(\hat{c}_2-X_i^{*\prime}\hat{\beta}),k=3\\ \Phi(\hat{c}_4-X_i^{*\prime}\hat{\beta})-\Phi(\hat{c}_3-X_i^{*\prime}\hat{\beta}),k=4\\ 1-\Phi(\hat{c}_4-X_i^{*\prime}\hat{\beta}),k=5\end{cases}$$

此概率量化了某消费者各种偏好的可能性。显然，如果某偏好（例如 k）的可能性越大，我们就越趋向于将 Y_i^* 预测为 $\hat{Y}_i^*=k$。所以，最简单的预测方法莫过于定义

$$\hat{Y}_i^*=\text{argmax}_{1\leqslant k\leqslant5}\,p_k(X_i^{*\prime}\hat{\beta},\ \hat{c})$$

简单地说，就是把消费者偏好预测为发生概率 $p_k(X_i^{*\prime}\hat{\beta},\ \hat{c})$ 最大的那个。在定序变量只有两个取值的情况下（如 1=不喜欢，2=喜欢），这等价于

$$\hat{Y}_i^*=\begin{cases}2,p_2(X_i^{*\prime}\hat{\beta},\ \hat{c})>0.5\\ 1,p_1(X_i^{*\prime}\hat{\beta},\ \hat{c})\leqslant0.5\end{cases}$$

这和前一章的0—1变量预测完全一样。现有的统计学理论已经证明，这样做的结果是极小化了错判概率 MCR$=m^{-1}\sum_{i=1}^{m} I(Y_i^* \neq \hat{Y}_i^*)$。

同0—1回归类似，以上结论非常简单而优美，而且非常有用。但是MCR隐含着一个假设，那就是不同类型的错误预测（如把1预测成2，2预测成3，或者3预测成4）所带来的损失都是一样的。这常常是一个合理的假设，如果在整个样本中各种偏好水平的消费者分布比较均匀可比。但是，如果该分布非常不均匀，情况就不同了，我们应该考虑加权的错判概率WMCR，应该对那些稀有的样本赋予更大的权重，而对另外的丰富的样本赋予较小的权重。对于定序数据，WMCR可以定义如下：

$$\text{WMCR}=\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{5}\frac{I(Y_i^* \neq \hat{Y}_i^* \mid Y_i^*=k)}{\pi_k}$$

以WMCR为标准，那么产生的预测规则为：

$$\hat{Y}_i^*=\text{argmax}_{1\leqslant k\leqslant 5}\left\{\frac{P_k(X_i^{*\prime}\hat{\beta},\hat{c})}{\pi_k}\right\}$$

细想一下，这个结论很符合人们的常识。如果没有任何解释变量，我们知道$Y_i^*=k$的先验概率为π_k。但是，在解释变量X_i^*的帮助下，如果发现新的概率$p_k(X_i^{*\prime}\hat{\beta},\hat{c})$大于先验$\pi_k$，可以认为该样本似乎更有可能取值$Y_i^*=k$，否则为其他。当有多个偏好水平都有$p_k(X_i^{*\prime}\hat{\beta},\hat{c})$大于$\pi_k$，谁的相对差异$p_k(X_i^{*\prime}\hat{\beta},\hat{c})/\pi_k$最大，谁就最有可能发生。

4.6 SAS编程

到此为止，逻辑回归的基本理论就介绍完了。我们回到本章的手机偏好案例，通过SAS程序具体分析如下：

```
proc genmod data=A0 DESCENDING;
    class Y X1;
    model Y =X1-X7 .
       /dist=multinomial link=cumprobit type3;
run;
```


SAS所产生的主要输出如下：

Class Level Information

Class	Levels	Values
Y	5	1 2 3 4 5
X1	4	波导 摩托罗拉 诺基亚 三星

从上可以看到，我们的数据主要涉及两个定性变量：一个是因变量Y（消费者偏好度），它有5个不同的取值水平，分别为1，2，3，4，5；另外一个定性变量是第一个解释变量X_1手机品牌，它有4个不同水平，分别为波导、摩托罗拉、诺基亚、三星。

Response Profile

Ordered Value	Y	Total Frequency
1	5	137
2	4	429
3	3	501
4	2	263
5	1	121

PROC GENMOD is modeling the probabilities of levels of Y having LOWER Ordered Values in the response profile table.

