2020 IEEE 20th International Conference on Software Quality, Reliability and Security Companion (QRS-C)

Best Seller Rank (BSR) to Sales: An empirical look at Amazon.com

Amit Sharma
Industrial and Systems Engineering
San Jose State University
San Jose, USA
amit.sharma01@sjsu.edu

Hongrui Liu
Industrial and Systems Engineering
San Jose State University
San Jose, USA
hongrui.liu@sjsu.edu

Hongwei Liu
Scientific Research Department
Tianjin Foreign Studies University
Tianjin, China
liuhongwei@tjfsu.edu.cn

***Abstract*— The market size of e-commerce continued to grow in recent years. Amazon, a leading technology company that provides reliable online marketplaces had attracted more than two million active sellers with over 600 million active products listed on it in a snapshot in 2019. The Amazon Best Sellers Rank (BSR) is an important indicator of how well the listed product is sold and it is a common interest to all sellers. This research investigates the algorithm for setting BSR on Amazon by developing a python-based web scrapper to collect hourly data on Amazon Web Services (AWS) server. A BSR to sales mapping study is performed using a machine learning regression algorithm. The study would help Amazon sellers to understand the mechanism of BSR and its impact on future sales.**

Keywords*— *Amazon, Best Seller Rank, e-commerce, online marketplace, forecasting

I. Introduction

Online buying and selling have become increasingly prevalent today because of the numerous advantages offered by it [1]. Not only the buyers can enjoy the convenience of buying anytime, anywhere through the online businesses, but also the online channel offers great opportunities to the sellers to sell their products directly to customers all over the world without the need for a physical store [2, 3, 4, 5]. The advantages are even more obvious in the current pandemic situation, when it is not convenient to shop at physical stores. At the same time, the online business has increased competition in nearly all markets [6]. It is critical important for the online sellers to understand the impact factors to sales so they can set strategies to attract more customers to maximize their sales. Many traditional forecast models and machine learning models have been utilized to help predict sales [7, 8]. However, forecasting sales for online products is very complex as there are so many factors that may affect the sales, such as price, number of reviews, and competitors' behaviors, and no single model has performed well across all the situations.

In this paper, we will conduct an empirical study on Amazon sellers to gain some insights on online sales with the purpose to help the sellers to identify the right items to scout when it comes to selling online. Amazon, a leading technology company that provides reliable online platform-based markets is attracting more and more customers [9, 10]. According to Statista, Amazon was responsible for 45% of US ecommerce spending in 2019 [11] and 9 out of 10 consumers price check a product on Amazon according to Consumer News and Business Channel (CNBC) [12]. There are currently more than 2.5 million sellers at Amazon and it is the desire of the sellers to promote their products appropriately to attract attentions of more customers to increase sales. The Amazon Best Sellers Rank (BSR) is a score that Amazon assigns to products based on its sales volume and historical sales data and it is updated hourly [13]. It essentially tells you how your product ranks compared to other products within the same category. The BSR also affects how a product will be displayed upon search queries and thus the future sales. Therefore, the BSR is a common interest to all sellers on Amazon marketplaces. It is a well know that Amazon BSR is calculated based on historical sales. However, the details of the algorithm that Amazon uses to assign BSR for each specific product and the impact of BSR to future sales remain little known to the sellers and researchers.

In this paper, we will conduct an empirical study to investigate how a BSR is assigned for a product. Moreover, we work with a seller on Amazon marketplace to collect the actual sales data so we can perform a predictive analysis on sales from the BSR.The research work includes three parts: (1) develop a python-based web scrapper to collect hourly data on Amazon Web Services (AWS) server, (2) analyze BSR fluctuations and get an insight on the possible impact factors such as sales, pricing and number of reviews, and (3) develop a machine learning regression model to study the mapping of the BSR to future sales. The objective of the research is to provide the Amazon sellers with some guidelines on how a BSR is assigned and a predictive model for sales based on BSR.

