微信小程序影响因素分析报告

第 12 组 - 2018 年 10 月

组长：王嬰　18210980068
组员：栗书敬　18210980007
张继丹　18210980019
张媛媛　18210980079
张云华　18210980080
柳素问　18210980090

微信小程序影响因素分析

摘要：本案例以微信小程序为研究对象，通过统计分析探究评价人数、上线时长、评分等因素对微信小程序人气的影响作用，建立了线性回归模型来刻画各因素与小程序人气之间的关联。结论表明评价人数、上线时长对于小程序人气都有显著影响，小程序各类别对人气的影响有所区别。

一、 背景介绍

近年来通讯社交、视频、游戏等主流行业应用活跃用户规模增长乏力，行业应用活跃率出现不同程度下降。移动智能终端用户平均安装与平均每日打开应用款数已连续两年出现下滑，存量时代移动应用对于用户的争夺更加激烈。在这样的背景下，各类轻应用陆续探索市场，2017 年 1 月 9 日，微信小程序正式上线。

小程序在 2017 年整体处于用户培育期，偏向于以公众号体系、微信支付为基础通道及能力的内部体系打造，开发者和用户都在探索小程序产品定位，公众号关联是刺激小程序用户规模增长的最有效手段。2017 年 12 月末发布的小游戏彻底引爆了小程序用户增长，疯狂的社交传播使得小程序活跃用户规模在 2018 年 Q1 突破了 4 亿，也让开发者对与小程序的流量潜力有了最直接的体验。

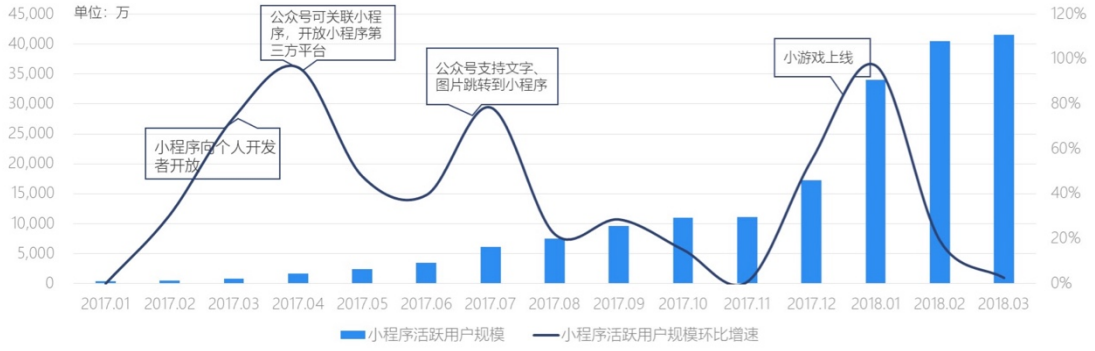

图 1-1：微信小程序活跃用户规模增长情况

小程序借小游戏的超高流量跨入新的阶段，在社交关系链为主导的大方向下走向了广告变现的主旋律，持续激发长尾价值。截至 2018 年 Q2 微信小程序总量突破 100 万，累计用户数破 6 亿。在众多的小程序中，哪些小程序是最有人气的？哪些小程序的用户活跃度最高？事实上，小程序人气的高低是多种因素综合作用的结果，到底有哪些因素在影响小程序的人气？不同小程序之间人气的巨大差异又是如何产生的？本报告采用课程提供的小程序相关数据，对小程序人气的影响因素展开研究。

二、 数据说明

本报告使用的是课程提供的微信小程序相关数据，共 1212 条记录。在数据预处理阶段，我们将原始数据中的发布时间的日期型变量转变为数字型，以 2017 年 6 月 30 日为截止日计算在线时长；将原本数据的 61 个分类水平进行整合，形成 16 个新的分类水平。此外，还利用原始数据中的小程序介绍这一文本信息构建了介绍字数、特征词比例[1]以及 TF-IDF 权重[2]这三个变量。模型构建所用数据共包含 8 个变量，因变量为小程序人气值，其他变量为自变量。具体的变量说明如表 2-1 所示。

