Ramakrishna Mission Residential College, Narendrapur

NAME: RAJATSUBHRA MISTRY

REGISTRATION NUMBER: A03-1122-0252-18

COLLEGE ROLL NUMBER: STUG/096/18

SUBJECT: STATISTICS HONOURS

ACADEMIC YEAR: 2020-21

PROJECT ON STATISTICS

Topic: Predicting **IMDb** rating of a cinema using **Multiple Linear Regression** AND Verifying whether all the factors are **significant or not**.

Contents:

Introduction:

Picking a film to watch is an emotional rollercoaster. First, you must deal with the crushing knowledge that none of your streaming services of choice have the film you want to watch. Then you narrow the field down to three films that you never really intended to watch but are the only half-decent options available.

At this point, paralysed by the thought of making the wrong decisions in life, you will Google the ratings of these films to find out if they're worth your time. Three hours later – unable to decide because of the conflicting information – you realise that it's too late to start watching an old film now anyway.

But why do the big film-ranking sites come up with such radically different options? Is **The Wizard of Oz** best film of all time, or is it **The Shawshank Redemption**?

To answer all these questions, let's look at how the biggest film-ranking sites come up with their ratings.

There are mainly three rating apps. **IMDb**, **Rotten Tomatoes**, and **Metacritic** are the three most popular ratings sites for movies. Here we ignore Metacritic ratings because their coverage is very small.

On the contrary, some people use newspaper to choose cinema. Newspapers also cover many cinemas and give them ratings by their known professional critics. There are so many peoples who watch cinemas by their relatives rating or choices. So here it completely depends on those audiences, whom they asked for.

Abstract:

IMDb is an online database of information related to films, television programs, home videos, video games, and streaming content online – including cast, production crew and personal biographies, plot summaries, trivia, ratings, and fan and critical

reviews. As of December 2020, IMDb has approximately 7.5 million titles (including episodes) and 10.4 million personalities in its database, as well as 83 million registered users.

Rotten Tomatoes is an American Review-Aggregation website for film and television. The company was launched in August 1998 by three undergraduate students at the University of California, Berkeley. It has almost 30 million registered users.

Film critics working for newspaper, magazine, broadcast media, and online publications, mainly review new releases, although also review older films. An important task for these reviews is to inform readers on whether they would want to see the film. A film review will typically explain the premise of the film before discussing its merits or flaws. The verdict is often summarised with a form of rating. Numerous rating systems exist, such as 5- or 4-star scales, academic-style grades, and pictograms, such as in the San Francisco Chronicle. These are the other rating systems generally used before watching films.

Problem Definition:

Before giving problem definition, we must know the rating systems of these different rating apps or newspaper.

On **IMDb**, all films are given an overall rating **out of ten**. In a roundabout way, these ratings are derived from votes submitted by **IMDb users, not movie critics**. All registered IMDb users can submit a single rating a number between one and ten – for any film on the website. These votes are then re-jigged so that certain demographics (newly registered users, for example) don't disproportionately influence the overall ranking of the film. This is my first column of the data set i.e., the dependent variable which I want to predict on basis of the other independent variables.

Rotten Tomatoes gives films a score **out of 100** based on the **averaged reviews of professional film critics**. If a film gets a rating of 60 or more it gets a 'fresh' red tomato on the site. Less than 60 and it gets a rotten tomato. A **red tomato score** indicates

its fresh status while a **green splat** indicates its rotten status. **Tomatometer** gives the **percentage of Approved Tomatometer Critics who have given this movie a positive review**. But my study doesn't bother about the percentage. **Tomatometer also takes the rating of this professional critics out of 10 and gives us the average**. **This is my 2nd column and 1st independent variable under study**.

The **Audience Score** is designated by a **popcorn bucket**. The score is the percentage of users who have rated the movie or show positively. There is also a section for Verified Ratings which includes those that have bought tickets. To receive a **full popcorn bucket**, at least 60% of users give a film or show a star rating of 3.5 or higher. A **tipped over popcorn bucket** indicates that less than 60% of users have given it a 3.5 or higher out of 5. It is clear from previous sentence that it takes the record of the rating of individuals out of 5 and shows us the average. I collected this rating and converted it to out of 10 structures. **This is my 2nd independent variable under study**.

We will discuss our 3rd independent variable of study in **Strategy and Sampling** part.

