

Carton and volume forecasting from picking lines

Jason Samuels



Thesis presented in partial fulfillment of the requirements for the degree of
MComm (Operations Research)
Department of Logistics Stellenbosch University, South Africa

Supervisor: Prof. SE Visagie

Date: March 2017

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

March 2017

Abstract

Supply chains consist of many stages and all these stages need to be managed. Being able to predict stock flow at any stage can be cost effective for the whole supply chain. In this thesis data from the Pepstores Ltd (PEP) distribution centre in Kuilsrivier, Cape Town are used to predict number of cartons and volume of stock that a hub in their supply chain owned by Pepkor logistics (PKL) will receive. These forecasts will help PKL to schedule delivery trucks and routes to stores with more accurate data and thus lower transportation costs.

Simple linear regression (SLR) and multiple linear regression (MLR) are used to predict cartons and volume, but heteroscedasticity is obtained in the residuals. Different types of transformations on the SLR model are introduced and used on dependent and independent variables. A logarithmic weighted transformation could overcome these problems and is thus used along with polynomial regression to predict the number of cartons and volume of stock. The carton prediction model uses a polynomial regression model with order 2 and the volume prediction model uses a SLR model on the logarithmic weighted variables. Accuracy tests show that the models predict the number of cartons and volumes of stock well. A case study on actual data to forecast volume and cartons is presented. These predictions were then compared to the actual values and the forecast that was sent to the hub from the DC over a two week period. It is concluded that PEP can use these models within their systems, but coefficients need to be reviewed periodically in order to take into account the different types of products.

Opsomming

Voorsieningskettings bestaan uit verskeie stappe en elke stap moet bestuur word. Die vermoë om voorraadvloei te kan bepaal by enige stap kan voordelig wees vir die hele voorsieningsketting. In hierdie tesis word data van Pepstores Bpk (PEP) se verspreidingsentrum in Kuilsrivier, Kaapstad gebruik om die aantal kartonne en die volume voorraad te voorspel wat na 'n Pepkor Logistics (PKL) verspreidingsentrum gestuur word. Hierdie voorspellings sal vir PKL help om roetes na takke met akkurater data te bepaal en dus kostes te bespaar.

Eenvoudige lineêre regressie en meervoudige lineêre regressie word gebruik om die kartonne en volume te bepaal, maar heteroskedastisiteit kom voor in die residue. Verskillende tipe transformasies word voorgestel op die eenvoudige lineêre regressiemodel en word gebruik op die afhanklike en onafhanklike veranderlikes. 'n Logaritmiëse geweege transformasie saam met polinomiëse regressie word gebruik om die aantal kartonne en volume voorraad te voorspel. Die kartonvoorspellingsmodel gebruik 'n polinomiëse regressie van orde 2 terwyl die volumevoorspellingsmodel 'n eenvoudige lineêre regressiemodel gebruik op die logaritmiëse geweege veranderlikes gebruik. Die akkuraatheid van vooruitskattings word getoets en die modelle voorspel die aantal kartonne en volume goed. 'n Gevallestudie met werklike data wat volumes en kartonne voorspel word aangebied. Die voorspellings is dan vergelyk met die werklike waardes en die voorspellings wat gestuur was van die PEP verspreidingsentrum na die sekondêre verspreidingsentrum. Daar word bevind dat PEP hierdie modelle kan gebruik in hulle stelsels, maar dat die koëffisiënte van die model moet gereeld hersien word om die veranderende verkoopspatrone van produkte in ag te neem.

Contents

1	Introduction	1
1.1	Supply chain	1
1.2	Hub-and-spoke supply chain networks	2
1.3	Distribution centers	4
1.3.1	Receiving	4
1.3.2	Storage	4
1.3.3	Order picking	5
1.3.4	Sorting	6
1.3.5	Shipping	6
1.4	Crossdocking	6
1.5	Industry background	7
1.6	Problem description	9
1.7	Thesis objectives	10
1.8	Thesis layout and organisation	10
2	Simple linear regression	13
2.1	Literature review	13
2.1.1	Background	13
2.1.2	Ordinary least squares method	14
2.1.3	Regression through the origin	19
2.2	Model assumptions	20
2.3	Model	20
2.4	Data	21
2.5	Results for the training	22
2.5.1	SLR results part 1	22
2.5.2	SLR results part 2	25
2.6	Conclusion	29

3	Multiple linear regression	31
3.1	Literature review	31
3.2	Models	32
3.2.1	Business unit model	32
3.2.2	Category model	35
3.2.3	A SKU count model	36
3.2.4	Clothing indicator model	39
3.2.5	Combined variable model	41
3.3	Conclusion	44
4	Transformation analysis	45
4.1	Literature on transformed models	45
4.1.1	Log-linear transformation	45
4.1.2	Linear-log transformation	46
4.1.3	Log-log transformation	47
4.1.4	Quadratic transformation	47
4.2	Logarithmic transformation	49
4.2.1	Log-linear model	49
4.2.2	Linear-log model	50
4.2.3	Log-log model	51
4.3	Quadratic transformation	53
4.4	Weighted regression model	55
4.5	Conclusion	57
5	Polynomial regression	59
5.1	Literature review	59
5.2	Model	60
5.3	Results	61
5.3.1	PRM1	61
5.3.2	PRM2	64
5.4	Conclusion	66
6	Carton and volume forecasting	69
6.1	Carton forecasting	70
6.2	Volume forecasting	72
6.3	Conclusion	74

7 Case study	77
7.1 Carton and volume forecast against actuals	77
7.2 Conclusion	79
8 Conclusion	81
8.1 Recommendations to PEP	82
8.2 Objectives achieved	82
8.3 Further studies and recommendations	82
A SLR analysis plots	89
A.0.1 Results part 1	89
A.0.2 Results part 2	89
B MLR analysis	95
B.1 MLR coefficient statistics	95
B.2 MLR regression statistics	102
B.3 MLR residual plots	105
C Heteroscedasticity test results	113
D Quadratic transformation plots and BP test results	115
D.1 BP test results	115
D.2 Transformation plots	117
E Polynomial plots	123
F Forecasting accuracy	127

List of Figures

1.1	Graphical representation of different parties involved in a typical supply chain.	1
1.2	A network with one supply node coloured in black and eight demand nodes with arrows directing the flow of stock.	2
1.3	Schematic representation of two hub-and-spoke supply chain networks.	3
1.4	The flow of stock between departments in a distribution center.	4
1.5	Two types of storage areas that may be used for order picking.	5
1.6	A schematic representation of the layout of the Durban DC owned by PEP.	7
1.7	A photo of the large storage area from the PEP Durban DC.	8
1.8	A photo of the pickers picking stock from bays on either side of the conveyor belt in the picking line.	8
1.9	A schematic representation of the layout of the hub in Kuilsrivier, Cape Town.	9
2.1	Examples of histogram and corresponding Q-Q plots.	16
2.2	Graphical representation of the residuals in regression and include a total and regression part.	17
2.3	An example of a Cook's distance plot that displays Cook's distance values for each observation.	19
2.4	Scatter plots showing the relationship between dependent and independent variables.	23
2.5	Residual histogram and Q-Q plot for Model 1 with units as independent variable and volume as dependent variable.	24
2.6	Fitted and standardised residual plots for the model with units as independent variable and volume as dependent variable.	25
2.7	Cook's distance plots for the 4 models before observations were removed.	26
2.8	Cook's distance plots for the 4 models after observations were removed.	27
2.9	Fitted and studentised residual plots for the model 1 after influential observations were removed.	28
2.10	Residual histogram and Q-Q plot for model 1 after influential observations were removed.	29
3.1	Studentised residual plot for the BU MLR model for BVV that is indicating heteroscedasticity in the residuals. The variance of the studentised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.	34
3.2	Studentised residual plot for the Category MLR model for sub model 1, VolVol that is indicating heteroscedasticity in the residuals. The variance of the studentised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.	36

3.3	Studendised residual plots for the models BSVV and CSVV indicates heteroscedasticity in the residuals. The variance of the studendised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.	38
3.4	Studendised residual plot for the BU and Category Clothing indicator MLR model for sub model 1, VolVol that is indicating heteroscedasticity in the residuals. The variance of the studendised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.	41
3.5	Studendised residual plots for the models BCVV and CCVV that is indicating heteroscedasticity in the residuals with a funnel shape to the right. The variance of the studendised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.	43
4.1	Quadratic equation plots with different signs for b and a	48
4.2	Scatter plots between assigned volume/units and predicted volume/cartons with a natural logarithmic for the log-linear transformation.	50
4.3	Scatter plots between assigned volume/units and predicted volume/cartons with a natural logarithmic for the linear-log transformation.	51
4.4	Scatter plots between assigned volume/units and actual volume/cartons with a natural logarithmic for the log-log transformation.	52
4.5	Scatter plots between assigned volume/units and predicted volume/cartons with a squared-linear transformation.	54
4.6	Scatter plots for the six weighted models. The top row represent the weighted model without any transformations, plots (c) and (d) are the weighted models with a natural logarithm and plots (e) and (f) is the weighted model with a fourth root transformation.	56
5.1	Regression analysis graphs for PRM1 of order 1 (SLR model)	63
5.2	Regression analysis graphs for PRM1 of order 2	64
5.3	Regression analysis graphs for PRM2 of order 1	66
6.1	Bar chart of the actual versus predicted cartons for model PRM1 of order 2 per day from 16 July 2014 until 25 July 2014.	71
6.2	Bar chart of the actual versus forecasted volume measured in cubic meters (m^3) for the M2 polynomial model of order 1 (SLR) per day from 16 July 2014 until 25 July 2014.	74
A.1	Residual histogram and Q-Q plots for the model with volume independent variables and volume as dependent variable.	89
A.2	Residual histogram and Q-Q plots for the model with units independent variables and cartons as dependent variable.	90
A.3	Residual histogram and Q-Q plots for the model with volumes independent variables and cartons as dependent variable.	90
A.4	Fitted and studendised residual plots for the SLR model 2.	90
A.5	Fitted and studendised residual plots for the SLR model 3.	91

A.6	Fitted and studendised residual plots for the SLR model 4.	91
A.7	Residual histogram and Q-Q plots for model 2 after influential observations.	91
A.8	Residual histogram and Q-Q plots for model 3 after influential observations.	92
A.9	Residual histogram and Q-Q plots for model 4 after influential observations.	92
A.10	Fitted and studendised residual plots for the SLR model 2 after influential observations were removed.	92
A.11	Fitted and studendised residual plots for the SLR model 3 after influential observations were removed.	93
A.12	Fitted and studendised residual plots for the SLR model 4 after influential observations were removed.	93
B.1	The studendised residual plots for the BU MLR model for sub models 2, 3 and 4 (BVU, BCV, BCU), that is indicating heteroscedasticity in the residuals. The variance of the studendised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.	105
B.2	The studendised residual plots for the Category MLR model for sub models 2, 3 and 4 (CVU, CCV, CCU), that is indicating heteroscedasticity in the residuals. The variance of the studendised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.	106
B.3	The studendised residual plots for the BU Sku count MLR model for sub models 2, 3 and 4 (BSVU, BSCV, BSCU), that is indicating heteroscedasticity in the residuals. The variance of the studendised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.	107
B.4	The studendised residual plots for the Categories Sku count MLR model for sub models 2, 3 and 4 (CSVU, CSCV, CSCU), that is indicating heteroscedasticity in the residuals. The variance of the studendised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.	108
B.5	The studendised residual plots for the BU Clothing indicator MLR model for sub models 2, 3 and 4 (BCIVU, BCICV, BCICU), that is indicating heteroscedasticity in the residuals. The variance of the studendised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.	109
B.6	The studendised residual plots for the Categories Clothing indicator MLR model for sub models 2, 3 and 4 (CCIVU, CCICV, CCICU), that is indicating heteroscedasticity in the residuals. The variance of the studendised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.	110
B.7	The studendised residual plots for the BU Combined variable MLR model for sub models 2, 3 and 4 (BCVU, BCCV, BCCU), that is indicating heteroscedasticity in the residuals. The variance of the studendised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.	111
B.8	The studendised residual plots for the Categories Combined variable MLR model for sub models 2, 3 and 4 (CCVU, CCCV, CCCU), that is indicating heteroscedasticity in the residuals. The variance of the studendised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.	112

D.1	Scatter plots between assigned volume/units and predicted volume/cartons with a linear-squared transformation.	117
D.2	Scatter plots between assigned volume/units and predicted volume/cartons with a squared-squared transformation.	118
D.3	Scatter plots between assigned volume/units and predicted volume/cartons with a fourth-linear transformation.	119
D.4	Scatter plots between assigned volume/units and predicted volume/cartons with a linear-fourth transformation.	120
D.5	Scatter plots between assigned volume/units and predicted volume/cartons with a fourth-fourth transformation.	121
E.1	Regression analysis graphs for M1 of order 3	124
E.2	Regression analysis graphs for M2 of order 2	126

Chapter 1

Introduction

Businesses in today's global market face fierce competition and are under pressure to meet their customers' demands on time. This forces businesses to invest in the development of their supply chain and manage it more effectively. In a typical supply chain, raw materials are procured and items are produced at one or more factories, shipped to warehouses for intermediate storage, and then shipped to retailers or customers [9].

This introductory chapter explains what a supply chain is and looks at the *hub-and-spoke* supply chain network that is used by many companies [5]. Distribution centers are also discussed, the study's objectives are given and the underlying problem for this thesis is explained.

1.1 Supply chain

A supply chain is a process and/or network where final usable goods are produced starting from raw materials, like natural resources (water, trees, gold, etc). The supply chain consists of different parties that have to ensure the successful delivery of the final goods to the user.

The various parties involved in a supply chain can be divided into 5 stages, the supplier, manufacturer, distributor, retailer and customer [13]. Figure 1.1 gives a schematic representation of a supply chain with the flow of goods from the supplier to the customer.

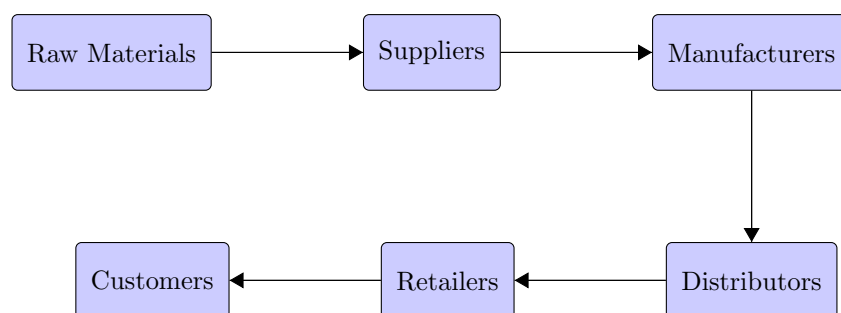


Figure 1.1: Graphical representation of different parties involved in a typical supply chain.

Suppliers typically have the ability to refine raw materials so that it would be ready for the manufacturers to use when they develop the final form of the product through potentially different

stages. The manufacturers usually make use of large facilities or factories that house the necessary machinery to manufacture the product in its final form.

Distributors mainly use *warehouses* and/or *distribution centers* to contribute their part of the supply chain where they temporarily store the products and then send it to retailers or directly to customers. Warehouses mainly focus on the storing of products, while distribution centres focus more on the distribution of products and the rapid movement of products through the facility [36]. Demand from customers usually vary over seasons and when demand is low, distribution centers and warehouses are used as buffers to slow down the flow of products [9].

The retailer receives products from distribution centers and sell it to the customers. Customers in turn may give information back to retailers and retailers then send this information to distribution centers and from distribution centers to manufacturers and then to suppliers. This upstream chain is usually referred to the value chain [36] with the supply chain referred to as the downstream chain.

Each stage itself can be broken up into different entities, like product development, manufacturing, operations and finance. Each stage is not necessarily limited to only one supplier, manufacturer, distributor, retailer or customer. A manufacturer can be supplied by many suppliers that supply different raw materials to develop a product, similarly distributors can be supplied with different products from many manufacturers.

1.2 Hub-and-spoke supply chain networks

In many supply chains, the concept of hub-and-spoke network design is used [5]. Consider a network of nodes and arcs in Figure 1.2 where the circles represent the nodes and the arrowed lines are the arcs.

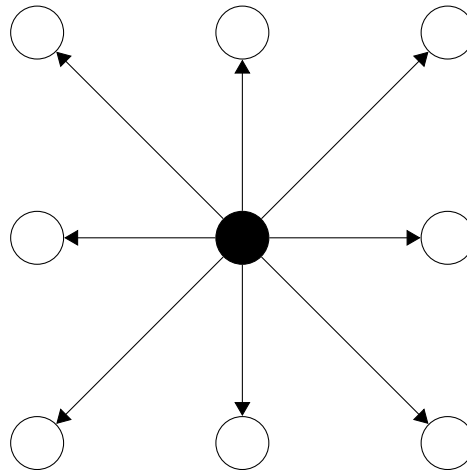


Figure 1.2: A network with one supply node coloured in black and eight demand nodes with arrows directing the flow of stock.

The hub-and-spoke supply chain network design can be constructed the same way where the black node is the hub also known as the supply node and the white nodes are the demand nodes or in the case of a supply chain the customers/retailers. The arrows point in the direction of stock flow.

In a supply chain, the hubs receive and rearrange stock that is sent to the spokes or the customers/retailers. Hub-and-spoke networks help to consolidate traffic from different origins and

send goods directly or via another hub to different destinations [2]. If the stock arrives at a customer via another hub, the stock is partitioned into smaller batches than from when it was received from the previous hub and then sent to the customers/retailers.

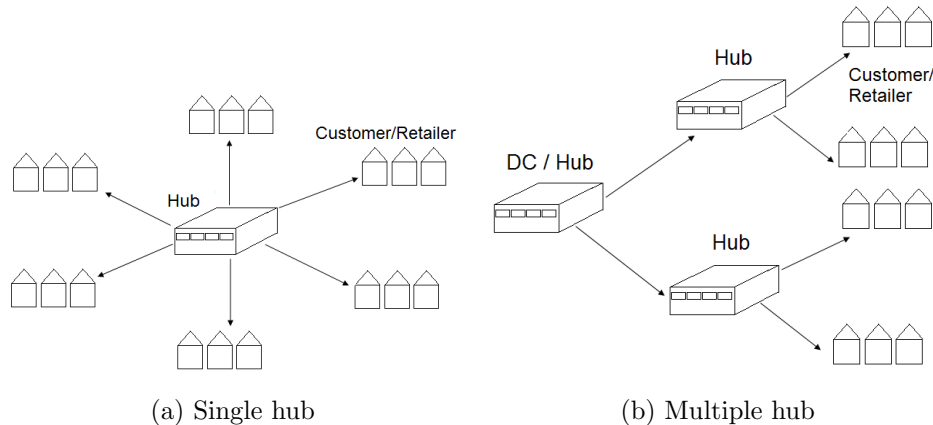


Figure 1.3: Schematic representation of two hub-and-spoke supply chain networks.

The two networks in Figure 1.3 depict two different hub-and-spoke layouts with the arrows pointing into the direction of the stock flow. Figure 1.3(a) depicts a hub-and-spoke network with one hub from where all customers/retailers get served. Figure 1.3(b) depict a network with two stages of hubs. Stock leave the DC (first stage), sent to separate hubs (second stage) and then sent to customers/retailers. Figure 1.3(b) can be seen as a network where products arrive at the DC from the manufacturers and then the stock must be sent to different areas to customers who are placed far from each other. The separate hubs then accommodate the nearby customers in their respective areas.

