Subscribers Forecasting of Netflix Based on Multiple Linear Models

Guanyao Wang†, Zhe Wang†, Yiming Xie*, †

Faculty of Arts and Science, University of Toronto, Toronto, Canada

*Corresponding author: elaineee.xie@mail.utoronto.ca

Abstract. Netflix is one of the world's leading entertainment services, with millions of subscribers all over the world. To make a prediction of the expected number of future subscribers is a meaningful and valuable research topic. However, there is no authoritative model to help make predictions. Therefore, this study will explore the relationship between possible factors and the number of future subscribers in terms of the data from Netflix over the years. Subsequently, some candidate multiple linear regression models are constructed and the "best" model is selected to help make predictions. The final model shows that the number of subscribers is related to four variables, i.e., a negative relationship with the price of the basic plan, a positive relationship with the price of the standard plan, the number of countries where Netflix is available and the medium level of annual world income. The data on the number of subscribers over the years shows an increase in subscribers every year, as well as the amount of growth varies from year to year. In other words, the increase in the price of the basic plan may lead to a decrease in the number of subscribers, while the increase in the price of the standard plan, in the number of countries where Netflix is available and in the medium level of annual world income may lead to an incline in it. These results shed light on guiding further exploration of having a practical method to predict the number of future subscribers for Netflix.

Keywords: Multiple linear regression model; Netflix Subscribers; COVID-19.

1. Introduction

Netflix is one of the world's leading entertainment services, with millions of subscribers all over the world. Almost 15% of Internet downlink traffic was attributed to Netflix. To be specific, it controlled 26.58% of the global market for video streaming services [1]. It offers three subscription plans of different prices, where members can choose the best suitable option for their needs. In the statistical field, predicting the number of future users is worth studying. As the market demand develops and varies frequently, Netflix regularly adjusts its prices to adapt to the market and understand more about consumer needs in order to serve its members better and attract more subscribers. However, it is unknown that whether the change in the prices will have an impact on subscribers. Therefore, it is essential for Netflix to know the impact ofthe change on subscribers so that Netflix can makes the next-year decisions directly [2-5].

One of the Previous research topics is impacts of binge watching on Netflix during the COVID-19 pandemic. In this study, we used the form of a questionnaire to explore the changes in the number of Netflix subscribers during the COVID-19. Variables considered in the research are the gender, age range, profession and the ownership of Netflix subscription. In the next two tables of the research, we explored the devices used for nonstop Netflix consumption and the satisfaction level of binge-watching on Netflix. In this section, the data is collected about the devices used by Netflix subscribers and their satisfaction by studying the number of subscribers obtained in the above table. Afterwards, it is concluded that more Netflix subscribers are more concerned about portability, and these subscribers are generally satisfied with Netflix [6-8].

Therefore, this study will construct a multiple linear regression model to help analyze future Netflix's subscribers count and its variation trend, by collecting data on yearly number of subscribers, with data related to the possible factors that may have influence subscribers. Subsequently, these selected factors will be utilized to construct some candidate multiple linear regression models and select the "best" model. The rest of the paper is organized as follows. The Sec. 2 will introduce the data set and the method which will be used, including an overview of the samples, variables, the

selection of models and procedures of the research. The Sec. 3 will show the specific empirical analysis for the progress of model construction, with references of the tables to provide an intuitive understanding. Moreover, an analysis for further research will be presented.

2. Data & Method

2.1 Samples

Netflix was launched in 1997. Initially, it only offered a single-rate rental service for DVD and Blue-ray discs in the United States. Then, beginning in 2011, it splits the subscription service into two different packages: one for streaming and the other for DVD rental. The turning point was in 2013, with the release of the drama series House of Cards, Netflix started offering specifically produced video content for its streaming service, which has gradually become the mainstream. In this age of streaming, sales of physical DVDs have dropped dramatically. During its 25-year history so far, Netflix had not started to offer three different prices of plans as nowadays until 2014. Therefore, only the data of streaming services between 2014 and 2022 will be discussed in this paper.