Newspaper, magazine, broadcast media, and online publications gives rating out of 5 or 4 stars. I took all the averages and converted it out of 10 structure. **This is my 4th column of independent variable under study**.

Without these rating apps, I took a sample of the audiences of Bengal, who never rated in IMDb or Rotten tomatoes. I told them to give rating of cinema, I mentioned out of 10. I took the average of all the response taken. This is my 4th column of independent variable.

As all are the audience rating apps or audience manipulated things, IMDb must depend upon the other rating apps, newspaper rating and audience rating (here I am talking about those audiences who never rate on any rating app). The question, I want to solve in this project is that are all these factors significant to predict IMDb? I have used 'Microsoft Excel' and 'Minitab' as my tool of analysis.

Strategy and Sampling Technique:

For regression purpose, I have shortlisted 60 films of different languages. I have collected IMDb rating, Tomatometer rating, audiencemeter rating and newspaper rating of these 60 films from internet.

Other than these data I have collected rating of each cinema in 10-star scale from almost 250 people over phone and response of 113 people by google form. I phoned my known contacts and some unknown contacts absolutely and noted down their ratings. It initially contains 96 cinemas. But there are too many people who didn't watch all the films, e.g. The Japanese Wife was seen by only 14 people out of 250. As this small size data will harm my calculations, I chose those 60 cinemas, which got the greater number of responses. I took the rounded off average of the collected data for each cinema and it made my 4^{th} column of independent variable.

On the other hand, I have designed a brief questionnaire and uploaded it on Google Forms and shared the link through various social media platforms to known associates.

The questions were as follows:

i. Name.
ii. Which language do you prefer most while watching films?
iii. Favourite genres of cinema.
iv. How much films do you watch in a month?
v. On basis of which you generally watch films?

I have collected a total of 114 responses, of which I made bar diagrams and pie chart for my initial calculations about their test and leisure time for cinema.

Concept:

Multiple Regression:

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable.

In linear regression, the relationships are modelled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function ($f(X_1, X_2, \ldots, X_n) = A_1X_1 + A_2X_2 + \ldots + A_nX_n$, type functions) of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

Key Takeaways

- Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.
- Multiple regression is an extension of linear (OLS) regression that uses just one explanatory variable.
- MLR is used extensively in econometrics and financial inference.

Formula:

$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon,$

Where,

$Y = (Y_1, Y_2, \ldots, Y_n)$ = Dependent (or Response) Variable

X_i = Explanatory variables, $\forall$ i = 1, 2, 3, …, p

α = Y intercept (constant term)

β_i = Slope coefficient for explanatory variable X_i, $\forall$ i = 1, 2, 3, …, p

ε = Error term of this model (also known as residuals)

The multiple regression model is based on the following assumptions:

- There is a linear relationship between the dependent variables and the independent variables.
- The independent variables are not too highly correlated with each other
- Y_i observations are selected independently and randomly from the population

The coefficient of determination (ρ-squared) is used to measure how much of the variation in outcome can be explained by the variation in the independent variables. ρ^2 always increases as more predictors are added to the MLR model. ρ^2 can only be between 0 and 1, where 0 indicates that the outcome cannot be predicted by any of the independent variables and 1 indicates that the outcome can be predicted without error from the independent variable.

$$\rho^2_{1234} = \left(1 - \frac{|R|}{|R_2|}\right)$$

Dataset:

Serial Number	IMDb Rating (Y)	Tomato Meter Rating (X_1)	Tomato Audience Rating (X_2)	Bengal's Rating (X_3)	Newspaper Rating (X_4)
1	8.2	7.2	8.2	8	7.7
2	8.2	6.5	8.6	8.2	6.9
3	8.4	7.4	8.8	8.9	9
4	8.4	7	9.2	8.8	7.5
5	8.2	7.7	9	8.2	8
6	7.8	7.4	7.8	8.4	8
7	8.2	7.6	8.8	8.5	9
8	8.1	6.9	8.2	8.1	7.5
9	7.8	7.7	8.2	8.4	7.6
10	8	7.3	8.2	7.7	6.6
11	7.6	5.1	8.8	7.8	5.6
12	8	5.3	8.8	8	5.9
13	8.2	8.2	8.6	8	7
14	8.2	7.3	8.6	7.8	6.7
15	8.3	5.8	8.8	8.6	6.9
16	7.9	7.9	8.2	8.3	8
17	8.1	7.4	8.2	8.2	7.5
18	7.4	4.7	7.6	7.9	6.6
19	7.3	5.2	8	7.5	5
20	8.2	7.9	8.2	7.9	7
21	7.4	6.2	7.8	8	5
22	8.3	7	8.4	8.9	9.2
23	8.4	6.5	8.4	8.7	8.8
24	8.6	9.3	9	9.3	9
25	7.5	6.8	8	7.7	6
26	7.3	6.2	7.6	7.1	7
27	8.2	7.1	8.4	8	8.2
28	8.1	7.5	9	8.4	7.5
29	7.6	6.9	8.2	8.2	7