从上可以看到，样本中有137个回答$Y=5$（非常喜欢），429个回答$Y=4$（比较喜欢），501个回答$Y=3$（一般喜欢），263个回答$Y=2$（比较不喜欢），而121个回答$Y=1$（根本不喜欢）。尤其值得注意的是，从其最后一句话还可以特别注意到，SAS模拟的是定序因变量的低端概率，而不是高端概率。这能够保证SAS所估计的参数和前一节讨论的模型一致。因此，我们才能够正确解读相应的参数估计结果，否则，你会发现所有参数估计的正负号都和预期相反。

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
X1	3	33.25	<.0001
X2	1	42.91	<.0001
X3	1	25.40	<.0001
X4	1	17.07	<.0001
X5	1	203.45	<.0001
X6	1	7.14	0.0075
X7	1	0.61	0.4355

上表展示的是各个因素的基于离差的似然比检验。为了保持和第 2 章一致，我们希望这里汇报的是一张第 2 型方差分析，遗憾的是 SAS 的 PROC GENMOD 没有 Type Ⅱ这个选项，因此我们只好用第 3 型方差分析。由于本模型没有涉及交互作用，因此无论第 2 型还是第 3 型方差分析，结果应该完全一样。从上表可以看到，除了最后一个解释变量 X_7 游戏数目以外，每一个因素都是高度显著的。其中值得注意的是，第一个因素 X_1 手机品牌，它有 4 个不同水平，因此相应的自由度为 3。除此以外，其他变量的自由度都是 1。最后，去掉那个不显著的解释变量 X_7 游戏数目，将模型重新拟合如下：

```
proc genmod data=A0 DESCENDING;
    class Y X1;
    model Y =X1-X6
       /dist=multinomial link=cumprobit type3;
    output out=A3 p=p;
run;
```

所产生的极大似然估计量汇报如下：

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept1		1	-2.4645	0.1295	-2.7183	-2.2106	362.08	<.0001
Intercept2		1	-1.3082	0.1207	-1.5447	-1.0717	117.52	<.0001
Intercept3		1	-0.2987	0.1169	-0.5280	-0.0695	6.53	0.0106
Intercept4		1	0.5259	0.1198	0.2911	0.7607	19.27	<.0001
X1	波导	1	-0.2794	0.0861	-0.4480	-0.1107	10.54	0.0012
X1	摩托罗拉	1	-0.0084	0.1014	-0.2072	0.1903	0.01	0.9338
X1	诺基亚	1	0.2102	0.0831	0.0474	0.3730	6.41	0.0114
X1	三星	0	0.0000	0.0000	0.0000	0.0000	.	.
X2		1	0.3897	0.0598	0.2725	0.5069	42.48	<.0001
X3		1	0.3075	0.0616	0.1866	0.4283	24.87	<.0001
X4		1	0.2534	0.0617	0.1325	0.3744	16.86	<.0001
X5		1	0.9026	0.0636	0.7780	1.0273	201.52	<.0001
X6		1	0.1868	0.0731	0.0436	0.3300	6.54	0.0106

上表第二至五行汇报了 4 个不同的截距项（Intercept1～Intercept4），它们就是对 c_1～c_4 是 4 个未知阈值的极大似然估计。因此，它们之间有一个天然的从小到大的关系。我们一般不对此做任何解读。通过对 X_1 手机品牌结果的分析可以看到，自动选取三星作为基准，因为其极大似然估计精确为

0。在给定其他功能一样的前提下，摩托罗拉同三星之间没有表现出任何有统计意义的区别。但是，诺基亚却具有明显的品牌优势，其相应的极大似然估计为 0.21。最后，如同预期一样，波导具有明显的品牌劣势，其相应的极大似然估计为－0.46。除了手机品牌以外，还有其他几个功能，即 X_2（能否拍照），X_3（能否收看电视），X_4（有无触摸屏），X_5（电话本能否多条记录），以及 X_6（有无 MP3）。和预期一样，具备这些功能显然能够提高消费者的满意度。因此，它们相应的极大似然估计都是正的，而且高度显著。相比较而言，X_5（电话本能否多条记录）是最重要的，因为它所对应的极大似然估计最大（0.90），这再次说明电话本功能对商务人士的重要性。