2.2 Variables

The dataset includes several possible variables which may be used to construct a multiple linear regression model to make a prediction of the number of future subscribers (*Subscribers*): years from 2014 to 2022 (*Year*), which work as the index of the dataset, the price of basic plan (*Basic*), standard plan (*Standard*) and premium plan (*Premium*), number of countries where Netflix is available (*Countries*) , and the world annual middle income (*Income*).

2.3 Models

To determine whether the quantities are associated and if the strength of association exists, a linear regression model is a useful tool. In this research, a full model is constructed first which assumes the response variable are related to all the independent variables. According to the summary output of the full model, the model is not fitted well, with some p-values larger than 0.05. Therefore, a transformation on the variables can be applied. While after the transformation, the summary output of the transformed model is still poor fitted. Then, it is essential to use R to help select the model, leading to a final model which contains *Basic*, *Standard*, *Countries* and *Income*. On this basis, the summary output for the final model shows the satisfied p-values resulting to a good fitted linear regression model.

2.4 Procedures

First of all, several packages need to be installed and use library() function to load them in RStudio: tidyverse, car, rms and ggplot2. Then, one needs to import the csv. file of the needed table and use read_csv() function to run it, which contains 9 rows and 7 columns: *Year*, *Basic*, *Standard*, *Premium*, *Countries*, *Income*, *Subscribers*, grouping by years from 2014 to 2022.

Assume all the predictors are associated with the response, a full model can be constructed by the lm() function first. Then, cor() function with using all complete pairs of those particular variables to compute is applied to create a correlation matrix to check the correlation between variables. A summary() function helps generate a summary output of the full model, where several p-values are larger than 0.05. In order to check whether multicollinearity exists or not, vif() function in car package in R Language is run. Nevertheless, the variance inflation factor of some of the variables are abnormally large, therefore, a log transformation log() can be applied to improve the full model. Unfortunately, some p-values are still not smaller than 0.05, so step() function is applied to help select the appropriate variables to construct the final model by lm() function. As a result, a new summary table of final model is generated to obtain the regression model.

3. Results & Discussion

3.1 Empirical analysis

In order to find the optimal model, it is necessary to construct a full model first, which includes all the predictors that may represent the true relationship with the response variable. Using a 95% confidence interval ($\alpha = 0.05$). Then, in order to check the correlation between variables, R is a helpful tool to help get a correlation matrix. As shown in Table 1, the correlation matrix shows that the dependent variable Subscribers is positively and strongly correlated with all of the five predictors [1]. Then, the R language is utilized to implement the parameter estimation of the regression model, with the help of lm() function to do the modeling process. As listed in Table 2, the model is poorly fitted. The parameters of Basic, Standard and Premium are not significant, as their p-values are greater than 0.05 which fails the test. In order to effectively optimize the model, the existence of multicollinearity should be tested first. With the help of R, the function vif() in the car package can be used to check the variance inflation factor (VIF). To be specific, the vif of *Countries* and *Income* are less than 5, while *Basic*, *Standard* and *Premium* are larger than 10, indicating the existence of multicollinearity. Therefore, the log transformation can be applied to try to make the highly skewed distributions less skewed.

Table 1. Correlation matrix

	Basic	Standard	Premium	Countries	Income	Subscribers
Basic	1.0000000	0.9421416	0.9373850	0.4282064	0.5880846	0.8802222
Standard	0.9421416	1.0000000	0.9891014	0.6201081	0.5192965	0.9759817
Premium	0.9373850	0.9891014	1.0000000	0.5799086	0.5679969	0.9751154
Countries	0.4282064	0.6201081	0.5799086	1.0000000	-0.1224886	0.6628210
Income	0.5880846	0.5192965	0.5679969	########	1.0000000	0.5815590
Subscribers	0.8802222	0.9759817	0.9751154	0.6628210	0.5815590	1.0000000

Table 2. Summary of the regression models

	Estimate	Standard Error	t value	Pr (>|t|)
(Intercept)	-689.95590	119.86212	-5.75600	0.0104
Basic	-25.78663	9.34269	-2.76000	0.0701
Standard	22.52889	7.10879	3.16900	0.0505
Premium	4.71840	4.55487	1.03600	0.37640
Countries	0.30775	0.07590	4.05500	0.0270
Income	0.16063	0.02819	5.69800	0.0107