In hub-and-spoke networks, minimising travelling costs and time is important in order to meet customer demands. Much research has been done on deciding how many hubs to assign to a network of nodes and also where in a network to place these hubs [2]. O’Kelly [37] presents a quadratic integer program for a p -Hub median problem (p -HMP) and shows the p -HMP is NP-hard. The p -HMP attempts to place p hubs in a network such that the total transport cost is minimised. The assumptions for the p -HMP is, (1) all the hubs are connected, (2) each spoke is assigned to a single hub and (3) the hubs have unlimited capacity. Another hub location problem is the uncapacitated hub problem (UHP) where the number of hubs is not pre-determined, but there exist a cost for each hub included in the network [2]. The UHP then tries to decide how many hubs to place in a network while minimising the total travelling distance and the costs of the hubs included. The p -HMP and UHP also refer to the p -center and p -median problems described in Tansel *et al.* [47]. In a p -center problem, the furthest point from the center point must be minimized, whereas with the p -median problem, the total distances from the median must be minimised. These methods all place hubs on a node of a graph, but with the absolute median problem, hubs can also be placed on arcs of a network [23].

Another transport cost optimisation problem is the vehicle routing problem (VRP) that is also used in hub-and-spoke design. This problem is considered after the hubs have been identified and placed in a network. The goal with the VRP is to find minimum cost routes serving a set of customers with known demands. There are also many variants to the VRP like the VRP with time windows [33]. Hubs use the concept of the VRP to schedule trucks and serve customers.

1.3 Distribution centers

A distribution center (DC) is a facility where products can be stored, rearranged and grouped to meet the customers' demands. Distribution centers can also be seen as the regrouping phase in a supply chain [36]. DCs are used between the manufacturer and retailers to manage the flowrate of products in the supply chain. The DC also acts as the hub in a supply chain network. Sometimes economic and seasonality changes may have a big influence on the flowrate of products.

Products usually arrive at DCs in large quantities. These large quantities are either directly sorted and distributed or first broken up into smaller quantities. A stock keeping unit (SKU) is defined as the smallest unit quantity of a unique product or according to Bartholdi and Hackman [9] the smallest handling item of a certain product a DC work with. A SKU can either be a full pallet of products, a box of products or a single piece of a product depending on the policy a DC uses.

When products are sorted, they have to go through various operations in a DC. A DC usually consists of receiving, storing, picking, sorting and shipping operations. Figure 1.4 gives a schematic representation of these operations and the flow of the stock.

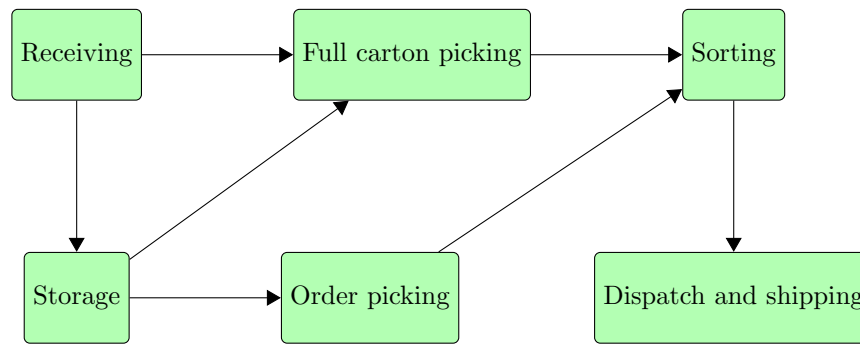


Figure 1.4: The flow of stock between departments in a distribution center.

1.3.1 Receiving

Receiving refers to the arrival of stock at a DC. Receiving includes the update of the inventory system used by the DC, the inspection of these products for any damage, incorrect counts and wrong descriptions.

The newly arrived products are temporarily stored in a goods received area of the DC and are then stored in bulk storage. Products may be stored as pallets or boxes and each product gets a storage area with a storage number for regular inventory control and easy retrieval. Deciding on storage locations for products are important at a later stage when retrieving the products from storage [9].

1.3.2 Storage

Storage involves the transfer of received stock to storage locations and can include repacking and physical movements [26]. Before the stock can be stored in the storage locations, each product or SKU must be assigned to a location in the storage area. The location of a SKU can have an effect on the time and cost of retrieving the stock. These storage locations get linked with its

corresponding SKU in the location in order for the workers to know where products are and how many locations are available for use.

1.3.3 Order picking

DCs receive *orders* from retailers for which products or SKUs they require. From the received retailer orders, *picking slips* are set up which contains the SKUs and information about the SKUs, like their quantity to be picked and location in storage. Depending on what type of layout and policy a DC uses, the stock for the orders will either be retrieved directly from the large storage area or from smaller storage areas, also known as fast pick areas that gets replenished from the large storage areas. Figure 1.5 depicts examples of a large and a small storage area.

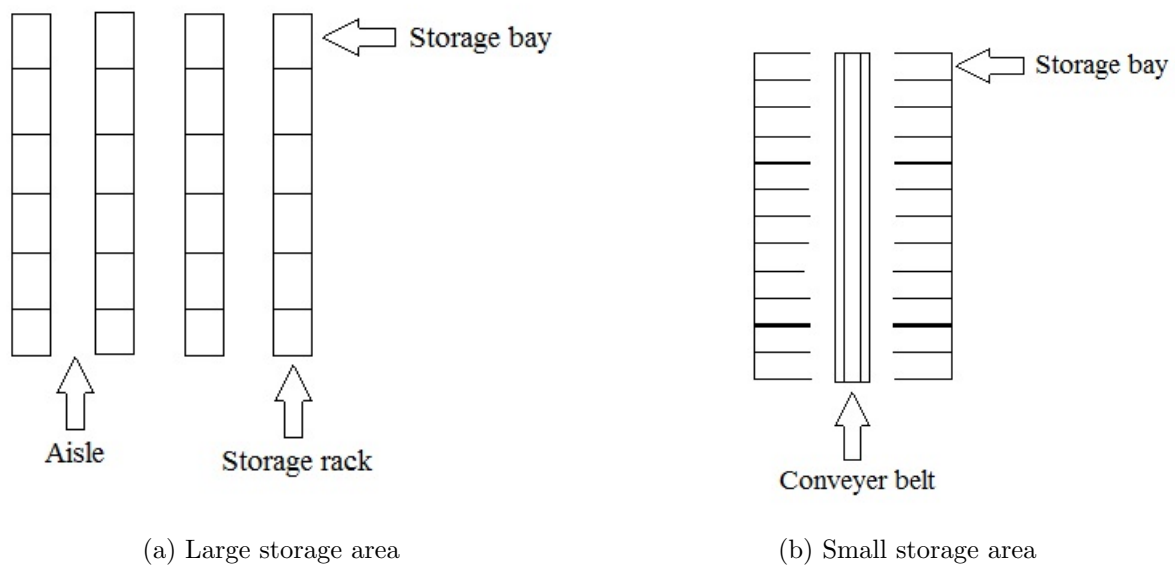


Figure 1.5: Two types of storage areas that may be used for order picking.

Figure 1.5(a) displays an example of a large storage area and Figure 1.5(b) a small storage area for picking. The large storage area consists of a floor to roof storage rack with storage bays that are placed next to each other with a gap between them. This gap is the aisle that workers use to retrieve stock from the storage bays. These storage areas can either be used for picking or just for storage that will replenish smaller storage areas. The small storage area on the right has storage bays on each side of a conveyor belt. Workers walk around the conveyor belt while they pick stock from the bays and place it on the conveyor belt from where it is conveyed to the sorting area. The DC under consideration in this project uses this type of storage for their picking. It gets replenished from a large storage area as shown in Figure 1.5(a). The DC refers to this small temporary storage area as picking lines.

The retailer/customer orders arrive at DCs in batches according to *time windows*. A time window can be an hourly time interval during which orders are received [26]. People called *pickers* do the picking operation also known as *order-picking* [9], where they retrieve the products from storage areas required by retailers for their orders. When an order arrives, it gets sequenced into a certain time slot or *picking shift* [26]. A picking shift can be seen as a period during a day when order picking takes place for a certain number of orders. Picking can also be done by custom made

machinery called automated picking [26]. The products of each order are stored in a container by pickers and send for sorting.

In full carton picking, products do not get broken down into smaller pieces as with the normal order picking. In full carton picking, the products are retrieved in their original form from the large storage area or directly from receiving. These products can be big and bulky items or small items contained in cartons, where the full cartons must be sent to the customer and not just a single item in the carton as it would be with piece picking.

1.3.4 Sorting

After the picking process, the picked orders must be sorted and batched into a single customer order. These batches of customer orders also get grouped in the appropriate routes in order for transportation vehicles to deliver the orders to the correct customers. Grouping and batching the orders save transportation costs and it makes it easier for the customers [9] to manage. The sorting can be a very labour-intensive task although walking long distances is not required. It is also the perfect time to check whether the orders have been picked accurately.

1.3.5 Shipping

The shipping area receives stock from sorting and loads it onto transportation vehicles. Each vehicle has an assigned route and the batched stock needs to be loaded into the correct vehicle. Depending on the system that a DC uses, the stock sometimes has to be staged if it needs to be loaded in reverse order of delivery or shipped long distances when trailers must be fully filled [9]. The stock also has to leave the DC in certain time windows in order to arrive on time at the customers or retailers and thus vehicles must be loaded by a certain time.

1.4 Crossdocking

Crossdocking normally refers to a warehouse or distribution center where the turnover time of stock is very fast. The stock arrives at a dock, gets sorted and immediately leaves the dock for its destination. In some cases, this turnover of stock can be measured in hours [9]. The reason for this fast turnover is because the customer's request is already known and there is no need to break the stock down into smaller pieces. Also, storage space in a dock would be a lot smaller than in a traditional warehouse or DC. The reason for crossdocking is to reduce transportation costs and for the consolidations of multiple shipments.

Docks normally have a long I-shaped design with many doors. Some doors are for unloading vehicles and others for loading stock into or onto vehicles. A variety of material handling methods can be used in a dock, like forklifts, palletjacks or conveyer belts. The design and layout in and around a dock can play a big role in the efficiency of a dock. Bartholdi and Hackman [9] provide some dock designs in practice and issues surrounding the designs of these docks and docks in general. Some of the issues mentioned by Bartholdi and Hackman [9] is the layout of the dock inside and the number of doors a dock has. More doors can lead to a dock being less efficient as the dock will have to be bigger and in turn workers need to do more. Stock that need to leave might travel long distances and pass doors that would have been more suitable for exit. Loading tend to take longer than unloading especially when there is a certain manner in which the stock must

be loaded into vehicles. Also the manner in which material is handled (for example palletjacks or forklifts) can also cause problems. The outside design and layout can cause congestion between vehicles and Bartholdi and Hackman [9] discuss the problems of T, U and H dock designs.

1.5 Industry background

Pepstores Ltd. (PEP) is a South African based clothing retail company established in 1965. PEP is part of a bigger holding company, Pepkor that owns other retailing companies in South Africa and abroad. Pepkor also own their own logistics company called Pepkor Logistics (PKL) in order to help the retailers to get the stock to their stores. PEP grew over the years and currently owns about 2000 stores in South Africa and neighbouring countries like Namibia, Botswana, Lesotho and Swaziland. Although PEP predominantly sell clothing, footwear and home (CFH) products, they also trade with cellular products like cellphones and airtime. The design and specifications of most of the CFH products are approved by employees of PEP and then ordered from suppliers in South Africa and predominantly in China to manufacture these products.

The suppliers ship the products to South Africa, where it is stored in PEP's DCs. PEP owns three DCs in South Africa, located in Durban, Cape Town and Johannesburg. Figure 1.6 shows the the layout of PEP's Durban DC. Within the DC, stock arrives at the receiving area and is then stored in the storage area. Figure 1.7 depicts the large storage racks in the Durban DC. The stock gets moved from the storage areas to picking lines where the order-picking process begins. Figure 1.8 shows how the pickers walk with trolleys between bays and the conveyor belt. They pick stock from bays and place it into cartons on the trolleys. Once the cartons on the trolleys are full, it gets placed on the conveyor belt. The picked stock is sent to the shipping area via the conveyor belts represented with wavy arrowed lines in Figure 1.6 and gets sorted before being loaded onto trucks and sent to the stores.

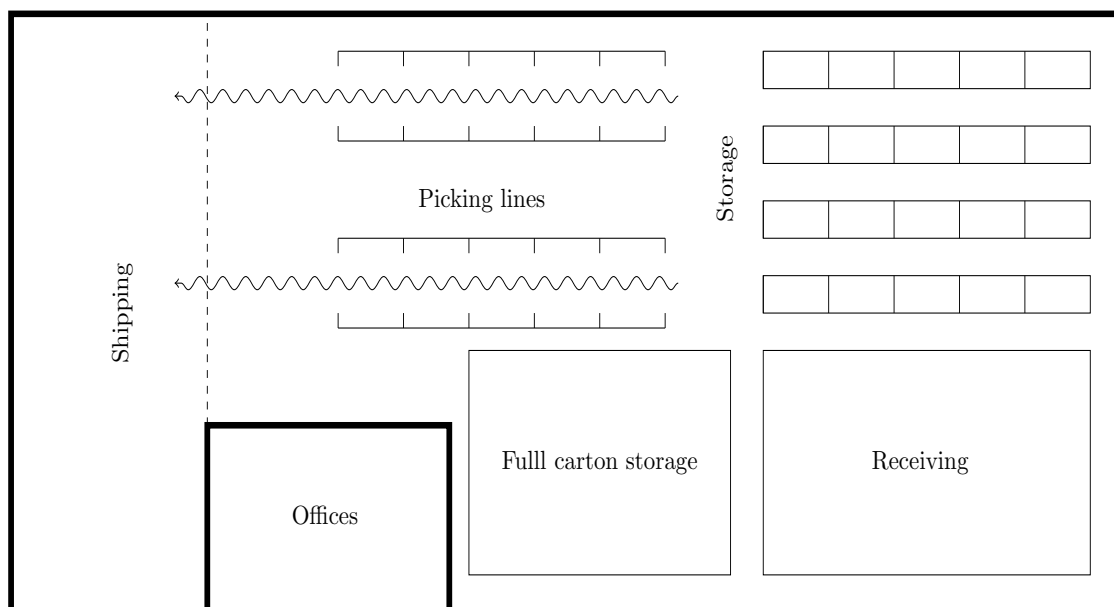


Figure 1.6: A schematic representation of the layout of the Durban DC owned by PEP.

The stock arrives at a store from a DC of PEP via a hub owned by PKL. The stock arrives at the hub in big volumes for many stores, located in the area serviced by the hub. The hub sorts the



Figure 1.7: A photo of the large storage area from the PEP Durban DC.



Figure 1.8: A photo of the pickers picking stock from bays on either side of the conveyor belt in the picking line.

stock according to the areas where it needs to go. These hubs act as crossdocking stations within the Pepkor supply chain as other Pepkor retailers, besides PEP also make use of these hubs.

Stock can also be transported between hubs and not just from the DC to the hub. In a case like this, stock for instance needs to travel from a hub in Cape Town to a hub in Johannesburg from where the stock is transported to the stores. Movement of stock between hubs is also called long-haul transportation, where the transportation between hub and store is called local transportation. Figure 1.9 contains a representation of the Cape Town hub situated in Kuilsrivier. There are doors to handle inbound and outbound stock separately. The stock for the local and long-haul transportation are stored in separate locations on the floor. These storage areas are surrounded by conveyor belts indicated by wavy lines. There is an outer and inner conveyor belt for each local and long-haul transport destination. When new stock arrives, it goes via conveyor

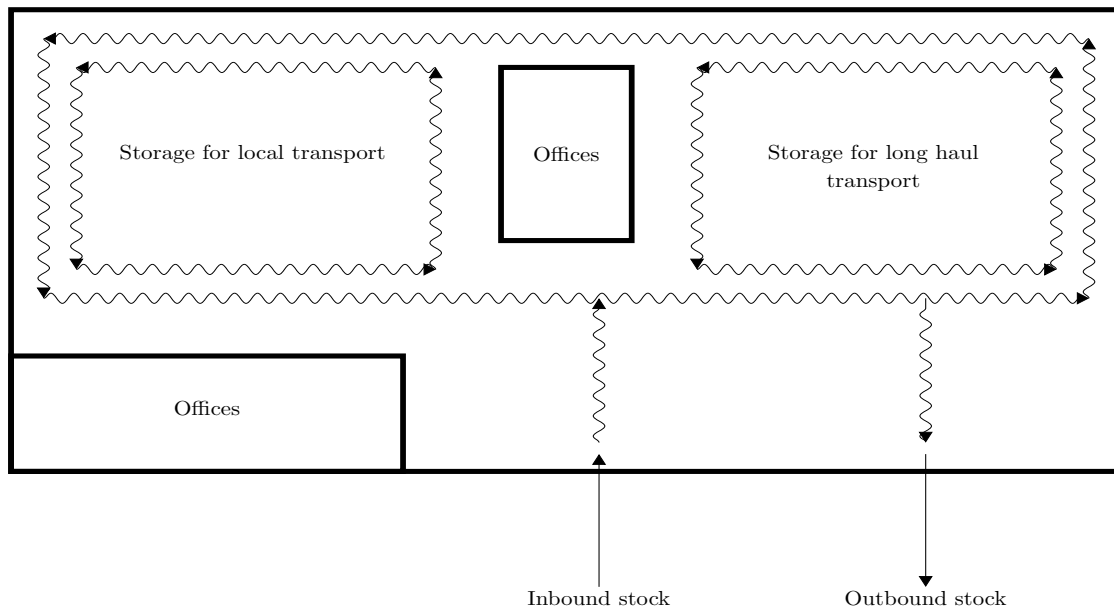


Figure 1.9: A schematic representation of the layout of the hub in Kuilsrivier, Cape Town.

belts on to the outer conveyer belt and workers push the stock to the inner conveyer belts to the correct storage area. When stock leaves, the stock gets placed on the outer belts and moves on the conveyor belts that leads to the trucks for loading.

The DCs and hubs share information with each other on how many cartons and volume (cubic meter) a hub will receive as well as the estimated delivery times. In this project, focus will be placed on the prediction of the of the number of cartons and volume per store. The prediction will be done by using regression modelling [8]. Regression is a mathematical model that establishes a relationship between one dependent variable and one or more independent variables. In the case of this thesis, the independent variables will be the assigned volume and quantity to a picking line. Separate regression models will be used to regress the dependent variables, volume and number of cartons sent to the hub.

1.6 Problem description

The hubs owned by PKL rent trucks that deliver the stock to the stores. These trucks are rented from third party companies at the beginning of each month. The hub decides how many trucks they need, based on historical data and their expectations for the coming month. Every day the hub employees need to set up a routing schedule that will determine which stores need to receive stock, routes that a truck should travel and the number of trucks needed to transport the stock to the stores.

The setup and scheduling of these routes play an important role in the profitability of a hub. Firstly, it is important for the hub to rent just enough trucks, not too many and not too few. If they rent too many trucks they could end up with trucks that are not utilized efficiently or trucks that are never used. If they rent too few, they can struggle to get the stock to stores fast enough and have problems with capacity in the hub where stock gets temporarily stored. Secondly, bad routing schedules can also lead to extra incurred costs. For instance, with a bad routing schedule, a store might need to be visited four times but with a better schedule the store might only need

to be visited twice, which can mean shorter routes and in turn cost savings.