下面我们尝试按照不同的标准对消费者偏好予以预测。同前一章一样，为了简便我们没有区分内外样本，这样会夸大实际中的预测精度。很遗憾的是，SAS 没有提供非常方便的预测选项，因此稍微有点麻烦。首先我们了解一下前面生成的数据 A3：

	Y	X1	X2	X3	X4	X5	X6	X7	Ordinal Level	Ordinal Level Value	Predicted Value
1	1	摩托罗拉	1	1	0	0	0	5	1	5	0.0378900436
2	1	摩托罗拉	1	1	0	0	0	5	2	4	0.2878046724
3	1	摩托罗拉	1	1	0	0	0	5	3	3	0.6517300724
4	1	摩托罗拉	1	1	0	0	0	5	4	2	0.8877472175
5	1	诺基亚	0	0	0	1	1	3	1	5	0.1220546238
6	1	诺基亚	0	0	0	1	1	3	2	4	0.496597353
7	1	诺基亚	0	0	0	1	1	3	3	3	0.8415702437
8	1	诺基亚	0	0	0	1	1	3	4	2	0.9660425631
9	1	诺基亚	0	1	1	0	0	7	1	5	0.0451952364
10	1	诺基亚	0	1	1	0	0	7	2	4	0.2956014218
11	1	诺基亚	0	1	1	0	0	7	3	3	0.6816878794
12	1	诺基亚	0	1	1	0	0	7	4	2	0.9026855639
13	1	波导	1	1	0	1	0	3	1	5	0.1263078685
14	1	波导	1	1	0	1	0	3	2	4	0.5048784337
15	1	波导	1	1	0	1	0	3	3	3	0.8465382612
16	1	波导	1	1	0	1	0	3	4	2	0.9675777914
17	1	波导	0	0	0	0	1	5	1	5	0.0052766885
18	1	波导	0	0	0	0	1	5	2	4	0.0806433775
19	1	波导	0	0	0	0	1	5	3	3	0.3477894652
20	1	波导	0	0	0	0	1	5	4	2	0.5676173836
21	1	波导	1	0	1	1	1	7	1	5	0.1559600685
22	1	波导	1	0	1	1	1	7	2	4	0.5576629502
23	1	波导	1	0	1	1	1	7	3	3	0.8758540816
24	1	波导	1	0	1	1	1	7	4	2	0.976100275
25	1	摩托罗拉	0	1	1	1	0	3	1	5	0.1564040057
26	1	摩托罗拉	0	1	1	1	0	3	2	4	0.5583946394
27	1	摩托罗拉	0	1	1	1	0	3	3	3	0.8762334252
28	1	摩托罗拉	0	1	1	1	0	3	4	2	0.9762044075

其中最右一列反映的是累计概率。然后，通过下面的程序获得 $p_k(X_i^{*\prime}\hat{\beta}, \hat{c})$：

```
data A3; set A3; ID=ceil(_n_/4); run;
proc sort data=A3; by ID Y; run;
proc transpose data=A3 out=A4;
    by ID Y;
    var p;
    id _level_;
run;
data A4;
    set A4; drop _name_ _label_;
    _1=1-_2; _2=_2-_3; _3=_3-_4; _4=_4-_5;
run;
```

结果存在 SAS 数据集合 A4 中，如下图所示：

	ID	Y	_5	_4	_3	_2	_1
1	1	1	0.038	0.230	0.384	0.236	0.112
2	2	1	0.122	0.375	0.345	0.124	0.034
3	3	1	0.045	0.250	0.386	0.221	0.097
4	4	1	0.126	0.379	0.342	0.121	0.032
5	5	1	0.005	0.075	0.267	0.320	0.332
6	6	1	0.156	0.402	0.318	0.100	0.024
7	7	1	0.156	0.402	0.318	0.100	0.024
8	8	1	0.038	0.230	0.384	0.236	0.112
9	9	1	0.058	0.281	0.385	0.198	0.078
10	10	1	0.034	0.219	0.382	0.244	0.121

