B站up主修炼手册

----基于“新站”第三方数据网站

李明欣 邵明炎 王子韦

2022.12

1.内容摘要

本项目主要利用“新站”第三方数据平台上B站知识区up主的粉丝数、up主基本信息、up 主近三个月的视频转评赞数据进行综合分析和预测。通过使用最小二乘估计、广义加权最小二乘估计、分位数估计等估计方法，我们分析发现性别、认证情况、收录时长、近90天稿件数、总获赞数、近90天视频质量、互动率以及赞粉比都是影响up主粉丝数的显著因素。基于本报告的分析结果，我们可以为up提供一些参考建议，帮助up主成功涨粉。

2 .研究目的

近几年，科普、知识分享类内容越来越兴盛，而 B 站作为最受年轻人喜爱的视频创作平台，也成为了知识类视频的一大阵地。同时，也有越来越多的机构、团队、个人加入 B 站知识区 up 主的阵营。Up 主的粉丝数与他们的视频质量、更新动力是息息相关的。我们从新站上可获取两类数据，一类是up的个人信息（名称、账号标签等），一类是up的视频信息（更新频率、平均三连数等）。我们利用回归分析，探索各种因素与粉丝数间的关系，从而为知识区 up 主提供一些科学的涨粉建议。

3 .数据相关说明

3.1 原始数据获取

本实验中所用的原始数据来自新站。我们从新站上获取的截至2022年12月12日的数据，数据包括：up主名称、up主性别、up主认证信息、账号标签、粉丝数、总获赞数、收录时间、赞粉比、互动率、近九十天视频更新数、近九十天视频平均三连数。

由于粉丝量太少的up主还未发展起来，且人数过多；而粉丝量过大如两百万以上的up主人数太少，粉丝量又远大于其他up主，且其粉丝量如此高具有很多偶然因素。为了保证数据具有分析意义，我们选择了粉丝数在五万以上，两百万以下的up主数据。

而我们收集数据时，发现部分up主90天内未更新视频，我们认为该类up主或随心更新，或永久断更，不具普适意义，故也不具分析意义，予以剔除。

综上，最后我们选出了2995条数据。

3.2 原始数据处理

在进行后续分析之前；为了使自变量更具直接的实际意义，并且为可以用

来进行统计分析的数值型变量或属性变量，我们对原始数据进行了初步处理。详细的处理过程如下：

3.2.1 up 主名称数据处理

由于 up 主名称为字符型变量却又不是属性变量，故我们将 up 主名称拆为两种变量：一是其名称长度，为数值型变量；二是名称中是否含有英文字母，为属性变量。

3.2.2 收录时间数据处理

收录时间指新站开始收录该 up 主数据的时间，由于该时间与我们的因变量粉丝数没有直接关联，我们用我们的数据采集时间减去收录时间，得到 up 主被新站收录的时长。

又由于新站建立时间为 2020 年，而许多老 up 主已经经营账号超过两年时间了，故收录时长无法完全反映每个 up 主正式成为 up 的时长。而我们根据原始数据画出了其对应的柱状图，发现绝大部分 up 主都已于 2020 年 5 月 20 日前入驻 B 站，而此后 up 主入驻 B 站的时间较为平均，故我们决定将数据分为长、短两类，探究改变量与粉丝数的关系。

长、短两类对应的收录时长区间如下：

类型名称	对应的收录时长区间
长	>=两年
短	<两年

3.2.3 稿件数数据处理

对于 up 主总的作品数，我们做其散点图，发现大部分 up 主的总作品数集中在 0-600，1000 以上的 up 主数量很少，故我们决定对变量进行分组，将其变为属性变量。

分组情况如下：

类型名称	对应取值范围
作品数小于等于 100	作品数=<100
作品数大于 100	100<作品数=<200
作品数大于 200	200<作品数=<300
作品数大于 300	300<作品数=<400
作品数大于 400	400<作品数=<600
作品数大于 600	600<作品数=<700
作品数大于 700	700<=作品数

对于 up 主近 90 天的稿件数，同理，对其进行如下分组：

类型名称	对应取值范围
稿件数小于等于 5	稿件数=<5
稿件数大于 5	5<稿件数=<10
稿件数大于 10	10<稿件数=<15
稿件数大于 15	15<稿件数=<20
稿件数大于 20	20<稿件数=<30
稿件数大于 30	30<稿件数=<50

稿件数大于 50	50<=稿件数

最终我们选取的自变量如下表所示：

变量名称			变量类型	取值范围
因变量	粉丝数		连续变量	2w~1000w
自变量	up 主名称	名称长度	连续变量	2~16
		名称是否含英文	属性变量	Y，N
	up 主认证情况		属性变量	未认证，个人认证，机构认证
	up 主成为会员情况		属性变量	非会员，大会员，年度大会员
	up 主性别		属性变量	男，女，保密
	被收录时长		属性变量	长、短
	总作品数		属性变量	分为了七类，详情见报告
	总获赞数		连续变量	1877~44184389
	账号标签		属性变量	\
	总充电人数		连续变量	0~100%
	近 90 天稿件数		连续变量	分为了七类，详情见报告
	近 90 天集均播放		连续变量	0~4182000
	近 90 天集均弹幕		连续变量	0~31000
	近 90 天集均评论		连续变量	0~9018
	近 90 天集均分享		连续变量	0~96000
	近 90 天集均获赞		连续变量	0~390000
	近 90 天集均投币		连续变量	0~238000
	近 90 天集均收藏		连续变量	0~143000
	近 90 天互动率		连续变量	0~0,6037