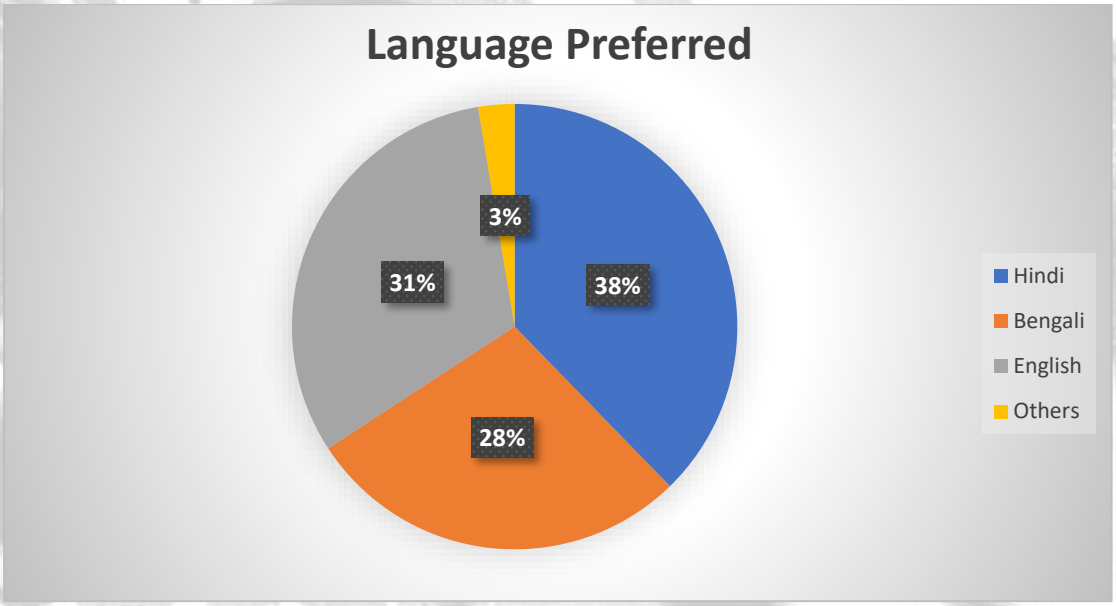
30	7.3	6.4	8	7.6	7
31	7.6	7.5	8.8	8.2	7.2
32	8.2	7.6	9.2	8.2	8
33	8.5	7.8	9	8.6	8.2
34	7.4	6.1	8.2	8.3	8
35	6.6	5.8	7.2	7.4	6.5
36	7.7	7	8.6	7.8	6.5
37	8.1	8.2	9	8.7	7.5
38	8.8	8.6	9.4	9.2	8
39	8.4	7.3	8.6	8.3	7.6
40	7.9	6.9	8.8	7.9	6.5
41	8.6	9.4	9	8.8	9.3
42	9	8.6	8.8	8.7	9.6
43	9.3	8.3	9.4	9	9
44	8.2	7.2	8.6	8.5	9.5
45	9.2	9.4	9.6	8.6	9.2
46	8.6	8.1	8.8	8.5	8
47	8.4	7.3	8.8	8.7	7.5
48	8.4	8.2	9	8.7	8.9
49	8.2	7.8	8	8.3	9
50	8.6	7.1	8.4	8.9	7.8
51	7.8	7.2	8.2	8	8
52	7.8	8	6.6	8.2	7.5
53	7	7.5	8.2	8.9	7.2
54	8.3	7.7	9	8.5	7.8
55	8.2	6.7	7.6	8.6	6.7
56	8.1	6.9	8.6	8.7	9
57	9	9.1	9.2	8.8	9.4
58	8.3	9.2	9	8.5	7
59	7.6	8.9	8.6	8.5	7.3
60	8.6	9.6	9.4	8.7	9.2