Table 3. Summary output for transform models

	Estimate	Standard Error	t value	Pr (>|t|)
(Intercept)	0.5548455	1.7541483	0.316	0.7725
Basic	-0.3375568	0.1367277	-2.469	0.0902
Standard	0.2943882	0.1040351	2.83	0.0662
Premium	-0.014443	0.0666592	-0.217	0.8424
Countries	0.0034449	0.0011108	3.101	0.0532
Income	0.0007943	0.0004126	1.925	0.1499

According to Table 3, the model seems to be poorly fitted. The constant term and the parameters of the five predictors are not significant, with p-values greater than 0.05 leading to the failure of the test. In order to optimize the model, the function step() works for choosing a model by AIC in a Stepwise Algorithm. The process is to iterate between forward and backwards selection until the model becomes unable to add or delete further variables. After the stepwise selection, given in Table 4, only *Basic*, *Standard*, *Countries* and *Income* are chosen as variables in the model. Their p-values

become smaller than 0.05 except for the intercept and Income. Therefore, the final model only includes *Basic*, *Standard*, *Countries* and *Income*. Its summary output is shown in Table 5. The relationships between *Subscribers* and *Basic*, *Standard*, *Countries* and *Income* are statistically significant since the p-values for these terms are less than the significance level of 0.05, indicating the pass of the test. Therefore, the final regression model is:

$$Subscribers = -727.43359 - 26.31336 \times Basic + 28.76522 \times Standard + 0.30522 \times Countries + 0.17029 \times Income \quad (1)$$

In addition, the resulting adjusted R^2 is 0.9939, which shows about 99.39% of total variability in the number of subscribers can be explained by the basic plan fee, the standard plan fee, the number of countries where Netflix is available and the middle income per capita. This suggests the model provides a good fit to the data.

Table 4. Summary output for the variables after step function

	Estimate	Standard Error	t value	Pr (>|t|)
(Intercept)	0.6695644	1.4595717	0.459	0.67024
Basic	-0.3359445	0.1191556	-2.819	0.04786
Standard	0.2752989	0.0482877	5.701	0.00468
Countries	0.0034526	0.000969	3.563	0.02352
Income	0.0007647	0.0003398	2.250	0.08760

Table 5. Summary output for the final models.

	Estimate	Standard Error	t value	Pr (>|t|)
(Intercept)	-727.43359	115.31107	-6.308	0.00323
Basic	-26.31336	9.41369	-2.795	0.04905
Standard	28.76522	3.81489	7.540	0.00166
Countries	0.30522	0.07655	3.987	0.01630
Income	0.17029	0.02685	6.344	0.00316

3.2 Model Interpretation

The final model shows that the number of subscribers is related to four variables. Specifically, the relationships can be summarized as a negative relationship with the price of the basic plan, a positive relationship with the price of the standard plan, the number of countries where Netflix is available and the medium level of annual world income. When the standard plan fee, the number of countries where Netflix is available, the middle income per capita and the basic plan fee is 0, the mean value of subscribers is -727.43359 million. When the standard plan fee, the number of countries where Netflix is available and the middle income per capita are fixed, as the basic plan fee increases one unit, the average of subscribers decreases 26.31336 million. When the basic plan fee, the number of countries where Netflix is available and the middle income per capita are fixed, as the standard plan fee increases one unit, the average of subscribers increases 28.76522 million. When the basic plan fee, the standard plan fee and the middle income per capita are fixed, as the number of countries where Netflix is available increases one unit, the average of subscribers increases 0.30522 million. When the basic plan fee, the standard plan fee and the number of countries where Netflix is available are fixed, as the middle-income per capita increases one unit, the average of subscribers increases 0.17029 million. In brief, the increase in the price of the basic plan may lead to a decrease in the number of subscribers, while the increase in the price of the standard plan, in the number of countries where Netflix is available and in the medium level of annual world income leads to an incline in it.

3.3 Analysis for further research

According to the final model, subscribers are only affected by the subscription fee of basic and standard plans. Based on the yearly change in the price of the two plans, an analysis of future research

is implemented: As listed in Table. 6, the count of subscription fees increased or not for every plan is two times for the basic plan and six times for the standard plan.

