

Project Compiled By :

3-14-11-0414

3-14-11-0417

3-14-11-0419

3-14-11-0422

3-14-11-0432

3-14-11-0434

3-14-11-0451



Prediction of opening
weekend box office
performance

INTRODUCTION:

The term '**Box office**' is used in the context of the film industry. It refers to the amount of business a particular movie makes .The movie industry is a business with a high profile, and a highly variable revenue stream.

A single movie can be the difference between millions of dollars of profits or losses for a studio in a given year.



It's not surprising, therefore, that movie studios are intensely interested in predicting revenues from movies; the popular nature of the product results in great interest in gross revenues from the general public as well.

The statistical analysis of the box office revenue is also highly important for the creative industries and also a source of interest for the fans. As we know, producing movie is one of the riskiest businesses, the movie industry being highly dynamic. The involvement of various factors makes it more ambiguous. Here our objective is to set up a model for predicting the box office success.

OBJECTIVE:

- We know that the opening weekend box office performance of a particular movie largely determines its total revenue. Hence for simplicity, we try to predict only the opening weekend performance of a movie.
- **Now we know that different type of films has different classes of audiences. Thus taking movies without considering their genre (type) will not be meaningful.**
- Hence our objective is to find different models for predicting the box office success for different types of movies.



DATA:

Source of data:

We have collected our data from the following websites:

- **[www.boxofficemojo](http://www.boxofficemojo.com)**
- **www.imdb.com**
- **www.metacritic.com**



The variables used for our analysis are as follows

RESPONSE: Our response is opening weekend
box office collection

For the prediction purpose ,we have taken into account the following
Covariates

1.Production Budget

2.No of screens

3.Metacritic rating

4.Star Power.

5.Release Timing

6.Sequel



Normality

- For any kind of analysis, we must check whether our response variable follows a normal distribution or not.

We first perform the Shapiro-Wilk's normality test.

- At first we applied Shapiro- Wilk's test on our response y (opening weekend box office collection). The p-value obtained in this test is $= 1.136e-13 (<0.05)$

- As our response is not normal we need to transform it to a normal variable taking some transformation.

- After checking some transformation we finally come to a conclusion that fourth-root transformation whose p-value after using shapiro-Wilks' Test comes out to be $=0.6187$.

Thus we choose fourth root of y as our response.

Choice of Movies

- We selected the time period 2012-13 for our analysis and we collect our data from the movies of this time interval. By simple random sampling method we selected 117 movies from all the movies of 2012-2013.
- Before doing so we at first divided the movies on the basis of genre.
since we know that different type of films has different classes of audience and hence the analysis of all the films taking together will not be meaningful.
- So we need to classify the selected films into a number of mutually exclusive and exhaustive genres. So we divide the movies
into five major genres namely
 - (A) action
 - (B) animation
 - (C) comedy
 - (D) drama
 - (E) thriller

Interpretation of the regression coefficients corresponding to the variables used in the model

— **Y**: opening weekend box office collection

X : 1, if sequel

: 0, otherwise

The model is of the form

$$Y = \alpha + \beta x$$

Now we need to give interpretation of **b** [estimate of β] obtained from the above model.

— **case 1:** (high positive value of **b**)

If the movie is a sequel then it has positive impact on opening weekend box office collection.

— **case 2:** (high negative value of **b**)

If the movie is a sequel then it has a negative effect on box office collection, i.e., box office collection is lesser if it is a sequel.

— **case 3:** (small value (-ve or +ve) of **b**)

Sequel has no impact on box office collection.

— [Similarly assigning a value 1 to **x** if a movie is released in a festive season, 0 otherwise, we can give similar kind of interpretation of **b** in case the indicator variable is *release time*]

ANALYSIS

Let us use the numbers 1 to 6 for the six factors($p=6$), the numbering is specified below:

- \mathbf{x}_1 : Number of screens
- \mathbf{x}_2 : Production budget
- \mathbf{x}_3 : Release Timing
- \mathbf{x}_4 : Sequel
- \mathbf{x}_5 : Star power
- \mathbf{x}_6 : Meta-critic Rating
- \mathbf{y}
: $\sqrt[4]{}$ Opening weekend boxoffice collection



ACTION

After sampling we get $n=26$.