Analysis of Our Model:

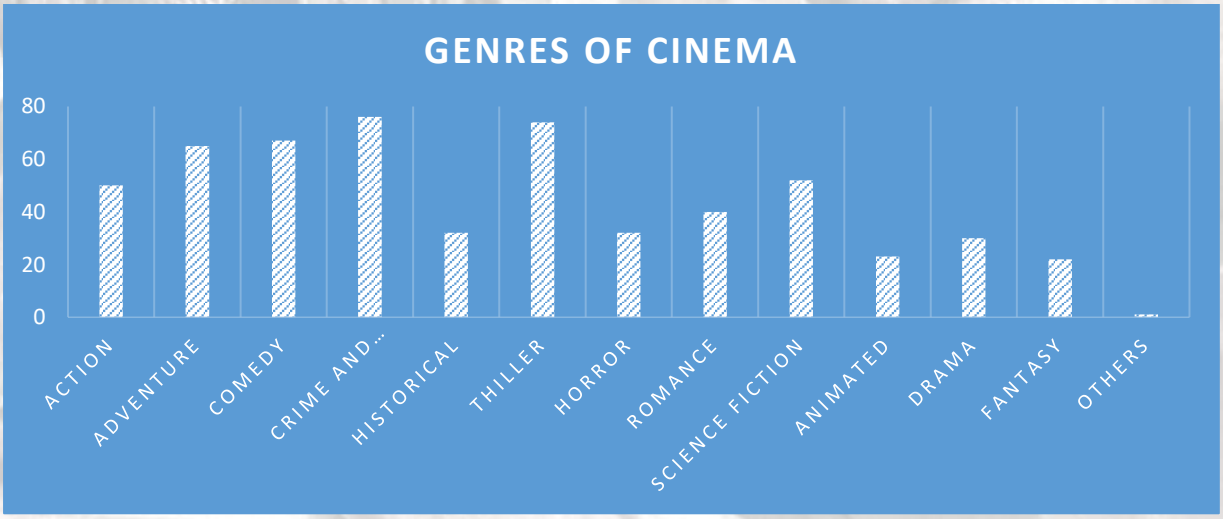
As stated earlier, I have collected 113 responses by making an google form to get some idea about film choices, genre choices and test of audiences.

The following are glimpses of the data received per question.

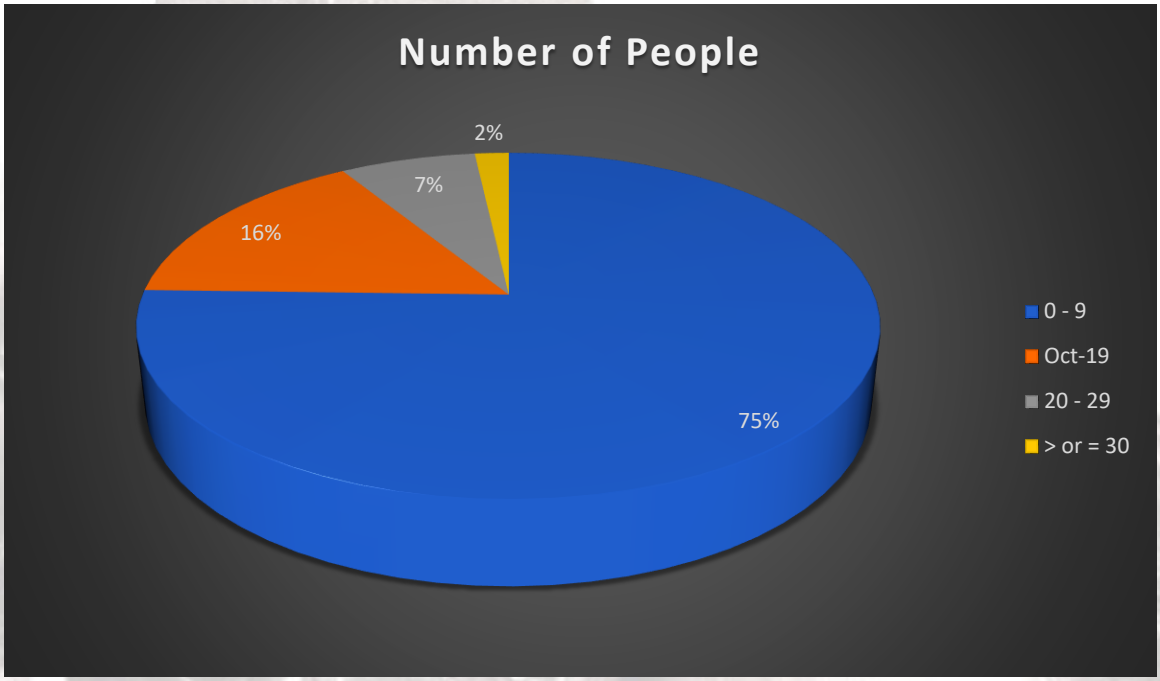
1. **Which language do you prefer watching films? (You can choose only one)**