4 .描述性分析

4.1 因变量 - 粉丝数

因变量知识区 up 主粉丝数为数值型连续变量，绘制直方图如下。

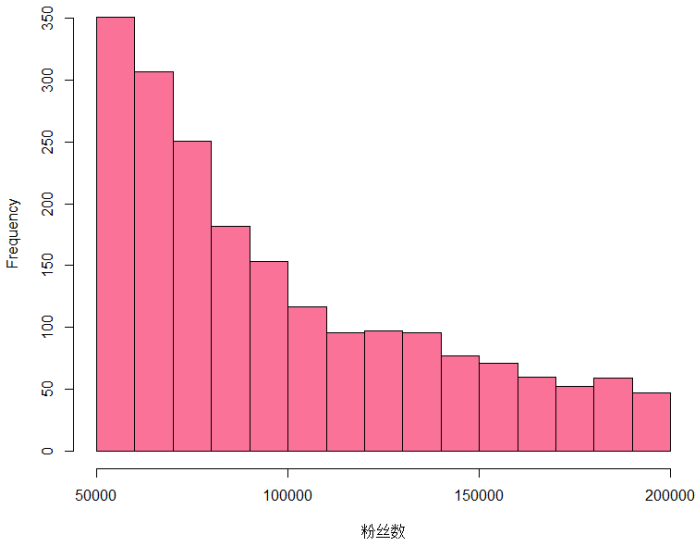

图 1：粉丝数直方图

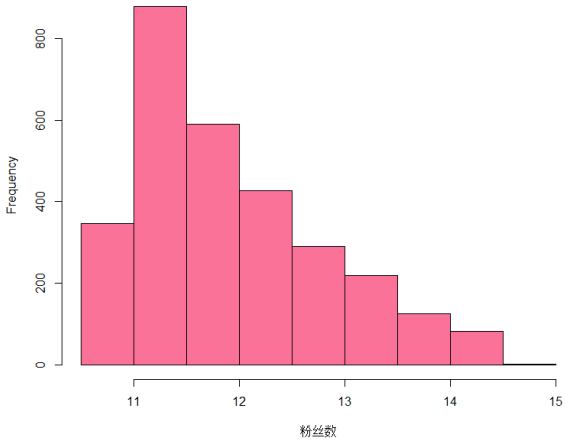

图 2：粉丝数取对数后直方图

可以看出，大部分 up 的粉丝数集中在 50000 到 500000，粉丝数 1000000 以上的 up 主相对来说人数很少。显然，粉丝数的分布整体呈陡降的趋势，不满足线性模型中对因变量正态性的要求。后续分析中我们将对这一问题做相关处理。而在进行后续分析前，我们先对粉丝数进行一次取对数变换。

4.2 自变量－up 主个人情况

4.2.1 up 主名称

由箱型图可以看出 up 主名称长度与粉丝数无明显的趋势性关系。而名称是否英文与不含英文的 up 主的粉丝数箱型图几乎一样，所以我们认为名称是否含英文对粉丝数无影响。

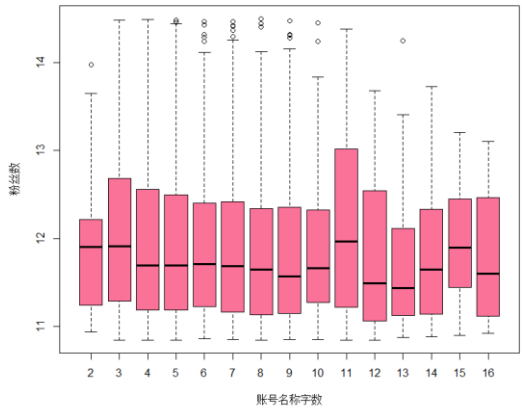

图 3：账号名称字数箱型图

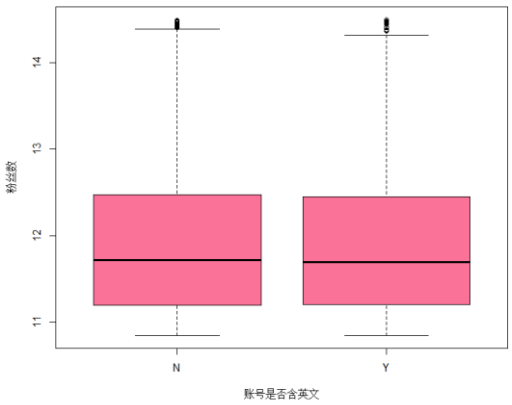

图 4：账号是否含英文箱型图

4.2.2 up 主其他个人信息

1）Up 主认证情况

通过箱型图可以看出，三类认证情况的 up 主粉丝数差别较大，其中个人认证的 up 主粉丝数较为分散，但总体最多，而未认证的 up 主粉丝数相对最少。

2）Up 主会员情况

由箱型图可以看出，是年度大会员的 up 主粉丝数相对来说最多，而非会员的 up 主粉丝数最少。

3）Up 主性别

由箱型图可以看出，男、女以及性别保密的 up 主，其粉丝数的中位数与集中区间都比较相近，其中女 up 主的粉丝数总体略少，而男 up 主的粉丝数略多，可见 up 主性别对其粉丝数是有一定影响的。

从图中还可看出，有许多粉丝数过大的异常值点，这是由总体样本量的粉丝数分布极其不均造成的，后续我们将解决这个问题。

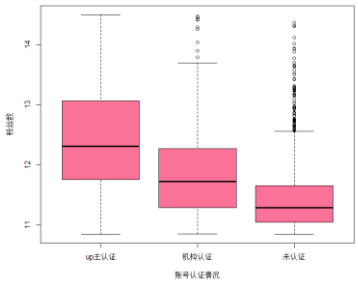

图 5：账号认证情况箱型图

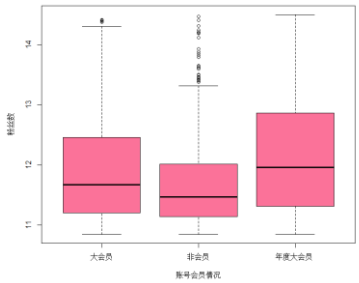

图 6：账号会员情况箱型图

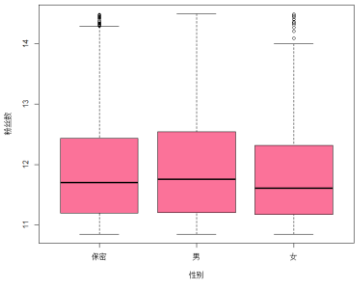

图 7：性别箱型图

4.2.3 被收录时长

由箱型图可以看出，被收录时间长的 up 主的粉丝数整体显著多于短的 up 主。而被收录时间短的 up 主，其由较多粉丝数过多的异常值点，说明也有部分 up 主，虽然成为 up 主时间较短，但可能由于视频质量高、更新快等原因涨粉迅速，这也是符合实际情况的。

