UP 主粉丝数的影响因素分析

Group 2 周子逸 徐智昱 逄一哲 祝尔康

摘要

本文基于逻辑回归课上学习的多分类数据分析方法，对 B 站 UP 主粉丝数的影响因素进行分析。本文从 ifans 网站抽取 B 站 UP 主的粉丝数、性别、分区等相关数据，将 UP 主按照粉丝数分为 4 个水平，通过 EDA 总结变量特点并初步筛变量，之后分别使用列联表卡方检验、对数线性模型、多分类逻辑回归模型对数据进行分析解释，探究了 UP 主粉丝数水平与性别、分区、视频数等变量的关系，并试图结合各个模型对有意愿成为 UP 主的同学提供有关建议。

目录

1 背景介绍

Bilibili 视频网站 (下文简称 B 站) 在近年愈发受到年轻人的欢迎，数量庞大的 UP 主群体为 B 站的视频生态的多样性做出了非常大的贡献。在庞大的 UP 主群体之下，UP 主的粉丝数也有显著的差异。本研究旨在通过研究各种影响因素与 UP 主的粉丝数量的分层不同带来的影响。

本次研究的数据来源于 ifans 网站，网站上提供了粉丝量、最近更新时间、分区、视频数、充电数、近 8 篇平均视频投币数、近 8 篇平均视频弹幕数、近 8 篇平均视频收藏数、近 8 篇平均视频点赞数、近 8 篇平均视频播放数、近 8 篇平均视频评论数、近 8 篇平均视频分享数数据、性别数据，本研究将所有变量均放入模型中进行研究，同时研究是否能够有减少变量的方法。本研究凭借网站上的粉丝数量分区，将粉丝量的分区分为 “<10 万”，“10 万 50 万”，“50 万 100 万”，“>100 万” 四个分区，作为后续分类型数据分析的基础。

由于不同分区的 UP 主人数差异较大，本研究采用回溯性研究方法，即在四个粉丝量的分区分别抽取 250 个 UP 主的数据进行研究分析。

2 数据爬取与预处理

分别选择粉丝数量为 10 万以内、10 万到 50 万、50 万到一百万以及粉丝数量一百万以上的 up 主，排序方式使用随机排序，各采样 500 个 up 主（网页查询上限就是 500）。然后利用 singleFile 插件将网页中需要爬取的内容选中，保存为 html 文件。接着，利用 rvest 对 html 文件进行爬取，并将四种粉丝量的 up 主的数据分别整理为 xlsx 格式并与原始网页进行比对，检查爬取过程中是否出现错误。然后，分别读入 4 个 xlsx 格式的文件，剔除缺失性别等信息的数据，然后在各个粉丝档次剩下的 up 主中随机采样，样本量为 250。最终将 1000 个 up 主的数据整理在一起，就得到了大作业所用的数据集。在数据爬取的过程中遇到了一些困难，例如 up 主昵称没法爬取，最后采用了手动复制的方法解决。

由于每个变量的单位相差较大，为了避免单位造成的影响，本研究将所有非分类型数据进行归一化处理。由于选择的解释变量均为非负数据，为保持这一性质，采用以下的归一化方法。

$$\tilde{x} = \frac{x - \alpha}{\beta - \alpha} \tag{1}$$

其中 α,β 分别为本解释变量中的最大值与最小值。

3 探索性数据分析

3.1 数据可视化

首先，作出各类数据之间相关系数的热量图：

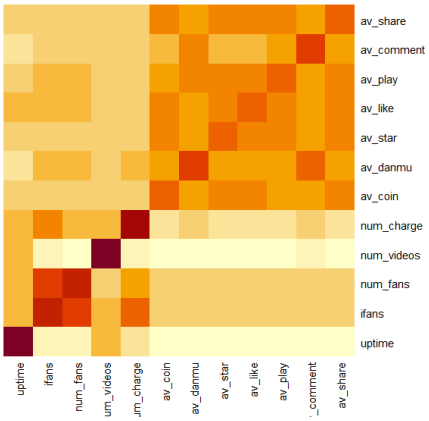

图 1: Heatmap for Correlation

由热量图可见，右上角各种平均的线性相关性较强。由于作出 Scatter Plot Matrix 时，如果变量过多，则容易显示不清楚，因此选择近 8 篇平均投币数和近 8 篇平均点赞数作为各种平均的代表绘制 Scatter Plot Matrix。

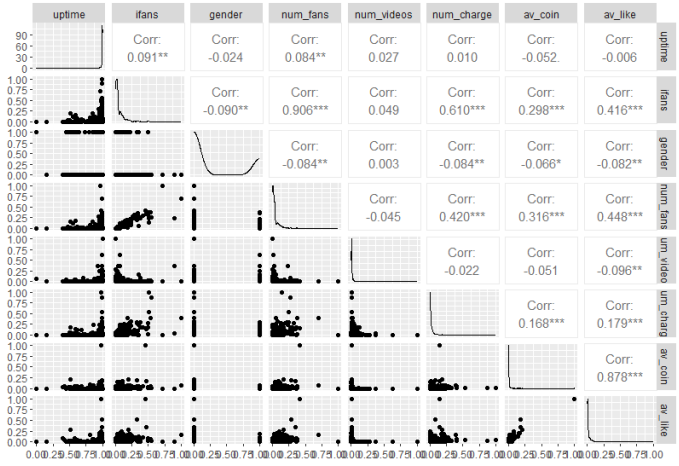

图 2: Scatter Plot Matrix

由 Scatter Plot Matrix 可见，各类数据间普遍具有一定的相关性，这说明对这些数据建立模型进行分析应该会有一定的效果。同时，由每个变量的核密度估计可见，各个解释变量是明显右偏的，这说明 B 站的 UP 主存在一定的少部分 UP 主占据了极大的资源的效应。

