Predicting Movie Box Office and Virtual Stock Market Price

Olivia Fan

1. Introduction

With the advent of rapid digitization, the movie industry has encountered an explosive growth of greater than 1000 movies produced per year; consequently, it becomes a crucial concern to investors whether the movie succeeds (Bhave et al. 2015). This zealous growth subsequently gives rise to virtual stock markets (VSMs), the world's largest of which, established in 1996 is the Hollywood Stock Exchange where unlimited number of consumers can trade thousands of entertainment securities (Karniouchina 2011). It comes to the fact that not only the success of the movie itself is at stake, but the inextricable relations of the movie's success with the VSM stock price could give rise to numerous hedging implications that would allow investors to make statistically informed decisions.

While extensive literature has constructed models predicting movie box office, the assessment of which in light of virtual market proves to be a fairly understudied domain recently. Older studies have found that despite arbitrage opportunities in VSMs, the predictive power of HSX is quite high (Karniouchina 2011). Within the movie box office models, a considerate amount of work relies on methods that lack interpretability, such as multi-layer back propogation neural network and ensemble learning (Lee et al. 2018). Researchers (Bhave et al. 2015) point out that accuracy can be improved by incorporating social factors on various online platforms, in addition to classical intrinsic factors of the movie itself. Therefore, this study aims to gauge insights into significant predictors of this multi-layered relationship with recent data (~2020) via statistical methods with greater interpretability such as ARIMA, Baysian Model Averaging and decision tree.

Study Design

The aim of this multifaceted study is three-fold, which contains three inextricably intertwined complex objectives. First, we would like to analyze factors that predict movie box office. Secondly, we would like to analyze factors that predict virtual market stock prices. Finally, we would also like to assess whether virtual markets are efficient predictors of new product success, with manifestation in box office. On top of this hierarchy of research questions, the nature of this time series data set lends itself to diverse methods such as the ARIMA (autoregressive integrated moving average) model, exponential smoothing, etc. The two sets of predictor variables of interest are (1) movie budget, genre, distributor, release date, number of theaters, MPAA rating, (2) trading volume, total volume held long, total volume held short, and IPO date. The response variables are domestic, international and worldwide box office, and stock price of the 9380 movies.

2. Data & Data Processing

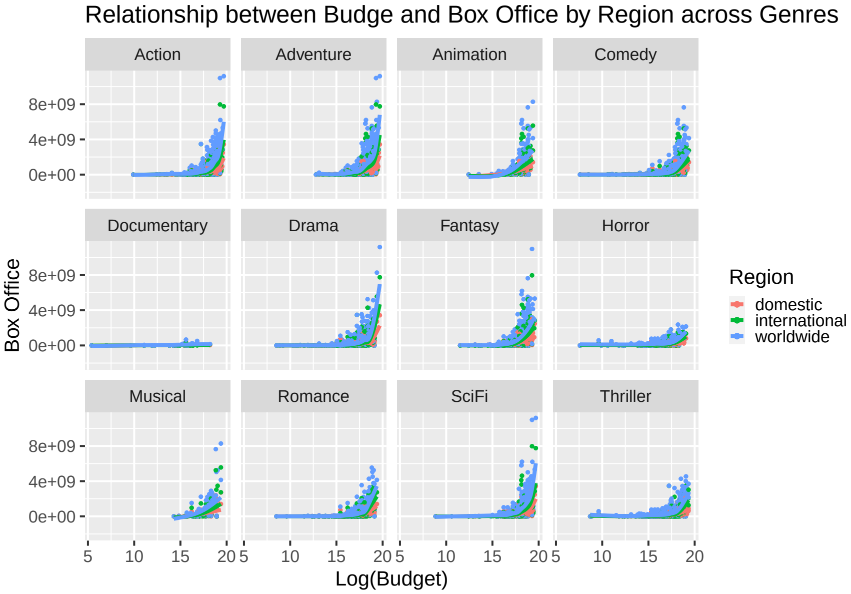
The data in this data set were scraped from two websites ("Hollywood Stock Exchange & Box Office Data" 2021): (1) Hollywood Stock Exchange (HSX.com), the world's leading virtual entertainment market which provides information on movie stock prices, (2) BoxOfficeMojo.com which tracks box-office revenue in a systematic way and provides the information on movie box office. The former HSX data source contains 325,640 daily domestic box office results (1995-2020) which includes the number of theaters exhibiting the movie release on this date and identifier of movie release; it also contains 16,968 movie releases, its identifier, budget, distributor name, domestic gross to date, international gross to date, worldwide gross to date, release date, widest release, genre and MPAA rating. The latter BoxOfficeMojo data source contains master movie data on 9,380 movies from HSX.com., i.e. genre, stock IPO date, release date, delist date, MPAA rating, number of theaters and distributor; it also contains 12,677,219 hourly movie stock prices (1997-2020) from

