Abigail Ahlquist, Apurv Singhdeo, Cameron Kerkemeyer, RJ Burjek

Alexandra Chronopoulou

STAT 425

19 October 2023

The film industry today draws in more revenue than ever before. From movie theater ticket sales to rented copies and streams, popular movies rake in millions of dollars in revenue each year from fans. With modern films, however, come costly budgets. Some movies cost companies millions of dollars to make, resulting in the production of high-quality, expensive films to be extremely risky. Given a variety of variables, we would like to study whether certain qualities of films have a higher impact on the amount of revenue that the film draws in. In the data set we are studying, we are given a list of films that were released between the years of 1986 and 2016. Along with each film, our data set contains the budget, company, country of origin, director, genre, gross, rating, release date, runtime, IMDb user rating score, star, votes, writer, and year the film was released. Using these variables in a multiple linear regression model, we will test to see if there is a correlation between any of these predictors and the gross revenue that a movie draws in.

At the start of our data analysis, we began by finding a count of unique responses for each column so that we could initially remove any predictors that may be too specific or varying for what we want to analyze. These are categorial predictors that do not have a large impact on our resulting regression. Through this, we found that in the 6820 entries there were 6731 unique names, 2759 unique directors, 4199 unique writers and 2403 unique release dates. Therefore, we thought it would be best to remove the columns for name, director, writer, and release date. These were categorical variables that were too specific, so we chose to remove them. We also chose to remove the release data because the data also includes a year column, so it would also be redundant. With these first columns removed, we were able to create a full model of the data and reduced models to compare in order to decide which model was most adequate.

We began testing parameters by creating reduced models for each and using the ANOVA table to discover whether or not a certain parameter was statistically significant. For all testing purposes, we decided that our significance level would be $\alpha = 0.05$. The first parameter we tested was company, and since this was a categorical variable we conducted a partial F-test. From the use of the ANOVA table, the F-Test resulted in a final p-value of 0.6285, which is greater than our significance level of 0.05. Therefore, we did not reject the null, meaning we chose to remove the company variable since it was not statistically significant.

Similarly, we tested the significance of the star parameter through the use of another partial F-test. In this we removed star and compared to the previous reduced model, which excluded company, using the ANOVA table, and it was found from the F-Test that the p-value was equal to 0.9999. Because this was greater than our significance level, we did not reject and

null and concluded it was not statistically significant. Therefore we chose to remove the star variable.

We then tested the country parameter following a similar procedure. Referencing the most recent reduced model, we found from the ANOVA table and the F-test that the p-value was equal to 0.02999. Since the p-value here is less than our significance level, the country parameter is statistically significant in estimating the gross income.

We then tested the runtime parameters by creating another reduced model that included all parameters from the previous reduced model, that excluded company and star, but excluded the runtime. From this, we used the ANOVA table and the F-test to find that the p-value was equal to 0.8759. Since this is greater than our significance level, we did not reject the null and decided to remove the runtime parameter from the model as well. This was surprising for us since we originally had thought this would be significant from general film industry consensus but this test showed the opposite.

Lastly, we tested for the parameter year against our most recent reduced model and found from the F-Test that the p-value was equal to 2.189e-07 which was much less than our significance level. Therefore, our year parameter must be kept in the model as it is statistically significant in estimating the gross income. All other parameters that had not yet been removed were found to be statistically significant in estimating the gross income as well. Therefore, our final MLR model in estimating the gross income included the parameters of budget, country, genre, rating, score, votes, and year.

After determining and solidifying our final model through the processes stated above, we then began to test for unusual observations. We began looking for high leverage points in our data by measuring how far a data point in the sample was from the center. From the initial observations we made, it was determined that we had many high leverage points. We then had to determine if the high leverage points were good or bad. We did this by computing the IQR and then the upper and lower bounds by using the first and third quartiles. Through the filtering of the high leverage points, we found that only observations 38, 86, 88, 124, 126, and 135 were “bad” high leverage points.