通过 Scatter Plot Matrix，作出两大分区变量粉丝量与性别和与这两个因素有显著作用的的变量之间的箱型图，观察这些因素对两大分类型变量的影响。

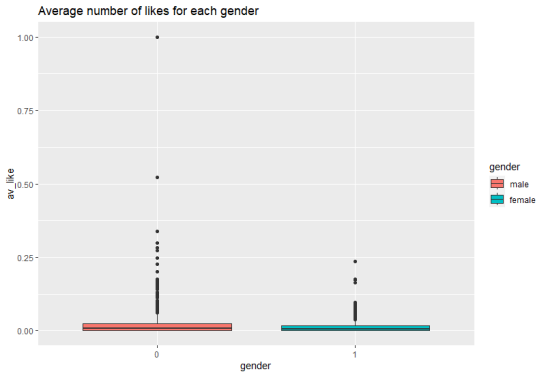

图 3: 各性别平均点赞数

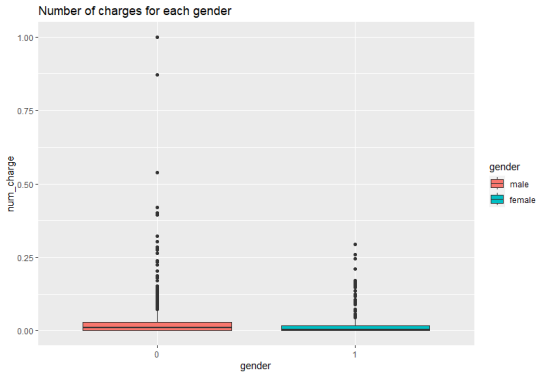

图 4: 各性别充电数

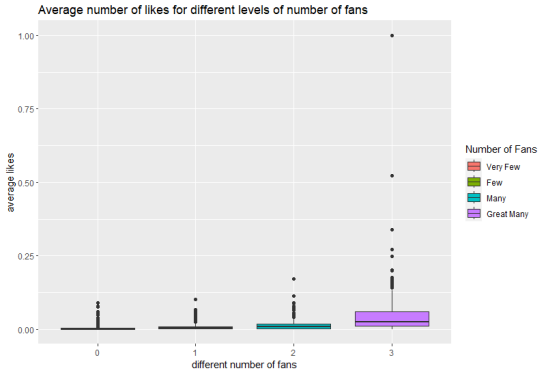

图 5: 各粉丝量平均点赞数

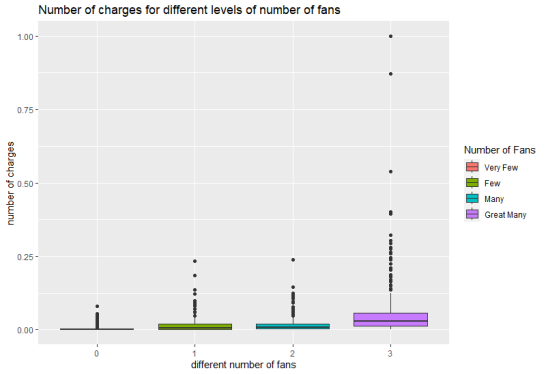

图 6: 各粉丝量充电数

由图可见，对于部分解释变量，其对不同性别确实有显著的影响；对于平均点赞数和充电数而言，均表现为男性的离群值点远大于女性的特点。而诸多解释变量对于粉丝量也有显著的作用，与上述分析相似的，对于粉丝量较少的三类，实际上两个变量的表现较为相似，而粉丝量大于 100 万的样本，在点赞量和充电数上，无论是平均值还是数量高的值都明显多于前三类。这映照了之前认为的少部分 UP 主获得了最多的关注的同时，也按时是否需要优化分区结构，能够获得更好的效果。

由于解释变量仍较多，并且从 Scatter Plot Matrix 可以看出部分解释变量之间存在较为显著的线性相关性，因此考虑是否能对解释变量进行一定程度的筛选。下面考虑利用主成分分析 (PCA) 和 LASSO 两种方法对是否能够减少解释变量数量进行分析。