HSX.com, along with total number of shares held short, shares held long and trading volume at the time stamp.

In data processing, we first filtered out extraneous information irrelevant to our analysis, such as the old BoxOfficeMojo id, the BoxOfficeMojo symbol, synopsis of the movie and the BoxOfficeMojo url. To facilitate further analysis, we transformed the dates from characters to the correct format. Because many movies are attributed multiple genres, in order to analyze the impact of genre on the response, we separated the list of genres into separate rows so that each contains one category Then we filtered out missing information such as phase (with over 96% missing) and release pattern (98% missing) in HSX data, as well as domestic opening (every row contains the identical 0) in BoxOfficeMojo data.

3. Exploratory Data Analysis

Objective 1: Predicting Movie Box Office

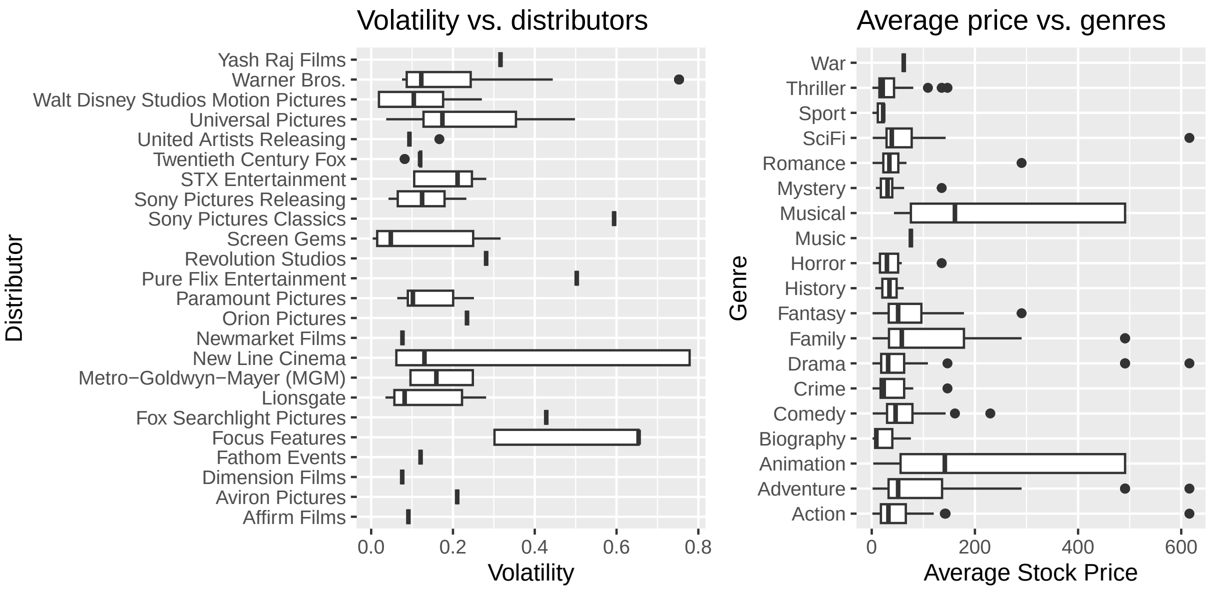


We observe that the relationship between budget and box office is vastly different across genres: Action, adventure, drama and sci-fi movies have a steep slope and generally high budget spans, with outliers which have exceedingly high budget and high box office. On the other hand, genres such as horror, thriller and romance have a much flatter slope, which corresponds to the industry knowledge that certain genres are more conducive to low-budget film making than others. According to New Review of Film and Television Studies (2011), horror and thriller movies are typically associated with such framework, and our further analysis corroborates this insight. By calculating the box office to budget ratio, we found that three out of the top 4 movie in terms of this cost effectiveness are horror or thriller movies.