- The multiple correlation coefficient ($r_{y.123456}$) = **0.924121**
- We now test for the significance of the multiple correlation coefficient. So we test
- **$H_0 : \rho_{y.123456} = 0$ vs $H_1 : \text{not } H_0$**
- The test statistic **F**, under H_0 , follows an **F** distribution with $df = (6, 19)$, the value of **F** comes to be **18.5228**
- We reject H_0 iff **$F > F_{\alpha; 6, 19} = 2.62832$** , where $\alpha=0.05$ is the desired level of significance.
- **Here $F > F_{\alpha; 6, 19}$**
- **Hence we reject H_0 .**
- **Thus we may conclude that in the light of the given data these factors all taken together have an effect on the response at 5% level of significance.**
- Now we try to find out those factors which can be used as predictors .



(1)The correlation coefficients between y and the factors are given below:

- $r_{y1}=0.889$; $r_{y2}=0.731$; $r_{y3}=-0.264$; $r_{y4}=0.558$;
 $r_{y5}=0.337$; $r_{y6}=0.255$
- We see that r_{y1} is **maximum**. So factor1, i.e., **No. of screens** is chosen as predictor.
- Here we need to check whether the effect of x_1 (i.e, **no of screens**) is significant or not.
- We have to test
- $H_0 : \rho_{y1}=0$ vs $H_1 : \text{not } H_0$
- The appropriate test statistic F , which under H_0 , follows an F distribution with $df=(1,24)$. Its value comes to be **90.4607**. We reject H_0 against H_1 iff $F > F_{\alpha;1,24} = 4.25$
- Here $F > F_{\alpha;1,24}$
- Hence we reject H_0 .
- Therefore ρ_{y1} is significantly different from 0.
- Thus we take number of screens as predictor at 5% level of significance.



(2) We need to calculate the 1st order partial correlation coefficients, which are given as follows:

- $r_{y2.1}=0.091$; $r_{y3.1}= -0.052$; $r_{y4.1}= 0.359$; $r_{y5.1}= -0.044$;
 $r_{y6.1}= 0.490$
- $r_{y6.1}$ is maximum, so we test whether it is significant.
- $H_0: \rho_{y6.1}=0$ vs $H_1: \text{not } H_0$
- The test statistic is given by F which under H_0 , follows an F distribution with $df=(1,23)$ whose value is **7.26**
- $F_{\alpha;1,23}=4.27$
- Since $F > F_{\alpha;1,23}$ so we reject H_0 . So factor 6 i.e. rating is worthwhile in predicting the response at 5% level of significance.



(3) Now we calculate the 2nd order partial correlation coefficients.


- $r_{y2.16} = 0.177$; $r_{y3.16} = 0.033$; $r_{y4.16} = 0.254$; $r_{y5.16} = -0.129$;
- here $r_{y4.16}$ is maximum therefore we check whether it is significant or not.
- The test statistic is F which under H_0 , follows an F distribution with $df=(1,22)$. Its value is **1.51**.
- $F_{\alpha;1,22} = 4.30$
- Since $F < F_{\alpha;1,22}$, so we accept H_0 at 5% level of significance and so conclude that sequel does not significantly affect the response and hence the remaining factors do not affect as well. So number of screens and metacritic rating are the factors that affect the response. We fit a final regression equation treating only these two factors as predictors.
- *The regression equation hence obtained is given by:*
 *$y = -19.5 + 0.0240 * \text{No of screens} + 0.318 * \text{rating}$*



ANIMATION

After sampling we get $n = 21$ movies.