其中 ID 列是我们人为生成的样本编号，第三列是真实的消费者偏好，而剩下几列是产生各种偏好的可能性大小。

```
data A5;
    set A4;
    Y1=1; probmax=_1;
    if _2> probmax then do; Y1=2; probmax=_2; end;
    if _3> probmax then do; Y1=3; probmax=_3; end;
    if _4> probmax then do; Y1=4; probmax=_4; end;
    if _5> probmax then do; Y1=5; probmax=_5; end;
run;
```

首先考虑极小化 MCR。因此，我们要在_1 到_5 的各个列中，寻找概率最大的一列。相应的 SAS 程序如上所示。接下来，对预测精度分析如下：

```
proc freq data=A5; table Y*Y1; run;
```

相应的 SAS 输出是：

Y　　Y1

频数 百分比 行百分比 列百分比	1	3	4	合计
1	28 1.93 23.14 23.73	72 4.96 59.50 9.84	21 1.45 17.36 3.49	121 8.34
2	49 3.38 18.63 41.53	145 9.99 55.13 19.81	69 4.76 26.24 11.48	263 18.13
3	34 2.34 6.79 28.81	303 20.88 60.48 41.39	164 11.30 32.73 27.29	501 34.53
4	7 0.48 1.63 5.93	188 12.96 43.82 25.68	234 16.13 54.55 38.94	429 29.57
5	0 0.00 0.00 0.00	24 1.65 17.52 3.28	113 7.79 82.48 18.80	137 9.44
合计	118 8.13	732 50.45	601 41.42	1451 100.00

由此可见完全准确预测的样本总数为：28＋303＋234＝565。相应的 MCR＝（1 451－565）/1 451＝61.1%。

```
data A5;
    set A4;
    _1=_1/8.34;_2=_2/18.13;_3=_3/34.53;_4=_4/29.57;_5=_5/9.44;
    Y1=1; probmax=_1;
    if _2> probmax then do; Y1=2; probmax=_2; end;
    if _3> probmax then do; Y1=3; probmax=_3; end;
    if _4> probmax then do; Y1=4; probmax=_4; end;
    if _5> probmax then do; Y1=5; probmax=_5; end;
run;
```

```
proc freq data=A5; table Y*Y1; run;
```

接下来考虑极小化 WMCR。根据上表结果，我们知道各个偏好的先验概率分别为 8.34%（k＝1），18.13%（k＝2），34.53%（k＝3），29.57%（k＝4），以及 9.44%（k＝5）。因此，相应的 SAS 程序如上所示。分析结果如下：

Y	Y1				
频数 百分比 行百分比 列百分比	1	2	3	5	合计
1	57 3.93 47.11 15.79	27 1.86 22.31 11.02	16 1.10 13.22 6.56	21 1.45 17.36 3.49	121 8.34
2	95 6.55 36.12 26.32	53 3.65 20.15 21.63	46 3.17 17.49 18.85	69 4.76 26.24 11.48	263 18.13
3	144 9.92 28.74 39.89	99 6.82 19.76 40.41	94 6.48 18.76 38.52	164 11.30 32.73 27.29	501 34.53
4	57 3.93 13.29 15.79	58 4.00 13.52 23.67	80 5.51 18.65 32.79	234 16.13 54.55 38.94	429 29.57
5	8 0.55 5.84 2.22	8 0.55 5.84 3.27	8 0.55 5.84 3.28	113 7.79 82.48 18.80	137 9.44
合计	361 24.88	245 16.88	244 16.82	601 41.42	1451 100.00

将该结果同 MCR 的结果对比发现，WMCR 的预测结果更加丰富。除了 4 以外，从 1 到 5 都有取值，而 MCR 的预测结果中没有 2 和 5。WMCR 正确判断的样本总数是：57＋53＋94＋113＝317，相应的错判概率是：(1 451－317)/1 451＝78.2%，明显高于 MCR 的 61.1%。对本案例而言，到底是 MCR 好，还是 WMCR 好，不是很清晰。如果我们能够大概测算出不同判断错误所带来的损失，那么也许会有一个结论。

4.7 总结讨论

本章通过一个市场调研案例，对 PROBIT 定序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，PROBIT 并不是该问题的关键。如果采用和前一章逻辑回归一样的链接函数（link function），所获得的就是 LOGIT 定序回归模型。有兴趣的读者可以查阅相关统计学教材以及 SAS 文档。