3.2 主成分分析

对归一化后的连续型解释变量进行主成分分析，所得肘图如下图：

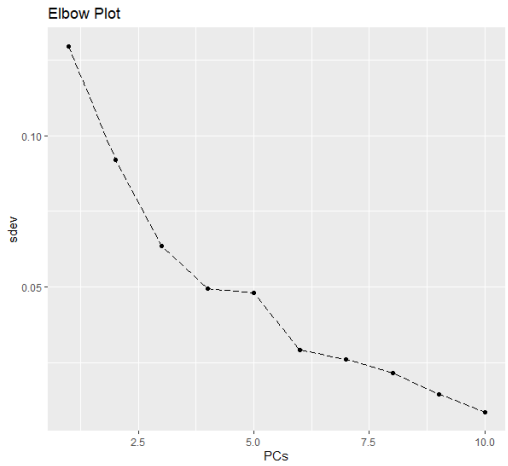

图 7: PCA 肘图

由肘图可见，实际上 PCA 的效果并不好，在拐点后方程下降速度仍较为显著，并且拐点处的方差仍较大。如果要保留 99% 的信息量，则可舍弃最后两个主成分；如果要保留 95% 的信息量，则可舍弃最后四个主成分。一定程度上起到了降维的效果。PCA 的旋转矩阵如下表:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
uptime	-0.17	0.97	0	0.11	-0.08	0.04	-0.03	0.01	-0.02	-0.01
videos	-0.04	0	-0.03	0.55	0.83	-0.04	-0.04	0	-0.01	0.01
charge	0.15	0.07	-0.93	-0.29	0.17	0.04	-0.01	0	-0.01	0
coin	0.23	0.07	0.08	-0.07	0.1	0.21	0.46	-0.36	0.33	-0.65
danmu	0.51	0.01	-0.17	0.57	-0.35	-0.19	0.37	0.3	-0.01	0.08
star	0.3	0.08	0.16	-0.2	0.16	0.09	0.31	-0.26	-0.78	0.17
like	0.32	0.12	0.14	-0.19	0.14	-0.13	0.09	-0.32	0.51	0.65
play	0.48	0.12	0.14	-0.19	0.11	-0.58	-0.47	0.06	-0.05	-0.35
comment	0.31	-0.03	-0.07	0.33	-0.21	0.49	-0.56	-0.44	-0.05	0.01
share	0.34	0.08	0.2	-0.23	0.2	0.56	-0.07	0.65	0.1	0.03

表 1: PCA 旋转矩阵

由旋转矩阵可见，实际上 PCA 的解释性较差。又由于本研究主要为解释性研究，因此如果采用 PCA 后的主成分进行分析，最后会导致解释性较差，所以将舍弃 PCA 部分的结果。

3.3 LASSO

利用 LASSO 将所有解释变量放入模型中，分别对粉丝量 (分类型) 作多分类变量逻辑回归与对粉丝量 (连续型) 作广义线性回归，得到均方误差随正则项的变化分别如下两图。

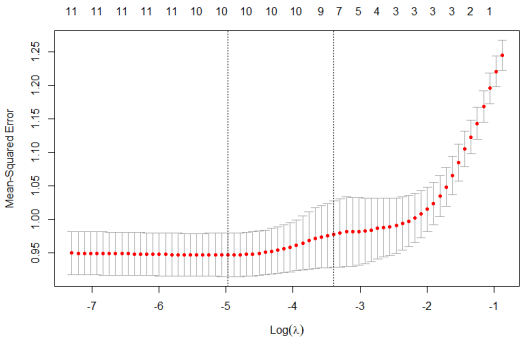

图 8: 多分类逻辑回归

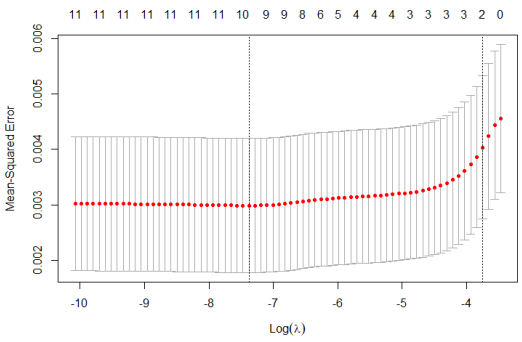

图 9: 连续型线性回归

所得在让均方误差最小的正则项系数与在 1 个标准差意义下的正则项系数如下表，其中进行多分类逻辑回归得到 4 个返回的 λ 值，筛选出效果较好的值对应的系数：

Name	Multinomial		Continuous	
	min	1se	min	1se
Intercept	8.76	5.26	-0.03	0.04
uptime	-8.16	-5.69	0.05	0
gender	0.08	0.15	0	0
videos	-1.08	0	0	0
charge	-54.93	-20.74	0.34	0.05
coin	0	0	-0.3	0
danmu	0	0	0.09	0
star	8.08	0	-0.11	0
like	-73.91	-21.17	0.66	0
play	-13.79	0	0.2	0.1
comment	-1.07	0	-0.06	0
share	22.47	0	-0.16	0

表 2: 正则后系数表

由上表可见，使用 LASSO 多类别逻辑回归和线性回归得到的系数有较为明显的区别，因此说明经过将粉丝量分类，得到了不同的回归结果，有不同的解释效应。因此，在进行多分类逻辑回归会得到直接线性回归不同的结果。由于两个 1 个标准差内的结果都省略了过多的变量，可能造成解释能力不足，因此在进行多分类逻辑回归时，考虑舍弃投币数和弹幕数两个变量。

4 列联表分析

为了探索 UP 主粉丝数水平与分区和性别的关系，我们对其构成的三维列联表尝试了卡方检验、累计逻辑回归和 log linear model，得到了三个有意义的结论。

4.1 列联表卡方检验

首先我们利用 pearson 检验分别对无分区变量的总表和考虑分区的三个子表测试变量间的独立性。在总表中得到的卡方值为 10^{-16}，说明粉丝数水平与性别显著相关；在动画、游戏区中的卡方值分别为 0.6172, 0.4575，这说明可以认为粉丝数水平和性别独立；而在知识分区中卡方值则为 0.02，这说明有 95% 以上的把握认为粉丝数水平与性别有关。

4.2 累积逻辑回归模型

为了更好地利用粉丝数水平作为次序变量的特点，我们采用了适用于这种情形的累积逻辑回归模型。下面多分类逻辑回归部分也会使用到该模型，区别在于后面针对的解释变量更多且大多为连续型，而本处的解释变量是性别和分区两个分类型变量。

模型的响应变量为 Fanslevel，是一个 ordinal 的变量，代表 up 主粉丝数水平类别，分为 4 类，“<10w”、“10w-50w””、“50w-100w”、“>100w” 分别规定为 $j = 1, 2, 3, 4$，这里 “>100w” 默认为 1，所以 j 只可能取 1,2,3。解释变量为性别和分区，其中性别变量 gender 是一个 binary 的变量，1 代表女性，0 代表男性；分区我们考虑了知识区、动画区和游戏区，使用 type1 和 type2 两个哑变量来表示分区变量，(type1,type2)=(0,0) 对应知识区，(type1,type2)=(1,0) 对应动画区，(type1,type2)=(0,1) 对应游戏区。全模型如下：

$$\text{logit}[P(fanslevel\ j|type)] = \alpha_j + \beta_g * gender + \beta_1 * type1 + \beta_2 * type2 + \beta_{g1} * gender * type1 + \beta_{g2} * gender * type2$$