(a) The multiple correlation coefficient

- $r_{y.123456} = 0.760855$
 - We now test for the significance of the multiple correlation coefficient. So we test
 - $H_0: \rho_{y.123456} = 0$ vs $H_1: \text{not } H_0$
 - The test statistic F , under H_0 , follows an F distribution with $df = (6, 14)$. Its value is **3.20771**
 - We reject H_0 iff $F > F_{\alpha; 6, 14} = 2.84773$
 - Here $F > F_{\alpha; 6, 14}$
 - Hence we reject H_0 .
 - Thus we may conclude that in the light of the given data these factors all taken together have an effect on the response at 5% level of significance.
 - Now we try to find out factors among those 6 factors which can be used as predictors .
- 

(1) The correlation coefficients between y and the factors are given below.

- $r_{y1}=0.636$; $r_{y2}=0.673$; $r_{y3}=-0.050$; $r_{y4}= 0.244$; $r_{y5}= 0.280$; $r_{y6}=0.332$
- We see that r_{y2} is maximum. So factor x_2 i.e. Production budget is chosen as predictor.
- Here we need to check whether the effect of x_2 (i.e. production budget) is significant or not.
- We have to test
- $H_0: \rho_{y2}=0$ vs $H_1: \text{not } H_0$
- The appropriate test statistic F , under H_0 , follows an F distribution with $df=(1,19)$. Its value is **15.7304**.
- $F_{\alpha;1,19}=4.38$
- Here $F > F_{\alpha;1,19}$
- Therefore we reject H_0 , i.e. ρ_{y2} is significantly different from zero.
- Thus production budget can be taken as a predictor.



(2) We need to calculate the 1st order partial correlation coefficients, which are given as follows:

- $r_{y1.2} = 0.416$; $r_{y3.2} = -0.156$; $r_{y4.2} = 0.348$; $r_{y5.2} = 0.011$; $r_{y6.2} = 0.050$
- Here $r_{y1.2}$ is maximum, therefore we need to check whether $r_{y1.2}$ is significantly different from zero or not.
- Here we will test
- $H_0: \rho_{y1.2} = 0$ vs $H_1: \text{not } H_0$
- $F = 3.76689$ $F_{\alpha;1,18} = 4.41387$
- Hence we accept the null hypothesis at $\alpha = 0.05$.
i.e. $\rho_{y1.2}$ is not significantly different from zero.
- So we conclude that number of screen does not significantly affect the response and hence the remaining factors do not affect as well. So production budget is the only factor that affects the response. We fit a final regression equation treating only this one factor as predictor.
- *The regression equation hence obtained is given by:*
 *$y = 56.8 + 0.184 * \text{production budget}$*



COMEDY

After sampling we get $n=22$ movies.

The multiple correlation coefficient is

- $r_{y.123456} = 0.906642$
- We now test for the significance of the multiple correlation coefficient. So we test
- $H_0: \rho_{y.123456} = 0$ vs $H_1: \text{not } H_0$
- The test statistic F , under H_0 , follows an F distribution with $df=(6,15)$. Its value is 11.5449 .
- We reject H_0 iff $F > F_{\alpha;6,15} = 2.79046$
- Here $F > F_{\alpha;6,15}$
- Hence we reject H_0 . i.e, $r_{y.123456}$ is significantly different from 0.
- Thus we may conclude in the light of given data that these factors all taken together have an effect on the response at 5% level of significance.
- Now we try to find out factors among those 6 factors which can be used as predictors.

(1)The total correlation coefficients are given by:

- $r_{y1}= 0.883$; $r_{y2}= 0.471$; $r_{y3}= 0.022$; $r_{y4}= 0.285$;
 $r_{y5}=-0.108$; $r_{y6}= -0.253$
- Since r_{y1} is maximum,we test for its significance. So we test
- $H_0: \rho_{y1}=0$ vs $H_1: \text{not } H_0$
- We get the value of the test statistic as:
- $F_{\text{obs}}=70.7808$; $F_{0.05;1,20}= 4.35124$
- Since $F_{\text{obs}}>F_{0.05;(1,20)}$ so we reject H_0 and conclude that X_1 i.e. Number of screens is a significant factor.



(2)Next we calculate the first order partial correlation coefficients between y and the other factors given X_1 . The values are given below.