In 2016, these two plans didn't increase their price, and the number of subscribers was 47.9 million. However, in 2017, the standard increased. Then, the rate of increase for the number of subscribers decreased by 3.53%. In 2018, these two plans didn't increase their price. Therefore, the rate of increase increased by 1.66%. In 2019, all two plans increased. Then, the rate of increase decreased by 3.68%, which is similar to the decrease in 2017.

Table 6. Annual quantity of subscribers

Year	Subscribers (unit: million)	Compared with Previous Year	Rate of Increase	Net Rate of Increase	If the Price of Basic Plan Increased	If the Price of Standard Plan Increased
2014	47.9	12.3	34.55%	34.55%	No	Yes
2015	62.7	14.8	30.90%	-3.65%	No	Yes
2016	79.9	17.2	27.43%	-3.47%	No	No
2017	99	19.1	23.90%	-3.53%	No	Yes
2018	124.3	25.3	25.56%	1.66%	No	No
2019	151.5	27.2	21.88%	-3.68%	Yes	Yes
2020	192.9	41.4	27.33%	5.45%	No	Yes
2021	219.7	26.8	13.89%	-13.44%	No	No
2022(April)	221.64	1.94	0.88%	-13.01%	Yes	Yes

Seen from the linear relationship of yearly subscribers and the year (Fig. 1), it shows a positive relationship which means that as the year grows, subscribers increase, too. Although the number of subscribers is constantly growing, this increase doesn't mean that the number of subscribers was not affected by the increase in the subscription fee of plans since the total population also increases yearly. Thus, it is evident that the increase in the price of each plan will decrease the number of subscribers and the rate of increase of subscribers.

Since 2020 till now, an important environmental factor has existed in COVID-19. There was a high increase from 2019-2020, not only in the number of subscribers but also in the rate of increase of subscribers (5.45%). These changes are mainly caused by COVID-19. During this pandemic, many people have lost their jobs [6], and there are also many restrictions on traveling [8]. Therefore, more and more people will spend time at home watching TV dramas and movies. That's why there will be a high increase here.

However, the increase rate decreased significantly in these two years (2021 and 2022). In 2021, there was no increase in the two plans. However, the rate of increase was still decreased by 13.44%. In 2022, all plans increased, which decreased 13.01% from last year.

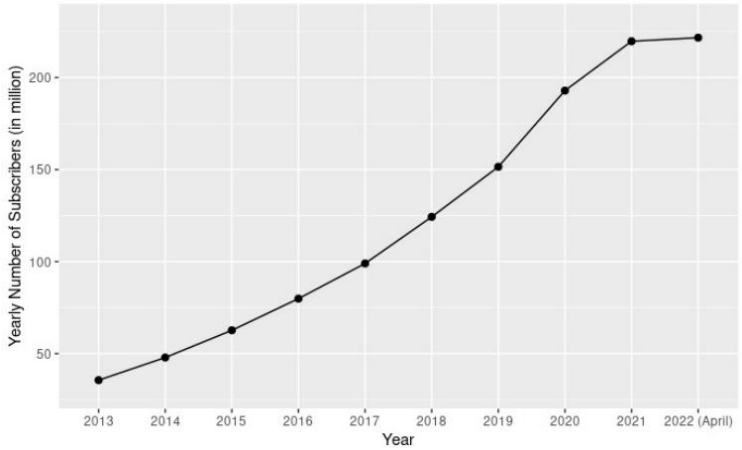

Fig. 1 The trend of annual subscribers.

Therefore, several possible factors below may affect the change in subscribers. As for the economic aspect, some people plan to give up subscribing to the membership to reduce expenses because the employer's income decreases and the price of daily things people need increases [9]. In addition, many production teams cannot work because of the epidemic. This results in reduced film and television production between 2020 and 2022. Since the output decreases, the quantity of new films and televisions is fewer than that of the year before COVID-19. Thus, there is nothing to watch for people in these two years. Besides, more and more people found that online film and television websites are profitable during the epidemic. Therefore, many new video software emerged in the market. This caused high market competition, and Netflix became less popular. According to the data, in 2022, Netflix only ranked 21st among all video websites in terms of total page views, while Netflix's competitors are not only YouTube, which ranks higher, but also many other video websites like imdb.com, which has also been listed as one of the potential competitors [10]. Regarding to the copyright, some websites on the market allow people to watch TV dramas or movies that need to pay on Netflix for free. Except for this, Netflix also lacks the copyright for some TV dramas or movies. These two phenomena are also one main reason for reducing subscription members. Here is an example, when the eighth season of the Game of Thrones, the hottest and most discussed season of this TV series, was broadcast in 2019, Netflix had no copyright to show this play then. Thus, people can only watch the Game of Thrones on the HBO platform. This caused many Netflix customers to join HBO members instead of Netflix members [11].