In the case of the PEP Cape Town DC and hub, the DC sends the hub a prediction on the total volume and number of cartons they will receive the following day. This forecast is a total forecast and the hub doesn't know the number of cartons and volume per store until they receive it in the hub. The total number of cartons normally gets multiplied by a cubic meter per carton value to calculate the total volume. The number of cartons is determined by the average number of cartons a picker can pick and pack during a certain period of time.

Only when the cartons and volume per store is known by the hub, can they start scheduling the trucks and routes. If the hub receive an accurate forecast per store earlier, they can then start scheduling trucks and routes earlier and also make changes to the schedules in order to improve efficiency as new information arrives. In this thesis a method to accurately estimate the number of cartons and volume per store per day is presented. This method use information on what will be sent to the picking lines in the next few days to estimate the carton/volume outputs that go to the hub.

1.7 Thesis objectives

This thesis presents models to assist with the forecasting of stock sent from a DC to a hub. To achieve this, the following objectives are set:

1. Provide a literature review on simple and multiple linear regression as well as, regression analysis. Implement simple and multiple linear regression by using real life picking line data to predict the number of cartons and volume that comes from a picking line.
2. Provide a literature review on regression transformations. Investigate transformations of the different input and output picking line data.
3. Use the coefficients calculated by the regression models in a training data set and apply it on a test data set in order to implement forecast accuracy tests by comparing the actual against the forecasted number of cartons and volumes.
4. Perform a case study on a new data set and compare its accuracy to real life data.
5. Analyse the results from the forecast accuracy tests and draw a conclusion on whether the models can be used by PEP for forecasting the number of cartons and volume generated during a wave of picking.

1.8 Thesis layout and organisation

In Chapter 2, simple linear regression is discussed and picking line data from PEP's Kuilsrivier DC in Cape Town are used to predict the number of cartons and volume from the picking lines. Regression analysis using the simple linear regression model is also provided. The third chapter introduces multiple linear regression modelling for the carton and volume prediction. The fourth chapter gives a literature review on different types of transformations in regression analysis and provide a possible transformation for both carton and volume forecasts where heteroscedasticity is minimised.

Chapter 5 provides a literature review on polynomial regression modelling and uses transformed data to implement polynomial regression and provide regression analysis. The sixth chapter test the forecast accuracy of the models used in Chapter 5. Chapter 7 provides a case study where the actual values are compared to forecasted values from the polynomial models and DC forecast. Chapter 8 concludes the thesis.

Chapter 2

Simple linear regression

Regression models attempt to find a good mathematical relationship between two or more variables. Regression modelling are used in many sectors like finance, agriculture, science and business [11]. Many types of regression models exist and in this chapter an introduction to simple linear regression is provided. A simple linear regression model for the problem under consideration in this thesis will be discussed and conclusions presented.

2.1 Literature review

The literature review is divided into three sections. The first section defines and provide a background on simple linear regression. The second section introduces a method called ordinary least squares that calculates parameters to a simple linear regression model and the third section introduces regression through the origin.

2.1.1 Background

Simple linear regression (SLR), consist of building a relationship between two variables. One is the dependent variable and the other is the independent variable. The relationship between the two variables can be represented by the formula for a line [6],

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad (2.1)$$

where y_j is the dependent variable or observed value, x_j is the independent variable, β_0 and β_1 are the intercept and slope respectively and ε_j is the error term which is assumed to be normally distributed with mean 0 and standard deviation of σ [44]. The parameters β_0 and β_1 are the population parameters while the predicted model is written as,

$$\hat{y} = b_0 + b_1 x, \quad (2.2)$$

where \hat{y} is the predicted value, b_0 and b_1 are the sample parameters. SLR is used in many fields. Chen [12] uses SLR in the field of psychology where people's response times (in a case where people have to deal with conflicting situations) are used as a dependent variable and the expected time under normal conditions are used as an independent variable. Papadopoulos *et al.* [38] use

simple linear regression to estimate the cost of building and maintaining water waste plants in Greece.

The next subsection describe how to find the the estimates of b_0 and b_1 from equation (2.2) via the ordinary least squares method, the influence of points on the estimates derived by least squares and also the model fit is discussed.

2.1.2 Ordinary least squares method

The coefficients, b_i , can be found via the *ordinary least-squares method* (OLS). It is the most used method in regression [29] for finding the parameters. The OLS determine values for all the b_i 's by minimising

$$\sum_{j=1}^n (y_j - \hat{y}_j)^2 = \sum_{j=1}^n (y_j - b_0 + b_1 x_{1j})^2, \quad (2.3)$$

where y_j is the actual value or observation, \hat{y}_j is the estimate in equation (2.2) and the equation sum over n , the number of observations. In order for a regression model to be viable, it need to adhere to the four properties of OLS. These 4 properties are described in the next four points.

1. The parameters b_0 and b_1 are linear estimators of the dependent variable y .
2. The parameters b_0 and b_1 are unbiased, meaning, $E(b_0) = \beta_0$ and $E(b_1) = \beta_1$. Thus, in repeated applications of a regression model, the b_0 and b_1 paramters will on average equal the population parameters β_0 and β_1 .
3. The OLS estimator of the error variance is unbiased, that is, $E(\hat{\sigma}^2) = \sigma^2$. Thus, in repeated applications of a regression model, estimated value on average will converge to the true value.
4. The $\text{var}(b_0)$ and $\text{var}(b_1)$ are less than the variance of any other unbiased estimator of β_0 and β_1 .

The OLS estimator of the error variance is unbiased

When parameters adhere to these properties, then it is said that an SLR model is a best linear unbiased estimator (BLUE) [22]. If it does not adhere to these properties then the inference testing can be biased and wrongly interpreted.

The difference $(y_j - \hat{y}_j)$ in equation (2.3) is also called the residuals or errors [51]. These residuals are also referred to as ordinary residuals while other types of residuals also exist, for instance standardised residuals and studendised residuals. Standardised residuals are given by

$$r'_j = \frac{r_j}{s_r \sqrt{1 - h_j}}, \quad (2.4)$$

where r'_j is the standardised residual, r_j is the ordinary residual, s_r is the error of variance obtained by $\sqrt{\frac{1}{n-2} \sum_{j=1}^n (y_j - \hat{y}_j)^2}$ and h_j is the leverage of point j . The leverage for a point can

be considered high or low. High leverage points can have a large effect on the estimates when removed or used in the model. The leverage of a point j is defined as,

$$h_j = \frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum_{k=1}^n (x_k - \bar{x})^2}. \quad (2.5)$$

The leverage of a point (also known as hat values [30]), capture how far an observation is from the mean. Observations far from the mean will have large hat values while those close to the mean will have small values. Studendised residuals are defined as

$$r_j^* = \frac{r_j}{s_{r(-j)}\sqrt{1-h_j}}, \quad (2.6)$$

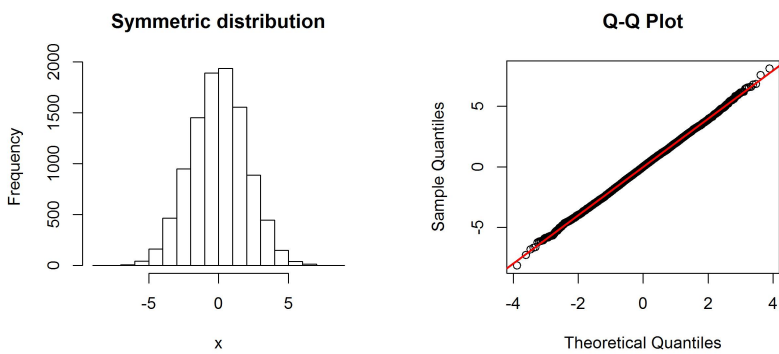
where $s_{r(-j)}$ is the error of variability if observation j is removed from the data. Standardised residuals and studendised residuals are preferred over ordinary residuals due to the difference in variability in the residuals. Standardised residuals and studendised residuals ensure all the residuals have the same variability. In this thesis the studendised residuals are used for the histogram plots and Q-Q plots.

Q-Q plots compare two distributions with each other by first calculating the quantiles of each distribution and then use it as the x and y axis. Each observation will fall into a quantile on both distributions and the line where $y = x$ is where a symmetric distribution is found. Figure 2.1 depicts the histogram of four distributions and the corresponding Q-Q plot. The x -axis of the Q-Q plots are quantiles for the normal distribution and the y -axis is the quantiles for the observed data. The red line is where $y = x$. The histogram in Figure 2.1(a) is a symmetric normal distribution and all the dots of the Q-Q plot falls on the red line where $y = x$. The symmetric distribution with fat tails in Figure 2.1(b) have the tails in the Q-Q plot move slightly away from the red line. The histograms in Figures 2.1(c) and 2.1(d) are positively and negative skewed respectively, where the Q-Q plot for the positively skewed plot have its tales going up away from the red line. The Q-Q plot for the negative skewed plot have its tales going down away from the red line.

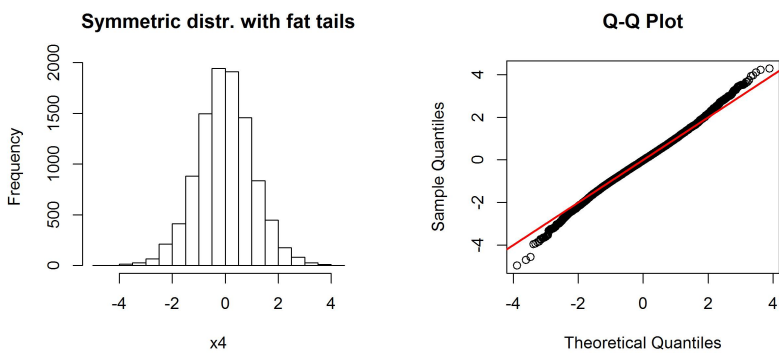
Both histograms and Q-Q plots are presented in the results section of this chapter. The other plots that will be included are fitted and residual plots. The histogram and Q-Q plots will be used to check for normality in the residuals. The fitted plot are used to see how the regression line fit the actual data. The studendised residual plots will be used to check whether heteroscedasticity are obtained in the residuals along with the Breusch-Pagan test for heteroscedasticity [22]. The Breusch-Pagan test regress the residuals of a regression model with the independent variable or variables and result in an auxiliary regression. The coefficient of determination or R^2 resulting from the auxiliary regression are multiplied by the number of observations that result in a χ^2 statistic with p degrees of freedom under the null hypothesis that homoscedasticity exist, where p is the number of independent variables. If n are very large the p value tend to be small and suggest that heteroscedasticity exists. In order to determine whether heteroscedasticity improved or has been alleviated, the residual plots and fitplots also need to be analysed along with the Breusch-Pagan test.

The partial derivates of equation (2.3) with respect to b_0 and b_1 and are taken and set equal to zero. Solving thes equations for b_0 and b_1 yields

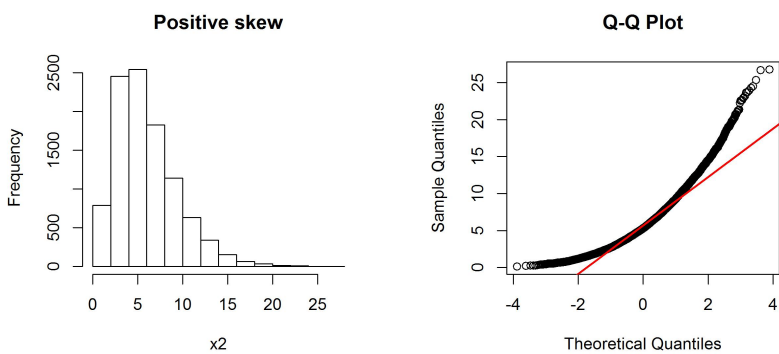
$$b_1 = \left(\sum xy - n\bar{x}\bar{y} \right) / \left(\sum x^2 - n\bar{x}^2 \right) \quad (2.7)$$



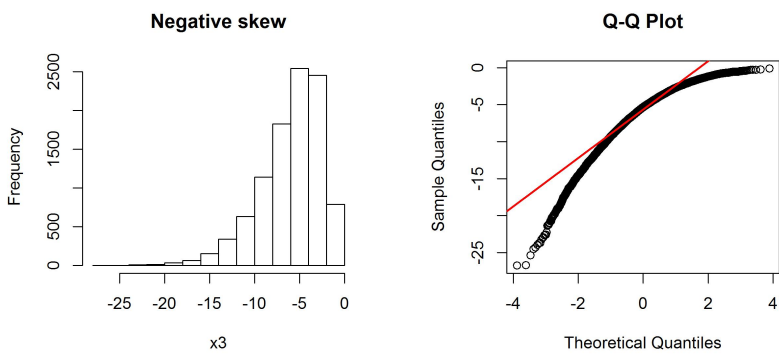
(a) Symmetric distribution



(b) Symmetric distribution with fat tails



(c) Positive skewed dstribution



(d) Negative skewed dstribution

Figure 2.1: Examples of histogram and corresponding Q-Q plots.

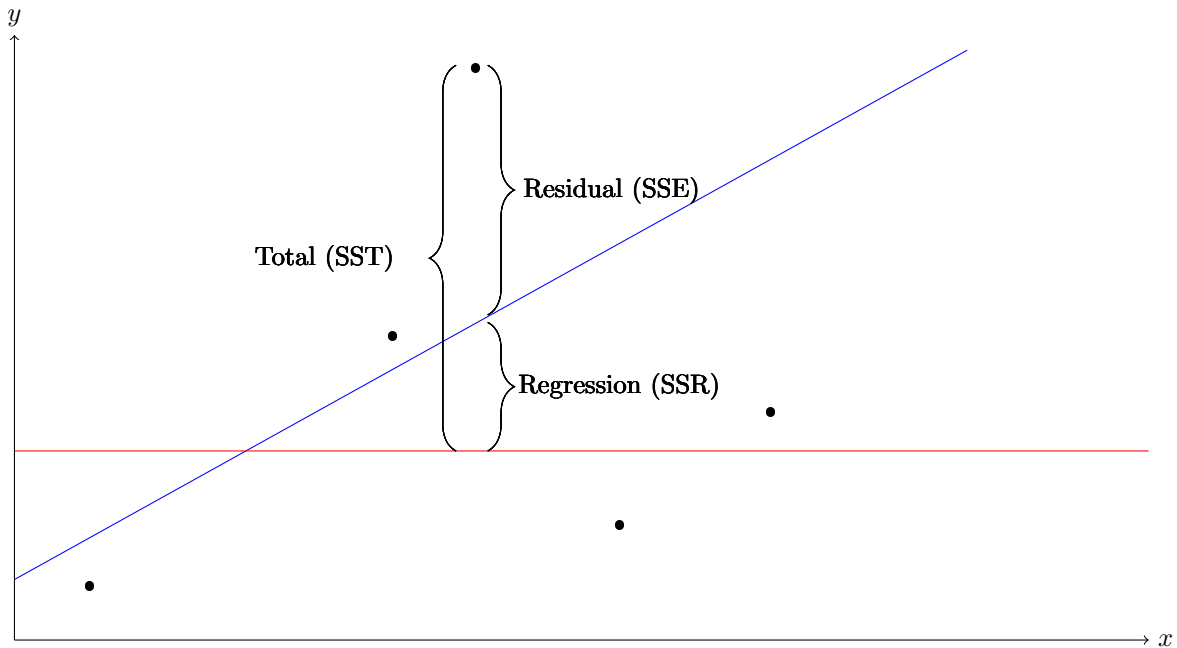


Figure 2.2: Graphical representation of the residuals in regression and include a total and regression part.

and

$$b_0 = \bar{y} - b_1 \bar{x}. \quad (2.8)$$

In equations (2.7) and (2.8), \bar{x} is the average of all the x -values and \bar{y} is the average of all the y -values [6]. OLS minimises the squared residuals and the b_i values obtained from this optimisation problem are unique. This leads to the best linear equation (if the squared errors are minimised) for the given points. Figure 2.2 shows the residual graphically and also include a regression and total part.

The regression part is the distance between the regression line and mean value of the dependent variable while the total is the residual plus regression. If the total sum of squared residuals are presented as SSE, sum of squared regression as SSR and sum of total as SST then

$$SST = SSR + SSE. \quad (2.9)$$

If both sides are divided by SST, it follows that

$$1 = \frac{SSR}{SST} + \frac{SSE}{SST}, \quad (2.10)$$

where $R^2 = \frac{SSR}{SST}$ and is the variation in the output variable explained by the regression model. The R^2 statistic must be a value between 0 and 1. If it is 0 then there exist no relationship between the output and input variable or the regression line fit the data poorly and if it is 1 then the regression line fit the data perfectly. The R^2 statistic is a goodness of fit measure for regression models and it is widely used in literature and mathematical software packages. Other ways of measuring the goodness of fit also exist, like the Akaike information criterion (AIC), Bayesian

information criterion (BIC) and Mallows' C_p statistics [7, 46]. The AIC can be calculated as

$$AIC = -2 \log L(\theta, y) + 2k, \quad (2.11)$$

while BIC is calculated as

$$BIC = -2 \log L(\theta, y) + k \log(n) \quad (2.12)$$

and Mallows' C_p is calculated as

$$C_p = SSE/(\sigma)^2 + 2k - n. \quad (2.13)$$

Within these equations n is the number of observations, k (for SLR, $k = 1$) is the number of variables, $\log L(\theta, y)$ is the log-likelihood, σ is the error variance and SSE is the sum of squared residuals. The R^2 measure will be used in the SLR models within this chapter. The calculated estimates or b_i values from the regression models will also be analysed, by focusing on the t and p -values of the estimates and determining whether the input/independent variables are sufficient estimators for the models used in this thesis.

OLS weighs each observation equally and this can cause influential observations or outliers to make the regression line fit badly. This can also cause the model not to be BLUE. Flaster [34] who regresses the weight of adults plotted a graph of residuals to identify outliers. In this thesis, outliers will be identified by using Cook's distance graphs. The Cook's distance statistic was developed by R. Dennis Cook in 1977 [31]. It provides the degree by which an observation influence the β estimates in a regression model and it is another measure of leverage. The $(1 - \alpha)$ confidence ellipsoid for the unknown vector, β for a given point \mathbf{b} must satisfy the inequality

$$\frac{(\beta - b^T)X^T X(\beta - b)}{ps^2} \leq F(p, n - p, 1 - \alpha), \quad (2.14)$$

where $s^2 = \frac{R^T R}{n-p}$, b is the point value, n is the number of observations, p is the number of variables and $F(p, n - p, 1 - \alpha)$ is an F -distribution with p degrees of freedom for the numerator and $n - p$ degrees of freedom for the denominator. The X is an $n \times p$ matrix that represent the known values of the independent variables. In order to find the degree of influence of a point on the β estimates, one can delete the point or observation from the data and get new β estimates. Cook's distance uses this technique with the left side of inequality (2.14) to calculate the influence of an observation. Cook's distance are mathematically defined as [14]

$$D_i = \frac{(\beta - \beta_{-i})^T X^T X(\beta - \beta_{-i})}{ps^2}, \quad (2.15)$$

where β_{-i} is the estimates calculated with observation i deleted and D_i is Cook's distance for point i . A critical value

$$c = \frac{4}{n - k - 2} \quad (2.16)$$

is calculated for Cook's distance statistic. Any observation with a bigger Cook's distance value than c is considered an outlier. Cook's distance plots indicate which observations can be considered

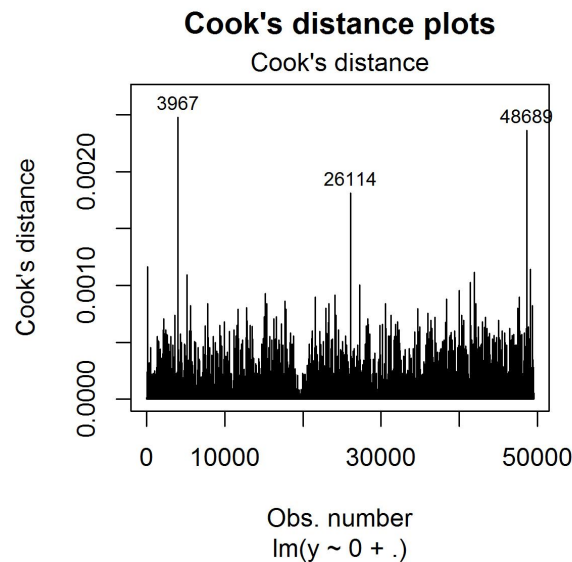


Figure 2.3: An example of a Cook's distance plot that displays Cook's distance values for each observation.

as outliers. An example of a Cook's distance plot can be seen in Figure 2.3. The observations are on the x -axis while Cook's distance are on the y -axis.

