The purpose of this project is to collect a comprehensive set of data on the critical & commercial performance of movies featuring the prominent star Samuel L. Jackson, and analyze this data set to answer the following questions:

* Has the passage of time significantly influenced the reception of Samuel L. Jackson’s movies?
* Have the movies with higher box office performance also been received better by audiences on IMDB?

All data used in this project was gathered using the program ParseHub, from the website IMDB (Internet Movie Database). It should be noted that the cleaned data still has a significant number of null values, particularly in the Metacritic and box office columns; the models used in this project only include the entries with no null values for the relevant variables. Additionally, I have attached an “Output” document which contains the R code, as well as the corresponding output, that I used in the analysis.

1. **Effect of time on domestic box office**

To answer the first question, we should examine three different response variables: domestic box office revenues, IMDB ratings, and Metacritic ratings. (For those who don’t know, IMDB ratings are generated by average user ratings posted to the website, while Metacritic ratings are an average of professional critical ratings posted to the Metacritic website).

First, we plot the domestic box office revenues of Jackson’s films against their release years:

A graph of a number of people

Description automatically generated

The plot appears by visual inspection to suggest that the domestic box office of Jackson’s films have generally increased over time, and the regression line supports this, with a slope of 7.274, suggesting that the domestic box office for these films increases, on average, by about $7.2 million each year. The adjusted R-square value is 0.1772, so the effect explains about 17.7% of the variation in domestic gross, which is fairly significant. However, we should note that the residuals appear to be somewhat inconsistent, and we should examine a residual plot:

A line graph with numbers and dots

Description automatically generated

The residuals do have a slight pattern, appearing to get a little further from 0 as the fitted values increase. The variation is most likely not large enough to cause a critical error, but if we were looking for more rigorous results, it would be worth using a Box-Cox transformation to normalize the residuals.

Just to be thorough, we can see that the same effect holds in a normal QQ plot:

A graph of a normal q-q plot

Description automatically generated

1. **Effect of time on IMDB rating**

Next, we will examine the potential changes in the IMDB ratings of Jackson’s films over time.

A line graph with numbers and symbols

Description automatically generated

As we can see, the IMDB ratings appear to be fairly randomly distributed, with the regression line having a slope of approximately zero. The summary output gives a slope of 0.001, with a P-value of 0.898, suggesting that time has had little effect on the audience reception of Jackson’s films.

1. **Effect of time on Metacritic score**

Finally, we will examine the potential change in the Metacritic ratings of Jackson’s films over time:

A graph of metacritic scores

Description automatically generated

The scatterplot shows what appears to be a notable decrease in the average Metacritic ratings of Jackson’s films over time. The regression summary gives the line a slope of -0.3712, suggesting that the average Metacritic rating decreases by about 0.37 each year, or 3.71 per decade. The P-value reported for this coefficient is 0.01371, suggesting that this effect is significant. We also get an adjusted R-square value of 0.04418, suggesting that about 4.4% of the change in the Metacritic ratings is explained by the passage of time; this is fairly small, but worth noting.

Just to be safe, we’ll check the residual plot to make sure our basic model assumptions are justified:

A graph of a number of dots

Description automatically generated

The residual plot shows no unusual patterns. To be safe, we’ll also check the normal QQ plot:

A graph of a normal q-q plot

Description automatically generated

Again, there is nothing concerning here.

1. **Effect of IMDB Rating on Domestic Gross**

To answer our second question, we’ll compare the IMDB rating (as a measure of audience reception) with the domestic box office gross (as a measure of commercial reception). We’ll use the IMDB rating as the explanatory variable, and the box office as the response variable.

First, the scatterplot:

A line graph with numbers and dots

Description automatically generated

Here, we can see a noticeable positive effect of the IMDB rating on the domestic gross. The regression summary gives the regression line a slope of 78.72, suggesting that the domestic gross increases, on average, by about $78 million for each full point of the average IMDB rating. The adjusted R-square value of 0.1935 suggests that about 19% of the variation in domestic gross is explained by the audience reception, as measured by the IMDB rating.

The residuals do seem to have a pattern, so we’ll check the residual plot and normal QQ plot:

A line graph with dots

Description automatically generated

A graph of a normal q-q plot

Description automatically generated

Like with the comparison of domestic box office against time, we see some notable abnormality in the residuals towards the higher end of the explanatory variable. This suggests that the model may not be entirely consistent for prediction with higher IMDB scores.

**Conclusions**

To answer our first question (“Has the passage of time significantly influenced the reception of Samuel L. Jackson’s movies?”):

The passage of time has had a significant positive effect on both commercial and critical reception of Samuel L. Jackson’s films; the commercial reception has increased significantly, while critical reception has deteriorated.

To answer our second question (“Have the movies with higher box office performance also been received better by audiences on IMDB?”):

Having a higher IMDB rating absolutely seems to predict a higher box office performance among Samuel L. Jackson’s films.

Future studies may wish to combine variables in order to produce a single model for some response variable (such as the domestic gross); however, it should be noted that since our dataset has a significant number of missing values; these are mostly films released straight-to-DVD or streaming, most of which have no box office revenue and some of which received no rating from Metacritic. Trying to combine variables would likely either result in a relatively small dataset after removing the rows with null values, or result in a biased model if we were to try and fill those values in the cleaning stage, since there isn’t anything resembling a clear-cut answer for how to handle these special cases.