

**BEST-SELLER PRICING ON AMAZON.COM: A PANEL VECTOR
AUTOREGRESSIVE APPROACH**

by
İREM BİLEN

Submitted to the Sabancı Graduate Business School
in partial fulfilment of
the requirements for the degree of
Master of Science in Business Analytics

Sabancı University
September 2020

BEST-SELLER PRICING ON AMAZON.COM: A PANEL VECTOR
AUTOREGRESSIVE APPROACH

Approved by:

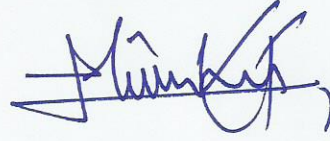
Prof. Abdullah Daşcı
(Thesis Supervisor)



Assist. Prof. Burak Gökgür



Assoc. Prof. Mümtaz Karataş



Date of Approval: September 1, 2020

İREM BİLEN 2020 ©

All Rights Reserved

ABSTRACT

BEST-SELLER PRICING ON AMAZON.COM: A PANEL VECTOR AUTOREGRESSIVE APPROACH

İREM BİLEN

BUSINESS ANALYTICS M.Sc. THESIS, SEPTEMBER 2020

Thesis Supervisor: Prof. Abdullah Daşcı

Keywords: Best-seller, Panel Vector Autoregression, Loss Leader, Pricing, PVAR

Amazon has created an ideal stop for one-stop shopping with its broad assortment of products sold by Amazon itself and other retailers. Its huge selection of products, big data-driven recommendation system, nice user interface, and many other factors entice consumers to shop there, and spend hours to discover items. A customer who visits Amazon.com is likely to buy unplanned items website recommends or that fulfill the condition for free delivery. High cross-selling potential of Amazon, and consumers' high impulse buying potential facilitate using loss leader strategy. It is known that Amazon.com sells best-seller books at below cost, but there is limited understanding of the factors that influence pricing decisions of this company. In this study, we observe how key market characteristics impact discounting decisions of Amazon and how all these variables affect each other in this marketplace. We conduct Panel Vector Autoregressive modelling on a panel time series dataset with 15500 observations on 5 endogenous variables (discount, sales rank, list price, customer review and number of sellers) and 1 exogenous variable (physical format) of 500 books for 31 days. By using Panel Vector Autoregressive modelling, we also take the impact of previous days' observations into consideration in explaining the relationship. Our results suggest that on Amazon.com discounts are deeper for books with better sales ranks, higher list prices, higher customer reviews, or lower number of sellers. We also demonstrate the effects of these variables to each other. Our study is among the few that observe dynamics of Amazon marketplace.

ÖZET

AMAZON.COM'DA LİSTE BAŞI KİTAP FİYATLANDIRMASI: BİR PANEL VEKTÖR OTOREGRESİF MODELLEME YAKLAŞIMI

İREM BİLEN

İŞ ANALİTİĞİ YÜKSEK LİSANS TEZİ, EYLÜL 2020

Tez Danışmanı: Prof. Abdullah Daşcı

Anahtar Kelimeler: Liste başı kitap, Panel Vektör Otoresresyon, Zararına satış,
Fiyatlandırma, PVAR

Amazon, hem kendisinin hem de diğer mağazaların satışını yaptığı geniş ürün çeşidiyle insanların tek bir yerden alışveriş yapabileceği ideal bir site yarattı. Onun geniş ürün seçenekleri, büyük veri tabanlı tavsiye sistemi, hoş kullanıcı arayüzü vb. faktörler müşterileri oradan alışveriş yapmaya, ürünleri keşfetmek için saatler harcamaya çekiyor. Amazon.com'u ziyaret eden bir müşteri; sitenin önerdiği veya bedava kargo koşulunu sağlayan almayı planlamadığı ürünleri de satın almaya meyilli oluyor. Amazon'un yüksek çapraz satış potansiyeli ve müşterilerinin dürtüsel satın alma potansiyeli Amazon'un fiyatlandırmada zararına satış stratejisini kullanmasına zemin hazırlıyor. Amazon'un liste başı kitapları zararına sattığı biliniyor ancak stratejisini belirlerken hangi faktörlerin etkili olduğu hakkında bilgi kısıtlı. Bu çalışmada anahtar pazar özelliklerinin Amazon'daki indirimlere ve birbirlerine etkisini gözlemliyoruz. Panel vektör otoresresif modelleme ile 500 kitabın 31 günlük 15500 gözlemine içeren beş endojen değişkeni (indirim, satış sıralaması, liste fiyatı, müşteri kritiği, ve mağaza sayısı) ve bir dışsal değişkeni (format) kapsayan panel zaman serisi veri setini inceliyoruz. Panel vektör otoresresif modelleme ile değişkenlerin önceki günlerde gözlenen değerlerinin diğer değişkenlere olan etkilerini de göz önünde bulunduruyoruz. Sonuçlarımız Amazon.com'da daha yüksek indirimlerin daha iyi satış sıralaması, daha yüksek liste fiyatı, daha yüksek müşteri kritiği olan ya da daha az sayıda mağaza tarafından satılan kitaplara olduğunu gösteriyor. Ayrıca, bu değişkenlerin birbirlerine olan etkilerini de gösterdik. Bizim çalışmamız Amazon pazarının dinamiklerini gözlemleyen az çalışmadan bir tanesi.

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisor Prof. Abdullah Daşcı. I am very lucky to have the chance to work under the supervision of him. His academic wisdom, continuous guidance and endless support is invaluable for me. I would also like to thank Prof. Cenk Koçaş for his advice and providing me with the comprehensive dataset I used in this research.

Furthermore, I would like to thank all the professors I met during my time in Sabancı University for all their contributions.

I would also like to express my sincere thanks to Prof. Mehmet Gençer for his endless support, and for providing me with the foundation and love of business analytics in my undergraduate years.

Finally, I would like to express my deepest gratitude to my family for everything.

*Dedicated
to my beloved family*

TABLE OF CONTENTS

LIST OF TABLES	x
LIST OF FIGURES	xii
LIST OF ABBREVIATIONS	xiii
1. INTRODUCTION	1
2. LITERATURE REVIEW	3
3. DATA DESCRIPTION	9
3.1. Dataset Description	9
3.2. Problems with the Data	10
3.3. Data Cleaning	11
3.3.1. Creating the Balanced Panel Data	14
4. METHODOLOGY	18
4.1. Modelling Approach	18
4.2. Unit Root Tests.....	19
4.3. Cointegration Test	19
4.4. Granger Causality Test	20
4.5. Vector Autoregressive Modelling.....	20
4.6. Impulse Response Functions	22
5. RESULTS	23
5.1. Data Analysis	23
5.1.1. Unit Root Tests	23
5.1.2. Cointegration Test	24
5.1.3. Granger Causality Test	24
5.1.4. Vector Autoregression Model.....	25
5.1.4.1. Model Selection.....	25
5.1.4.2. Results of VAR Model.....	26

5.1.5.	Impulse Response Functions	27
5.2.	Analysis with Panel Data Adjustment	33
5.2.1.	Unit Root Tests	33
5.2.2.	Cointegration Test	33
5.2.3.	Granger Causality	34
5.2.4.	Panel Vector Autoregression Model	35
5.2.4.1.	Model Selection.....	35
5.2.4.2.	Results of the PVAR Model	35
5.2.5.	Results of the Generalized Impulse Response Functions	36
5.3.	Analysis on Panel Data with Modified Variables	42
5.3.1.	Unit Root Tests	42
5.3.2.	Cointegration Test	43
5.3.3.	Granger Causality Test	43
5.3.4.	Panel Vector Autoregression Results	44
5.3.4.1.	Model Selection.....	44
5.3.4.2.	Results of PVAR Model with Modified Panel Data ..	45
5.3.5.	Results of Generalized Impulse Response Functions	46
5.4.	Findings and Discussion	52
6.	CONCLUSION	54
	BIBLIOGRAPHY.....	57

LIST OF TABLES

Table 3.1. Descriptive Statistics of Numerical Variables in Raw Data	9
Table 3.2. Descriptive Statistics of the Numerical Variables of the Data with No Missing Values	12
Table 3.3. Descriptive Statistics of Physical Format and Our Classification	13
Table 3.4. Clustering of Number of Sellers	15
Table 3.5. Descriptive Statistics of ISBN	16
Table 3.6. Descriptive Statistics of Numerical Variables in Balanced Panel Data	17
Table 3.7. Descriptive Statistics of Categorical Variables in Balanced Panel Data	17
Table 5.1. Unit Root Test Results	24
Table 5.2. VAR Stability Condition Check	26
Table 5.3. Same Day and Cumulative Effects on Variables (from GIRF estimates)	28
Table 5.4. Same Day and Cumulative Effects on Variables (from GIRF estimates) cont.	29
Table 5.5. Same Day and Cumulative Effects on Variables (from GIRF estimates) cont.	30
Table 5.6. Unit Root Test Results on Panel Dataset	33
Table 5.7. Cointegration Test Results on Panel Dataset	34
Table 5.8. Same Day and Cumulative Effects on Panel Time Series Vari- ables (from GIRF estimates)	37
Table 5.9. Same Day and Cumulative Effects on Panel Time Series Vari- ables (from GIRF estimates)-cont.....	38
Table 5.10. Same Day and Cumulative Effects on Panel Time Series Vari- ables (from GIRF estimates)-cont.....	39
Table 5.11. Unit Root Test Result of Clustered Number of Sellers Variable	42
Table 5.12. Panel Time Series Cointegration Test Results	43

Table 5.13. Same Day and Cumulative Effects on Panel Time Series Variables (from GIRF estimates)	47
Table 5.14. Same Day and Cumulative Effects on Panel Time Series Variables (from GIRF estimates)-cont.....	48
Table 5.15. Same Day and Cumulative Effects on Panel Time Series Variables (from GIRF estimates)-cont.....	49
Table 5.16. Accuracy Metrics of Vector Autoregressive Models	52

LIST OF FIGURES

Figure 3.1. Elbow method to choose number of clusters for retailers variable	14
Figure 5.1. Johansen Cointegration Test	24
Figure 5.2. Model Selection for VAR model	25
Figure 5.3. Stability of VAR Model	26
Figure 5.4. Discount GIRF	30
Figure 5.5. List Price GIRF	31
Figure 5.6. Customer Review GIRF	31
Figure 5.7. Number of Sellers GIRF	32
Figure 5.8. Sales Rank GIRF	32
Figure 5.9. Model Selection for PVAR model	35
Figure 5.10. Stability of PVAR Model	36
Figure 5.11. Discount Panel GIRF	39
Figure 5.12. List Price Panel GIRF	40
Figure 5.13. Customer Review Panel GIRF	40
Figure 5.14. Number of Sellers Panel GIRF	41
Figure 5.15. Sales Rank Panel GIRF	41
Figure 5.16. Optimal Lag Length Selection	45
Figure 5.17. Stationarity and Stability of PVAR Model	46
Figure 5.18. GIRF Graph of Discount	49
Figure 5.19. GIRF Graph of List Price	50
Figure 5.20. GIRF Graph of Customer Review	50
Figure 5.21. GIRF Graph of Number of Sellers	51
Figure 5.22. GIRF Graph of Sales Rank	51
Figure 6.1. Summary Diagram of the Findings	56

LIST OF ABBREVIATIONS

ADF Augmented Dickey-Fuller	19
AIC Akaike Information Criterion	25
GIRF Generalized Impulse Response Functions.....	22
IRF Impulse Response Functions	22
ISBN International Standard Book Number	10
KPSS Kwiatkowski-Phillips-Schmidt-Shin	19
PP Phillips-Perron	19
PVAR Panel Vector Autoregression	21
SC Schwarz Information Criterion	25
VAR Vector Autoregression	19

1. INTRODUCTION

Amazon has been the pioneer in the online market and has been considered Walmart of the web. Its online sales were more than seven times bigger than online sales of Walmart (Kotler & Armstrong, 2016). The inventor of the big data-driven recommendation system has always been successful in attracting consumers to its website with a personalized website for each customer (Kotler & Armstrong, 2016). Amazon, by surpassing Walmart, is now considered as the largest retailer in the world according to Forbes' Global 2000, a list that measures the biggest public companies by a score consisting of revenues, profits, assets, and market value of these companies (Debter, L., 2019).

Amazon.com is a unique marketplace where both Amazon and third-party sellers sell the same products. According to Zhu & Liu, 2018, Amazon competes with the sellers that sell in this marketplace and cause them to leave the market. In addition, Amazon follows other competing websites with Amazon's Competitive Intelligence arm by regularly purchasing merchandise in order to analyze and compare speed of delivery, service quality, and assortment (Kotler & Armstrong, 2016).

Although the competition among online retailers are more severe than retailers in offline markets, the competition is not perfect competition. Price dispersion among retailers of the same product is observed. With its personalized website, big brand name, broad range of assortment, and by making consumers Amazon addicts with its Amazon Prime program, it entices consumers to its website (Kotler & Armstrong, 2016). Therefore, Amazon has a very big potential of cross-selling and impulse buying. According to Kotler & Armstrong, 2016, the discovery effect of Amazon's website lures customers to discover and stay for a while to learn products, alternatives, and other customers' opinions.

Baye, Morgan & Scholten, 2004; Clay, Krishnan & Wolff, 2001 showed that in the online market, loss leader pricing is effective when complementary goods are existent. In addition, according to Kocas, Pauwels & Bohlmann, 2018, there exists an asymmetric cross-selling potential among sellers, and retailers with larger average

basket sizes like Amazon.com can benefit from loss leader pricing strategy. In fact, Amazon has been accused of predatory pricing by many publishers and book sellers for destroying their industry (Kotler & Armstrong, 2016).

In this research, it is of our interest to observe what factors impact Amazon's discounting decisions on books. By observing a dataset with 15500 observations consisting of 31 days for each 500 books on six important market characteristic variables, the relationship among discount, sales rank, list price, customer reviews, and number of sellers observed. In this study, since we observe a market where a homogeneous product is offered by all the sellers, the relationships of the factors can be easily seen. By using Panel Vector Autoregressive Modelling, we not only observe the effect of variables for the day of observation, but also the previous days' effect of each variable on other variables.

2. LITERATURE REVIEW

Started its online business in 1995, the online pioneer Amazon was selling only books. Today, it still sells books but in huge amounts and with a broad range of other products such as electronics, clothing, toys, movies, housewares, groceries, jewelry, etc. (Kotler & Armstrong, 2016). In 2019, this company made \$11.588 billion in profits (CNBC, 2019). One may wonder what the recipe of Amazon to become such a successful company is. If you ask Jeff Bezos, it is "Obsess over customers" (Kotler & Armstrong, 2016).

It was the first company to create personalized stores by analyzing customer data on past purchases, browsing histories, and patterns of similar customers. The big data-driven customer interface is unbeatable in creating a highly satisfying online buying experience (Kotler & Armstrong, 2016).

According to Kotler & Armstrong, 2016, prices have been the weapon of both Amazon and Walmart in their online supremacy battle. Walmart is fighting with aggressive pricing; however, when the prices on these companies' websites are compared, it can be seen that this war raging across lots of products.

In the online market, competition is more severe than offline markets; however, it is nowhere near to perfect competition (Li, Tang, Huang & Song, 2009). Price dispersion among retailers that sell same product is observed by many researchers, and according to Zhao, Zhao & Deng, 2015, this expresses market efficiency.

In the literature, various factors that lead to price dispersion online and offline are examined. In our model, we include key market characteristics that can be summarized from Pan, Ratchford & Shankar, 2004; Zhao et al., 2015 to observe how they affect each other and discounting of the products. These are item price level (list price in our model), number of sellers of the same product, and product popularity (sales rank and customer reviews in our case).

Product price levels can be an important factor of price dispersion and Stigler, 1961 stated that expensive products would have lower price dispersion than cheap

products which can be because of consumer motivation to search more for the best price. Consumers' search for the best prices can force sellers to set prices at a competitive level (Zhao et al., 2015).

Number of sellers is also an important factor that affects the extent of price dispersion. According to the findings of Stigler, 1961; Wang & Li, 2020, a larger number of sellers increases search costs of consumers and this causes additional price dispersion across retailers. On the other hand, increased competition with the increase in number of sellers may decrease price dispersion because sellers would need to set prices competitively when there are lots of alternatives for consumers (Stiglitz, 1987). According to the study of Zhao et al., 2015, increase in search costs outweighs for consumers but competition outweighs for sellers. Thus, they highlighted two contrasting findings on impacts of number of sellers to price dispersion.

Furthermore, seller reputation is a key factor in price dispersion among homogeneous products. According to Wang & Li, 2020, since consumers do not examine the physical products, the seller quality indicators and trustworthiness are important for them. Results of Wang & Li, 2020 showed that store reputation leads to online price dispersion, and although sellers strategically price their products, realization of sales is related to reliance of consumers to sellers. They suggested that new stores may establish reputation and survive online with substantial price discounts, advertising, or reward programs. Zhao et al., 2015 showed that price dispersion of listing prices is affected by heterogeneous seller reputation, and their results showed that sellers make pricing decisions according their reputation.

We also include format of the books as an exogenous variable since format of the book may change the pricing decisions of both publishers and Amazon. According to Barrot, Becker, Clement & Papies, 2015, price elasticity for paperback books tend to be more negative than for hardcover books, and pricing decisions are made accordingly. They stated that usually hardcover version of a book is released before paperback version, and while paperback books have more utilitarian character, hardcover books have more hedonic character. Thus, paperback books attract a different customer segment which is more sensitive to price changes.

