MALAYSIAN FILM AUDIENCES INFLUENTIAL FACTORS AND FORECAST USING MULTINOMIAL LOGISTIC REGRESSION MODELS

Noraini Abdullah
Diana Hassan
Suhaimi Salleh
Zainodin Haji Jubok

ABSTRACT

Watching films has been one of the most popular leisure activities. These activities might lead to provide information and pseudo experiences. As required by the Malaysian film industry, an analysis was performed to identify the significant factors that might have influenced audiences' perceptions and film watching activities, thus increasing the frequency of film watching. Multinomial Logistic Regression (MLR) without interactions with independent variables were employed in this study. MLR model is a more appropriate model with regard to regression models since the categorical dependent variable is nominal with more than two levels. The data were obtained from distribution of questionnaires including Sabah, Sarawak, Johor, Pahang, Kedah and Selangor. The study was conducted to examine the significant factors based on the frequency of watching categories, such as 1) once or nil in a month, 2) twice a month, 3) three to four times in a month and 4) more than five times per month. From the frequency of watching categories (1, 2, 3 & 4), the frequency of watching more than five times per month (4) was referred to as the reference group, while the other categories had exhibited 255 models each. Statistical tests, modelling procedures and models' goodness-of-fit tests were carried out on a total of 765 models in this study. Forecasting based on MAPE was done using the best models from each category. Findings showed that the best models from all the respective categories (1, 2 & 3) had two common significant factors on the dependent variable. The results also showed that the best model from Cat 1 had the least MAPE (6.57%) thus indicated it was excellent to be used for forecasting and esimation of missing values. Based on this, it is suggested that to attract more audiences, less films should be produced in a year, however, the allocated budget for film making should be focussed on producing films which conformed to the identified significant factors that would attract more audiences. By using the best model, the number of audiences can thus be forecasted, and the expected revenue for the film industry can thus be predicted.

Keywords: Film watching, audiences, multinomial logistic regression, categories, significant factors.

INTRODUCTION

Film has a unique cultural value because it is a universal medium, remarkably accessible and inclusive with its appeal traversing eras and intersection national and phonetic boundaries. Film is able to confront people with the real world, whilst also speaking to their imaginations. Film can be informative and reveal essential truths about the human condition. According to Kusumarasdyati (2004) and Luo (2004), films can catch the learners' interest and it can positively affect student motivation to learn. Film also has immediacy and when viewed at the cinema provides an immersive experience and accessible. It has the ability to influence viewer's attitude and perception especially to an avid viewer. They often use the word "obsession" to portray their association with film. Cinema offers more than stories told in light and sound, seen once and soon forgotten for avid audiences. They are exceptionally visited by cinemagoers that frequently go to film celebrations and seasons. Avid audiences are attracted to free silver screen,s and film is integral in their social life. Compared to non avid audiences, they seldom come to the cinema and usually based on their interest to a certain movie or influenced by box-office movies. It is common for avid audiences to cite a particular film as the formative influence on their development.

Since cinematography, watching movies has been one of the most popular leisure activities of all times. These activities might lead to provide information and pseudo experiences, particularly in the absence of an individual's experiences. An analysis of the perception of film audiences is required on the Malaysian film industry; especially the film industry has a business and economic impact on Malaysia. This study thus aspires to examine the factors that might have influenced viewers' perceptions towards increasing the frequency of film watching. Factors relevant would include Encouragement, Source of information, Gratification, Film Genre, Perception on Malaysian Films, Medium-watching, Film Production and Attractions-watching. In addition, this study would also highlight the procedures in obtaining these significant factors that might have influenced film audiences' perceptions, and hence, increased in the frequency of film viewing activities.

LITERATURE REVIEW

According to Miller (1999), the ability of film to influence perception because it provides information and pseudo-experiences, particularly in the absence of an individual's own experience. Dyna (2012) stated that many researches had been done, but none of the researchers are able to develop parsimonious consumer decision making model that explain cinema decision making process involving many factors. Mohammadian & Habibi (2012) had discovered that only four influential factors in attracting Iranians to go to the cinemas; they were product, price, places and promotional factor. While, Mustafa (2009) had studied that seven factors had helped the Egyptian audiences in determining their choice of films such as movie stars, directors, trailers, general advertising, word of mouth, movie genre and reviews.