The paper is organized as follows. In Section II, the data information and how they are collected are introduced. In Section III, data analysis and investigation findings on BSR and the predive analysis on sales are illustrated. Conclusions and future work are discussed in Section IV.

978-1-7281-8915-4/20/$31.00 ©2020 IEEE
DOI 10.1109/QRS-C51114.2020.00104

II. DATA COLLECTION

The most important variable in this analysis is the BSR. The BSR information can be found on amazon product page. An example is shown in Fig. 1.

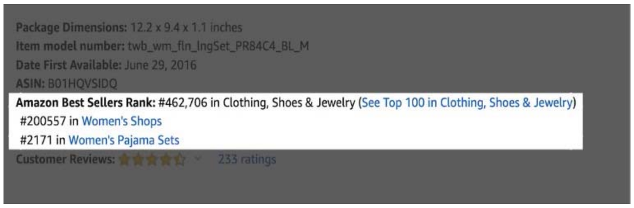

Fig. 1. Product details from Amazon.com

As it can be seen from the example in Fig. 1, a product has one main category (Clothing, Shoes and Jewelry) and multiple subcategories. The product is ranked differently under different categories. It is ranked at #462,706 in the main category and #200,557 and #2171 under the two subcategories respectively. The category "Clothing, Shoes and Jewelry" is selected for the study as the actual sales data we can collect falls into this category. We used only the main product category as it demonstrated the highest fluctuations. There are currently 85,864,736 products listed under Clothing, Shoes and Jewelry section, which means the highest BSR a product can have is 85,864,736 and the lowest is 1 under this category.

The data we collected includes the BSR, the price, number of reviews, number of questions asked for 10 different ASINs from Clothing, Shoes & Jewelry category. ASINs, also called Amazon Standard Identification Numbers are unique blocks of 10 letters and/or numbers that identify items sold on Amazon. Data for the 10 products was collected every hour for a period of 3 months. To collect the data a web scrapper made on python (Scrapy) was scheduled on amazon AWS (EC2) server which would collect the data every hour and store it in a MySQL database. A data set of around 10,000 data points was created with an expectation of yielding a good model for the regression analysis.

III. DATA ANALYSIS AND RESULTS

A. Analysis on BSR

The collected data was plotted in Fig. 2 to Fig. 7 to get a visualization of the correlations between BSR and other factors. The time series sales vs. BSR is shown in Fig.2, the hourly sales vs. BSR is shown in Fig. 3 and Fig. 4 is a closer look of Fig. 3, the number of reviews vs. BSR is shown Fig. 5, the fluctuation of BSR of a new product is shown in Fig. 6, and the BSR of a product with a constant review is shown in Fig. 7.