Li et al., 2009, with their least-squares dummy variable panel data model on a longitudinal dataset of 27030 price observations over one year on Australian online DVD market, found that the price dispersions of popular titles category are smaller than random titles category. They explained that the severe competition among retailers and lower searching costs for consumers may push them to sell the books at similar discount levels with the other booksellers or at the distributor suggested prices to set the optimum pricing strategy. If retailers have small number of these

items in their inventory, they may give big discounts to increase operation cash flow.

If a retailer can assure that its price is the lowest price for a best selling item and have the ability to sell large inventory of these items at the same price, it may dominate the market. Kocas et al., 2018 stated that cross-selling potential is asymmetric across retailers and this should be taken into account when modelling. By constructing a model which assumes asymmetry among retailers, they found that retailers with larger average basket sizes can benefit from loss leader pricing and attract more traffic and profits by capitalizing on cross-selling efforts.

By using machine learning techniques random forest, neural networks, and boosted gradient trees, Bodoh and Boehnke and Hickman, 2017 explained the price dispersion of e-bay price listings for different categories of homogeneous products. They found that random forest is the best technique and they could explain around 26% of the price dispersion within each category. They found that the reliability and professionalism in the posts are important for explaining price dispersion.

Amazon is a company that sells almost every type of product at competitive prices with excellent personalized service that is unbeatable on such a mass level (Kotler & Armstrong, 2016).

In traditional store setting, Inman, Winer & Ferraro, 2009 stated that in general, consumers should be motivated to shop as many aisles as possible and in particular, be exposed to lots of product categories and in-store displays. Visiting few aisles, more frequent trips to stores, paying by cash, limited time spent in the store, using lists are reducing the likelihood of making unplanned purchases. In online setting, Amazon.com can provide huge sizes of assortments, creating a spot for one-stop shopping in the convenience of internet. Shoppers can access the stores from anywhere they are and spend as much time as they want. Then, in the time consumers visit the store, amazon.com should encourage consumers to shop from various categories and expose them to as many products as possible to increase impulse buying. Internet retailers, including Amazon.com ensures this with tricks such as interactive displays which people can zoom in or spin the product photograph, or recommendation systems containing social influence features based on what "other people" bought, etc. (Thomas, L., 2019).

Personal influence carries great importance, especially for expensive, highly visible, or risky products. Recommendations from people, online consumer opinions are things that consumers care about. People check Amazon's customer reviews and "Customers who bought this also bought..." section before deciding to purchase an item (Kotler & Armstrong, 2016).

The searches consumers make online, the sites they visit, how and what they purchase are information that are pure gold to marketers. In order to use consumer data and use consumer information in target ads personalized for each customer, they mine that gold. This is why consumers see the ad of the item they leave in shopping cart without buying on Amazon.com in the news or sports websites (Kotler & Armstrong, 2016).

Stilley, Inman & Wakefield, 2010 stated that consumers have a mental budget and this budget has space for unplanned purchases in the store. They created the term in-store slack which indicates the consumers' room for unplanned purchases. Number of aisles visited and impulsiveness of the consumer affect what he/she does with in-store slack. They showed that savings on planned purchases increase the quantity of items purchased, and for highly impulsive shoppers, savings on purchases of unplanned products affect the purchase of more unplanned products.

Kocas et al., 2018 labeled the cross-selling in which customers who are interested in a best seller book and impulse buy other items as conversion, and the cross-selling in which consumers interested in other products also purchase best seller book as inclusion. They found that for products with higher conversion to inclusion ratios like best-seller books or seasonal items, the depth and the frequency of discounts are higher. They also stated that as conversion to inclusion ratio goes up, price of any best-seller item decreases.

Lee & Ariely, 2006 found that consumers begin their shopping processes with ill-defined shopping goals and these goals become more concrete as consumers go through the process. This lack of concreteness at the beginning of their shopping process may be used as an opportunity to manipulate the goals of consumers by giving them conditional promotions. In our case, conditional promotions such as buy 6, get 1 free; or free delivery for orders \$25 or more may also have effects on buying other products.

Dhar, Huber & Khan, 2007 discussed that psychological factors may lead to purchase of additional items, and defined shopping momentum effect which is continuing purchasing of unplanned products after an initial purchase. Promotions and discounts thus can attract both consumers who shop only for discounted items and the consumers who end up with large baskets because of psychological factors.

Walters & Jamil, 2002 measured the cross-category discounted item purchasing of consumers by analyzing shopping basket level data, and found that 39% of all the items purchased were on discount, and consumer search behaviors and household income affected purchasing of cross-category discounted item purchasing.

The website, amazon.com, lures customers to discover and stay for a while in the website to learn products and purchase alternatives as well as reading product reviews (Kotler & Armstrong, 2016).

The attractiveness of the items that are added to the assortment is important to increase the likelihood of purchasing (Koelemeijer & Oppewal, 1999). Amazon.com has huge variety of products and large assortments that appeal any consumer. When a loss leader best-seller book is added to the products to be sold, anyone can find appealing products to buy with the book.

Loss leading strategy is to put deep discounts on some products, sometimes selling them below cost to attract customers (Hess & Gerstner, 1987). To survive competition, all retailers have to have loss leaders to entice consumers to visit their stores by undertaking the trip to the store (Lal & Matutes, 1994). In retail markets with retailers that use promotions to attract consumers, promotions are important factors for competitive dynamics. By featuring deeply discounted items, additional store traffic and increased sales and profits will be generated for retailers (Gauri, Ratchford, Pancras & Talukdar, 2017).

It is of great interest to academicians and managers to have an understanding on how consumers react to retail promotions and what type of promotions are useful. In the study conducted by Gauri et al., 2017 with a store level dataset for 55 weeks and 24 stores, the effects of discounts on store performance metrics such as store traffic, sales per transaction, and profit margin are examined. They found that promotional discounts are beneficial on store performance metrics. Store traffic increases in response to discounts, especially when categories discounted are high penetration, high frequency items, and increased traffic facilitates profits. In addition, discounts increase sales per transaction especially when discounts are put on branded items. However, discounting a large proportion of a category leads to lower store margins.

Competitive forces may be a factor in determining prices at different retail formats, since consumer spending at a retailer may imply less spending at a competitor retailer. Location, ambiance, information, assortment, delivery are services consumers pay for because of lower search and transportation costs, and other benefits. Loss leader promotions exploit the within-format differences in location, information, and tendency to search among consumers (Kopalle, Biswas, Chintagunta, Fan, Pauwels, Ratchford & Sills, 2009), and can be used as a strategy against competitors.

According to Hayashida and Hoshino, 2020, loss leading increases store level sales, and loss leading in one product leads to positive effects on other categories of items. They also found that loss leading is more effective when stores compete locally, and

has the effect of reducing prices locally which benefits consumers.

Lal & Matutes, 1994 found that retailers sell products below marginal cost to increase store traffic and earn profit from other products. They also state that loss leaders may be the items which huge number of consumers buy, and are difficult to stockpile. Stockpiling may be hard for goods that are perishable, frequently consumed, or require large space to store. Thinking this way, amazon.com can offer loss leader discounts on best sellers since a best seller book is bought by large number of consumers and cannot be stockpiled since having one book is enough to consume many times.

Amazon has sold top ten best selling hardback books as loss leaders at less than \$10 each, and put very low prices on e-books to win customers for Kindle. It has been accused of predatory pricing by many publishers and book sellers for destroying their industry; however, proving Amazon's loss leader pricing being purposefully predatory instead of being good competitive marketing is not easy. This can be considered as selling below cost and a healthy competition instead of predatory pricing (Kotler & Armstrong, 2016).

In our study, we would like to observe the factors that influence Amazon's pricing decisions and the factors' relationships within each other in Amazon marketplace.

3. DATA DESCRIPTION

3.1 Dataset Description

Our dataset represents a time series dataset with 19 variables which has 847,403 observations of 7334 books, and runs from June 1, 2011 to August 10, 2011, a total of 71 days. These books are listed under New Releases - Coming Soon on amazon.com, meaning they are not available for shipment at the beginning of their data collection, but can be preordered. With this dataset, we can observe how their price and sales rank evolve over time starting prior to their shipment date. The descriptive statistics of numerical variables in our dataset are presented in Table 3.1.

Table 3.1 Descriptive Statistics of Numerical Variables in Raw Data

Variable	Observations	Missing	Mean	Median	Std. Dev.	Minimum	Maximum
listprice	520612	326791	29.31	21.95	24.90243	6.99	779.48
price	834024	13379	29.45	14.95	95.37559	0.01	4271
You Save	520612	326791	8.26	6.4	8.122728	0.01	429.49
You Save %	520578	326825	29.74	32	8.016844	1	77
ABRank	613439	233964	1316078.37	529057	1929478	1	10517303
retailers	360369	487034	20.68	19	11.88517	1	111
avg_cus_review	366043	481360	4.173	4.2	0.618474	1	5
numberoffike	847137	266	3.8	0	49.33375	0	3714
total reviews	366050	481353	64.25	16	192.5995	0	4708
5 Star reviews	366050	481353	34.19	7	127.3028	1	3083
4 Star reviews	366050	481353	12.98	4	30.16258	0	435
3 Star reviews	366050	481353	6.99	2	16.18238	0	258
2 Star reviews	366050	481353	4.89	1	13.88149	0	246
1 Star reviews	366050	481353	5.4	0	22.03623	0	808

3.2 Problems with the Data

There are some problems that prevent us to perform the analysis correctly. In this section, we will present these problems.

Missing Values

First of all, the dataset has many missing values which prevent data analysis. For example, for some of the books, the rank information is missing, for some of them the list price is unknown which obstruct calculation of the discount. These missing values should be handled in order to conduct the data analysis

Naming

The names of the books may also create a problem in the analysis because there are misspellings for the same book. In the name column, the authors name and the books name differ for some books. This problem can be fixed by taking ISBN as the panel identifier.

Unbalanced Number of Observations

Another problem we encounter is having unbalanced number of observations within and across books. For the same book, there may be one observation for June 1st, and three observations for July 2nd at the same time. In addition, some books are not observed in some of the days. For example, observations of a book start from June 1st, but for another book it is June 8th. Thus, there may be 90 observations for a book and 116 observations for another one. We need to have same number of observations for each day and for each book.

3.3 Data Cleaning

Data cleaning is considered as the most important step of data analysis. If the data is not well-prepared, the outcomes of the analysis would be useless. Thus, before starting our analysis, we will prepare our data to obtain correct and meaningful results.

In our analysis, list price, current price, sales rank, number of sellers and the customer rating of these books are the necessary variables. By using list price and current price information, we can calculate discount on these books. The information on these variables are crucial for our analysis, and the data containing this information should be well prepared.

First thing to do is dealing with the missing values. Missing values prevent us from performing Vector Autoregression. As it can be seen from Table 3.1, there are 326,791 missing values in the listprice column, 13379 in the price column, and 233,964 in the ABRank column. We observe that wherever the price column has missing values, listprice value of that observation is also missing. We delete the observations with missing values in listprice and ABRank columns. Doing so, we also get rid of missing values of the price column.

After the missing values in the listprice and ABRank columns are deleted, only missing values in the retailers and avg_cus_review are left. In this dataset, the missing values in the avg_cus_review and retailers columns are not entered because the book was not on sale on these dates and there were no sellers of these books on these dates. Thus, imputing 0 for the missing observations would be meaningful.

We can now calculate discounts offered. We calculate discount as *1-standardized price*, and standardized price by diving current sales price by the list price.

Next, we deleted the unnecessary variables from our dataset. Only ISBN13, listprice, discount, ABRank, retailers, date, and physical_format variables are left.

Descriptive statistics of the numerical values in the dataset with no missing values can be seen in Table 3.2. The descriptive statistics of physical_format and ISBN13 can be seen in the Table 3.3 and Table 3.5 respectively. It can be seen that number of observations of the books are not the same with each other. We need to crate a balanced sample with same number of observations for each book and same number of observations for a book in each day.

Table 3.2 Descriptive Statistics of the Numerical Variables of the Data with No Missing Values

	listprice	discount	ABRank	retailers	avg_cus_review
Observations	419147	419147	419147	419147	419147
Missing	0	0	0	0	0
Mean	25.21742	0.305871	1110851	11.49296	1.861292
Median	19.99	0.320213	442281	0	0
Std. Dev.	20.15595	0.077574	1698997	14.62705	2.113769
Minimum	6.99	0.000278	1	0	0
Maximum	779.48	0.772886	10472007	111	5

Physical Format

Physical format of the books is the exogenous variable of our analysis. However, not all the entries of physical format column necessarily different from each other. We organize the entries as a categorical variable with four categories. These categories are Hardcover, Paperback, Audiobook, and Others. The classification can be seen in Table 3.3. In this table, it is observed that books with Audiobook MP3 Audio Unabridged MP3 CD Library Binding format has only two, with Abridged Audiobook Unabridged Audio CD format has 57 observations, and with Mass Market Paperback format has 71 observations. The reason for these low numbers is wrong data entry. When we observe the data, the book with ISBN 978-1452652931 is a book with Audiobook MP3 Audio Unabridged MP3 CD format but for two observations, it entered as Audiobook MP3 Audio Unabridged MP3 CD Library Binding. The same also can be observed for the books with ISBN 978-1442344228 and ISBN 978-1554888931. Although the former is entered as Audiobook Unabridged Audio CD in the majority of the observations, it is entered as Abridged Audiobook Unabridged Audio CD for 57 observations, and the latter is in the Paperback format in the majority of the observations but it is entered as Mass Market Paperback for 71 observations. In our classification, we corrected all these wrong entries in the data.

Table 3.3 Descriptive Statistics of Physical Format and Our Classification

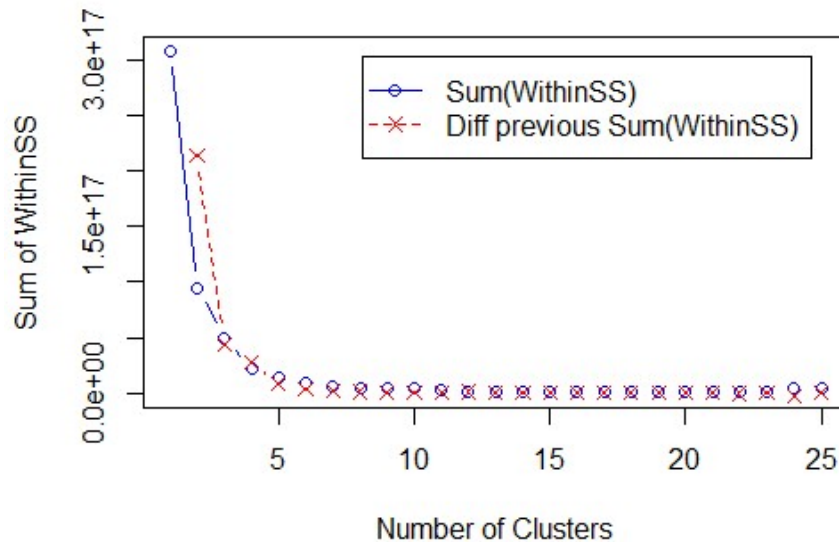
Physical_Format	Count_CleanedData	Consolidated
Paperback	167854	Paperback
Hardcover	149776	Hardcover
Audiobook CD Unabridged Audio CD	25155	Audiobook
Audiobook MP3 Audio Unabridged Audio CD	20400	Audiobook
Calendar	16260	Others
Audiobook Unabridged Audio CD	8858	Audiobook
Large Print Paperback	6394	Paperback
Abridged Audiobook CD Audio CD	4603	Audiobook
Large Print Hardcover	4596	Hardcover
Audiobook Audio CD	3375	Audiobook
Library Binding	2307	Hardcover
Abridged Audiobook Audio CD	1724	Audiobook
Board book	1244	Hardcover
Audiobook CD Audio CD	1010	Audiobook
Abridged Audio Cassette	463	Deleted
Cards	410	Others
Audiobook Unabridged MP3 CD	400	Audiobook
Audio CD	398	Audiobook
Unabridged Audio CD	359	Audiobook
Unabridged Audiobook Audio CD	348	Audiobook
Deluxe Edition Hardcover	348	Hardcover
Audiobook MP3 Audio Unabridged MP3 CD	341	Audiobook
CD-ROM	328	Deleted
Unabridged Paperback	273	Paperback
Unabridged Audio Cassette	266	Deleted
Game	197	Others
Leather Bound	152	Hardcover
Unabridged Hardcover	129	Hardcover
Abridged Audiobook Audio Cassette	116	Deleted
Abridged Audio CD	116	Audiobook
Audiobook MP3 Audio Unabridged Preloaded Digital Audio Player	116	Audiobook
DVD-ROM	116	Deleted
Deckle Edge Hardcover	116	Hardcover
Deluxe Edition Paperback	116	Paperback
Misc. Supplies	115	Others
Abridged Audiobook CD Unabridged Audio CD	93	Audiobook
Large Print Library Binding	86	Hardcover
Mass Market Paperback	71	Paperback
Import Paperback	59	Paperback
Abridged Audiobook Unabridged Audio CD	57	Audiobook
Audiobook MP3 Audio Unabridged MP3 CD Library Binding	2	Audiobook

K-means Clustering of Retailers Variable

The retailers column indicates number of sellers of each book for the date of observation. In this variable, there are many zero values, and taking number of sellers information as a continuous variable may be misleading. Thus, we make k-means clustering to decide how many categories we should divide this variable. The elbow method will show us the optimal number of clusters to choose.