Next, we searched for outliers in our model. This is necessary in order to find observations that do not fit the model. To do this, we obtained the studentized residuals and sorted them into decreasing order. We then proceeded to compute the critical t-value using Bonferroni correction. By using this method, we got a t-value of approximately –4.487. By taking the absolute value of our t-value, we were left with 4.487. Comparing our alpha value of 0.05 to our calculated t-value, we can see that the t-value is much greater. This indicates that the observation's studentized residual is higher than the critical value of the t-distribution with Bonferroni correction. We can then consider the observation made to be an outlier. Thus, given the information stated, we were able to conclude that we do have many outliers in the data due to ten of the studentized residuals being greater than the t-value of 4.487. Therefore, we concluded

that we would flag for potential problems with our current model. We would then seek a new alternative model if issues were to arise.

After completing the analysis of the outliers in our model, we moved our attention to possible influential observations in our analysis. These observations are those that, if removed, can greatly affect the analysis of our regression. By using Cook's Distance, we are able to conclude that there are no influential points in the data. This is due to none of the points having a distance greater than or equal to one. This also was seen in the half-normal plot of the Cook's Distances.

It is also necessary to check the variance, normality, and collinearity of the model as well. After taking observations of our model, it could be seen that the variance reflects a cone shape indicating that the variance is not constant. We then analyzed the normality and found it to be a rough normal distribution. Further analysis of the QQ-plot confirmed that the model is not normally distributed due to the lack of linearity. Lastly, we found our collinearity value to equal approximately 104.7887. Because of this high collinearity, there are possible issues that could arise with our model.

Finally, we determined the confidence interval of our model. We did so by producing a randomized sample of two rows from the original data. This produced a different result each time the code ran, but regardless of which rows from the sample were chosen, it was clear that the model does not effectively estimate the gross income for each given movie within our data set. Our regression model is, however, somewhat useful when it comes to making predictions for films that are not in our current data set. To test our model's effectiveness when it comes to predicting gross revenue, we took two movies and input their corresponding data for the predictors used in our regression model. The two movies we found prediction intervals for were "The Batman" and "Everything Everywhere All at Once," which are two films from the year 2022. For "The Batman," we calculated a prediction interval with a lower bound of 210,351,847 and an upper bound of 345,078,394. Next, the prediction interval we calculated for "Everything Everywhere All at Once" had a lower bound of 27,148,864 and an upper bound of 161,820,699. The total gross income for "The Batman" ended up being 772,245,583 and did not fall within our prediction interval. However, the total gross income for "Everything Everywhere All at Once" did fall within the bounds of our prediction interval, having a gross value of 141,129,020. One possible explanation for "The Batman" having a much higher total gross income than our interval predicted could be the fact that the film was released in the year 2022. All the films in our data set were released between the years of 1986 and 2016, so using our model to predict the gross income for a film outside of that range could result in poor prediction intervals. We did, however, see "Everything Everywhere All at Once" have a gross income that fell within its prediction interval, so this only seems to be an issue for some films over others. Nevertheless, because "The Batman" had a gross income much higher than its predicted value, it could suggest that our model is only effective for films that were released between the years 1986 and 2016. With an ever-changing entertainment industry and factors such as inflation, it is very likely that

different variables outside of our data set have resulted in higher gross income values generated by movies after the year 2016.

Overall, through this case study, we were able to determine which parameters produced a multiple linear regression model intended to estimate the gross income made by a movie in the years ranging from 1986 to 2016. We did so through a process of trial and error and then proceeded to investigate any possible issues with our model. In doing so, we found that the model unfortunately did not follow normality assumptions, and the regression variables did not have a constant variance. Regarding unusual observations, our model did prove to have high leverage points and outliers, and it did not have any high influential points. Using our model, we were able to produce confidence intervals for two randomly chosen movies and prediction intervals for two recent movies decided on by our team. The model does not produce a confidence interval that correctly estimates the gross income, but it was able to produce one prediction interval that was successful in estimating the possible range for the gross income.