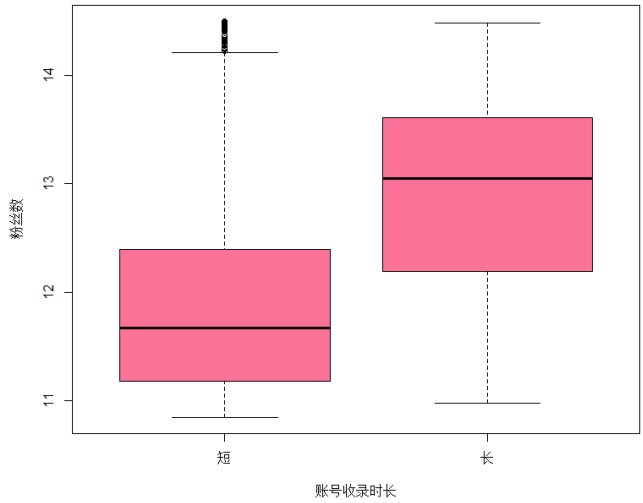

图 8：账号收录时长箱型图

4.2.4 总作品数

从图中可以看出，总作品数小于 400 时， up 主的粉丝数随作品数的变多而变多，由于 up 主吸引粉丝主要是靠产出视频，所以这是符合实际的。但当视频数过多如 400 以上时，粉丝数反而不会变多，联系实际，作品数过多会出现视频质量降低的情况，或者该 up 主的视频并非原创，从而粉丝数也不会太多。且我们还可以看到作品数小于等于 100 的 up 主粉丝数较少，但该部分还有较多粉丝数过大的异常值点，结合实际，这是由于部分 up 主视频精益求精，导致更新周期较长，但好的视频质量仍未该类 up 主吸引了不少粉丝。

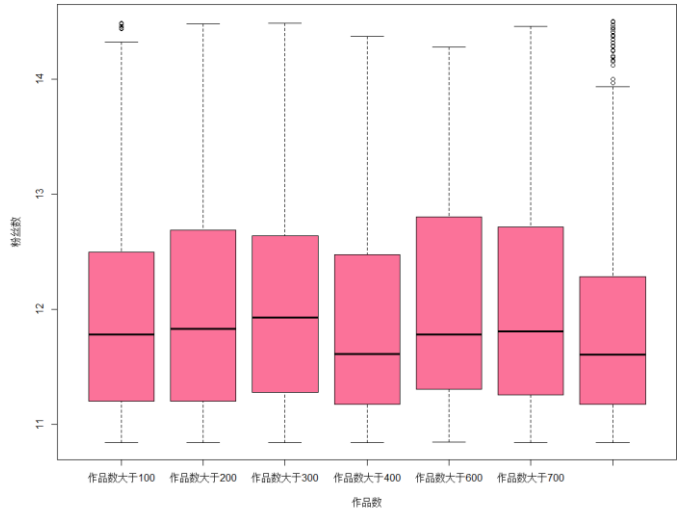

图 9：作品数箱型图

4.2.5 账号标签

账号标签为离散型属性变量，可以反映在知识区这个大类下，up 主发布的视频内容主要为哪几类知识分享。

我们对此绘制了词云图，可以看到知识区 up 主的细分类别还是非常丰富

的，但绝大多数 up 主都集中在校园学习、日常、人文历史、社科人文、科学科普方面，其中制作校园学习内容相关的 up 主数量又是明显最多的。

图 10：账号标签词云图

4.3 自变量 - up 近 90 天视频表现

4.3.1 近 90 天稿件数

近 90 天稿件数反映了 up 主近期的更新频率，该部分规律与总作品数相似。稿件数小于 15 时，up 主的粉丝数随稿件数变多而升高，但当稿件数超过 15 时，粉丝数反而随作品数的变多而下降。