As it can be seen from Figure 3.1, dividing the retailers variable into 5 clusters gets us very close to total within cluster sum of squares. According to this analysis, we divide our data as Table 3.4.

Figure 3.1 Elbow method to choose number of clusters for retailers variable



3.3.1 Creating the Balanced Panel Data

We created the dataset without any missing values but the number of observations per book is still not balanced. Different books have different number of observations, as it can be seen in Table 3.5. This is because there are different number of observations per day and some of the books are not observed in some of the days. Making our dataset as a balanced panel dataset would enable us to eliminate the effects of different books have on the results of our models. In order to make our dataset a balanced panel dataset, we used stratified random sampling method.

Table 3.4 Clustering of Number of Sellers

Value	Cluster
0	1
2 - 7	2
8 - 21	3
22 - 34	4
35+	5

Stratified Sampling

The data should be representative of the cleaned dataset. In order to ensure that, we took a stratified sample from the dataset. Looking at our exogenous variable, format, we have 419,147 observations, and 158,754 of them are hardcover, 174,767 of them are paperback, 68,200 of them are audiobook, and 17426 of them are in other format. We need a sample of 500 books with the same proportion of our dataset. In other words, the dataset has 38% hardcover, 42% paperback, 16% audiobook, and 4% of other format. Thus, our 500 books will be consisting of 190 hardcover, 208 paperback, 81 audiobook, and 21 other format.

The descriptive statistics of balanced panel data can be seen in Tables 3.6 and 3.7.

Table 3.5 Descriptive Statistics of ISBN

Row Labels	Count of ISBN13
978-0099540762	1
978-0738575780	2
978-0805089318	2
978-1455815517	2
978-1455816309	2
978-1455821594	2
978-1455821686	2
978-0755352586	3
978-0814776384	3
978-0719073397	4
978-0738575216	4
978-0738576305	4
978-1441794161	4
978-1615640898	4
.	.
.	.
.	.
978-0230106666	33
978-0230115088	33
978-0547745008	33
978-0738575353	33
978-0738579702	33
978-0738582481	33
978-0738583204	33
978-0738584904	33
978-0738587554	33
978-1419364112	33
978-1423809494	33
978-9380028569	33
978-0061686566	34
978-0143304708	34
.	.
.	.
.	.
978-8857200569	116
978-8857208305	116
978-9380741246	116
978-9380741253	116
978-9626342688	116
978-0061980978	117
978-0230111639	117
Grand Total	419147

Table 3.6 Descriptive Statistics of Numerical Variables in Balanced Panel Data

Variable	listprice	discount	ABRank	avg_cus_review
Observations	15500	15500	15500	15500
Missing	0	0	0	0
Mean	27.64034065	0.299525327	1171109.123	1.576445161
Median	19.99	0.32016008	421628.5	0
Std. Dev.	37.91527691	0.083067413	1809020.022	2.081147656
Minimum	10	0.020701169	1	0
Maximum	779.48	0.77083947	10453559	5

Table 3.7 Descriptive Statistics of Categorical Variables in Balanced Panel Data

Row Labels	Count of format	Row Labels	Count of numberofsellers
1	5890	1	7447
2	6448	2	765
3	2511	3	2432
4	651	4	3355
blank	0	5	1501
Grand Total	15500	Grand Total	15500

4. METHODOLOGY

Vector autoregressive modelling has been one of the most widely used method to explain the relationship between multiple time series. This study tends to assess the performance of non-panel time series and panel time series models in explaining the relationship between our variables. We use vector autoregressive modelling for our purposes. The steps of our analyses and their details can be found in this chapter.

4.1 Modelling Approach

We follow the steps of persistence modelling as done by Kocas et al., 2018; Trusov, Bucklin & Pauwels, 2009 using our time series dataset with 6 variables and 15500 observations. Afterwards, we follow the same steps by introducing our dataset as a panel time series dataset with 500 cross-sections, 31-day observations of each five endogenous variables for each cross-sections, and one exogenous variable. In this part, we use the same techniques that are modified for panel data analysis. Finally, we change physical format and number of sellers variables of our panel dataset to make improvements on our analysis. The steps of persistence modelling can be summarized as:

- Testing for stationarity with Unit Root Tests
- Testing for stationarity of series with Cointegration Test
- Testing for Endogeneity with Granger Causality Test
- Vector Autoregression Analysis
- Estimating Impulse Response Functions

4.2 Unit Root Tests

First of all, we test stationarity of our time-series variables. Stationarity of a time series means that the mean and autocovariances of the series are time invariant, in other words, do not depend on time.

If the time series variables of the datasets are not stationary, the results would not be meaningful. Stationarity of variables of a VAR model and VAR model itself are crucial and restrictive in analyses (Lütkepohl, 2005). To ensure stationarity of our variables, we perform unit root tests on each of our variables.

In order to test stationarity of a time series, unit root tests such as Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test (Kwiatkowski, Phillips & Shin, 1992), Phillips-Perron (PP) test (Perron, 1988), Augmented Dickey-Fuller (ADF) test (Said & Dickey, 1984) can be used. In our analysis, we test stationarity of our time series variables with ADF test.

For panel data analysis, unit roots of time series variables are tested with common root (Breitung, 2001; Levin, Lin & Chu, 2002), and individual root (Choi, 2001; Im, Pesaran & Shin, 2003; Maddala & Wu, 1999) tests. We use Levin, Lin, and Chu Unit Root Test which assumes a common root process so that autoregressive coefficients are identical across cross-sections but allows the lag order for the difference terms to vary across cross-sections.

4.3 Cointegration Test

We test stationarity also with Cointegration Tests. Although we conduct unit root tests for each of the variables, cointegration tests test stationarity based on Vector Autoregression.

Johansen Cointegration test (Johansen, 1991) can be used for time series data. This test is an improvement of Engle-Granger Cointegration Test (Engle & Granger, 1987) in which the presence of unit roots is tested using ADF test, and the difference of Johansen Cointegration test is that more than one cointegrating vectors can be detected. In this test, the null hypothesis is there is no cointegration so we need to reject the null hypothesis since if the time series are cointegrated, then the residuals

will be stationary.

For panel data, Pedroni Cointegration Test (Pedroni, 1999), Kao Cointegration Test (Kao, 1999), or Fisher-type Cointegration Test (Maddala & Wu, 1999) can be used. We use Johansen test for analyzing our time series dataset, and Kao test for analyzing our balanced panel time series datasets. Kao Cointegration Test is an Engle-Granger based test which is extended for panel data analysis. The null hypothesis is there is no cointegration.

4.4 Granger Causality Test

We test endogeneity using Granger Causality test (Granger, 1969). This test assesses how much of the current value of a variable can be explained by past values of this variable, and whether adding lagged values of other variables would be statistically significant. If there is no Granger causality between the variables, then Applying VAR model and interpretation of the model will not be useful in explaining the relationship between these variables.

Note that, the statement *Discount Granger causes sales rank* does not imply that sales rank is the effect or result of discount. With this test, precedence and information content are measured; the common use of the term *causality* is not indicated.

When we adjust our data as a panel data, we use Stacked (Common Coefficients) Granger causality test, so it is assumed that all coefficients are the same across all cross-sections. This test treats the panel data as one large stacked set of data and test Granger causality in the standard way but does not let data from one cross-section to enter the lagged values of data from the next cross-section.

4.5 Vector Autoregressive Modelling

To model VAR using our non-panel time series data, we use the VAR model as Kocas et al., 2018 used in analysing their first dataset.

A k-variate VAR(p) model (Lütkepohl, 2005) can be shown as follows,

$$Y_{it} = c + Y_{it-1}A_1 + Y_{it-2}A_2 + \dots + Y_{it-p+1}A_{p-1} + Y_{it-p}A_p + X_{it}B + e_{it}$$

In which $i \in \{1, 2, \dots, N\}$, $t \in \{1, 2, \dots, T_i\}$; and $c = (c_1, \dots, c_k)'$ is a fixed (kx1) vector of intercept terms, $(Y_{it})'$ is a (kx1) vector of dependent variables, $(X_{it})'$ is a (ℓ x1) vector of exogenous covariates, and e'_{it} is (kx1) vector of errors. The (kxk) matrices $A_1, A_2, \dots, A_{p-1}, A_p$ and (ℓ xk) matrix B' are parameters to be estimated.

In our model, we take $p = 1$ since it is the optimal lag length. Thus, our VAR(1) model can be represented as follows,

$$\begin{bmatrix} Discount_t \\ Rank_t \\ Sellers_t \\ Listprice_t \\ Cust.Rev_t \end{bmatrix} = \begin{bmatrix} \alpha_D \\ \alpha_R \\ \alpha_S \\ \alpha_L \\ \alpha_C \end{bmatrix} + \begin{bmatrix} \delta_D \\ \delta_R \\ \delta_S \\ \delta_L \\ \delta_C \end{bmatrix} \times Format + \sum_{j=1}^J \begin{bmatrix} \phi_1^1 \\ \phi_1^2 \\ \phi_1^3 \\ \phi_1^4 \\ \phi_1^5 \end{bmatrix} \times \begin{bmatrix} Discount_{t-j} \\ Rank_{t-j} \\ Sellers_{t-j} \\ Listprice_{t-j} \\ Cust.Rev_{t-j} \end{bmatrix} + \begin{bmatrix} \varepsilon_{D,t} \\ \varepsilon_{R,t} \\ \varepsilon_{S,t} \\ \varepsilon_{L,t} \\ \varepsilon_{C,t} \end{bmatrix}$$

However, Holtz-Eakin, Newey & Rosen, 1988 has stated that individual heterogeneity is an important feature of disaggregate data, and it is not suitable to apply standard techniques for vector autoregressions to panel data. This is why we use their proposed set of procedures for estimating and testing vector autoregressions with our dataset after introducing our dataset as a panel dataset in the second analysis, and modify the format and number of sellers variables in the third analysis.

The PVAR(p) model (Holtz-Eakin et al., 1988) can be represented as,

$$Y_{it} = c + Y_{it-1}A_1 + Y_{it-2}A_2 + \dots + Y_{it-p+1}A_{p-1} + Y_{it-p}A_p + X_{it}B + u_i + e_{it}$$

In which $i \in \{1, 2, \dots, N\}$, $t \in \{1, 2, \dots, T_i\}$; and $c = (c_1, \dots, c_k)'$ is a fixed (kx1) vector of intercept terms, $(Y_{it})'$ is a (kx1) vector of dependent variables, $(X_{it})'$ is a (ℓ x1) vector of exogenous covariates, u'_{it} is a (kx1) vector of dependent variable specific panel fixed-effects, and e'_{it} is (kx1) vector of white noise error. The (kxk) matrices $A_1, A_2, \dots, A_{p-1}, A_p$ and (ℓ xk) matrix B' are parameters to be estimated.

The term e_{it} satisfies the orthogonality condition, that means lagged values of Y_{it} qualify as instrumental variables, i.e. they can be used for explaining the error term.

Our 5 endogenous variable, 1 exogenous variable PVAR(6) model is as follows:

$$\begin{bmatrix} Discount_t \\ Rank_t \\ Sellers_t \\ Listprice_t \\ Cust.Rev_t \end{bmatrix} = \begin{bmatrix} \alpha_D \\ \alpha_R \\ \alpha_S \\ \alpha_L \\ \alpha_C \end{bmatrix} + \begin{bmatrix} \delta_D \\ \delta_R \\ \delta_S \\ \delta_L \\ \delta_C \end{bmatrix} \times Format + \sum_{j=1}^J \begin{bmatrix} \phi_1^1 \phi_2^1 \phi_3^1 \phi_4^1 \phi_5^1 \\ \phi_1^2 \phi_2^2 \phi_3^2 \phi_4^2 \phi_5^2 \\ \phi_1^3 \phi_2^3 \phi_3^3 \phi_4^3 \phi_5^3 \\ \phi_1^4 \phi_2^4 \phi_3^4 \phi_4^4 \phi_5^4 \\ \phi_1^5 \phi_2^5 \phi_3^5 \phi_4^5 \phi_5^5 \\ \phi_1^6 \phi_2^6 \phi_3^6 \phi_4^6 \phi_5^6 \end{bmatrix} \times \begin{bmatrix} Discount_{t-j} \\ Rank_{t-j} \\ Sellers_{t-j} \\ Listprice_{t-j} \\ Cust.Rev_{t-j} \end{bmatrix} + \begin{bmatrix} u_{D,t} \\ u_{R,t} \\ u_{S,t} \\ u_{L,t} \\ u_{C,t} \end{bmatrix} + \begin{bmatrix} \varepsilon_{D,t} \\ \varepsilon_{R,t} \\ \varepsilon_{S,t} \\ \varepsilon_{L,t} \\ \varepsilon_{C,t} \end{bmatrix}$$

4.6 Impulse Response Functions

The response of a variable to an impulse in another variable in a system can be traced with impulse response functions (IRF). We may call that a variable is causal for another, if there is a reaction of the variable to an impulse in the other variable (Lütkepohl, 2005).

Sims, 1980 has stated that vector autoregressive model parameters are not interpretable on their own, and effect sizes and significance of the relationships should be determined through the analysis of impulse response functions.

In order to understand causal relationships between our variables, and interpret the parameters and their significance, we use Generalized Impulse Response Functions (GIRF).

As Pesaran & Shin, 1998 has described, generalized impulses construct an orthogonal set of innovations, and GIRF are derived by applying a variable specific Cholesky factor.

GIRF can be used for both non-panel and panel datasets. We use GIRF for all analyses. In addition, we calculate cumulative elasticities as the sum of all GIRF coefficients significantly different from 0 at the 95% significance level in the way Kocas et al., 2018; Trusov et al., 2009 have done.

5. RESULTS

In this chapter, we present the results of all the steps of our analysis. Mainly, we perform three different models. Firstly, in Chapter 5.1, we make the analysis as Kocas et al., 2018. We follow the persistence modelling steps (Kocas et al., 2018; Trusov et al., 2009) and take the variables as continuous time series without distinguishing the books as panel data. In Chapter 5.2, we adjust our dataset as a panel dataset, and perform all the steps Kocas et al., 2018 followed with the same methods tailored for analyzing panel data. Finally, in Chapter 5.3, we modify our exogenous variable *physical_format*, and endogenous variable *retailers* to perform a better model.

5.1 Data Analysis

In this section, we present the results of the analysis performed by following the persistence modelling steps as Kocas et al., 2018 to explain the relationship between the variables.

5.1.1 Unit Root Tests

In order to test whether the time series variables have unit root or not, we use ADF test. According to the test results on each variable, there is no unit root in these variables. Since all of them are stationary datasets, there is no need to make any adjustments or take their differences. The results of the tests are presented in Table 5.1.

Table 5.1 Unit Root Test Results

Augmented Dickey Fuller Test		
Alternative Hypothesis: Stationary		
Data	Dickey-Fuller	p-value
discount	-14.212	0.01
abrank	-16.491	0.01
listprice	-19.164	0.01
retailers	-17.355	0.01
avg_cus_review	-17.287	0.01

5.1.2 Cointegration Test

We use Johansen Cointegration Test in order to test stationarity based on VAR model. It is clear in Figure 5.1 that we reject the null hypothesis that there is no cointegration between these variables.

Figure 5.1 Johansen Cointegration Test

Unrestricted Cointegration Rank Test (Maximum Eigenvalue)				
Hypothesized No. of CE(s)	Eigenvalue	Max-Eigen Statistic	0.05 Critical Value	Prob.**
None *	0.022739	356.1712	30.43961	0.0001
At most 1 *	0.019754	308.9519	24.15921	0.0001
At most 2 *	0.016398	256.0283	17.79730	0.0001
At most 3 *	0.014997	233.9919	11.22480	0.0001
At most 4 *	0.000596	9.225470	4.129906	0.0028

Max-eigenvalue test indicates 5 cointegrating eqn(s) at the 0.05 level

* denotes rejection of the hypothesis at the 0.05 level

**Mackinnon-Haug-Michelis (1999) p-values

5.1.3 Granger Causality Test

Granger causality of the time series variables up to 8 lags suggests that sales rank Granger causes and Granger caused by discount at any lags up to 8 lags at 95 % significance level. In addition, discount Granger causes customer review at any lags

at 95 % significance level but the reverse is not supported. Sales rank also Granger causes retailers at any lag. These variables can be used to explain each other, Vector Autoregression modelling can be performed with our dataset.

5.1.4 Vector Autoregression Model

The ADF Unit Root tests, Johansen Cointegration test, and Granger Causality test show that applying Vector Autoregression Model would give meaningful and useful results in explaining the relationship between our time series variables. First step is to select the optimal lag length, second step is to perform the analysis.

5.1.4.1 Model Selection

The lag length that is optimal can be selected by looking at the information criterion such as Akaike Information Criterion (AIC) or Schwarz Information Criterion (SC). As it is presented in Figure 5.2, the optimal lag length for our model is selected as lag 1 according to both AIC and SC.