Hosmer & Lameshow (2000) had developed the model-building strategies for logistic regression, meanwhile, Kutner *et al.* (2008) had stated that logistic regression could be considered as a nonlinear regression. The logistic regression model consists of the non-linear response function and a random error term, which the random error term is assumed to be independent normal distributed with constant variance, written as $\varepsilon \sim N(0,\sigma^2)$. Logistic regression is used to predict the binary response when the dependent variable is dichotomous. It also applied to predict the outcome of categorical dependent variable based on the explanatory variables.

According to Habshah *et al.* (2011), multinomial logit regression model (MLR) is more appropriate than binary logit model when the response variable is normal with more than two levels. The MLR analysis is usually used to identify the significant factor that contribute to a particular or interest event. In the medical field, Kutner *et al.* (2005) conducted an MLR model to determine the risk factors of diseases that they were interested in, while Wang (2005) had conducted a study for anomaly intrusion detection using the MLR technique.

Zainodin & Khuneswari (2010) have conducted a study which developed the model building procedure to identify the best Multiple Binary Logit (MBL) model for coronary heart diseases. First step was to list out All Possible Models in Phase 1; then in Phase 2, the selected model were obtained after running the Multicollinearity and Coefficent test. Next, the best MBL model was identified in Phase 3 and in the final phase, the goodness-of-fit test was then carried out using the Deviance test so as to examine the appropriateness of the best model.

METHODOLOGY

Factor Analysis

Factor analysis usually carried out by using the statistical tool in SPSS and Microsoft Excel to investigate the relationship between the variables. According to Bartholomew (1980), factor analysis is an effective tool in reducing the dimensionality of the multivariate data. In factor analysis, there are three procedures have to go through, such as interpretation of the correlation matrix, extraction of the intial factor and rotation of the solution. Factors could be determined by using factor analysis based on the assumption that correlations were derived from scores that produced linear relationship (Child, 2006). The results of factor analysis suggested that the variable had no significant effect if the communality of a variable was too low, thus, eliminating the variable from the analysis.

Multinomial Logistic Regression (MLR)

Multinomial Logistic Regression (MLR) is used in this study to analyze the significant factors on movie watching activities. Logistic regression can form a best fitting equation or function using the maximum likelihood method which maximizes the probability of classifying the observed data into the appropriate category, given the regression coefficients, especially when there is a mixture of numerical and categorical independent variable(s), thus modelling requires fewer assumptions and is more statistically robust. In other words, logistic regression is necessary when the independent variables are categorical or a mix of continuous with categorical and the dependent variable is categorical. Thus, MLR model is a more appropriate model with regard to other regression models since the categorical dependent variable is nominal with more than two levels. The category frequency of viewing used was: Cat 1) once and never in a month, Cat 2) twice a month, Cat 3) three to four times in a month, and Cat 4) more than five times per month. From the frequency of viewing, Cat 4) category of more than five times per month was referred to as the reference/control group.

According to Zainodin & Khuneswari (2010), the general multinomial logistic regression model is given by: $Y_i = \Omega_0 + \Omega_1 W_1 + \Omega_2 W_2 + ... + \Omega_k W_k + u_j$ …(1) where 'Y_i' is the categorical dependent variable, 'W_j' denotes the j-th variable, 'Ω_0' is the constant term of the model, 'Ω_j' is j-th coefficient of independent variable Wj, for j=1,2, …, k, 'k' is the number of the single independent variables, (k+1) is the number of parameters, and 'u_j' is the error term, for j=1,2,…,k. This study was conducted to examine the significant factors on models without interactions based on the category frequency of viewing, Y_i, with i=1, 2, 3 & 4.

Data Preparation

Questionnaires were distributed to 1337 respondents in the states of Sabah, Sarawak, Johor, Selangor, Kedah and Pahang. The raw data collected were in the field of social sciences with regard to film audiences and the frequency of film watching. Data preparation was done before any statistical analyses that involved process of cleaning and organizing the data, factor analysis, and dummy transformation (Noraini *et al*., 2017). Data transformation were carried out resulting in 1277 samples, and further partitioned at 85% for modelling (n=1085), 10% for prediction (n=128), and 5% for estimating the missing values (n=64) using the best model. However, in this paper the multinomial logistic modelling procedures with Mean Average Percentage Error (MAPE) used for forecasting, together with the estimation of missing values were illustrated. Detailed illustrations on the Coefficient test were depicted too.

