公众号推文阅读量影响因素分析

一、背景介绍

自 2011 年出现以来，微信作为社交媒体，其发展火爆而迅速。与此同时，微信公众号的数量也随着微信用户的增加不断上涨，呈现出百花齐放的景象。微信公众号之所以能够迅速在互联网+的时代脱颖而出也得益于其突出的优势。

首先，通过微信公众号开展微信信息推广传播活动的成本非常低，很适合进行低成本宣传。其次，微信公众账号让用户的分类更加多样化，可以通过后台的用户分组和地域控制实现精准的消息推送。同时，微信公众号传递信息更加多元化。可以推送文字、图片、语音、视频、图文消息五个类别的内容，用户之间可以选择自己感兴趣的内容接受信息。因此，越来越多的人开始尝试经营微信公众号并且从公众号中了解更多的信息。

艾媒咨询的研究报告显示，截至 2016 年，中国微信公众号数量就已超过 1200 万个。微信用户平均每天会阅读 5.86 篇公众号文章，由此表明公众号内容已经成为人们获取信息的主要途径之一。

然而，随着微信公众号的蒸蒸日上，广大公众号运营主和广告商也在其中“内卷”了起来，它们利用用户打赏、接取商业性文章、收取广告费等形式扩大盈利，得到收益。因此每个公众号运营者都在思考如何将自己的公众号打造的更好，从中一战成名。

鉴于实际意义和数据的可获得性，本文选取公众号每次发文能够获得的总阅读量来反映公众号的价值，该指标不仅代表了公众号的传播阅读量的多少，还侧面反映了公众号的活跃粉丝数，对公众号的运营有着巨大影响。我们试图进行数据分析、建立回归模型的方式来探究公众号阅读量的影响因素，从而为公众号运营提供中肯建议。

二、数据来源和说明

本文的数据一共有 998 个不同的公众号，20 个变量，具体变量说明见表 1。

表 1 变量说明

变量名称			变量类型	取值范围	备注
因变量	总阅读量		连续变量	18.87～780.11	平均每次发文总阅读量（单位：千）
自变量	公众号属性	公众号类型	分类变量	个人、企业、不明	
		公众号类别	分类变量	搞笑娱乐等 12 个	
	名称	名称长度	连续变量	2～16	公众号所含字数
		名称是否包含英文	分类变量	不含英文、含英文	
	头像	头像是否含文字	分类变量	不含文字、含文字	
		头像彩色	分类变量	彩色、黑白	
	发文频次	日均发文次数	连续变量	0～8	平均每天发文次数
		单次篇数	连续变量	1～8	平均每次发文的篇数
	发文类型	原创占比	连续变量	0～1	3 月全部所发文章的原创占比
		视频占比	连续变量	0～1	3 月全部所发文章的视频占比
	发文时间	时段 1 占比	连续变量	0～1	从 18 点，每四个小时一个时间段
		时段 2 占比	连续变量		
		时段 3 占比	连续变量		
		时段 4 占比	连续变量		
		时段 5 占比	连续变量		
		时段 6 占比	连续变量		
	文章标题	标题长度	连续变量	5～55	最近 10 篇文章，标题包含的平均字数
		标题标点指数	连续变量	0～1	最近 10 篇文章含有特定标点的标题占比
		标题正向情感得分	连续变量	0～1	最近 10 篇文章标题的正向情感得分

其中，我们选取的因变量是公众号每次发文所能获得的总阅读量。自变量一共有 19 个，主要涉及公众号属性、公众号名称、公众号头像、发文频次、发文类型、发文时间以及文章标题等 7 个维度。

三、描述性分析

将数据读入后，我们对数据进行描述性分析，利用 R 语言进行编程，得到结果。

（一）因变量：总阅读量

我们首先关注因变量总阅读量的分布，得到结果如图 1：

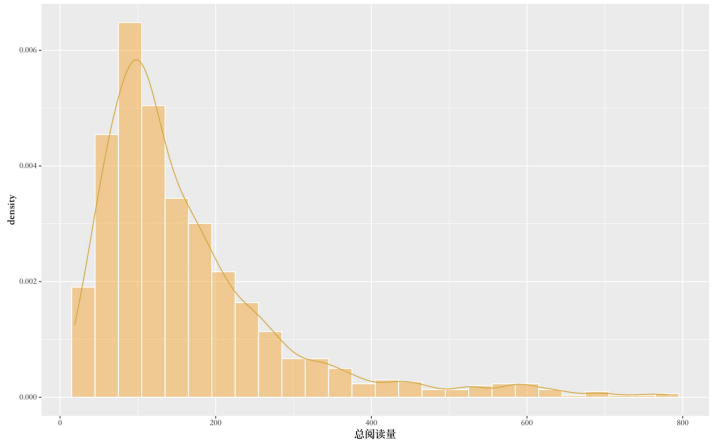

图 1 总阅读量的分布

如图 1 所示，上面是这 998 个公众号的平均总阅读量的分布。可以看出大多数公众号的总阅读量集中在 50,000～200,000 之间，后面呈现出“长尾”的分布状态。因此，在建模的时候我们需要对因变量进行对数变换，以保证其正态分布。将数据进行对数变换后，我们得到分布如图 2：