表 2-1:数据变量说明表

变量类型		变量名	详细说明	取值范围	备注
因变量		人气	定量变量	[710, 977582]	
自变量	外部因素	评价人数	单位：人次	[0, 36]	
		评分	单位：分	[0, 5]	好评率与评分线性相关，选取评分
	内部因素	特征词比例	定量变量	[0, 1]	
		TF-IDF 权重	定量变量	[0, 9.7]	
		介绍字数	单位：个	[4, 588]	
		上线时长	单位：天	[19, 185]	以 2017-6-30 为截止日，计算上线时长
		分类	定性变量：共 16 个水平	办公、出行、工具、购物等	

三、 描述性分析

在对小程序的影响因素进行模型探究之前，首先对各变量进行描述性分析，以初步判断人气的影响因素，为后续研究做铺垫。

（一） 因变量：人气值

在本案例中，人气值的最小值为 710，所对应的小程序是超搞笑 GIF 和去哪儿全球购，评价人数均为 1，所属类别分别为社交类和购物类。人气最大值为 977582，所对应的小程序是摩拜单车，所属类别为出行交通类，人气值高的原因主要是由于发布时间较早以及小程序很好的起到了替代 APP 的作用，用户需求较大。

通过绘制对数人气分布直方图（图 3-1）可以看到，各种小程序人气值差异较大。具体来说，人气的均值为 13837.05，中位数为 8986，中位数与均值差异较大。这一现象符合

[1] 特征词比例：每一个样本中特征词的个数占这一样本所有词的比重。
[2] TF-IDF 权重：每个样本数据中词的 TF-IDF 权重的之和。

我们对于小程序人气值判断的基本认知：即存在少数程序用户量远高于其他，因此拉高了整体人气的平均值。

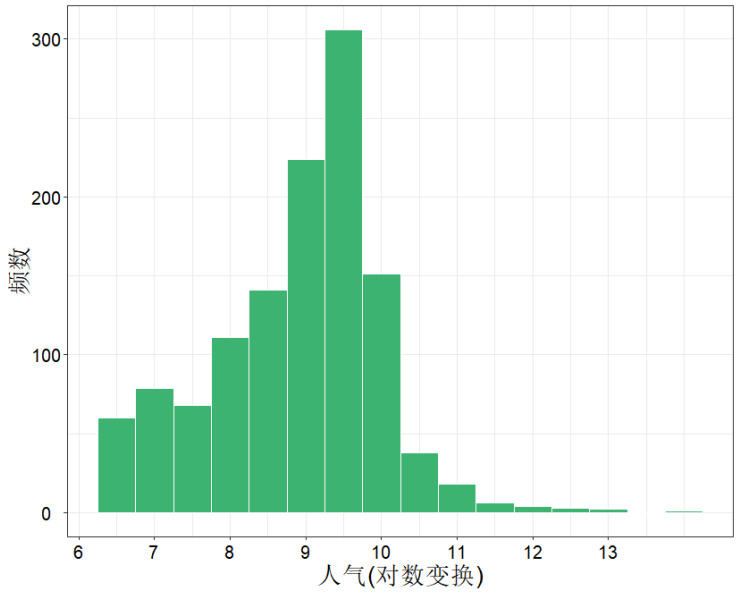

图 3-1：小程序人气对数分布直方图

（二） 自变量：外部因素

外部因素即评价因素，具体包括评价人数和评分两个变量。从图 3-2 可以看出，样本数据中的小程序评价人数较少，1212 个数据中共有 711 个数据中评价人数和评分均为 0，占比达到 58.67%。存在“评价人数”数据的小程序中，评价人数也集中在 1-10 人之间，评价人数最多为 36 人，对应小程序为当当购物。存在评分变量的 501 个数据中，绝大多数评分集中在 4.0-5.0 之间，其中 4.0-4.9 分占比为 30.14%，5.0 分占比达到 63.67%。评分呈高分集中分布现象，说明绝大多数小程序用户体验较好，得到使用者的认可。

