“拍照赚钱”的任务定价

摘要

在当今信息时代的大背景下，网络成为人们生产生活必不可少的一部分，而网络共享更是成为当下一种最流行的生活方式，“拍照赚钱”就是利用了互联网实时共享的特点，通过用户下载 APP，注册成为 APP 会员领取任务并且完成而获得相应酬金的过程。本文主要研究了在类型 APP 的大数据中对任务点的地理位置、任务点周边会员数量，会员信誉度，及任务定价之间的函数关系。在合理假设的基础上利用 MATLAB、Excel 等数学工具对附件中的大数据进行处理，针对问题一和问题二选用了不同的回归模型进行解答。

针对问题一：我们分析了各种可能影响任务定价的因素，对任务点周边平均会员信誉、任务点周边平均任务开始时间、任务点周边平均预定任务限额、任务点周边会员数量、任务价格、人均（年）GDP 进行系统分析，通过其拟合图形的趋势来大体预估出可能影响任务定价的因素，将数据标准化后进行总体拟合和优化拟合，通过得出的拟合函数初步得出定价规律。应用回归模型引用 sigmoid 函数决定该模型的决策边界，并且计算相应参数，用控制变量法提取数据对计算结果进行检验，保证模型的准确性，通过所得出的参数合理分析在不同因素影响下任务未完成的原因。

针对问题二：以问题一的多元回归数学模型为理论依据主要讨论会员数量、会员信誉、任务价格对任务完成率的具体影响。通过算法得出任务完成率的分析表，比较表中的完成率大小情况来对问题一的回归模型系数进行修正，通过修正后的模型所得出数据与未修正模型数据的前后对比发现完成率有明显提高，验证了模型的合理性和可行性。

针对问题三：引用聚类模型，对所有任务点按照相近距离和相近价格进行区域聚类，考虑到不同地区密集程度不同并具有一定差异的实际情况，故限定每个聚类中的点数，据此形成类包，对每个类包内的数据处理后进行拟合，作为改变定价模型的依据。

针对问题四：对附件三的数据进行预处理。通过问题一二三的数据结论比较，得到模型二与模型三并无明显差异，但对比原方案，效果有显著提升。因此，我们选择采用模型二，并且利用问题二中的定价方案解决问题四。

关键词：多元回归模型 逻辑回归模型 聚类模型分析 sigmoid 函数 决策边界

一、问题的重述

1.1 问题背景

“拍照赚钱”是移动互联网下的一种自助式服务模式。用户下载 APP，注册成为 APP 的会员，然后从 APP 上领取需要拍照的任务赚取 APP 对任务所标定的酬金。这种基于移动互联网的自助式劳务众包平台，为企业提供各种商业检查和信息搜集，相比传统的市场调查方式可以大大节省调查成本，而且有效地保证了调查数据真实性，缩短了调查的周期。因此 APP 成为该平台运行的核心，而 APP 中的任务定价又是其核心要素。如果定价不合理，有的任务就会无人问津，而导致商品检查的失败。

1.2 问题提出

1、研究附件一中项目的任务定价规律，分析任务未完成的原因。

2、为附件一中的项目设计新的任务定价方案，并和原方案进行比较。

3、实际情况下，多个任务可能因为位置比较集中，导致用户会争相选择，一种考虑是将这些任务联合在一起打包发布。在这种考虑下，如何修改前面的定价模型，对最终的任务完成情况又有什么影响？

4、对附件三中的新项目给出你的任务定价方案，并评价该方案的实施效果。

二、基本假设

（1）不考虑任务之间难易程度的差异。

（2）忽略天气、地理位置等客观条件对任务造成的额外差异。

（3）样本数据真实，能够反映具体情况。

三、符号约定

符号	符号说明
x_1	会员数量
x_2	平均会员信誉
x_3	任务开始时间
x_4	类包中包涵的任务点
a_0	常数
a_n（n=1、2……n）	参数
R^2	拟合度
F、p、S^2	检验值
y	定价
∂	参照系数
x_i	任务数量
$\overline{x}_i$	任务数量平均值
$\hat{x}_i$	标准化任务数量

四、问题分析

4.1 问题一的分析与模型建立

为分析 APP 任务定价的合理性，作者主要选取广州、佛山、东莞、深圳四个地区为集中采样点，根据附件一给出的关于任务位置、定价和完成情况的大数据分析，可以看出任务的完成情况与任务所在的地区，该地区的经济发展情况（以该地区的人均 GDP 为指标）及任务点周边的会员数量及会员信誉度、会员限额有着直接的联系。

提取附件二中的信息数据，会员的位置、信誉值、配额等影响因素联系附件一中每个任务的位置、定价和完成情况计算出“拍照赚钱”在不同地区供求关系的不同，从而进一步讨论上述条件对定价的影响情况。提取数据中任务所在的点，以点为讨论对象，拟合出在一定半径的区域中，会员数量与定价的相关关系。以“广东省统计局”（官方）中的四个采样点城市的年人均 GDP 值为讨论依据，拟合出人均 GDP 对该地区定价的相关关系。最终通过“回归模型”计算得出相应的影响因素与定价的最终关系。通过影响因素与定价的关系，对影响任务完成情况的几个因素进行分析计算，讨论出正负相关性，再通过数据论证，最终得出影响任务完成的几个因素。

4.1.1 影响定价因素结构图

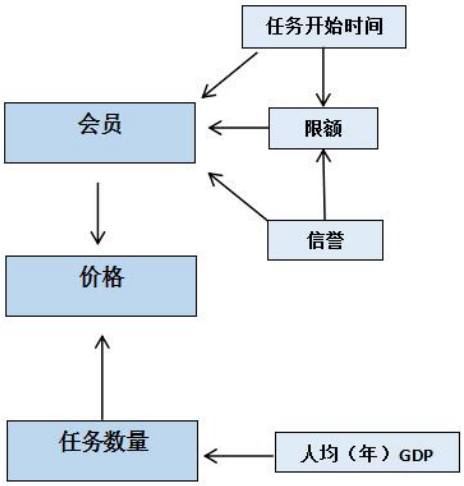


图 1：影响定价因素结构图

4.1.2 任务价格与任务点周边平均会员信誉、任务开始时间、预定任务限额的相关关系

利于 matlab，绘制成如下图表。matlab polyfit 编辑 polyfit 函数是 matlab 中用于进行曲线拟合的一个函数。其数学基础是最小二乘法曲线拟合原理。曲线拟合：已知离散点上的数据集，即已知在点集上的函数值，构造一个解析函数[1]（其图形为一曲线）使在原离散点上尽可能接近给定的值。（下文拟合方法相同）