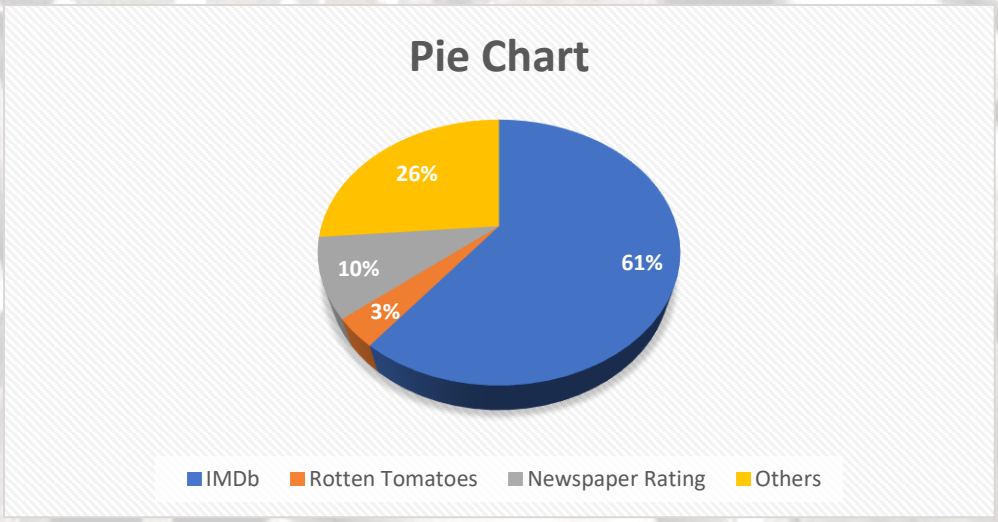
2. **Your favourite genres of cinema. (You can select more than one)**



3. **How much films do you watch in a month? (You can choose only one)**



4. **On basis of which you generally watch films? (You can choose only one)**



These are some charts regarding my data.

Now, we will concentrate on **Multiple Linear Regression**, our primary topic of interest. Firstly, our all data are treated as continuous by the corresponding rating sites like IMDb, rotten tomatoes. It is also evident that average of the rating taken over phone on 10 scale is also continuous because the way I collected the data was very similar to IMDb. Our 4th variable is also continuous as I took the average rating of all newspapers. Secondly, we have to see how much our response variable depends our predictor variables i.e., to check whether $\rho^2_{1.2345}$ is significant.

Here, $\rho^2_{1234} = (1 - \frac{|R|}{|R_2|}) = 0.6683$, where R is the correlation matrix, R_2 is the correlation matrix of predictors only. Here ρ^2_{1234} is significantly greater than 0 i.e., response variable is clearly depending on predictor variables by almost 67%.

Now,

Our regression equation,

$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$, here every notation is pre-defined.