- $r_{y2.1} = 0.037$; $r_{y3.1} = -0.083$; $r_{y4.1} = 0.085$; $r_{y5.1} = 0.038$; $r_{y6.1} = 0.404$
- Since $r_{y6.1}$ is maximum we test for its significance. So we test
- $H_0: \rho_{y6.1} = 0$ vs $H_1: \text{not } H_0$
- We get the value of the test statistic as:
- $F_{\text{obs}} = 3.70598$; $F_{0.05;1,19} = 4.38075$
- Since $F_{\text{obs}} < F_{0.05;1,19}$, so we accept H_0 and conclude that meta-critic rating is not a significant factor and hence the remaining factors do not affect the response as well. So number of screens is the only factor that affects the response. We fit a final regression equation treating only number of screens as predictor.

The fitted regression equation is given by:

$$y = 6.9 + 0.0158 * \text{screens}$$



THRILLER

After sampling we get $n=24$ movies.

Here the multiple correlation coefficient is given as

- $r_{y.123456}=.8746$.
- We know test for the significance of the multiple correlation coefficient. So we test
- $H_0: \rho_{y.123456}=0$ vs $H_1: \text{not } H_0$
- The value of the test statistic is given by:
- $F_{\text{obs}}=9.20$
- $F_{0.05;6,17}=2.69866$
- Since $F_{\text{obs}} > F_{0.05;(6,17)}$ so we reject H_0 and hence conclude that the multiple correlation coefficient differs significantly from 0.
- Thus we may conclude in the light of the given data that these factors all taken together have an effect on the response at 5% level of significance.
- Now we try to find out factors which can be used as predictors .



(1) We now calculate the correlation coefficients of the response with the different factors. The values are given below:

- $r_{y1} = 0.84770$; $r_{y2} = 0.28704$; $r_{y3} = -0.12778$; $r_{y4} = 0.44658$; $r_{y5} = -0.00634$; $r_{y6} = -0.22033$
- Since r_{y1} is maximum we test for its significance. So we test
 $H_0: \rho_{y1}=0$ vs $H_1: \text{not } H_0$
- We get the value of the test statistic as:
- $F_{\text{obs}} = 56.1792$
- $F_{0.05;1,22} = 4.30095$
- Clearly we reject null hypothesis at 5% level of significance.
- We include x_1 i.e. number of screens as a predictor in our regression equation.



(2) Now we calculate the first order partial correlation coefficients between response and the other factors when x_1 is given. The values are given below.

- $r_{y2.1} = -0.09800$; $r_{y3.1} = -0.20500$; $r_{y4.1} = 0.24100$; $r_{y5.1} = 0.22100$; $r_{y6.1} = 0.11600$
- Since $r_{y4.1}$ is maximum we test for its significance. So we test
- $H_0: \rho_{y4.1} = 0$ vs $H_1: \text{not } H_0$
- We get the value of the test statistic as:
- $F_{\text{obs}} = 1.29605$
- $F_{0.05;1,21} = 4.32479$
- Since $F_{\text{obs}} < F_{0.05;1,21}$ so we accept H_0 and conclude that sequel is not a significant factor and hence the remaining factors do not affect the response as well. So number of screens is the only factor that affects the response.

We fit a final regression equation treating it as the only factor. It is given by:

$$y = 21.5 + 0.0155 * \text{No of screens}$$



DRAMA

After sampling we get $n=24$ movies. For this genre in our data, none of the movies have sequels and hence we do not consider it as a factor affecting the response. So we have five factors($p=5$) which are denoted as follows:

- \mathbf{x}_1 : Number of screens
- \mathbf{x}_2 : Production budget
- \mathbf{x}_3 : Release Timing
- \mathbf{x}_4 : Star power
- \mathbf{x}_5 : Meta-critic Rating
- $\mathbf{y} : \sqrt[4]{\text{Opening weekend boxoffice collection}}$



Here the multiple correlation coefficient is given as

- $r_{y.12345} = 0.917660$
- We now test for the significance of the multiple correlation coefficient. So we test
- $H_0: \rho_{y.12345} = 0$ vs $H_1: \text{not } H_0$
- The test statistic F , under H_0 , follows an F distribution with $df=(6,18)$. Its value is **15.9873**
- We reject H_0 iff $F > F_{\alpha;6,18} = 2.66130$
- Here $F > F_{\alpha;6,18}$, hence we reject H_0 at 5% level of significance.
- i.e. $r_{y.12345}$ is significantly different from zero.
- Thus we may conclude that in the light of the given data these factors all taken together have an effect on the response at 5% level of significance.
- Now we try to find out those 5 factors which can be used as predictors .