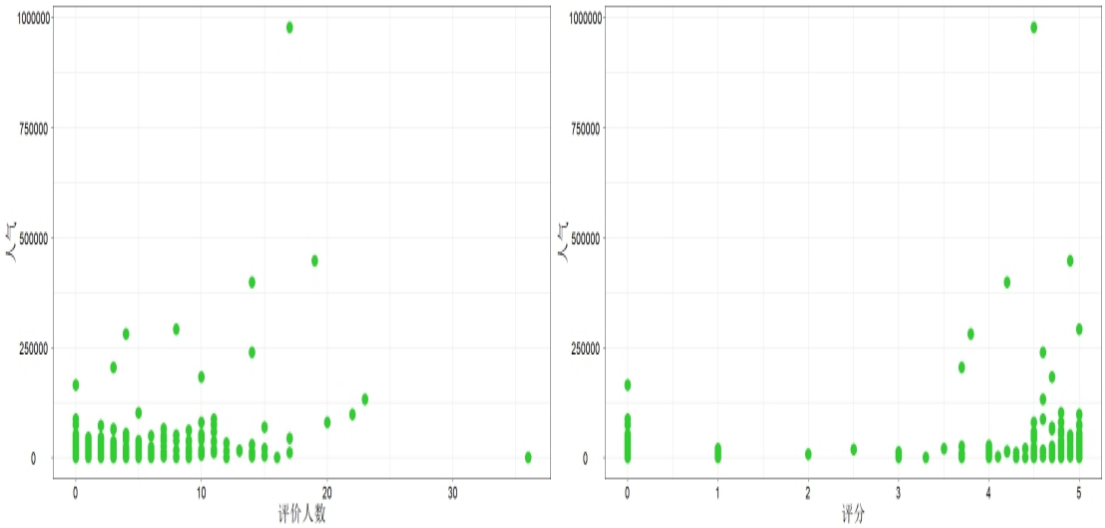

图 3-2：小程序人气与评价人数、评分的散点图

（三） 自变量：内部因素

内部因素包括小程序上线时长、所属类别、介绍字数等变量。微信小程序于 2017 年 1 月 9 日正式上线，因此 2017 年 1 月新上线时间短发布的小程序数量最多，即上线时长集中在 150-172 天。仅 2017 年 1 月 9 日当天新上线小程序占比达到样本总量的 15.92%，1 月份整月上线小程序数量为样本总量的 1/3。从图 3-3 中也可以看出，2017 年 1 月份发布的小程序累计人气值最高。

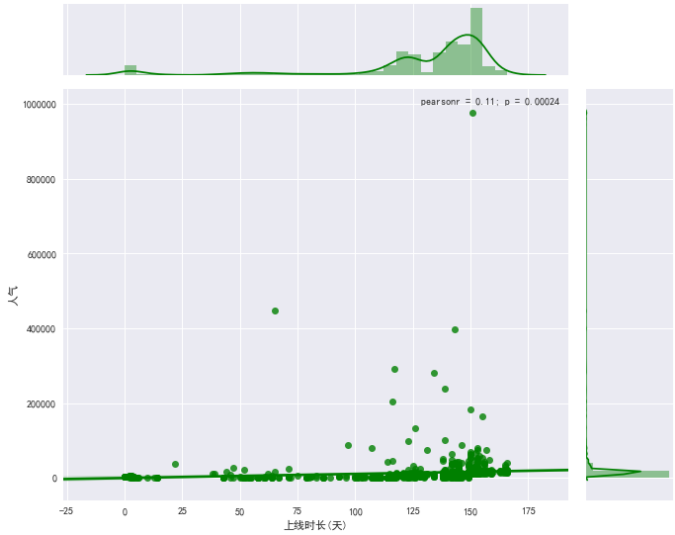

图 3-3：小程序人气与上线时长的散点图

从图 3-4 的结果上来看，在 16 个分类水平下按照中位数大小划分，音乐类小程序人气最高，视频类的人气最低，宽度代表类别的样本个数。

进一步分析类别间的差异可知，“工具”、“社交”类数量最多，都超过 150 个；“体育”数量最少，仅为 19 个。不同类型的小程序在市场上的占有率差异较大，且大多集中在少数的几种类型。这在一定程度上也能反映出市场对不同类型小程序的需求差异。