For i^{th} individual,

$Y_i = \alpha + \beta_{1i} X_{1i} + \beta_{2i} X_{2i} + \ldots + \beta_{pi} X_{pi} + \varepsilon_i$, for all i = 1, 2, 3, …, 60

With,

Mean Vector ($\underline{X}$) =

8.093
7.373
8.52
8.33
7.66

And, Dispersion Matrix (Σ) =

0.2636	**0.3420**	**0.1998**	**0.1547**	**0.3689**
0.3420	1.1543	0.3045	0.2766	0.7408
0.1998	0.3045	0.3176	0.1397	0.2591
0.1547	0.2766	0.1397	0.2028	0.3310
0.3689	0.7408	0.2591	0.331	1.2004

Here Y depends on X_1, X_2, X_3, X_4

Now, $f(X_{(1)}) = \alpha + \beta' X_{(1)}$, where $\alpha = \bar{Y} - \underline{\beta' \bar{X}_{(1)}}$ -), $\beta = \sigma^{-1}_{(2)}\sigma_{(1)}$

Where $f(X_{(2)})$ is the part of X_1 explained by the Multiple Linear Regression of Y on (X_1, X_2, X_3, X_4)

Here,

α = 1.748

$$\beta = \begin{pmatrix} \mathbf{0.0661} \\ \mathbf{0.3688} \\ \mathbf{0.2065} \\ \mathbf{0.1300} \end{pmatrix}$$

$f(Y) = X_{1234} = \alpha + \beta' X_{(1)}$

$= 1.748 + 0.0661X_1 + 0.3688X_2 + 0.2065X_3 + 0.1300X_4$

Now $Y = X_{1234} + \varepsilon_{1234}$

So our Multiple Linear Regression of Y on (X_1, X_2, X_3, X_4) is,

$Y = 1.748 + 0.0661X_1 + 0.3688X_2 + 0.2065X_3 + 0.13X_4 + \varepsilon_{1.2345}$

With $E(\varepsilon_{1234}) = 0$ and $V(\varepsilon_{1234}) = 8.74 * 10^{-8}$ i.e. very negligible.

Tests to find significant predictors:

I. Test on individual Regression Coefficient(T-test):

The t test is used to check the significance of individual regression coefficients in the multiple linear regression model. Adding a significant variable to a regression model makes the model more effective, while adding an unimportant variable may make the model worse. The hypothesis statements to test the significance of a particular regression coefficient, βj, are:

H_0: βj=0 against H_1: βj≠0

For β_j, the test statistic for this test is based on the t distribution is,

$T_0(\beta_j) = \frac{\widehat{\beta_J}}{SE(\widehat{\beta_J})}$, where **$T_0(\beta_j)$** is the T-Value obtain in following chart for X_j

Now, we would accept the null hypothesis at 5% level if the test statistic lies in the acceptance region: $-t_{0.025,\ 58} < \mathbf{T_0(\beta_j)} < t_{0.025,\ 58}$

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.749	0.848	2.06	0.044	
X_1	0.0661	0.0513	1.29	0.203	1.91
X_2	0.3689	0.0882	4.18	0.000	1.55
X_3	0.206	0.132	1.57	0.123	2.21
X_4	0.1300	0.0537	2.42	0.019	2.18

For β_1 the test looks like **H_0: $\beta_1=0$ against H_1: $\beta_1\neq0$**

The test statistic for this test is based on the t distribution is,

$$\mathbf{T_0(\beta_1)} = \frac{\widehat{\beta_1}}{SE(\widehat{\beta_1})}$$

From the chart of **coefficients**, it is obtained that, $\mathbf{T_0(\beta_1)}$ = 1.29

Now, we would accept the null hypothesis at 5% level if the test statistic lies in the acceptance region: $-t_{0.025,\ 58} < \mathbf{T_0(\beta_1)} < t_{0.025,\ 58}$

And, $t_{0.025,\ 58}$ = 2.301

As **$-2.301 < T_0(\beta_1) = 1.29 < 2.301$**, We accept null hypothesis H_0: $\beta_1 = 0$ at 5% level of significance. And **our conclusion is X_1(Tomato Meter Rating) predictor is not significant in predicting Y(IMDb Rating).**

For β_2, $T_0(\beta_2) = 4.18 > 2.301$, and We reject null hypothesis H_0: $\beta_2 = 0$ at 5% level of significance.

For β_3, $-2.301 < T_0(\beta_3) = 1.57 < 2.301$, We accept null hypothesis H_0: $\beta_3 = 0$ at 5% level of significance.

For β_4, $T_0(\beta_4) = 2.42 > 2.301$, and We reject null hypothesis H_0: $\beta_4 = 0$ at 5% level of significance.