对于全模型，我们求得各参数分别为：$\alpha_1 = 1.844, \alpha_2 = 3.467, \alpha_3 = 4.346, \beta_g = 0.049, \beta_1 = 0.145, \beta_2 = 0.270, \beta_{g1} = -0.006 = 7, \beta_{g2} = -0.087$。但是只有截距项和 β_1, β_2 较为显著。我们利用 drop1 函数比较 AIC 来减少模型中的变量数，并利用 anova 来检验丢弃变量操作的合理性。最终我们选定只有 type1 和 type2 两个解释变量的模型

$$\text{logit}[P(fanslevel\ j|type)] = \alpha_j + \beta_1 * type1 + \beta_2 * type2$$

模型残差为 14.700，与全模型做方差分析得到 p 值为 0.513，说明简化模型并未损失太多信息。简化后的各参数分别为：$\alpha_1 = 1.858$, $\alpha_2 = 3.481$, $\alpha_3 = 4.360$, $\beta_1 = 0.144$, $\beta_2 = 0.248$。该模型中性别和性别的交互项都不显著与我们此前的列联表卡方检验结果相呼应，表明性别在各个分区看来不是一个粉丝数水平的显著决定因素，此前不考虑分区的结果正是明显的辛普森悖论现象。

对于 β_1 的理解为：$\text{logit}[P(fanslevel\ j|type1 = 1)] - \text{logit}[P(fanslevel\ j|type1 = 0)] = 0.144$，即 fanslevel 关于 type1 的 odds ratio 为 1.155。这表明在动画区的某一粉丝数水平 (水平 j) 以下的概率对应的 odds，为在知识区的同一粉丝数水平 (水平 j) 以下的概率对应 odds 的 1.155 倍，这反映了在动画区聚集在相对低级粉丝数水平的 up 主比例高于知识区。同理，

对于 β_2 的理解为：$\text{logit}[P(fanslevel\ j|type2 = 1)] - \text{logit}[P(fanslevel\ j|type2 = 0)] = 0.248$，即 fanslevel 关于 type1 的 odds ratio 为 1.281。这表明在游戏区的某一粉丝数水平 (水平 j) 以下的概率对应的 odds，为在知识区的同一粉丝数水平 (水平 j) 以下的概率对应 odds 的 1.281 倍，这反映了在游戏区聚集在相对低级粉丝数水平的 UP 主比例高于知识区。因为 $\beta_2 > \beta_1$，我们可以进一步认为在游戏区聚集在相对低级粉丝数水平的 UP 主比例高于动画区，这意味着游戏区竞争更为激烈。

观察实际的各 UP 主分布情况如下表所示，可以看出实际情况确实如此，但与低粉丝数 UP 主比例高相对应的是游戏区各水平 UP 主的数目都明显多于动画、知识分区，这一点将在 log linear model 中进一步体现。

动画区粉丝数占比	<10w	10w-50w	50w-100w	>100w
所占比例	0.881	0.093	0.0148	0.011

表 3: 动画区各粉丝数水平占比

游戏区粉丝数占比	<10w	10w-50w	50w-100w	>100w
所占比例	0.892	0.085	0.013	0.010

表 4: 游戏区各粉丝数水平占比

知识区粉丝数占比	<10w	10w-50w	50w-100w	>100w
所占比例	0.865	0.015	0.0148	0.011

表 5: 知识区各粉丝数水平占比

4.3 对数线性模型

为充分利用 UP 主粉丝数是 count 型变量的特点以及分区和性别两个解释变量之间可能存在的关系，考虑使用 log linear model 进行建模。模型如下:

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ki}^{ZX} + \lambda_{ijk}^{XYZ} \tag{2}$$

其中 X, Y, Z 分别对应粉丝数水平 fanslevel，性别和分区，各参数含义如下:

参数名	参数含义
λ	截距项，代表基线水平。这里基线为 50w-100w 粉丝量的动画区男性 up 主
λ_i^X	单独代表 X(粉丝量水平) 对人数的作用效果，“<10w”“10w-50w”“50w-100w”“>100w” 分别规定为 $j = 1, 2, 0, 3$，对应为 little,middle,big,superbig，这里以 big 为基线水平。
λ_j^Y	单独代表 Y(性别) 对人数的作用效果，$j = 0, 1$ 分别代表男性，女性
λ_k^Z	单独代表 Z(分区) 对人数的作用效果，动画、游戏、知识区分别对应 $Z = 0, 1, 2$。动画区为基线水平。
λ_{ij}^{XY}	代表粉丝量水平、性别变量的协同作用效果
λ_{jk}^{YZ}	代表性别、分区变量的协同作用效果
λ_{ki}^{ZX}	代表粉丝量水平、分区变量的协同作用效果
λ_{ijk}^{XYZ}	代表粉丝量水平、性别和分区变量的协同作用效果

表 6: 对数线性模型的参数解释

认为下标出现 0 的系数对应为 0，如 $\lambda_0^X = \lambda_j^Y = \lambda_k^Z = \cdots = 0$。由于全模型过于复杂，我们利用 drop1 函数在 AIC 减小不是很大的情况下删除了三交互项 (λ_{ijk}^{XYZ}) 和性别与粉丝数水平的二交互项 (λ_{ij}^{XY})，最终模型仅剩 15 个参数，参数估计如下：

