Multidimensional linear regression analysis to predict the change in one-day video views on Bilibili website

Jixing Lu[1, *] and Yifan Xia[2]

[1] School of Mathematical Sciences, Nankai University, Tianjin, China

[2] Department of Mathematics, Shanghai University, Shanghai, China

* Corresponding author: 2212103@mail.nankai.edu.cn

Abstract. This paper utilizes a multidimensional linear regression analysis model to investigate video playback volume changes within Bilibili's film, cartoon, anime, auto-tune remix-themed content, game, and entertainment partitions. We employ crawler software and a webpage loop with a nested collection of video links to extract and sequence titles. Technical abbreviations are defined when first presented. We utilize crawler software to extract various data sets from online video links. We establish a web page cycle and collected video links within nested collections to accomplish this. Subsequently, they systematically accessed each link to extract information such as titles, playback, likes, coins, favorites, and video synopses. Finally, the data 6,075 are obtained and filtered down to 1,691 representative data points using playback, likes, coins, favorites, and other data sets from the previous day. The previous day's playback, likes coins, favorites, and other data are also collected. After conducting a multidimensional linear analysis, we evaluate the Durbin-Watson test, VIF, and R^2 to determine the accuracy of fit. Our findings indicate that collecting videos and playback from the previous day positively impacts playback for the following day. However, pursuing likes andcoins excessively is susceptible to restriction by the system.

Keywords: Linear regression analysis; Bilibili; Durbin-Watson test.

1. Introduction

Computer networks have developed by leaps and bounds in recent decades, bringing about tremendous changes in human life. Computer networks have become an indispensable part of today's society. In this era of information technology, people use computer networks almost everywhere in their daily lives. Daily communication, online entertainment, network payment and other advanced functions make people's lives more convenient.

Since entering the new century, China's Internet has developed very rapidly. According to the 52nd Statistical Report on the Development of the Internet in China [1], as of June 2023, 1.079 billion people in China were using the Internet, and the penetration rate had reached 76.4 per cent. It can be said that the Internet is already very popular in China. And online video, as a new type of entertainment means on the Internet, has gained wide popularity among Internet users because of its rich content, low threshold, interactivity and other characteristics. According to the 52nd Statistical Report on the Development of the Internet in China, as of June 2023, there were 1.044 billion Internet users using online video, accounting for 96.8% of the total number of Internet users, which is larger than the number of users of online payment, online shopping and other Internet industries that are also very hot at the moment.

Among the many online video platforms, BiliBili is one of the more special ones. In contrast to other online video platforms that provide professional content (PGC), BiliBili has taken a different approach by establishing a video community that focuses on user-generated content (UGC). It allows users to act not only as recipients or consumers of content, but also as content creators. This has increased user participation, and people are very willing to make a video and upload it to BiliBili to share their interesting life stories, opinions and ideas, and so on.

When a video is uploaded on BiliBili, the site does a basic push of the video. If the video is well received, it will be pushed more often, which means more people will see it. If the video is poorly received or doesn't meet certain norms, it may be pushed less frequently and be subject to traffic

restrictions. The number of plays of a video is very important for video uploaders, which may determine how much they earn. Therefore, there is a need for a way to predict the playback amount a video will receive after it is posted, so that uploaders can see if their videos match up with them, see if the content is in line with today's trends, and use this to decide what to create next.

Regarding to the prediction of online video playback amount, there is some research. Since online video itself is popular for a relatively short period of time, its research mainly focuses on the last ten years.

In the paper written by Liu in 2012 [2], the seasonal autoregressive integrated moving average (SARIMA) model in time series forecasting was used to predict a set of video playbacks on the online video platform of PPTV. However, the popularity of network video at that time is completely incomparable with that of now, so the reference is not significant.

In a paper written by Wang in 2021[3], the author studies the problem of predicting the playback of short videos, and it is for the study of the playback time series after the video is online. The Network Vector Autoregressive (NAR) model used in this paper. In order to target the diversity of trends in that time series, this paper uses the K-means clustering algorithm for clustering, and the trends in playback volume are grouped into three categories.

There are two papers written in 2022 which both studied pop-up videos on BiliBili. The former uses ELM as the theoretical framework to construct a model of the factors influencing the playback amount of pop-up videos [4], and by capturing the video data from BiliBili website, the model is validated by a combination of hierarchical regression and fsQCA methods, and the stable playback is studied [5]. The latter combines feature selection and different algorithms to construct random forest model, XGBoost model and LSTM model respectively to predict the playback amount of pop-up videos, and study the short-term playback volume in 24 hours, but it only studied the videos that are in the section of daily life.

Multidimensional linear regression is a usual way to predict what may happen in the future based on the existing data. It has been widely used in scientific researches. Wei et al use Multidimensional linear regression to predict air quality in Guilin during spring festival [6]. Xiao et al use Multidimensional linear regression to predict ethanol concentration from gas mixture spectrum [7]. Wang et al use Multidimensional linear regression to predict pulmonary tuberculosis based on Baidu index [8].

2. Research methods

2.1. Model instruction

Multiple regression analysis, or multiple regression for short. That is, a regression equation is used to portray the relationship between a dependent variable and multiple independent variables. It can usually be written as:

$$Y = XB + \varepsilon,$$

where Y is the column vector of the dependent variable $\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$. $X = \begin{pmatrix} 1 & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{np} \end{pmatrix}$, coefficient matrix $B = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$. $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ is the column vector of errors. The coefficient matrix can be solved according to the least squares method $B = (X^T X)^{-1} X^T Y$.

Using the strong interactivity of Bilibili, the behavioural information is formed from browsing, likes, coins and favourites. They have undergone principal component analysis for multidimensional linear regression analysis in order to carry out the prediction of video playback amount. From the B-site platform, the variable metrics to be crawled were identified based on the hypothetical model of

the article. Using the Octopus crawler software's network cycle scrolling, cycle page flip and web page nesting function, crawling rules were set and multiple web page nested crawling were carried out to obtain video information data, video creator information data, and viewer interaction information data.

The theoretical basis is to take the video playback amount, likes, coins and favourites on the first day of the video as the dependent variable, and the video playback amount after 24h as the independent variable, and after multidimensional linear regression modelling, a series of tests such as the Durbin Watson test, R^2 test and residual analysis are used to determine the degree of model fit and the accuracy of the predicted values.

2.2. Data collection

The authors use Octopus crawler software to collect data on Bilibili, respectively, in the film, cartoon, anime, auto-tune remix-themed content, game, entertainment 6 video sub-districts by page information flip crawling. The specific steps are to enter a single video sub-district, set the cycle of scrolling and capture the current page of the video link, through the cycle of nesting sequentially open the link to capture the playback amount, coins, Likes, favourites, and other corresponding data. Further, we capture the author information in the profile and the video title, which is to facilitate the subsequent collation and analysis of the data.

2.3. Data collation

After using Octopus software to collect data, a total of 6,075 Bilibili video data were collected. Each piece of data contains 6 variables covering likes, coins, favourites, plays, titles, authors. After using Octopus software to collect the data, the data with a release date of less than 3 days from the date of collection (27th to 30th September 2023) were screened out. Meanwhile, videos released by official accounts of government units, media organizations, higher education institutions, enterprises and institutions were excluded, in order to exclude the analysis bias caused by the large increase of video releases in a short period of time due to holidays, emergencies, and so on, as well as the soaring of pop-ups and likes. Since many data are almost never changed throughout the day, and the relevant variables we need (likes, coins, etc.) are also maintained at 0, they are not representative and are not included in the statistics. Duplicates are also de-duplicated. A certain amount of data is taken from each playback interval. After removing invalid and missing values, finally there are 1,691 pieces of data. The 1691 data covers videos from 500 to 1,000,000 plays, so it is representative for the multidimensional linear regression analysis. We use Videoviews1 to represent the number of plays on the first day and Videoviews2 to represent the number of plays on the second day.

Table 1. Model test

R	R*R	Adjusted R*R	Error in standard estimates	change				Durbin-Waston test
				Change of R*R	Change of F	DOF	Significance Change	
0.985	0.971	0.971	15802.002	0.971	13912.652	4	0.000	2.147

R-square ratio is used to judge the fit of the data. It was observed that the R-square ratio reached 0.971 meaning that 97.1 per cent of the data can be explained by model.

Durbin-Waston is used to test serial correlation where the random error term has a first-order autoregressive form, that is, it is used for autocorrelation test, and the value of Durbin-Waston is between 1 and 4 and close to 2, which indicates that there is a more obvious serial correlation and there is a higher probability that observations are independent of each other.

Table 2. Coefficient

Name	Unnormalized coefficient		Standard coefficient	t	Significa-nce	Confidence interval		Collinearity statistics	
	B	Stan-dard error	Beta			lower limit	upper limit	Toleran-ce	VIF
Const-ant	961.990	401.208		2.398	0.017	175.071	1748.909		
Videoviews1	1.203	0.010	1.004	125.443	0.000	1.184	1.222	0.272	3.670
Like	-1.869	0.172	-0.151	-10.862	0.000	-2.206	-1.531	0.090	11.128
Coins	-3.363	0.606	-0.055	-5.549	0.000	-4.552	-2.175	0.176	5.683
Favori-tes	6.820	0.441	0.188	15.475	0.000	5.955	7.684	0.118	8.484

Variance inflation factor VIF=$\frac{1}{1-R^2} = Tolerance^{-1}$, R^2 is negative correlation coefficient between the current variable and other variables. So only Tolerance is needed to be considered here. Notice that the Tolerance of Likes is less than 0.1, that is, there is covariance, but considering that in the Bilibili website, the video playback depends largely on the system's push mechanism, and the Avalon mechanism of Bilibili, that is, the system will automatically limit the flow of a part of the video to control the amount of its likes, favourites, and coins, and the Like and the favourites each decide the amount of the likes, favourites and coins. The system will automatically limit the streaming of some videos to control the amount of likes, favourites and coins.

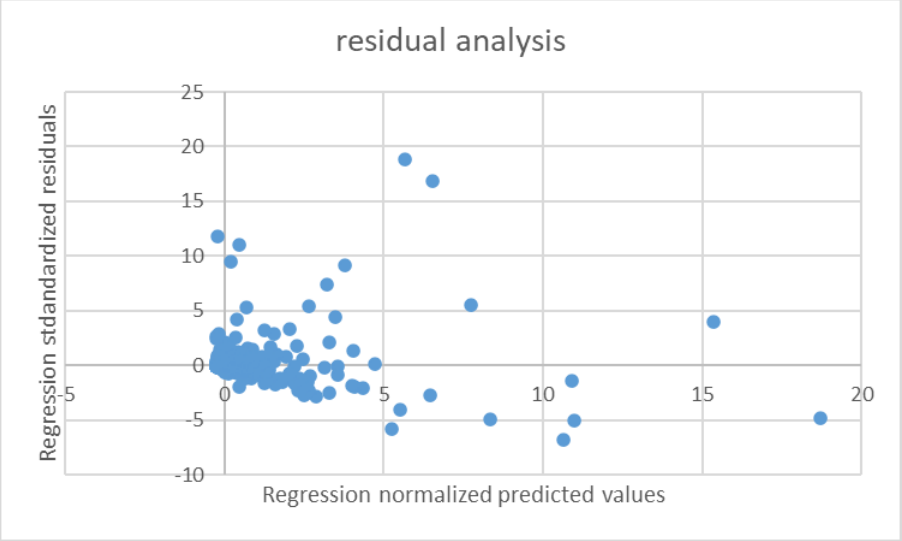


Fig. 1 Residual analysis

In the scatterplot of regression standardized predicted values and regression standardized residuals, only 12 data exceed 5, and the residuals of the rest of the data are in the vicinity of 0. Therefore, the difference between the predicted values and the true values are relatively satisfying.

3. Conclusion

Through the analysis of the model, video playback amount on the second day is mainly positively correlated with video playback amount and the number of favourites on the first day, the high video playback amount determines the degree of attention of the video in the webpage to increase the click rate, and the number of favourites means that users will watch the video many times to increase the number of plays, which is also very consistent with the reality of life. The number of likes and coins is negatively correlated, which can be explained by the fact that Bilibili mainly pushes videos with low number of likes and coins. If a video receives a lot of likes and coins, then the video uploader can gain a huge amount of revenue, which can be seen as a way for Bilibili to save money. Therefore, if the video author expresses the desire for coins in the video and receives a large number of likes and

coins, it may be counterproductive, and the pursuit of allowing viewers to favourite the video more to increase the number of plays so that the video is pushed multiple times with the platform to obtain more plays.

4. Discussions

There are some limitations in this study, so the following outlooks are proposed, which can be studied in more depth in the future.

There are 18 sub-districts in Bilibili, and the author mainly crawls the data in the six sub-districts which are popular among young people, but in recent years, there is a considerable audience in the life district, the technology district, and even the education and knowledge district which is affected by the epidemic. Since 2020, educational model has been catalyzed by the COVID-19 pandemic, leading colleges and universities to initiate unprecedented large-scale online teaching programs [9]. Therefore, the other sub-districts are still to be studied.

This model is a general multidimensional linear regression, but time series models as well as non-linear regression models can also be considered. Some other variables such as the number of comments, the number of author followers, and the number of subscribers may also be important factors reflecting the amount of playback, which are not covered in this model. The IEVP model combines an estimate with the average weight of content viewed and the number of video plays during a specific period to predict future video plays during a target time period [10].

The peak period may lead to playback growth is small due to the competition, and the trough period may lead to the same problem because there are fewer users online, which will inevitably affect the next day's growth of playback, so the time period for the release of the video is also a very important factor.

The above factors, all of which can create errors in the study, coupled with the authors' errors in the process of collecting the data, can be addressed in subsequent studies to further investigate the above issues.

Acknowledgements

All the authors contributed equally and their name were listed in alphabetical order.

References

[1] China Internet Network Information Center. 52nd Statistical Report on the Development of the Internet in China, 2023.

[2] Liu Yan. Design and implementation of video playback prediction software based on SARIMA model. Chengdu: University of Electronic Science and Technology of China, 2013

[3] Wang Zhe. Short video playback research based on network vector autoregressive modelling. Shanghai University of Finance and Economics, 2022.

[4] Chen Minghong, Huang Jiale, Fang Shishen, Huang Hanhui. Influencing Factors and Configuration Effect of the View Counts of Bullet-screen Videos. Library Tribune, 2022, 42(6): 150-160.

[5] Yang Li, Qin Jiangtao. Short-term Play Volume Forecast and Analysis of Factors Influencing Pop-up Videos. Computer & Digital Engineering, 2022, 50(09): 2012-2017+2078.

[6] Wei Pengzhi, Xie Shaofeng, Zhou Zhihao, Zhang Wei, Mo Zhixiang, Huang Liang ke. Prediction of PM2.5 concentration under holiday effect. Journal of Guilin University of Technology, 2022, 42(02):455-460.

[7] Xiao Xiongliang, Chen Changmin. Establishment and Verification of Multivariate Linear Regression Model for Prediction of Ethanol Concentration. Infrared Technology, 2021, 43(12): 1228-1233.

[8] Wang Yue, Zhou Haitao, Yue Tingyu, Chen Wei, Hu Bin. Research of prediction model of pulmonary tuberculosis based on Baidu index. Disease Surveillance.

[9] Huan Luo, Research on the Development of Online Education in Colleges and Universities under the background under the Background of Internet and Education, Southeast University, 2023.

[10] Zhiyi Tan, Research on video heat prediction algorithm for online video services. Shanghai Jiao Tong University, 2022.