title	ratio	genres
Paranormal Activity	12890.3867	Horror, Mystery, Thriller
The Blair Witch Project	4143.9850	Horror, Mystery
Tarnation	2690.9727	Biography, Dcoumentary
The Gallows	429.6441	Horror, Mystery, Thriller

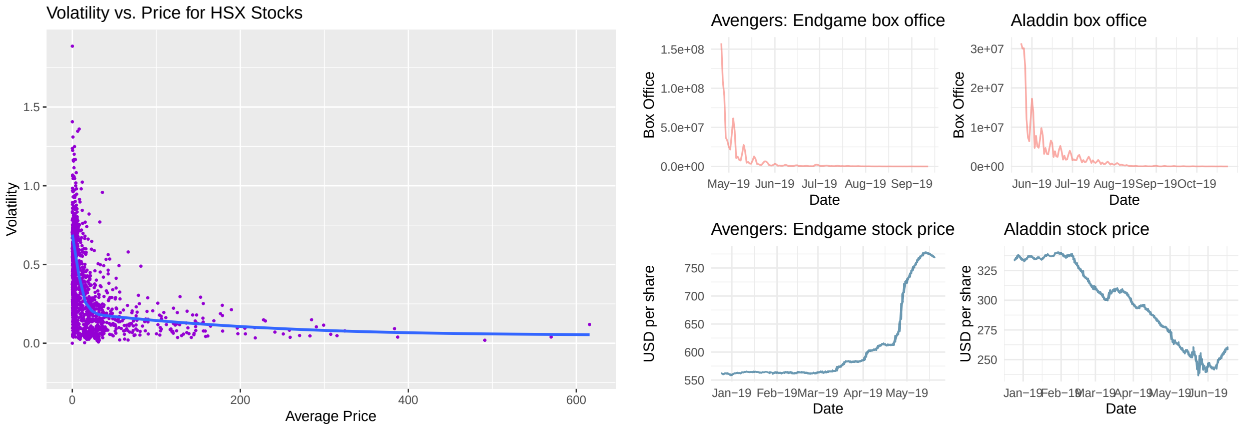
Objective 2. Predicting Stock Price

The second objective is again two fold. We first want to predict (1) the price, and (2) the volatility of the stock in order to both gauge insights into its average performance, and assess the risk in the investment. We measure the stock price as the average over time. To measure volatility, we use the standard deviation of its prices over time to quantify the rate of fluctuations, as suggested by the Corporate Finance Institute (CFI 2023).



We observe distributor's significant impact on volatility: As we might have expected, the stock of the "big name" production companies such as Walt Disney, Sony Pictures, Fox and Warner Brothers tend to have low volatility and tight range; Whereas, New Line Cinema has exceedingly variable volatility spanning from under 0.1 to around 0.8. Similarly we observe genre's impact on stock price: Musical and Animation tend to have the highest stock prices albeit a wide range, whereas history, sport and horror movies have consistently low average stock prices.

Objective 3. Relating virtual markets to product success



We observe an inverse association between average price and volatility. Comparing the time series movement of box office and stock price for movies Avenger and Aladdin, we notice that for both movies, significant movements in stock price (albeit different directions, Avenger rose in stock price while Aladdin declined) preceded significant declines in box office, hinting the predictive power of VSM on box office. One issue down the road is that the box office and the stock price come from two different websites and therefore do not have a one-to-one correspondence, leading to a portion of data missing, which could be mitigated through further web scraping or data imputation.

Statistical Analysis Plan

4. Aims & Hypotheses

Aim 1 What are the factors that affect the fluctuations of movie box office over time? Specifically, does the daily theater count or widest release correlate more strongly with the oscillations in movie box office - in other words, does a movie achieve success through continuous rapport or does a transient success suffice? How are budget, genre, runtime and distributor associated with a movie's box office? Do daily theater count correlate with, or wildest release?