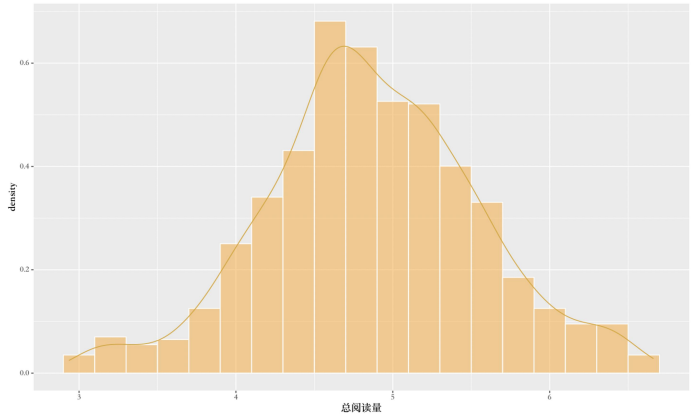

图 2 对数变换后总阅读量的分布

由图 2 可以观察到，对数变换后的因变量分布基本服从正态分布，因此可以用对数变换后的数据进行建模。

（二）自变量：公众号属性

我们对第一个维度公众号属性进行分析，这个维度包含两个自变量——公众号类型和公众号类别，我们用箱线图进行分析，得到结果图如下：

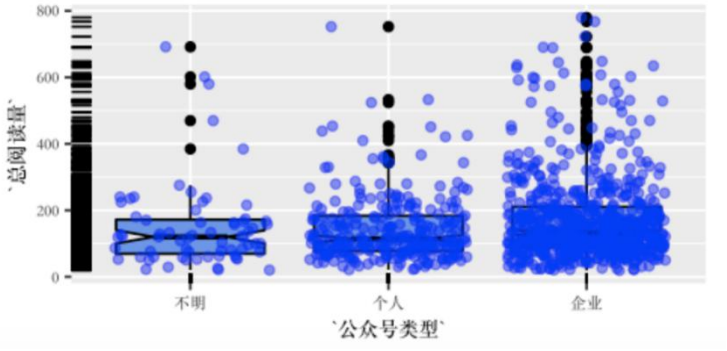

图 3 公众号类型与总阅读量箱线图

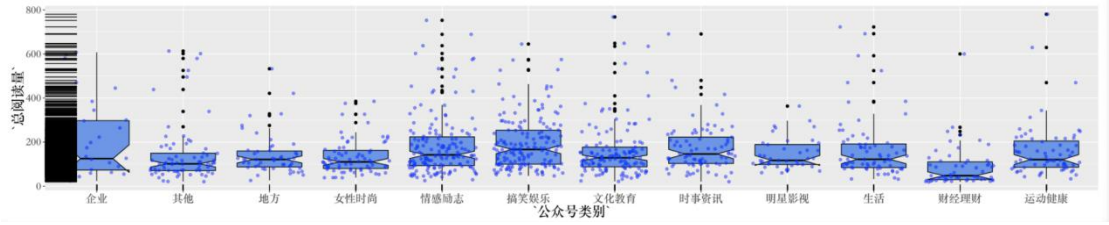

图 4 公众号类别与总阅读量箱线图

由图 3 可以看出企业类公众号的总阅读量要比不明类型和个人类型多一些，这可能是由于企业的宣传力度更大一些，内容更加丰富，且具有一定可信度。由图 4 可以看出在公众号类别中，搞笑娱乐类公众号的总阅读量最多、其次是时事资讯、情感励志、生活、运动健康类，而财经理财类的公众号阅读量最少。这可能是由于人们在浏览公众号文章时，更倾向于得到放松的心情，而搞笑娱乐往往更加吸引人且可以得到开心的情绪，因此浏览量最高。而贴近生活类的文章也是人们所必需了解的，这些公众号可以得到更多知识，因此生活类也是比较热门的类别。而财经理财类更偏向于一些经商、自律的人群，可能无法涵盖大部分人最需要了解的方面，因此阅读量偏少。

（三）自变量：公众号名称

我们对第二个维度公众号名称进行分析，这个维度包含两个自变量——名称长度、名称是否包含英文，我们继续用箱线图进行分析，得到结果图如下：

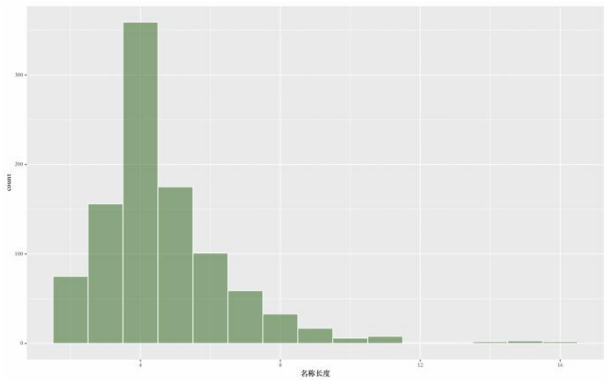

图 5 名称长度的分布

由图 5 可以看出公众号所含字数在 2～6 的分布最多。因此令名称长度大于 6 为长，小于等于 6 的时候为短。画出箱线图如下：

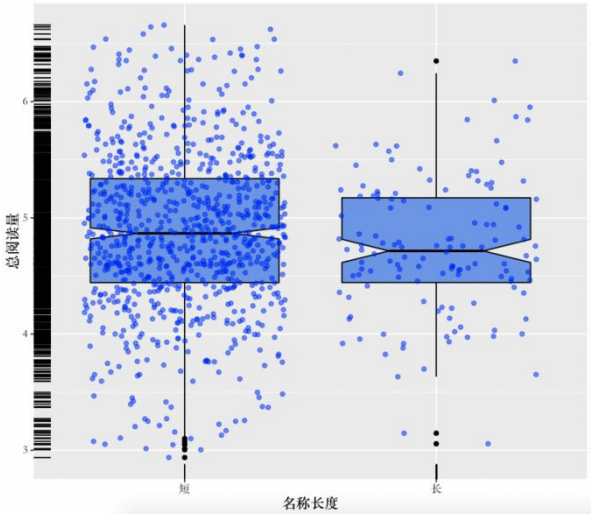

图 6 名称长度箱线图

由箱线图 6 可以看出名称长度短的阅读量更高一点，这可能是由于名称长度短更精简，更让人有耐心看完，也更加吸引人，而名称长度长的公众号名称本就在数量少略少于长度短的公众号。