Model-Building Procedures

Summary of the Four-Phase Model-Building Procedures carried out in this study is shown in Figure 1 below. Further explaination and illustration on the modelling procedures can be referred in: (Noraini *et al*., 2016; Zainodin *et al*., 2011; Diana *et* *al*. 2017), while the multicollinearity test based on the VIF approach (Zainodin *et al*., 2015). All the four phases of the model-building procedures were carried out on all the models. The Mean Average Percentage Error (MAPE) on the best model would be calculated to validify its forecasting performance.

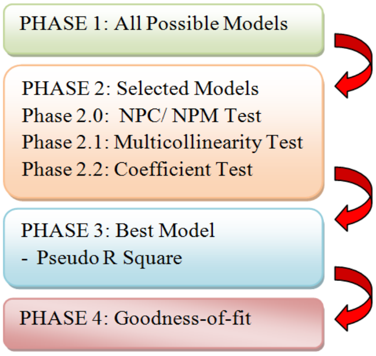

Figure 1: Four-Phase Model-building Procedures in Multinomial Logistic Regression (Noraini *et al*., 2017)

RESULTS AND DISCUSSIONS

The scree plot shown in Figure 2 depicted the graf of eigenvalues versus the components. The results from the scree plot suggested that the first two components with eigenvalues greater than 1.0 were chosen. Table 1 displayed the rotated component matrix. Eight categorical independent variables from Component 1 and 2 respectively that had the highest absolute value correlation greater than 0.50 were chosen for next step in the model building procedures. From Phase 1 in Figure 1, all possible models obtained were 255 from each category (1, 2 and 3), giving a total of 765 models.

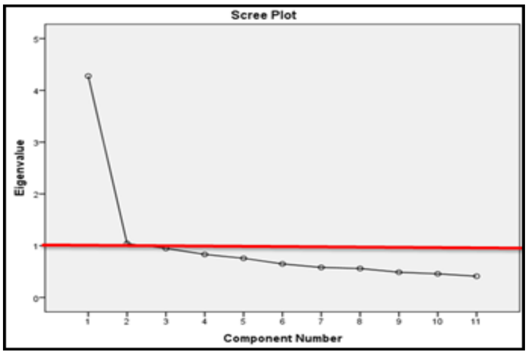

Figure 2:Scree Plot

Table 1: Rotated Component Matrix

	Component	
	1	2
Source of Information	.815	
Encouragement	.739	
Medium of Watching	.560	
Gratification	.442	.466
Film Production	.399	.301
Perception on Malay Film		.774
Public Perception		.635
Attraction	.374	.608
Theme	.395	.581
Genre	.490	.562
Showtime	.325	.478

According to Zainodin *et al.* (2011), a model with free from multicollinearity source variable and free from insignificant variable can be written as $M_{a.b.c}$, where 'Ma' denotes the parent model of the MLR model, 'b' denotes the number of variable/s removed due to multicollinearity and 'c' denotes the number of variable/s eliminated due to variable insignificance. For illustration purposes, model M13 was taken as the parent model from Category 1. Phase 2.0 of the modelling procedures involved the removal of highly correlated variables which had R^2 more than 0.95, the near perfect colinearity (NPC) and near-perfect multicollinearity (NPM) from the models. Next, Phase 2.1 of the multicollinearity test would sequentially remove multicollinearity source variable/s with VIF values more than 5.0 starting with the highest, until there would be no more variables of VIF values more than 5.0 remained in the model as shown in Table 2 below.

Table 2: Phase 2.0-2.1 on Model M13.0.0

PHASE 2.0

IV	R^2
D1	0.22122
D2	0.00921
D3	0.21014
D4	0.22402
D5	0.23529

PHASE 2.1

Model	Coefficients	
	Collinearity Statistics	
	Tolerance	VIF
D1	.779	1.284
D2	.991	1.009
D3	.790	1.266
D4	.776	1.289
D5	.765	1.308

Phase 2.2 of the Coefficient test would eliminate sequentially the insignificant variable/s with the highest p-value of the selected model. Table 3 showed the removal of insignificant variable, G3 via the coefficient test of Phase 2.2 on model 13.0.0. After removal, the model then became M13.0.1.