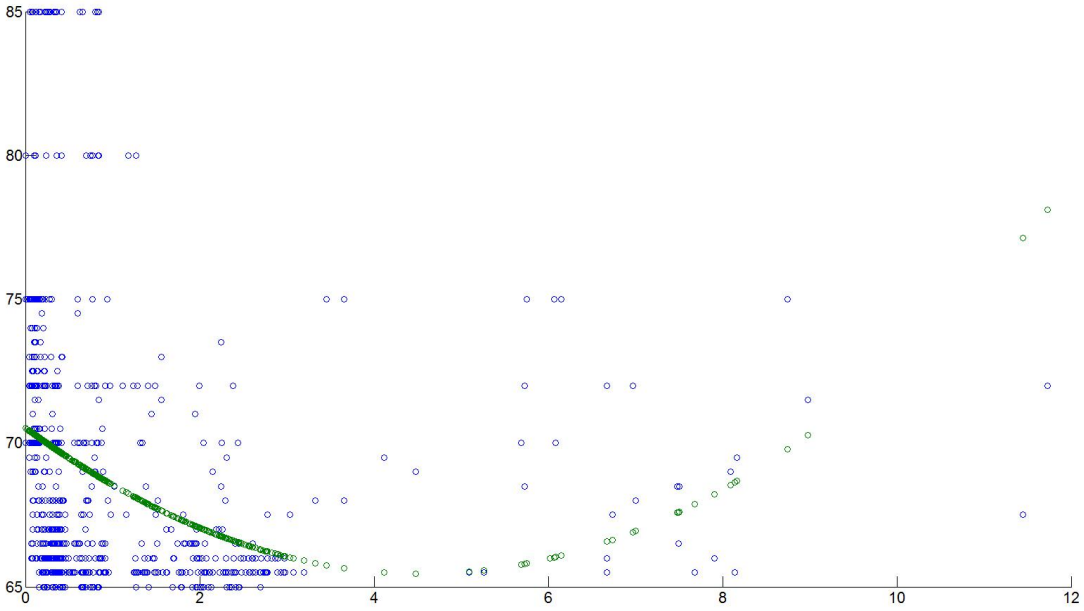


图 2：任务价格与任务点周边平均会员信誉标准化拟合图

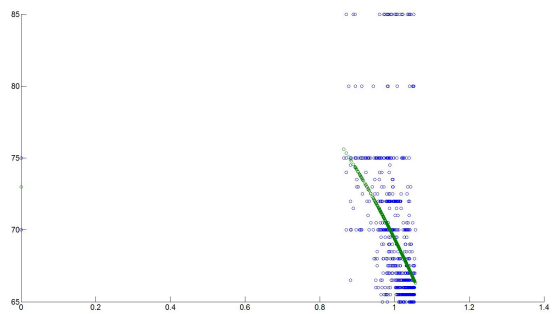


图 3：任务价格与任务点周边平均任务开始时间标准化拟合图

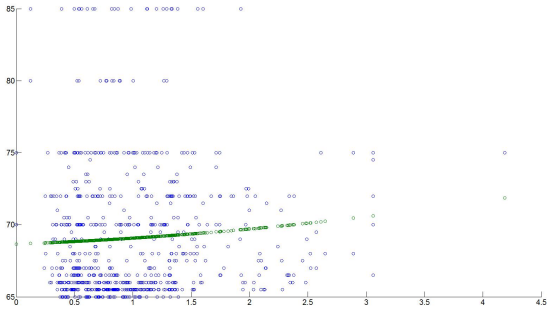


图 4：任务价格与任务点周边平均预定任务限额标准化拟合图

由图 1、图 2、图 4 三张图的拟合曲线趋势来判断，任务价格与任务点周边平均会员信誉，与周边任务平均开始时间有比较明显的关系，图二中，剔除 4~12 中的少量散点，从图中的标准化曲线趋势中可以看出，在 0~4 的范围内，任务价格与周边平均会员信誉度呈现反比例趋势，图三中，可以看出任务价格与周边任务开始时间呈反比例关系，而图四的拟合程度较低，可以看做任务价格与任务点周边平均预定任务限额没有明显关系。

4.1.3 任务价格与任务点周边会员数量相关关系分析

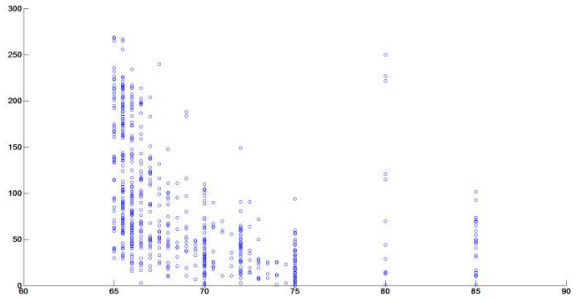


图 5：任务价格与任务周边会员数量关系图

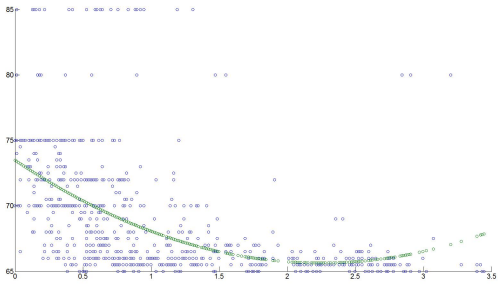


图 6：任务价格与任务周边会员数量标准化拟合图

图 5 中，散点图的横坐标代表任务标价，纵坐标代表任务周边会员数量。由散点图分布情况，可以看出任务比较集中的价格同样大致位于 65~75 的价格区间内，并且整体上会员数量越多时所对应的任务价格呈现递减趋势。

图 6 中，标准化任务数量 x_i 使 $\hat{x}_i = x_i - \bar{x}_i (i = 1,2,3.....n)$（以下标准化方法相同）将标准化后的任务量作为散点图的横坐标，将任务价格作为纵坐标，用 matlab 拟合出任务数量与任务价格的标准化关系图，剔除 2~3.5 后的少量散点，从图中的标准化曲线趋势可以明显看出，在 0~2 的范围内，任务价格与任务周边会员数量呈反比例变化趋势，会员数量越多，该地区的任务价格就越低。