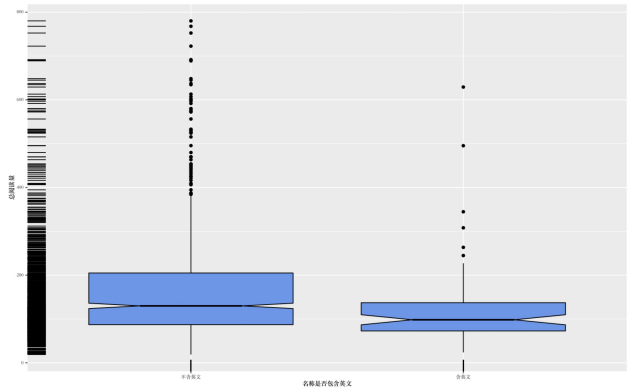

图 7 名称是否包含英文

由箱线图 7 可以看出公众号文章不含英文的阅读量比含英文的阅读量要高

一些，这可能由于目前一些人群，例如一些孩子、老人以及英文水平较弱的人更加倾向于看中文的文章，可以省时且看的更加明白。

（四）自变量：公众号头像

公众号的头像往往决定了用户对其的第一印象。我们在公众号头像这一维度的两个变量进行分析，分别是是否含文字以及颜色是彩色还是黑白。

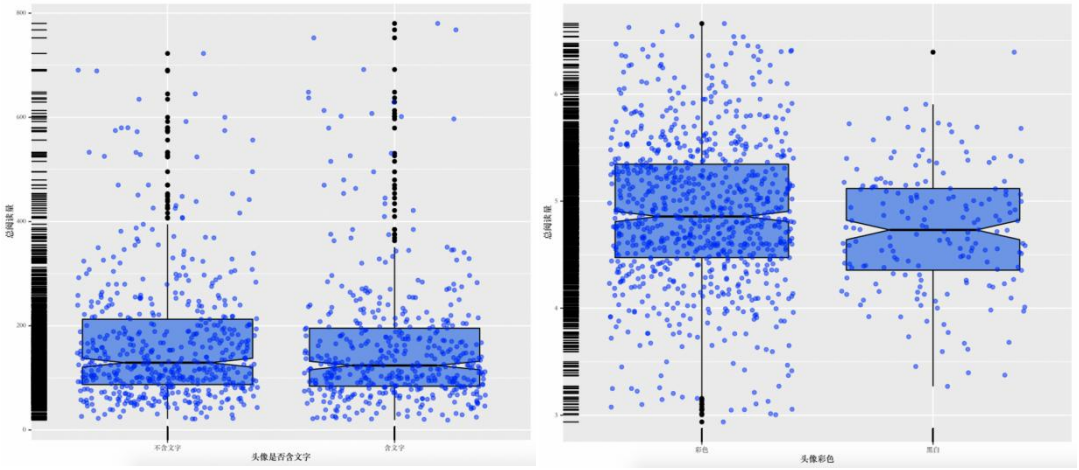

图 8 公众号头像箱线图

由箱线图图 8 所示，公众号的头像是否含有文字对于总阅读量的影响并不大，但彩色头像的公众号要比黑白头像的公众号总阅读量更高一些，这可能是因为彩色头像色彩更加丰富，更能够带给用户视觉冲击感，比黑白头像的公众号更能在第一时间吸引用户眼球，从而引起用户的兴趣。

（五）自变量：发文频次

我们对发文频次这一维度的单次篇数进行分析。首先，我们看单次篇数的分布情况，如图 9。

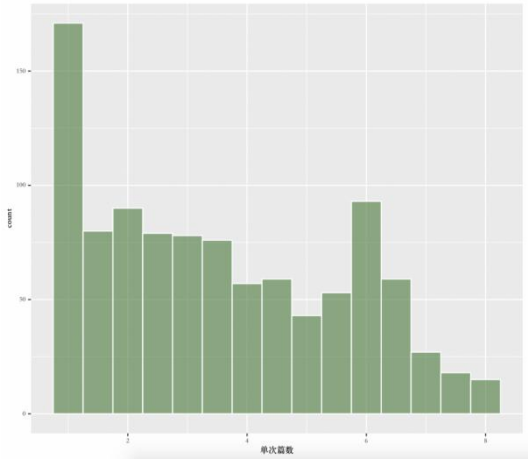

图 9 单次篇数的分布

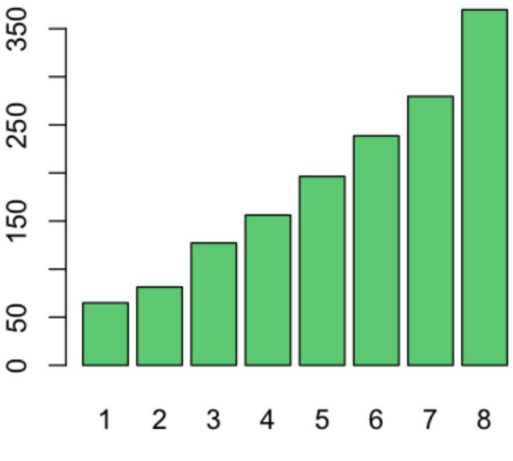

图 10 单次篇数与平均阅读量的关系图

其次，分析与平均总阅读量的关系。由图 10 可以得到，单次篇数与平均总阅读量呈正相关，随着单次篇数的增多，阅读量也会增多，这是由于篇数的增多可以增加文章的宣传能力，从而多方面吸引客流量。

（六）自变量：发文类型

我们对发文类型这一维度的两个变量——原创占比与视频占比进行分析。

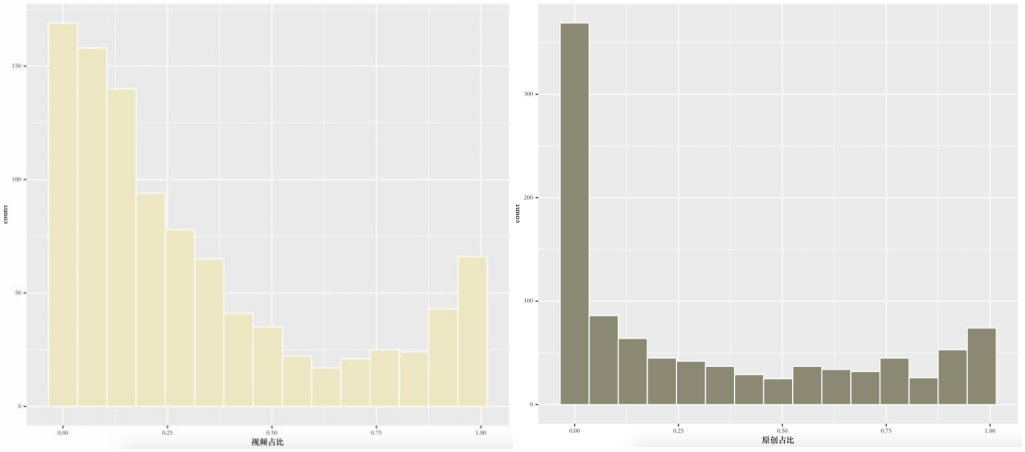

图 11 发文类型的分布

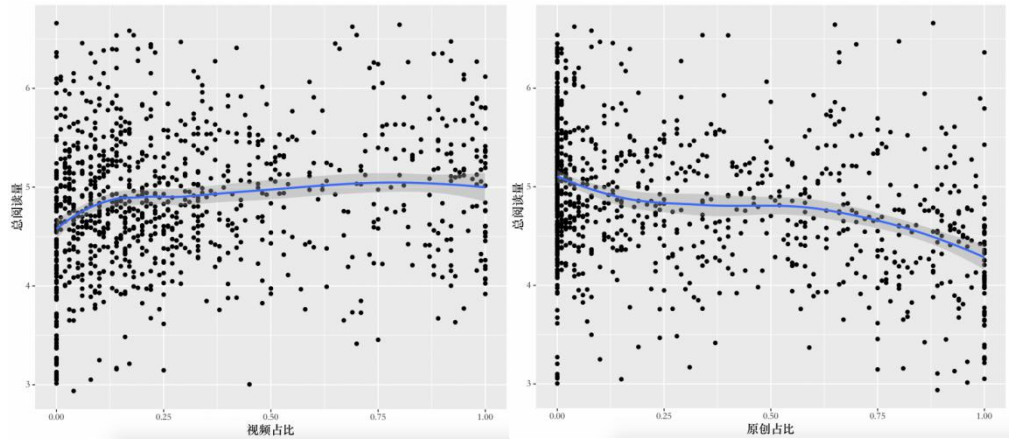

图 12 发文类型的散点图

由图可以看出视频占比集中在 0～0.2 之间，和原创占比集中在 0～0.1 之间。可以得到视频和原创作品在公众号中并不算主流，这可能是由于原创和视频的制作更有难度，因此更多人选择文字形式的文章。由散点图可以得到视频占比对总阅读量影响不大。而随原创占比的提高，总阅读量呈下降趋势，可能由于大部分原创类公众号较为小众，传播性不广，很难被人群看到。

（七）自变量：发文时间

我们对发文时间这一维度的 6 个时段进行分析。从晚上 18 点开始，每 4 个小时是一个时间段。由下图得到，40.2%的公众号都集中在第一时段发文，也就是晚上 18 点～22 点之间。这一时间段正是人们结束了一天的学习工作之后放松的时候，公众号在该时间段发文更有可能被读者阅读。

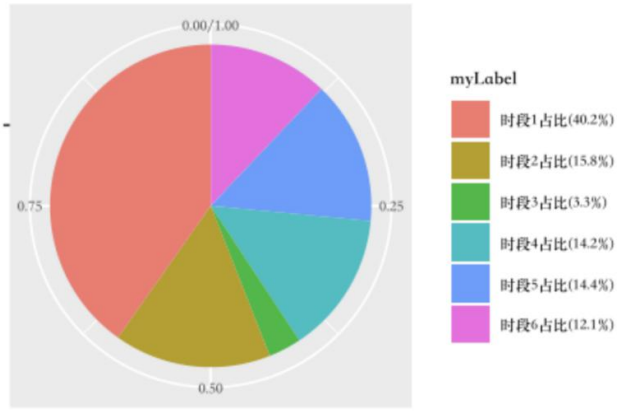

图 13 时间段占比

其次，我们根据公众号在 6 个时段的发文比例，通过信息学领域中的熵变量来衡量其发文的规律性。熵的取值越高说明发文时段越分散，熵的取值越小说明发文时段越集中。

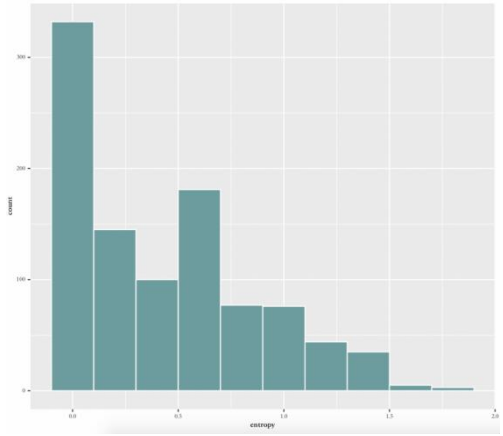

图 14 发文规律性的分布图

从图 14 中我们看到，大部分公众号的发文时段都比较规律。这一方面可能是由于公众号由专人管理，有固定的工作时间，另一方面可能是因为规律的发文更能培养人们的阅读习惯。

（八）自变量：文章标题

标题是文章的门面，人们往往根据标题来判断是否要阅读该文章，因此标题的一些特性很可能会影响其阅读量。我们对文章标题这一维度的三个变量进行分析，分别为标题长度、标题标点指数和标题正向情感得分。

通过绘制每个公众号平均标题长度的分布图如图 15，我们发现大部分标题集中在 15~25 个字符之间。因此，我们以 15、25 为分界线，将标题长度划分为“短”“适中”和“长”三个类别，并绘制了不同类别下公众号总阅读量的箱线图。如下右图：

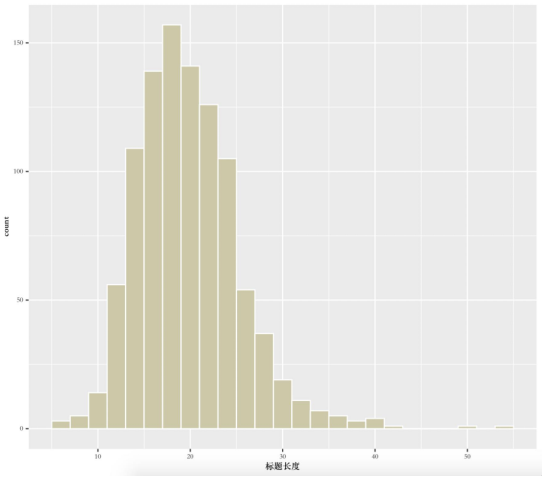

图 15 平均标题长度的分布图

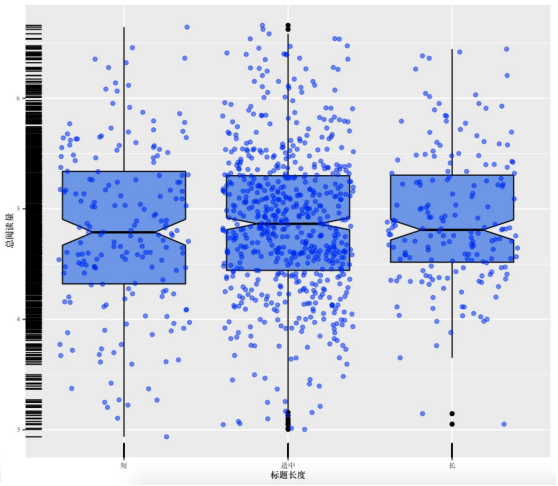

图 16 平均标题长度的分布图

我们发现，标题长度适中的公众号，其总阅读量略高于其他两种类型，因此适当的标题长度可以增加阅读了。此外，很多文章标题都会使用感叹号、问号、省略号等营造大消息、大新闻的效果来吸引读者，因此我们计算了包含感叹号、问号、省略号等特殊标点符号的文章标题在 10 个标题中所占比例，把这个定义为标点指数。