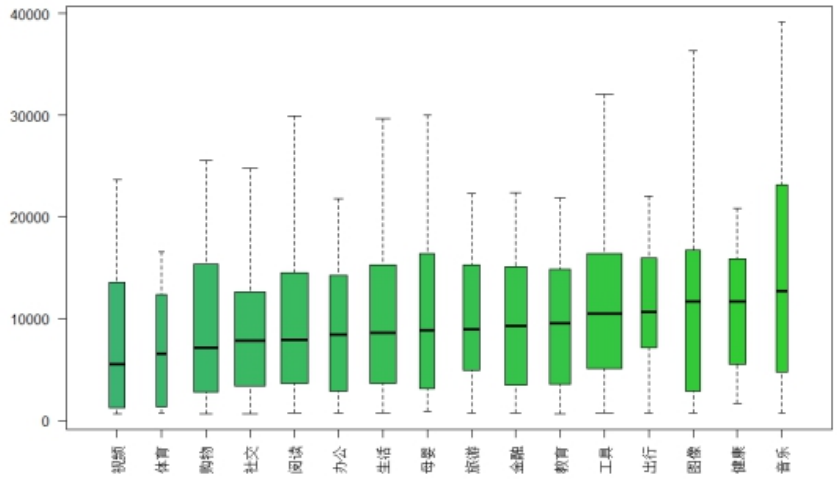

图 3-4：小程序人气与分类的箱线图

提取小程序介绍中词频前 200 的进行词云图展示。选取高人气的出行类与人气值相对较低的购物类进行对比分析。出行类词云关键词为：程序、查询、服务、出行、代驾、公交、预定等，展现了出行类小程序应用性强的特点。购物类词云关键词为：购物、生活、商品、优质、正品、服务、产品。

图 3-5：小程序出行类与购物类的词云图

综上，通过对本案例数据的描述性分析，可以推测：对小程序人气值可能会产生影响的因素包括：评价因素（评价人数、评分）和内部因素（在线时长，所属类别，介绍字数，特征词比例等）；从影响作用来看，预测评价人数、在线时长因素相比其他可能影响作用更为明显一些。

四、 模型建立

为了更深入地分析各因素对小程序人气的影响，本案例将建立小程序人气值关于外部评价因素和内部因素的回归模型，使用定量化的方式更为精细地刻画两大方面各个因素的影响作用大小。

（一） 线性回归模型

首先，对数据建立简单的线性回归模型，结果发现，简单线性回归模型的 F 检验虽然拒绝原假设，但是调整 R^2 仅为 0.16，模型的拟合程度较差，且残差的分布明显不符合正态性。因此考虑对因变量人气值做对数变换处理，进行模型优化。

对因变量做对数变换后，再次建立线性模型并进行诊断。估计结果如表 4-1 所示。从表 4-1 可以看出，对数变换后，模型建立是显著的（通过 F 检验），模型的拟合程度有一定上升，调整 R^2 提高到 0.34。但评分、特征词比例、TF.IDF 权重、介绍字数等变量对因变量人气值的影响不显著，因此需要进一步对自变量进行筛选。为了解决这一问题，考虑采用 STEP 逐步回归分析，进一步筛选自变量。

表 4-1:线性回归结果

变量	回归系数	P 值	备注
评价人数	0.0945	<0.001	
评分	0.0029	0.8310	
特征词比例	-0.2233	0.4077	
TF.IDF 权重	0.0028	0.9482	
介绍字数	0.0010	0.2839	
上线时长	0.0162	<0.001	
办公	0.5059	<0.05	
出行	0.5865	<0.05	
工具	0.6653	<0.01	
购物	0.3597	0.1131	
健康	0.8554	<0.001	
教育	0.4836	<0.05	
金融	0.4682	<0.05	
旅游	0.4753	<0.1	基准组：体育
母婴	0.6862	<0.01	
社交	0.4846	<0.05	
生活	0.6216	<0.01	
视频	-0.0614	0.8068	
图像	0.4041	0.1094	
音乐	0.4399	0.1180	
阅读	0.4071	<0.1	
F 检验	P 值<0.0001	调整的 R^2	0.3442

z

（二） STEP 逐步回归

通过 STEP 逐步回归分析,估计结果如表 4-2 所示，可以看出，评分、特征词比例、TF.IDF 权重等变量被筛选剔除，筛选后的变量包括评价人数、介绍字数、上线时长、分类等都会对因变量人气值有一定的影响作用。模型通过了 F 检验，模型的拟合程度略有提高（调整 R^2=0.3454）。进一步进行模型诊断，诊断结果如图 4-1 所示。