Another method for calculating the b_i estimates is by the method of least absolute deviation (LAD). This method is more robust than the OLS method and does not get influenced by outliers as much as OLS. The equation for LAD can be written as,

$$\sum_{j=1}^n |y_j - \hat{y}_j| = \sum_{j=1}^n |y_j - b_0 + b_1 x_{1j}|. \quad (2.17)$$

Instead of calculating the squared residuals, LAD calculates the absolute values of the residuals. LAD does not give a unique solution and need to use linear programming to find estimates [21]. LAD is not always sufficient when solving real life problems [50] and is generally difficult to solve due to discontinuous derivatives. Wang *et al.* [50] uses neural networks to solve a LAD problem for estimating time delay in bioelectric signals. Abdelmalek *et al.* [1] uses the LAD estimation in image restoration. The regression model implemented in this chapter will use the OLS method.

The discussion of SLR and OLS in this section include the intercept term in the model. Sometimes the non zero intercept is not needed in the model and it may be assumed that the intercept is zero. The next section gives an overview on regression modelling with an intercept of zero.

2.1.3 Regression through the origin

Regression through the origin (RTO) is not well documented and has a lot of controversy behind it [18]. Forcing a regression through the origin is essentially placing a constraint on the model and can cause the model fit not to be good and also negatively affect the accuracy of the model. In some cases regression with a zero intercept is needed. Eisenhauer [18] mentions the lagging of observation in order to correct serial correlation leads to an RTO model and correcting observations for heteroscedasticity also leads to an RTO model. There also exist other cases when RTO is

appropriate. Chambers and Dunstan [10] uses a regression through the origin to predict total cane harvested, gross value of cane and total farm expenditure by using the size of the area assigned for cane planting as an independent variable. In this case if no area has been assigned or used then there would be no harvesting output. Adelman and Watkins [3] regress the value of oil and gas while dropping the constant term as no oil and gas will lead to a value of 0. The RTO model is also investigated in this thesis, because if the independent variable, which will be the volume or units sent to a picking line are 0 then the dependent variable, which will be the number of cartons or volume coming off the picking line, should be 0.

Barreto and Maharry [8] provide an algorithm for RTO models that make use of least median squares (LMS). The LMS method have a similar formula as OLS, where instead of using the sum of squared residuals, the sum gets replaced by the median and may be formulated as [44]

$$\text{minimise median}\{(y_1 - (b_0 + b_1x_1)), (y_2 - (b_0 + b_1x_2)), \dots, (y_n - (b_0 + b_1x_n))\}. \quad (2.18)$$

Kayhan and Gunay [28] takes the algorithm from Barreto and Maharry [8] further and create a similar algorithm for RTO models with more than one independent variable. These algorithms find different values for the b_i estimates and consider the b_i parameters as the optimal solution which generates the set of residuals with the smallest median. In this thesis OLS is used for the RTO models.

2.2 Model assumptions

In the proposed SLR model, the following assumptions are made:

1. The stock assigned to a picking line is fixed and already stored in the picking line bays.
2. The pickers will not short or over pick the units or volumes assigned to a picking line for a store.

2.3 Model

An SLR model is proposed where the volume and number of cartons that will come from a picking line need be to predicted. The volume on a picking line that need to be picked will be called assigned volume and the units on a picking line that needs to be picked will be called assigned units. The volumes that will come from a picking line will be referred to as the predicted volume and the cartons from a picking line will refer to the predicted cartons. The assigned volume and units are also the independent variables, while the predicted volume and cartons are dependent variables.

There are 4 models, where two models predict volume and the other two predict cartons coming off a picking line. Two of the models use assigned units as independent variable and two assigned volume as independent variable. The variables for the 4 models are summarised in Table 2.1.

In each model, y'_{mj} is the dependent variable, x'_{mj} is the independent variable and the volumes are measured in cubic meters (m^3). The corresponding mathematical formulation for the SLR model

Variable	Description
y_{1j}^v	the predicted volumes from a picking line for observation j and model 1
x_{1j}^u	the assigned units to a picking line for observation j and model 1
y_{2j}^v	the predicted volumes from a picking line for observation j and model 2
x_{2j}^v	the assigned volumes to a picking line for observaton j and model 2
y_{3j}^c	the predicted cartons from a picking line for observation j and model 3
x_{3j}^u	the assigned units to a picking line for observation j and model 3
y_{4j}^c	the predicted cartons from a picking line for observation j and model 4
x_{4j}^v	the assigned volumes to a picking line for observaton j and model 4

Table 2.1: The variables for the 4 models where volume is measured in cubic meters (m^3).

for the four models are

$$y_{mj}^l = a_m x_{mj}^l + \varepsilon_{ml} \quad (2.19)$$

where a_m is the coefficient to the independent variables and ε_{ml} is the error term. The equation does not contain an intercept as zero inputs will have to lead to zero outputs. OLS is used to calculate the coefficient a_m for each model. Table 2.2 shows the variables and corresponding model formulation for each of the 4 models. The models are also numbered and each model will be referred to by its number further in the chapter.

Model 1 has predicted volume as dependent variable and assigned units as independent variable, model 2 has predicted volume as dependent variable and assigned volume as independent variable, model 3 has predicted cartons as dependent variable and assigned units as independent variable and model 4 has predicted cartons as dependent variable and assigned volume as independent variable.

Model nr.	Dependent variable	Independent variable	Dependent variable description	Independent variable description	Model formulation
1	y_{1j}^v	x_{1j}^u	volume	units	$y_{1j}^v = a_1 x_{1j}^u$
2	y_{2j}^v	x_{2j}^v	volume	volume	$y_{1j}^v = a_2 x_{2j}^v$
3	y_{3j}^c	x_{3j}^u	cartons	units	$y_{2j}^v = a_3 x_{3j}^u$
4	y_{4j}^c	x_{4j}^v	cartons	volume	$y_{2j}^v = a_4 x_{4j}^v$

Table 2.2: The 4 SLR models numbered from 1 to 4.

2.4 Data

Picking line data from 7 June 2014 until 31 July 2014 from PEP's Kuilsrivier DC near Cape Town were gathered from their reporting system. The data contains the store numbers, picking line numbers, the total volumes produced per store per picking line, the total cartons produced per store per picking line, the total assigned volumes per store per picking line and the total assigned units per store per picking line. Each store and picking line combination are an observation.

PEP calculate assigned volumes by first determining the number of units in a carton that arrive at a DC and then measure the m^3 of the carton to determine the produced values. The carton m^3 is then divided by the units in the cartons to get a volume per unit. The assigned volumes in the data is then a sum of volume for all SKUs that needs to be picked for an observation. The produced volumes from a picking line are determined by the dimension scanner used in the DC. The dimension scanner measures the dimension of a carton and calculate the m^3 for each carton given the dimensions. A total of 58 observations out of 65657 observations were removed from the data due to an assigned cubic meters of 0. Table 2.3 shows an example of observation number 1 with store number 102 and picking line number 7123. The picking line number is a sequentially generated number by PEPs warehouse system. Every time a picking line is scheduled a new number is generated in order to distinct between picking lines and for historical use.

Observation nr.	Store nr.	Picking line nr.	Produced volume	Produced cartons	Assigned volume	Assigned units
1	102	7123	0.41	2	0.26	26

Table 2.3: An example of the data with one observation.

The data are split into two sets, a training set where the regression coefficients for the four models are determined with the SLR model and a test set will be used to determine the forecast accuracy of the regression models. The coefficients determined in the training sets will be used in the test sets and the actual volumes and cartons will be compared to the predicted volumes and cartons. The training set contains 55186 observations and include data from picking lines that were built from 7 June 2014 until 15 July 2014. The test set has 10471 observations containing picking line data from 16 July 2014 until 31 July 2014. The regression model coefficients and results with the training data set are performed in the software package, R version 3.0.2 [42]. The coefficients are retrieved from the software package R version 3.0.2 and exported to the software package Excel [24] where the forecast accuracy is calculated with the test set.

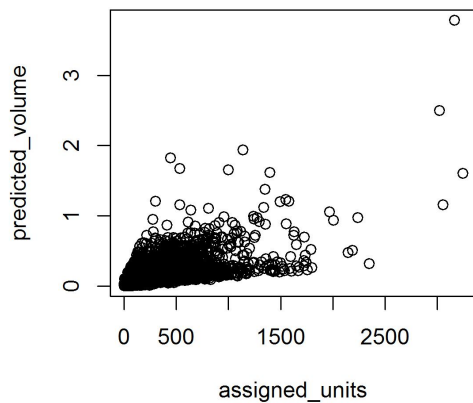
2.5 Results for the training

This section provides feedback on the results obtained on all four models with SLR. Figure 2.4 contains graphical representations that displays the relationship between independent variables, assigned units/volume and dependent variables, predicted cartons/volumes.

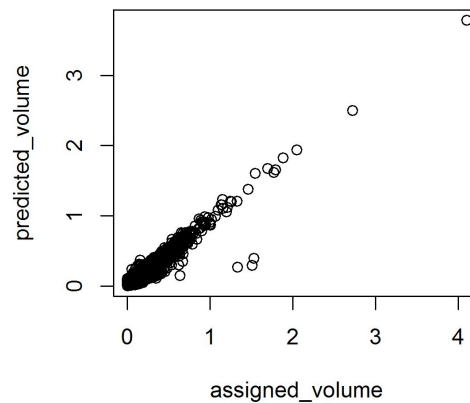
These plots indicate that there is a positive relationship between the dependent (predicted volumes/cartons) and independent variables (assigned volumes/units). It also indicate that the variance of the dependent variables increase as the value of the independent variables increase and a transformation to the data might be needed. The change of variance is less of a problem in Figure 2.4(b). Before doing any transformations, the regression models are applied to the untransformed data.

2.5.1 SLR results part 1

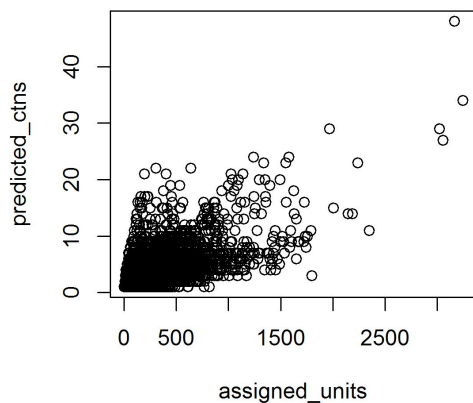
This subsection contains the results to the 4 models without any transformation and/or removal of observations. Table 2.4 provide the estimates, standard error, t and p -values for the 4 SLR models.



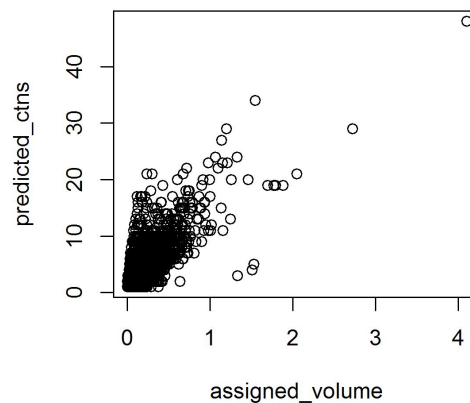
(a) Predicted volume vs assigned units



(b) Predicted volume vs assigned volume



(c) Predicted cartons vs assigned units



(d) Predicted cartons vs assigned volume

Figure 2.4: Scatter plots showing the relationship between dependent and independent variables.

Model	Dependent variable	Indendent variable	Estimate	Std. Error	t value	$\Pr(\geq t)$
1	Volume	Units	0.00	1.28E-06	404.235	0
2	Volume	Volume	0.98	0.000785	1245.72	0
3	Cartons	Units	0.01	2.89E-05	387.79	0
4	Cartons	Volume	19.49	0.040801	477.975	0

Table 2.4: A summary of the coefficients obtained from the four SLR models.

From this table the p -values are all zero which indicates that the assigned units and assigned volume are good estimates for the dependent variables, produced cartons and produced volume. The models with units as independent variables have estimates less than 1 and models with volume as independent variable, have larger estimates. This is expected as volume (most of the time less than 1) will be small compared to units that will be at least one. The standard error of the

estimates or standard deviation of the errors is small, which also indicate that the models predicts well.

Model	1	2	3	4
Dependent/Independent	Volume/units	Volume/volume	Cartons/units	Cartons/volumes
Variables	y_{1j}/x_{1j}	y_{1j}/x_{2j}	y_{2j}/x_{1j}	y_{2j}/x_{2j}
R^2	0.773	0.964	0.748	0.824
RSE	0.048	0.019	1.139	0.954

Table 2.5: Regression measurements for the 4 models of the SLR model.

Table 2.5 provide the R^2 and root square error (RSE) for the four models. The first column gives the measurement description while the rest of the columns provide the statistics for each model. The models with assigned volume (models 2 and 4) produce bigger R^2 values than the models with assigned units (models 1 and 3). The models where cartons are predicted have larger RSEs than those models that predict volume. This is caused by the larger values of cartons and shows there is a tendency to over or under predict by approximately 1 carton. There is an indication that by using these models, predicting the cartons will be less accurate than predicting volumes and by using assigned volume to predict cartons or volumes seems to be a better option over assigned units as the assigned volume models, 2 and 4, has more favourable R^2 and RSE values.

Figure 2.5 gives the residual histogram and Q-Q plot for Model 1. The histogram indicate that the residuals are not normally distributed and there are outliers on the tails of the distribution of the residuals.

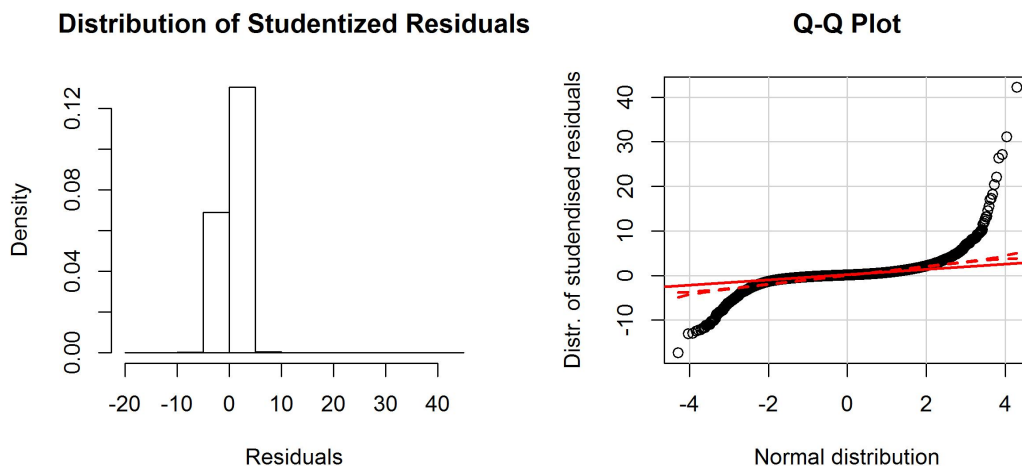


Figure 2.5: Residual histogram and Q-Q plot for Model 1 with units as independent variable and volume as dependent variable.

The fitted plot and residual plots are shown in Figure 2.6 for model 1. The fitted line in the fitted plot, Figure 2.6(a), does not fit the data well. As the independent variable increase, the variance of the dependent variable increase. The residual plot in Figure 2.6(b) shows that the residuals form a funnel shape to the right.

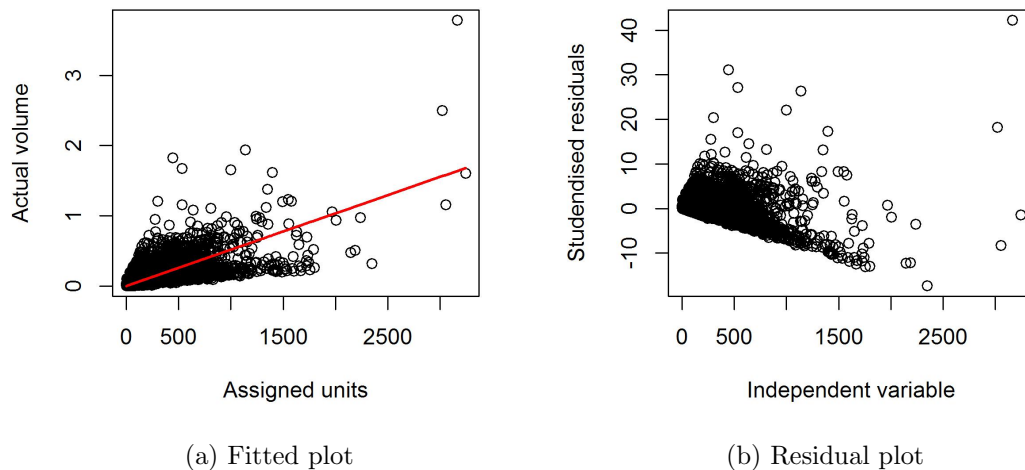


Figure 2.6: Fitted and standardised residual plots for the model with units as independent variable and volume as dependent variable.

The BP test results are displayed in Table 2.6 for all 4 models. Model 1 has a χ^2 equal to 3074.56. The p -values are all 0, suggesting heteroscedasticity exist in the residuals for all 4 models. The residual plot in Figure 2.6(b) support this assumption with a funnel shape to the right formed by the residuals against the independent variable that is units.

Breusch-Pagan Test	Model 1 Volume/units	Model 2 Volume/volume	Model 3 Cartons/units	Model 4 Cartons/volume
χ^2	3074.56	1080.22	3900.69	3074.56
Degrees of freedom	0	0	0	0
p -value	0	0	0	0

Table 2.6: The Breusch-Pagan tests for the four models.

The histograms, Q-Q plots, fitted plots and residual plots for the other models are shown in Appendix A, Figure A.1–A.6. These plots show similar patterns in the residuals as with Model 1. The fitted plot of Model 2 shown in Figure A.4(a) has a fitted line that fit the data better than the other models. The R^2 is also higher than the other models at 0.964 shown in Table 2.5 for model 2. Although there seem to be a better fit, the residual plot in Figure A.5(b) still seem to support the BP test in Table 2.6, that assume heteroscedasticity exists. These results shows that heteroscedasticity does exist in the data and the histograms and Q-Q plots shows that the residuals are not normally distributed, which means the BLUE was not achieved.