Our Conclusion: X_1(Tomato Meter Rating) and X_3(Bengal's Rating) predictor are not significant in predicting Y(IMDb Rating), while X_2(Tomato Audience Rating) and X_2(Newspaper Rating) predictor is significant in predicting Y(IMDb Rating).

II. P-Value Test:

If P-value of β_j coefficient is less than 0.05, we will accept that the factor X_j is significant in predicting Y.

Key Results: The predictors X_2 and X_4 have p-values that are less than the significance level of 0.05. The p-value for X_1 and X_3 is greater than 0.05.

Our Conclusion: X_1(Tomato Meter Rating) and X_3(Bengal's Rating) predictor are not significant in predicting Y(IMDb Rating), while X_2(Tomato Audience Rating) and X_2(Newspaper Rating) predictor is significant in predicting Y(IMDb Rating).

From 2 types of tests, it is evident that, we can remove X_1 and X_3 as it is not significant in predicting Y.

So, in our new regression equation, we will predict Y by X_2 and X_4, after removing unusual observations.

Fits and Diagnostics for Unusual Observations

Obs	Y	Fit	Resid	Std Resid	
52	7.800	7.379	0.421	1.68	X
53	7.000	8.041	-1.041	-3.61	R
59	7.600	8.212	-0.612	-2.09	R

R: Large residual
X: Unusual X

Regression equation of Y on X_2 and X_4 after removing 52nd, 53rd and 59th observation,

Regression Equation

$Y = 2.052 + 0.5451\ X_2 + 0.1836\ X_4$

And, Unusual observations are

Fits and Diagnostics for Unusual Observations

Obs	Y	Fit	Resid	Std Resid
31	7.6000	8.1707	-0.5707	-2.06 R
34	7.4000	7.9905	-0.5905	-2.13 R
35	6.6000	7.1700	-0.5700	-2.18 R
53	8.2000	7.4248	0.7752	2.87 R