我们进一步分析标题标点指数，首先，绘制标点指数分布图，如下左图：

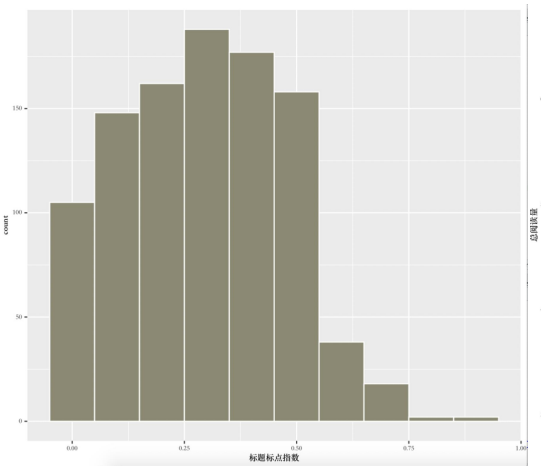

图 17 标题标点指数的分布图

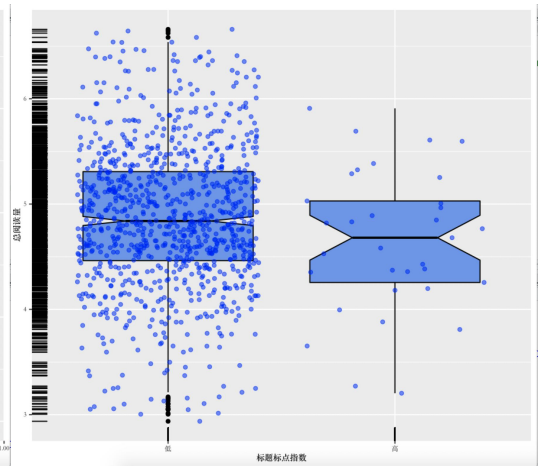

图 18 标题标点指数的箱线图

通过标点指数的分布，发现 0.6 左右的分布差异较大，因此将大于 0.6 的定义为“高”，反之为“低”。从下面的箱线图可以看到，标点指数低的总阅读量更高一些，这是由于更多的新闻等阅读量较高的公众号更倾向于应用较严谨的语言，因此标点用的会更少更谨慎一些。

除标点指数外，我们希望更细致的去探索标题表达情感的情况。由情感分布图图 19 可以看出在 0.4～0.6 之间标题正向情感得分最多，因此我们将得分离散化。小于 0.4 的得分记作“低”，大于 0.6 的得分记作“高”，0.4～0.6 的得分记作“适中”，而后画出箱线图。

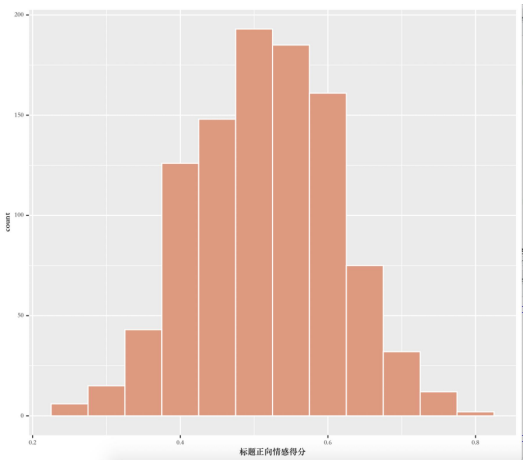

图 19 标题正向情感得分分布图

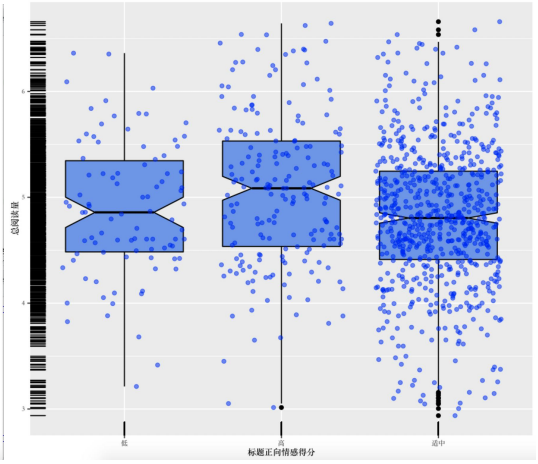

图 20 标题正向情感得分箱线图

由图 20 看出正向情感得分高的阅读量更高一些，正向情感得分低和适中的阅读量相差不多，这是由于人们更喜欢阅读正能量的文章，从中获取力量，因此正向情感更高的阅读量也更多。

四、数据建模

（一）回归分析

为了进一步探究上述因素与总阅读量之间的关系，我们建立线性回归模型，其中因变量为取对数后的总阅读量，自变量包括公众号类型、公众号类别等原始自变量以及名称长度、标题长度等离散化后的自变量，回归模型的结果如下表：

表 2 回归结果（因变量：总阅读量）

变量	估计值	t 统计量	p 值	显著性
截距项	10.11886545	2.776104949	0.005607886	**
公众号类型个人	-0.036145531	-0.611067637	0.54129871	
公众号类型企业	0.066225476	1.188785122	0.234816687	
公众号类别地方	0.251870557	2.924983023	0.003525381	**
公众号类别搞笑娱乐	0.479437033	6.812915576	1.68E-11	***
公众号类别明星影视	0.565639524	6.296112466	4.63E-10	***
公众号类别女性时尚	0.470894393	5.97103893	3.31E-09	***
公众号类别其他	0.310958962	3.996909511	6.91E-05	***
公众号类别企业	0.62327051	6.314326573	4.13E-10	***
公众号类别情感励志	0.482177232	6.949982566	6.72E-12	***
公众号类别生活	0.358582898	4.542973241	6.25E-06	***
公众号类别时事资讯	0.248916264	3.132840158	0.001783443	**
公众号类别文化教育	0.279351107	3.975232245	7.56E-05	***
公众号类别运动健康	0.374856186	4.733246203	2.54E-06	***
名称长度长	-0.052420809	-1.174032911	0.2406715	
名称是否包含英文含英文	-0.032925404	-0.576472839	0.564430055	
头像是否含文字含文字	-0.044213183	-1.54377152	0.122971545	
头像彩色黑白	-0.08614889	-2.365004622	0.018226864	*
日均发文次数	-0.013021855	-0.538872426	0.590099173	
单次篇数	0.25547329	31.09207978	1.34E-147	***
原创占比	0.09707736	2.061514061	0.039521402	.
视频占比	-0.020853654	-0.431207753	0.666413624	
时段 1 占比	-6.489876241	-1.780761713	0.075265901	.
时段 2 占比	-6.543611959	-1.794370433	0.073066979	.
时段 3 占比	-6.462248227	-1.773853333	0.076402708	.

时段 4 占比	-6.637974469	-1.821832259	0.068789884	*
时段 5 占比	-6.549727327	-1.79655739	0.072718565	.
时段 6 占比	-6.581226868	-1.806380265	0.071170399	.
标题长度适中	-0.008737506	-0.233677066	0.815285255	
标题长度长	0.110950206	2.258415487	0.024142378	*
标题标点指数高	-0.220094432	-2.850417969	0.004459174	**
标题正向情感得分高	0.086130763	1.538218628	0.12432298	
标题正向情感得分适中	-0.0332106	-0.687173281	0.492138647	

注： ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1‘ ’ 1

从回归结果可以看出，对总阅读量有显著影响的变量有：公众号类别、头像是否为彩色、单次发文篇数、原创占比、文章标题长度、文章标点指数。回归模型的调整后 R^2 为 0.6153，表明该模型的拟合效果较好。同时，我们对模型进行了回归诊断，如图 21，发现因变量服从正态分布，呈现一个较好的结果，因此我们可以认定模型是正确的。

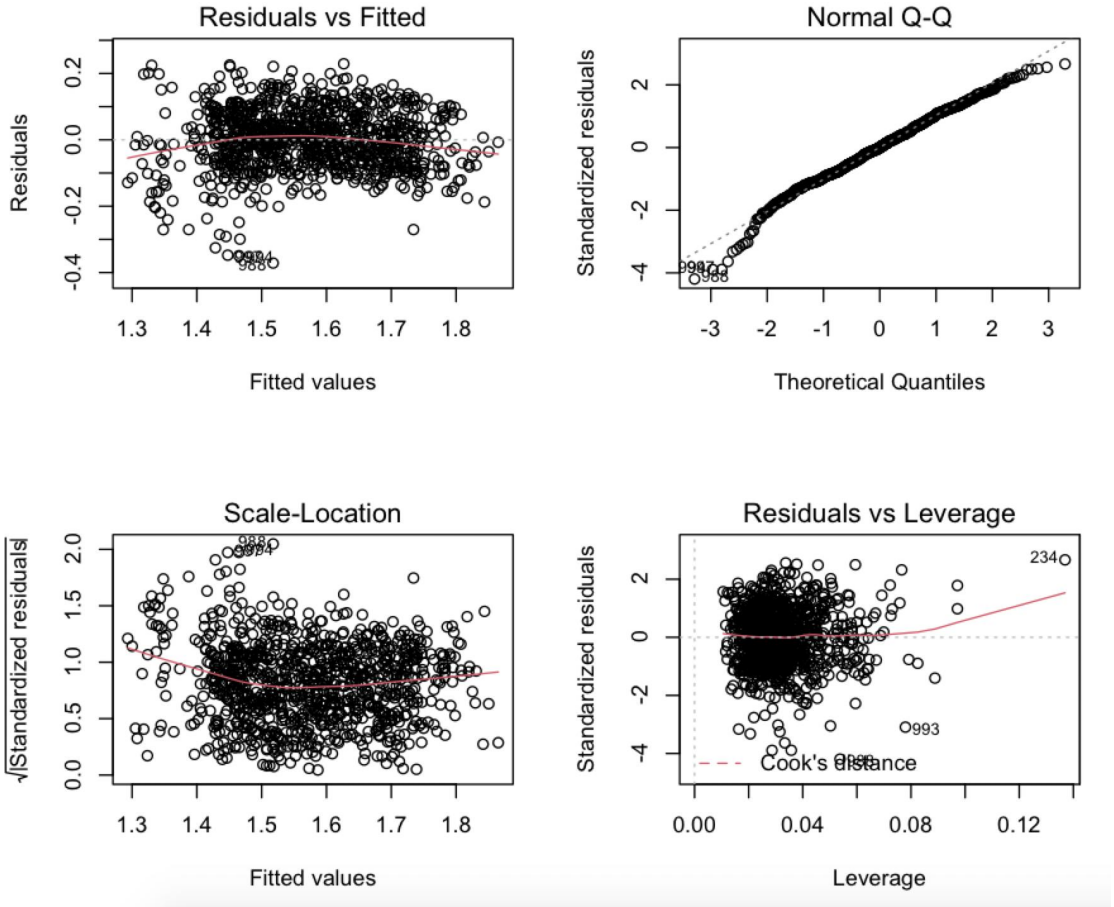

图 21 模型的回归诊断

下面，我们对回归模型的结果进行解读。我们发现，搞笑娱乐类公众号的总阅读量最多、其次是时事资讯、情感励志、生活、运动健康类，说明人们更喜欢

有趣的文章。头像为彩色的公众号其总阅读量更高，说明彩色头像更能在推文的公众号列表中被注意到。单次发文篇数越多，总阅读量越高，这非常符合情理，说明公众号可以通过多推文章来扩大受众；原创占比越高，其总阅读量越低，这是由于大部分原创类公众号较为小众，传播性不广，很难被人群看到。在控制了其他变量以后，同文章标题长度适中的公众号的公众号有更多的总阅读量，这可能是因为适中长度的标题能描述清楚文章的内容和思想且不繁琐，更加让读者有耐心看完整；而文章标点指数越高，总阅读量越低，说明通过使用标点符号来吸引眼球的套路走不通。最后，标题的正向情感得分越高，总阅读量越高，说明人们对于有着明显情感倾向的文章更有好奇心，更想要阅读。