附录 4A　分析报告

手机商务消费者偏好度研究

1. 研究目的

通过分析商务消费者的问卷数据，了解商务人士对手机不同功能的偏好度，为高端商务手机的准确市场定位提供帮助。

2. 背景介绍

手机市场竞争非常激烈，为了确立自己在市场中的相对优势地位，开发新功能是手机制造商常用的手段。因为与开发新产品相比，在现有产品的基础上增加新功能所需要的投入少，且承担的风险较小，因此不失为一种既快速又有效的方式。当企业决定为现有产品增加新功能时，往往面对众多选择，例如，下一个升级换代产品是增加拍照功能还是增加收音机或者 MP3 播放器功能？对这个问题的正确回答依赖于目标客户。不同的目标客户对不同新功能的评价喜好是不一样的，而这种喜好将直接决定消费者的支付意愿，对产品的实际销售价格产生影响。因此，研究消费者对新功能的偏好就有着特别重要的意义。我们尤其感兴趣的是，当消费者面对不同功能取舍的时候（如有拍照功能但是没有触摸屏），哪些功能更为消费者看重。

我们的数据来源于问卷调查，被调查对象是北京大学光华管理学院的 MBA 学生和高级经理培训班学员。对于每一个被调查者，我们为其呈现几款具体的手机描述，然后要求被调查者对其偏好程度打分（5 度李克特量表）。在实际调查中，有可能一个被调查者评价了若干款手机，因此会产生多条观测。

3. 指标设计

在为被调查者描述产品的时候，我们考虑了以下几个产品属性。首先是品牌，这里考虑了两个知名的国际品牌（诺基亚、摩托罗拉），一个亚洲地区的知名品牌（三星），以及一个本土知名品牌（波导）。考虑的手机功能包

括：拍照功能、收看电视功能、触摸屏、电话本多条记录、MP3 功能和游戏数目。详细情况如表 4—3 所示。将这七个要素按不同方式组成 12 个产品组合，如表 4—4 所示。

表 4—3　　变量描述

变量类型	变量含义	变量名	变量水平
因变量	对该产品的偏好程度	Y	1=根本不喜欢；2=比较不喜欢；3=一般喜欢；4=比较喜欢；5=非常喜欢
自变量	手机品牌	X_1	诺基亚、摩托罗拉、三星和波导
	能否拍照	X_2	共两种（1=能，0=不能）
	能否收看电视	X_3	共两种（1=能，0=不能）
	有无触摸屏	X_4	共两种（1=有，0=无）
	电话本能否多条记录	X_5	共两种（1=能，0=不能）
	有无 MP3	X_6	共两种（1=有，0=无）
	游戏数目	X_7	数值型

表 4—4　　产品设计方案

品牌	能否拍照	能否收看电视	有无触摸屏	电话本能否多条记录	MP3	游戏数目
诺基亚	不能	不能	无	能	有	3
	能	不能	有	不能	有	5
	不能	能	有	不能	无	7
波导	能	能	无	能	无	3
	不能	不能	无	不能	有	5
	能	不能	有	能	有	7
摩托罗拉	不能	能	有	能	无	3
	能	能	无	不能	无	5
	不能	不能	无	能	无	7
三星	能	不能	有	不能	无	3
	不能	能	有	不能	有	5
	能	能	无	能	有	7

对于表 4—4 中的每一款产品邀请多个被调查者互相独立地评估他们对该产品的偏好程度。由于被调查人群是北京大学光华管理学院的 MBA 学生和高级经理培训班学员，他们能够代表高端商务人士。

4. 描述分析

本研究涉及的解释变量比较复杂，数据类型比较丰富，因此，需要根据

相应的数据特点采用不同的描述分析方法。

首先考虑第一个解释变量 X_1 手机品牌。该因素是一个定性变量，具有 4 个不同的水平，适合做列联表分析，如表 4—5 所示。

表 4—5　　消费者偏好同手机品牌的关系

消费者偏好	波导	摩托罗拉	诺基亚	三星	合计
1	37	24	34	26	121
2	80	64	53	66	263
3	98	138	132	133	501
4	109	108	116	96	429
5	28	30	35	44	137
合计	352	364	370	365	1 451