R: Large residual

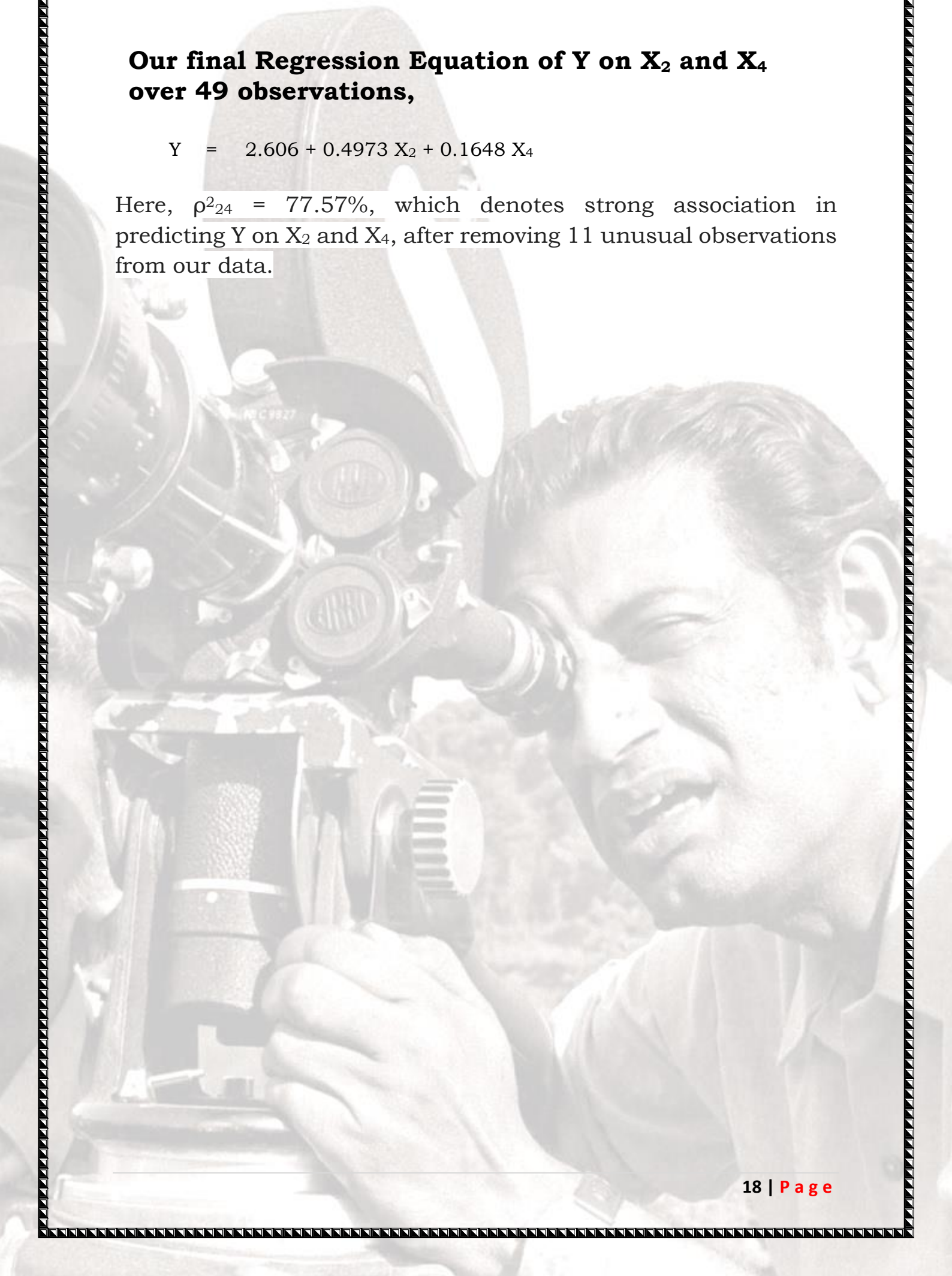
Here, $\rho^2_{24} = 69.87\%$,

Continuing this for more couple of stages, we are getting,

Our final Regression Equation of Y on X_2 and X_4 over 49 observations,

$$Y = 2.606 + 0.4973\ X_2 + 0.1648\ X_4$$

Here, $\rho^2_{24} = 77.57\%$, which denotes strong association in predicting Y on X_2 and X_4, after removing 11 unusual observations from our data.



Conclusions:

As was argued in the introduction, predicting the rating of a movie or a TV show starting from some of its measurable quantitative features, which was the goal of our study, is not an easy task. This is reflected in the performance of the models we have analysed, which seems to support the intuitive idea that the quality of a movie does not depend entirely on its other rating apps or site.

After performing the above-mentioned tests it is evident that in predicting IMDb Rating System, Tomatometer Rating and Bengal's Rating of the cinema is not significant at all. The rejection of t-test at 5% level of significance assures us to reject those.

So, given a Tomato Audience Rating and Newspaper rating, we can predict IMDb Rating sufficiently.

It is evident from our survey that more than 75% people watch only 0-9 films in a month. As we all know that, less quantity demands better quality, the films should be worth to watch. Our project result ensures that you can easily believe on IMDb rating.

Bibliography:

1. Advanced Theory of Statistics (Vol – 3) (Kendall & Stuart)
2. An Introduction to Multivariate Statistical Analysis
3. Applied Multivariate Statistical Analysis (C. R. Rao)
4. Linear Statistical Inference and its applications (Johnson & Wichern)

Links for Further Study:

https://www.imdb.com/

https://www.rottentomatoes.com/

https://en.wikipedia.org/wiki/Linear_regression

Acknowledgement:

I want to express my sincere thanks and gratitude to my professors of Department of Statistics, Ramakrishna Mission Residential College, Narendrapur, Dr Dilip Kumar Sahoo, Dr Parthasarathi Chakrabarti, Sri Tulsidas Mukherjee, Sri Palas Pal, Sri Subhadeep Banerjee for helping me to complete this project.

I would also like to thank my classmates who helped me whenever I needed their help to complete my project.

I also want to express my gratitude who contributed directly or indirectly in my project.

Certificate:

Certified that the project titled **"Predicting IMDb rating of a cinema using Multiple Linear Regression AND Verifying whether all the factors are significant or not"** is bonafide work of **Mr. Rajatsubhra Mistry (Roll No. - STUG/096/18, Registration No. - A03-1122-0252-18)** who carried out the project work under my supervision in the academic year 2020-21. Certified further that, to the best of my knowledge the work reported herein does not form part of any other project or dissertation on the basis of which a degree or award was confirmed on an earlier occasion or any other candidate.

Signature of Professor

PALASH PAL

Department of Statistics

Ramakrishna Mission Residential College, Narendrapur

Date: