

UP 主粉丝数的影响因素分析

CDA 小组

摘要

这里是摘要

目录

1	背景介绍	2
2	数据预处理	2
3	探索性数据分析	2
3.1	数据可视化	2
3.2	主成分分析	4
3.3	LASSO	4
4	多分类逻辑回归分析	6
4.1	变量选取	6
4.2	建立模型	8
4.3	模型解释	8
4.4	数据可视化	9
4.5	模型改进	9
5	研究结论	11
6	研究缺陷与改进	11
6.1	分区方式	11
6.2	多分类回归与线性回归	11

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 数据预处理

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。

$$\tilde{x} = \frac{x - \alpha}{\beta - \alpha} \quad (1)$$

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:

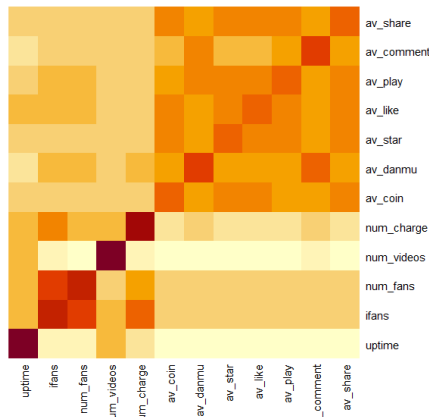


图 1: Heatmap for Correlation

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。

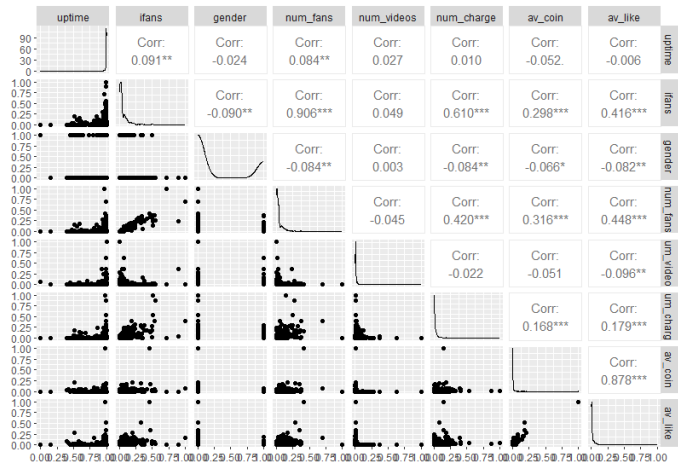


图 2: Scatter Plot Matrix

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。

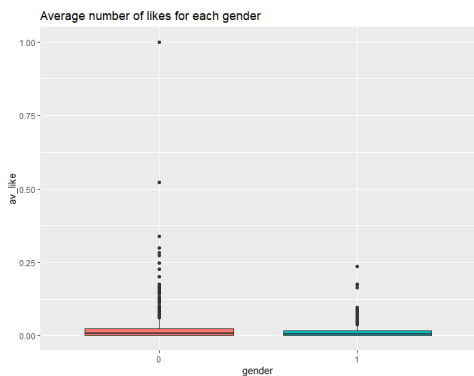


图 3: 各性别平均点赞数

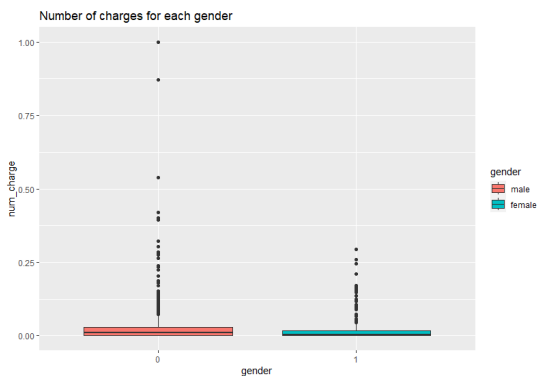


图 4: 各性别充电数

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

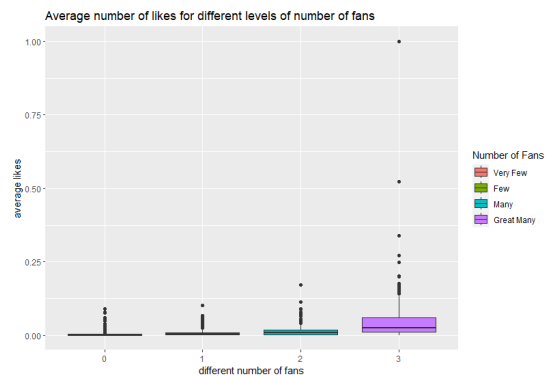


图 5: 各粉丝量平均点赞数

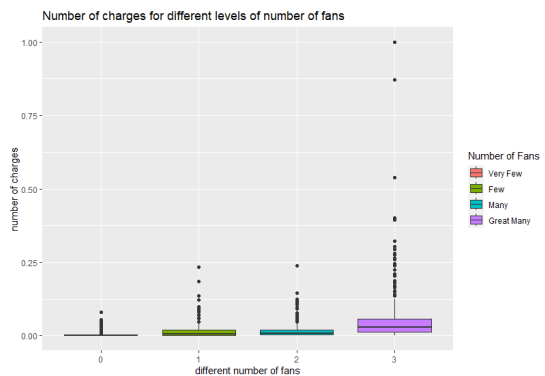


图 6: 各粉丝量充电数

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：

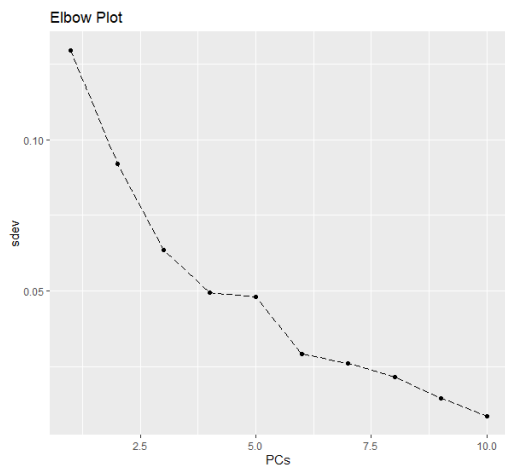