从表 4—5 最后一行第二列可以看出，在样本中一共有 352 人测试了波导手机，占整个样本 1 451 的 24.26%。然后从同一列的第二行可以看出，在 352 个波导样本中，有 37 个人给出了“1=根本不喜欢”的答案。这部分样本（37 个）占整个波导样本（352 个）的 10.51%。该比率是 4 个不同品牌中最高的，明显地高于摩托罗拉以及三星。同样的规律也出现在第三行中。根据第三行最后一个数据，我们知道在整个样本中一共有 263 个样本给出了“2=比较不喜欢”这个答案。其中有 80 个来自波导，是所有品牌中最多的，占 263 个样本的 30.42%。所有分析都似乎表明波导在我们的被测样本中是一个比较弱势的品牌。这和我们的目标人群是高端商务人群有关，如果面对另外一类消费人群，结论可能完全不一样。

接下来，我们分析五个功能变量，它们分别是能否拍照、能否收看电视、有无触摸屏、电话本能否多条记录，以及有无 MP3。结果如表 4—6 所示。

表 4—6　　各个手机功能对消费者偏好的影响

消费者偏好	能否拍照	能否收看电视	有无触摸屏	电话本能否多条记录	有无 MP3
1	0.43	0.44	0.47	0.21	0.54
2	0.38	0.43	0.43	0.39	0.50
3	0.48	0.49	0.54	0.42	0.45
4	0.58	0.53	0.55	0.62	0.52
5	0.58	0.64	0.38	0.85	0.60

在表 4—6 中，对每一个变量对比其首个和最后一个数据。从第三列的第二行我们知道在回答 $Y=1$（根本不喜欢）的样本中，具备拍照功能的样

本只占到43%。但是，同样的比例在回答$Y=5$（非常喜欢）的样本中，占到58%。这说明更好的满意度（即$Y=5$）和更高的拥有拍照功能比例之间有一定的相关性。同样的规律也出现在X_3（能否收看电视），X_4（有无触摸屏），X_5（电话本能否多条记录），还有X_6（有无MP3）中。其中，尤其以X_5（电话本能否多条记录）表现最为明显。其最后一个和首个数据之差高达85%−21%=64%。这非常清楚地告诉我们，高端商务人群十分看中电话本的多条记录功能，而该功能的实现是极其容易的。

最后对X_7（游戏数目）做描述分析。由于该变量具有数值意义，因此均值、方差等传统描述统计量均适用。具体结果如表4—7所示。从中可以大概看出，不同的偏好分组（即不同的Y取值）所对应的游戏数目X_7的均值是非常类似的。这似乎提示我们，游戏多少对商务人士不重要。

表4—7　　游戏数目对消费者偏好的影响

消费者偏好	均值	标准差	最大值	中位数	最小值
1	5.066	1.340	7	5	3
2	5.061	1.512	7	5	3
3	4.904	1.582	7	5	3
4	5.037	1.724	7	5	3
5	5.000	1.940	7	5	3

5. 模型分析

在描述分析的基础上，我们通过PROBIT定序回归模型对各个因素同消费者偏好之间的关系做了模型分析。各个因素的卡方检验如表4—8所示。

表4—8　　各个因素卡方检验结果

变量名称	自由度	统计量	p-值
手机品牌	3	33.25	<0.000 1
能否拍照	1	42.91	<0.000 1
能否收看电视	1	25.40	<0.000 1
有无触摸屏	1	17.07	<0.000 1
电话本能否多条记录	1	203.45	<0.000 1
有无MP3	1	7.14	0.007 5
游戏数目	1	0.61	0.435 5

根据表4—8结果可以看到，除了游戏数目之外，所有因素都在5%的水平下高度显著。因此，我们剔除游戏数目这个因素后重新拟合模型，获得极大似然估计如表4—9所示。

表4—9　　各个因素卡方检验结果

变量名称	水平	参数估计	卡方统计量	p-值
手机品牌	波导	−0.279 4	10.54	0.001 2
	摩托罗拉	−0.008 4	0.01	0.933 8
	诺基亚	0.210 2	6.41	0.011 4
	三星	0.000 0	NA	NA
能否拍照		0.389 7	42.48	<0.000 1
能否收看电视		0.307 5	24.87	<0.000 1
有无触摸屏		0.253 4	16.85	<0.000 1
电话本能否多条记录		0.902 6	201.52	<0.000 1
有无MP3		0.186 8	6.54	0.010 6