Figure 5.2 Model Selection for VAR model

VAR Lag Order Selection Criteria
 Endogenous variables: DISCOUNT ABRANK LISTPRICE RETAILERS AVG_CUS_REVIEW
 Exogenous variables: C HARDCOVER
 Date: 08/11/20 Time: 17:03
 Sample: 1 15500
 Included observations: 15486

Lag	LogL	LR	FPE	AIC	SC	HQ
0	-399618.7	NA	1.79e+16	51.61161	51.61655	51.61325
1	-298792.9	201560.4	3.97e+10*	38.59330*	38.61058*	38.59903*
2	-298770.8	44.12381	3.97e+10	38.59368	38.62331	38.60349
3	-298749.0	43.56498	3.97e+10	38.59409	38.63607	38.60799
4	-298737.8	22.44639	3.98e+10	38.59587	38.65019	38.61386
5	-298726.6	22.36445	3.98e+10	38.59765	38.66432	38.61973
6	-298711.4	30.32965	3.99e+10	38.59891	38.67793	38.62509
7	-298701.9	19.03446	4.00e+10	38.60091	38.69227	38.63118
8	-298692.6	18.49651	4.01e+10	38.60294	38.70665	38.63730
9	-298682.3	20.51885	4.01e+10	38.60484	38.72090	38.64329
10	-298670.2	24.03465	4.02e+10	38.60651	38.73491	38.64905
11	-298660.3	19.72191	4.03e+10	38.60846	38.74921	38.65509
12	-298653.9	12.90042	4.04e+10	38.61086	38.76395	38.66157
13	-298601.9	103.4741*	4.02e+10	38.60737	38.77281	38.66218
14	-298589.7	24.34147	4.03e+10	38.60902	38.78681	38.66792

* indicates lag order selected by the criterion
 LR: sequential modified LR test statistic (each test at 5% level)
 FPE: Final prediction error
 AIC: Akaike information criterion
 SC: Schwarz information criterion
 HQ: Hannan-Quinn information criterion

5.1.4.2 Results of VAR Model

The Vector Autoregression Model with 1 lag satisfies the stationarity condition and no root lies outside the unit circle when we check the VAR stability condition as presented in Table 5.2 and Figure 5.3.

Table 5.2 VAR Stability Condition Check

Roots of Characteristic Polynomial

Endogenous Variables: discount abrank listprice retailers avg_cus_review

Exogenous Variables: c hardcover

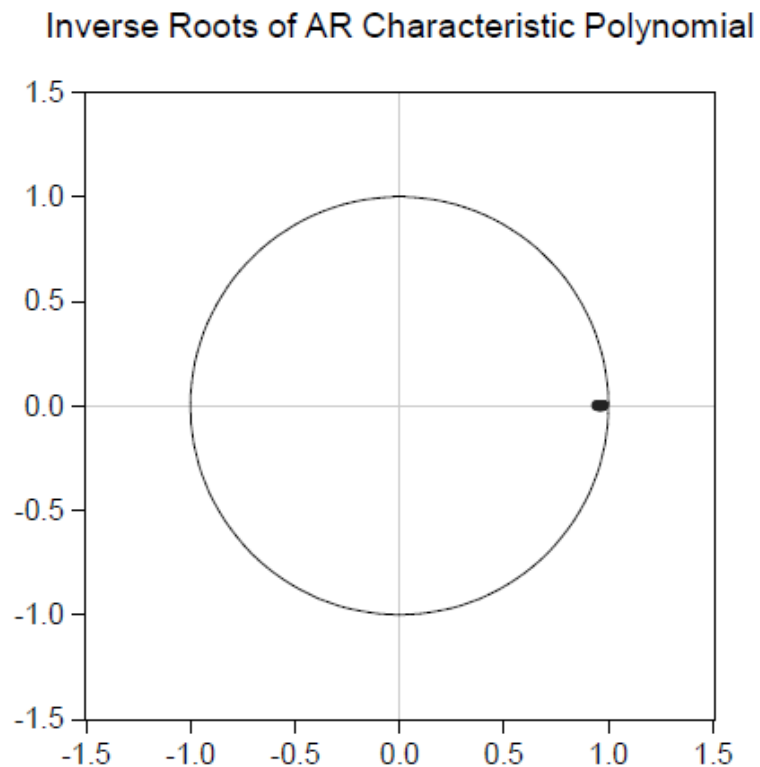
Lag specification: 1 1

Root	Modulus
0.975127	0.975127
0.965916	0.965916
0.958784-0.001003i	0.958785
0.958784+0.001003i	0.958785
0.944851	0.944851

No root lies outside the unit circle.

VAR satisfies the stability condition.

Figure 5.3 Stability of VAR Model



The VAR model explains 0.945925 of the variance in discount, 0.922412 of sales rank, 0.941596 of list price, 0.914517 of number of sellers, and 0.921475 of customer reviews. The AIC and SC are 38.58954 and 38.60681 respectively.

Since the VAR estimates cannot be interpreted on their own, in the next step we will be interpreting Impulse Response function estimates.

5.1.5 Impulse Response Functions

The results of Generalized Impulse Response functions can be seen on Tables 5.3, 5.4, 5.5, and Figures 5.8, 5.6, 5.4, 5.5, 5.7. In these tables we present the same day and cumulative effects, as well as the estimates and standard errors of GIRF on these variables. Cumulative elasticities are computed as the sum of impulse response coefficients.

The GIRF on discount shows that all the variables are significant in explaining this variable. According to these results, higher discounts are applied to books with better sales rank, higher list price, higher customer review, or less number of sellers.

The results also show that sales rank and number of sellers do not have effect on list price at lag 14 at 95% significance level.

The depth of discount does not significantly affect number of sellers at lag 14 at 95% significance level.

As customer review increases, sales rank gets better at lags 1 and 7. However, customer review does not significantly affect sales rank at lag 14. The higher the list prices, the lower the sales rank at lags 1 and 7. As number of sellers increases, the sales rank gets better. Deeper discounts also have impact on decrease in sales rank.

Table 5.3 Same Day and Cumulative Effects on Variables (from GIRF estimates)

DISCOUNT	Response Estimate	Standard Error	Cumulative Elasticity	Z value	p-value
Sales rank (1 day)	-2.52E-03	1.50E-04	-0.005078	-16.77333333	0
Sales rank (7 days)	-2.69E-03	2.50E-04	-0.018343	-10.768	0
Sales rank (14 days)	-2.67E-03	3.80E-04	-0.037206	-7.026315789	1.06026E-12
List price (1 day)	3.65E-03	1.50E-04	0.007114	24.36	0
List price (7 days)	2.61E-03	2.30E-04	2.18E-02	11.36521739	0
List price (14 days)	1.72E-03	3.50E-04	3.63E-02	4.911428571	4.52076E-07
Customer review (1 day)	0.001117	0.00016	0.002278	6.98125	1.46283E-12
Customer review (7 days)	0.001311	0.00025	8.58E-03	5.244	7.85661E-08
Customer review (14 days)	0.001376	0.00038	1.81E-02	3.621052632	0.000146703
Number of sellers (1 day)	-1.39E-03	0.00015	-0.002756	-9.253333333	0
Number of sellers (7 days)	-1.25E-03	0.00026	-9.24E-03	-4.792307692	8.24369E-07
Number of sellers (14 days)	-0.001055	0.00039	-1.72E-02	-2.705128205	0.003413902
LIST PRICE	Response Estimate	Standard Error	Cumulative Elasticity	Z value	p-value
Sales rank (1 day)	2.55E-01	7.36E-02	0.502214	3.459891304	0.000270197
Sales rank (7 days)	2.21E-01	1.21E-01	1.66E+00	1.824139639	0.034065482
Sales rank (14 days)	2.00E-01	1.80E-01	3.11E+00	1.112177531	0.133030913
Discount (1 day)	1.73E+00	7.30E-02	3.363751	23.75808662	0
Discount (7 days)	1.18E+00	1.11E-01	1.01E+01	10.65143551	0
Discount (14 days)	7.16E-01	1.67E-01	1.65E+01	4.281566049	9.27913E-06
Customer review (1 day)	0.551059	0.07355	1.085798	7.492304555	3.38618E-14
Customer review (7 days)	4.57E-01	0.12043	3.52E+00	3.794561156	7.39524E-05
Customer review (14 days)	0.360254	0.18016	6.33E+00	1.999633659	0.022769918
Number of sellers (1 day)	-2.58E-01	7.36E-02	-0.512262	-3.507486413	0.000226181
Number of sellers (7 days)	-2.31E-01	1.24E-01	-1.72E+00	-1.871931671	0.030608029
Number of sellers (14 days)	-0.195916	0.18352	-3.19E+00	-1.067545772	0.142862728

Table 5.4 Same Day and Cumulative Effects on Variables (from GIRF estimates)
cont.

CUSTOMER REVIEW	Response Estimate	Standard Error	Cumulative Elasticity	Z value	p-value
Sales rank (1 day)	-4.21E-02	4.68E-03	-0.082991	-8.988675214	0
Sales rank (7 days)	-3.61E-02	7.46E-03	-0.272594	-4.83847185	6.54206E-07
Sales rank (14 days)	-3.11E-02	1.08E-02	-0.504307	-2.88894052	0.001932711
Discount (1 day)	3.37E-02	4.68E-03	0.069864	7.205555556	2.88991E-13
Discount (7 days)	4.52E-02	6.84E-03	2.80E-01	6.61125731	1.90534E-11
Discount (14 days)	5.15E-02	9.99E-03	6.25E-01	5.150650651	1.29792E-07
List Price (1 day)	0.035072	0.00468	0.068556	7.494017094	3.34177E-14
List Price (7 days)	2.63E-02	0.00683	2.14E-01	3.849633968	5.91472E-05
List Price (14 days)	0.018143	0.01002	3.64E-01	1.810678643	0.035095305
Number of sellers (1 day)	2.34E-01	4.49E-03	0.459317	52.11670379	0
Number of sellers (7 days)	1.86E-01	7.53E-03	1.47E+00	24.73041169	0
Number of sellers (14 days)	0.142124	0.01092	2.59E+00	13.01501832	0
NUMBER OF SELLERS	Response Estimate	Standard Error	Cumulative Elasticity	Z value	p-value
Sales rank (1 day)	-6.15E-01	3.50E-02	-1.23411	-17.60869814	0
Sales rank (7 days)	-6.16E-01	5.50E-02	-4.332037	-11.20472555	0
Sales rank (14 days)	-5.78E-01	7.84E-02	-8.511603	-7.378522205	8.00471E-14
Discount (1 day)	-3.14E-01	3.51E-02	-0.599141	-8.957269099	0
Discount (7 days)	-1.61E-01	5.05E-02	-1.64E+00	-3.192777998	0.000704556
Discount (14 days)	-4.06E-02	7.29E-02	-2.26E+00	-0.557354555	0.288642604
List Price (1 day)	-0.123179	0.03512	-0.24887	-3.507374715	0.000226276
List Price (7 days)	-1.34E-01	0.05044	-9.04E-01	-2.650872324	0.004014209
List Price (14 days)	-0.135469	0.07307	-1.85E+00	-1.853961954	0.031872302
Customer Review (1 day)	1.75E+00	3.37E-02	3.422788	52.08907363	0
Customer Review (7 days)	1.30E+00	5.43E-02	1.06E+01	23.87424047	0
Customer Review (14 days)	0.908342	0.07825	1.81E+01	11.60820447	0

Table 5.5 Same Day and Cumulative Effects on Variables (from GIRF estimates)
cont.

SALES RANK	Response Estimate	Standard Error	Cumulative Elasticity	Z value	p-value
Number of Sellers (1 day)	-7.09E+04	4.03E+03	-140286.6	-17.60734101	0
Number of Sellers (7 days)	-6.13E+04	6.55E+03	-462940.7	-9.359933776	0
Number of Sellers (14 days)	-5.02E+04	9.43E+03	-847094.4	-5.321471674	5.14656E-08
Discount (1 day)	-6.56E+04	4.03E+03	-1.32E+05	-16.28155675	0
Discount (7 days)	-6.99E+04	5.89E+03	-4.77E+05	-11.87503611	0
Discount (14 days)	-6.92E+04	8.60E+03	-9.66E+05	-8.049919415	0
List Price (1 day)	1.40E+04	4047.65	27512.47	3.45975072	0.000270338
List Price (7 days)	1.16E+04	5885.59	8.89E+04	1.962476149	0.024853537
List Price (14 days)	9.83E+03	8631	1.62E+05	1.138983084	0.127355105
Customer Review (1 day)	-3.63E+04	4.04E+03	-69191.85	-8.989975663	0
Customer Review (7 days)	-1.87E+04	6.39E+03	-1.89E+05	-2.926551285	0.001713715
Customer Review (14 days)	-6139.946	9264.57	-2.66E+05	-0.66273405	0.253750449