4.1.4 任务价格与任务数量相关关系分析

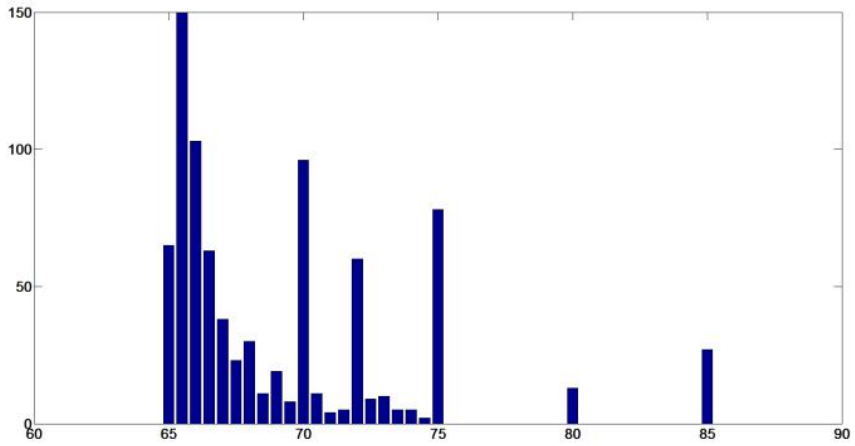


图 7：任务数量与任务价格分布直方图

图七中，直方图的横坐标代表任务标价，纵坐标代表任务数量。由直方图分布情况，可以看出任务比较集中的价格位于 65~75 的价格区间内，而低于 65 与高于 75 的任务则只占少数，可以看出当任务价格在合理范围内时，任务数量相对最多。

4.1.5 任务所在地区（年）GDP 用与任务数量相关关系分析

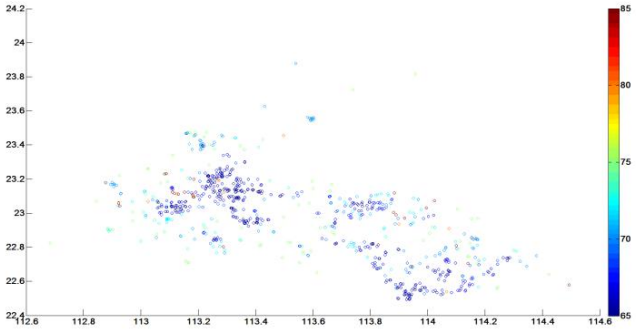


图 8：任务所在地区人均（年）GDP 与任务数量关系图

图 9：取样城市任务点卫星观测图

图 8 中，卫星图的横轴代表任务点的经度，纵轴（左）代表任务点的纬度，纵轴（右）代表任务价格的分布，用颜色加以区分。由卫星图的颜色分布情况，可以看出任务比较集中的价格同样大致位于 65~75 的价格区间内。从图九可以看出任务比较集中的区域位于四个取样城市的中心城区（一般认为中心城区的人均（年）GDP 相对其他地区较高），所以当任务所在地区的人均（年）GDP 较高时，任务的数量相对较多。

4.1.6 任务所在地区（年）GDP 与任务价格相关关系分析

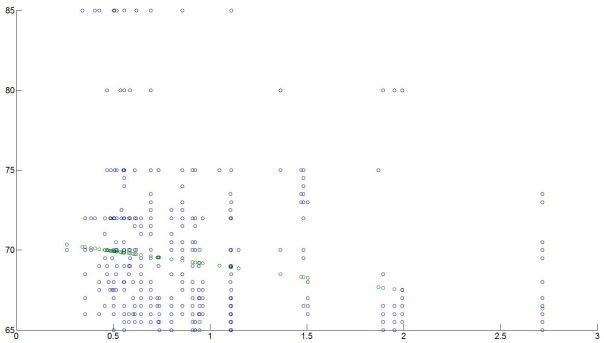


图 10：任务价格与任务所在地区（年）人均 GDP 标准化拟合图

图 10 中，将标准化后的任务所在地区（年）GDP 作为图像的横坐标，将任务价格作为纵坐标，拟合图形的拟合度较低，仅能看出拟合曲线有轻微的下降趋势，由此得出，任务所在丢（年）GDP 与任务价格并无明确关系。

针对上述条件的分析，我们可以得出，在我们所讨论的任务点周边平均会员信誉、任务开始时间、预定任务限额、任务点周边会员数量、任务数量、任务所在地区（年）GDP 六个条件中，预定任务限额与任务所在地区（年）GDP 对任务价格并没有明显影响，故使用其他四个量对最终结果进行拟合。

4. 1. 7 最终拟合结果分析

针对上述条件的分析，我们以任务点周边 6km 半径内会员数量、平均会员信誉、任务开始时间为拟合函数的三个未知量 x_1、x_2、x_3

将模型计做：　　$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3$

利用 MATLAB 的统计工具箱得到的结果如表一

表 1：模型（1）的计算结果

参数	参数估计值	参数置信区间
a_0	79. 5061	[75. 5763, 83. 4361]
a_1	−2. 1445	[−2. 5163, −1. 7727]
a_2	−0. 3934	[−0. 5694, −0. 2173]
a_3	−7. 8576	[−11. 9544, −3. 7607]
$R^2 = 0.2513$	$F = 92.9530$	$p < 0.0001$　　$s^2 = 15.3033$

由表 1 可知，所建立模型的拟合度仅 0.2513，拟合度较低，并且观察图六、图三发现拟合曲线的趋势比较缓慢，所以将一次拟合变为二次拟合再进行计算。

将模型计做：　　$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4{x_2}^2 + a_5{x_3}^2$

利用 MATLAB 的统计工具箱得到的结果如表 2

表 2：模型（2）的计算结果

参数	参数估计值	参数置信区间
a_0	73. 2197	[68. 9064, 77. 5330]
a_1	−0. 9840	[−1. 4755, −0. 4925]
a_2	−1. 0151	[−1. 4580, −0. 5723]
a_3	29. 1354	[16. 6241, 0. 1486]
a_4	0. 0904	[0. 0322, 0. 1486]
a_5	−31. 3688	[−41. 5039, −21. 2327]
$R^2 = 0.3021$	$F = 68.4295$	$p < 0.0001$ $S^2 = 14.5024$