Table 3: Phase 2.2 Coefficient Test on Model M13.0.0

M13.0.0	*B*	*Std. Error*	*Wald*	*df*	P-Value	Exp(B)	Decision H_0	Action
Intercept	1.2846	0.1887	46.3387	1	0.0000E+00			
D1	1.3779	0.2188	39.6738	1	0.0000E+00	3.9666	Reject	
D2	0.0537	0.1583	0.1149	1	7.3468E-01	1.0551	Accept	
D3	0.1067	0.1779	0.3599	1	5.4857E-01	1.1126	Accept	
D4	0.1037	0.1726	0.3612	1	5.4786E-01	1.1093	Accept	
D5	0.3009	0.2042	2.172	1	1.4054E-01	1.3511	Accept	
G1	-0.5012	0.1772	8.0028	1	4.6706E-03	0.6058	Reject	
G2	-0.5967	0.2012	8.7936	1	3.0229E-03	0.5506	Reject	
G3	0.0058	0.1726	0.0011	1	9.7298E-01	1.0059	Accept	Removed
G4	-0.1357	0.1652	0.6753	1	4.1121E-01	0.8731	Accept	
G5	-0.1835	0.1677	1.1974	1	2.7384E-01	0.8324	Accept	
G6	-0.0359	0.169	0.0451	1	8.3184E-01	0.9647	Accept	
G7	-0.4079	0.1903	4.5949	1	3.2067E-02	0.665	Reject	
G8	0.1694	0.1911	0.7853	1	3.7552E-01	1.1846	Accept	

Further removal of insignificant variables, G6 and D2 on model M13.0.1 would result in model M13.0.2 being obtained as shown in Table 4.

Table 4: Phase 2.2 Coefficient Test on Model M13.0.1 => M13.0.2

M13.0.1	*B*	*Std. Error*	*Wald*	*df*	P-Value	Exp(B)	Decision H_0	Action
Intercept	1.2854	0.1871	47.2006	1	0.0000E+00			
D1	1.3782	0.2186	39.7501	1	0.0000E+00	3.9678	Reject	
D2	0.0539	0.1582	0.1159	1	7.3354E-01	1.0553	Accept	
D3	0.1071	0.1775	0.3644	1	5.4605E-01	1.1131	Accept	
D4	0.1037	0.1726	0.3609	1	5.4802E-01	1.1092	Accept	
D5	0.3016	0.2032	2.2037	1	1.3768E-01	1.3521	Accept	
G1	-0.5019	0.1762	8.1138	1	4.3931E-03	0.6054	Reject	
G2	-0.5947	0.1927	9.5279	1	2.0237E-03	0.5517	Reject	
G4	-0.1346	0.1621	0.6899	1	4.0619E-01	0.874	Accept	
G5	-0.1833	0.1675	1.1964	1	2.7404E-01	0.8325	Accept	
G6	-0.0359	0.169	0.0452	1	8.3157E-01	0.9647	Accept	Removed
G7	-0.4076	0.1901	4.5972	1	3.2024E-02	0.6652	Reject	
G8	0.1691	0.191	0.7843	1	3.7584E-01	1.1842	Accept	

M13.0.2	*B*	*Std. Error*	*Wald*	*df*	P-Value	Exp(B)	Decision H_0	Action
Intercept	1.2793	0.1848	47.9439	1	0.0000E+00			
D1	1.3759	0.2183	39.73	1	0.0000E+00	3.9585	Reject	
D2	0.0523	0.1581	0.1097	1	7.4051E-01	1.0537	Accept	Removed
D3	0.1072	0.1775	0.3646	1	5.4596E-01	1.1131	Accept	
D4	0.1067	0.172	0.3844	1	5.3525E-01	1.1126	Accept	
D5	0.297	0.202	2.161	1	1.4156E-01	1.3458	Accept	
G1	-0.5062	0.175	8.3701	1	3.8145E-03	0.6028	Reject	
G2	-0.5979	0.1921	9.6866	1	1.8561E-03	0.55	Reject	
G4	-0.1367	0.1618	0.7137	1	3.9822E-01	0.8722	Accept	
G5	-0.1933	0.1608	1.4442	1	2.2946E-01	0.8243	Accept	
G7	-0.4109	0.1896	4.6966	1	3.0223E-02	0.6631	Reject	
G8	0.1712	0.1908	0.8054	1	3.6948E-01	1.1867	Accept	

Subsequent removals of insignificant variables using the Coefficient test were performed until the p-values of all variables that remained in the model were less than 0.05. The resulting model M13.0.10 was obtained after the removal of 10 insignificant variables. Table 5 showed the final table of the Coefficient test of Phases 2.2 of the MLR modelling procedures with all the p-values being less than 0.05.

Table 5: Final Table of Phase 2.2 Coefficient Test on Model M13.0.10

M13.0.10	*B*	*Std. Error*	*Wald*	*df*	P-Value	Exp(B)	Decision H_0
Intercept	1.1839	0.0994	141.8674	1	0.0000E+00		
D1	1.4863	0.2036	53.2839	1	0.0000E+00	4.4207	Reject
G1	-0.5064	0.1694	8.9401	1	2.7898E-03	0.6026	Reject
G2	-0.6563	0.185	12.5888	1	3.8810E-04	0.5188	Reject

Category 1 had obtained 42 selected models before proceeding to Phase 3 of choosing the best model of the model-building procedures based on the pseudo R-square criteria as shown in Table 6.

Table 6: The corresponding selection criteria based on Pseudo R-square Category 1

NP	Model	Selected Model	(k+1)	Cox and Snell R^2	Nagelkerke R^2	McFadden R^2
6	M1	M1.0.4	2	0.0450	0.0689	0.0435
6	M2	M2.0.4	2	0.0078	0.0119	0.0074
↓	↓	↓	↓	↓	↓	↓
12	M15	M15.0.9	3	0.0562	0.0861	0.0546
↓	↓	↓	↓	↓	↓	↓
19	M55	M55.0.19	6	0.0974	0.1491	0.0967
↓	↓	↓	↓	↓	↓	↓
17	M75	M75.0.9	8	0.0777	0.1189	0.0763
↓	↓	↓	↓	↓	↓	0.0951
22	M100	M100.0.15	7	0.1066	0.1632	0.1064
↓	↓	↓	↓	↓	↓	↓
23	M134	M134.0.14	9	0.0616	0.0942	0.0600
↓	↓	↓	↓	↓	↓	↓
29	M175	M175.0.21	8	0.0893	0.1368	0.0884
↓	↓	↓	↓	↓	↓	↓
30	M189	M189.0.19	11	0.1264	0.1935	0.1276
↓	↓	↓	↓	↓	↓	↓
29	M200	M200.0.19	10	0.0743	0.1137	0.0729
↓	↓	↓	↓	↓	↓	↓
35	M222	M222.0.24	11	0.1265	0.1936	0.1277
↓	↓	↓	↓	↓	↓	↓
36	M240	M240.0.23	13	0.1105	0.1692	0.1106
41	M247	M247.0.28	13	0.1364	0.2088	0.1385
42	M253	M253.0.30	12	0.1172	0.1793	0.1176
			Maximum	0.1364	0.2088	0.1385

The best model was chosen based on the model that having the majority maximum value of pseudo Rsquare, namely, the criteria based on Cox & Snell, Nagelkerke and McFadden. Results in Table 6 showed that model M247.0.28 was chosen as best model from category 1(Cat1) where it had the maximum values of the pseudo R-square criteria. Similar procedures were carried out for Category 2 (Cat 2) and category 3 (Cat 3) respectively. The factors of the best models from Cat 2 and Cat 3 respectively were shown in Table 7.

Table 7: Best Models From Category 1, 2 and 3

Category	Factors of Best Model
1	M247.0.28 : Y_1= f(D_1, S_4, M_1, M_2, M_3, K_5, P_2, P_4, P_6, G_1, T_1, T_5)
2	M250.0.26 : Y_2= f(D_1, D_3, D_5, S_3, S_4, S_5, M_1, M_2, M_3, K_5, G_6, G_8, T_5, F_3)
3	M14.0.6 : Y_3= f(D_1, T_1, T_3, T_5)

The common significant factors of the MLR best models (Table 7- highlighted in green) were from Encouragement-Interest (D_1), and Attractions of Watching-Friends Influence (T_5). Goodness-of-fit (GOF) tests, namely the Pearson and Deviance Tests were carried out to examine the goodness or appropriateness of the best model M247.0.28 from category 1, in fitting the data. The Deviance test with statistics (670.1612) at significant p-value of 0.1837 ($p>0.05$) had shown that model M247.0.28 is an appropriate model. The likelihood ratio test carried out on all the factors in the best model also showed that they were significant with p-values less than 0.05.