从图 4-1 可以看出，Cook 距离表现正常，表明没有异常点。残差的波动情况有所缓解，但模型仍可能存在异方差的问题。为了解决这一问题，采用加权最小二乘法进一步修正异方差。

表 4-2:逐步回归结果

变量	回归系数	P 值	备注
评价人数	0.0958	<0.001	
介绍字数	0.0011	<0.1	
上线时长	0.0162	<0.001	
办公	0.5121	<0.05	基准组：体育
出行	0.5933	<0.05	
工具	0.6693	<0.05	
购物	0.3674	0.1036	
健康	0.8621	<0.001	
教育	0.4886	<0.05	
金融	0.4681	<0.05	
旅游	0.4813	<0.05	
母婴	0.6969	<0.01	
社交	0.4928	<0.05	
生活	0.6288	<0.01	
视频	-0.0548	0.8268	
图像	0.4081	0.1051	
音乐	0.4564	0.1023	
阅读	0.4112	<0.1	
F 检验	P 值<0.0001	调整的 R^2	0.3454

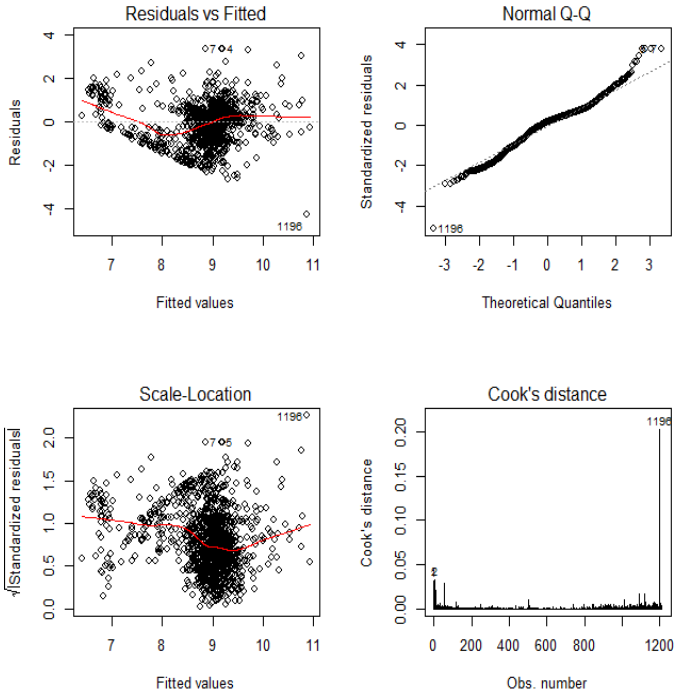

图 4-1：逐步回归后的模型诊断图

（三） 加权最小二乘法修正异方差

加权最小二乘法进一步处理后,回归结果如表 4-3 所示，诊断结果如图 4-2 所示。

从图 4-2 可以看出，Cook 距离表现正常，表明没有异常点。通过加权最小二乘法修正处理，使得异方差得到消除。在 99%置信区间下可以不拒绝原假设，即认为不存在异方差。从表 4-3 可以看到，模型仍然是显著的（通过了 F 检验），模型的拟合程度略微上升（调整 R^2=0.3537）。

