Analysis of public satisfaction on the webcast in Taiyuan city, Shanxi Province

Jianqin Sun[1, a], Zhe Liu[1, b]

[1]School of Statistics, Shanxi University of Finance and Economics, Taiyuan, Shanxi, China

[a]907399246 @qq.com, [b]776977959@qq.com

Keywords: Webcast; Chi-Square Test; Logistic Regression; SPSS; Python

Abstract: Webcast is to watch films on different platforms through the network system at the same time. It mainly composes of the living client, live web page, and management background. It is not only an emerging social network mode but also a brand new social media. This paper takes internet live broadcast user group in Taiyuan city as the research object, and researches on watching the webcast and their satisfaction with webcast for users over ten years old. The research content is divided into two parts, namely, users' habits of watching webcast and users' comments on the webcast. In the research of these two parts, the user groups were refined, and the statistical charts were used comprehensively, supplemented by chi-square test and logistic regression model, to study the characteristics of watching the webcast by user groups of different genders and ages, as well as the user satisfaction of webcast.

1. Introduction

According to the statistical report on the development of China's Internet network released by China Internet network information center (CNNIC) in February 2019, by December 2018, the number of webcast users had reached 397 million, with a user utilization rate of 47.9%, down 6.8% from the end of 2017. However, the future development of "webcast" is still optimistic by many people. Founder securities predict that the scale of China's live broadcast market will exceed 60 billion yuan in 2020. The advantage of the webcast is to catch the trend and spread a large amount of information in a short time, which significantly meets people's demand for information. However, at the same time, the hype also makes people hate webcast. Then, what factors are related to the public's satisfaction with webcast? In this paper, disordered multiple classification logistic regression was used for empirical analysis.

2. Data sources and fundamental statistical analysis

The data used in this paper are from the questionnaire survey on the current situation of network "vitality" or "crisis" webcast market. The survey scope is Taiyuan residents, and 365 questionnaires are investigated by sampling method. The proportion of male and female users is moderate, and the user group is concentrated between 10 and 30 years old, as shown in table 1.

Table 1 Sample structure information table

Category		Subtotal	Total	Proportion(%)	Total(%)
Gender	Man	175	365	47.55	100
	woman	190		52.45	
Age	10-20 years old	169	365	46.3	100
	20-30 years old	123		33.7	
	30-40 years old	34		9.32	
	Over 40 years old	39		10.68	
Whether or not to watch webcasts	Yes	269	365	73.91	100
	No	96		26.09	

According to the results of the survey, we found that the most common webcast platforms (top

  DOI: 10.25236/mfssr.2019.157

three) used by the respondents are: bilibili, douyu, huya. Therefore, according to the results of this survey, we use python software to conduct crawler on douyu, one of the platforms.

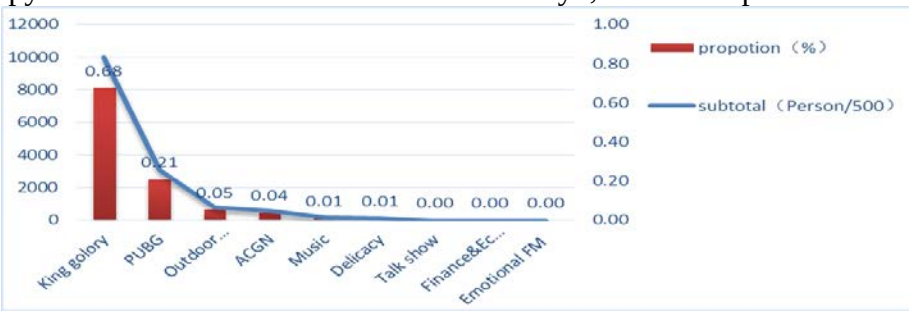

Fig.1 Crawler data of webcast platform

Fig.1 is the data of webcast content and online users on July 30, 2018, obtained by a python software crawler. It can be seen from the figure that the types of webcast content watched by users involve various games, beauty makeup, and quadratic elements, etc., and the number of online viewers of different types varies greatly. Therefore, it is necessary to conduct a detailed analysis.

3. Empirical analysis

(1) Preference for webcast content

The analysis of live content preferences, we according to a different gender, different ages live content preferences chi-square and draw the following conclusion: different gender in the online game competition, network program, science, and technology education, outdoor reality, cooking and eating, clothing, beauty makeup has a significant different preference.