Figure 5.4 Discount GIRF

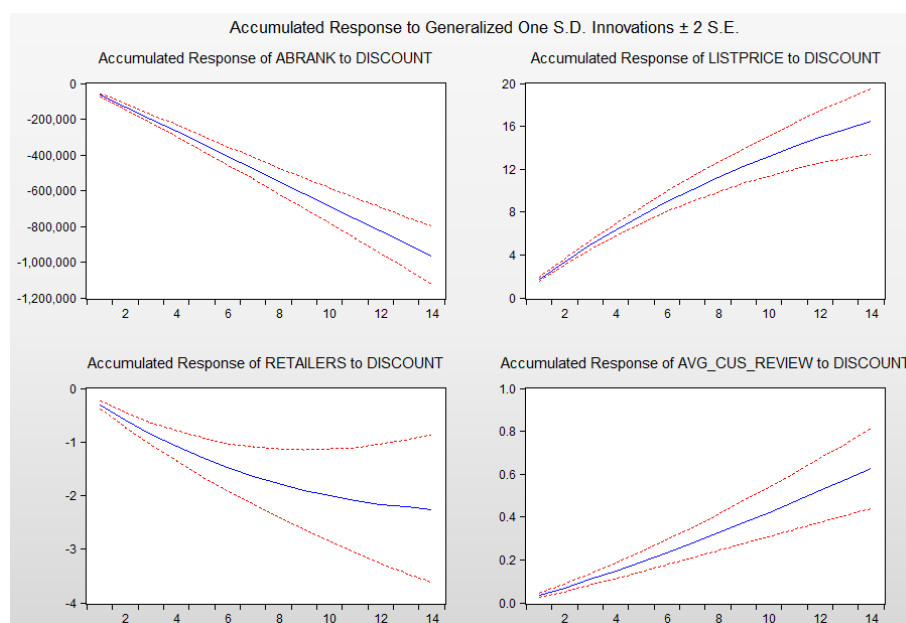


Figure 5.5 List Price GIRF

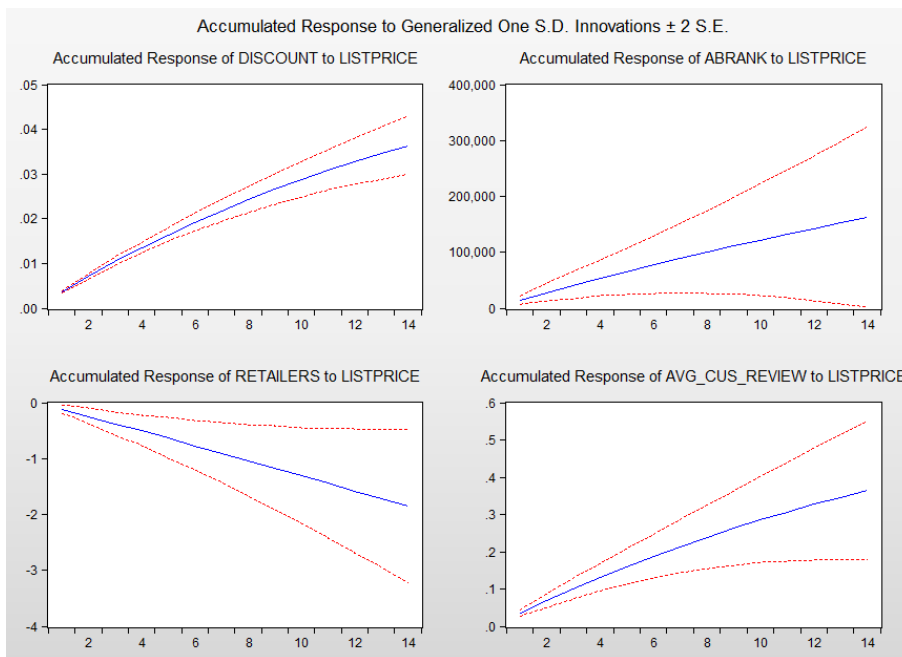


Figure 5.6 Customer Review GIRF

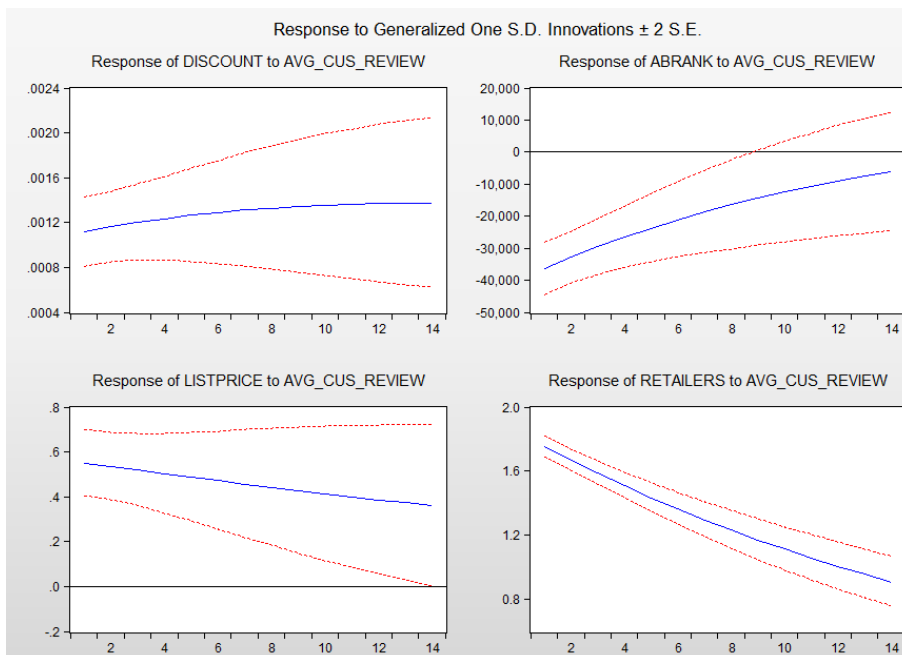


Figure 5.7 Number of Sellers GIRF

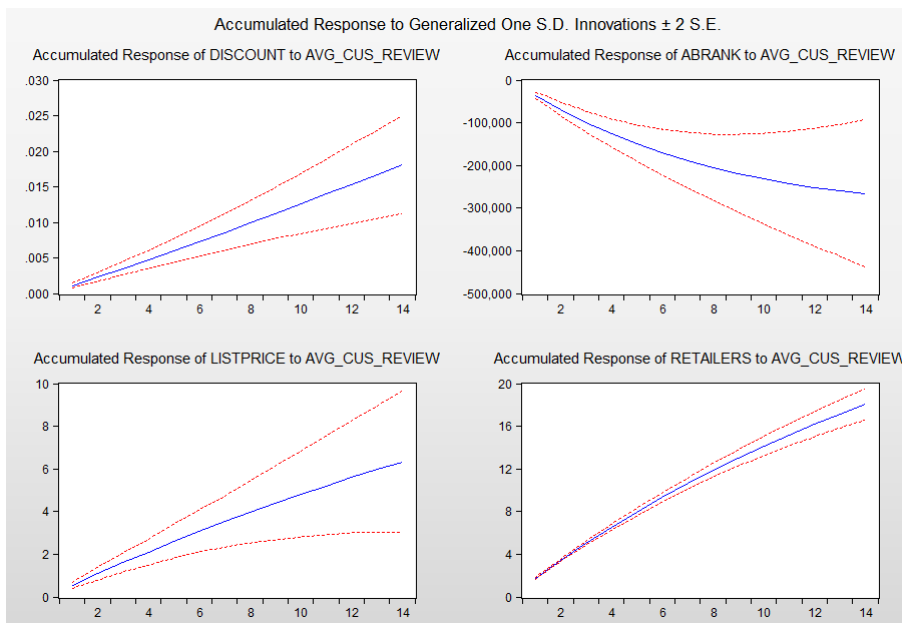
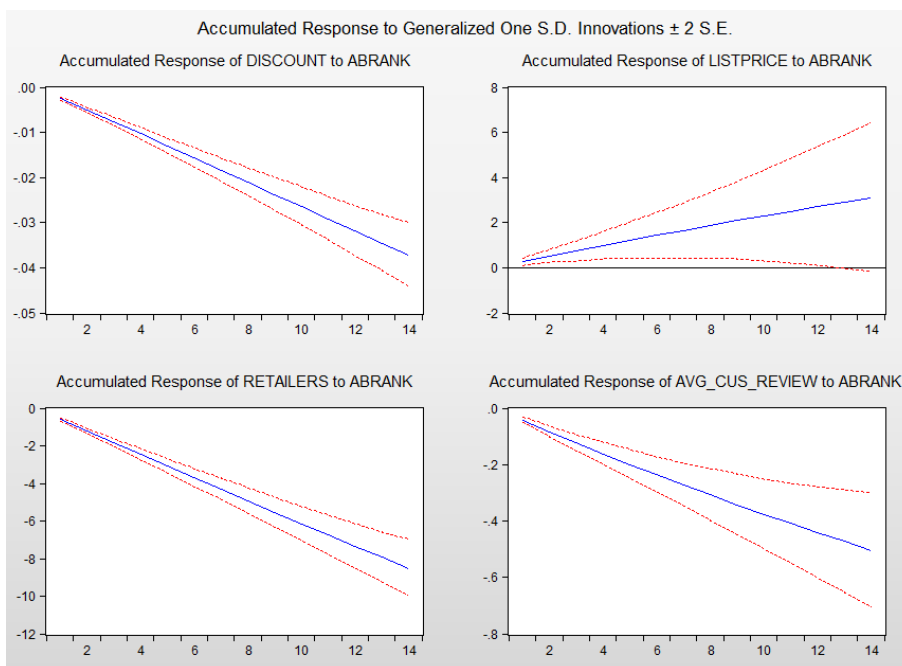


Figure 5.8 Sales Rank GIRF



In the next section, we will follow the same steps as in this section, with the same dataset. The main difference is that we will take this dataset as a panel time series dataset and the differences between the books will also be considered. The methods we will use in the modelling steps are also methods that are tailored for panel data analysis.

5.2 Analysis with Panel Data Adjustment

In this section, we take our dataset as a panel dataset but we do not make any modifications on variables of physical format and number of sellers. In the following section, we will make adjustments on these two variables and present the results.

5.2.1 Unit Root Tests

When we take the dataset as a panel dataset, the unit root tests indicate that according to SBIC, the variables discount, listprice, retailers, avg_cus_review, and ABRank are stationary thus, we do not need to make any changes. The results of common root Levin, Lin, Chu Unit Root Test can be seen in Table 5.6.

Table 5.6 Unit Root Test Results on Panel Dataset

Variable	Statistic	p-value
discount	557.071	0.0019
abrank	-3.74069	0.0001
listprice	-1.74201	0.0408
retailers	-2.97865	0.0014
avg_cus_review	-2.55967	0.0052

5.2.2 Cointegration Test

According to Kao Residual Cointegration test on panel data, p-value is 0.0006, it is less than 0.05, that is we can reject null hypothesis that there is no cointegration at the 95% significance level. There is cointegration which means there is stationarity.

Next step is to do Granger causality tests on these variables. The results of the Cointegration test of our panel data can be seen in Table 5.7.

Table 5.7 Cointegration Test Results on Panel Dataset

Kao Residual Cointegration Test		
Series: discount abrank listprice avg_cus_review retailers		
Sample: 7/01/2011 7/31/2011		
Included observations: 15500		
Null Hypothesis: No cointegration		
Trend assumption: No deterministic trend		
	t-statistic	p-value
ADF	3.227752	0.0006
Residual variance	0.000110	
HAC variance	8.94E-05	

5.2.3 Granger Causality

We look at Granger causality up to 8 lags between the variables. Granger causality tests on our variables show that they contain information that helps predicting other variables in the model. Performing VAR model with these variables would give meaningful results.

Specifically, discount Granger causes sales rank, number of sellers, and customer review at any lag up to 8 lags. Discount is Granger caused by sales rank at lags 3,4,5,6,7; at any lag by retailers; and at lags 6,7,8 by listprice at the 95% significance level. Customer review Granger causes discount at lags 5 and 6 at 90% significance level.

Sales rank Granger causes customer review at any lag but the reverse is not supported. Sales rank also Granger causes number of sellers at any lag but Granger caused by number of sellers at lags 4,5,6,7,8.

List price Granger causes number of sellers at lags 6,7,8. Customer review Granger causes and Granger caused by listprice at lags 4,5,6,7,8.

5.2.4 Panel Vector Autoregression Model

Since the Unit Root, Cointegration, and Granger Causality tests show that the variables have no unit root, and can be used to explain and predict each other, we can continue with our next step, Panel Vector Autoregression Analysis.

5.2.4.1 Model Selection

In order to decide the number of lags to be included in our analysis, we perform model selection. According to the results, taking 6 lags for SC, and 11 lags for AIC are the best ways to perform PVAR analysis. Since we make all our analyses with SC, we take lag 6 for our analysis as Figure 5.9.

Figure 5.9 Model Selection for PVAR model

Lag	LogL	LR	FPE	AIC	SC	HQ
0	-219527.7	NA	1.87e+16	51.65592	51.66421	51.65875
1	-104322.1	230221.4	31736.45	24.55461	24.58363	24.56451
2	-104191.8	260.1772	30960.00	24.52984	24.57959	24.54681
3	-104080.9	221.3293	30340.62	24.50963	24.58011	24.53368
4	-104011.8	137.8513	30027.39	24.49925	24.59046	24.53037
5	-103987.7	48.16889	30033.32	24.49945	24.61138	24.53765
6	-102986.3	1995.253	23868.68	24.26971	24.40237*	24.31498
7	-102969.0	34.45320	23911.96	24.27152	24.42491	24.32386
8	-102945.9	45.83337	23923.04	24.27198	24.44610	24.33140
9	-102927.1	37.35289	23958.08	24.27345	24.46830	24.33994
10	-102906.5	41.00705	23982.74	24.27447	24.49005	24.34804
11	-102603.1	602.7954	22461.91*	24.20896*	24.44527	24.28960*
12	-102597.9	10.24068	22567.03	24.21363	24.47067	24.30134
13	-102577.0	41.43302*	22588.92	24.21460	24.49237	24.30939
14	-102561.0	31.82649	22636.55	24.21670	24.51520	24.31857

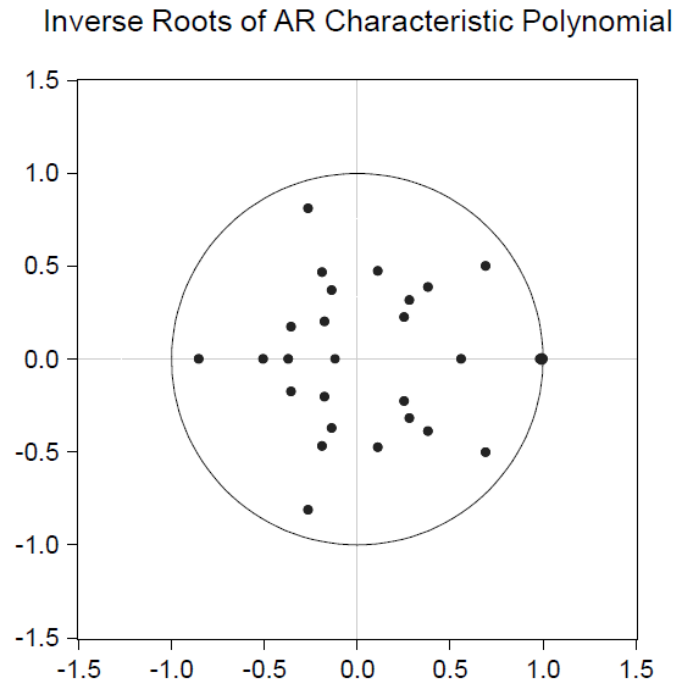
* indicates lag order selected by the criterion
 LR: sequential modified LR test statistic (each test at 5% level)
 FPE: Final prediction error
 AIC: Akaike information criterion
 SC: Schwarz information criterion
 HQ: Hannan-Quinn information criterion

5.2.4.2 Results of the PVAR Model

We perform PVAR analysis with 6 daily lags since it is selected as the optimal. At this lag, PVAR model is stable (stationary), as this can be seen in Figure 5.10. According to this figure, and the results, no root lies outside the unit circle, and

stability condition is satisfied. The graph reports the inverse roots of the characteristic AR polynomial, and explains that the estimated PVAR is stationary if all roots have modulus less than one, and are in the unit circle on the graph.

Figure 5.10 Stability of PVAR Model



At 6 lags, our PVAR model explains 0.983013 of variance in discount, 0.970508 in sales rank, 0.999998 in listprice, 0.984095 in number of sellers, and 0.978647 in customer review. The AIC and SC statistics are 23.67436, and 23.76951 respectively.

When PVAR model is used, we can see a good improvement compared to VAR model. Lower AIC and SC statistics, and better R-squared values are good indicators of this improvement.

5.2.5 Results of the Generalized Impulse Response Functions

The results of Generalized Impulse Response Functions can be seen in Tables 5.8, 5.9, 5.10, and Figures 5.11, 5.12, 5.13, 5.14, 5.15. The standard errors of these functions are computed by Monte Carlo error estimates with 100 repetitions.

Table 5.8 Same Day and Cumulative Effects on Panel Time Series Variables (from GIRF estimates)

DISCOUNT	Response Estimate	Standard Error	Cumulative Elasticity	Z value	p-value
Sales rank (1 day)	-3.05E-05	8.50E-05	-0.00026	-0.358823529	0.359863555
Sales rank (7 days)	-0.000848	0.0002	-0.004299	-4.24	1.1176E-05
Sales rank (14 days)	-0.000854	0.0002	-0.010165	-4.27	9.77365E-06
List price (1 day)	1.05E-03	1.20E-04	0.002116	8.783333333	0
List price (7 days)	5.89E-04	1.70E-04	0.0064	3.464705882	0.000265406
List price (14 days)	7.57E-04	1.30E-04	0.011194	5.823076923	2.8887E-09
Customer review (1 day)	-8.03E-05	0.00011	-0.000253	-0.73	0.232695092
Customer review (7 days)	3.57E-04	0.00022	0.000834	1.622727273	0.052323859
Customer review (14 days)	0.000348	0.00023	0.003328	1.513043478	0.065134308
Number of sellers (1 day)	-3.92E-04	9.10E-05	-0.000876	-4.307692308	8.24833E-06
Number of sellers (7 days)	5.53E-05	0.00022	-0.001445	0.251363636	0.40076649
Number of sellers (14 days)	-0.000276	0.00023	-0.002359	-1.2	0.11506967
LIST PRICE	Response Estimate	Standard Error	Cumulative Elasticity	Z value	p-value
Sales rank (1 day)	3.41E-05	4.70E-04	-0.000343	0.072553191	0.471080838
Sales rank (7 days)	6.46E-04	1.15E-03	0.001197	0.56173913	0.287146885
Sales rank (14 days)	4.67E-04	1.02E-03	0.003919	0.457843137	0.323532571
Discount (1 day)	5.73E-03	6.50E-04	0.011487	8.82	0
Discount (7 days)	3.29E-03	1.07E-03	0.035247	3.071962617	0.001063282
Discount (14 days)	4.26E-03	1.00E-03	0.061948	4.262	1.01303E-05
Customer review (1 day)	-9.89E-05	0.00048	-0.000294	-0.206041667	0.418379182
Customer review (7 days)	-3.39E-03	0.00112	-0.014419	-3.030357143	0.001221324
Customer review (14 days)	-0.003301	0.00099	-0.033072	-3.334343434	0.000427505
Number of sellers (1 day)	-0.000838	0.00052	-0.001555	-1.611538462	0.053531206
Number of sellers (7 days)	-1.23E-04	0.00124	-0.003915	-0.099193548	0.460492299
Number of sellers (14 days)	1.43E-05	0.00105	-0.004394	0.013619048	0.494566954

Table 5.9 Same Day and Cumulative Effects on Panel Time Series Variables (from GIRF estimates)-cont.

CUSTOMER REVIEW	Response Estimate	Standard Error	Cumulative Elasticity	Z value	p-value
Sales rank (1 day)	-2.42E-03	2.85E-03	-0.004426	-0.849473684	0.197808884
Sales rank (7 days)	-6.11E-03	6.17E-03	-0.029923	-0.99076175	0.160900965
Sales rank (14 days)	-1.36E-02	6.72E-03	-0.103644	-2.029613095	0.021197941
List price (1 day)	-5.14E-04	2.51E-03	-0.001663	-0.204780876	0.418871664
List price (7 days)	2.10E-02	5.37E-03	0.108281	3.909497207	4.62442E-05
List price (14 days)	1.45E-02	3.89E-03	0.185654	3.735475578	9.36804E-05
Discount (1 day)	-0.00227	0.00297	-0.007106	-0.764309764	0.222341334
Discount (7 days)	9.56E-03	0.00654	0.019466	1.4617737	0.071901617
Discount (14 days)	0.017528	0.00612	0.114599	2.864052288	0.002091295
Number of sellers (1 day)	5.86E-02	0.00316	0.132328	18.54905063	0
Number of sellers (7 days)	9.73E-02	0.00587	0.580601	16.57103918	0
Number of sellers (14 days)	0.094049	0.00691	1.251711	13.6105644	0
NUMBER OF SELLERS	Response Estimate	Standard Error	Cumulative Elasticity	Z value	p-value
Sales rank (1 day)	-2.83E-03	1.61E-02	-0.020471	-0.175573466	0.430314525
Sales rank (7 days)	-8.34E-02	4.26E-02	-0.246223	-1.954831144	0.025301502
Sales rank (14 days)	-1.88E-01	4.69E-02	-1.26398	-4.003836317	3.11617E-05
List price (1 day)	-2.70E-02	1.67E-02	-0.057507	-1.614217443	0.053240135
List price (7 days)	-9.21E-02	3.62E-02	-0.322129	-2.543839779	0.005482068
List price (14 days)	-6.14E-02	2.82E-02	-0.868833	-2.176620616	0.014754443
Customer review (1 day)	0.363838	0.01928	0.788909	18.87126556	0
Customer review (7 days)	4.13E-01	0.0497	2.922907	8.319738431	0
Customer review (14 days)	0.395475	0.05721	5.762149	6.912690089	2.37776E-12
Discount (1 day)	-6.87E-02	0.01608	-0.161891	-4.272636816	9.65875E-06
Discount (7 days)	2.22E-03	0.04274	-0.309589	0.051848386	0.479324751
Discount (14 days)	0.046238	0.04387	-0.12266	1.053977661	0.145946572

Table 5.10 Same Day and Cumulative Effects on Panel Time Series Variables (from GIRF estimates)-cont.