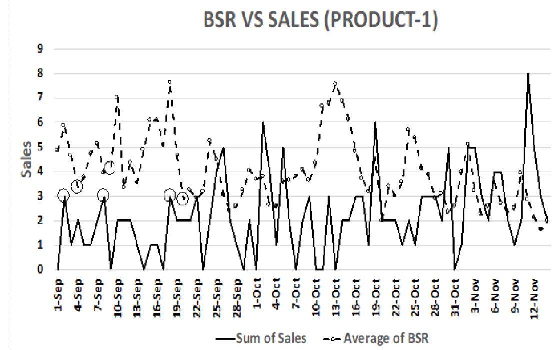

Fig. 2. Sales vs BSR over time

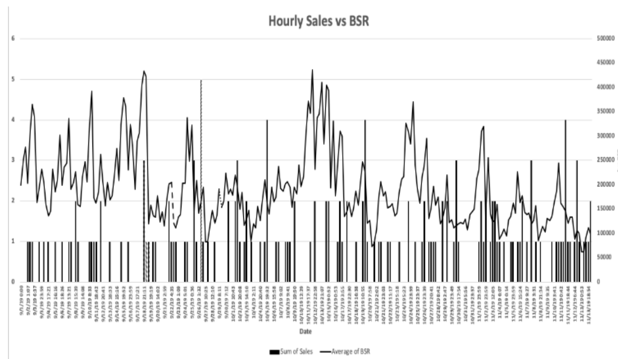

Fig. 3. Hourly Sales vs BSR on time

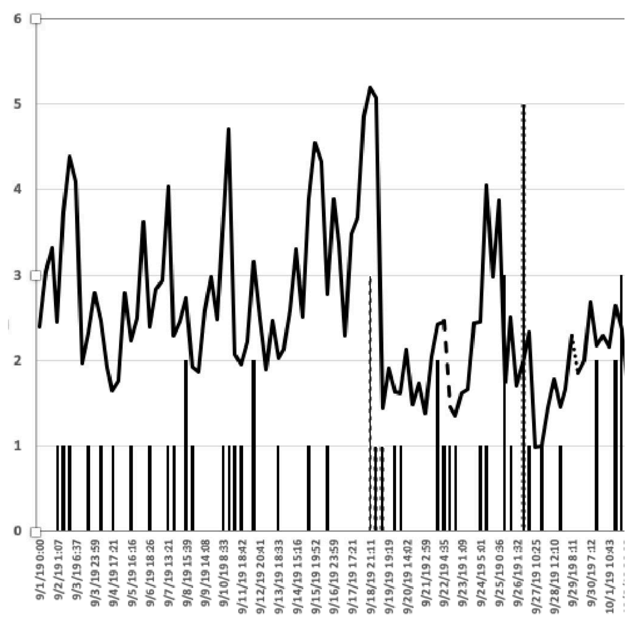

Fig. 4. Closer view of Fig. 3

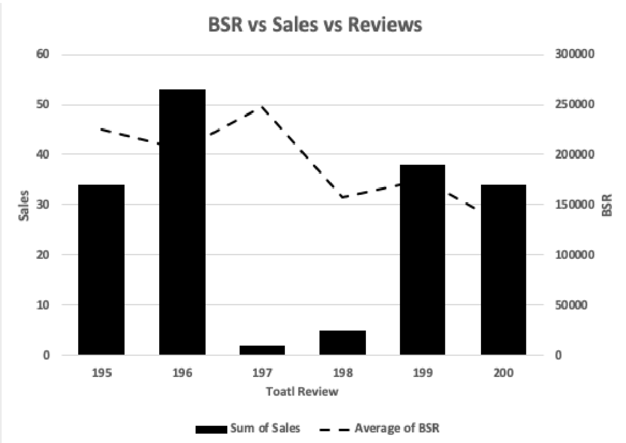

Fig. 5. BSR vs Sales vs Review on time

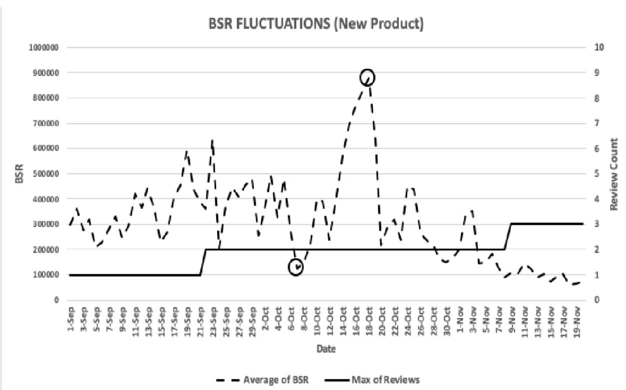

Fig. 6. BSR fluctuations of a new product

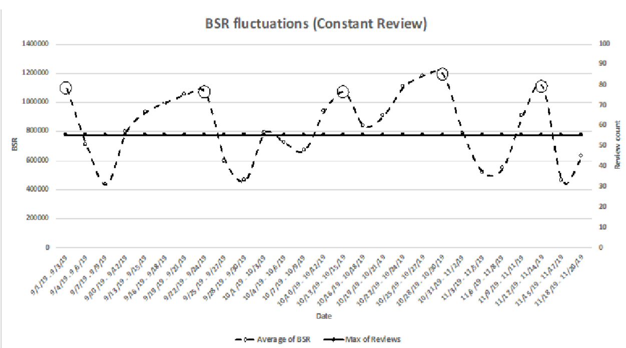

Fig. 7. BSR fluctuations for a constant review product

We observed the following scenarios from our data analysis and from the plots in Fig. 2 to Fig 7.