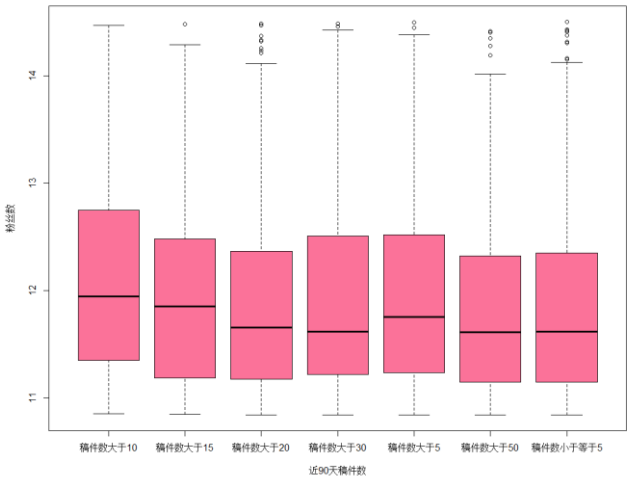

图 11：近 90 稿件数箱型图

4.3.2 近 90 天互动率

改变量反映了近期 up 主与粉丝以及其他用户的互动情况，值越大说该 up 主的互动越频繁。通过散点图我们并不能看出互动率与粉丝数有明显线性关系。

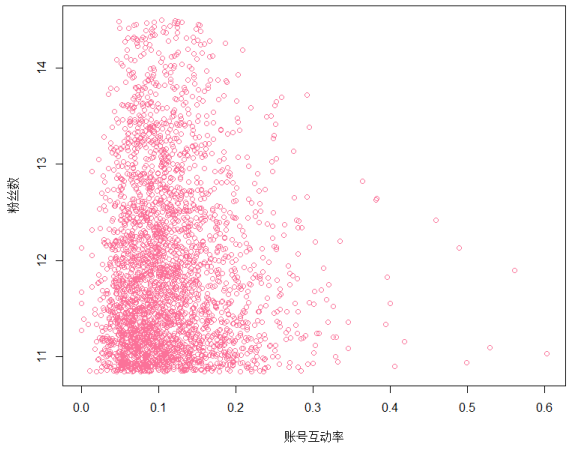

图 12：账号互动率-粉丝数散点图

4.3.3 近 90 天视频集均表现

近90天视频集均表现包含近90天集均播放量、弹幕数、评论数、获赞数、投币数与收藏数。我们绘制他们与粉丝数以及彼此之间的散点图。

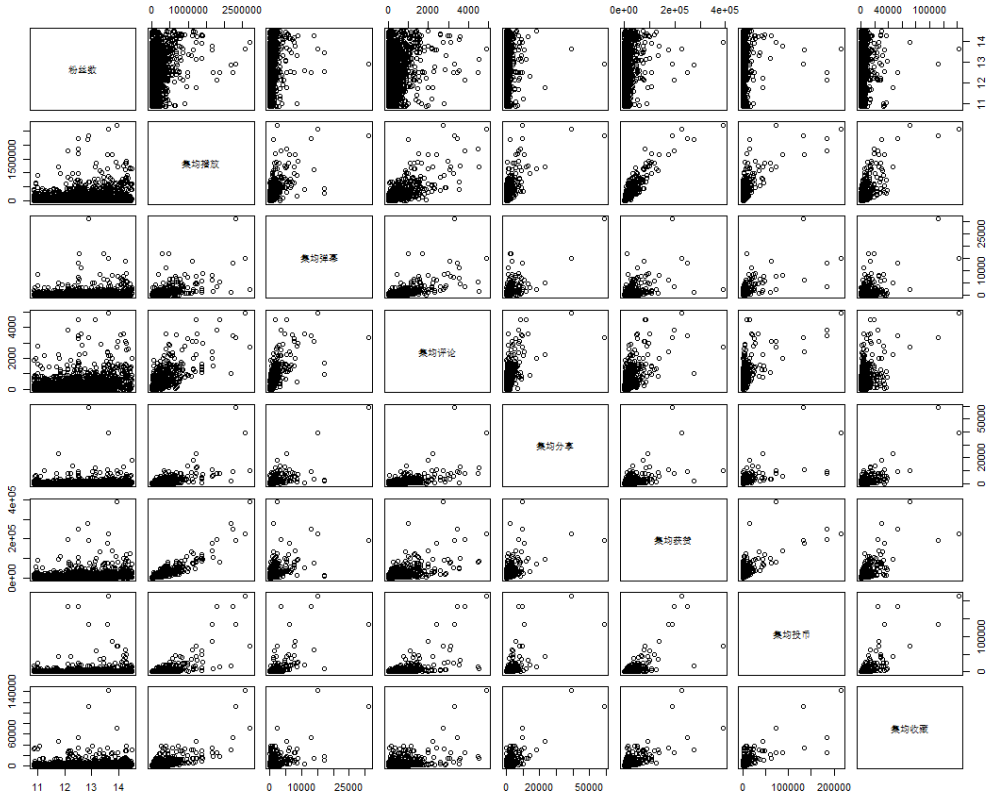

图 13：n*n 散点图

我们看到这些自变量间有较强的的线性性，这会导致进行回归时模型有较强的复共线性，会破坏模型有效性。我们在此处使用熵权法，获得这七个变量各自的权重，将其加和到一起作为新的回归自变量，并将该变量命名为“视频质量”。并且我们发现取该变量的对数，该变量与粉丝数间的线性性更为明显。

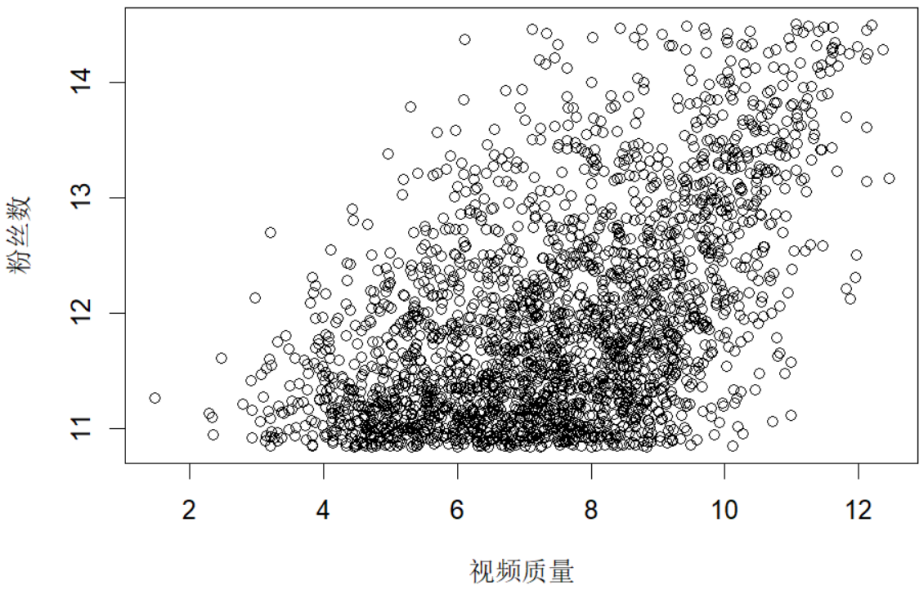

图 14：视频质量-粉丝数散点图

5 .模型建立

5.1 全模型分析

通过上述描述性分析我们对我们的自变量与因变量有了基础的了解，并对部分自变量进行了调整。下面我们利用得到的最终数据进行全模型回归，得到的结果如下表所示：

变量名	估计值	P 值	显著性水平
截距	10.029183	< 2e-16	***
是否含英文 Y	0.008056	0.716838	
性别男	-0.40106	0.029299	*
性别女	-0.024238	0.320290	
会员情况非会员	0.033674	0.115166	
会员情况年度大会员	0.006345	0.759099	
认证情况机构认证	-0.043291	0.196432	
认证情况未认证	-0.242297	< 2e-16	***
收录时长长	0.099875	0.081554	.
名称字数	0.003716	0.351670	
作品数大于 200	-0.041785	0.141410	
作品数大于 300	0.001383	0.970820	
作品数大于 400	-0.044216	0.215728	
作品数大于 600	0.021627	0.714928	
作品数大于 700	-0.019018	0.652850	
作品数小于等于 100	-0.019937	0.350311	
稿件数大于 15	0.029971	0.433932	