2.5.2 SLR results part 2

The histograms and Q-Q plots also showed that some influential observations exists that skew the estimates. The first attempt to lessen the problem of heteroscedasticity and normality in the residual distribution is by removing some of the influential observations or outliers. These

observations are identified by determining the Cook's distance for each observation and if the Cook's distance is greater than the critical value as calculated in equation (2.16), the observation are removed from the training set. Once all these observations are removed, new SLR results are obtained. Figure 2.7 shows the Cook's distance plots for the 4 models before influential outliers are removed. Each line on the plot indicate the Cook's distance of an observation. The observations with a larger Cook's distance than the critical value are indicated by the observation number.

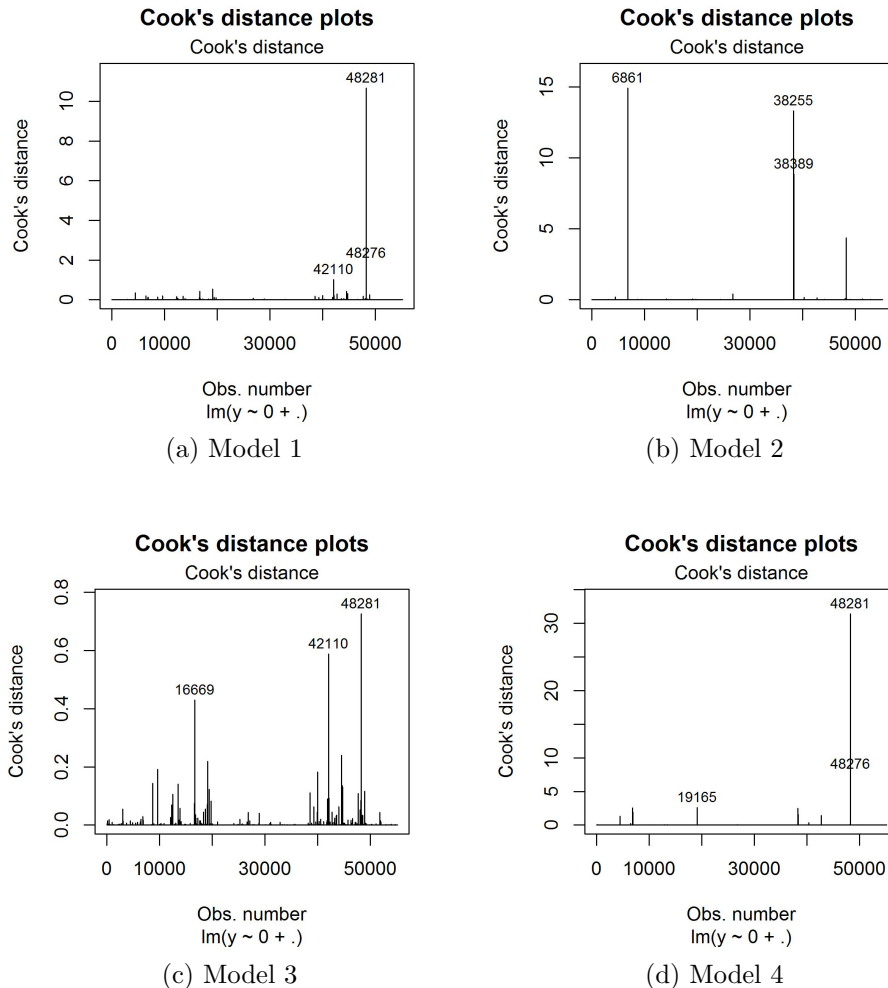


Figure 2.7: Cook's distance plots for the 4 models before observations were removed.

Figure 2.8 displays the Cook's distance plots after the influential observations are removed. The Cook's distance values for these observations are more close to each other than before the observations has been removed. Some observations has Cook's distance values larger than the critical value because the critical value decreased due to less observations. A total of 5 625 observations were removed and a new training set were obtained. SLR modelling of the 4 models were implemented on this new training set and the coefficient results are shown in Table 2.7.

The coefficient estimates and t -values does not differ significantly from the coefficient estimates before outliers has been removed. The coefficients are more significant to the model where observations were removed with estimates and t -values that are slightly higher than the results in Table 2.4.

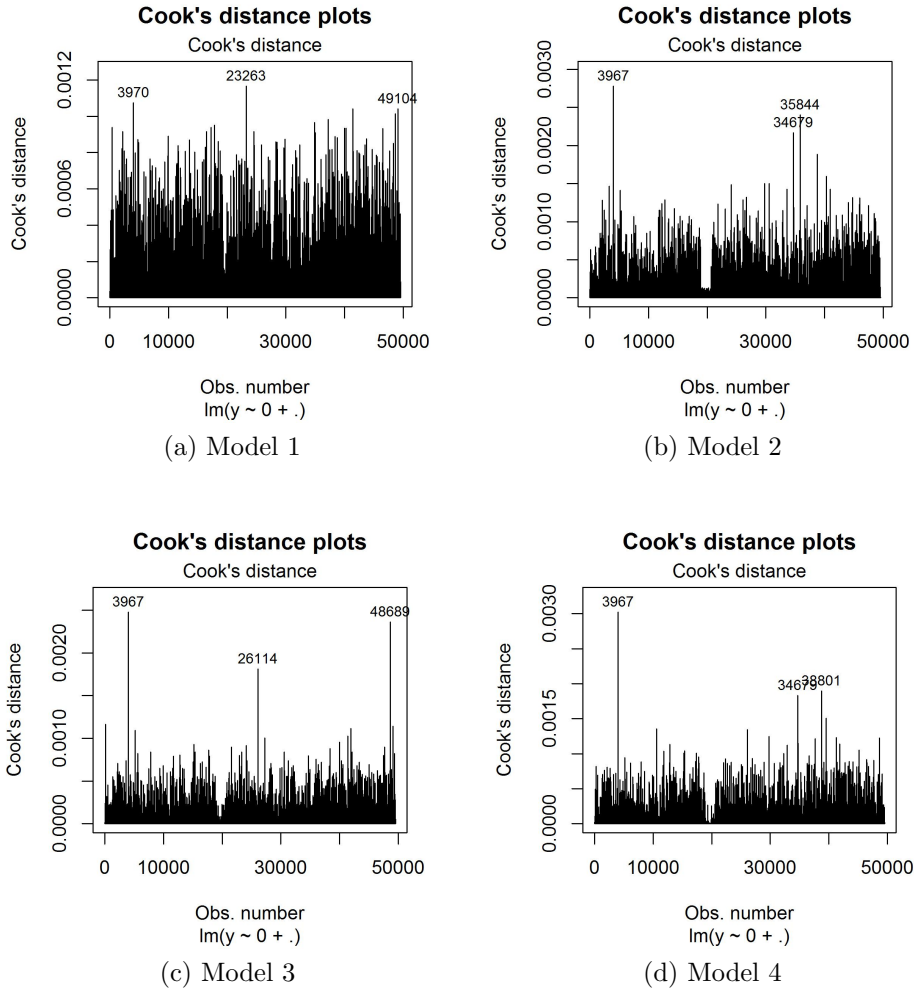


Figure 2.8: Cook's distance plots for the 4 models after observations were removed.

Model	Dependent variable	Indendent variable	Estimate	Std. Error	t value	$\Pr(\geq t)$
1	Volume	Units	0.000561	1.05E-06	534.55	0
2	Volume	Volume	1.04	0.000754	1381.53	0
3	Cartons	Units	0.0129	3.02E-05	428.34	0
4	Cartons	Volume	23.18	0.0458	505.99	0

Table 2.7: A summary of the coefficients obtained from the 4 SLR models after influential observations were removed.

The R^2 shown in Table 2.8 also increased from the previous results in Table 2.5 on all 4 models. The RSE values also decreased indicating better fit of the model. The fitted plot for Model 1 are displayed in Figure 2.9(a). This plot shows there is not many observations that could be considered as influential outliers. There is a diamond shape to the manner the dependent and independent variable relates to each other. This also cause the residuals to change variance as the independent variable increases as shown in Figure 2.9(b). The fitted plot and studendised residual plot for Model 2 after influential observations are removed displays the same problem. This can be seen in Figure A.10(a). The fitted plots for Models 3 and 4 form lines caused by the integer

carton values and the fitted line does not fit the data very well in Figures A.11(a) and A.12(a).

Model	1	2	3	4
Dependent/Independent	Volume/units	Volume/volume	Cartons/units	Cartons/volumes
Variables	y_{1j}/x_{1j}	y_{1j}/x_{2j}	y_{2j}/x_{1j}	y_{2j}/x_{2j}
R^2	0.852	0.975	0.787	0.837
RSE	0.027	0.011	0.785	0.686

Table 2.8: Regression measurements for the 4 models of the SLR model after influential observations were removed.

Figure 2.10 shows the histogram and Q-Q plots for model 1 after the influential observations were removed. The residuals are not normally distributed, but is positively skewed. The studentised residual distributions of the other models is also not normally distributed and positively skewed. These plots can be seen in Figures A.7–A.9. Table 2.9 shows the BP tests for the models without influential observations. The χ^2 values are significantly smaller than the BP test results for the SLR models with influential observations, which means heteroscedasticity did improve, but the p -values are still 0 indicating heteroscedasticity. Because the histogram and Q-Q plots suggest the residuals are not following a normal distribution the models without influential observations does not adhere to the normality assumption of regression modelling.

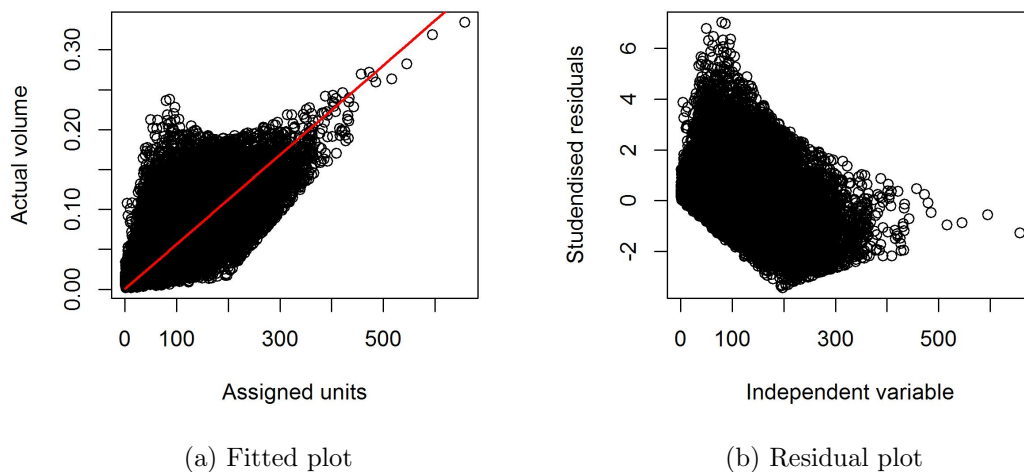


Figure 2.9: Fitted and studentised residual plots for the model 1 after influential observations were removed.

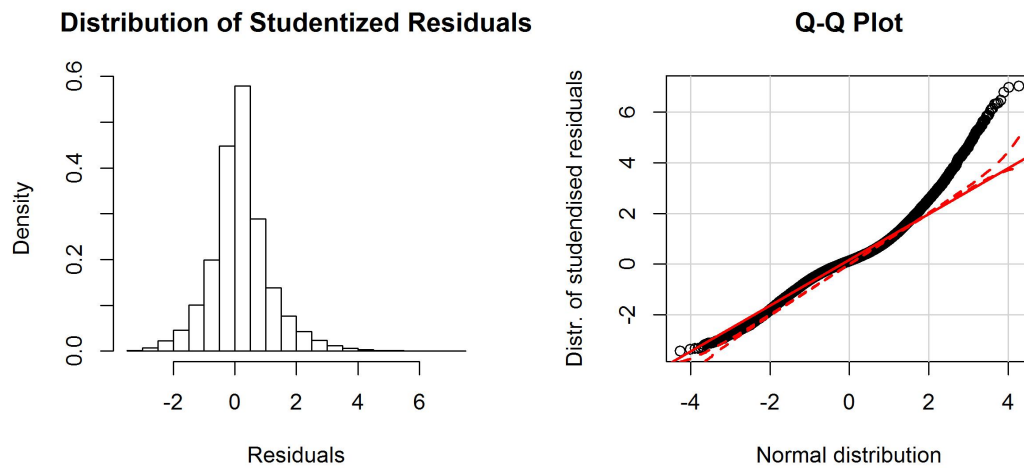


Figure 2.10: Residual histogram and Q-Q plot for model 1 after influential observations were removed.

Breusch-Pagan Test	Model 1 Volume/units	Model 2 Volume/volume	Model 3 Cartons/units	Model 4 Cartons/volume
χ^2	426.14	637.86	149.20	79.69
Degrees of freedom	0	0	0	0
p -value	0	0	0	0

Table 2.9: The Breusch-Pagan tests for the 4 models with influential observations removed.

2.6 Conclusion

The plots that shows the relationship between the independent variables and dependent variables shows a positive correlation. As the assigned units and volume increase so does the predicted cartons and predicted volume. These plots also suggested that heteroscedasticity exists. By implementing SLR, it can be confirmed by means of the residual plots and BP test that there does exist heteroscedasticity in the residuals. It was also found that the models with volume as independent variable provided a better fit. Furthermore the models are not BLUE, because heteroscedasticity exist in the residuals and the distribution of the residuals is not a normal distribution. In the next chapter multiple linear regression models are introduced.

Chapter 3

Multiple linear regression

The SLR models in the previous chapter displays heteroscedasticity. In this chapter a multiple linear regression (MLR) approach will be taken. This chapter are divided into three sections, namely a literature review on MLR, a discussion of the MLR models in the second section and the final section will contain a conclusion.

3.1 Literature review

Multiple linear regression (MLR) models build a relationship between two or more variables. The equation for an MLR model is

$$\hat{y}_j = b_0 + b_1x_{1j} + b_2x_{2j} + \dots + b_nx_{nj} \quad (3.1)$$

where \hat{y}_j is the predicted variable, b_0 the intercept and the other b_i 's are the coefficients of the dependent variables x_{ij} for $i = 1, \dots, n$. This equation can be seen as a plane in a multidimensional space. The b_i parameters can be found via the OLS method and the same BLUE assumptions apply in MLR as in SLR.

Heteroscedasticity must also be tested for in MLR models, where the distribution of the residuals are compared with the fitted values. Another problem that can occur in MLR is *multicollinearity*. Multicollinearity occurs when there exist strong relationships between independent variables. These relationships can cause variables to be insignificant or have unexpected coefficient signs [51].

If many variables are available that could be used in a model but there are uncertainty which variables will be useful for the model, there are different techniques on how to determine the best variables from a list of variables. Wilson *et al.* [51] mentions that the AIC in equation (2.11) and BIC (2.12) can be used to determine what variables to use from a list of independent variables. For instance, if a variable is added to a model and the BIC and AIC increase then the variable should not be used in the model. If a variable is removed from the model and the BIC and AIC decrease then the variable should not be included. This technique of choosing variables are called stepwise regression [4]. Steel and Uys [46] uses the Mallows' C_p (2.13) to select variables but states, the AIC, BIC and Mallows' C_p cannot be used in isolation, but the significance of variables should also be taken into account when doing stepwise regression.

Ghani *et al.* [20] forecast fish landing in Terengganu, Malaysia and first use a Pearson correlation to determine which variables to consider by analysing the Pearson correlation between variables. Independent variables with small correlations with the dependent variable is removed and combinations of variables with large correlations is also removed. Stepwise regression were further used on the rest of the variables, where a variable are added to the model, the fit are analysed along with the statistical significance. If a variable is statistically significant and the fit improved then the variable remains in the model. Uyanik *et al.* [49] forecast student test results by using 5 variables and decide to keep the variables based on variance increase factor (VIC) values of the independent variables and also condition indexes.

3.2 Models

The SLR models in Chapter 2 uses the total assigned volumes and total assigned units to predict cartons and volume. PEP has their own product hierarchy with different levels where products are divided into. The lowest level are SKU level and the highest level are company level, containing all products. One of these levels are business unit (BU). There are 11 BUs in PEP, namely, Essentials, Babies, FMCG, Home, Footwear, Kids clothing and accessories, Adult clothing and accessories, Cellular, Pepclub, Pepmarket and Sundries.

The Essentials BU contains schoolwear, umbrellas, stationary and eyewear, Babies have all the baby clothing and accessories like bottles and teethers, FMCG have products like soap, sweets, shampoo, deoderants, etc. The Home BU contains products like, cutlery, vases, placemats, containers, etc., where Footwear has all shoes for both women and men, while the Kids and Adult clothing and accessories BUs contains outer and underwear for both kids and adults. The Cellular BU contain products like cellphones, airtime, starter packs and Dstv's, Pepclub have all the financial products like funeral cover and money transfers. Pepmarket BU contains products like food, appliances and tools. The Sundries BU contains all the staff clothing.

The models described in this Chapter will only use, Babies, Kids and Adults clothing and accessories, Home, FMCG and Essentials BUs. Products from the Sundries, Cellular and Pepclub BUs does not get distributed as often as products from the other BUs and form only a small percent of the total distributions. All the footwear products are distributed from the Johannesburg and Durban DC and thus not included as well, because the data that are used are from the Kuilsrivier, Cape Town DC.

3.2.1 Business unit model

The BU model divide the assigned volumes and units into smaller volumes and units for each observation based on the 6 BUs. Just as in the SLR case, there will be four models. The first model is where volumes are assigned to a picking line and volumes are predicted. The second model is where units are assigned to the picking line and volumes predicted. The third is where volumes are assigned and cartons predicted and in the fourth, units are assigned and cartons are predicted. Table 3.1 provide a summary of each of the four BU sub models with a short description in column 4 for predicted volumes/cartons and assigned volumes/units. Model 1 will refer to BVV, model 2 to BVU, model 3 to BCV and model 4 will refer to as BCU and this notation will be used throughout this chapter.

Sub model	1	2	3	4
Predicted	Volume	Volume	Cartons	Cartons
Assigned	Volume	Units	Volume	Units
Model desc. BU	BVV	BVU	BCV	BCU
Category	CVV	CVU	CCV	CCU
BU SKU Count	BSVV	BSVU	BSCV	BSCU
Category SKU Count	CSVV	CSVU	CSCV	CSCU
BU Clothing Indicator	BCIVV	BCIVU	BCICV	BCICU
Category Clothing Indicator	CCIVV	CCIVU	CCICV	CCICU
BU Combined Variable	BCVV	BCVU	BCCV	BCCU
Category Combined Variable	CCVV	CCVU	CCC	CCCU

Table 3.1: This table provide a summary description of the four sub models used for the MLR model.

The BU model are modelled as,

$$y_j = \beta_0 + \sum_{i=1}^k \beta_i x_{ij} + \varepsilon_j \quad (3.2)$$

where y_j represent the dependent variable that are observed cartons or observed volumes, x_{ij} are the independent variables that are assigned volumes or assigned units for each BU i and observation j where $j = 1, \dots, n$ and $k = 6$ represent the total number of BUs. The parameter β_0 which is the intercept will equal 0 that will force the model through the origin and the β_i s are coefficients for each independent variable x_{ij} and ε_j is the error for observation j .

BU model results

Table 3.2 shows the coefficients and corresponding estimates, standard errors, t -values and p -values for each variable that are a BU for model BVV.