- BSR is updated only after the payment for the product sold on Amazon is cleared by the bank, i.e. within 3 days as shown in Fig. 2.
- BSR weights more heavily on recent sales than on past sales. This will be explained in a scenario later in this section.
- BSR tends to be order dependent but it is independent of the number of units ordered. As observed from Fig. 3 and Fig. 4, the change in BSR that is represented by the dashed line is associated with a sum of three orders: one order of 3 unit and two orders of 1 unit each; the change in BSR that is represented by the dotted line is associated with a single order of 5 units. The total sales of both scenarios are the same 5 units, whereas their impact to the BSR are different. The change of BSR that is associated with 3 orders is bigger than the one that is associated with one order.
- Products with more reviews tend to have a better BSR as shown in Fig. 5, a steady decrease in BSR is observed as the review count increases.
- New items do not have a BSR; when a new item starts to have a BSR, it tends to get a larger BSR swings even in the range on 10,000 to a Million in a week. An example of such a case is shown in Fig. 6.
- One sale of a very popular product may not influence its BSR much at all, but one sale of a lower volume product may significantly improve that product's BSR. This will be explained in a scenario later in this section.
- Children of same family are bundled with the parent to give a single BSR for all, e.g. different colors of a dress sold on Amazon share a single BSR.
- Two products of the same category can never have a same BSR allocated to them.
- The max BSR of products with constant number of reviews tend to be the same as the max BSR in the previous periods or worse when there are no new sales. An example is shown in Fig. 7.

It is easy to see that BSR reflects how well a product is sold in the past few months and also in the past few hours. However, one cannot tell what the current BSR indicates on the sales exactly unless a detailed BSR data set is available.

A few scenarios are used to illustrate the fluctuations of BSR as follows.

1) The BSR of a product can change siginificantly by its recent sales – the sales and BSR of products A and B in Clothing, Shoes and Jewelry department are shown in TABLE I and TABLE II.

TABLE I. COMAPRISON FOR FLUCTUATIONS IN BSR WITH NO SALES FOR PRODUCT B FOR 30 DAYS PERIOD

	Product A	**Product B**
Sales Per Day	70	0
Sales last 30 days	2100	0
Total all time sales	12000	1
Current BSR	750	2,050,000

TABLE II. COMAPRISON FOR FLUCTUATIONS IN BSR WITH PRODUCT B SALES GREATER THAN PRODUCT A FOR 30 DAYS PERIOD

	Product A	Product B
Sales Per Day	70	90
Sales last 30 days	2100	250
Total all time sales	12000	251
Current BSR	750	780

It is observed from TABLE II that when product B starts to pick up on sales recently (because of an advertisement) compares to the data in TABLE I, the BSR for product B improved significantly from 2,050,000 to 780. Products A and B now have similar BSRs even though Product B has only 250 units sold whereas Product has sold 2100 units which is nearly 10 times more than B. This shows that the BSR is more dependent on the recent sales than historical sales.

2) Products may have similar number of sales and still have very different BSRs – the comparison of the sales and BSR of products C and D in Clothing, Shoes and Jewelry department are shown in TABLE III and TABLE IV.

Products C and D in TABLE III have 4 units difference in sales but have 2500 difference in BSRs. When 4 units of product C were sold in the past 6 hours as shown in TABLE IV, the BSR of product C improved from the previous 15000 to 7000. This again verifies that a small change in sale can change BSR significantly as there are large number of products listed on amazon. While in TABLE IV, even though products C and D have the same number of sales (300 units) in a month, they have very different BSRs.