Table 2 Chi-square analysis of ranking of live broadcast content and gender and age

Category	Gender		Age	
	X^2□	P□	X^2	P
Online game	63.981	0.000**	125.963	0.000**
Network program	19.372	0.013*	73.705	0.000**
Science and technology education	24.019	0.002**	51.753	0.015*
Outdoor reality show	15.542	0.049*	35.577	0.304
Cooking and eating seeds	35.109	0.000**	34.878	0.522
Animation and Cosplay	8.101	0.524	52.899	0.034*
Live talk show (telling ghost stories, interpreting movies)	12.492	0.187	50.952	0.05
Beauty makeup clothing	70.422	0.000**	53.908	0.009**
Others	20.515	0.009**	34.34	0.356

(2) Basic features of watching the webcast

According to the results of the survey, it is concluded that the primary purpose of users to watch webcast is pure entertainment, followed by "to satisfy their interests and hobbies" and "idle and boring," respectively accounting for 42.26% and 38.49%. Among them, people of different genders and ages watch webcast mainly from 18:00 to 22:00. Also, both gender and age are related to the viewing time, and the elderly watch more live broadcasts from 6:00 to 10:00.

(3) Evaluation of live broadcast content

➢ Model establishment

In the survey, it was found that 50.55% of the respondents rated the current webcast content as "normal," accounting for more than half. 41.33% were satisfied, 4.43% were "dissatisfied," and 3.69% were "very dissatisfied." It can be seen that users generally do not have a high evaluation of the current webcast content. To explore whether satisfaction is correlated with age and gender, we adopted the chi-square test.

H_0: Satisfaction has nothing to do with age H_1: Satisfaction is related to age

Table3 Tabulation of degree age

Degree	Age					
	Under 10	10-20	20-30	30-40	Over 40	Total
Highly Dissatisfaction	0	2	3	3	2	10
Dissatisfaction	0	5	6	1	0	12
Normal	2	71	47	11	6	137
Satisfaction	1	38	21	8	6	74
Very satisfactory	0	8	14	5	12	39
Total	3	124	91	28	26	272

Pearson chi2(16) = 40.5341 P= 0.001

Table 3 is the chi-square test information table, P-value =0.001<0.05, when the significance level =0.05, we can reject the null hypothesis, that is, satisfaction is related to age, and do regression analysis.

H_0: Satisfaction has nothing to do with gender H_1: Satisfaction is related to gender

Table4 Tabulation of degree gender

Degree	Gender		
	Man	Woman	Total
Highly Dissatisfaction	7	3	10
Dissatisfaction	6	6	12
Normal	56	81	137
Satisfaction	44	30	74
Very satisfactory	19	20	39
Total	132	140	272

Pearson chi2(4) = 8.6085 P = 0.072

Table 4 is the chi-square test information table, and the P-value is 0.072>0.05. When the significance level is 0.05, we cannot reject the null hypothesis, and there is no evidence that satisfaction is not related to gender. Therefore, logistic regression analysis was conducted on satisfaction and age to explore the relationship between them further.

- Model setting and parameter estimation

$P(Y=j)$ is the probability of rating J of public evaluation of webcast, j=1, 2, 3, 4, 5. $\frac{P(Y=j)}{P(Y=1)}$

Is the public's evaluation of webcast as grade j, and the evaluation of webcast as grade 1, that is, the evaluation of the chance ratio of very unsatisfactory (abbreviated as OR value), where Y = 1 is used as a reference classification; j=1, 2, 3, 4, 5, $P(Y<j)$ is the cumulative probability, that is, the public's evaluation of webcast is the sum of the probability of grade j and below-grade j. Based on the above definition of symbols, the disordered multivariate logistic model of public evaluation of webcast content can be set as follows:

$$\ln[\frac{P(Y \le j)}{1-P(Y \le j)}] = \alpha_j + \beta_1 x_{1\text{-}1} + \beta_2 x_{1\text{-}2} + \beta_3 x_{1\text{-}3} + u$$

Where u is the random error term representing the random influencing factors outside the model. Since the evaluation content is very dissatisfied, general, satisfied, and very satisfied, the categories of dependent variables are ordered, so we give priority to the ordered multi-classification model.

Table 5. Model Fitting information

Model	-2 Log likelihood	Chi-Square	df	Sig
Only intercept	89.167			
Final	74.445	14.722	4	.005

As can be seen from the likelihood ratio test in table 5, compared with the invalid model with only constant terms, the logarithm likelihood ratio of -2 decreased from 89.167 to 74.445, and the

chi-square value of the likelihood ratio was 14.722. P<0.005 passed the chi-square test, indicating the significance of adding independent variable x.