SALES RANK	Response Estimate	Standard Error	Cumulative Elasticity	Z value	p-value
Discount (1 day)	-875.8686	2436.16	-3054.938	-0.359528356	0.359599934
Discount (7 days)	-7.57E+03	5340.29	-26335.2	-1.418132536	0.078076037
Discount (14 days)	-1.39E+04	5246.13	-105275.7	-2.658313843	0.003926636
List price (1 day)	1.80E+02	2.51E+03	-496.1521	0.071602035	0.47145931
List price (7 days)	-9.52E+01	4.54E+03	-2544.293	-0.020956062	0.491640353
List price (14 days)	-5.07E+02	3.09E+03	-4601.808	-0.163866529	0.434918109
Customer review (1 day)	-2457.613	2892.43	-4693.997	-0.849670692	0.197754099
Customer review (7 days)	-1.99E+01	4968.12	-10897.08	-0.004007152	0.498401382
Customer review (14 days)	5851.014	4877.59	13221.31	1.19957069	0.115153058
Number of sellers (1 day)	-4.63E+02	2636.63	-3208.673	-0.175682747	0.430271595
Number of sellers (7 days)	-7.52E+03	5006.14	-38721.9	-1.502080046	0.066538219
Number of sellers (14 days)	-7135.917	5151.01	-90555.65	-1.385343263	0.082973765

Figure 5.11 Discount Panel GIRF

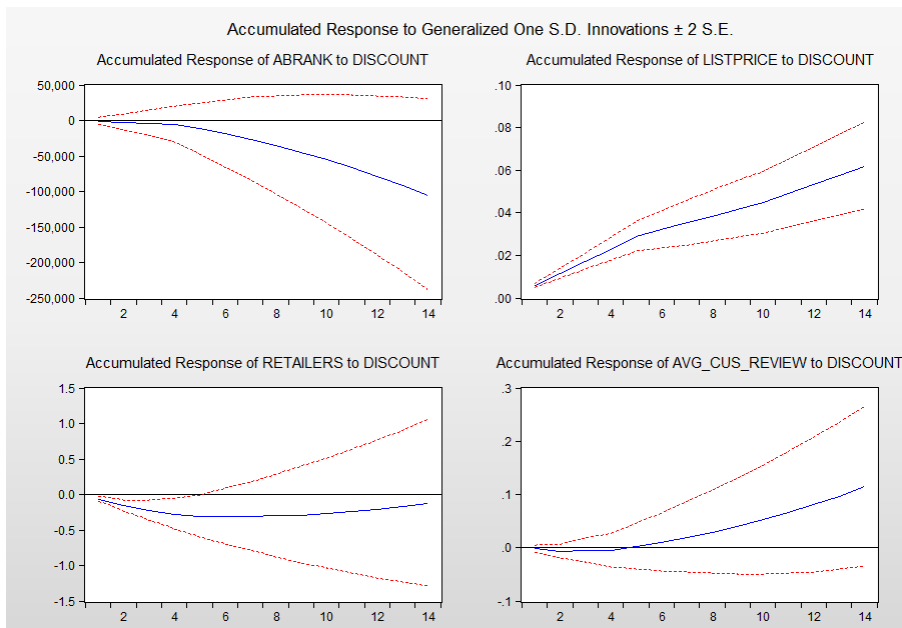


Figure 5.12 List Price Panel GIRF

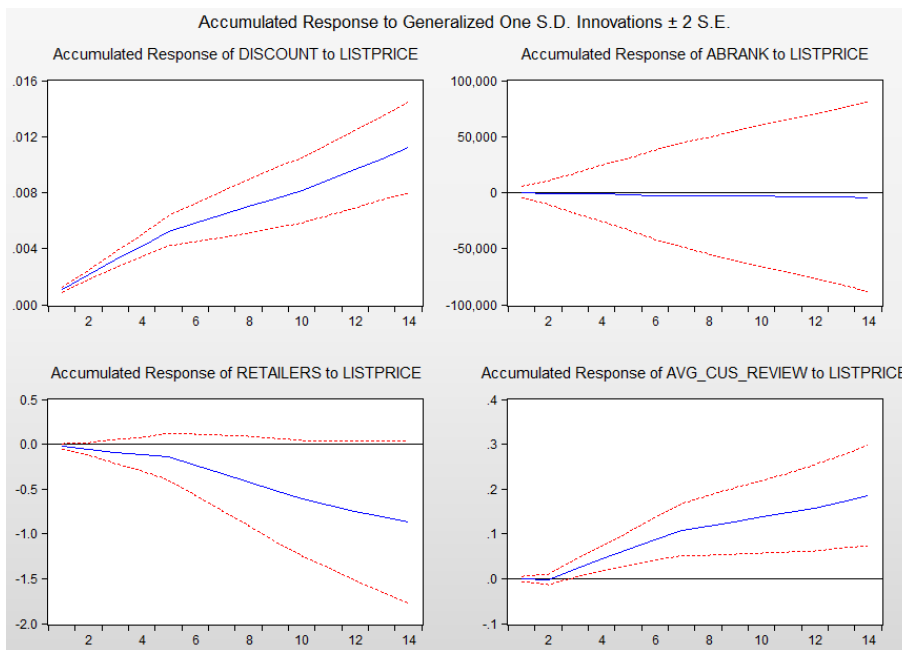


Figure 5.13 Customer Review Panel GIRF

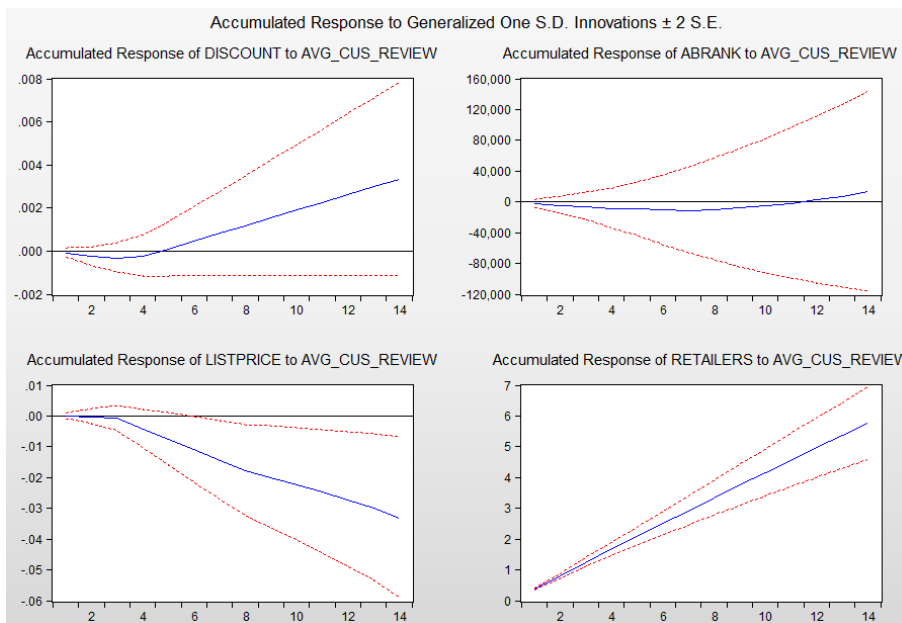


Figure 5.14 Number of Sellers Panel GIRF

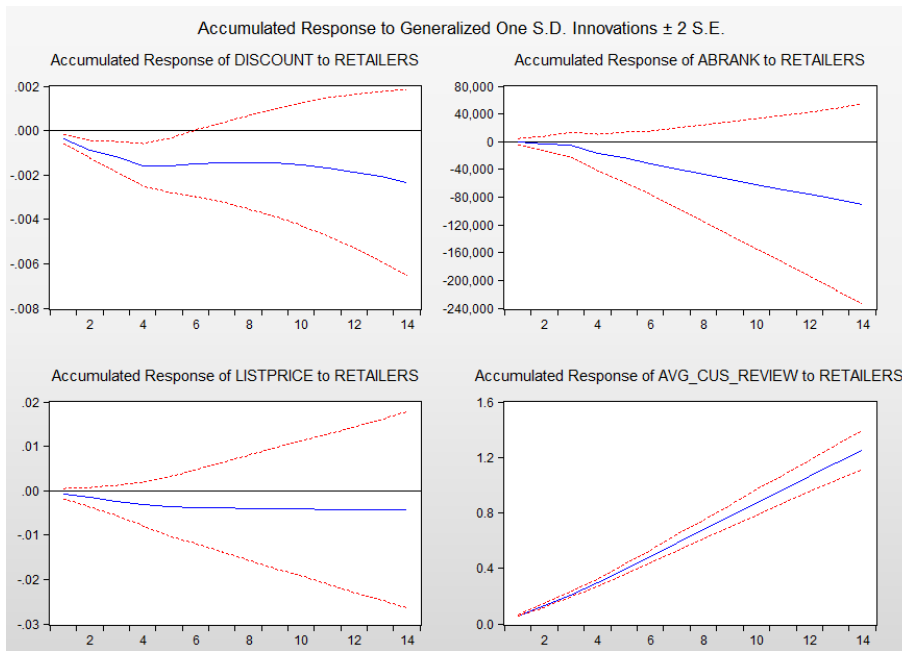
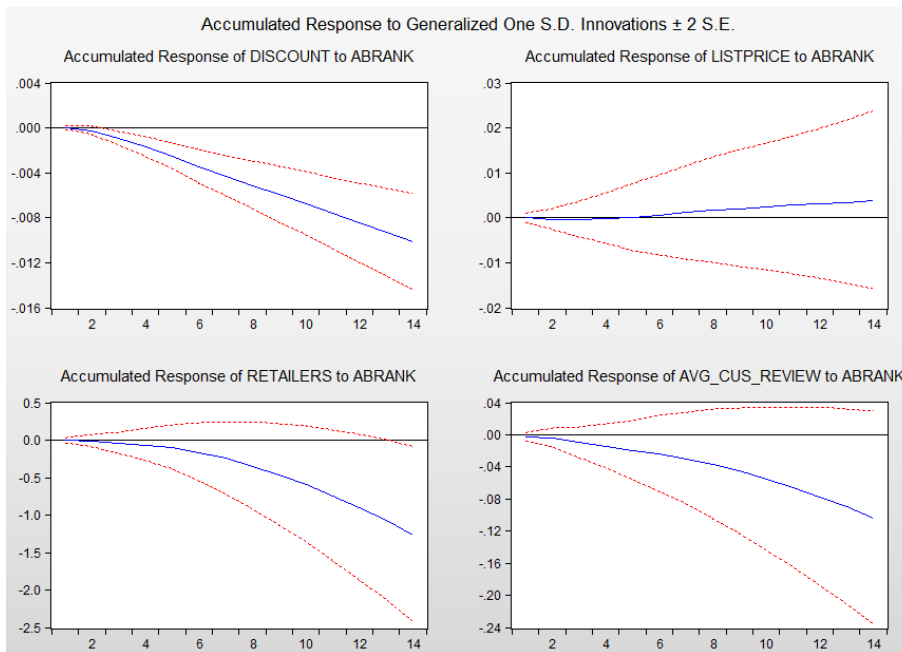


Figure 5.15 Sales Rank Panel GIRF



According to GIRF estimates, as listprice increase, deeper discounts are realized. At lag lengths 7 and 14, sales rank impacts discount, i.e. the better the sales rank, the deeper the discounts.

Lower sales rank affects higher customer reviews, and the higher the list price, the higher customer review at lags 7 and 14. As discounts gets deeper, customer reviews increase at lag 14.

As number of sellers increase, higher customer reviews are observed. In addition, the higher the customer review, the higher the number of sellers. Deeper discounts affect sales ranks to get better, and high number of sellers impacts better sales ranks (lags 7,14).

5.3 Analysis on Panel Data with Modified Variables

In this section, we analyze the panel data with modified physical format and number of sellers variables. We follow the same steps with the previous section and use the methods that are developed for analyzing panel data.

5.3.1 Unit Root Tests

Since there is no change in the variables discount, abrank, listprice, and avg_cus_review, the unit roots of them are as in Table 5.6. Only the retailers variable is modified according to k-means clustering and Elbow methods as Figure 3.1. The unit root test results of this variable is as in Table 5.11.

Table 5.11 Unit Root Test Result of Clustered Number of Sellers Variable

Panel Unit Root Test		
Exogenous Variables: Individual Effects		
Automatic selection of maximum lags		
Method: Levin, Lin & Chu t		
Null: Unit root (assumes common unit root process)		
Variable	Statistic	p-value
numberofsellers	-9.48945	0.0018

From these results, we can understand that there is no unit root on number of sellers variable. We do not need to take difference of this variable, this variable is mean stationary. Next step is to test for cointegration.

5.3.2 Cointegration Test

Kao Residual Cointegration Test Results can be seen in Table 5.12. The null hypothesis is there is no cointegration between these variables.

Table 5.12 Panel Time Series Cointegration Test Results

Kao Residual Cointegration Test		
Series: discount abrank listprice numberofsellers avg_cus_review		
Sample: 7/01/2011 7/31/2011		
Included observations: 15500		
Null Hypothesis: No cointegration		
Trend assumption: No deterministic trend		
	t-statistic	p-value
ADF	4.054788	0.0000
Residual variance	0.00011	
HAC variance	8.96E-05	

At 95% significance level, there is cointegration between these variables, i.e. there is no unit root.

5.3.3 Granger Causality Test

According to the results of Granger causality test up to 8 lags between these variables, our time series variables can be used to predict each other.

On the basis of the results of Granger causality test, discount Granger causes sales rank, number of sellers, and customer review at any lag at 95% significance level.

List price Granger causes discount and number of sellers at lags 6, 7, 8 with 95% significance level. Customer review also Granger caused by list price from lag 3 to 8 at 95% significance level.

Discount is Granger caused by sales rank from lag 3 to 8 with 95% significance level, and at lag 2 with 90% significance level. In addition, sales rank Granger causes number of sellers and customer review at any lag at 95% significance level.

Number of sellers Granger causes discount and customer review at any lag at 95%

significance level. Sales rank is Granger caused by number of sellers at lags 4 and 5 with 95% significance level, and at lags 6 and 8 with 90% significance level.

Customer review Granger causes discount at lags 5 and 6, and number of sellers at lag 8 with 90% significance level. At 95% significance level, it Granger causes number of sellers at lags from 2 to 6, and list price from 4 to 8.

Since these variables drive one another, we can estimate PVAR model of these variables by taking them as endogenous variables of our model.

5.3.4 Panel Vector Autoregression Results

All our variables are mean stationary, and the variables can be used to explain each other according to Unit Root, Cointegration, and Granger Causality tests. In this section, we perform Panel Vector Autoregression by taking our time series variables as endogenous variables and modified format variable as exogenous variable.

5.3.4.1 Model Selection

We decide optimal lag length of the PVAR model as Figure 5.16. According to AIC, the optimal lag length is 11 but according to SC, it is 6. Since we perform our analyses according to SC, we take 6 as the lag length of our PVAR model.