TABLE III. COMAPRISON FOR FLUCTUATIONS IN BSR WITH PRODUCT C SALES THE SAME AS PRODUCT D FOR 6 HOURS

	Product C	Product D
Total monthly sales	245	249
BSR	15000	12500
Sales in last 6 hrs.	0	0

TABLE IV. COMAPRISON FOR FLUCTUATIONS IN BSR WITH PRODUCT C SALES GREATER THAN PRODUCT D FOR 6 HOURS.

	Product C	Product D
Total monthly sales	300	300
BSR	7000	11000
Sales in last 6 hrs.	4	1

B. Predicting Sales Using Regression

In this section, we predict sales from the BSR data using a machine learning linear regression algorithm. The machine learning algorithms deal with building models that can learn from data, and be revised automatically according to data patterns learned. In addition, they are more flexible by accounting for other influencing parameters therefore able to have a better fit to the data [14, 15, 16]. In this paper, we only use one factor, the BSR, to predict sales and the linear regression is a good fit for our analysis. This method is a linear approach to modeling the relationship between dependent variable and independent variable and is described in equation (1).

$$Y = aX + b, \quad (1)$$

where Y is a dependent variable or predicted variable, X is an independent variable, a is regression coefficient and b is the error term or noise. In our case, the dependent variable or predicted variable Y represents the sales and the independent variable X represents the BSR. The coefficient will be obtained from the machine learning linear regression algorithm, which is a standard function in R.

To begin with the regression analysis, the raw data was first grouped into a period of 3 days to. The reason of doing so is that an order is considered as one data point when it is placed on Amazon, whereas Amazon updates BSR only after bank clears the transaction which takes up to 3 days. The data was then divided into training and testing data sets. We used 80% of the 3 months data to train a linear regression mode and the remaining 20% to test model accuracy. To increase the model accuracy and minimize the residual error, different transform functions were applied to the predicted variable. Fig. 8 shows the sum of sales, Fig. 9 shows the square root of sales and Fig. 10 shows the logarithm of sales.