Aim 2 What are the factors that affect HSX stock average price, and volatility over the span of time?

Aim 3 To what extent does the HSX stock prices predict the movie box office?

Primary Hypothesis The daily theater count correlates more strongly with daily movie box office than the widest release. While genre, and distributor have a significant effect on the box office, the budget and runtime do not have significant effect.

Secondary Hypothesis The HSX stock average price correlates negatively with the volatility over the span of time, and HSX stock prices move synchronously with fluctuations in box office.

5. Baseline Univariate Model: ARIMA

To establish a baseline model for movie box office over time independent of the covariates, we first examine a Auto-Regressive Integrated Moving Average (ARIMA) model, also known as Box-Jenkins approach (Kotu and Deshpande 2019). As a combination of two models, the auto-regressive and the moving average models, the ARIMA model helps us predict the future forecast via lagged observations and an integrated moving average. We take the time series box office of the movie Deep Sea as an example, and visualize the time series data:

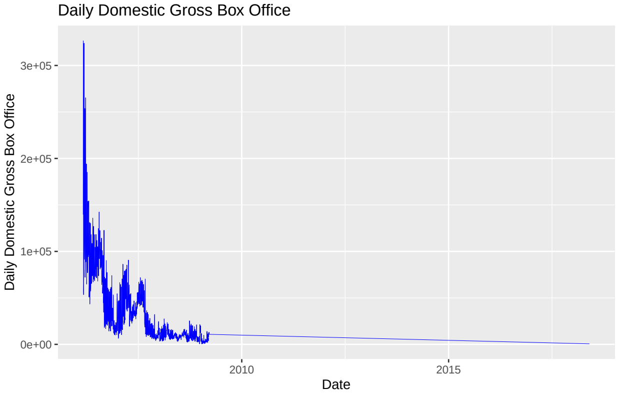


Figure 1: Time Series Visualization for Deep Sea Box Office

In order to perform any successive modeling, by model assumption, requires data to be stationary. That is, the mean, variance, and covariance of the series should be constant with respect to time, and there should not be white noise. Therefore, we take the difference between the log value of the daily domestic gross box office to stationarize the data, and demonstrate later through the Dickey Fuller Test that the data meets model assumptions (in the section below). By the same token, we can stationarize the HSX stock price data before fitting ARIMA to forecast future prices and perform the model assessment, diagnostics and sensitivity analysis in the following sections.

5.5. ARIMA: Model Assumptions, Sensitivity Analysis & Validation

5.5.1 Model Assumptions

5.5.1.1 Stationary: Dickey-Fuller test We conduct the Dickey-Fuller test to assess the stationary principle: The Dickey-Fuller test returns a p-value of 0.01, resulting in the rejection of the null hypothesis and accepting the alternate, that the data is stationary.

By by stationary it means that the properties of the series doesn't depend on the time when it is captured. A white noise series and series with cyclic behavior can also be considered as stationary series.

5.5.1.2 Univariate We assess box office as the univariate response variable, which aligns with ARIMA's assumptions that data should be univariate, since ARIMA works on a single variable.

5.5.2 Sensitivity Analysis: ACF/PACF There are primarily two hyperparameters in the model that we can tune to perform sensitivity analysis, MA (moving-average) and AR (auto-gression) coefficients. The ACF (Auto-Correlation Function) gives us values of any auto-correlation with its lagged values which will help us determine the number of MA coefficients in our ARIMA model, while the PACF (Partial Auto-Correlation Function) finds correlation of the residuals with the next lag value which helps us identify the number of AR coefficients in our ARIMA model. In the ACF graph below, the curve drops significantly after the first lag, which indicates a moving average component of MA(1). We can tune the MA and AR coefficients to achieve sensitivity analysis.

The standard ARIMA models expect as input parameters 3 arguments, p which standards for the number of lag observations, d which is the degree of differencing, as well as q which is the size of the moving average window. This study will tune the parameters via cross validation, as well as sensitivity analysis.