Variable	Estimate	Std. Error	t -value	p -value
BU1 - Babies	0.974	0.002	429.397	0
BU2 - Kids	0.998	0.006	156.521	0
BU3 - Adults	0.637	0.004	146.554	0
BU4 - Home	0.776	0.009	86.368	0
BU5 - FMCG	1.039	0.002	453.032	0
BU10 - Ess	1.022	0.001	727.603	0

Table 3.2: This table contains the coefficients and corresponding estimates, standard errors, t -values and p -values for the BU MLR model BVV.

According to the p -values, all the variables are significant to the model with p -values equal to 0 and it can be concluded that the estimate values differ from 0. The standard errors is also small, almost equal to 0. These results suggest that the variables are sufficient variables for the model. Table 3.3 shows the R^2 , Adjusted R^2 , residual standard error (RSE) and BP test for the BU MLR model BVV. The R^2 and adjusted R^2 are close to 1 at 0.97 indicating the model fit the data well. The RSE is also small indicating the model produces errors that is very small. The BP test

indicate the existence of heteroscedasticity in the residuals, with a p -value of 0, that means the null hypothesis that homoscedasticity exists are rejected and heteroscedasticity are assumed.

R^2	Adj. R^2	RSE	Breusch-Pagan test		
			χ^2	DF	p -value
0.97	0.97	0.02	3707.58	5	0

Table 3.3: This table contains the R^2 , adjusted R^2 , RSE and Breusch-Pagan heteroscedasticity test results for the model BVV.

The plot in Figure 3.1 support the assumption that heteroscedasticity exist as stated with the BP test. The variance of the residuals seem to increase as the fitted values increase and there are also observations that can be considered as outliers as some of the residual values are far from 0. The smallest residual value are less than -60. Thus the model BVV, does not alleviate the existence of heteroscedasticity and will not be a sufficient model for predicting volumes.

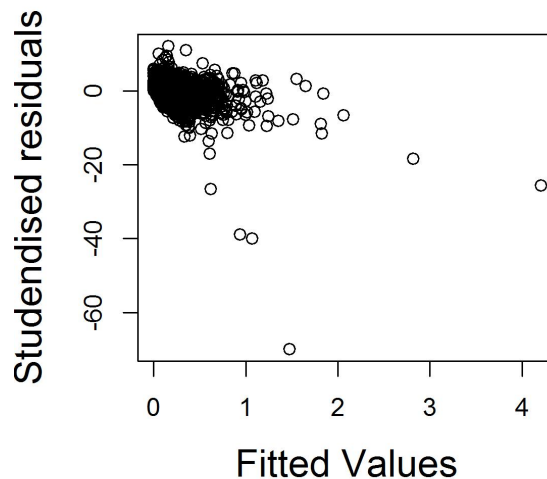


Figure 3.1: Studendised residual plot for the BU MLR model for BVV that is indicating heteroscedasticity in the residuals. The variance of the studendised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.

The results for the other models (BVU, BCV, BCU) show similar results and are displayed in Appendix B. Tables B.1–B.3, displays the coefficient statistics with standard errors, t -values and p -values for these models. The p -values are all 0 with small standard errors for the coefficients. The R^2 and adjusted R^2 in Table B.25 are also high and close to 1 indicating good model fit. The BP test results suggests heteroscedasticity exists and the residual plots in Figure B.1 support this assumption for models BVU, BCV and BCU. These plots shows that the variance of the residuals increase as the fitted values increase as the residuals form a funnel shape to the right.

3.2.2 Category model

In this Category model, the assigned variables are divided into 7 categories similar to the BU model. These 7 categories are, Accessories, Clothing, FMCG, Underwear, Home, Luggage and Stationary and is it not part of PEPs product structure. Accessories includes small products like, earrings, bracelets, necklaces, curtain hooks and gliders. Clothing includes all outerwear clothing for babies, kids and adults. FMCG includes all products sold in the FMCG BU. Underwear includes all underwear for kids and adults, the Home category includes all products from the Home BU except the kitchen textiles that is part of the Accessories category. Luggage includes all suitcases and bags, while Stationary includes all pens, pencils, books, etc.

This model will also have 4 sub models as summarised in Table 3.1 with the descriptions in row 5. The Category model is modelled similar to the BU model,

$$y_j = \beta_0 + \sum_{i=1}^k \beta_i x_{ij} + \varepsilon_j \quad (3.3)$$

where y_j represent the dependent variable that are observed cartons or observed volumes, x_{ij} are the independent variables that are assigned volume or assigned units for each Category i and observation j , $k = 7$ represent the total number of Categories and $j = 1, \dots, n$ are the number of observations. The parameter β_0 is the intercept and the β_i s are coefficients for each independent variable and ε_j is the error for the observation j . The intercept are forced throught the origin thus β_0 will equal 0.

Category model results

Table 3.4 contains the 7 category coefficients with the corresponding estimate values, standard error, t and p -values for model, CVV. The p -values are all 0 and the standard errors is small, indicating the category variables are significant to the model.

Variable	Estimate	Std. Error	t -value	p -value
CAT1 - Accessories	0.83174	0.003	304.129	0
CAT2 - Clothing	0.887	0.002	458.817	0
CAT3 - FMCG	1.02984	0.002	455.54	0
CAT4 - Underwear	1.21338	0.003	382.654	0
CAT5 - Home	0.67534	0.01	68.451	0
CAT6 - Luggage	1.01492	0.003	334.157	0
CAT7 - Stationary	1.70496	0.028	60.367	0

Table 3.4: This table contains the coefficients and corresponding estimates, standard errors, t -value and p -values for the Category MLR model for model CVV.

Table 3.5 displays the R^2 , Adjusted R^2 and BP test for the Category MLR model for sub model 1, CVV. The R^2 and adjusted R^2 are very high and close to 1 with a small RSE showing the good fit of the model. The BP test does show the existence of heteroscedasticity with a p -value of 0 and the residual plot in Figure 3.2 support this assumption. As the fitted values increase, the variance of the residuals also seem to increase.

The other models show similar results in Appendix B. The coefficient results are displayed in Tables B.4–B.6, the regression statistics in Table B.26 and residual plots in Figure B.2. All the

R^2	Adj. R^2	RSE	Breusch-Pagan test		
			χ^2	DF	p -value
0.97	0.97	0.02	1564.81	5	0

Table 3.5: This table contains the R^2 , adjusted R^2 , RSE and Breusch-Pagan heteroscedasticity test results for the Category model for the Category MLR model for sub model 1, VolVol.

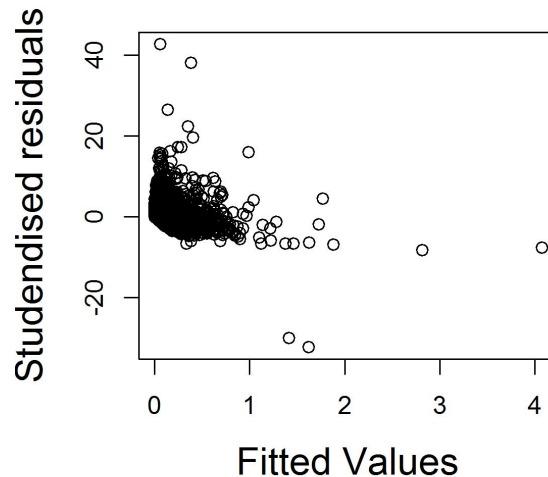


Figure 3.2: Studentised residual plot for the Category MLR model for sub model 1, VolVol that is indicating heteroscedasticity in the residuals. The variance of the studentised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.

p -values are 0 with small coefficient standard errors. The t -value for category 7, Stationary, tend to be lower than for the other variables. This shows that stationary does not have a significant influence on the predicted value as the other categories. The R^2 and adjusted R^2 are also high with small RSE's, although the RSE for cartons tend be higher at close to 1. The BP test still show that heteroscedasticity exist, just as in the BU models. The residual plots in Figure B.2 support the assumption that heteroscedasticity exists in the models, with a change in the variance of the residuals as the fitted values increase. Thus the BU and Category models does not show a clear indication that heteroscedasticity was sufficiently alleviated, compared to the SLR models in Chapter 2 and cannot be used and other models need to be explored.

3.2.3 A SKU count model

The SKU count model add a variable to the BU and Category models that are the number of SKUs that need to be picked for a store on a picking line or observation and will be called a sku count variable. When less SKUs need to be picked for a store, then the chances of the underutilisation of cartons could increase. With the combination of the BU and Category variables, the SKU count model then have a combination of the assigned volume or units to be picked and the number of

SKUs to be picked. The SKU count does not include the number of times a SKU must be picked. The SKU count model are defined as,

$$y_j = \beta_0 + \sum_{i=1}^k \beta_i x_{ij} + \beta_s s_j + \varepsilon_j \quad (3.4)$$

where y_j represent the dependent variable that are observed cartons or observed volumes, x_{ij} are the independent variables that are assigned volumes or assigned units for each BU or Category i and observation j , k represent the total number of BUs ($k = 6$) or Categories ($k = 7$) and $j = 1, \dots, n$ are the number of observations. The parameter β_0 is the intercept that equal 0 and the β_i s are coefficients for each independent variable. The β_s is the coefficient for the sku count variable s_j that indicates the number of SKUs to be picked in an observation and ε_j is the error for the observation j .

SKU count model results

The detailed results of the models BSVV and CSVV will be discussed. Tables 3.6 and 3.7 show the coefficients and corresponding estimate values, standard errors, t -values and p -values for the models BSVV and CSVV respectively. As in previous results the coefficients are all significant to the model, with small standard errors.

Variable	Estimate	Std. Error	t -value	p -value
BU1 - Babies	0.91214	0.002	432.07	0
BU2 - Kids	0.92656	0.006	160.687	0
BU3 - Adults	0.50269	0.004	123.063	0
BU4 - Home	0.65468	0.008	80.387	0
BU6 - FMCG	1.02072	0.002	493.23	0
BU10 - Ess	0.85809	0.002	448.951	0
SKU count	0.00063	0	114.282	0

Table 3.6: This table contains the coefficients and corresponding estimates, standard errors, t -value and p -values for the BU Sku count MLR model for model BSVV.

Variable	Estimate	Std. Error	t -value	p -value
CAT1 - Accessories	0.74342	0.003	264.483	0
CAT2 - Clothing	0.82119	0.002	409.817	0
CAT3 - FMCG	1.01083	0.002	469.744	0
CAT4 - Underwear	0.86728	0.005	165.312	0
CAT5 - Home	0.55101	0.009	58.237	0
CAT6 - Luggage	0.98581	0.003	340.371	0
CAT7 - Stationary	1.51743	0.027	56.573	0
SKU count	0.0006	0	80.413	0

Table 3.7: This table contains the coefficients and corresponding estimates, standard errors, t -value and p -values for the Category Sku count MLR model for model CSVV.

The R^2 , Adjusted R^2 , RSE and BP test results for the models BSVV and CSVV are shown in Table 3.8. The R^2 and Adjusted R^2 are close to 1 with small RSEs indicating good model fit.

The BP test also suggest heteroscedasticity with a p -value of 0. The plots in Figure 3.3 shows the studentised residuals for the BU and Category SKU count models. The variance of the residuals does seem to increase as the fitted values increase. There is also signs of influential observations, where some of the residual values are far from 0 and less than -60.

Model	R^2	Adj. R^2	RSE	Breusch-Pagan test		
				χ^2	DF	p -value
BSVV	0.98	0.98	0.02	3985.20	6	0
CSVV	0.97	0.97	0.02	1078.50	6	0

Table 3.8: This table contains the R^2 , adjusted R^2 , RSE and Breusch-Pagan heteroscedasticity test results for the BU and Category Sku count MLR models for sub model 1, VolVol. The R^2 and adjusted R^2 are close to 1 indicating good model fit.

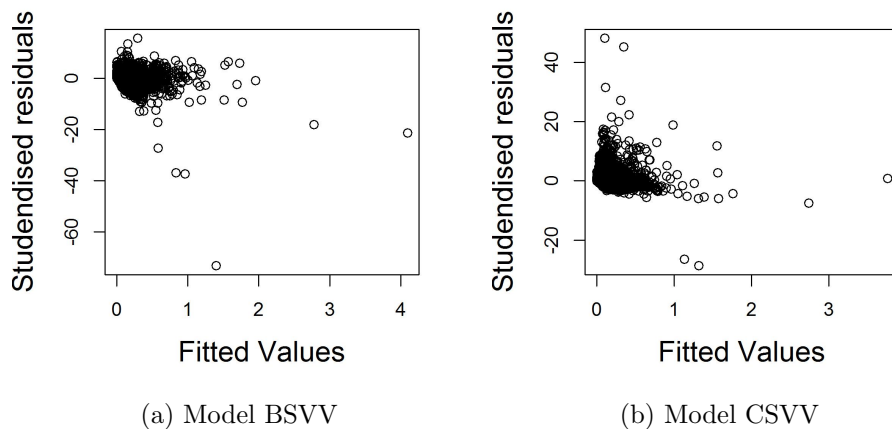


Figure 3.3: Studentised residual plots for the models BSVV and CSVV indicates heteroscedasticity in the residuals. The variance of the studentised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.

The coefficient statistic results to the other models are shown in Tables B.7–B.12, with the regression statistics in Tables B.27 and B.28 for the BU and category models. In the model BSVU, the Home BU's coefficient became insignificant to the model with a p -value of 0.64 shown in Table B.7. The stationary category have very small negative values in models CSVU and CSCU and indicates that the stationary category does not contribute to the predicted volumes or cartons as the other categories when units are the independent variable. The t -values for the stationary category is also smaller compared to the other categories in the Category SKU count model for each model. Under model CSCU in Table B.12 the Underwear category has a large p -value at 0.564 with a small estimate value indicating that for model CSCU the Underwear category does not affect the predicted cartons as much as the other categories.

The R^2 and adjusted R^2 values are still high at about 0.8 for the SKU count models. The BP test still suggest heteroscedasticity and the plots in Figures B.3 and B.4 also show signs of

heteroscedasticity in the residuals. The SKU count models does not seem to solve the problem of heteroscedasticity and another model will also be implemented in the next subsection.

3.2.4 Clothing indicator model

The clothing indicator model adds a binary dummy variable to each observation in the BU and category models. If the value is 1 then it means clothing products need to be picked for the observation and 0 if clothing products does not need to be picked for the observation. Clothing products tend to have larger volumes than the other products and due to the flexibility of clothing materials it can be folded and compressed to fit a carton and in turn improve carton utilisation. The smaller products like earrings, bracelets, gliders and other accessories, has a fixed form and small volumes compared to clothing. By only having to pick small items like accessories and other fixed formed products can lead to inefficient use of cartons. The assumption made in this model is, if clothing needs to be picked for a store on a picking line then the carton utilisation can improve.

In the category model, all clothing items were removed from the observations and the clothing indicator was added. The clothing category in the Category models, falls away. This ensures that there are not at any multicollinearity present in the models. The regression equation for the clothing indicator are

$$y_j = \beta_0 + \sum_{i=1}^k \beta_i x_{ij} + \beta_t t_j + \varepsilon_j \quad (3.5)$$

where y_j represent the dependent variable that are observed cartons or observed volume, x_{ij} are the independent variables that are assigned volumes or assigned units for each BU or Category i and observation j , k represent the total number of BUs or Categories and $j = 1, \dots, n$ are the number of observations. The parameter β_0 is the intercept and the β_i s are coefficients for each independent variable, β_t is the coefficient for the binary dummy variable t_j that indicate whether clothing need to be picked for the observation j or not and ε_j is the error for the observation j .

Clothing indicator results

This subsection discusses the results to the clothing indicator models BCIVV and CCIVV. The coefficient estimates, heteroscedasticity tests and residual plots of the other models are shown in Appendix B. Tables 3.9 and 3.10 displays the coefficient estimates of the BU and Category clothing indicator models BCIVV and CCIVV. All the variables in both models are significant according to Tables 3.9 and 3.10, but the dummy variables' estimates are very small and close to 0. For the model BCIVV the dummy estimate is 0.0289 while for the Category model it is 0.0289. This shows that the clothing indicator, does not add significant value to the predicted volumes.

The R^2 , adjusted R^2 , RSE and BP test results are displayed in Table 3.11 for the BU and Category models with the clothing indicator. The R^2 and Adjusted R^2 are still high at 0.89 but lower than the BU model and Category model without the clothing indicator and lower than the SKU count model. The RSE increased slightly from 0.02 to 0.03 from the previous models. Thus for the clothing indicator the the fit is still good, but worse than the previous models.

The BP tests in Table 3.11 still suggests heteroscedasticity, and the χ^2 values are higher than in previous models, thus heteroscedasticity seem to be more of a problem in this model. Figure 3.4

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
BU1 - Babies	1.28465	0.01	125.105	0
BU2 - Kids	1.04113	0.012	88.044	0
BU3 - Adults	0.56738	0.008	69.208	0
BU4 - Home	0.80564	0.017	48.455	0
BU6 - FMCG	1.0673	0.004	250.043	0
BU10 - Ess	1.15358	0.004	274.657	0
Clothing Dummy	0.02894	0	135.558	0

Table 3.9: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the BU Clothing indicator MLR model BCIVV.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
CAT1 - Accessories	0.87823	0.005	163.117	0
CAT3 - FMCG	1.08451	0.004	247.026	0
CAT4 - Underwear	1.21034	0.007	174.827	0
CAT5 - Home	0.66323	0.019	34.475	0
CAT6 - Luggage	1.09261	0.006	185.151	0
CAT7 - Stationary	1.15475	0.055	21.014	0
Clothing Dummy	0.02835	0	122.486	0

Table 3.10: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the Category Clothing indicator MLR model CCIVV.

Model	R^2	Adj. R^2	RSE	Breusch-Pagan test		
				χ^2	DF	<i>p</i> -value
BCIVV	0.89	0.89	0.03	4340.37	6	0
CCIVV	0.89	0.89	0.03	1180.61	6	0

Table 3.11: This table contains the R^2 , adjusted R^2 , RSE and Breusch-Pagan heteroscedasticity test results for the BU and Category Clothing indicator MLR models, BCIVV and CCIVV.

displays the studentised residual plots for models BCIVV and CCIVV. These plots indicate that heteroscedasticity exist, as there is a funnel shape to the right in both plots.

The coefficient results for models BCIVU, BCICV, CCIVU and CCICU in Tables B.13, B.14 and B.16 and B.17 shows that the clothing indicator dummy variable is not influential on the predicted volumes and cartons than the other coefficients, as the dummy variables has a smaller estimates compared to the other coefficients. The estimate of the dummy variable on model BCICU are higher than other estimates of the other variables and shows that the clothing count dummy variable has a significant influence on the predicted cartons when assigned units are the independent variable. The dummy variable is also significant to all the models with a *p*-value of 0.