由表 2 可知，所建立模型的拟合度和 F 值都比模型（1）有所改进，并且所有回归系数的置信区间都不含零点，说明模型二是完全可用的。

4. 1. 8 任务未完成原因分析

引用逻辑回归模型分析任务未完成原因，需要将 Hypothesls 的输出界定在[0~1]之间

$$0 \le h_\theta(x) \le 1$$

显然线性回归无法做到，故引入一个函数 g 令回归的 Hypothesls 表示为

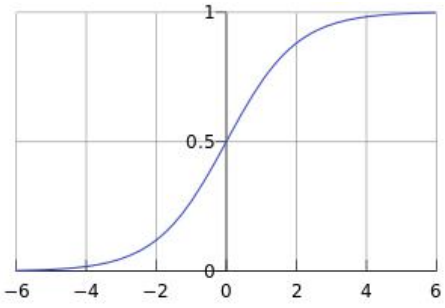
$$h_\theta(x) = g(\theta^T x)$$

g 的具体表达为 $$g(z) = \frac{1}{1+e^{-z}}$$

综合上述两式，得到的逻辑回归模型的数学表达式为

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$ （其中是 θ 参数）

得到函数图像 $g(z)$



由于在讨论任务完成情况时，我们用 1 来表示任务已完成，用 0 来表示任务未完成，故进行分类

$$P(y=0|x;\theta)+P(y=1|x;\theta)=1$$

$$P(y=0|x;\theta)=1-P(y=1|x;\theta)$$

由$g(z)$的图像可以看出，当$g(z)\geq 0.5$时，$z\geq 0$

对于$h_\theta(x)=g(\theta^T x)\geq 0.5$时$\theta^T x\geq 0$此时意味着预估$y=1$

反之当预测$y=0$时$\theta^T<0$

所以我们可以认为 $\theta^T x=0$ 是区分 0 与 1 的决策边界，当它大于 0 或小于 0 时，回归模型分别预测不同的结果[2]。

将任务点周边会员数、会员平均信誉、任务所在地区人均（年）GDP，任务价格的大数据带入回归模型中，分别计算出参数θ的值

表 3：因素对应θ表

因素	θ
会员数	-0.7092
会员平均信誉	0.5171
人均（年）GDP	-1.0125
任务价格	1.8593

由表 3 中的数据可以看出，会员数与会员平均信誉与人均（年）GDP 对任务完成情况呈负相关，任务价格与会员平均信誉对任务完成情况呈正相关，为了论证结论的严密性，我们用控制变量法从数据库中寻找四组分别能够反映以上变量对完成情况产生影响的数据。通过数据见的对比直观反映出任务未完成的原因。

表 4：信誉标准值值数据对比表

编号	取样点纬度	取样点经度	完成情况 1/0	会员数量标准值	信誉标准值	价格标准值	GDP 标准值
1	22.6354	114.2264	1	0.550479111	0.920477722	0.962223281	1.109233561
2	22.73803	114.2641	0	0.550479111	0.27026051	0.969458043	1.109233561
3	22.54808	113.9453	1	1.063598756	1.698532443	1.023352565	1.003201365
4	22.56470	113.9820	0	0.985678126	0.865982134	1.033652131	1.03316482
5	22.52394	113.9434	0	0.866532954	0.33659288	1.365984312	0.998656564
6	22.55101	113.9567	1	0.89945454	0.844769532	1.469952301	0.92236632

以 12、34、56 各为一对比组，根据数据观测，很明显可以看出，再其余影响因素完全相同的情况下，信誉高的任务完成，信誉值低的任务未完成，所以在任务的其余条件几乎相同的情况下，任务点周边的会员平均信誉度越高，这个任务越容易被完成，若在一个点周边会员的平均信誉度较低，就有可能导致任务无人问津的情况。

表 5：会员标准值数量数据对比表

编号	取样点纬度	取样点经度	完成情况 1/0	会员数量标准值	信誉标准值	价格标准值	GDP 标准值
1	22.59957	114.1304	0	1.689842852	0.370127606	0.962223281	1.109233561
2	22.737827	114.2859	1	0.448064392	0.402589707	0.947753758	1.109233561
3	22.65027	113.9366	0	0.332164742	0.645879432	1.332064685	1.109233561
4	22.73859	113.8184	1	2.112365655	0.569778446	1.033648519	1.109233561
5	23.03480	113.0880	1	1.655499764	0.778452661	0.996535241	0.910913381
6	23.03489	113.0901	0	0.311246494	0.745698514	1.022348979	0.910913381

以 12、34、56 各为一对比组，根据数据观测，很明显可以看出，再其余影响因素完全相同的情况下，会员数量多的任务未完成，会员数量少的任务完成。由此可以看出，如果在某个任务周边会员数量过多，导致该任务地区竞争力过高，会员过早承包任务，而又无力完成，导致任务无法完成。

表 6：价格标准值数据对比表

编号	取样点纬度	取样点经度	完成情况 1/0	会员数量标准值	信誉标准值	价格标准值	GDP 标准值
1	22.64154	114.0719	0	2.240321962	0.751437277	0.947753758	1.009233561
2	23.03098	113.3158	1	2.189114603	0.749260186	1.340518997	0.910913381
3	23.030043	113.1298	0	2.185647811	0.658965301	0.911246467	0.910913381
4	22.914524	113.6760	1	2.036469746	0.685422009	1.566648103	1.109233561
5	22.856156	114.1539	1	1.655587942	0.379854641	1.225546985	1.109233561
6	23.009444	113.0925	0	1.325468743	0.330125402	0.903322155	0.985556002

以 12、34、56 各为一对比组，根据数据观测，很明显可以看出，再其余影响因素完全相同的情况下，价格高的任务被完成，价格低的任务未被完成，由此可见，在其余条件均相同的情况下，会员更愿意选取任务定价高的任务去完成，而低价任务周边若是高价任务的数量较多，则导致低价任务无人问津的情况。