Figure 5.16 Optimal Lag Length Selection

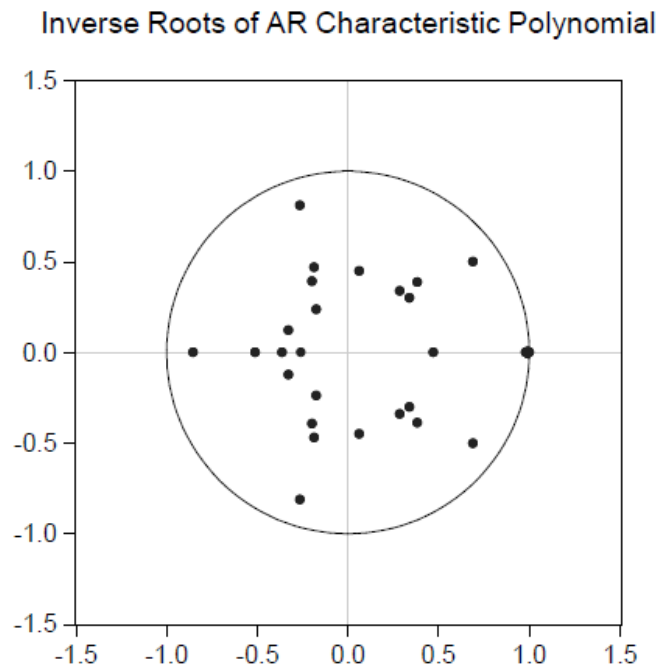
VAR Lag Order Selection Criteria						
Endogenous variables: DISCOUNT ABRANK LISTPRICE NUMBEROFSELLERS AVG_CUS_REVIEW						
Exogenous variables: C FORMAT						
Date: 08/11/20 Time: 21:30						
Sample: 7/01/2011 7/31/2011						
Included observations: 8500						
Lag	LogL	LR	FPE	AIC	SC	HQ
0	-199952.5	NA	1.87e+14	47.05000	47.05829	47.05283
1	-88221.07	223278.9	718.2000	20.76613	20.79515	20.77604
2	-88091.48	258.8144	700.7412	20.74152	20.79127	20.75850
3	-87985.44	211.6484	687.5065	20.72246	20.79293	20.74651
4	-87918.76	133.0256	680.7962	20.71265	20.80386	20.74377
5	-87900.82	35.76193	681.9285	20.71431	20.82625	20.75251
6	-86895.93	2002.211	541.5107	20.48375	20.61641*	20.52902
7	-86883.41	24.91973	543.1039	20.48669	20.64008	20.53903
8	-86859.94	46.72011	543.2988	20.48704	20.66117	20.54646
9	-86841.10	37.46126	544.0876	20.48849	20.68335	20.55499
10	-86822.00	37.97792	544.8428	20.48988	20.70546	20.56345
11	-86515.70	608.4869	509.9486*	20.42369*	20.66000	20.50433*
12	-86510.23	10.85205	512.2980	20.42829	20.68533	20.51600
13	-86488.28	43.55875	512.6656	20.42901	20.70677	20.52379
14	-86467.85	40.52022*	513.2170	20.43008	20.72858	20.53194

* indicates lag order selected by the criterion
 LR: sequential modified LR test statistic (each test at 5% level)
 FPE: Final prediction error
 AIC: Akaike information criterion
 SC: Schwarz information criterion
 HQ: Hannan-Quinn information criterion

5.3.4.2 Results of PVAR Model with Modified Panel Data

Our PVAR model with 6 lags shows stationarity as in Figure 5.17, no unit root lies outside the unit circle.

Figure 5.17 Stationarity and Stability of PVAR Model



Our PVAR model explains 0.983016 of variance in discount, 0.970494 in sales rank, 0.999998 in list price, 0.965501 in number of sellers, and 0.978644 in customer review.

The AIC and SBIC statistics are 19.83525 and 19.93040 respectively. The lower AIC and SC statistics show us that this model is a better fit than the PVAR model without number of sellers and format modification.

In order to interpret our model, we look at GIRF estimates in the next step.

5.3.5 Results of Generalized Impulse Response Functions

We present the results of GIRF estimates in Tables 5.13, 5.14, 5.15 and in Figures 5.18, 5.19, 5.20, 5.21, 5.22. For computing standard errors, Monte Carlo estimates are used.

Table 5.13 Same Day and Cumulative Effects on Panel Time Series Variables (from GIRF estimates)

DISCOUNT	Response Estimate	Standard Error	Cumulative Elasticity	Z value	p-value
Sales rank (1 day)	-2.64E-05	9.10E-05	-0.000255	-2.90E-01	0.385866085
Sales rank (7 days)	-8.35E-04	1.70E-04	-0.004246	-4.91E+00	4.51302E-07
Sales rank (14 days)	-8.35E-04	1.90E-04	-0.010006	-4.39E+00	5.54535E-06
List price (1 day)	1.05E-03	8.80E-05	2.12E-03	1.20E+01	0
List price (7 days)	6.00E-04	1.80E-04	6.44E-03	3.33E+00	0.00042906
List price (14 days)	7.72E-04	1.40E-04	1.13E-02	5.51E+00	1.751E-08
Customer review (1 day)	-6.94E-05	8.90E-05	-2.20E-04	-7.80E-01	0.217761579
Customer review (7 days)	3.96E-04	0.00018	1.03E-03	2.20E+00	0.013903448
Customer review (14 days)	0.000413	0.00018	3.92E-03	2.29E+00	0.010882491
Number of sellers (1 day)	-2.87E-04	9.10E-05	-7.41E-04	-3.15E+00	0.00080567
Number of sellers (7 days)	-3.01E-05	0.0002	-1.52E-03	-1.51E-01	0.440185075
Number of sellers (14 days)	-0.000477	0.00025	-3.57E-03	-1.91E+00	0.028195608
LIST PRICE	Response Estimate	Standard Error	Cumulative Elasticity	Z value	p-value
Sales rank (1 day)	2.54E-05	6.10E-04	-0.000364	4.16E-02	0.483393104
Sales rank (7 days)	6.35E-04	1.18E-03	0.001109	5.38E-01	0.295241721
Sales rank (14 days)	4.40E-04	1.02E-03	0.003705	4.31E-01	0.333098753
Discount (1 day)	5.73E-03	4.80E-04	1.15E-02	1.19E+01	0
Discount (7 days)	3.28E-03	1.09E-03	3.52E-02	3.01E+00	0.001313755
Discount (14 days)	4.25E-03	9.90E-04	6.18E-02	4.29E+00	8.89716E-06
Customer review (1 day)	-9.13E-05	0.00052	-2.58E-04	-1.76E-01	0.430313167
Customer review (7 days)	-3.38E-03	0.00097	-1.43E-02	-3.49E+00	0.000245548
Customer review (14 days)	-0.003257	0.00091	-3.27E-02	-3.58E+00	0.000172376
Number of sellers (1 day)	-6.32E-04	0.00057	-1.18E-03	-1.11E+00	0.133764291
Number of sellers (7 days)	-1.25E-04	0.00122	-3.91E-03	-1.02E-01	0.459196171
Number of sellers (14 days)	0.000115	0.00121	-3.46E-03	9.50E-02	0.462141002

Table 5.14 Same Day and Cumulative Effects on Panel Time Series Variables (from GIRF estimates)-cont.

CUSTOMER REVIEW	Response Estimate	Standard Error	Cumulative Elasticity	Z value	p-value
Sales rank (1 day)	-2.43E-03	2.73E-03	-0.004576	-8.89E-01	0.186998587
Sales rank (7 days)	-6.29E-03	6.54E-03	-0.0306	-9.61E-01	0.168196908
Sales rank (14 days)	-1.40E-02	6.67E-03	-0.106285	-2.09E+00	0.018216844
List price (1 day)	-4.74E-04	2.70E-03	-1.55E-03	-1.76E-01	0.430321561
List price (7 days)	2.11E-02	5.64E-03	1.09E-01	3.74E+00	9.2114E-05
List price (14 days)	1.47E-02	4.19E-03	1.87E-01	3.51E+00	0.000220639
Discount (1 day)	-1.96E-03	2.52E-03	-0.006477	-7.79E-01	0.218116108
Discount (7 days)	1.03E-02	6.57E-03	0.023527	1.57E+00	0.058454066
Discount (14 days)	1.86E-02	6.72E-03	0.125478	2.77E+00	0.002779088
Number of sellers (1 day)	5.10E-02	0.00282	1.18E-01	1.81E+01	0
Number of sellers (7 days)	8.77E-02	0.00584	5.26E-01	1.50E+01	0
Number of sellers (14 days)	0.084241	0.00732	1.13E+00	1.15E+01	0
NUMBER OF SELLERS	Response Estimate	Standard Error	Cumulative Elasticity	Z value	p-value
Sales rank (1 day)	-2.83E-03	2.42E-03	-0.005655	-1.17E+00	0.121033746
Sales rank (7 days)	-1.13E-02	4.72E-03	-0.038865	-2.38E+00	0.008545464
Sales rank (14 days)	-2.19E-02	5.19E-03	-0.161999	-4.21E+00	1.26914E-05
List price (1 day)	-2.97E-03	2.67E-03	-5.23E-03	-1.11E+00	0.132750491
List price (7 days)	-1.11E-02	5.04E-03	-3.39E-02	-2.21E+00	0.013507205
List price (14 days)	-5.72E-03	3.68E-03	-9.11E-02	-1.55E+00	0.059985979
Customer review (1 day)	0.04617	0.00256	9.54E-02	1.80E+01	0
Customer review (7 days)	4.99E-02	0.00538	3.50E-01	9.27E+00	0
Customer review (14 days)	0.047974	0.00559	6.94E-01	8.58E+00	0
Discount (1 day)	-7.34E-03	0.00234	-1.85E-02	-3.14E+00	0.000855398
Discount (7 days)	-1.52E-03	0.00552	-3.98E-02	-2.74E-01	0.391866906
Discount (14 days)	0.00077	0.00486	-4.24E-02	1.58E-01	0.43705654

Table 5.15 Same Day and Cumulative Effects on Panel Time Series Variables (from GIRF estimates)-cont.

SALES RANK	Response Estimate	Standard Error	Cumulative Elasticity	Z value	p-value
Discount (1 day)	-7.58E+02	2.61E+03	-2949.998	-2.90E-01	0.385757421
Discount (7 days)	-7.56E+03	4.26E+03	-26132.29	-1.78E+00	0.037878127
Discount (14 days)	-1.41E+04	4.46E+03	-105762.3	-3.16E+00	0.000799006
List price (1 day)	1.34E+02	3.19E+03	-6.16E+02	4.20E-02	0.483267151
List price (7 days)	-1.62E+02	4.69E+03	-3.03E+03	-3.46E-02	0.486216047
List price (14 days)	-6.22E+02	3.38E+03	-5.58E+03	-1.84E-01	0.426950732
Customer review (1 day)	-2464.642	2769.09	-2.46E+03	-8.90E-01	0.186718216
Customer review (7 days)	-2.51E+02	4335.97	-1.13E+04	-5.80E-02	0.476880851
Customer review (14 days)	5529.483	4645.66	1.09E+04	1.19E+00	0.116974657
Number of sellers (1 day)	-3.17E+03	2703.06	-7.59E+03	-1.17E+00	0.120217255
Number of sellers (7 days)	-9.28E+03	5408.16	-5.28E+04	-1.72E+00	0.043076005
Number of sellers (14 days)	-7749.308	6087.71	-1.12E+05	-1.27E+00	0.10151913

Figure 5.18 GIRF Graph of Discount

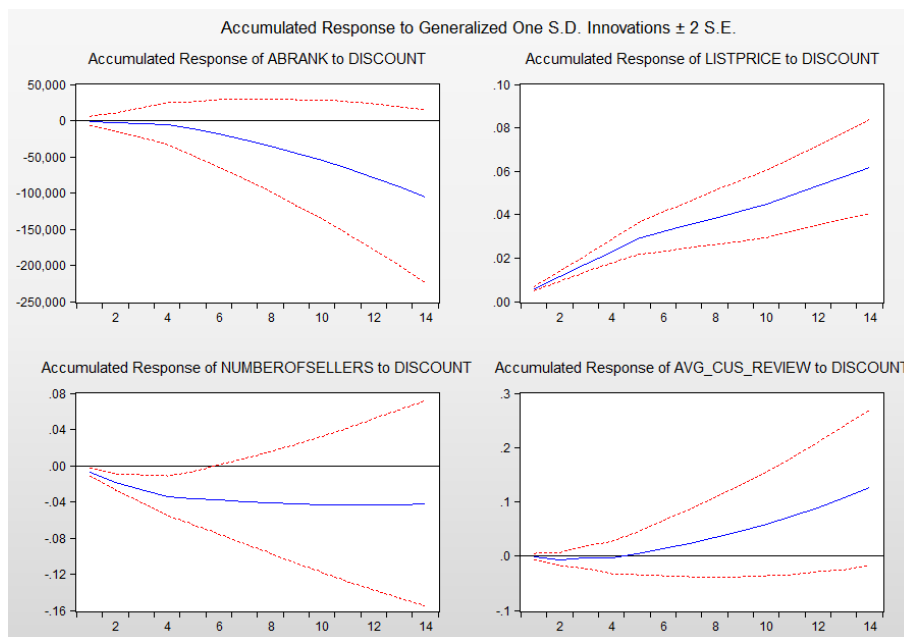


Figure 5.19 GIRF Graph of List Price

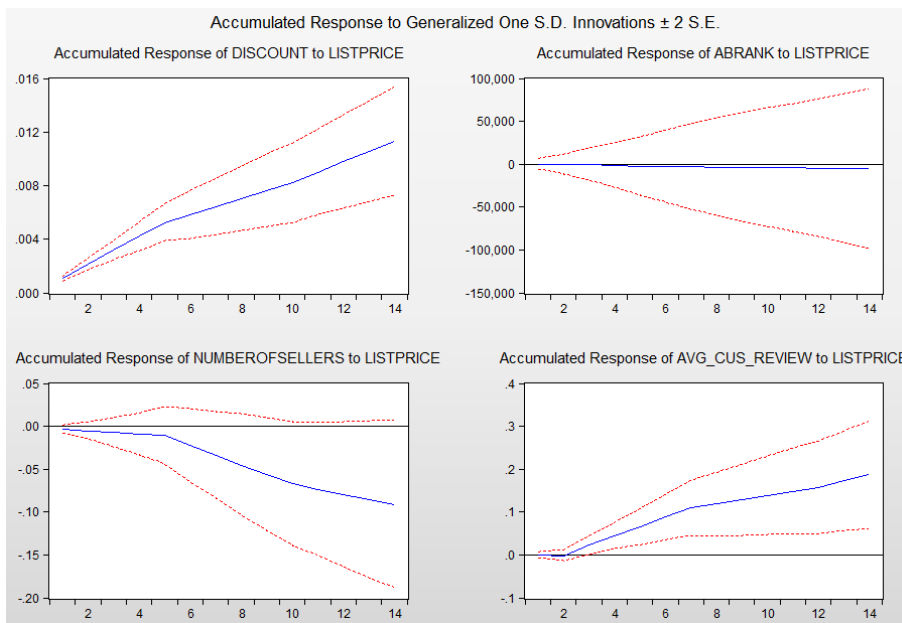


Figure 5.20 GIRF Graph of Customer Review

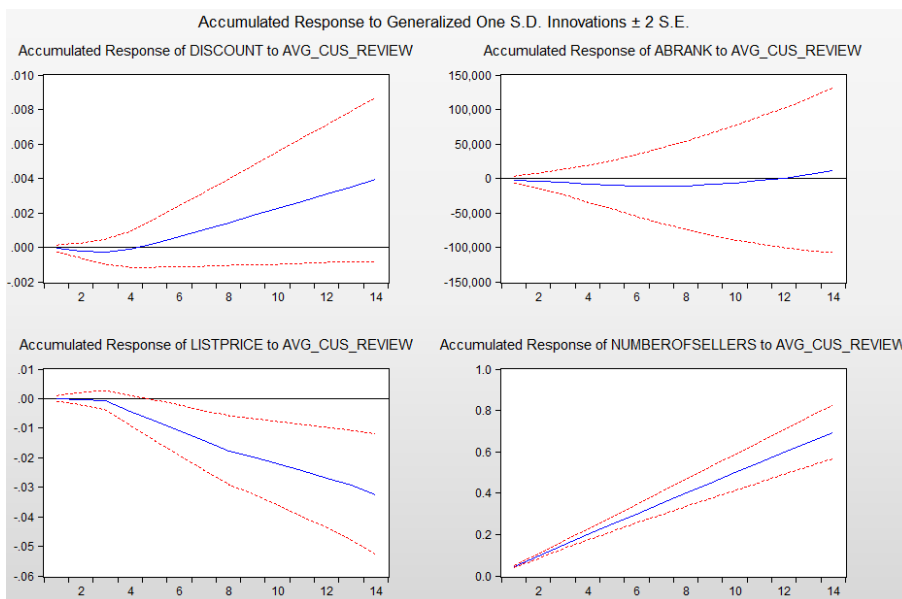


Figure 5.21 GIRF Graph of Number of Sellers

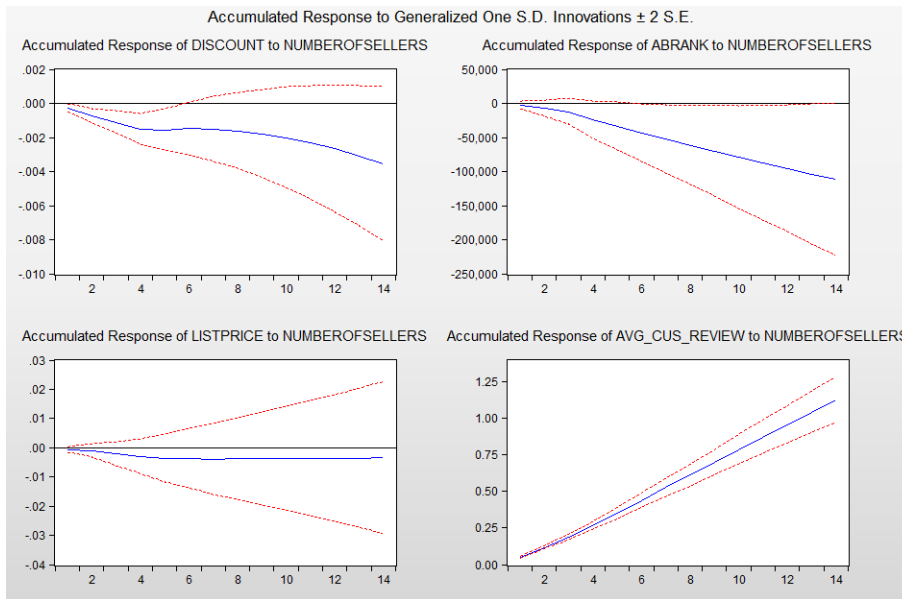
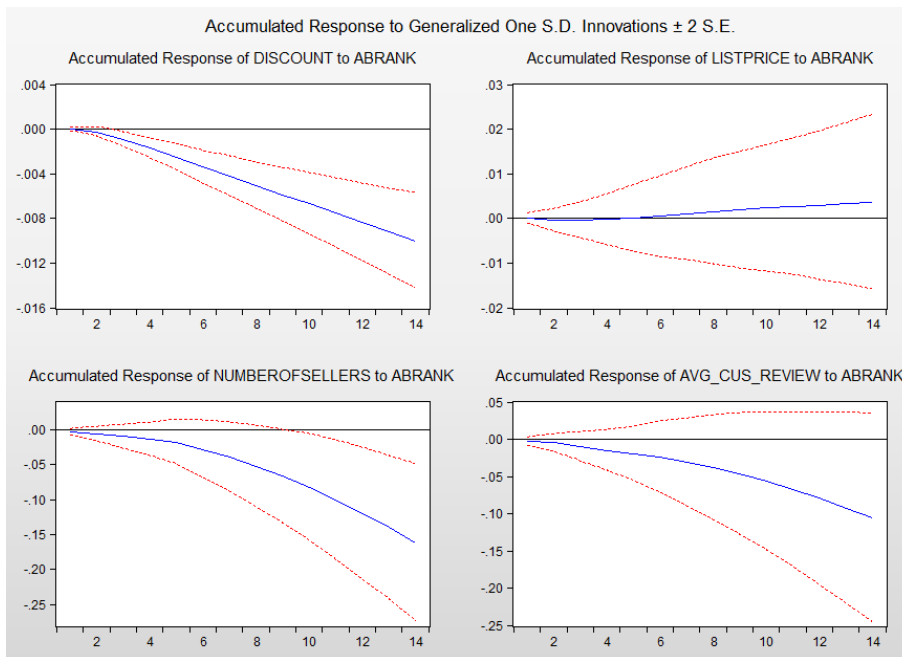


Figure 5.22 GIRF Graph of Sales Rank



Our GIRF results show that the better the sales rank, the deeper the discounts. Previous times' sales ranks impact depth of discounts (lags 7 and 14). In addition, at 95% significance level, the lags 7 and 14 of list prices and customer reviews impacts discounts. The higher the list prices, or the customer reviews, the deeper the discounts. The lower number of sellers also impacts higher discounts.