（二）多重共线性分析

我们接下来分析可能的多重共线性，由 R 程序可以得到表 3:

表 3 多重共线分析

	GVIF	自由度	GVIF^(1/(2*Df))
公众号类型	1.168653758	2	1.03973213
公众号类别	2.941324633	11	1.050261407
名称长度	1.349288736	1	1.161588884
名称是否包含英文	1.381069903	1	1.175189305
头像是否含文字	1.14870449	1	1.071776325
头像彩色	1.059247281	1	1.029197397
日均发文次数	1.289089395	1	1.135380727
单次篇数	1.566061332	1	1.251423722
原创占比	1.540824956	1	1.241299704
视频占比	1.244026852	1	1.115359517
时段 1 占比	11867.88426	1	108.9398194
时段 2 占比	6819.61626	1	82.58096791
时段 3 占比	1739.89178	1	41.71201002
时段 4 占比	6655.801983	1	81.58309864
时段 5 占比	5897.015459	1	76.79202731
时段 6 占比	4226.640975	1	65.01262166
标题长度	1.402650675	1	1.184335542
标题标点指数	1.429507877	1	1.19562029
标题正向情感得分	1.137895223	1	1.066721718

由表 3 知，6 个时段占比的 GVIF 值很大，因此它们之间存在多重公线性。我们又计算了这 6 个变量的 Kappa 值，结果为 55709.79，印证了多重共线性。我们接下来，利用逐步回归法筛选变量，发现时段 3 占比被筛选出去。

五、结论与建议

通过上述结果，我们可以为经营公众号的人们提供一定的可信数据，让他们更加了解如何经营公众号可以最大程度的增加阅读量，比如文章标题可以写的长一些吸引眼球等。同时，我们可以进一步为各公众号进行分析，为其提供增加阅读量的定制化服务。首先，可以对公众号当前状况进行全方位诊断，然后根据回归模型结果，为公众号提升阅读量提出具体建议。

对于正在经营公众号的人群，我们也提出以下建议：可以在选择在人们阅读量大的时间段进行发文（18～22 点），适当增加公众号的乐趣吸引更多读者。同时要注意头像黑白、单次篇数较少、标题长度不够长等问题加以改进，这些都是可能影响阅读量的因素，而无需从文章标点上入手。