稿件数大于 20	0.009813	0.784746	
稿件数大于 30	0.040130	0.281216	.
稿件数大于 5	-0.050884	0.083351	.
稿件数大于 50	0.065143	0.091732	***
稿件数小于等于 5	-0.094236	0.000979	***
Log(获赞数)	0.103772	< 2e-16	***
Log(视频质量)	0.059031	3.37e-15	***
互动率	-0.002077	0.987949	
赞粉比	-0.909254	< 2e-16	***

在最终的模型分析之前，为了确保模型分析结果的合理性，我们首先对模型的正态性以及同方差假设进行检验，得到的结果如图所示。

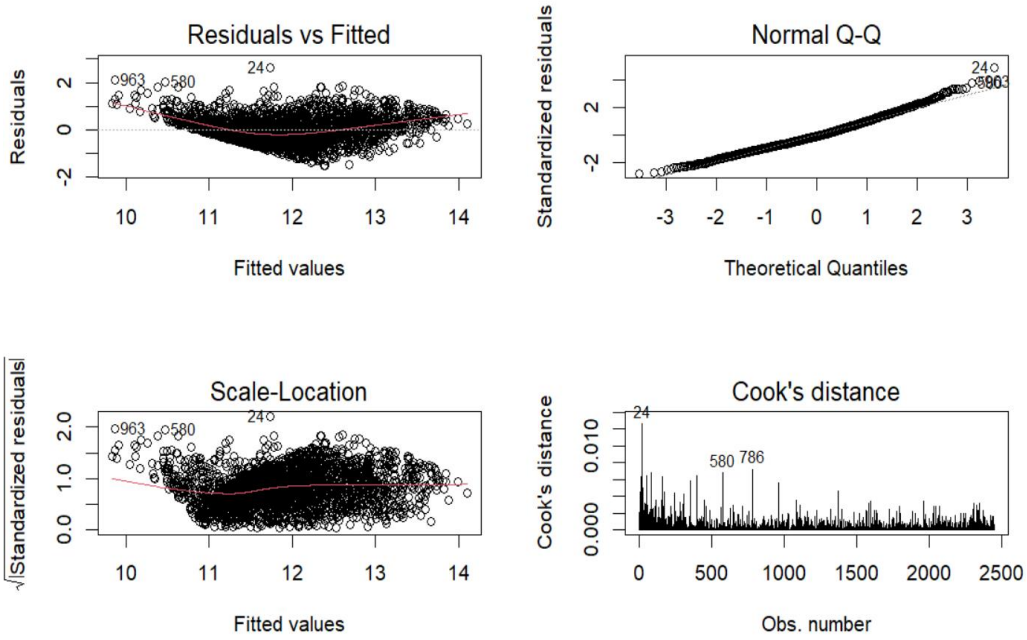

图 15：全模型回归诊断图

从残差图中我们可以看出(y_i, e_i)均落在[-2,2]的水平区域内，但呈现中间略低两端略高的特征，我们在 R 中用 ncvTest 函数对其进行检验得到的 P 值如下，可以看到 P=0.0008，在显著性水平为 0.05 的情况下我们有充分的理由拒绝原假设，即原模型存在异方差性。

P 值
0.00085681

从 Q-Q 图中我们可以看到其趋近于落在 y=x 线上，但其尾部明显偏离了直线，我们在 R 中用 shapiro.test 函数对其进行检验，得到的 P 值=1.365e-05，我们有充分的理由拒绝原假设，即原模型不满足正态性假设。

P 值
1.365e-05

接下来，我们对原模型的复共线性进行检验，结果如下，可以看到设计矩阵条件数为 164.1977，因此我们认为原模型存在中等程度的复共线性。

Kappa
164.1977

为了确保接下来分析的正确性，我们对模型进行上述问题的修正。

5.2 异方差和正态性问题解决

5.2.1 Box-Cox 变换

Box-Cox 变换可以明显改善数据的正态性以及方差相等性，因此我们首先采用该方法对原模型进行修正。我们通过画出λ与对数似然函数的图像来找到最优的λ值

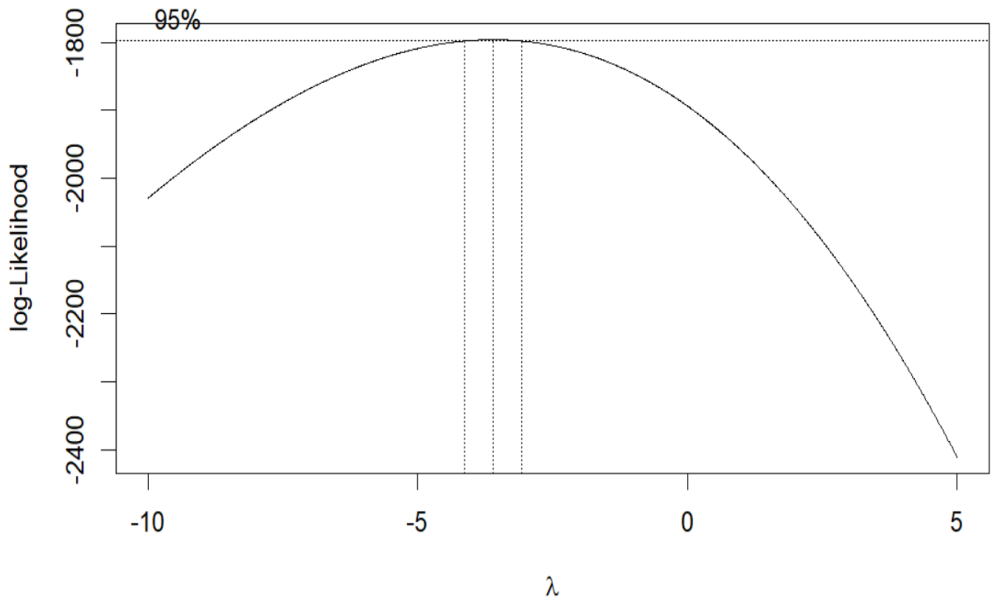

图 16：λ-对数似然函数

如图所示，我们可以看到图中存在唯一的λ值使得似然函数达到最大值，因此我们可以近似地认为该λ值即为变换参数的最优选择。我们再次对 Box-Cox 变换后的因变量与自变量进行方差分析，结果如下图：