图 7: PCA 肘图

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：

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。

	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

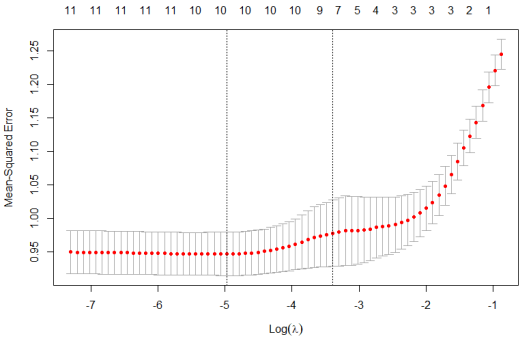


图 8: 多分类逻辑回归

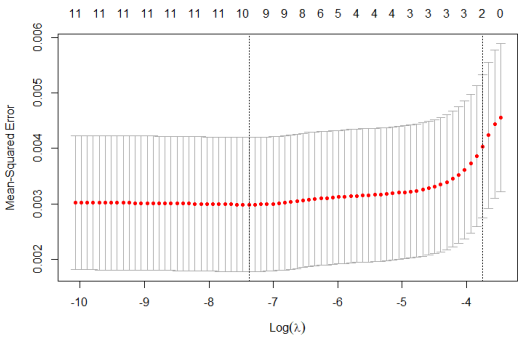


图 9: 连续型线性回归

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 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。

4.1 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 3: 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

变量的含义如下表所示，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
<code>gender</code>	binary	性别, 1 代表女性, 0 代表男性
<code>num_videos</code>	continuous	视频数的多少, 单位是个
<code>num_charge</code>	continuous	充电数的多少, 单位是个
<code>av_star</code>	continuous	最近 8 个视频的平均收藏量, 单位是个
<code>av_like</code>	continuous	最近 8 个视频的平均点赞数, 单位是个
<code>av_play</code>	continuous	最近 8 个视频的平均播放数, 单位为次
<code>av_comment</code>	continuous	最近 8 个视频的平均评论数, 单位是个
<code>av_share</code>	continuous	最近 8 个视频的平均分享数, 单位是个
<code>av_coin</code>	continuous	最近 8 个视频的平均投币量, 单位是个
<code>av_danmu</code>	continuous	最近 8 个视频的平均弹幕数量, 单位是个