λ	λ_1^X	λ_2^X	λ_3^X	λ_1^Y	λ_1^Z	λ_2^Z	λ_{11}^{XZ}
5.0	4.1	1.8	-0.3	-0.7	0.7	0.4	0.1
λ_{21}^{XZ}	λ_{31}^{XZ}	λ_{12}^{XZ}	λ_{22}^{XZ}	λ_{32}^{XZ}	λ_{11}^{YZ}	λ_{12}^{YZ}	
0.002	-0.02	-0.2	-0.05	-0.06	-5.4	-0.2	

表 7: 对数线性模型的拟合参数

所有变量参数中，截距项为 λ；所有粉丝数水平变量 λ_1^X，λ_2^X，λ_3^X；所有性别变量 λ_1^Y；所有分区变量 λ_1^Z λ_2^Z；小粉丝量水平和知识分区的二交互项 λ_{12}^{XZ}；所有性别和分区的二交互项 λ_{11}^{YZ}，λ_{12}^{YZ} 均效应显著。利用 R 进行检验，该模型拟合效果很好，最大的标准化残差只有 2.93。但若进一步删除不显著的变量，1/3 的拟合结果会出现绝对值大于 3 的标准残差。

fits1	adjresids1
8699.59	-0.74
4256.41	0.74
918.57	0.39
449.43	-0.39
146.38	-0.06
71.62	0.06
109.45	1.25
53.55	-1.30
20412.80	0.91
5824.20	-0.91
1946.60	-1.39
555.40	1.38
309.65	0.16
88.35	-0.16
227.96	0.85
65.04	-0.87
10632.87	-0.83
4366.13	0.83
1290.92	-0.60
530.08	0.59
215.51	1.20
88.49	-1.23
151.71	2.72
62.29	-2.93

图 10: 对数线性模型拟合结果

三交互项参数的不显著说明模型三个变量具有同质性，即任意两个变量的 odds ratio 不受第三个变量的影响；性别与粉丝数水平二交互项参数的不显著说明在给定分区的情况下两个变量间是独立的，这与累计逻辑回归的结论是一致的；粉丝量水平与知识分区的二交互项中仅有小粉丝量水平和知识分区的二交互项显著，且参数绝对值普遍较小，这说明在给定性别的情况下，除了知识区的小粉丝量和大粉丝量水平 UP 主数目有明显差异，其他都区别不大。

继续观察各参数的正负，我们可得到以下结论：

(1) 小、中粉丝量水平 (<10w 和 10w-50w) 参数为正，超大粉丝量水平 (>100w) 参数为

负。这说明性别、分区固定时，粉丝数越多的 up 主对应的人数越少，呈单调递减。

(2) $\lambda_1^Y < 0$，这反映了男性 UP 主数量明显多于女性 UP 主。

(3) $\lambda_1^Z = 0.749 > \lambda_2^Z = 0.387 > 0$，这说明游戏分区人数 > 知识分区人数 > 动画分区人数。

(4) 小粉丝量水平和知识分区的二交互项系数为 $-0.186108 < 0$，说明：(i) 固定分区为知识区和性别时、<10w 粉丝数水平相比 50w-100w 水平 UP 主约为 0.83 倍；(ii) 或者固定性别、固定粉丝数水平为 <10w，知识分区该类 UP 主人数约为动画分区的 0.83 倍。

(5) 性别和游戏分区的二交互项系数为 $-0.539 < 0$，说明：(i) 固定性别为女和粉丝数水平时，游戏分区该类 UP 主的数目约为动画分区的 0.58 倍；(ii) 固定分区为游戏分区并固定粉丝数水平时，女性 UP 主中的该类别 UP 主为男性中该类 UP 主的 0.58 倍。

(6) 性别和知识分区的二交互项系数为 $-0.175 < 0$，说明：(i) 固定性别为女和粉丝数水平时，知识分区中该类 UP 主数目为动画分区中的 0.84 倍；(ii) 固定在知识分区并固定粉丝数水平，女性 UP 主中该类别 UP 主数目为为男性中该类 UP 主的 0.84 倍。

5 多分类逻辑回归分析

为了分析性别以及点赞数、充电数这些连续型变量与 UP 主粉丝数分区的关系，我们采用多分类逻辑回归模型进行建模分析。

5.1 变量选取

根据前面部分可知，我们按照 “<10 万”，“10 万 50 万”，“50 万 100 万”，“>100 万” 的标准将 up 主按照粉丝数分成了 4 类，我们将这 4 个 UP 主量级分别记为小 UP 主，中 UP 主，大 UP 主和超级 UP 主，得到响应变量如下：

响应变量	变量类型	变量含义
fans_cat	Multicategory	粉丝数的分类，small, middle, big, super big 四类

表 8: 多分类逻辑回归响应变量

对于解释变量，我们一共有 1 个二分类的性别变量和 9 个归一化后的连续型变量，各个变量的含义如下标所示，根据 LASSO 部分的分析，我们可以去掉 av_coin 和 av_danmu 两个变量，选择余下 8 个解释变量作为建模最初的 8 个变量。

解释变量	变量类型	变量含义
gender	binary	性别, 1 代表女性, 0 代表男性
num_videos	continuous	视频数的多少, 单位为个
num_charge	continuous	充电数的多少, 单位为个
av_star	continuous	最近 8 个视频的平均收藏量, 单位为个
av_like	continuous	最近 8 个视频的平均点赞数, 单位为个
av_play	continuous	最近 8 个视频的平均播放数, 单位为次
av_comment	continuous	最近 8 个视频的平均评论数, 单位为个
av_share	continuous	最近 8 个视频的平均分享数, 单位为个
~~av_coin~~	continuous	最近 8 个视频的平均投币量, 单位为个
~~av_danmu~~	continuous	最近 8 个视频的平均弹幕数量, 单位为个