表 7：GDP 标准值数据对比表

编号	取样点纬度	取样点经度	完成情况 1/0	会员数量标准值	信誉标准值	价格标准值	GDP 标准值
1	22.70529	114.1403	0	1.1905711	2.220485317	1.029909457	1.109233561
2	22.52395	113.9434	1	1.075354542	2.7310849	0.947753758	0.717233538
3	22.604959	113.8575	0	2.033652019	0.675442135	1.333265089	1.254697467
4	22.750342	113.5835	1	2.155879465	0.731245668	1.235648785	0.799658552
5	23.132447	113.3398	1	1.226468756	0.779885468	0.966355874	0.733561438
6	23.137822	113.3913	0	1.245764986	0.744124656	0.954678102	1.099656452

以 12、34、56 各为一对比组，根据数据观测，很明显可以看出，再其余影响因素完全相同的情况下，任务所在地区 GDP 高，任务未被完成，任务所在地区 GDP 低，任务被完成。因为 GDP 是国民经济核算的核心指标，也是衡量一个国家的总体经济状况重要指标。所以可以用 GDP 值的大小来衡量一个地区的经济发展情况[3]，因此，GDP 高的地方经济较发达，可能导致比较少会员会关注及执行“拍照赚钱任务”，而 GDP 较低的地区，有比较多的会员会选择“拍照赚钱”的方式提高生活质量，所以任务被完成的可能性高。

4.2 问题二的分析与模型建立

以问题一建立的多元回归数学模型为依据，讨论会员数量，会员信誉，任务价格和 GDP 对任务完成率的具体影响，由于在无特殊情况的条件下，地区 GDP 的值较难发生改变，故在问题二中，我们只讨论会员数量，会员信誉任务价格对任务完成率的具体影响，为了避免情况的特殊性，我们选了 500 组数据进行运算（具体运算程序下文附上），根据统计出的数据，计算出当所讨论的三个变量发生改变时任务完成情况增加的概率，通过比较概率大小，设计新的任务定价方案。

4.2.1 算法介绍

```
for j = 1:10;
    c = 0;
    d = 0;
    for i = 1:500;
        a = mean(compl == 1);
        theta1 = theta + [rand * 0.7092;0;0;0];
        p1 = predict(theta1, X);
        b1 = mean(p1 == 1);
        if b1 >= a
            c = c + 1;
        else
            d = d + 1;
        end
    end
    cal = c/500;
    dal = d/500;
    cal10 = cal10 + cal;
    dal10 = dal10 + dal;
end
calc = cal10 / 10;
dalc = dal10 / 10;
```

在程序中输入 500 组数据，对数据进行循环，在循环中不断把这 500 组数据中任务完成率提高的数据提取并统计，将统计结果与 500 进行比较，算出完成率提高的概率，通过概率的大小来判断改进方案。

4.2.2 因数系数导致完成率提高分析表

表 8：因数系数导致完成率提高分析表

	会员数量		会员平均信誉		任务价格	
会员数量	+ + 83.4%	− − 0%	− + 64.72%	− − 0%	− + 13.24%	− − 0%
会员平均信誉	+ + 96.86%	+ − 11.62%	+ + 64.10%	− − 0%	− + 42%	− − 0%
任务价格	+ + 99.58%	+ − 76.88%	+ + 98.78%	+ − 87.52%	+ + 93.64%	− − 0%
+ + +	99.9%		− − −		0%	

4.2.3：基于完成率提高分析表进行任务定价

由表八中的完成率提高百分比数据可以明显看出：当会员数量、会员平均信誉与任务价格均提高时，完成率提高的概率最高，达到 99.9%，但是由于客观情况，会员平均信誉度很难在短时间内大幅提高，所以主要影响任务定价的因素是会员数量和任务价格。根据表八中的数据，我们可以看出当会员数量和任务价格同时提高时，完成率提高的概率达到 99.58%，故提出建议，大力宣传“拍照赚

钱”APP，用商业手段引导更多群众注册会员，提高会员数量，在保证盈利的条件下，适当提高任务价格，诱导更多的会员参与完成任务，从而提高会员的平均信誉。通过此种方法同时提高三个因数，大大提高任务完成概率。

根据模型（2）得出原定价公式：

$$y = 73.2197 - 0.9840x_1 - 1.0151x_2 + 29.1354x_3 + 0.0904x_2^{\ 2} - 31.3688x_3^{\ 2}$$

根据模型（2）修正定价公式：

$$y = 73.2197 - 0.1245x_1 - 0.6234x_2 + 29.1354x_3 + 2.5684x_2^{\ 2} - 31.3688x_3^{\ 2}$$

根据模型（2）得出原参照系数公式（精准度为 65.4611%）：

$$\partial = -0.6165x_1 + 0.1133x_2 + 1.4159x_3$$

根据模型（2）修正参照系数公式：（精准度为 79.0698%）

$$\partial = -1.6371x_1 + 1.5888x_2 + 3.0465x_3$$

4.2.4：新旧模型比较

下表是引用新旧参照系数公式得出的数据对比图，最后一栏的参照系数是通过 $y = \frac{1}{1+e^{-x}}$ 公式，根据拟合比对得出当 y 为 0.63 时真实数据的完成度最高，故令 y =0.63 时逆推出 x 的取值，即 x 为参照系数，当 $x \geq 0.5322$ 时，我们默认任务完成。