表 4-3:修正异方差后的线性回归模型回归结果

变量	回归系数	P 值	备注
评价人数	0.1234	<0.001	
介绍字数	0.0002	0.4359	
上线时长	0.0239	<0.001	
办公	0.4342	<0.05	基准组：体育
出行	0.5392	<0.01	
工具	0.6005	<0.001	
购物	0.4269	<0.05	
健康	0.7048	<0.001	
教育	0.4068	<0.05	
金融	0.4391	<0.05	
旅游	0.3719	<0.05	
母婴	0.5386	<0.01	
社交	0.3608	<0.05	
生活	0.4668	<0.01	
视频	0.0156	0.9511	
图像	0.5309	<0.05	
音乐	0.3594	0.1173	
阅读	0.3616	<0.05	
F 检验	P 值<0.0001	调整的 R^2	0.3537

通过表 4-3 可以得到如下结论：在控制其他因素不变的情况下，

- 不同类别小程序之间的人气值有所区别。体育类小程序人气值最低，健康类小程序人气值最高、比体育类平均高 70.48%；
- 评价人数对人气值有正向影响作用，评价人数每增加 1 人，人气值平均增长 12.34%。
- 小程序上线时长的增长会带来人气值的增长，上线时长每增长 1 天，人气值平均增长 2.39%。

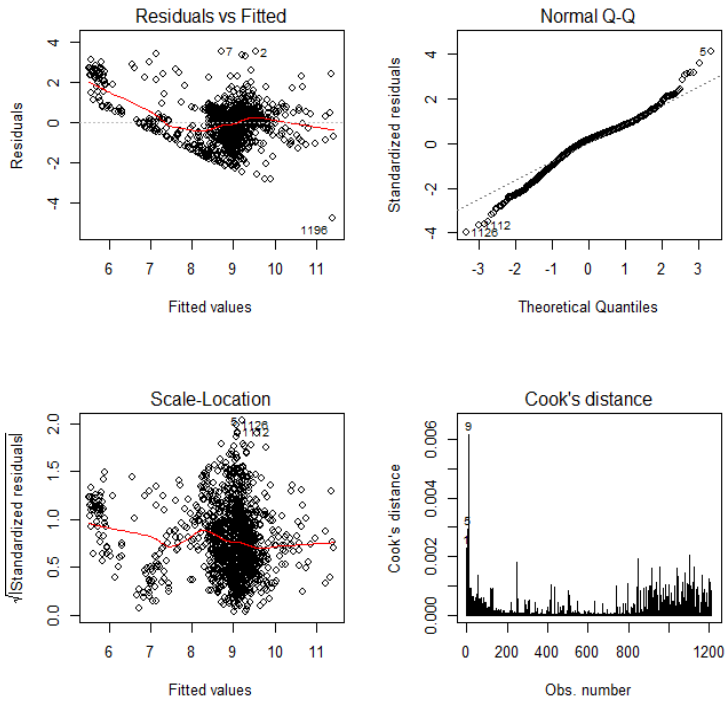

图 4-2：修正异方差后的回归模型诊断图

五、 结论与建议

本案例对 2016 年 12 月 27 日-2017 年 6 月 11 日之间发布的部分微信小程序数据进行统计分析，得到如下结论：

- 影响小程序人气值的主要因素有：（1）外部因素：评价人数；（2）内部因素：上线时长、所属类别。
- 小程序所属类别的人气值有所区别。工具类、健康类的小程序人气值最高，音乐、视频、体育类相比而言人气值较低。

由于微信小程序近一年的飞速发展，影响小程序人气的因素不断增多，因此在未来的研究中可以考虑在模型中加入更多可能的影响因素，比如小程序关联度（app、公众号等）、小程序的开发者属性（头部应用所属公司、一般人气公司、个人等）、小程序流量入口多样性（种类、个数等）。

第一次作业小组分工

一、项目分工：

分工	姓名
数据预处理、描述性分析	王嬰、栗书敬
模型建立、回归分析	张云华、张媛媛、张继丹
报告撰写、PPT 制作、课堂汇报	柳素问

二、任务量：

姓名	任务量
王嬰	数据清洗，提取文本变量信息；数据可视化展示
张云华	回归模型变量选择，变量形式选择和模型修正
柳素问	报告撰写、PPT 制作和课堂展示
张媛媛	对回归出现的异方差，非正态等问题尝试了多种方法
栗书敬	数据描述性分析，提供文本分析思路
张继丹	参与回归建模，对数据分析提供意见