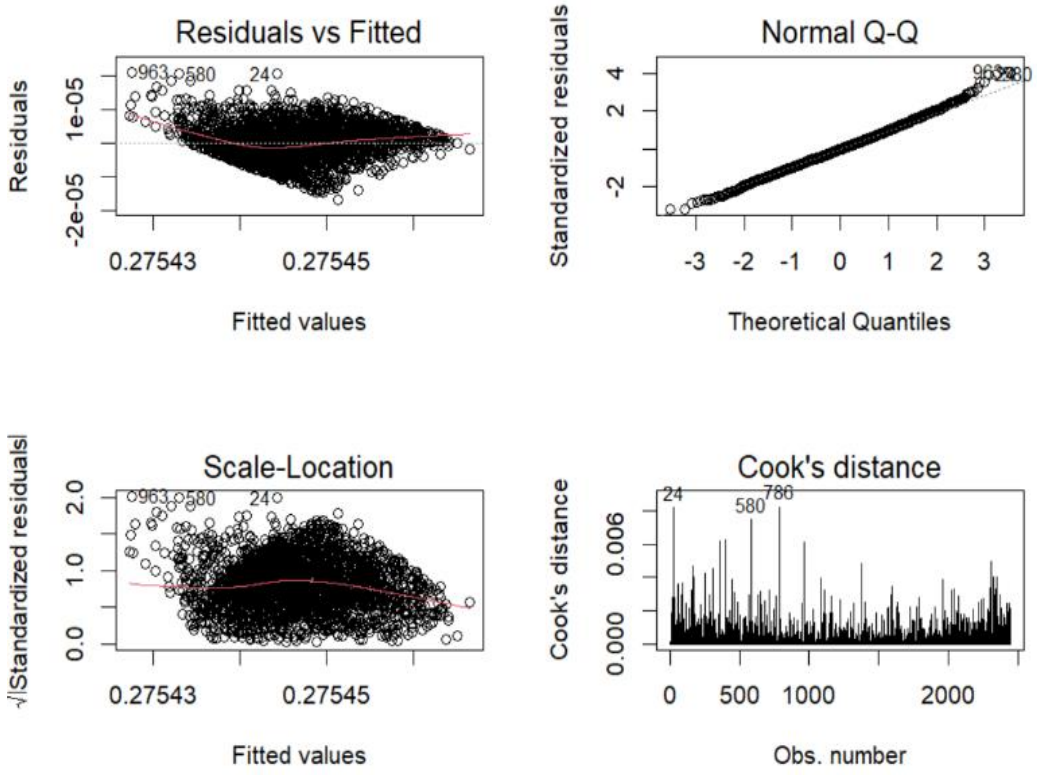

图 16：Box-Cox 变换后模型诊断图

可以看到残差项的分布区间明显缩小，Q-Q 图的尾部偏离直线部分也有所改善，我们通过 R 语言来进一步验证。

我们通过 ncvTest 函数检验异方差所得的 P 值由 0.0008 变为 0.01277，对正态性假设检验所得的 P 值变为 0.005276，虽已有极大改善，在显著性水平为 0.05 的情况下我们依然拒绝原假设。

P 值
0.01277

P 值
P=0.05276

下一步我们将对异方差和非正态性作出进一步的改进。

5.2.2 广义加权最小二乘法

经过上述修正后，我们发现残差图仍存在中间略低两端略高的问题，因此我们可以在经典的最小二乘回归法上改进，对方差大的赋予较小权重，对方差小的赋予较大权重，使原模型的异方差误差项转换为同方差误差项，从而使加权变换后的模型满足上述一系列假定。加权后的模型诊断图如下。

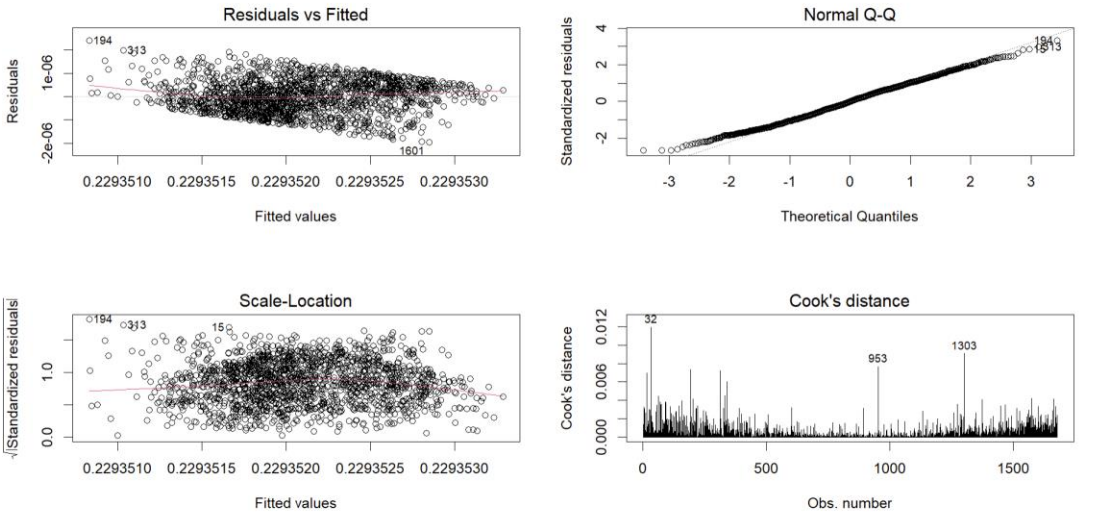

图 16：广义加权最小二乘法模型诊断图

通过对上述诊断图进行可视化分析发现残差项分布有所改善，其同方差检验所得的 P 值为 0.06，在显著性水平为 0.05 的情况下接受原假设，即我们认为原模型已不存在异方差。