表 9：参照系数前后对比图

	纬度	经度	完成情况	会员数	会员信誉	价格	参照系数
原方案	22.63539828	114.2263772	1	0.550479111	0.920477722	0.962223281	1.127331698
	22.73802844	114.264097	0	0.550479111	0.27026051	0.969458043	1.063905787
	22.54808582	113.9453119	1	1.063598756	1.698532443	1.023352565	0.985699989
	22.56470734	113.9820093	0	0.985678126	0.865982134	1.033652131	0.953993263
	22.52394954	113.9434416	0	0.866532954	0.33659288	1.365984312	1.438015595
	22.55101305	113.9567446	1	0.89945454	0.844769532	1.469952301	1.622504127
	22.59955905	114.130363	0	1.689842852	0.370127606	0.962223281	0.362559284
	22.73782038	114.2858901	1	0.448064392	0.402589707	0.947753758	1.111306262
	22.65027837	113.9366241	0	0.332164742	0.645879432	1.332064985	1.754468964
	22.73859039	113.8184493	1	2.112365655	0.569778446	1.033648519	0.22582541
	23.03480604	113.0880386	1	1.655499764	0.778452661	0.996535241	0.47857733
	23.03489605	113.0901446	0	0.311246494	0.745698514	1.022348979	1.340148097
	22.64153534	114.0719117	0	2.240321962	0.751437277	0.947753758	0.0459039
	23.03098143	113.3157936	1	2.189114603	0.749260186	1.340518997	0.633342874
	23.03004352	113.1298659	0	2.185647811	0.658965301	0.911246467	0.017442766
	22.91452474	113.6760975	1	2.036469746	0.685422009	1.566648103	1.040391765
	22.85615682	114.1539907	1	1.655587942	0.379854641	1.225546985	0.75761954
	23.009444	113.0925428	0	1.325468743	0.330125402	0.903322155	0.499265567
	22.70528903	114.1403154	0	1.1905711	2.220485317	1.029909457	0.975842704
	22.52394954	113.9434416	1	1.075354542	2.7310849	0.947753758	0.98840039
	22.60495971	113.8575896	0	2.033652019	0.675442135	1.333265089	0.710551164
	22.75034263	113.5835226	1	2.155879465	0.731245668	1.235648785	0.503305559
	23.1324472	113.3398257	1	1.226468756	0.779885468	0.966355874	0.700506318
	23.13782219	113.3913207	0	1.245764986	0.744124656	0.954678102	0.668023934

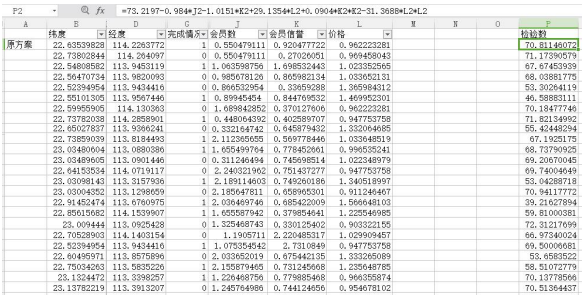
	纬度	经度	完成情况	会员数	会员信誉	价格	参照系数
新方案	22.63539828	114.2263772	1	0.550479111	0.920477722	0.962223281	3.492678879
	22.73802844	114.264097	0	0.550479111	0.27026051	0.969458043	2.481654473
	22.54808582	113.9453119	1	1.063598756	1.698532443	1.023352565	4.07505441
	22.56470734	113.9820093	0	0.985678126	0.865982134	1.033652131	2.911239972
	22.52394954	113.9434416	0	0.866532954	0.33659288	1.365984312	3.277648875
	22.55101305	113.9567446	1	0.89945454	0.844769532	1.469952301	4.347882491
	22.59955905	114.130363	0	1.689842852	0.370127606	0.962223281	0.753030234
	22.73782038	114.2858901	1	0.448064392	0.402589707	0.947753758	2.793440134
	22.65027837	113.9366241	0	0.332164742	0.645879432	1.332064985	4.540521406
	22.73859039	113.8184493	1	2.112365655	0.569778446	1.033648519	0.596120395
	23.03480604	113.0880386	1	1.655499764	0.778452661	0.996535241	1.562531537
	23.03489605	113.0901446	0	0.311246494	0.745698514	1.022348979	3.789810326
	22.64153534	114.0719117	0	2.240321962	0.751437277	0.947753758	0.413584286
	23.03098143	113.3157936	1	2.189114603	0.749260186	1.340518997	1.69051619
	23.03004352	113.1298659	0	2.185647811	0.658965301	0.911246467	0.244952401
	22.91452474	113.6760975	1	2.036469746	0.685422009	1.566648103	2.527887314
	22.85615682	114.1539907	1	1.655587942	0.379854641	1.225546985	1.626778922
	23.009444	113.0925428	0	1.325468743	0.330125402	0.903322155	1.106549305
	22.70528903	114.1403154	0	1.1905711	2.220485317	1.029909457	4.716442285
	22.52394954	113.9434416	1	1.075354542	2.7310849	0.947753758	5.466016593
	22.60495971	113.8575896	0	2.033652019	0.675442135	1.333265089	1.805642837
	22.75034263	113.5835226	1	2.155879465	0.731245668	1.235648785	1.396816869
	23.1324472	113.3398257	1	1.226468756	0.779885468	0.966355874	2.175233203
	23.13782219	113.3913207	0	1.245764986	0.744124656	0.954678102	2.051250233

根据新旧两张图的参照系数对比情况，我们可以看出在依照新的任务定价方案修正参照系数公式后，参照系数普遍提高，并且均大于 0.5322，说明理论该改进方案能使任务完成率大大提升。