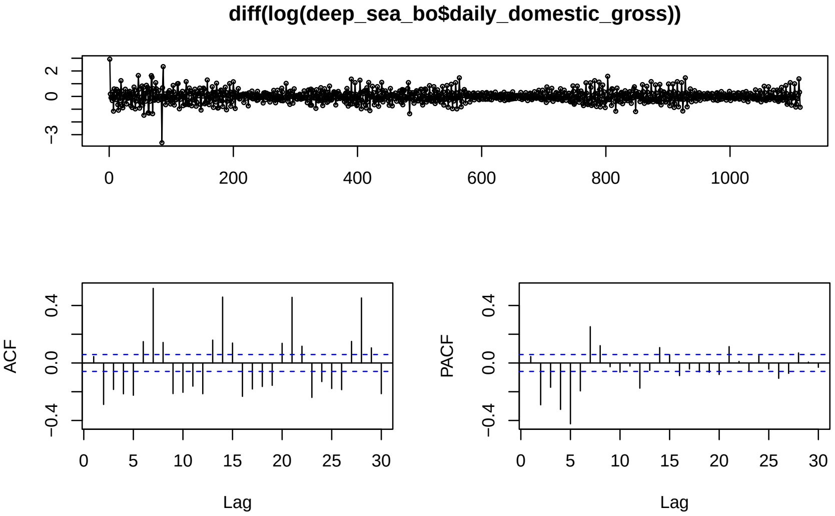


Figure 2: ACF and PACF for the Deep Sea Box Office Model

Additionally, the study also aims to perform sensitivity analysis based on seasonality, which compares the robustness of the model over the seasonal span.

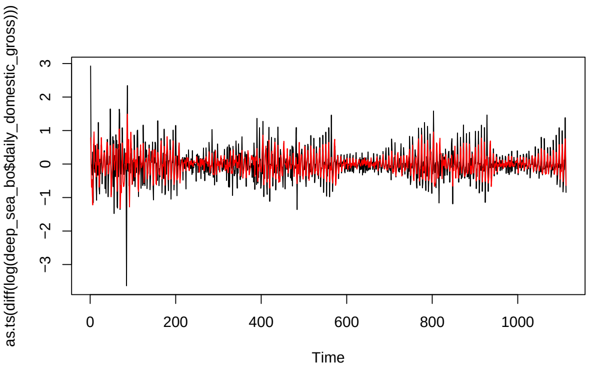


Figure 3: Training Fit for ARIMA Model

5.5.3 Preliminary Results & Validation As preliminary results, we obtained a with p (AR coefficient) of 5, d (Integrated value) of 0, and q (MA) value of 2 which obtains an AIC value of 759.91, and BIC value of 800.03. The graph above demonstrates that the model is a close fit to the training data. Splitting past data into a training set (pseudo future data), we can examine performance on this pseudo future data to achieve cross validation.

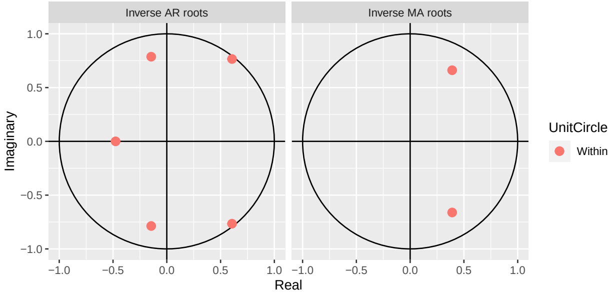


Figure 4: Diagnostic Plot for ARIMA Model

5.5.4 Model Diagnostics Plotting the characteristic roots for the model fitted, we see that they are all inside the unit circle, as we would expect because R ensures the fitted model is both stationary and invertible.