通过对X_1手机品牌结果的分析可以看到，自动选取三星作为基准，因为其极大似然估计精确为0。在给定其他功能一样的前提下，摩托罗拉同三星之间没有表现出任何有统计意义的区别。但是，诺基亚却具有明显的品牌优势，其相应的极大似然估计为0.21。最后，如同预期一样，波导具有明显的品牌劣势，其相应的极大似然估计为−0.28。除了手机品牌以外，还有其他几个功能，即X_2（能否拍照），X_3（能否收看电视），X_4（有无触摸屏），X_5（电话本能否多条记录），以及X_6（有无MP3）。和预期一样，具备这些功能显然能够提高消费者的满意度。因此，它们相应的极大似然估计都是正的，而且高度显著。相比较而言，X_5（电话本能否多条记录）是最重要的，因为它所对应的极大似然估计最大（0.90），这再次说明电话本功能对商务人士的重要性。

6. 预测评估

接下来，我们考虑用建立的PROBIT定序回归模型做预测。首先考虑最优化整体错判概率MCR。实际分析结果如表4—10所示，相应的整体错判概率为MCR=61.1%。

表 4—10　　极小化 MCR 的预测结果

真实 \ 预测	1	3	4	合计
1	28	72	21	121
2	49	145	69	263
3	34	303	164	501
4	7	188	234	429
5	0	24	113	137
合计	118	732	601	1 451

我们也考虑了极小化加权错判概率 WMCR。实际分析结果如表 4—11 所示，相应的整体错判概率为 78.2%。

表 4—11　　极小化 WMCR 的预测结果

真实 \ 预测	1	2	3	5	合计
1	57	27	16	21	121
2	95	53	46	69	263
3	144	99	94	164	501
4	57	58	80	234	429
5	8	8	8	113	137
合计	361	245	244	601	1 451

对本研究而言，对于不同的错判情形，我们无法给出一个合理的错判损失函数。因此，无法给出一个一般化的建议，到底是表 4—10 中的 MCR 结果好，还是表 4—11 中的 WMCR 结果好。所以，我们都予以汇报，仅供参考。

7. 总结讨论

本研究分析了 1 451 个关于手机产品的消费者偏好调研数据，分析了手机品牌以及 5 种不同的手机功能因素。从中发现手机品牌以及手机能否支持强大的电话本功能最为商务人士看重。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 4B　课后习题

信用卡用户信用评级

1. 研究目的

通过对某银行信用卡用户的历史数据分析，了解信用卡用户逾期不还的行为特征，为将来新客户办卡的信用评级提供参考。

2. 数据介绍

某年度随机抽取的 8 372 个信用卡用户。因变量是该用户最近的逾期状态，有 8 个不同取值，分别为：0＝没有逾期，1＝逾期 1～30 天，2＝逾期 31～60 天，3＝逾期 61～90 天，4＝逾期 91～120 天，5＝逾期 121～150 天，6＝逾期 151～180 天，7＝逾期 180 天以上。解释变量有：信用卡使用率，该指标的具体定义没有给出，简单地说，该指标越高，说明相应的信用卡使用的频率越高；信用卡额度；住房贷款月供（有可能是 0）；最近一次逾期之前的历史逾期赖账次数；同一个客户开户的次数。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 4.csv”中。

3. 作业要求

- 问题理解：请参阅相关信用卡评级资料，了解信用评级对信用卡业务的重要性。
- 做完整的 PROBIT 定序回归分析，包括参数估计、假设检验，以及预测评估。
- 将分析结果汇总成如附录 4A 所示的简短研究报告。

附录 4C　R 程序演示

首先通过下面的 R 程序读入数据：

```
> a=read.csv("D:/商务数据分析与应用/案例数据/第4章.csv",header=T)
> a[c(1:5),]
  Y     X1 X2 X3 X4 X5 X6 X7
1 3 诺基亚  0  0  0  1  1  3
2 4 诺基亚  1  0  1  0  1  5
3 4 诺基亚  0  1  1  0  0  7
4 4   波导  1  1  0  1  0  3
5 3   波导  0  0  0  0  1  5
```

描述品牌和偏好度的关系：

```
> table(a$Y,a$X1)

   波导 摩托罗拉 诺基亚 三星
1    37       24     34   26
2    80       64     53   66
3    98      138    132  133
4   109      108    116   96
5    28       30     35   44
```