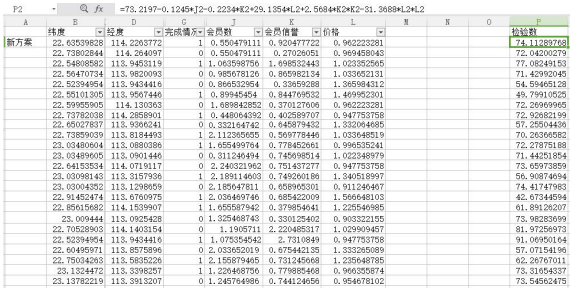
4.2.5：新旧定价方案比较

下表是引用新旧定价公式得出的数据对比图：

表 10：任务定价先后对比图



	纬度	经度		完成情况	会员数	会员信誉	价格					检验数
原方案	22.63539828	114.2263772		1	0.550479111	0.920477722	0.962223281					70.81146072
	22.73802844	114.264097		0	0.550479111	0.27026051	0.969458043					71.17390579
	22.54808582	113.9453119		1	1.063598756	1.698532443	1.023352565					67.67453939
	22.56470734	113.9820093		0	0.985678126	0.865982134	1.033652131					68.03881775
	22.52394954	113.9434416		0	0.866532954	0.33959288	1.365984312					53.39264119
	22.55101305	113.9567446		1	0.89945454	0.844769532	1.469952301					46.58883111
	22.59955905	114.130363		0	1.689842852	0.370127606	0.962223281					70.18477446
	22.73782038	114.2858901		1	0.448064392	0.402589707	0.947753758					71.82134992
	22.65027837	113.9306241		0	0.332164742	0.645879432	1.332064685					55.42448294
	22.73859039	113.8184493		1	2.112365655	0.569778446	1.033648519					67.1925175
	23.03480604	113.0880386		1	1.655499764	0.778452661	0.996535241					68.73790925
	23.03489605	113.0901446		0	0.311246494	0.745698514	1.022348979					69.20670045
	22.64153534	114.0719117		0	2.240321962	0.751437277	0.947753758					69.74004649
	23.03098143	113.3157936		1	2.189114603	0.749260186	1.340518997					53.04288718
	23.03004352	113.1298659		0	2.185647811	0.658965301	0.911246467					70.94117772
	22.91452474	113.6760975		1	2.036469746	0.685422009	1.566648103					39.21627894
	22.85615682	114.1539907		1	1.655587942	0.379854641	1.225546985					59.81000381
	23.009444	113.0925428		0	1.325468743	0.330125402	0.903322155					72.31217699
	22.70528903	114.1403154		0	1.1905711	2.220485317	1.029909457					66.97340024
	22.52394954	113.9434416		1	1.075354542	2.7310849	0.947753758					69.50006681
	22.6049971	113.8575896		0	2.033652019	0.675442135	1.333265089					52.6583522
	22.75034263	113.5835226		1	2.155879465	0.731245668	1.235648785					58.51072779
	23.1324472	113.3398257		1	1.226468756	0.779885468	0.966355874					70.13778566
	23.13782219	113.3913207		0	1.245764986	0.744124656	0.954678102					70.51364437



	纬度	经度		完成情况	会员数	会员信誉	价格					检验数
新方案	22.63539828	114.2263772		1	0.550479111	0.920477722	0.962223281					74.11289768
	22.73802844	114.264097		0	0.550479111	0.27026051	0.969458043					72.04200279
	22.54808582	113.9453119		1	1.063598756	1.698532443	1.023352565					77.08249153
	22.56470734	113.9820093		0	0.985678126	0.865982134	1.033652131					71.42992045
	22.52394954	113.9434416		0	0.866532954	0.33959288	1.365984312					54.59465128
	22.55101305	113.9567446		1	0.89945454	0.844769532	1.469952301					49.79910525
	22.59955905	114.130363		0	1.689842852	0.370127606	0.962223281					72.26969965
	22.73782038	114.2858901		1	0.448064392	0.402589707	0.947753758					72.92682199
	22.65027837	113.9306241		0	0.332164742	0.645879432	1.332064685					57.25504439
	22.73859039	113.8184493		1	2.112365655	0.569778446	1.033648519					70.26366582
	23.03480604	113.0880386		1	1.655499764	0.778452661	0.996535241					72.27875188
	23.03489605	113.0901446		0	0.311246494	0.745698514	1.022348979					71.44251854
	22.64153534	114.0719117		0	2.240321962	0.751437277	0.947753758					73.65973859
	23.03098143	113.3157936		1	2.189114603	0.749260186	1.340518997					56.90874694
	23.03004352	113.1298659		0	2.185647811	0.658965301	0.911246467					74.41747983
	22.91452474	113.6760975		1	2.036469746	0.685422009	1.566648103					42.67344594
	22.85615682	114.1539907		1	1.655587942	0.379854641	1.225546985					61.89126207
	23.009444	113.0925428		0	1.325468743	0.330125402	0.903322155					73.98283699
	22.70528903	114.1403154		0	1.1905711	2.220485317	1.029909457					81.97256973
	22.52394954	113.9434416		1	1.075354542	2.7310849	0.947753758					91.06950164
	22.6049971	113.8575896		0	2.033652019	0.675442135	1.333265089					57.07154196
	22.75034263	113.5835226		1	2.155879465	0.731245668	1.235648785					62.26767011
	23.1324472	113.3398257		1	1.226468756	0.779885468	0.966355874					73.31654337
	23.13782219	113.3913207		0	1.245764986	0.744124656	0.954678102					73.54562475

跟据新旧两张图的任务定价对比情况，可以明显看出根据新的定价公式得出的定价方案，并且根据检验，该定价方案成立。

4.3 问题三的分析与模型建立

引用聚类模型，对所有任务点按照相近距离和相近价格进行区域聚类，考虑到不同地区密集程度不同并具有一定差异的实际情况，故限定每个类包内的点数，据此形成类包，对每个类包进行拟合，作为改变定价模型的依据。

4.3.1：类包形成依据及分析

```
for i = 1:300
    ra = randi([1,835],1,1);
    if com(ra) ~= 0
        n = 1;
        jingn = jing(ra);
        wein = wei(ra);
        pricen = price(ra);
        GDPn = GDP(ra);
        for j = 1:835
            if (price(ra) - 5 < price(j) < price(ra) + 5) && ( dis( ra,j ) < 5 )
                n = n + 1;
                com(j) = 0;
%                 jing(i) = jing(i);
%                 wei(i) = wei(i);
%                 price(i) = price(i);
                jingn = jingn + jing(i);
                wein = wein + wei(i);
                pricen = pricen + price(i);
                GDPn = GDPn + GDP(i);
                if n > 35
                    break;
                end
            end
        end
        if pricen ~= 0
            new(i,1) = jingn / n ;
            new(i,2) = wein / n ;
            new(i,3) = pricen / n;
            new(i,4) = GDPn / n;
            new(i,5) = n;
%             new(i,1) = median(jingm);
%             new(i,2) = median(wein);
%             new(i,3) = median(price);
        end
    end
end
```

图 11：类包形成程序图

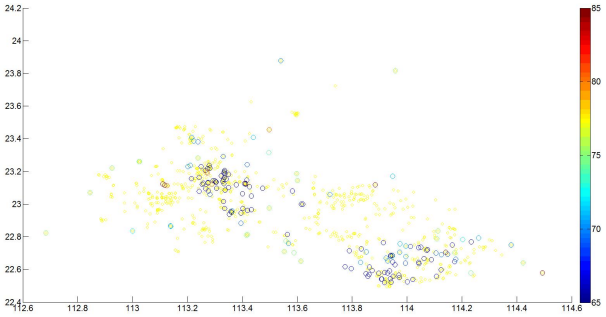


图 12：类包形成图

根据图 11 中的程序显示，基于对一些数据的观察，我们设定所有五公里范围内和±5元 价格差的任务点进行区域聚类，考虑到不同地区密集程度不同并具有一定差异的实际情况，故限定每个类包内的点数不得超过 35 个，以上述为聚类程序依据。运行程序需后得到图 12，其中彩色点代表任务分布点，空心圆点代表类包聚合后的点，聚合的依据是类包内所有任务点经纬度平均值。