In all the models the stationary category is not as significant to the models compared to the other categories in the models for clothing indicator and was also identified in previous category models. The R^2 and adjusted R^2 displayed in Tables in B.29 and B.30 have values around 0.8, with an

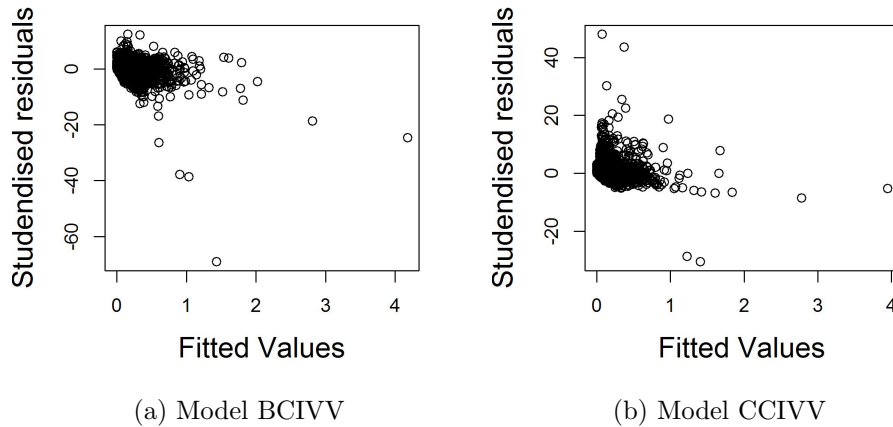


Figure 3.4: Studendised residual plot for the BU and Category Clothing indicator MLR model for sub model 1, VolVol that is indicating heteroscedasticity in the residuals. The variance of the studendised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.

increase in RSEs. This suggest that the fit is not as good as in the SKU count model, but the BP test has smaller χ^2 values than the previous models, which indicates that heteroscedasticity lessen. The residual plots in Figure B.5 and B.6 suggest that heteroscedasticity is still a problem with the clothing indicator model. Thus other models need to be considered.

3.2.5 Combined variable model

This subsection provide the combined variable model where both SKU count and Clothing indicator are added to the BU and Category models. The clothing category are removed as a variable from the Category model and all clothing products are removed from the observations. This combined variable model are modelled as,

$$y_j = \beta_0 + \sum_{i=1}^k \beta_i x_{ij} + \beta_s s_j + \beta_t t_j \varepsilon_j \quad (3.6)$$

The dependent variable y_j represent observed cartons or observed volumes, x_{ij} are the independent variables that are assigned volumes or assigned units for each BU or category i and observation j , k represent the total number of BUs or categories and $j = 1, \dots, n$ are the number of observations. The parameter β_0 is the intercept and the β_i s are coefficients for each independent variable, β_s and β_t are the coefficients for the variables s_j and t_j respectively that displays the number of SKUs in an observation and indicate whether clothing needs to be picked for the obervation or not and ε_j is the error for the observation j .

Combined variable model results

The models BCVV and CCVV for both BU and Category models will be discussed in this section with the results of the other sub models in Appendix B. Tables 3.12 and 3.13 shows the coefficient estimates, standard errors, t -value and p -value of models BCVV and CCVV respectively. In both

tables all the variables has a p -value of 0 indicating that all the variables are significant to the model. The additional variables has very small estimates compared to the other variables. The SKU count estimate is 0.00089 and 0.00127 for the model BCVV and CCVV, while the clothing dummy has estimates of 0.01633 and 0.01471 for models BCVV and CCVV respectively. This indicate that these additional variables does not add any significant value to the predicted volumes.

Variable	Estimate	Std. Error	t -value	p -value
BU1 - Babies	1.16269	0.01	115.718	0
BU2 - Kids	0.95633	0.011	83.537	0
BU3 - Adults	0.44206	0.008	54.516	0
BU4 - Home	0.66646	0.016	41.312	0
BU6 - FMCG	1.05101	0.004	255.472	0
BU10 - Ess	0.96001	0.005	192.858	0
SKU count	0.00089	0	66.591	0
Clothing dummy	0.01633	0	58.469	0

Table 3.12: This table contains the coefficients and corresponding estimates, standard errors, t -value and p -values for the BU Combined dummy variable model, BCVV.

Variable	Estimate	Std. Error	t -value	p -value
CAT1 - Accessories	0.72904	0.006	131.441	0
CAT3 - FMCG	1.04297	0.004	246.21	0
CAT4 - Underwear	0.62338	0.01	59.443	0
CAT5 - Home	0.45926	0.019	24.683	0
CAT6 - Luggage	1.03836	0.006	182.457	0
CAT7 - Stationary	0.90401	0.053	17.172	0
SKU count	0.00126	0	72.152	0
Clothing dummy	0.01471	0	50.554	0

Table 3.13: This table contains the coefficients and corresponding estimates, standard errors, t -value and p -values for the Category Combined dummy variable model, CCVV.

Tables 3.14 shows the regression statistics, R^2 , adjusted R^2 , RSE and BP test results for models BCVV and CCVV. Comparing Table 3.14 to Table 3.11, there is not a significant difference. The R^2 and adjusted R^2 increase by 0.01, the RSEs stayed the same, the χ^2 for the BP test dropped for both the BU model and Category model. Thus according to the BP test, heteroscedasticity improved in these BCVV and CCVV.

Model	R^2	Adj. R^2	RSE	Breusch-Pagan test		
				χ^2	DF	p -value
BU	0.90	0.90	0.03	4032.87	7	0
Category	0.90	0.90	0.03	1037.15	7	0

Table 3.14: This table contains the R^2 , adjusted R^2 , RSE and Breusch-Pagan heteroscedasticity test results for the BU and Category Combined variable models, BCVV and CCVV

The plots in Figure 3.5 shows the studentised residual plots for the BU and Category combined variable models. Both plots have a funnel shape form to the right that suggest heteroscedasticity.

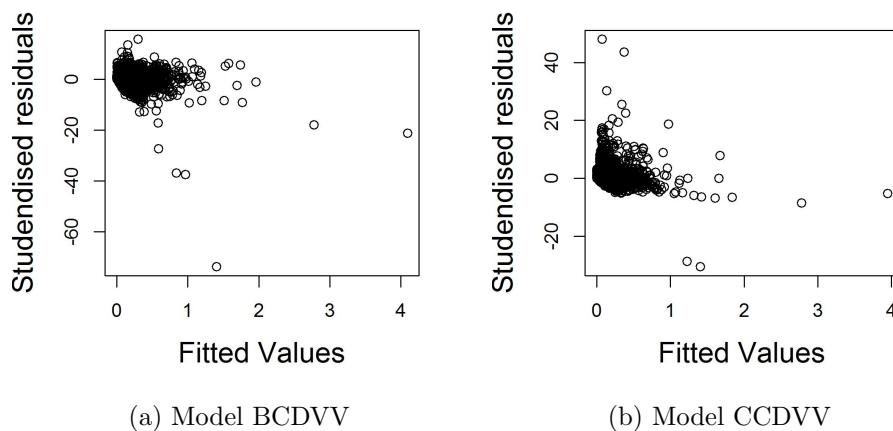


Figure 3.5: Studentised residual plots for the models BCVV and CCVV that is indicating heteroscedasticity in the residuals with a funnel shape to the right. The variance of the studentised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.

The results to the other sub models shows similar results. Tables B.19 – B.21 shows the coefficient estimates, standard errors, t -value and p -value for models BCVV, BCCV and BCCU. All the variables in these tables are significant to the model. In models BCVV and BCCV the additional variables, plays less of an important role in the prediction of volumes and cartons as their estimate in these models are small compared to the other variables. In BCCU the estimates does influence the dependent variable as it has larger estimates than the other variables.

The Category model's coefficient statistics are shown in Tables B.22 – B.24. The category models with CCVV and CCCV also has estimates to the additional variables that is small compared to other variables, but in CCCU the additional variables' estimates are large compared to the other variables. The stationary category has very small negative values as estimates compared to the other variables in sub models, 2 and 4. The stationary show small contributions to the dependent variables as in previous category models as well. This means that by categorising products according to Accessories, Clothing, FMCG, Underwear, Home, Luggage and Stationary, the stationary category does not add much value to the dependent variable in sub models 2 and 4. In sub model 1, the stationary category has larger estimates and more influence on the dependent variable, but the RSE is also high compared to the other models for stationary.

The regression statistics for models 2, 3 and 4 are shown in Tables B.31 and B.32 for BU and Category respectively. These results does not differ much from the previous models' results. The R^2 and adjusted R^2 stayed at about 0.8 for the models and the RSE for model 2 changes by 0.01, while the RSE for Ctns stays at about 1. The BP test results does not differ much from each other and all indicate the existence of heteroscedasticity with p -values of 0.

The residual plots B.8 and B.7 also suggest heteroscedasticity as there exist a funnel shape to the right in the residuals for both figures. The combined variable approach also does not give a satisfactory model and in the next chapters other models will be discussed.

3.3 Conclusion

The MLR models discussed in this chapter result in high R^2 values and small RSEs and the coefficients on most of the models are significant with p -values < 0.05 . The BP tests on all the MLR models still suggests heteroscedasticity exists and the χ^2 on these MLR models does not differ significantly from the SLR models. On all the models there are a change in the variance of the studentised residuals as the fitted values increase, which support the assumption from the BP tests that heteroscedasticity exist. The MLR models does not provide models that significantly improve on the regression statistics and heteroscedasticity than the SLR models and thus other types of models will be investigated in subsequent chapters.

Chapter 4

Transformation analysis

All the regression models considered in previous chapters, displays heteroscedasticity in their residuals. Heteroscedasticity can normally be fixed by transforming the variables [22] and in this chapter different types of transformations are described and used on the variables from the models in Chapter 2. The chapter is divided into four sections. The first section contains a literature review on transformed regression models and in the subsequent sections, different transformations are implemented on four regression models.

4.1 Literature on transformed models

Figure 2.4 shows the variance of the predicted volume increase as the assigned volume and units increase. Although less so in the case where volume are the independent variable and predicted volume the dependent variable. The next subsections describe different types of transformations.

4.1.1 Log-linear transformation

The log-linear transformation transforms the dependent variable with a logarithm. The log-linear model can be written as

$$\ln y_j = b_0 + b_1 x_j, \quad (4.1)$$

and the predicted value can then be found by

$$\hat{y}_j = e^{b_0 + b_1 x_j}. \quad (4.2)$$

OLS can be applied to equation (4.1) to calculate the estimate of b_0 and b_1 . The same linear regression assumptions as discussed in Chapter 2 holds for equation (4.1). The only difference between equation (2.2) and (4.1) are the different values that is used for the dependent variable.

The interpretation of the model's coefficients is also different to the standard regression model. Log transformations works with percentages, where in equation (4.1), 1 unit change in x leads to a $100(e^{b_1} - 1)$ change in y . Suppose

$$\ln y_1 = b_0 + b_1 x \quad (4.3)$$

and

$$\ln y_2 = b_0 + b_1(x + 1). \quad (4.4)$$

By calculating equation (4.4) minus equation (4.3), it follows that

$$\ln(y_2) - \ln(y_1) = \ln\left(\frac{y_2}{y_1}\right) = b_1. \quad (4.5)$$

Taking the exponential on both sides, subtracting by 1 and then multiplying by a 100 on both sides of equation (4.5), it follows that

$$100\left(\frac{y_2}{y_1} - 1\right) = 100(e^{b_1} - 1), \quad (4.6)$$

where $100\left(\frac{y_2}{y_1} - 1\right)$ is the percentage change in y . Miller *et al.* [25] uses a weighted log-linear regression model in order to forecast radioactivity under a certain depth in ground. Pearl *et al.* [40] use a log-linear model to predict CA 125 levels in patients having Ovarian cancer after they were treated with taxol. The CA 125 levels drops at an exponential rate. It initially drop fast and then after a while the levels drops at a slower rate.

4.1.2 Linear-log transformation

This transformation keeps the dependent variable the same and transforms the independent variable with a logarithm. The resulting regression equation is

$$y_j = b_0 + b_1 \ln x_j. \quad (4.7)$$

The formula for the predicted value stays the same. The linear-log transformation can be interpreted as 1% change in x is associated with $b_1/100$ unit change in y . This can be shown in a similar way to the log-linear transformation. Suppose,

$$y_1 = b_0 + b_1 \ln(x_1) \quad (4.8)$$

and

$$y_2 = b_0 + b_1 \ln(1.01x_1), \quad (4.9)$$

where $1.01x_1$ is a 1% increase in x_1 . Subtracting equation (4.8) from (4.9) gives

$$y_2 - y_1 = b_1 \ln\left(\frac{1.01x_1}{x_1}\right) \approx b_1/100. \quad (4.10)$$

Frank *et al.* [19] uses a linear-log transformation to regress the heart rate of cyclist against time as the cyclist have to increase power output after every two minutes.

4.1.3 Log-log transformation

Log-log transformations gets the logarithm of both dependent and independent variables. The resulting regression equation is,

$$\ln y_j = b_0 + b_1 \ln x_j. \quad (4.11)$$

The formula for the predicted value becomes

$$\hat{y}_j = e^{b_0 + b_1 \ln(x_j)}. \quad (4.12)$$

Log-log transformations can be interpreted as 1% change in x is associated with a $100(1.01^{b_1} - 1)\%$ change in y . Suppose

$$\ln y_1 = b_0 + b_1 \ln x_1 \quad (4.13)$$

and

$$\ln y_2 = b_0 + b_1 \ln(1.01x_1) \quad (4.14)$$

where $1.01x_1$ is a one percentage increase in x_1 . Subtracting equation (4.13) from (4.14), yields

$$\ln y_2 - \ln y_1 = b_1 \ln \left(\frac{1.01x_1}{x_1} \right), \quad (4.15)$$

which can be simplified to

$$\ln \left(\frac{y_2}{y_1} \right) = b_1 \ln(1.01). \quad (4.16)$$

By taking the exponential on both sides, subtracting 1 and multiplying by 100 on both sides of equation (4.16), the percentage change in y can be found as

$$100 \left(\frac{y_2}{y_1} - 1 \right) = 100(1.01^{b_1} - 1). \quad (4.17)$$

Xiao *et al.* [52] proves that there is not a significant difference between a linear regression model on log transformed data compared to a non-linear model without log transformed data, but rather states that the residuals can give an indication what type of transformed model must be used. The next subsection describe another transformation method called quadratic transformation.

4.1.4 Quadratic transformation

The quadratic transformation gets the square-root of the dependent variable and can be written as

$$\sqrt{y_j} = b_0 + b_1 x_j, \quad (4.18)$$

where the predicted value is calculated as,

$$\hat{y}_j = (b_0 + b_1x_j)^2 = b_0^2 + 2b_0b_1x_j + b_1^2x_j^2. \quad (4.19)$$

This is also the equation of a parabola. By replacing $b_0^2 = c$, $2b_0b_1 = b$ and $b_1^2 = a$, equation (4.19) can be re-written as

$$\hat{y} = c + bx + ax^2. \quad (4.20)$$

The turning point of the parabola can be calculated as $\frac{-b}{2a}$. Assuming x is positive, if b and a is positive, then the y increase as x increase but increase at a slower rate as the x value increases. If b is positive and a negative then the y increases until the point where $x = \frac{-b}{2a}$ and then start to decline. If b is negative and a is positive the y decrease until $x = \frac{-b}{2a}$ and then start to increase. If both b and a are negative then the y decrease, but decrease at a slower rate as the x value increases. Figure 4.1 give examples of different values of a and b and show how the curve responds to the different values of a and b . A quadratic transformation/regression is also a special case of polynomial regression [48] with degree 2.

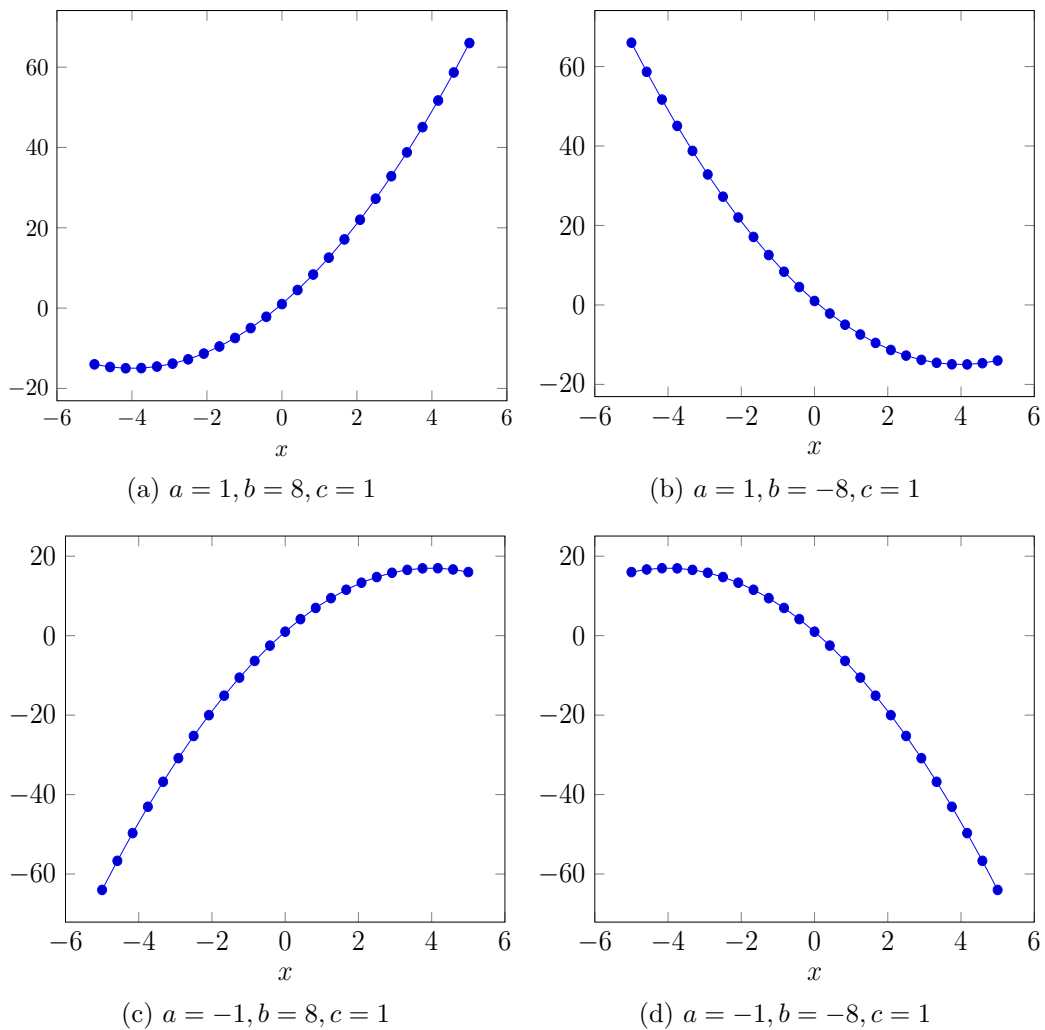


Figure 4.1: Quadratic equation plots with different signs for b and a .

In the next three sections variables are transformed with different transformations. Scatter plot analysis and BP tests are performed in order to choose the best linear relationship.

4.2 Logarithmic transformation

In this section the log-linear, linear-log and log-log transformations will be implemented on the four models described in the Chapter 2. Each subsection provide a different transformed model and graphical analysis of the scatter plots that depicts the relationship between the variables. There are four variables described in Table 2.1 with y_{ij} the dependent variable and x_{ij} the independent variable in each of the four models.

4.2.1 Log-linear model

Substituting the variables of equation (4.1) with the y and x variables in Table 2.1, the plots in Figure 4.2 are obtained. The plots are scatter plots displaying the relationship between the independent variable and natural logarithmic dependent variable. The natural logarithm on the dependent variable causes the dependent variable to initially increase at a high rate and slow down as the independent variable increases. The plots 4.2(c) and 4.2(d) with $\ln(\text{predicted ctns})$ as dependent variable, form lines in the plots due to the integers of the carton values. The variance of the dependent variables $\ln y$ on all four plots changes as the independent variable increase which indicate signs of heteroscedasticity. The transformation causes the relationship between the dependent and independent variables not to be linear and a fitted line would not fit data in the scatter plots well.