附录 R 语言程序代码

```
##读入数据
rs <- read.csv('/Users/yuqinhan1229/Desktop/data.csv')
##导入所需包
library(ggplot2)
library(car)
##总阅读量的分布
###ggplot 包画图
ggplot(rs, aes(x=总阅读量)) + geom_histogram(aes(y=..density..),
                                              color= 'white',
                                              fill = 'orange',
                                              alpha=.5,   ##设置透明度
                                              binwidth = 30,
                                              center = 0)+
geom_density(color='goldenrod',alpha=.5) +
  theme(text=element_text(family="Songti SC",size=12,face = "bold"))
##因变量对数变换
rs$总阅读量 <- log(rs$总阅读量)
ggplot(rs, aes(x=总阅读量)) + geom_histogram(aes(y=..density..),
                                              color= 'white',
                                              fill = 'orange',
                                              alpha=.5,
                                              binwidth = 0.2,
                                              center = 0)+
geom_density(color='goldenrod',alpha=.5) +
  theme(text=element_text(family="Songti SC",size=12,face = "bold"))



##探索不同的公众号类型、类别与总阅读量的关系
###不同的公众号类型与总阅读量的关系
rs <- read.csv('data.csv') #重新读入数据
ggplot(rs, aes(x=公众号类型, y=总阅读量)) +
  geom_boxplot(fill="cornflowerblue",
               color="black", notch=TRUE)+
  geom_point(position="jitter"##位置分散
```

```
                , color="blue", alpha=.5##透明度
  )+
  geom_rug(color="black")+
  theme(text=element_text(family="Songti SC",size=8,face = "bold"))
###不同的公众号类别与总阅读量的关系
options(repr.plot.width=20, repr.plot.height=4)
ggplot(rs, aes(x=公众号类别, y=总阅读量)) +
  geom_boxplot(fill="cornflowerblue",
                   color="black", notch=TRUE)+
  geom_point(position="jitter"##位置分散
                , color="blue", alpha=.5##透明度
  )+
  geom_rug(color="black")+
  theme(text=element_text(family="Songti SC",size=10,face = "bold"))


##探索名称与阅读量之间关系，查看名称长度的分布情况，查看是否包含英文与平均阅读量的关系
###查看名称长度的分布情况
options(repr.plot.width=4, repr.plot.height=4)
ggplot(rs, aes(x=名称长度)) + geom_histogram(
  color= 'white',
  fill = 'darkgreen',
  alpha=.5,
  binwidth = 1,
  center = 0) + theme(text=element_text(family="Songti SC",size=10,face = "bold"))
###查看是否包含英文与平均阅读量的关系
options(repr.plot.width=4, repr.plot.height=4)
ggplot(rs, aes(x=名称是否包含英文, y=总阅读量)) +
  geom_boxplot(fill="cornflowerblue",
                   color="black", notch=TRUE)+
  geom_point(position="jitter"##位置分散
                , color="blue", alpha=0##透明度
  )+
  geom_rug(color="black")+
  theme(text=element_text(family="Songti SC",size=8,face = "bold"))
```

```
##使用箱线图探索头像是否包含文字与头像是否彩色与总阅读量的关系
###头像是否包含文字与总阅读量的关系
options(repr.plot.width=4, repr.plot.height=4)
ggplot(rs, aes(x=头像是否含文字, y=总阅读量)) +
  geom_boxplot(fill="cornflowerblue",
                    color="black", notch=TRUE)+
  geom_point(position="jitter"##位置分散
                , color="blue", alpha=.5##透明度
  )+
  geom_rug(color="black")+
  theme(text=element_text(family="Songti SC",size=8,face = "bold"))
###头像是否彩色与总阅读量的关系
options(repr.plot.width=4, repr.plot.height=4)
ggplot(rs, aes(x=头像彩色, y=总阅读量)) +
  geom_boxplot(fill="cornflowerblue",
                    color="black", notch=TRUE)+
  geom_point(position="jitter"##位置分散
                , color="blue", alpha=.5##透明度
  )+
  geom_rug(color="black")+
  theme(text=element_text(family="Songti SC",size=8,face = "bold"))


##探索单次发文篇数和平均总阅读量的关系
###单次发文篇数
options(repr.plot.width=4, repr.plot.height=4)
ggplot(rs, aes(x=单次篇数)) + geom_histogram(
  color= 'white',
  fill = 'darkgreen',
  alpha=.5,
  binwidth = 0.5,
  center = 0) + theme(text=element_text(family="Songti SC",size=8,face = "bold"))
rank = ceiling(rs$单次篇数)
rss= tapply(rs$总阅读量,rank,mean)
barplot(rss,col='springgreen3',ylabel='总阅读量',xlabel='单次篇数')
```

```
##探索发文类型（原创、视频）、发文时间、标题情感等其它因素对阅读量影响
###发文类型
par(mfrow=c(1,2))
#原创
ggplot(rs, aes(x=原创占比)) + geom_histogram(
   color= 'white',
   fill = 'lemonchiffon4',
   binwidth = 0.07,
   center = 0) + theme(text=element_text(family="Songti SC",size=8,face = "bold"))
#视频
ggplot(rs, aes(x=视频占比)) + geom_histogram(
   color= 'white',
   fill = 'lemonchiffon2',
   binwidth = 0.07,
   center = 0) + theme(text=element_text(family="Songti SC",size=8,face = "bold"))

###信息熵+饼状图探索发文时间与阅读量的关系
rs_2 = rs[,c(13:18)]
mean = colMeans(rs_2, na.rm=TRUE)
dt = data.frame(mean)
B <- c("时段 1 占比", "时段 2 占比", "时段 3 占比", "时段 4 占比","时段 5 占比","
时段 6 占比")
A = round(as.vector(dt$mean),3)
myLabel = as.vector(B)
myLabel = paste(myLabel,"(", round(A *100/ 1, 2), "%)", sep = "")
p = ggplot(dt, aes(x = "", y = A, fill = myLabel)) + #创建坐标轴
   geom_bar(stat = "identity") +
   coord_polar(theta = "y")      +
   labs(x = "", y = "", title = "") +
   theme(text=element_text(family="Songti SC",size=8,face = "bold"))
print(p) #显示饼图
rs_3 = rs_2 * log(rs_2)
rs_3[is.na(rs_3)] =0
rs_2$entropy <- abs(rowSums(rs_3[, 1:6])) ##计算信息熵
ggplot(rs_2, aes(x=entropy)) + geom_histogram(
```

```
    color= 'white',
    fill = 'cadetblue',
    binwidth = 0.2,
    center = 0) + theme(text=element_text(family="Songti SC",size=8,face = "bold"))

###标题长度与阅读量的关系
options(repr.plot.width=4, repr.plot.height=4)
ggplot(rs, aes(x=标题长度)) + geom_histogram(
    color= 'white',
    fill = 'lemonchiffon3',
    binwidth = 2,
    center = 0) + theme(text=element_text(family="Songti SC",size=8,face = "bold"))
rs <- read.csv('data.csv')
###将长度离散化
attach(rs)
rs$标题长度[标题长度<=15]="短"
rs$标题长度[标题长度>15 &  标题长度< 25]="适中"
rs$标题长度[标题长度>= 25]="长"
detach(rs)
options(repr.plot.width=7, repr.plot.height=4)
ggplot(rs, aes(x=标题长度, y=总阅读量)) +
    geom_boxplot(fill="cornflowerblue",
                         color="black", notch=TRUE)+
    geom_point(position="jitter"##位置分散
                       , color="blue", alpha=.5##透明度
    )+
    geom_rug(color="black")+
    theme(text=element_text(family="Songti SC",size=8,face = "bold"))

##同理操作标题标点指数
options(repr.plot.width=4, repr.plot.height=4)
ggplot(rs, aes(x=标题标点指数)) + geom_histogram(
    color= 'white',
    fill = 'lemonchiffon4',
    binwidth = 0.1,
    center = 0) + theme(text=element_text(family="Songti SC",size=8,face = "bold"))
rs <- read.csv('data.csv')
```

```
attach(rs)
rs$标题标点指数[标题标点指数>0.6]="高"
rs$标题标点指数[标题标点指数<=0.6]="低"
detach(rs)
options(repr.plot.width=7, repr.plot.height=4)
ggplot(rs, aes(x=标题标点指数, y=总阅读量)) +
  geom_boxplot(fill="cornflowerblue",
                        color="black", notch=TRUE)+
  geom_point(position="jitter"##位置分散
                      , color="blue", alpha=.5##透明度
  )+
  geom_rug(color="black")+
  theme(text=element_text(family="Songti SC",size=8,face = "bold"))

##情感正向得分与阅读量的关系
ggplot(rs, aes(x=标题正向情感得分)) + geom_histogram(
  color= 'white',
  fill = 'darksalmon',
  binwidth = 0.05,
  center = 0) + theme(text=element_text(family="Songti SC",size=8,face = "bold"))



##对数线性回归
rs$总阅读量 <- log(rs$总阅读量)
rs.fit <- lm(总阅读量~ 公众号类型 + 公众号类别 + 名称长度 +
名称是否包含英文+头像是否含文字+头像彩色+日均发文次数+单次篇数 +
原创占比+视频占比+时段 1 占比+时段 2 占比+时段 3 占比+时段 4 占比+时段 5
占比+时段 6 占比+
  标题长度+标题标点指数+标题正向情感得分,data=rs)
summary(rs.fit)
##模型的回归诊断
opar<-par(no.readOnly=TRUE)
par(mfrow = c(2, 2))
plot(rs.fit)
par(opar)
###检验多重共线性
vif(rs.fit)
```

```
####计算 kappa
XX<-cor(rs[, 13:18])
kappa(XX,exact=TRUE)
eigen(XX)
rs.d<-rs[, c(2,13:18)]
lm<-lm(总阅读量~.,data=rs.d)
lm_step<-step(lm,direction="both")
summary(lm_step)
###删掉时段 3 占比后的模型
rs.fit2 <- lm(总阅读量~ 公众号类型 + 公众号类别 + 名称长度 +
                        名称是否包含英文+头像是否含文字+头像彩色+日均发文
次数+单次篇数 +
                        原创占比+视频占比+时段 1 占比+时段 2 占比+时段 4 占比
+时段 5 占比+时段 6 占比+
                        标题长度+标题标点指数+标题正向情感得分,data=rs)
summary(rs.fit2)
```