3.4 Limitation

In general, it is essential to check the assumptions of linearity, independence, homoscedasticity and normality for a regression model. However, not all assumptions are satisfied for this model. According to the residual plot illustrated in Fig. 2, the blue line is not horizontal, while the standard residuals are randomly dispersed around the horizontal axis with no systematic pattern. This indicates that the linearity assumption is valid where there is some linearity between the response variable and the predictors. Next, as the residuals do not appear in clusters, the independence assumption satisfies. Then, since there is no spread of the residuals which leads to validation of homoscedasticity assumption, indicating a constant variance. According to the Normal Q-Q plot depicted in Fig. 3, the residual points are discretely distributed, with deviating from the Q-Q line, which indicates a violation of the normality assumption. To sum up, all the linearity, independence and homoscedasticity assumptions satisfy, except for the violation of the normality assumption.

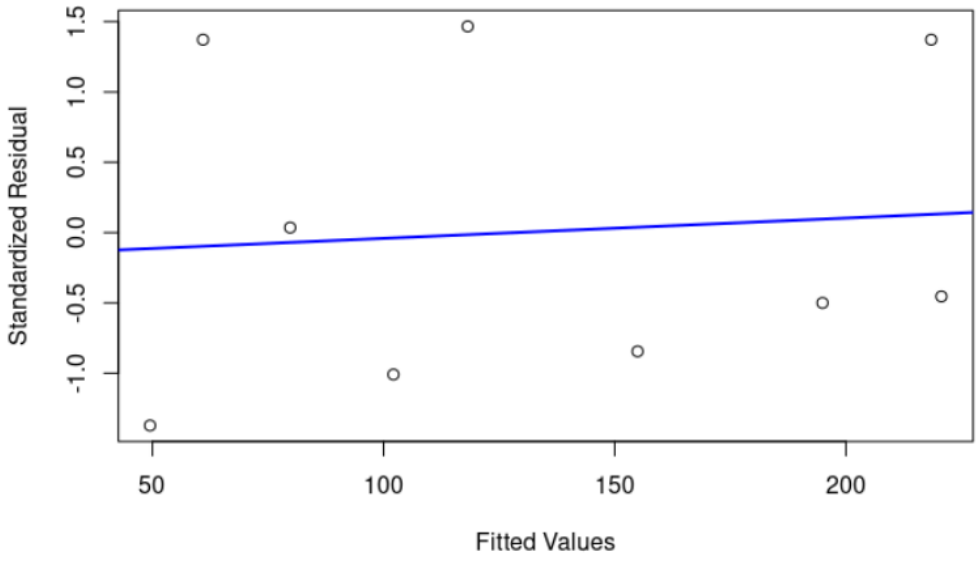

Fig. 2 Residual plot between residuals and fitted values

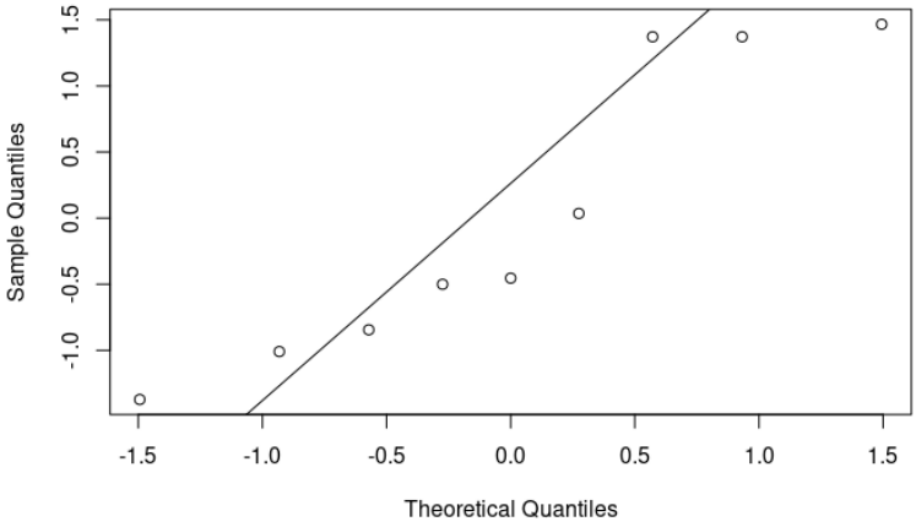

Fig. 3 Q-Q plots of residuals in the full model

Moreover, as the sample size is only 9 which is not large enough, leading the variance inflation factor (VIF) to be abnormally large, which may lead to the multicollinearity problem. In order to handle the problem, more data needs to be collected, or it is also feasible that at least one of the predictors can be removed from the model. Nevertheless, there is no obvious evidence showing which specific predictor should be removed, and the ability of the prediction of the model will be reduced if a wrong predictor is removed.

4. Summary

In conclusion, this paper investigates a method to forecast the number of future subscribers of Netflix based on constructing a regression model. Specifically, constructing a multiple linear regression model with the help of R language to help predict the number of future subscribers of Netflix. According to the analysis, COVID-19 is the main reason for the change in subscribers from 2020 to 2022. More precisely, several sources support that income, films and television production, market competition, and copyright, all caused by COVID-19, are the main factors affecting subscribers. Nevertheless, there still have some