Sales rank does not significantly affect and affected by list price but the lower the discounts, the lower the list price. This may be because of the pricing and discount

offerings of the publishers.

Discounts affect sales rank, i.e. the higher the discounts, the better the sales ranks. List price and customer review do not significantly affect sales rank. Number of sellers impacts sales rank at lag 7, the higher the number of sellers, the better the sales ranks.

Customer review is affected by list price at lags 7 and 14; by sales rank, and discount at lag 14; and by number of sellers at all lags. At these lags, better sales ranks, higher list prices, deeper discounts, or higher number of sellers have impact on higher customer reviews.

Higher customer reviews also affect higher number of sellers. Furthermore, the lower the list price; the better the sales rank; or the lower the discounts; the higher the number of sellers.

Some of the relationships that are significant for VAR model are not significant for PVAR model. The effect of sales rank on list price, and vice versa; the effect of customer reviews to sales rank are not significant anymore.

5.4 Findings and Discussion

We present the accuracy metrics of our Vector Autoregressive models in Table 5.16. These statistics present evidence that our final model is the closest to the reality since its AIC is the lowest.

Table 5.16 Accuracy Metrics of Vector Autoregressive Models

	VAR model	PVAR model	PVAR model with changed variables
R-squared			
Discount	0.945925	0.983013	0.983016
Sales Rank	0.922412	0.970508	0.970494
List Price	0.941596	0.999998	0.999998
Number of Sellers	0.914517	0.984095	0.965501
Customer Review	0.921475	0.978647	0.978644
AIC	38.58954	23.67436	19.83525
SC	38.60681	23.76951	19.9304

To summarize the findings of the best model, discounts are deeper for books with better sales ranks, higher list prices, higher customer reviews, or lower number of sellers.

Sales rank is better for books with deeper discounts or higher number of sellers. The other variables do not show significant relationship with sales rank.

Customer review is higher for books with deeper discounts, better sales ranks, higher list prices, or higher number of sellers.

Number of sellers is higher for books with lower discounts, better sales ranks, lower list prices, or higher customer reviews.

6. CONCLUSION

In this study, we observed the factors that impact discounting decisions of Amazon.com, and the dynamics in this marketplace. Although it is known that this company uses complex pricing strategies, there are few studies, and limited understanding of dynamics of this marketplace. We observed book market in which a homogeneous product is offered by all the sellers.

The PVAR model we constructed with the dataset containing modified variables explained the relationship the best, and it was the closest to the reality according to AIC and SC when compared with VAR and PVAR models with the datasets containing non-modified variables. We therefore showed that for an unaggregated panel dataset, using models that are tailored for analyzing panel data is crucial in explaining relationships.

Our findings on discounting strategy of Amazon can be summarized as follows:

- Discounts are higher for books with better sales rank. Amazon puts deeper discounts on best seller books.

Gauri et al., 2017 found that when there are deep discounts, consumers buy more from items that can be stored, and profitability decreases. In addition, according to Lal & Matutes, 1994, loss leader items should be items that are difficult to stockpile, such as frequently bought items, items with high storage costs, etc. Best seller books are books that large number of consumers buy but cannot be stockpiled since a book is consumed once. We can infer from our results that Amazon considers this type of product as a good loss leader which increases profits and/or traffic.

- More discounts are put on books with higher list prices.

According to Zhao et al., 2015, consumers' search for best prices can force sellers to set prices at competitive levels. When prices are higher, consumers search for the best alternative, and they are more willing to bear search costs (Stigler, 1961). Since Amazon discounts books with higher list prices more

according to our results, we can infer that Amazon put itself forward in the competition and draw customers to the website with its pricing strategy.

- Deeper discounts are placed when customer reviews are higher.

Consumers do not have the opportunity to examine the physical products and they directly rely on the retailer when shopping online. Therefore, the trustworthiness of the seller is something consumers seek for, and they rest their retailer decisions on quality indicators such as customer reviews (Wang & Li, 2020). By discounting books with higher customer reviews, Amazon gains consumer trust, and drive more traffic into its website.

- When there are lower number of sellers, Amazon puts more discounts.

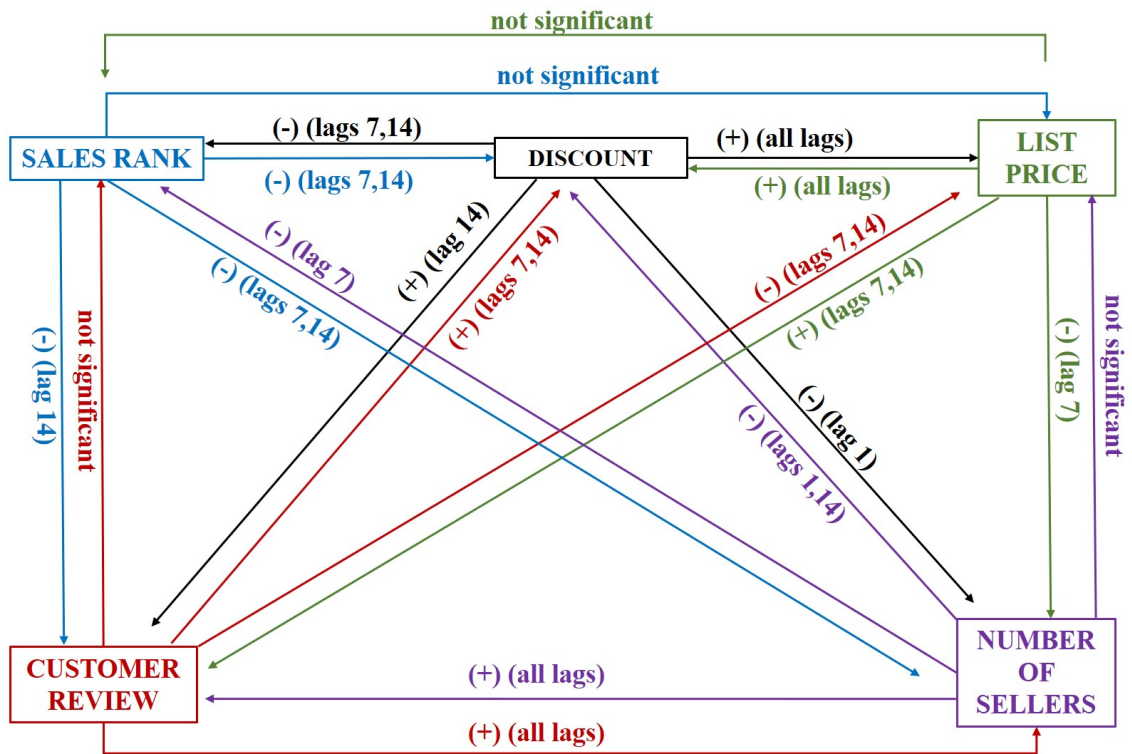
As Stigler, 1961; Wang & Li, 2020 suggest, the search costs of consumers increase when there are more sellers of the same good, and they prefer well-known ones (Wang & Li, 2020). Amazon can rely on its famous brand name when there are lots of sellers and still can attract customers. When consumers can see the alternatives, i.e. when there is lower number of sellers, Amazon discounts more to be the best alternative, according to our results.

Our findings also represent the impacts of these market characteristics within each other. The dynamics can be presented as follows:

- When more discounts are put on a book, the sales rank gets better. In addition, as more retailers sell the book, the sales rank of the book gets better.
- Customer review increases for books with deeper discounts. Better sales rank also affects higher customer reviews. If a book is sold by more sellers, customer reviews get higher. Higher list prices also affect higher customer review.
- The higher number of sellers is observed for books with lower discounts. Better sales ranks impact number of sellers of a book to be higher. Number of sellers is also higher for books with lower list prices. If a book has higher customer reviews, the number of sellers are higher.

The figure that summarizes our findings can be presented as in Figure 6.1.

Figure 6.1 Summary Diagram of the Findings



*The lags in parentheses are the lag lengths where the accumulated responses are significant at 95% significance level.

**If the arrow is from e.g. Discount to e.g. Sales Rank, the Accumulated Response of Discount to Sales Rank is considered.

Our research has focused on explaining dynamics on Amazon marketplace. We hope that this study would be a guide for marketers and researchers who are seeking to understand effects of key market characteristics within each other and pricing decisions in this marketplace.

BIBLIOGRAPHY

- Barrot, C., Becker, J. U., Clement, M., & Papiés, D. (2015). Price elasticities for hardcover and paperback fiction books. *Schmalenbach business review*, 67(1), 73–91.
- Baye, M. R., Morgan, J., & Scholten, P. (2004). Price dispersion in the small and in the large: Evidence from an internet price comparison site. *The Journal of Industrial Economics*, 52(4), 463–496.
- Bodoh and Boehnke and Hickman (2017). Using Machine Learning to Explain Violations of the Law of One Price. Available at SSRN: <https://ssrn.com/abstract=3033324>.
- Breitung, J. (2001). *The local power of some unit root tests for panel data*. Emerald Group Publishing Limited.
- Choi, I. (2001). Unit root tests for panel data. *Journal of international money and Finance*, 20(2), 249–272.
- Clay, K., Krishnan, R., & Wolff, E. (2001). Prices and price dispersion on the web: evidence from the online book industry. *The Journal of Industrial Economics*, 49(4), 521–539.
- CNBC (2019). Amazon.com Inc (AMZN:NASDAQ) Income Statement. <https://www.cnbc.com/quotes/?symbol=AMZN&qsearchterm=am&tab=financials>, Last accessed on August,23 2020.
- Debter, L. (2019). Amazon Surpasses Walmart As the World’s Largest Retailer. <https://www.forbes.com/sites/laurendebter/2019/05/15/worlds-largest-retailers-2019-amazon-walmart-alibaba/#28f1838a4171>, Last accessed on August,21 2020.
- Dhar, R., Huber, J., & Khan, U. (2007). The shopping momentum effect. *Journal of Marketing Research*, 44(3), 370–378.
- Engle, R. & Granger, C. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 55(2), 251–276.
- Gauri, D. K., Ratchford, B., Pancras, J., & Talukdar, D. (2017). An empirical analysis of the impact of promotional discounts on store performance. *Journal of Retailing*, 93(3), 283–303.
- Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, 37(3), 424–438.
- Hayashida and Hoshino (2020). Cross-category Sales Maximization in the Supermarket Industry: Identification from a Natural Experiment in a Loss Leader Category. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3615539, Last accessed on August,23 2020.
- Hess, J. D. & Gerstner, E. (1987). Loss leader pricing and rain check policy. *Marketing Science*, 6(4), 358–374.
- Holtz-Eakin, D., Newey, W., & Rosen, H. S. (1988). Estimating vector autoregressions with panel data. *Econometrica: journal of the Econometric Society*, 56(6), 1371–1395.
- Im, K. S., Pesaran, M. H., & Shin, Y. (2003). Testing for unit roots in heterogeneous panels. *Journal of econometrics*, 115(1), 53–74.

- Inman, J. J., Winer, R. S., & Ferraro, R. (2009). The interplay among category characteristics, customer characteristics, and customer activities on in-store decision making. *Journal of Marketing*, *73*(5), 19–29.
- Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica: journal of the Econometric Society*, *59*(6), 1551–1580.
- Kao, C. (1999). Spurious regression and residual-based tests for cointegration in panel data. *Journal of Econometrics*, *90*(1), 1–44.
- Kocas, C., Pauwels, K., & Bohlmann, J. D. (2018). Pricing best sellers and traffic generators: the role of asymmetric cross-selling. *Journal of Interactive Marketing*, *41*, 28–43.
- Koelemeijer, K. & Opperwal, H. (1999). Assessing the effects of assortment and ambience: a choice experimental approach. *Journal of Retailing*, *75*(3), 319–345.
- Kopalle, P., Biswas, D., Chintagunta, P. K., Fan, J., Pauwels, K., Ratchford, B. T., & Sills, J. A. (2009). Retailer pricing and competitive effects. *Journal of retailing*, *85*(1), 56–70.
- Kotler, P. & Armstrong, G. (2016). *Principles of Marketing*. Pearson Education Limited.
- Kwiatkowski, P. C. B., Phillips, P. S., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, *54*, 159–178.
- Lal, R. & Matutes, C. (1994). Retail pricing and advertising strategies. *Journal of business*, *67*(3), 345–370.
- Lee, L. & Ariely, D. (2006). Shopping goals, goal concreteness, and conditional promotions. *Journal of Consumer Research*, *33*(1), 60–70.
- Levin, A., Lin, C. F., & Chu, C. S. J. (2002). Unit root tests in panel data: asymptotic and finite-sample properties. *Journal of econometrics*, *108*(1), 1–24.
- Li, H., Tang, F. F., Huang, L., & Song, F. (2009). A longitudinal study on australian online dvd pricing. *Journal of Product and Brand Management*, *18*(1), 60–67.
- Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science and Business Media.
- Maddala, G. S. & Wu, S. (1999). A comparative study of unit root tests with panel data and a new simple test. *Oxford Bulletin of Economics and statistics*, *61*(S1), 631–652.
- Pan, X., Ratchford, B. T., & Shankar, V. (2004). Price dispersion on the internet: a review and directions for future research. *Journal of Interactive Marketing*, *18*(4), 116–135.
- Pedroni, P. (1999). Critical values for cointegration tests in heterogeneous panels with multiple regressors. *Oxford Bulletin of Economics and statistics*, *61*(S1), 653–670.
- Perron, P. (1988). Trends and random walks in macroeconomic time series. *Journal of Economic Dynamics and Control*, *12*, 297–332.
- Pesaran, H. H. & Shin, Y. (1998). Generalized impulse response analysis in linear multivariate models. *Economic letters*, *58*(1), 17–29.
- Said, S. E. & Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, *71*, 599–607.

- Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48(1), 1–48.
- Stigler, G. (1961). The economics of information. *Journal of Political Economy*, 69(3), 213–225.
- Stiglitz, J. E. (1987). Competition and the number of firms in a market: Are duopolies more competitive than atomistic markets? *Journal of political economy*, 95(5), 1041–1061.
- Stilley, K. M., Inman, J. J., & Wakefield, K. L. (2010). Planning to make unplanned purchases? the role of in-store slack in budget deviation. *Journal of consumer research*, 37(2), 264–278.
- Thomas, L. (2019). How online stores trick you into impulse buying. <https://www.futurity.org/impulse-buying-2047282/>, Last accessed on August,21 2020.
- Trusov, M., Bucklin, R. E., & Pauwels, K. (2009). Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site. *Journal of marketing*, 73(5), 90–102.
- Walters, R. & Jamil, M. (2002). Measuring cross-category specials purchasing: theory, empirical results, and implications. *Journal of Market-Focused Management*, 5(1), 25–42.
- Wang, W. & Li, F. (2020). What determines online transaction price dispersion? evidence from the largest online platform in china. *Electronic Commerce Research and Applications*, 42.
- Zhao, K., Zhao, X., & Deng, J. (2015). Online price dispersion revisited: How do transaction prices differ from listing prices? *Journal of Management Information Systems*, 32(1), 6261–290.
- Zhu, F. & Liu, Q. (2018). Competing with complementors: An empirical look at amazon.com. *Strategic Management Journal*, 39(10), 2618–2642.