Table 6 Ordered logistic regression

degree	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
age	0.376	0.122	3.09	0.002	0.138	0.614	***
cut1	-2.271	0.451	.b	.b	-3.156	-1.386	
cut2	-1.440	0.386	.b	.b	-2.196	-0.684	
cut3	1.371	0.358	.b	.b	0.669	2.072	
cut4	2.871	0.399	.b	.b	2.088	3.654	
Mean dependent var		3.441	SD dependent var		0.919		
Pseudo r-squared		0.014	Number of obs		272.000		
Chi-square		9.560	Prob > chi2		0.002		
Akaike crit. (AIC)		673.473	Bayesian crit. (BIC)		691.502		

*** *p<0.01, ** p<0.05, * p<0.1*

Table 6 shows the results of model fitting. Independent variables have passed the significance test, and the fitting effect is preliminaries considered to be good.

➢ Model test

Hypothesis 1: the dependent variable is unique and is an ordered multi-categorical variable. Hypothesis 2: there is one or more independent variables, which can be continuous, ordered multi-classification, or disordered classification variables. Hypothesis 3: there is no multicollinearity between independent variables. (variance expansion factor VIF is less than 10). Hypothesis 4: the model satisfies the "proportional advantage" hypothesis (P > and 0.05 of parallel line test, indicating that the parallelism hypothesis is valid). The independent variable is the age of the public, which is a disordered classification variable and satisfies hypothesis 2. Since there is only one independent variable age, multicollinearity does not exist, satisfying hypothesis 3. Hypothesis 4 requires us to do a parallelism test to prove that it is true. Parallelism test: H_0 =Non-proportional odds model can better explain the relationship between different levels of outcome variables.

Table 7 Parallelism test

Likelihood-ratio test	LR chi2(3) =	22.86
(Assumption: A nested in B)	Prob > chi2 =	0.0000

The calculation results are shown in Table 7, and the P-value is 0.0000, which rejects the original hypothesis, indicating that the data do not satisfy the parallelism hypothesis. We adopted the disordered multi-classification logistic model.

➢ Model improvement

Table 8 Model fitting information

Model	-2 Log likelihood	Chi-Square	df	Sig
Only intercept	86.008			
Final	50.973	35.034	12	.000

It can be seen from the likelihood ratio test in table 8 that compared with the invalid model with only constant terms, the logarithm likelihood ratio of -2 decreased from 86.008 to 50.973, and the chi-square value of the likelihood ratio was 35.034. P<0.000 passed the chi-square test, indicating the significance of adding independent variable table 9 is the result of parameter estimation, "not satisfied" section for the first chaotic classification logit regression model parameter estimation, "general" for the second disorder more classification logit model parameter estimation, the model will be satisfaction = 1 = 4 (i.e., very dissatisfied), age (over 40) as the variable reference level, so all the parameters in model estimation, the default is 0.

Table 9 Model fitting results

Satisfaction		B	S.E.	Wald	Sig	Exp(B)	95%C.I. for Exp(B) Lower	Upper
Dissatisfaction	intercept	19.334	1.155	280.365	.000	-	-	-
	10-20years old	20.251	1.426	201.685	.000	623418559.900	38107364.310	10198834470.000
	20-30years old	20.028	1.354	218.784	.000	498734847.900	35102446.100	7086014684.000
	30-40years old	18.236	.000	-	-	83122474.650	83122474.650	83122474.650
	over40years old	0[b]	-	-	-	-	-	-
Normal	intercept	1.099	.816	1.810	.178	-	-	-
	10-20years old	2.471	1.087	5.171	.023	11.833	1.407	99.550
	20-30years old	1.653	1.011	2.675	.102	5.222	.721	37.850
	30-40years old	.201	1.044	.037	.848	1.222	.158	9.467
	over40years old	0[b]	-	-	-	-	-	-
Satisfaction	intercept	1.099	.816	1.810	.178	-	-	-
	10-20years old	1.846	1.092	2.856	.091	6.333	.745	53.870
	20-30years old	.847	1.024	.685	.408	2.333	.314	17.346
	30-40years old	-.118	1.061	.012	.912	.889	.111	7.107
	over40years old	0[b]	-	-	-	-	-	-
Very satisfaction	intercept	1.792	.764	5.504	.019	-	-	-
	10-20years old	-.405	1.099	.136	.712	.667	.077	5.749
	20-30years old	-.251	.994	.064	.800	.778	.111	5.457
	30-40years old	-1.281	1.057	1.469	.225	.278	.035	2.204
	over40years old	0[b]	-	-	-	-	-	-

According to table 9, we write the specific formula of the model.

$$\ln\left[\frac{P(Y\le 2)}{P(1\text{-}Y\le 2)}\right]=19.334+20.251x_{1\text{-}1}+20.028x_{1\text{-}2}+18.236x_{1\text{-}3}$$