描述分析其他功能性指标：

```
> MU2=tapply(a$X2,a$Y,mean)
> MU3=tapply(a$X3,a$Y,mean)
> MU4=tapply(a$X4,a$Y,mean)
> MU5=tapply(a$X5,a$Y,mean)
> MU6=tapply(a$X6,a$Y,mean)
> result=cbind(MU2,MU3,MU4,MU5,MU6)
> result
        MU2       MU3       MU4       MU5       MU6
1 0.4297521 0.4380165 0.4710744 0.2148760 0.5371901
2 0.3840304 0.4334601 0.4296578 0.3878327 0.5019011
3 0.4830339 0.4890220 0.5369261 0.4231537 0.4530938
4 0.5780886 0.5314685 0.5547786 0.6153846 0.5151515
5 0.5766423 0.6350365 0.3795620 0.8540146 0.5985401
>
```


描述分析游戏的数目：

```
> N=tapply(a$X7,a$Y,length)
> MU=tapply(a$X7,a$Y,mean)
> SD=tapply(a$X7,a$Y,sd)
> MIN=tapply(a$X7,a$Y,min)
> MED=tapply(a$X7,a$Y,median)
> MAX=tapply(a$X7,a$Y,max)
> result=cbind(N,MU,SD,MIN,MED,MAX)
> result
   N        MU       SD MIN MED MAX
1 121 5.066116 1.339997   3   5   7
2 263 5.060837 1.512072   3   5   7
3 501 4.904192 1.582025   3   5   7
4 429 5.037296 1.723534   3   5   7
5 137 5.000000 1.940285   3   5   7
```

计算因变量频数：

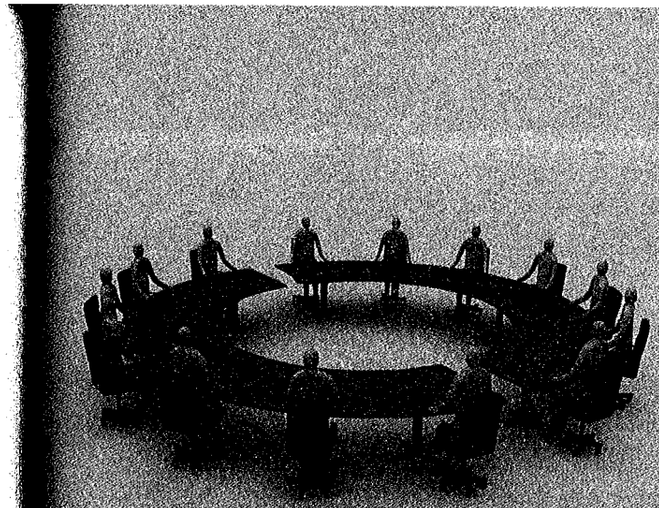
```
> table(a$Y)

  1   2   3   4   5
121 263 501 429 137
```

极大似然估计：

```
> library(MASS)
> probit1=polr(as.factor(Y)~as.factor(X1)+X2+X3+X4+X5+X6+X7,method="probit",Hess=T,data=a)
> summary(probit1)
Call:
polr(formula = as.factor(Y) ~ as.factor(X1) + X2 + X3 + X4 +
    X5 + X6 + X7, data = a, Hess = T, method = "probit")

Coefficients:
                              Value Std. Error t value
as.factor(X1)摩托罗拉      0.28048    0.09608  2.9193
as.factor(X1)诺基亚        0.48908    0.08529  5.7342
as.factor(X1)三星          0.27652    0.08613  3.2106
X2                         0.39145    0.05984  6.5419
X3                         0.31159    0.06188  5.0357
X4                         0.25499    0.06176  4.1291
X5                         0.90094    0.06362 14.1606
X6                         0.20205    0.07563  2.6714
X7                        -0.01373    0.01761 -0.7799

Intercepts:
    Value   Std. Error t value
1|2 -0.3036  0.1313    -2.3125
2|3  0.5214  0.1295     4.0247
3|4  1.5312  0.1331    11.5057
4|5  2.6876  0.1418    18.9589

Residual Deviance: 3984.573
AIC: 4010.573
```

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