4.3.2：拟合结果分析

将模型计做：$y = a_0 + a_1 x_2 + a_2 x_3 + a_3 x^2 + a_4 x_3^{\ 2} + a_5 x_4$

利用 MATLAB 的统计工具箱得到的结果如表十一

表 11：模型（3）的计算结果

参数	参数估计值	参数置信区间
a_0	72.5979	[67.2454, 77.9503]
a_1	−2.7791	[−4.3076, −1.2505]
a_2	40.6221	[22.4488, 58.7955]
a_3	0.5078	[0.1397, 0.8760]
a_4	−40.7318	[−56.3428, −25.1208]
a_5	−0.9909	[−1.4889, −0.4930]
$R^2 = 0.3899$	$F = 20.4543$	$p < 0.0001$ $S^2 = 14.6958$

由表 11 可知，所建立模型的拟合度和 F 值都在可行范围内，并且所有回归系数的置信区间都不含零点，说明模型三是完全可用的。

4.3.3 完成情况影响分析

①根据模型（2）得出修正定价公式：

$$y = 73.2197 - 0.1245 x_1 - 0.6234 x_2 + 29.1354 x_3 + 2.5684 x_2^{\ 2} - 31.3688 x_3^{\ 2}$$

②根据模型（3）得出打包发布定价公式：

$$y = 72.5929 - 2.7791x_2 + 40.6221x_3 + 0.5078{x_2}^2 - 40.7318{x_3}^2 - 0.9909x_4$$

由$z = \theta^T x$以及$g(z) = \frac{1}{1+e^{-z}}$

将①的计算数据带入$z = \theta^T x$中

根据$g(z) = \frac{1}{1+e^{-z}}$得出①的完成度为57.23%

用同样的方法计算得出②的完成度为56.63%

根据计算的完成度结果分析将这些任务联合在一起打包发布与题二中提出的修正系数定价方案在完成度上没有明确区别。

4.4 问题四的分析与结论

对附件三的数据进行预处理。通过问题一二三的数据结论比较，得到模型二与模型三并无明显差异，但对比原方案，效果有显著提升。因此，我们选择采用模型二，并且利用问题二中的定价方案解决问题四。

将模型计做：$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4{x_2}^2 + a_5{x_3}^2$

令$z = \theta^T x$得出$g(z) = \frac{1}{1+e^{-z}}$

将数据带入，得出其完成度为67.92%

根据前后完成度的对比情况可以看出，该方案相比原方案能够大大增加任务的完成度。

五、模型评价及改进

优点：

（1）回归模型分析可以准确的计量各个因素之间的相关程度与回归拟合程度的高低，提高预测方程式的效果。

（2）回归分析法在分析多因素模型时更加简单和方便。

（3）在图像处理方面，我们使用了 Matlab 作图，拟合数据的变化趋势以及散点图的拟合程度，使结果更加清晰、更具有条理性和更加直观。

缺点：

（1）

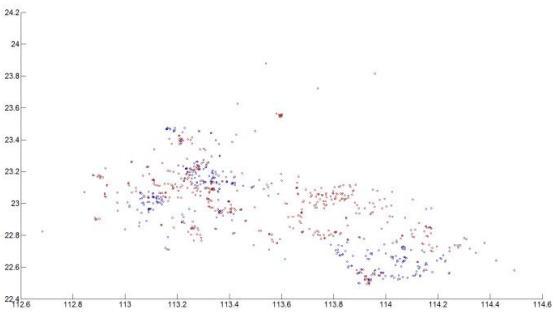


图 13：任务完成情况实际分布图

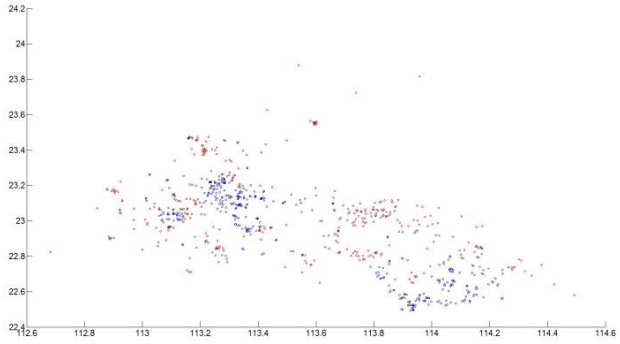


图 14：任务点完成情况拟合分布图

图 13 和图 14 分别表示任务完成情况的实际和拟合图形，用红蓝两色对任务是否完成加以区分，红色代表该点的任务完成，蓝色代表该点的任务未完成，通过实际图形和拟合图形的直观比较可以看出，我们所建立的模型在某些地区（广州）与实际图形有比较明显的差异。所以模型善不完善，适用区域有限。

（2）模型的建立没有完全考虑影响因素之间的交互性，建立背景太过理想化。

（3）所建立的回归模型拟合程度较低，还有待改进。

改进：可以更全面的考虑影响任务定价规律的原因，应当考虑影响因素之间的交互性。应当针对任务所在地区的差异进行分类讨论，增大模型的适用范围。

六、参考文献

【1】
https://baike.baidu.com/item/matlab%20polyfit/10186675?fr=aladdin

【2】Coursera 公开课笔记：斯坦福大学机器学习第六课“逻辑回归(Logistic Regression)”

【3】国家统计局.2015 年 1 季度我国 GDP（国内生产总值）初步核算情况

七、附录

```
function  g  =  sigmoid(z)

g  =  zeros(size(z));

g  =  (1+exp(-z)).^-1;

end
```

```
function  p  =  predict(theta,  X)

m  =  size(X,  1);

p  =  zeros(m,  1);

k  =  find(sigmoid(  X  *  theta)  >=  0.62  );
p(k)  =  1;

end
```

```
function  [J,  grad]  =  costFunctionReg(theta,  X,  y,  lambda)
m  =  length(y);

J  =  0;
grad  =  zeros(size(theta));

h  =  sigmoid(X  *  theta);
J  =  1/m  *  ((-y'*log(h))-((1-y)'*log(1-h)))  +  lambda  /  (2*m)
*  sum(theta(2:end).^2);
grad_1  =  theta(2:end);
grad  =  1/m  *  X'  *  (h  -  y)  +  lambda  /  m  *  [0;grad_1];

end
```