6. Multivariate Model: Vector Autoregression (VAR)

While the baseline ARIMA model gauges insights into the changes into the fluctuations of the box office over the span of time in and of itself, since we are essentially interested in the the factors that predict the box office or stock price, we resort to the VAR (Vector Autoregression) model which is essentially a generalization of the univariate autoregressive ARIMA model. A VAR model is a type of multivariate time series model that can capture the dynamic interactions between multiple time series variables via the assumption that each variable is a function of its own past values as well as the past values of other variables in the system (Stock and Watson 2001). For movie box office, we are primarily interested in genre, distributor, MPAA rating, budget, daily theater count, widest release and runtime which we plan to include in the model. For the former three categorical variables, we plan to use one-hot encoding to convert them to factor levels. Taking the box office data of the movie Titanic as an example, we explore the time series visualization of the quantitative variables below:

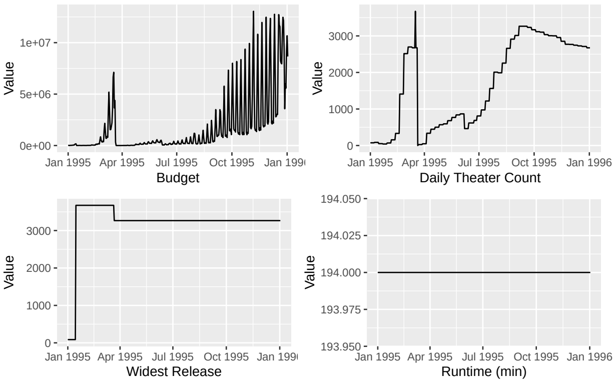


Figure 5: Time Series Plots of Quantitative Variables for Titanic

By the same token, for HSX stock prices, we are primarily interested in genre, distributor, number of shares short, number of shares long, total trading volume which we plan to include in the model. For the former categorical variables, we plan to use one-hot encoding to convert them to factor levels.

6.5 VAR: Model Assumptions, Sensitivity Analysis & Validation

6.5.1 Model Assumptions

6.5.1.1 Stationary Principle In the same token as ARIMA, since VAR is essentially a generalization of ARIMA in the multivariate case, we would also like to assess whether the variables under study are stationary. We use the Philips Perron test to assess the stationary principle, which finds that the response variable (daily domestic gross box office) along with all the predictor variables of interest above (daily theater count, wildest release, and runtime in minutes) having p values of 0.01, 0.05, 0.018 and 0.01 respectively. Therefore, we reject the null hypothesis which suggests that the data is stationary.

6.5.2 Preliminary Results & Validation We fit a preliminary model using the daily theater count, widest release and daily domestic gross box office which obtained an adjusted R^2 value of 0.93. After this, we will select the optimal lag order behind the VAR we will be using, which is 8 from the model output. Lastly, we will run diagnostics tests for autocorrelation, heteroscedasticity, normality and stability. By the same token, we plan to fit the VAR model for stock price using the number of shares long, the number of shares short and the trading volumne as predictors. Splitting past data into training set to create a pseudo future dataset from the dataset given, we plan to examine performance on future data via cross validation.

6.5.4 Diagnostics

6.5.4.1 Non-autocorrelated Residuals We first assess whether our model meets the assumption that the residuals should be non-autocorrelated, based on our assumption that the residuals are white noise and thus uncorrelated with the previous periods. We run the Breusch-Godfrey test for serially correlated errors to obtain a p value of 0.01, therefore see that the residuals do not show signs of autocorrelation. However, in case that is a chance that if we change the maximum lag order, there could be a sign of autocorrelation. Therefore, this study aims to experiment with multiple lag orders which we will confirm through sensitivity analysis.

6.5.4.2 ARCH Effects: Heteroscedasticity Another aspect to consider is the presence of heteroscedasticity, essentially clustered volatility areas in a time series dataset known as ARCH effects, which is common is time series data such as stock prices where massive rises or declines could be seen (Stock and Watson 2001). Through the ARCH test, we obtain a p value of less than $2.2e^{-16}$ under degrees of freedom of 540, which signifies no degree of heteroscedasticity as we reject the null hypothesis.

6.5.4.3 Normality The VAR normality test has three components: the Jarque-Bera test, the Kurtosis test, and the Skewness test. All of the three tests give us a p value of less than $2.2e^{-16}$. Therefore, based on all the three results, it appears that the residuals of this particular model are normally distributed.

6.5.4.4 Stability Finally, we perform the stability test through the CUSUM test which assesses the stability of the covariates in the time series VAR model via a plot of the sum of recursive residuals (Stock and Watson 2001). The diagnostic plot indicates structural breaks if at any point in the graph, the sum goes out of the red critical bounds. As we can see from the diagnostic plot below, while neither daily theater count nor widest release presents a structural break, the daily domestic gross box office slightly exceeds the critical bounds.