表 9: 多分类逻辑回归解释变量

为初步地观察解释变量与 UP 主粉丝数分类的关系，我们在四个分类中各随机抽取 75 个 UP 主，绘制出这 300 个 UP 主的并行坐标图如下图所示，图中每一条线就是一个 UP 主在各解释变量的坐标，UP 主的粉丝数分类由线的颜色和线型表示。从图中我们可以看出超级 UP 主的充电数，近期视频的充电数、平均点赞数、播放量和评论数都趋向于最大，而有一些小 UP 主的近期平均收藏量在抽取的 UP 主中可以达到最大。

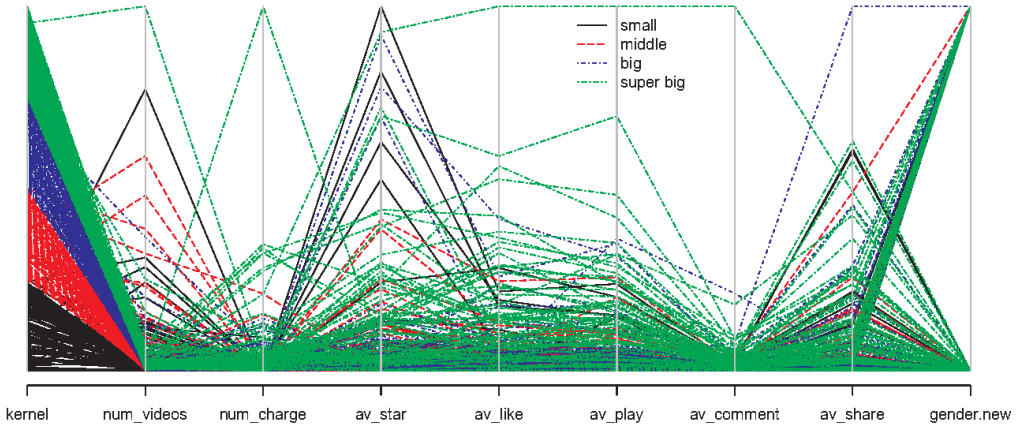

图 11: 并行坐标图

5.2 建立模型

由于响应变量 Y 是 ordinal 的, 记响应变量 fans_cat 为 Y, 四个粉丝量级与 Y 的对应关系为：small(Y=1)<middle(Y=2)<big(Y=3)<super big(Y=4)，我们对各水平的累积概率的 logit 采用相同的斜率，即采用“平行”的累积逻辑回归模型，模型如下:

$$\text{logit}(P(Y \le j)) = \alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p,\ j = 1, 2, 3 \tag{3}$$

将响应变量和 8 个解释变量放入 R 中得到最初模型，其 Anova 如下图所示，可以看到性别、视频数量和近期视频平均评论数这 3 个解释变量对应的系数均不显著。

```
Response: fans_cat
           LR Chisq Df Pr(>Chisq)
gender        1.460  1  0.2269955
num_videos    0.562  1  0.4533397
num_charge  155.058  1  < 2.2e-16 ***
av_star      10.988  1  0.0009171 ***
av_like      29.383  1  5.939e-08 ***
av_play      19.841  1  8.417e-06 ***
av_comment    0.358  1  0.5496833
av_share     18.951  1  1.341e-05 ***
```

图 12: Anova of first model

```
Response: fans_cat
           LR Chisq Df Pr(>Chisq)
num_charge  168.658  1  < 2.2e-16 ***
av_star      11.893  1  0.0005636 ***
av_like      28.195  1  1.097e-07 ***
av_play      21.509  1  3.522e-06 ***
av_share     21.003  1  4.585e-06 ***
```

图 13: Anova of final model

采用 Backward Elimination 的变量选取方法，并以 AIC 为准则，可以依次去除近期评论数, 视频数, 性别这 3 个解释变量，将剩余 5 个解释变量拟合得到最终模型，从图 12 可以看到最终模型中 5 个解释变量均显著。

5.3 模型解释

利用 R 中的 vglm 函数得到最终的估计模型如下:

$$\text{logit}(\hat{P}(Y \leq j)) = \hat{\alpha}_j - 29.4num\_charge + 16.8av\_star - 48.6av\_like - 19.4av\_play + 15.6av\_share$$

其中 $\hat{\alpha}_1 = -0.2, \hat{\alpha}_2 = 1.3, \hat{\alpha}_3 = 2.9$

我们注意到估计模型中 β_{num_charge}, $\beta_{av\_like}, \beta_{av\_play}$ 均小于 0，说明 num_charge，av_like，av_play 高的 up 主更倾向于是量级高的 UP 主，而 β_{av_star}, β_{av_share} 均大于 0，说明 av_star 和 av_share 高的 UP 主倾向于是量级低的 UP 主，这似乎与直觉不太相符，但由于这只是近 8 个视频的指标，所以我们可以猜测这些视频可能只是通过“玩梗”等方式在短期获得了较大的关注，或者是新 UP 主。

结合整个建模过程的分析，我们可以得到以下结论:

(1) 首先性别变量不显著，说明 UP 主粉丝数量级可能与性别关系不大.

(2) 一些高质量视频的正向收益指标 (点赞数，充电数) 显著，而总视频数并不显著，说明相较于视频数量，UP 主粉丝数量级与其视频质量的关系更大。说明想成为一名大 Up 主，视频在精不在多.