P 值
0.060236

其正态性检验的 P 值为 0.00555，经过广义加权最小二乘后该模型的正态性仍需改善。

P 值
0.00555

5.3 复共线性问题解决

如果该模型自变量间存在复共线性，那么最小二乘估计很可能失效，因此我们将采取一些措施来解决复共线性问题。首先我们采用 AIC 原则对其进行逐步回归，选取出最优模型，选取出来的变量为剩余变量、性别、认证情况、收录时长、名称字数、稿件数、Log(获赞数)、Log(视频质量)、互动率 、赞粉比。

变量名	vif
性别	1.021912
认证情况	1.070212
收录时长	1.019737
Group2	1.039826
Log(或赞数)	1.170057
Log(视频质量)	1.468578
赞粉比	1.464904

我们使用方差膨胀因子 vif 对其复共线性进行检验，结果如下：

P 值
0.026909

我们发现 aic 原则筛选后的模型已经不存在复共线性。

Shapiro .test
0.06268

通过以上 P 值可知，正态性质仍不良好，但已基本满足同方差性，并且不存在复共线性，接下来我们将探寻更加稳健的估计。

5.4 分位数回归

与最小二乘法不同，分位数回归通常是求残差的绝对值的加权求和最小来估计参数。相比于均值回归，分位数回归具有以下优点：

1) 分位数回归对数据分布的情况掌握的更全面客观。

2) 使用分位数回归，离群点对于数据整体的影响要比较使用均值回归小的多。

所以我们也可以说分位数回归更加稳健。

3) 分位数回归对于误差项更具有普适性。

在本案例中，我们发现经过上述处理的数据集仍不满足正态性假设，具有有偏性，回归效果未达到预期，因此我们考虑使用更加稳健的回归模型，并基于不同分位数给出不同粉丝区间影响 up 主粉丝量的因素及影响大小。

5.4.1 中位数回归

5.4.1.1 中位数回归模型结果

我们选定 0.5 为分位点，进行全模型回归，得到结果如下：

变量名	估计值	P 值	显著性水平
截距	10.14506	0	***
是否含英文 Y	0.01640	0.51934	
性别男	-0.04332	0.06112	*
性别女	-0.00905	0.69336	
会员情况非会员	0.04300	0.08216	*
会员情况年度大会员	0.02849	0.24068	
认证情况机构认证	-0.05942	0.17990	
认证情况未认证	-0.35027	0	***
收录时长长	0.09468	0.18281	
名称字数	0.00123	0.76625	
作品数大于 200	-0.05696	0.08282	*
作品数大于 300	-0.03482	0.49207	
作品数大于 400	-0.08422	0.03488	*
作品数大于 600	-0.11732	0.47942	

作品数大于 700	-0.01713	0.71660	
作品数小于等于 100	-0.03189	0.17318	
稿件数大于 15	0.01783	0.75952	
稿件数大于 20	-0.02545	0.54591	
稿件数大于 30	-0.01342	0.76258	
稿件数大于 5	-0.03281	0.34640	
稿件数大于 50	-0.01320	0.77941	
稿件数小于等于 5	-0.02142	0.54560	
Log(获赞数)	0.11082	0	***
Log(视频质量)	0.02924	0.00002	***
互动率	-0.14736	0.40752	
赞粉比	-0.08628	0	***

5.4.2 中位数回归与最小二乘法比较

变量名	估计值（0.5）	P 值（0.5）	估计值	P 值
截距	10.14506	0***	10.029183	< 2e-16***
是否含英文 Y	0.01640	0.51934	0.008056	0.716838
性别男	-0.04332	0.06112*	-0.40106	0.029299*
性别女	-0.00905	0.69336	-0.024238	0.320290
非会员	0.04300	0.08216*	0.033674	0.115166
年度大会员	0.02849	0.24068	0.006345	0.759099
机构认证	-0.05942	0.17990	-0.043291	0.196432
未认证	-0.35027	0***	-0.242297	< 2e-16***
收录时长长	0.09468	0.18281	0.099875	0.081554*
名称字数	0.00123	0.76625	0.003716	0.351670
作品数大于 200	-0.05696	0.08282*	-0.041785	0.141410
作品数大于 300	-0.03482	0.49207	0.001383	0.970820
作品数大于 400	-0.08422	0.03488*	-0.044216	0.215728
作品数大于 600	-0.11732	0.47942	0.021627	0.714928
作品数大于 700	-0.01713	0.71660	-0.019018	0.652850
作品数小于等于 100	-0.03189	0.17318	-0.019937	0.350311
稿件数大于 15	0.01783	0.75952	0.029971	0.433932
稿件数大于 20	-0.02545	0.54591	0.009813	0.784746
稿件数大于 30	-0.01342	0.76258	0.040130	0.281216
稿件数大于 5	-0.03281	0.34640	-0.050884	0.083351*

稿件数大于50	-0.01320	0.77941	0.065143	0.091732*
稿件数小于等于5	-0.02142	0.54560	-0.094236	0.000979***
Log(获赞数)	0.11082	0***	0.103772	< 2e-16***
Log(视频质量)	0.02924	0.00002***	0.059031	3.37e-15***
互动率	-0.14736	0.40752	-0.002077	0.987949
赞粉比	-0.08628	0***	-0.909254	< 2e-16***

由上表可知，对于两种回归方法显著的自变量大体相同，且符号大体一致，说明变量的影响及相关性基本相同。但存在中位数回归中显著而最小二乘回归中不显著的变量，如非会员与作品数，我们认为主要是由于异常值的存在，导致均值偏离。由于分位数回归稳健性的特点，我们认为实际上会员情况会影响 up 主的粉丝数，与箱型图中结果一致。同时，也存在最小二乘回归中显著而中位数回归不显著的变量，如收录时长、稿件数等。