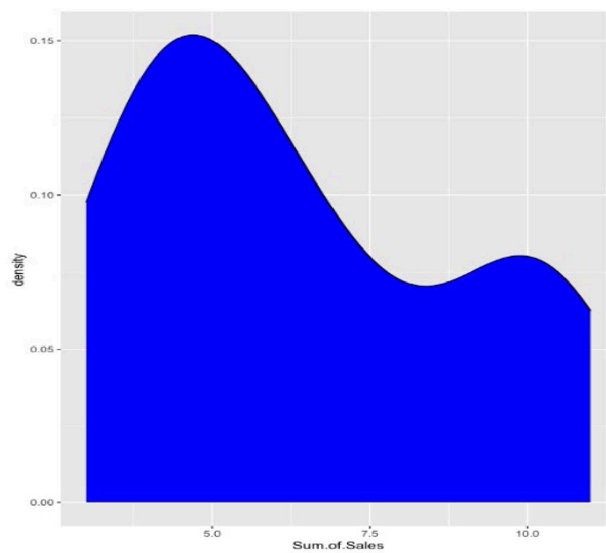

Fig. 8. Sales density – sum of sales

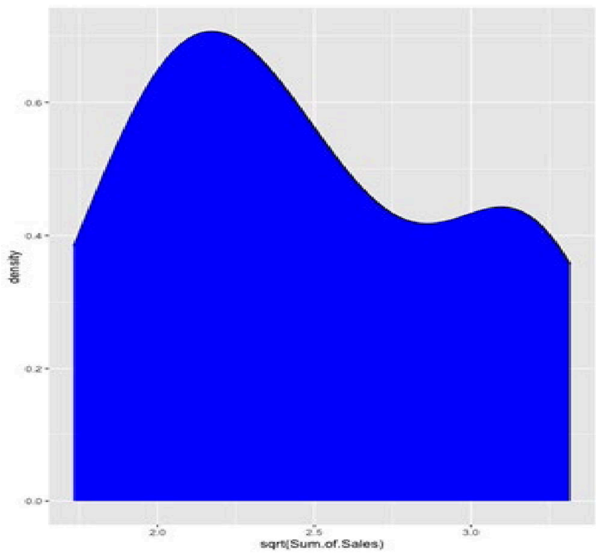

Fig. 9. Sales density - square root of sales

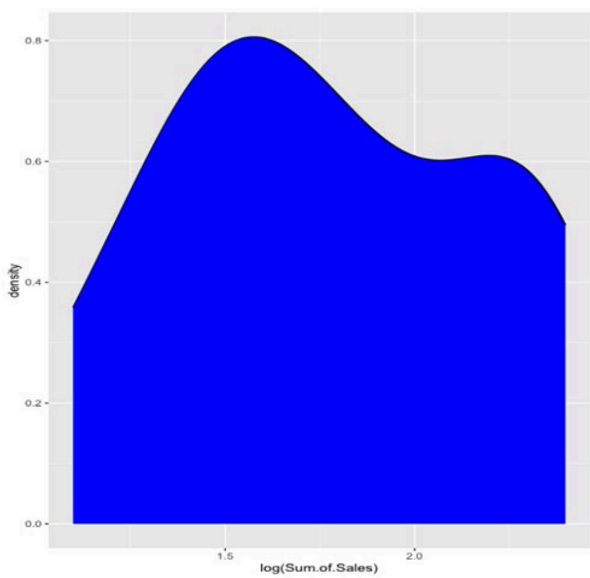

Fig. 10. Sales density – log transformation of sales

The logarithm of sales is used in the linear regression algorithm to train the model as it is the closest approximation to a normal distribution compare to the other two. After this, a regression model to map BSR to Sales was performed. The results are shown below:

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.50860 -0.12182  0.04371  0.16284  0.40846