MEAN ABSOLUTE PERCENTAGE ERROR (MAPE)

According to Ren & Glasure (2009), MAPE is probably the most widely used forecasting accuracy measurement. MAPE has important and desirable features including reliability, unit-free measure, ease of interpretation, clarity of presentation, support of statistical evaluation and the use of all the information concerning the error (Juan *et al.,* 2013). MAPE obeys the following mathematical expression given by:

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|\hat{y}_t - y_t|}{y_t} \text{x}100\%$$

where, n is the size of sample, $\hat{y}_t$ is the value predicted by the model at time of point *t*, and y_t is the value observed at time of point *t*.

According to Tayman & Swanson (1999), MAPE satisfactorily would meet at least four of the aforementioned criteria, but would be less satisfactory in meeting the validity criterion when used in evaluating the accuracy of population forecast. Beside that, Lewis (1982) had drawn up a table as shown in Table 8 that contained typical MAPE values for industrial and business data and their interpretation.

Table 8: Interpretation of typical MAPE values

MAPE	Interpretation
<10	Highly accurate forecasting
10 - 20	Good forecasting
20 – 50	Reasonable forecasting
>50	Inaccurate forecasting

Thus, the MAPE of each best model from the 3 categories were then calculated and compared. It can be seen below that model M247.0.28 from category 1 (Cat 1) has the least MAPE (6.57%) , hence this model would be the best to be used for forecasting of the film audiences with the frequency of going to watch films minimally once a month. However, for category 2 (Cat 2) with the frequency of going twice to watch films in a month, the best model M250.0.26 was also highly accurate to be used for forecasting film audiences.

MAPE (%): Category 1= $\frac{1}{128}\left(\sum\left|\frac{A_t-F_t}{A_t}\right|\right) = \frac{8.4109}{128} = 0.0657 = 6.57\%$

MAPE (%) : Category 2= $\frac{1}{128}\left(\sum\left|\frac{A_t-F_t}{A_t}\right|\right) = \frac{11.0285}{128} = 0.0862 = 8.62\%$

MAPE (%) : Category 3= $\frac{1}{128}\left(\sum\left|\frac{A_t-F_t}{A_t}\right|\right) = \frac{30.4706}{128} = 0.2381 = 23.81\%$

Table 9: Summary of MAPE for each categories

Model	Category	MAPE (%)	Remark
M247.0.28	Cat 1	6.57%	Highly accurate forecasting
M250.0.26	Cat 2	8.62%	Highly accurate forecasting
M14.0.6	Cat 3	23.81%	Reasonable forecasting

These findings were similar to Noraini *et al*. (2017), however, the detailed illustration on the coefficient test and the missing values estimation were not shown. In this paper, the best model was further tested for its robustness in the estimation of variables that were identified as missing values earlier during partitioning.

MISSING VALUE ANALYSIS

The main purpose of the missing value analysis is to check the accuracy of the best model by using it to estimate the missing values, and then compute the absolute error. The actual value and the estimated value of the missing value are then compared. The model gets more accurate when the difference between the estimated value and the actual value gets smaller. In this paper, the accuracy of the best model M247.0.28 for Cat 1 was investigated and further illustrated as follows.

The outcome of the missing value analysis can be simulated by first simply selecting the random number. By using the random selection in Microsoft Excel, the data partitioning were randomly selected. Firstly, a series of random number was generated by using the excel function RAND () function and then sort them in ascending order. After generating the random numbers, 64 samples were selected for partition 3, given that repeating numbers were ignored. Since there were a total number of 1277 cases were analyzed, the series of random number were then multiplied by 1277 and transformed them into an integer number by using INT( ) function in Microsoft Excel. Table 10 below displayed some number of samples that were used to generate random number for this partition analysis. The purpose of generating the random number series was to make sure the sample was selected randomly.