(3) 从拟合模型的系数上看，充电量、点赞量、播放量高的更倾向于是量级高的 UP 主，而近期视频收藏量，分享量高的更可能是“昙花一现”。

5.4 数据可视化

为了便于数据可视化，我们系数最显著的解释变量充电量 num_charge 单独拟合累积逻辑回归模型得到概率估计图如下，左侧是各个粉丝量级的概率分布，右侧是粉丝数量级的累积概率。

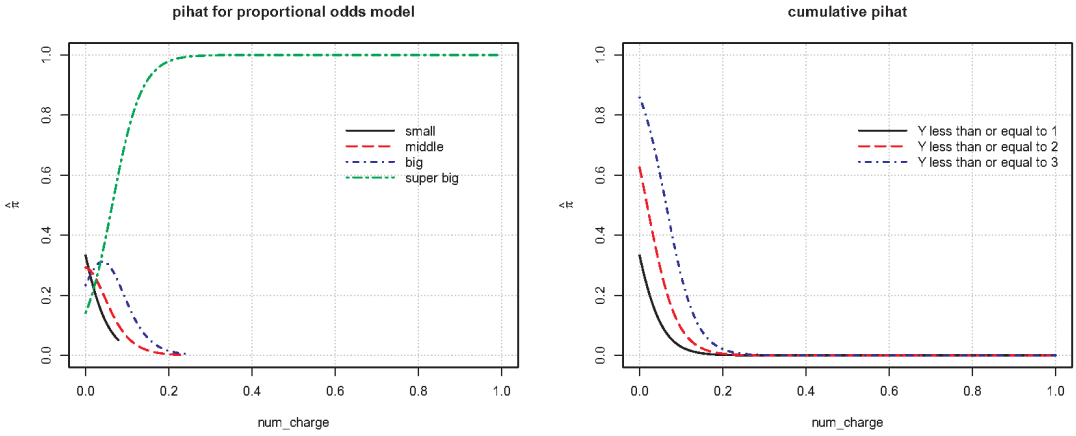

图 14: 以充电量为解释变量的累积逻辑回归模型概率分布图和累积概率图

从左图中我们可以看出充电量在超级大 UP 主和非超级 UP 主之间分布非常不均匀，归一化后的充电量约为 0.23 以上时，UP 主是超级 UP 主的概率就为 1 了；从右图可以看出并且随着充电量的增加，小，中，大量级对应的累积概率都递减，说明充电量稍微大一些，就很可能是超级 UP 主了，换言之，归一化后 0.23 充电量以上的 UP 主中基本都是超级 UP 主，这也反映了一种超级 UP 主和非超级 UP 主之间的“贫富差距”大的两极分化现象。

事实上，我们本身抽取 UP 采用回溯性研究，每个量级的 UP 主都抽取了 250 位，但实际上 100 万粉以上的超级 UP 主的数量相较于 100 万粉以下的小中大 UP 主的数量是少得可怜的 (列联表部分给出数据)，但在抽取数据中充电量 23% 值最大值以上都是超级 UP 主，而充电量本身就是 UP 主获得收益的一个主要渠道之一，可见少数的头部 UP 主获得远多于非头部 UP 主的收益，这体现了 B 站 UP 主的一种金字塔效应。

5.5 模型改进

在多分类逻辑回归分析部分，我们使用了“平行”的累积回归模型，对于最初的 8 解释变量模型，AIC=2273.2；对于最后的 5 解释变量模型，AIC=2269.1。我们发现最终得到的模型的 AIC 依旧在 2000 以上，说明模型并不是很理想。我们知道，在使用“平行”的累积回归模型时，我们假定了各解释变量对响应变量各水平的累积变量的影响是一致的，这个假设是否太强了，导致模型拟合不佳? 我们因此尝试了“非平行”的累积逻辑回归模型，当解释变量为 8 个时，“非平行”逻辑回归模型的 $3 \times (8+1) = 27$ 的参数均显著，但是注意到“非平行”模型的 Log-likelihood 为 NA，经过查阅资料和分析发现，在“非平行”模型情况下出现了累积概率“交叉”的情况，此时各粉丝量水平概率估计值出现负值，因此导致模型失效，因此“非平行”模型在本数据中不适用。

我们可以采取多项累积回归模型作为“非平行”累积回归模型的一种替代，看看是否复杂一些的模型有更好的效果，模型如下：

$$\log(\pi_j/\pi_{small}) = \alpha_j + \beta_{j1}x_1 + \cdots + \beta_{jp}x_p, j = middle, big, super\ big \tag{4}$$

经过 Anova 分析 (下图左) 发现模型的视频数变量不显著，建立去除后视频数的模型，进行卡方检验，p-value=0.44，说明可以保留简便的模型，即去除视频数的模型，我们由此得到一个 7 解释变量的模型，得到的 Anova 表 (下图右) 中所有变量均显著。