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.825e+00  1.739e-01  16.246 2.30e-11 ***
Min.of.BSR  -8.029e-06  1.331e-06  -6.032 1.74e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.224 on 16 degrees of freedom
Multiple R-squared:  0.6946,    Adjusted R-squared:  0.6755
F-statistic: 36.39 on 1 and 16 DF,  p-value: 1.743e-05
```

Fig. 11. Regression model results

As we can see from the Fig. 11, R-squared value is 69.46%. This indicates that the model explains 69.46% of variability in the predicted/response data around its mean. It can also be seen in the Residual vs Fitted plot in Fig. 12 that the points are scattered around the line showing the variability between actual and predicted sales data. Normal Q-Q plot in Fig. 13 shows that the data is normal to some extent but has some skewness in it, this generally happens when we have a small data set. The Scale-Location plot in Fig. 14 shows to some extent that the variance is equally divided over the data set i.e. homoscedasticity holds true. The Residual vs Leverage plot in Fig. 15 shows that all the cases are well inside Cook’s distance lines.

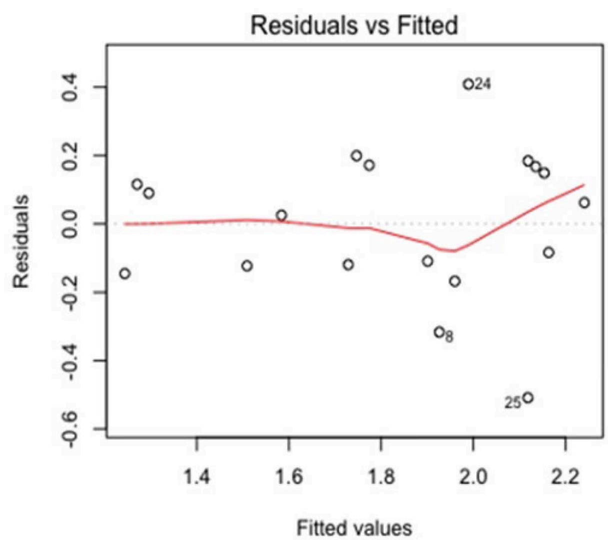

Fig. 12. Regression model results: residual vs. fitted

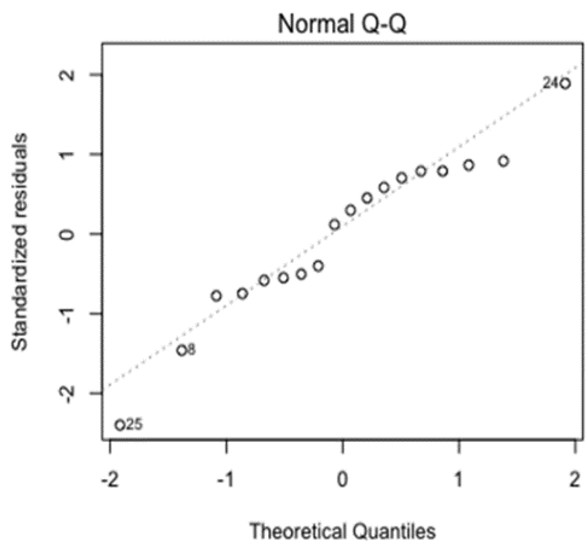

Fig. 13. Regression model results: normal Q-Q plot

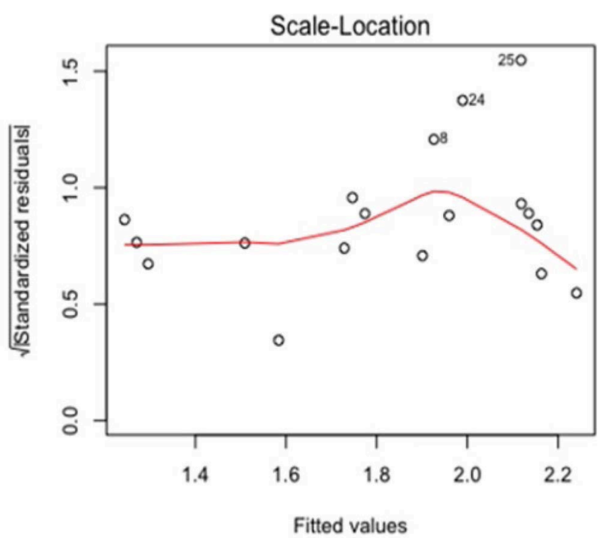

Fig. 14. Regression model results: scale-location

IV. Conclusion and Future Work

In the paper, we have conducted an empirical study on how a BSR is assigned on Amazon. The findings can provide some insights to the Amazon sellers on how to improve their BSR. Moreover, a regression model is developed to predict sales from the BSR data. This model will be helpful for sellers to get a rough estimated sale of their products based on their current BSR so sellers can manage their inventory or to scout new products accordingly. This model can be improved in the future to use a larger dataset mapping a huge number of BSR to its sales for a category. Accurate sales prediction is a challenging task in ecommerce, and it depends on many other factors other than the BSR [17]. A more comprehensive machine learning model or artificial intelligent model such Long short-term memory (LSTM) can used to incorporate more factors to predict sales with a better accuracy.

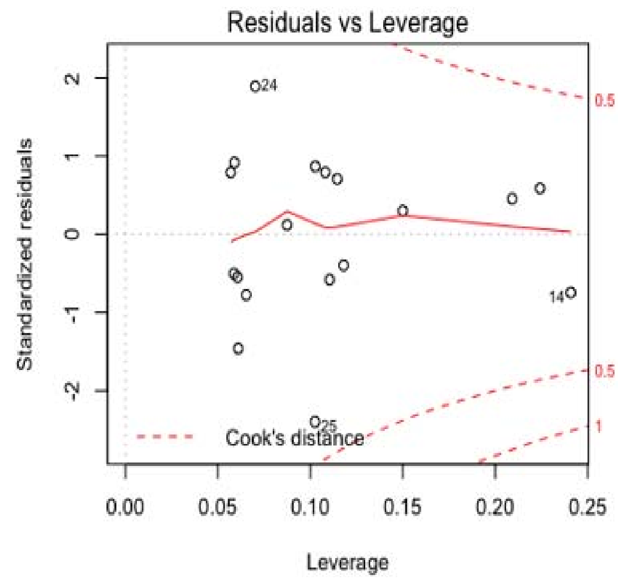

Fig. 15. Regression model results: residual vs. leverage

Acknowledgment

The work is sponsored in part by the National Natural Science Foundation of China (NSFC) (61877044).

References

[1] G. D. Gregory, L. V. Ngo, and M. Karavdic, “Developing e-commerce marketing capabilities and efficiencies for enhanced performance in business-to-business export ventures,” Industrial Marketing Management, Vol. 78, 2019, pp. 146-157.