Table 10: Generating Random Number for Missing Value Analysis

No	RAND ()	(1277*RAND())+1	Integer Number
	Random Number	Random Number	
1	0.6924	680.0354	680
2	0.4136	1157.0551	1157
3	0.4221	79.7548	80
4	0.3456	236.3279	236
5	0.2740	843.0095	843
6	0.5681	400.6450	401
7	0.2861	69.0399	69
8	0.0537	165.4271	165
9	0.5110	389.0017	389
10	0.1457	1119.6400	1120
11	0.7028	797.6243	798
12	0.1234	639.8650	640
13	0.0926	1089.4959	1089
14	0.0333	522.3292	522
15	0.3593	466.3013	466
16	0.1714	1244.5786	1245

$\hat{Y} = -1.0267 - 1.3248C^2 + 0.4594D^2 + 0.5578E^2 + 0.4368F^2 - 0.4577G^2 - 0.4148H^2 + 0.5043I^2 + 0.3594J_2 - 0.3704K_2 - 0.3697L_2 - 1.081M_2$Based on the best model equation for Cat 1:

Error! Digit expected. Then, substituting the actual value for each variable into the best model equation to get the similarity value as shown in Table 11. Using this value, the estimated value of the missing value could thus be obtained.

Table 11: Computation the Estimated Value

Missing	Similarity	EXP()	Estimate
T5	2.7876	16.2420	0.9420

Next, comparison between the actual value and estimated value was done to get the standard error as shown in Table 12. The absolute error for preditors T5 are less than 10%, which indicated that the best model CAT1, model M247.0.28 is highly accurate. The absolute standard error of the estimated variable T_5 (%) was given by | 1 - 0.9420| = 0.0580 = 5.80%, and the variable D_1 (%) was | 0 - 0.0779| = 0.0779 = 7.79 % respectively. Since the missing values estimation were less than 10%, it could be seen said that the model was appropriately robust and highly accurate in its use to forecast film audiences.

Table 12: Computation the Absolute Standard Error

Estimated	Actual Value	Estimated Value	Differences Actual Value and Estimated Value	Standard Error (%)
T5	1	0.9420	0.0580	5.80%

CONCLUSION

This study had introduced the concepts and procedures in mathematical modelling using the multinomial logistic regression technique in the field of social sciences, specifically business. In this paper, Multinomial Logistic Regression (MLR) models had identified common influential factors that would affect the sustainability of the Malaysian film industry as Encouragement due to interest and Attraction of Watching from friends influence. These factors gave positive and direct contribution to the increase in the frequency of film watching. In other words, these factors implied that more encouragement due to interest and influence from friends would lead to the higher frequency of film goers going to the cinemas, and indirectly, to the number of audiences. Thus, it can be suggested to film publishers, producers and policy makers not to neglect these factors since these will help to ensure the survival of the Malaysian film industry and viewability appeals besides towards producing high quality films for box offices.

ACKNOWLEDGEMENT

The authors would like to thank the FINAS and Universiti Malaysia Sabah for had partially funded this research under the grant number SGK0249-SS-2015.

REFERENCES

Al-Rodhan, Nayef, R.F., & Stoudmann, G. (2006). Definitions of Globalization: A Comprehensive Overview and A Proposed Definition. *Geneva Centre for Security Policy*.

Ballou, D.P. & Tayi, G.K. (1999). Enhancing Data Quality in Data Warehouse Environments. Communications of ACM, 42: 73-78.

Child, D. (2006). *The Essentials of Factor Analysis*. 3rd Edition. New York: Continuum International Publishing Group

Dasu, T. & Johnson, T. (2003). Exploratory Data Mining and Data Cleaning. John Wiley & Sons.

Devore, J.L. (2012). *Probability and Statistic for Engineering and the Science*. 8th edition. Canada: Brooks/Cole, Cencage Learning.

Gujarati, D.N. & Porter, D.C. (2009). *Basic Econometrics*. 5th Edition. Singapore: McGraw-Hill Inc.