表 4: 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。

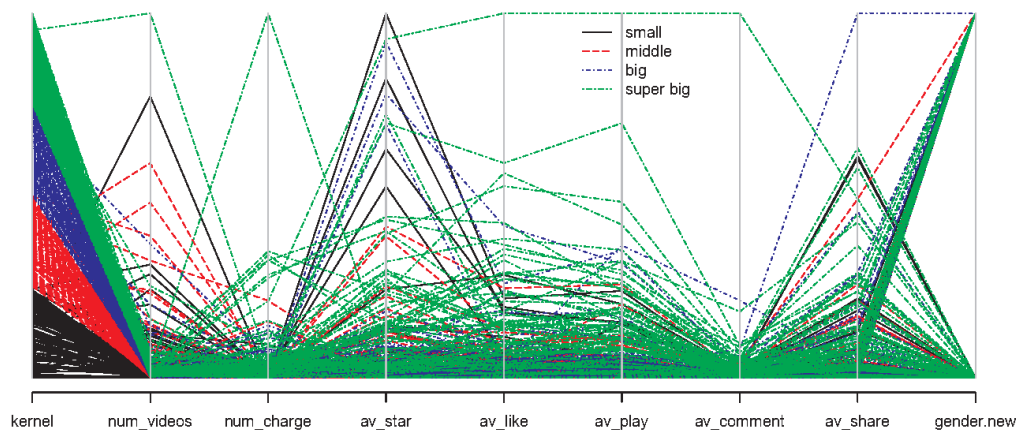


图 10: 并行坐标图

4.2 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:

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

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。

	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 11: 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 12: 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。

4.3 模型解释

利用 R 中的 vglm 函数得到最终的估计模型如下:

$$\text{logit}(\hat{P}(Y \leq j)) = \hat{\alpha}_j - 29.4\text{charge} + 16.8\text{star} - 48.6\text{like} - 19.4\text{play} + 15.6\text{share}$$

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 $\beta_{\text{num_charge}}, \beta_{\text{av_like}}, \beta_{\text{av_play}}$ 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 $\beta_{\text{av_star}}, \beta_{\text{av_share}}$ 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”。

4.4 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。

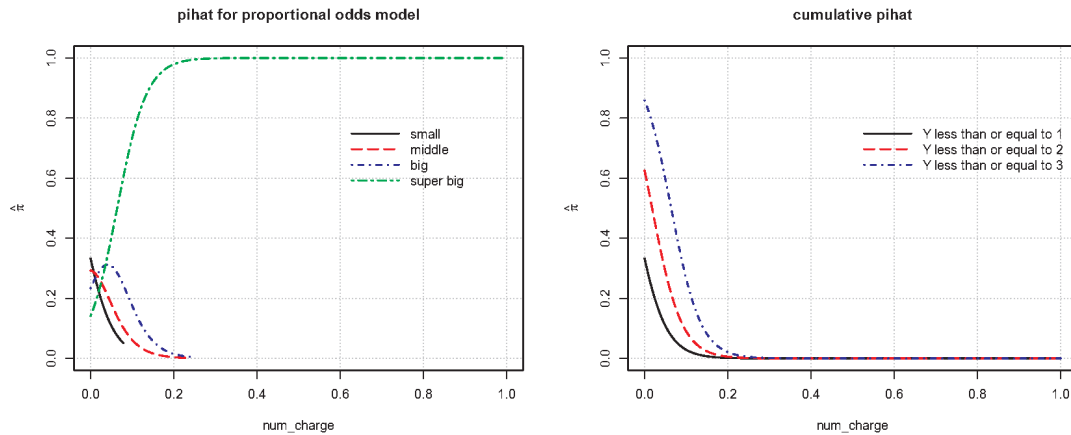


图 13: 以充电量为解释变量的累积逻辑回归模型概率分布图和累积概率图

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。

4.5 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：

$$\log(\pi_j/\pi_{small}) = \alpha_j + \beta_{j1}x_1 + \cdots + \beta_{jp}x_p, j = middle, big, super\ big \quad (3)$$

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。

Analysis of Deviance Table (Type II tests)

Response: fans_cat

	LR	Chisq	Df	Pr(>Chisq)
gender	8.843	3	0.0314491	*
num_videos	2.673	3	0.4448494	
num_charge	179.730	3	< 2.2e-16	***
av_star	14.746	3	0.0020469	**
av_like	71.304	3	2.244e-15	***
av_play	34.662	3	1.436e-07	***
av_comment	18.405	3	0.0003628	***
av_share	46.429	3	4.596e-10	***

图 14: Anova of first model

Analysis of Deviance Table (Type II tests)

Response: fans_cat

	LR	Chisq	Df	Pr(>Chisq)
gender	8.749	3	0.0328177	*
num_charge	179.832	3	< 2.2e-16	***
av_star	14.806	3	0.0019904	**
av_like	69.992	3	4.285e-15	***
av_play	34.532	3	1.529e-07	***
av_comment	18.651	3	0.0003228	***
av_share	46.092	3	5.421e-10	***

图 15: Anova of final model

得到的多项逻辑回归模型估计系数如下：

	Intercept	gender	charge	star	like	play	comment	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 5: 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。

5 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。

6 研究缺陷与改进

6.1 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”。

6.2 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。