[2] https://www.forbes.com/sites/jefffromm/2019/01/04/marketing-convenience-to-the-modern-consumer/#211e2ca8127f

[3] https://assets.kpmg/content/dam/kpmg/xx/pdf/2017/01/the-truth-about-online-consumers.pdf

[4] F. Rahimzadeh and M. Heydari, “A review of ecommerce competitive advantages in international trade,” Uct Journal Of Management And Accounting Studies 5(4), 2017, pp. 79-85.

[5] S. Sabou, B. Avram-Pop and L. A. Zima, “The impact of the problems faced by online customers on ecommerce,” Studia Universitatis Babes-Bolyai Oeconomica, 2016, Vol. 62: Issue 2.

[6] R. Agnihotri, R. Dingus, M.Y. Hu and M.T. Krush, “Social media: Influencing customer satisfaction in b2b sales,” Industrial Marketing Management 53, 2016, pp. 172– 180.

[7] E.T. Bradlow, M. Gangwar, P. Kopalle and S. Voleti, “The role of big data and predictive analytics in retailing,” Journal of Retailing 93(1), 2017, pp. 79–95.

[8] R. Valero-Fernandez, D. J. Collins, K. P. Lam, C. Rigby and J. Bailey, "Towards accurate predictions of customer purchasing patterns," 2017 IEEE International Conference on Computer and Information Technology (CIT), Helsinki, 2017, pp. 157-161, doi: 10.1109/CIT.2017.58.

[9] F. Zhu and Q. Liu, “Competing with complementors: An empirical look at Amazon.com,” Strategic Management Journal 39, no. 10, October 2018, pp. 2618–2642.

[10] P. D. Culpepper and K. Thelen, “Are we all Amazon primed? consumers and the politics of platform Power,” Comparative Political Studies, 2020, Vol. 53, issue 2, pp. 288-318.

[11] https://www.repricerexpress.com/amazon-statistics/

[12] https://www.cnbc.com/2016/09/27/amazon-is-the-first-place-most-online-shoppers-visit.html

[13] https://www.amazon.com/gp/help/customer/display.html?nodeId=GGGMZK378RQPATDJ

[14] A. Aluko & H. Liu, “A comparative study of traditional forecasting methodologies vs. machine learning algorithms”. Proceedings of the Institute of Industrial and Systems Engineers (IISE), Orlando, FL, May 2019

[15] J.V. Hansen and R.D. Nelson, “Neural networks and traditional time series methods: a synergistic combination in state economic forecasts,” IEEE Transactions on Neural Networks, Vol. 8, No. 4, July 1997

[16] R. Adhikari and R. K. Agrawal, “An introductory study on time series modeling and forecasting,” Ithaca: Cornell University Library, arXiv.org, 2013. Retrieved from https://arxiv.org/pdf/1302.6613.pdf

[17] B. Singh, P. Kumar, N. Sharma and K. P. Sharma, "Sales forecast for amazon sales with time series modeling," 2020 First International Conference on Power, Control and Computing Technologies (ICPC2T), Raipur, India, 2020, pp. 38-43