Habshah, M., Sarkar, S.K., and Rana, S. (2011). Adequacy of Multinomial Logit Model with Nominal Responses over Binary Logit Model. *Trends in Applied Science Research,* **6**(8):900-909

Hair, J.F., Anderson, R.E., Tatham, R.L. & Black, W.C. (1995). *Multivariate data analysis with reading*. 5th edition. New Jersey: Prentice Hall inc.

Haralambos, M. & Holborn, M. (2004). *Sociology Themes and Perspectives*. (6th Edition). London: Haper Collins Publishers.

Hosmer, D.W. & Lameshow, S. (2000). *Applied Logistic Regression*. 2nd Edition. Canada: John Wiley & Sons, Inc.

Huck, S.W. (2012). *Reading Statistic and Research*. 6th edition. New Jersey: Prentice Hall.

Jagdish. M. Rathod & Kosta, Y.P. (2009). Development of Conical Horn Feed for Parabolic Reflector Antenna.*International Journal of Engineering and Technology*, **1**(1): 71-74.

Juan J.M.M., Alfonso, P.P., Albert, S.A.,& Berta , C.B. (2013). Using the R-MAPE index as a resistant measure of forecast accuracy. *Psicothema,* **25** (4): 500-506.

Kutner, M.H. Nachtsheim, C. J. and Neter, J. (2008). *Applied Linear Regression Models.* 4th Edition. Singapore: McGraw-Hill Companies, Inc

Lewis, C. D. (1982). *Industrial and business forecasting methods.* London: Butterworths.

Noraini Abdullah & Zainodin H.J. (2013). Multicollinearity Remedial Techniques in Model-Building. *Matematika* (UTM), 29(1b): 105-115.

Noraini Abdullah, Suhaimi Salleh, Ramzah Dambul & Diana Hassan. (2014). Behavioural and Viewing Patterns of Cinema-Goers in Malaysia. *Proceedings the 2nd World Conference on Islamic Thought & Civilisation: Rise and Fall of Civilisation-Contemporary States of Muslim Affairs*, 18-19 August, Perak, Malaysia, pp. 876-883. ISBN: 978-967-5480-10-2.

Noraini Abdullah, Diana Hassan, Suhaimi Salleh & Zainodin H.J. (2017). Significant Factors and Prediction of Malaysian Film Viewers Using Multinomial Logistic Regression. *Proceedings International Conference on Business, Tourism & Technology (ICBTT)*, 1-2 April, Port Dickson, Negeri Sembilan, Malaysia. eISBN:978-967-14841-0-4.

Ramanathan, R. (2002). Introductory Econometrics with Application. 5th Edition. South-Western: Thomson Learning: Ohio.

Ren, L., & Glasure, Y. (2009). Applicability of the revised mean absolute percentage errors (MAPE) approach to some popular normal and non-normal independent time series. *International Advances in Economic Research, 15,* 409-420

Tayman, J., & Swanson, D.A. (1999). On the validity of MAPE as a measure of population forecast accuracy. *Population Research and Policy Review, 18,* 299-322

Zainodin H.J., & Khuneswari, G. (2010). Model-Building Approach in Multiple Binary Logit Model for Coronary Heart Disease. *Malaysian Journal of Mathematical, Science,* **4**(1):109-136

Zainodin, H.J., Noraini, A. & Yap, S.J. (2011). An Alternative Multicollinearity Approach in Solving Multiple Regression Problem. *Trends in Applied Science Research,* 6 (11): 1241 -12

Zainodin, H.J., Noraini, A. & Yap, S.J. (2014). Number of Parameters Counting in a Hierarchically Multiple Regression Model. *Science International,* 2 (2): 37-43.

Noraini Abdullah
Faculty of Science & Natural Resources
Universiti Malaysia Sabah, 84000 Kota Kinabalu, Sabah, Malaysia
Email: norainiabdullah.ums@gmail.com

Diana Hassan
Department of Postgraduate Studies
Universiti Malaysia Sabah, 84000 Kota Kinabalu, Sabah, Malaysia
Email: dianabintihassan@gmail.com

Suhaimi Salleh
Faculty of Humanities, Arts & Heritage,
Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah
E-mail: suhaimi.binsalleh@gmail.com

Zainodin Haji Jubok
Co-Founder & President of Lex Capital Sdn.Bhd.(1197516-D),Menara Maxis, KLCC
E-mail: zzainodin22@gmail.com