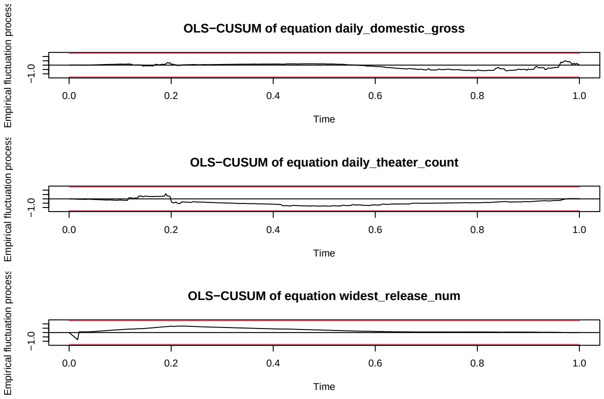


Figure 6: CUCUM Test for VAR Model

6.5.5 Sensitivity Analysis: Lag Structure Apart from the aforementioned multiple lag orders, this study also aims to perform sensitivity analysis by varying the lag structure in the VAR model. According to previous study (Hafer and Sheehan 1989) which examines the effect of lag structure on forecasting accuracy, the forecasting accuracy of the VAR model varies dramatically across simple ad hoc rules, versus statistical criteria such as mean square error and Bayesian rules. This study aims to explore various lag structures such as Bayesian rules and MSE to perform sensitivity analysis.

7. Predictability between Two Time Series: CCF

Given the relationship between two time series, we decide to use the cross correlation function (CCF) model to identify lags of the fluctuations in HSX stock prices that might be useful predictors of movie box office. Since the CCF model gives us informative information on the order of prediction between movie box office and HSX stock prices through the set of sample correlations, we can also identify which variable is leading and which is lagging. We can perform similar model assumption checks for stationary principle, as well as cross validation through creation of a pseudo training future dataset, as well as aim to perform sensitivity analysis through varying the width of the moving average window.

8. Citations

Bhave, Anand, Himanshu Kulkarni, Vinay Biramane, and Pranali Kosamkar. 2015. "Role of Different Factors in Predicting Movie Success." In *2015 International Conference on Pervasive Computing (ICPC)*, 1–4. https://doi.org/10.1109/PERVASIVE.2015.7087152.

CFI. 2023. "Volatility: A Measure of the Rate of Fluctuations in the Prices of a Security over Time," January. https://corporatefinanceinstitute.com/resources/capital-markets/volatility-vol/.

Hafer, R. W., and Richard G. Sheehan. 1989. "The Sensitivity of VAR Forecasts to Alternative Lag Structures." *International Journal of Forecasting* 5 (3): 399–408. https://doi.org/https://doi.org/10.1016/0169-2070(89)90043-5.

"Hollywood Stock Exchange & Box Office Data." 2021. https://www.kaggle.com/datasets/zeegerman/hollywood-stock-exchange-box-office-data.

Karniouchina, Ekaterina V. 2011. "Are Virtual Markets Efficient Predictors of New Product Success? The Case of the Hollywood Stock Exchange*." *Journal of Product Innovation Management* 28 (4): 470–84. https://doi.org/https://doi.org/10.1111/j.1540-5885.2011.00820.x.

Kotu, Vijay, and Bala Deshpande. 2019. "Chapter 12 - Time Series Forecasting." In *Data Science (Second Edition)*, edited by Vijay Kotu and Bala Deshpande, Second Edition, 395–445. Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-814761-0.00012-5.

Lee, Kyuhan, Jinsoo Park, Iljoo Kim, and Youngseok Choi. 2018. "Predicting Movie Success with Machine Learning Techniques: Ways to Improve Accuracy." In *Information Systems Frontiers*, 20:1–1. 577-588. https://doi.org/10.1007/s10796-016-9689-z.

Stock, James H., and Mark W. Watson. 2001. "Vector Autoregressions." *Journal of Economic Perspectives* 15 (4): 101–15. https://doi.org/10.1257/jep.15.4.101.