Under the condition of invariable in controlling other variables, aged from 10 to 20 to 20 to 30 years old, users satisfied with the live content never changes to occur very dissatisfied with the logarithmic ratio increased by 20.251, age from 20 to 30 years old to 30 to 40 years old, users satisfied with the live content never changes to occur very dissatisfied with the logarithmic ratio increased by 20.028, age from 30 to 40 years old to 40 years old above, users satisfied with the content never changes to occur very dissatisfied with the logarithmic ratio increased by 18.236.

$$\ln\left[\frac{P(Y\le 3)}{P(1\text{-}Y\le 3)}\right]=1.099+2.471x_{1\text{-}1}+1.653x_{1\text{-}2}+0.201x_{1\text{-}3}$$

Under the condition of invariable in controlling other variables, aged from 10 to 20 to 20 to 30 years old, user to broadcast content from general changes to occur very dissatisfied with the logarithmic ratio increased by 2.471, age from 20 to 30 years old to 30 to 40 years old, user to

broadcast content from general changes to occur very dissatisfied with the logarithmic ratio increased by 1.653, age from 30 to 40 years old to 40 years old or above, the user to broadcast content from general changes to occur very dissatisfied with the logarithmic ratio increased by 0.201.

$$\ln\left[\frac{P(Y \le 4)}{P(1\text{-}Y \le 4)}\right] = 1.099 + 1.846x_{1\text{-}1} + 0.847x_{1\text{-}2} - 0.118x_{1\text{-}3}$$

Since all the parameters at the satisfactory level are not significant, the formula is not explained.

$$\ln\left[\frac{P(Y \le 5)}{P(1\text{-}Y \le 5)}\right] = 1.792 - 0.405x_{1\text{-}1} - 0.251x_{1\text{-}2} - 1.281x_{1\text{-}3}$$

Under the condition that other variables remain unchanged when the age changes from 10-20 years old to 20-30 years old, the logarithmic occurrence ratio of users' satisfaction changes from very satisfied to very dissatisfied increases by 0.405, and other results are the same. As can be seen from the results of the formula, there is a negative relationship between age and the princess's satisfaction with the webcast. As people get older, their ability to accept new things becomes slower, such as the product of modern technology. (webcast)

4. Conclusions

From the perspective of gender and age of webcast users, this report studies the current situation of the webcast and draws the following conclusions:

1) Different genders and ages have different preferences for webcast content. Male students tend to compete in online games, while female students pay more attention to beauty, clothing, and cooking. 10 to 30 years old users online gaming competitions, users are over the age of 30 prefer to watch the education of science and technology, this is due to the different age groups, the tasks facing the distinction that having essence, 10 to 30 years old crowd had mostly children's burden, more the pursuit of self-feeling, while more than 30 users need to consider their children's education, focus is different, lead to different tastes.

2) Different age to watch the webcast time, 10 to 20 years old at 14:00 to 18:00 to watch the webcast of the proportion of about 10% more than the other ages, 20 to 30 years old are prone to 18:00 to 22:00 watch live, because after 18:00 usually go off work time, people will have more free time to watch the webcast, and users over the age of 40 like at 6 PM - 10:00 to watch webcast of the reason for this is that people over the age of 40 pay more attention to health, early to bed and early to rise is emphasized. So there is a certain relationship between viewing time and age.

3) Most users watch the webcast for pure entertainment. This is because nowadays, with the fast pace of life and the great pressure of study and work, webcast provides users with leisure and entertainment functions, so that many users use webcast for entertainment to relieve pressure.

4) Users' comments on the webcast are mostly average, but the type of live broadcast content can meet most people's needs. There is no significant relationship between gender and satisfaction, but age can affect satisfaction.

References

[1] Hosmer D W, Hosmer T , Cessie S L , et al. A comparison of goodness-of-fit tests for the logistic regression model[J]. Statistics in Medicine, 1997, 16(9):965-980.

[2] Yi Liu, Jiawen Peng, and Zhihao Yu. 2018. Big Data Platform Architecture under The Background of Financial Technology: In The Insurance Industry As An Example. In Proceedings of the 2018 International Conference on Big Data Engineering and Technology (BDET 2018). ACM, New York, NY, USA, 31-35.

[3] Berkson, Joseph. Some Difficulties of Interpretation Encountered in the Application of the Chi-Square Test[J]. Publications of the American Statistical Association, 1938, 33(203):526-536.

[4] Sharpe D. Your Chi-Square Test Is Statistically Significant: Now What?.[J]. Practical Assessment Research & Evaluation, 2015, 20(8):10.

[5] Kastrin A , Peterlin B , Hristovski D . Chi-square-based Scoring Function for Categorization of MEDLINE Citations[J]. Methods of Information in Medicine, 2010, 49(4):371-378.