limitations and shortcomings for this study. In order to improve the whole prediction model, a more extensive sample size data is indispensable. In the future, there is still a need to find and improve the model to satisfy more assumptions and influential variables. Overall, these results offer a guideline for predicting subscribers of Netflix in future years by constructing a multiple linear regression model.

References

[1] Kumar J., Gupta A., Dixit, S. Netflix: SVoD entertainment of next gen. Emerald Emerging Markets Case Studies, 2020, Vol. 10 No. 3.

[2] Natalie Sherman, Streaming Netflix's subscribers and profits plummet for the first time in a decade Five reasons behind. April 21, 2022. Retrieved from: https://www.bbc.com/zhongwen/simp/business-61161567

[3] kkvb, R Language-Case Analysis of Multiple Linear Regression. November 30, 2017. Retrieved on July 25, 2022. Retrieved from: https://zhuanlan.zhihu.com/p/31037989

[4] Similarweb, netflix.com. Retrieved on July, 2022. Retrieved from: https://www.similarweb.com/website/netflix.com/

[5] Daniel Ruby, Netflix Subscribers 2022 — How Many Subscribers Does Netflix Have. May 1, 2022. Retrieved on July 24, 2022. Retrieved from: https://www.demandsage.com/netflix-subscribers/

[6] Rahman K.T. and Arif M. Z. U. Impacts of Binge-Watching on Netflix during the COVID-19 pandemic, South Asian Journal of Marketing, 2021, Vol. 2 No. 1, pp. 97-112.

[7] Antipova A. Analysis of the COVID-19 impacts on employment and unemployment across the multi-dimensional social disadvantaged areas. Social sciences & humanities open, 2021, 4(1): 100224.

[8] Chen X, Qiu Y, Shi W, et al. Key links in network interactions: Assessing route-specific travel restrictions in China during the Covid-19 pandemic. China Economic Review, 2022, 73: 101800.

[9] Liang X, Rozelle S, Yi H. The impact of COVID-19 on employment and income of vocational graduates in China: Evidence from surveys in January and July 2020. China Economic Review, 2022: 101832.

[10] Hitesh Bhasin, Netflix Competitors Analysis. May 31, 2020. Retrieved from: https://www.marketing91.com/netflix-competitor-analysis/#:~:text=%20Netflix%20Competitors%20Analysis%20%201%20Amazon%20Prime,service%20provider%20with%20headquarters%20in%20El...%20More%20

[11] Jillian Bell, Netflix bound by copyright rules, contracts with studios. January 15, 2016. Retrieved from: https://www.cbc.ca/news/business/canadian-netflix-copyright-law-1.3406074