5.4.3 其他分位点回归

同时，为了得到不同粉丝区间影响 up 主粉丝量的因素及影响大小，我们分别选定 0.25，0.75 为分位点，进行全模型回归，得到结果如下：

变量名	估计值（0.25）	P 值（0.25）	估计值（0.75）	P 值（0.75）
截距	10.09053	0***	10.47249	0***
是否含英文 Y	-0.01102	0.67091	-0.00285	0.90200
性别男	-0.03681	0.01980*	-0.04766	0.03448*
性别女	0.00901	0.75863	-0.04287	0.13138
会员情况非会员	0.02432	0.15019	0.05124	0.06621*
会员情况年度大会员	0.01631	0.35606	0.00524	0.83323
认证情况机构认证	-0.02358	0.59984	-0.08487	0.03537*
认证情况未认证	-0.19639	0***	-0.37864	0***
收录时长长	0.20921	0.02789*	0.02517	0.77086
名称字数	0.00442	0.24250	0.00594	0.22867
作品数大于200	-0.09234	0.00030***	-0.02867	0.50086
作品数大于300	-0.02767	0.38308	0.00965	0.81751
作品数大于400	-0.02190	0.47222	-0.05658	0.13736
作品数大于600	-0.04249	0.38758	0.01109	0.92247
作品数大于	-0.03632	0.30356	-0.04365	0.41747

700				
作品数小于等于100	-0.02825	0.20947	-0.00491	0.85134
稿件数大于15	-0.07706	0.08148*	-0.01571	0.73125
稿件数大于20	-0.07456	0.01049*	-0.03893	0.27740
稿件数大于30	-0.03602	0.26332	-0.02362	0.58818
稿件数大于5	-0.04996	0.11163	-0.06727	0.03875*
稿件数大于50	-0.07338	0.02123*	0.06495	0.21761
稿件数小于等于5	-0.07539	0.00890**	-0.06109	0.06579*
Log(获赞数)	0.09391	0***	0.10654	0***
Log(视频质量)	0.02178	0***	0.02108	0.00828**
互动率	-0.17981	0.11739	0.11515	0.53566
赞粉比	-0.08291	0***	-0.09185	0**

在上四分位数回归与下四分位数回归中，与0.25分位相比，0.75分位回归中，机构认证显著，且为负值，即说明当粉丝达到一定数量后，大众更愿意关注up主个人账号而非机构账号，同时，未认证的负影响更大。此外，作品数不再显著，即对于已经有一定基础的up主来说，只靠视频数量已不再能帮助吸引新的粉丝群体。

6 .回归结果及解读

6.1 回归结果

通过上述检验，我们选定加权最小二乘估计经 aic 原则筛选后的模型作为最终优化后的模型，判别系数为 0.6，结果可信，得到的方差分析以及模型分析结果如下：

变量名	P 值	显著性水平
性别	0.02882	*
认证情况	< 2.2e-16	***
收录时长	0.12305	.
稿件数	4.734e-06	***
Log(获赞数)	< 2.2e-16	***
Log(视频质量)	< 2.2e-16	***
赞粉比	< 2.2e-16	***

估计值

变量名	估计值	P 值	显著性水平
截距	2.293e-01	< 2e-16	***
性别男	-1.002e-07	0.008595	**

性别女	-3.177e-08	0.520957	
认证情况机构认证	-6.253e-08	0.367685	
认证情况未认证	-4.924e-07	< 2e-16	***
收录时长长	1.680e-07	0.123054	
稿件数大于 15	5.128e-08	0.537202	
稿件数大于 20	1.403e-08	0.850433	
稿件数大于 30	9.020e-08	0.225923	
稿件数大于 5	-9.892e-08	0.102678	
稿件数大于 50	1.496e-07	0.047303	*
稿件数小于等于 5	-1.992e-07	0.000752	***
Log(获赞数)	2.215e-07	< 2e-16	***
Log(视频质量)	1.300e-07	< 2e-16	***
赞粉比	-2.001e-06	< 2e-16	***
判决系数 0.6234			

6.2 模型解读

经我们一系列的改善后，我们认为最后的模型对自变量和因变量之间的关系有一定的解释的能力。通过对各个自变量所对应的 t 检验的 p 值进行比较，在 0.05 的显著水平上，我们可以认为：性别、认证情况、收录时长、近 90 天稿件数、总获赞数、近 90 天视频质量、互动率以及赞粉比都是重要的影响因子，而其他因子对 up 主粉丝数的影响效果可以忽略。

根据估计值表的结果，我们可以得出以下结论：

1) 性别对 up 主粉丝数有一定影响，在知识区，性别保密的 up 主相对来说更容易吸引粉丝。在 B 站性别保密一般有如下几种情况：up 主自己未去认证性别、up 主视频不包含真人，性别成迷、up 主为一个团队，无法选出性别。说明在知识区，粉丝并不关注 up 主性别，只要认真输出好内容，就可以吸引到粉丝。

2) 个人认证的 up 主相对于未认证以及机构认证的 up 主能够吸引更多粉丝。故而一个知识区个人账号的 up 主如果想要认真经营自己的账号，可以对自己的账号申请认证，这样更能给人以靠谱、权威的感觉。

3) 收录时长长的 up 主粉丝相对多，这说明想要成为一名 up 主，不可以急躁，要沉下心来，不能指望刚发几个视频就能拥有很多粉丝，初步阶段粉丝不多也不要焦虑，只要坚持输出好内容，粉丝数总会涨起来的。

4) 近 90 天稿件数越多，up 主的粉丝数越多。这说明想要成为一名好的 up 主，更新视频要勤快。

5) 获赞数越多、视频质量越高的 up 主，粉丝数也越多。且进行回归时，这三个变量都取了对数，这说明获赞数与视频质量对粉丝数的影响是成指数级的。这说明在 B 站，在知识区，想要涨粉，up 主最重要的还是得持续产出

高质量作品。

6) 赞粉比一项的结果与我们的直观感受有偏差，我们猜测是赞粉比为符合变量，我们的因变量是作为赞粉比的分母存在，故回归结果显示二者间存在负相关性。

7. 改进与不足

目前我们的模型仍存在一定的不足，我们认为可以从以下几个方面进行改进：

1）在数据获取方面，由于 B 站的发展，其直播行业与专栏方面都受到了一定的关注，因此相应 up 主的粉丝数与视频质量的相关性可能较差，在视频质量方面，由于短视频行业的兴盛，我们曾试图将视频时长列为自变量，但由于技术的限制以及数据的体量，我们并未纳入考虑。

2）针对出现的非正态性问题，我们对自变量进行了多次变换，导致后续计算的复杂性，同时不能直观的衡量系数之间的关系。