(1) We now calculate the correlation coefficients of the response with the different factors. The values are given below:

- $r_{y1} = 0.87220$; $r_{y2} = 0.72326$; $r_{y3} = 0.29704$; $r_{y4} = 0.01933$; $r_{y5} = 0.17287$
- Since r_{y1} is maximum we test for its significance. So we test
- $H_0: \rho_{y1} = 0$ vs $H_1: \text{not } H_0$
- We get the value of the test statistic as:
- $F_{\text{obs}} = 69.9474$;
- $F_{0.05;1,22} = 4.30095$
- Since $F_{\text{obs}} > F_{0.05;(1,22)}$ so we reject H_0 and conclude that x_1 i.e. Number of screens is a significant factor.



(2) Next we calculate the first order partial correlation coefficients between y and the other factors given x_1 . The values are given below.

- $r_{y2.1}=0.45717$; $r_{y3.1}=0.42012$; $r_{y4.1}=0.32998$; $r_{y5.1}=0.12146$
- Since $r_{y2.1}$ is maximum we test for its significance. So we test
- $H_0: \rho_{y2.1}=0$ vs H_1 :not H_0
- We get the value of the test statistic as:
- $F_{\text{obs}}=5.54882$;
- $F_{0.05;1,21}=4.32479$
- Since $F_{\text{obs}} > F_{0.05;1,21}$ so we reject H_0 and conclude that x_2 i.e. Production budget is a significant factor.



(3) Now we calculate the 2nd order partial correlation coefficients between the response and the other factors given x_1 and x_2 . The values are given:

- $r_{y3.12}=0.29648$; $r_{y4.12}=0.22606$; $r_{y5.12}= -0.04607$
- Since $r_{y3.12}$ is maximum, we test for its significance. So we test
- $H_0: \rho_{y3.12}=0$ vs $H_1: \text{not } H_0$
- We get the value of the test statistic as:
- $F_{\text{obs}}=1.92743$
- $F_{0.05;1,20}=4.35124$
- Since $F_{\text{obs}} < F_{0.05;1,20}$ so we accept H_0 and conclude that release timing is not a significant factor and hence the remaining factors do not affect the response as well. So number of screens and production budget are the only factors that affect the response. We fit a final regression equation treating only these two variables as predictors. We get it as:

The regression equation is given by

$$y=30+0.104*\text{No of screens}+.198*\text{production budget}$$



Conclusion

From our analysis , we find the following genre wise regression model for predicting opening weekend collection. We also predict our response from some random movies taken from the year 2014 and compare them with the original values.

Action

Regression equation:

$$y = -19.5 + 0.0240 * \text{no of screens} + 0.318 * \text{critic rating}$$

Animation

Regression equation:

$$y = 56.8 + 0.184 * \text{production budget}$$



Comedy

Regression equation:

$$y = 6.9 + 0.0158 * \text{no of screens}$$

Drama

Regression equation

$$y = 30.0 + 0.0104 * \text{no of screens} + 0.198 * \text{production budget}$$

Thriller

Regression equation:

$$y = 21.5 + 0.0155 * \text{no of screens}$$



Genre	Movie	Predictor 1	Predictor 2	Y observed	Y predicted
Action	Need For Speed	No of screens =3115	Metacritic Rating =40	64.9984	67.98
Animation	Mr. Peabody & Sherman	Production Budget =145 million	-	75.3334	83.4800
Comedy	Muppet most wanted	No of screens =3194	-	64.2163	57.3652
Drama	300:rise of an empire	production budget =110 million	No of screens =3470	81.9211	87.8680
Thriller	Non stop	No of screens =3090	-	73.3049	69.3950