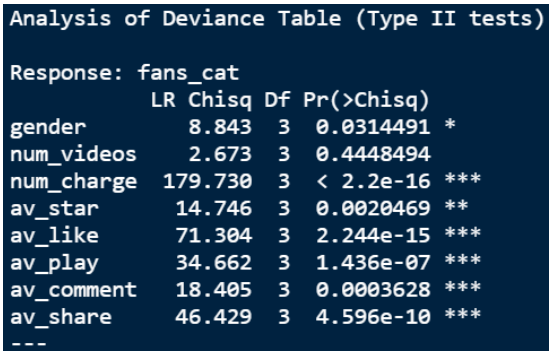

图 15: Anova of first model

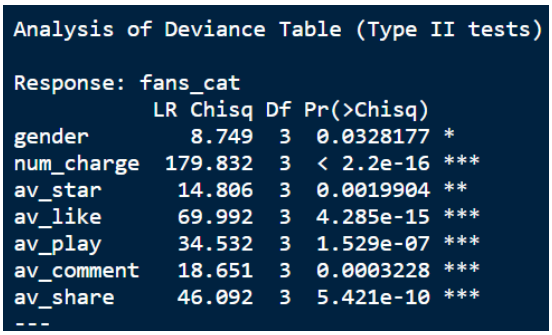

图 16: Anova of final model

得到的多项逻辑回归模型估计系数如下：

	Intercept	gender	num_charge	av_star	av_like	av_play	av_comment	av_share
$\log(\frac{\hat{\pi}_{middle}}{\hat{\pi}_{small}})$	-0.7	-0.5	58	-12	84	8.0	-5.1	-22
$\log(\frac{\hat{\pi}_{big}}{\hat{\pi}_{small}})$	-1.3	0.01	61	-27	121	18	13	-42
$\log(\frac{\hat{\pi}_{superbig}}{\hat{\pi}_{small}})$	-2.7	-0.1	81	-28	129	39	1.5	-46

表 10: 多分类逻辑回归解释变量

从上表中我们可以得到分析如下：

(1) 由于解释变量是归一化的，系数的绝对值可以一定程度反映解释变量对响应变量不同分类之间 odds 的影响，在 3 个水平下，性别和近期平均评论数的系数绝对值均较小，说明它们对响应变量影响较小，这与累积逻辑回归模型部分的结果是一致的。

(2) 充电数和近期平均点赞数的系数在 3 个 level 下为正且较大，说明随着充电数和近期平均点赞数的增大，UP 主位于中、大、超级 UP 主相较于小 UP 主的 odds 均增大，并且这种增大的效应比较显著。同时充电数和近期平均点赞数的系数满足 super big>big>middle，说明这种 odds 的增大效应随着 UP 主量级的增大而增大。

(3) 近期平均收藏量，近期平均分享量的系数为负，随着近期平均收藏量，近期平均分享量的增大，UP 主位于中、大、超级 UP 主相较于小 UP 主的 odds 均减小，这也与直觉不太相符，我们在累积逻辑回归部分已试图分析过其原因。

最终得到的多项逻辑回归的 AIC=2208.4，略优于“平行”的逻辑回归模型，但也并不太理想。并且从可解释性上看，似乎引入这种更为复杂的模型并没有给我们带来更多的可解释性，因此综合各方面看，我们认为采用“平行”逻辑回归模型来刻画 UP 主粉丝数的分类与各视频指标的关系是最优的。

6 研究结论

经过实际情况与以上的数据分析，我们可以总结出三点成为 B 站头部 up 主的秘诀。

(1) 首先是要尝试自己擅长的领域，因为想要成为一名顶尖 up 主更多取决于视频质量的高低而非视频数量的多少，而在自己擅长的领域更容易做出高质量的视频。

(2) 如果不知道成为哪个区的 up 主更容易收获粉丝，我们建议你可以从游戏区开始，因为游戏区的投递视频门槛相对较低，游戏选择合适的话更容易吸引对应玩家群体观看。

(3) 要保证视频质量持久稳定而非“昙花一现”，这样才能收获长期支持的真爱粉，足够多的真爱粉使视频保持长期持久的关注度，这将是成为头部 up 主的基础。

7 研究缺陷与改进

7.1 分区方式

在列联表研究中，我们只选择了三个分区为代表进行研究。我们的本意是想要建立有关全部分区的研究，但是受制于我们抽取数据的方式，如果直接进行更多分区的列联表研究，会有同一个 UP 主同时设计多个分区的情况，尤其在粉丝量大的部分，重复比例会相当严重，从而造成研究的结果不准确。因此我们尽量选择了重合较少的 UP 主进行分类数据处理，让分析的结果尽量准确。

因此，在后续可以考虑优化数据获取的方式，如爬下所有数据后进行去重 (但我们抽取数据的网站对其有限制) 等更加优化的抽取数据方式，来得到更多分区下的联表研究。

同时，从 EDA 中可以看出，可能粉丝量的分类存在一定不合理的方式，因为在粉丝量较小的情况下，UP 主的各个指标没有表现出明显的差异，因此放入模型中会让模型较为不显著，在回归时也会降低回归的效益。因此后续可以通过调整如何对粉丝量进行分区，来找到 UP 主的粉丝量的更显著的“台阶”。

7.2 多分类回归与线性回归

从 LASSO 的结果可以看出，多分类回归和线性回归得到的系数是不同的，因此这提示我们在进行粉丝量分区处理时，出现了一定的信息损失，或者说变异。从解释性上，可以说是由于粉丝量上一层与在全区间意义下的粉丝量的提升相关的变量存在不同重要性的影响因素。但是这种信息造成的偏差也是值得注意的。

进行了线性回归分析后，发现直接的线性回归对数据的拟合并不好，因此为更好地拟合相关的数据，可以考虑效益更加强的其它方法进行数据的处理。

当然，也应该发现的是其实二者的相似性还是比较强的，比如线性回归被正则的参数在多分类逻辑回归中的效益也较低，最后在筛选变量后被去除。因此也一定程度上说明了此种分类方式后的逻辑回归和不分类的线性回归有较强的相同的效益。

分工与致谢

在本次大作业中，周子逸负责数据的爬取和清洗，徐智昱负责数据预处理，EDA 以及研究缺陷和改进部分，逄一哲负责列联表分析部分，祝尔康负责多分类逻辑回归部分及论文整合，4 位同学都参与了 pre 的准备以及论文的撰写。

最后，感谢王江典老师和王梓涵助教对本次大作业的指导和帮助，感谢您指引我们走进多分类数据分析的殿堂，并在百忙之中为我们的研究提供指导性意见。

附录

代码，数据集，绘图均在如下 GitHub 链接中：
https://github.com/zekkr/CDA_Weiyang