Table 4.1 shows the BP test results of the 4 models with the log-linear transformation. The table contains the model number which refer to the same model numbers as in Table 2.2. In this table the dependent variables, predicted volume and cartons are transformed by a natural logarithm and the variables that correspond to the model number is also shown in Table 4.1. The χ^2 , degrees of freedom and p -value of the BP test results for each model is shown in the last three columns. The p -value of the BP test results are 0 for all 4 models indicating heteroscedasticity exist in the residuals.

Model	Dependent variable	Independent variable	Breusch-Pagan test		
			χ^2	Degrees of freedom	p -value
1	$\ln(\text{actual volume})$	assigned units	353.89	1	0
2	$\ln(\text{actual volume})$	assigned volume	285.29	1	0
3	$\ln(\text{actual cartons})$	assigned units	1048.99	1	0
4	$\ln(\text{actual cartons})$	assigned volume	874.69	1	0

Table 4.1: Breusch-Pagan heteroscedasticity test results for the log-linear transformations.

Due to the large number of observations, the p -values are all close to or equal to zero and the χ^2 of the different transformations will be compared to each other in order to see any improvement on heteroscedasticity.

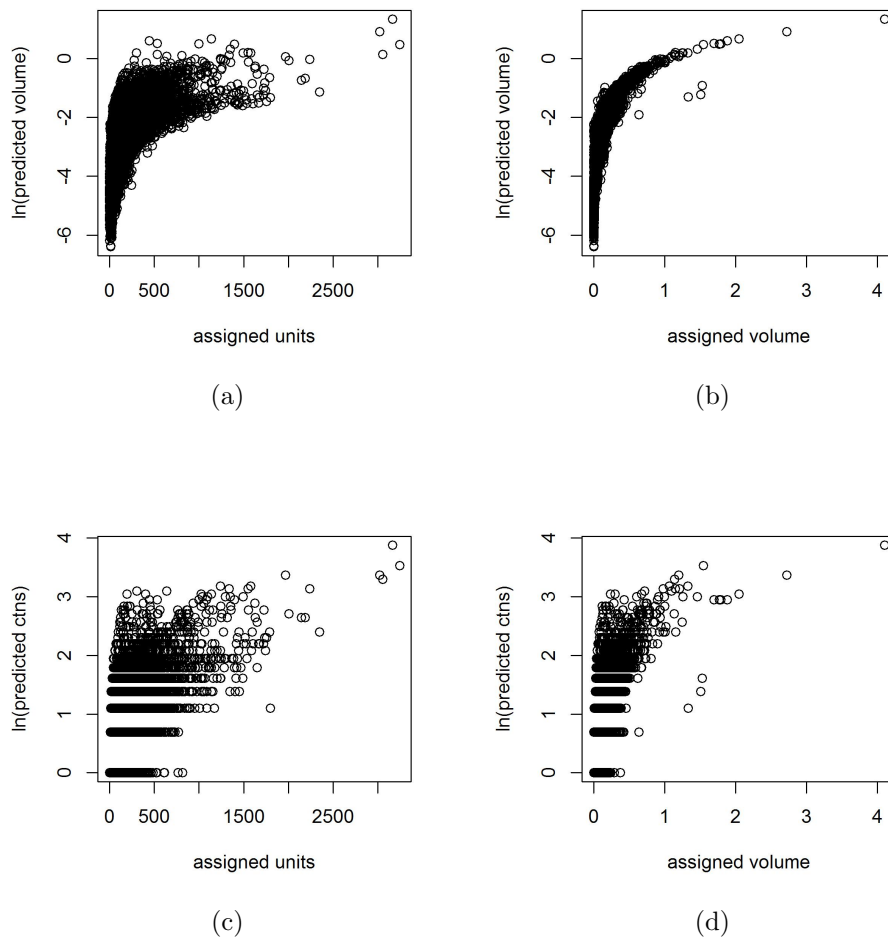


Figure 4.2: Scatter plots between assigned volume/units and predicted volume/cartons with a natural logarithmic for the log-linear transformation.

4.2.2 Linear-log model

Substituting the same set of variables into Equation (4.7) yields the scatter plots displayed in Figure 4.3. The plots in Figure 4.3 shows that the dependent variable initially increases slowly, but faster as the transformed independent variable increases. Heteroscedasticity still exist as the variance of the dependent variable in all four plots increases as the independent variable increases. According to the χ^2 values in Table 4.2, for the linear-log models, the heteroscedasticity did not improve as the χ^2 values are about 10 times the χ^2 values of the log-linear model. This increase in χ^2 values are due to the shape formed by the scatter plot. A normal regression line would not fit the data well.

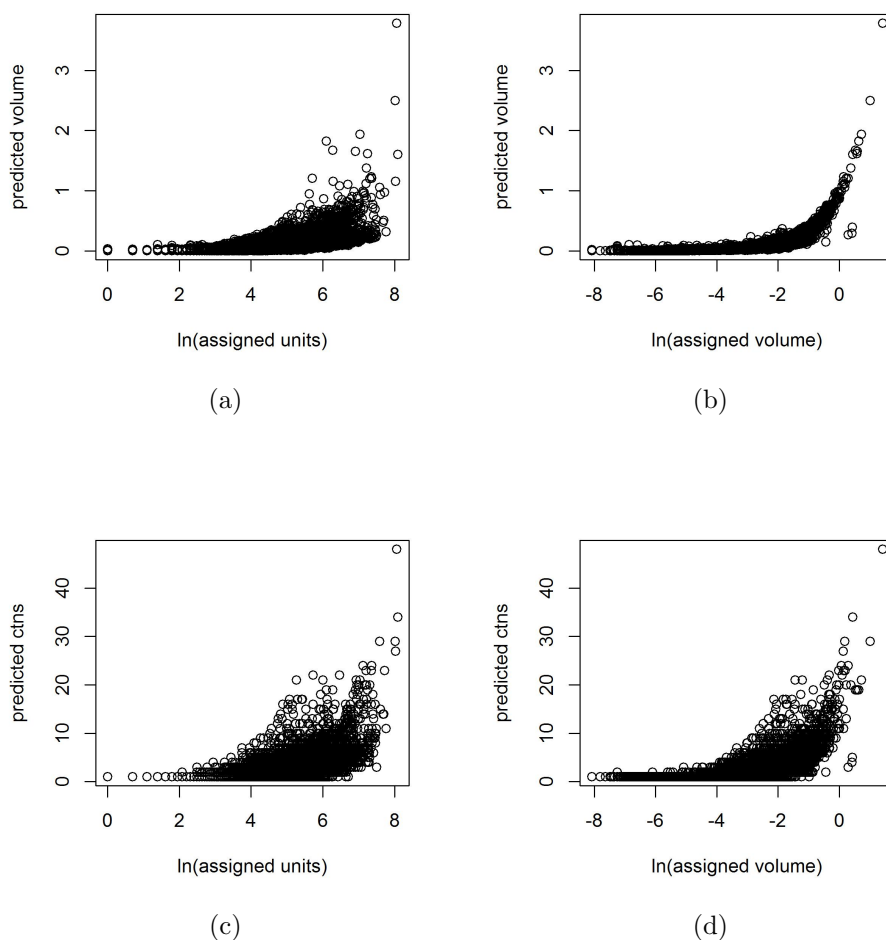


Figure 4.3: Scatter plots between assigned volume/units and predicted volume/cartons with a natural logarithmic for the linear-log transformation.

Model	Dependent variable	Independent variable	Breusch-Pagan test		
			χ^2	Degrees of freedom	p -value
1	actual volume	ln(assigned units)	7107.09	0	0
2	actual volume	ln(assigned volume)	8236.50	0	0
3	actual cartons	ln(assigned units)	10087.93	0	0
4	actual cartons	ln(assigned volume)	8976.70	0	0

Table 4.2: Breusch-Pagan heteroscedasticity test results for the linear-log transformations.

4.2.3 Log-log model

If the set of variables in Table 2.1 is substituted in equation (4.11), then scatter plots in Figure 4.4 are obtained. The variance of the dependent variable does not change as much as in the linear-log and log-linear transformations, but there is still some change in the variance particularly in Figures 4.4(c) and 4.4(d) where predicted cartons are used. A regression line might also fit the

data better than the previous two transformations. Table 4.3 shows the BP tests for the four models with a log-log transformation. The χ^2 values improved compared to the linear-log model, but it only improved on model 4 compared to the log-linear transformation.

Model	Dependent variable	Independent variable	Breusch-Pagan test		
			χ^2	Degrees of freedom	p -value
1	ln(actual volume)	ln(assigned units)	500.53	1	0
2	ln(actual volume)	ln(assigned volume)	3223.07	1	0
3	ln(actual cartons)	ln(assigned units)	2753.70	1	0
4	ln(actual cartons)	ln(assigned volume)	347.07	1	0

Table 4.3: Breusch-Pagan heteroscedasticity test results for the log-log transformations.

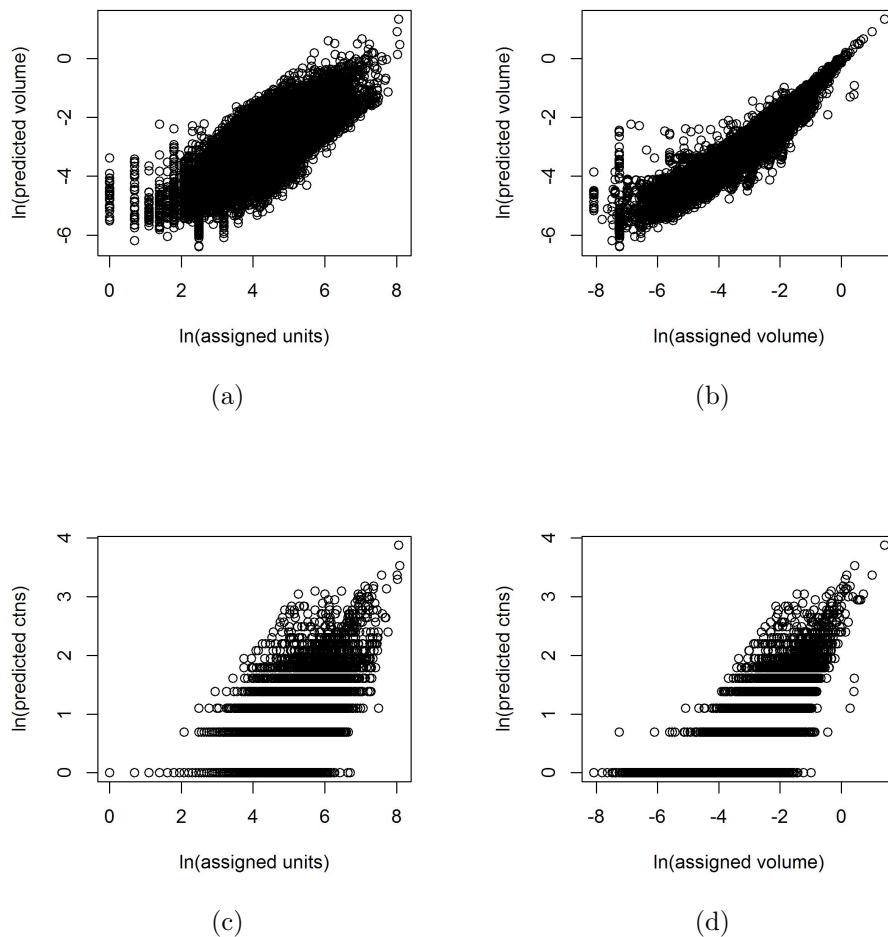


Figure 4.4: Scatter plots between assigned volume/units and actual volume/cartons with a natural logarithmic for the log-log transformation.

The log-linear model provide the most improvement in terms of heteroscedasticity according to the BP test, but the transformed models is not linear. The log-log model for predicting volumes has

the closest linear form of the three logarithmic transformations. The logarithmic transformations does not give a definitive transformation that fully satisfies the regression assumptions that will ensure a good model for predicting cartons and volume. The next section describes six different quadratic transformations on the variables in Table 2.1.

4.3 Quadratic transformation

This section introduces six quadratic transformations. Three of the transformations uses a square root to transform the dependent and/or independent variables while the other three uses a fourth root transformation to transform the dependent and/or independent variables. The six transformations are listed in Table 4.4. These transformed models have an intercept value included as the relationship between transformed variables might not go through the origin, even though the original model does. If a transformed model with an intercept is converted back, then the model should still go through the origin. The y and x variables in the equations of Table 4.4 are the dependent and independent variables respectively. This results into to 24 different types of relationships between dependent and independent variables as y can represent predicted cartons or volume and x assigned volume or assigned units.

Method	Transformation	Regression equation
Squared-linear	Transform the dependent variable with a square root	$y^{1/2} = b_0 + b_1x$
Linear-squared	Transform the independent variable with a square root	$y = b_0 + b_1x^{1/2}$
Squared-Squared	Transform both variables with a square root	$y^{1/2} = b_0 + b_1x^{1/2}$
Fourthroot-linear	Transform the dependent variable with a fourth root	$y^{1/4} = b_0 + b_1x$
Linear-fourthroot	Transform the independent variable with a fourth root	$y = b_0 + b_1x^{1/4}$
Fourthroot-fourthroot	Transform both variables with a fourth root	$y^{1/4} = b_0 + b_1x^{1/4}$

Table 4.4: Six versions of the quadratic model where the dependent and/or independent variable gets transformed by either a fourth root or by a squared root.

The scatter plots shown in Figure 4.5 displays the relationships between the variables for the squared-linear transformation. The variance of the dependent variable increase on all the plots as the independent variable increases. The relationship between the dependent and independent variables is also not linear. Table 4.5 shows the BP test results for the squared-linear transformations. The p -values are all 0, suggesting heteroscedasticity exist in the residuals. The BP tests for the other quadratic transformations are shown in Tables D.1–D.5. The p -values in these tables are all 0, indicating that heteroscedasticity exists. The linear-fourth model has the smallest χ^2 values showed in Table D.4, compared to the other transformations, which means the linear-fourth transformation alleviate heteroscedasticity more than the other quadratic transformations. The plots in Figure D.1–D.5 displays the scatter plots of the other quadratic transformations. The relationships between the variables as in Figure 4.5 is not linear and the variance of the dependent variable changes as the independent variable increase, which leads to heteroscedasticity. The scatter plot in Figure D.5 that displays the relationship of the squared-squared transformations, displays a plot that seems to be the closets to a linear relationship.

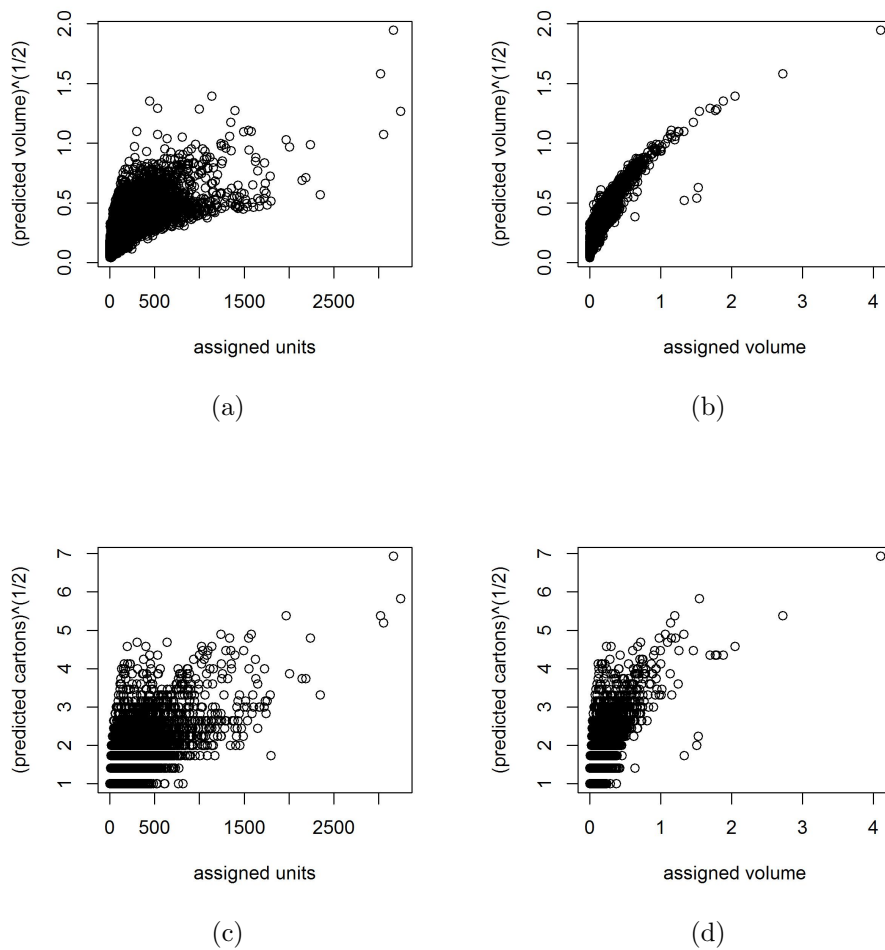


Figure 4.5: Scatter plots between assigned volume/units and predicted volume/cartons with a squared-linear transformation.

Model	Dependent variable	Independent variable	Breusch-Pagan test		
			χ^2	Degrees of freedom	p -value
1	$(\text{predicted volume})^{1/2}$	assigned units	11408.41	1	0
2	$(\text{predicted volume})^{1/2}$	assigned volume	7922.58	1	0
3	$(\text{predicted cartons})^{1/2}$	assigned units	6693.95	1	0
4	$(\text{predicted cartons})^{1/2}$	assigned volume	9228.39	1	0

Table 4.5: Breusch-Pagan heteroscedasticity test results for the squared-linear transformations.

The quadratic transformation does improve on heteroscedasticity compared to the standard regression model in Chapter 2, but a perfect linear relationship does not hold anymore. The next section weighs the variables and with a combination of logarithmic and quadratic transformations.

4.4 Weighted regression model

Gujarati [22] mentions a technique to remove heteroscedasticity by dividing the dependent and independent variables by a square root of the independent variable. This technique is called weighted least squares regression (WLS). Each observation gets a weight that is equal to $w_j = 1/\sqrt{x_j}$ where x_j is the independent variable and w_j represent the weight. A WLS regression equation are $w_j y_j = w_j(b_0 + b_1 x_j)$.

The weighted models in this section multiply the dependent variables, which is the predicted volumes and predicted cartons and the independent variable assigned units, by the assigned volume. Thus, the assigned volume for each observation acts as a weight for each observation. This also cause the number of models to reduce from 4 in the previous sections to 2. The idea of this weighted transformation is to remove the integer problem from predicting cartons and fix heteroscedasticity.

The logarithmic and quadratic transformation sections showed that when both dependent and independent variables are logarithmic or quadratic transformed, then a more linear relationship are obtained between the transformed variables. The weighted variables are also transformed by a logarithm and by a fourth root to add another 4 models to the analysis. In total for the weighted regression model, six models will be analysed. These models are presented in Table 4.6 where, y_1 is predicted volume, y_2 predicted cartons, x_1 assigned units and x_2 assigned volume. When implementing the six models in Table 4.6 the scatter plots in Figure 4.6 are obtained.

Model description	Regression equation
Volume weighted	$x_2 y_1 = b_0 + b_1 x_2 x_1$
Carton weighted	$x_2 y_2 = b_0 + b_1 x_2 x_1$
Log volume weighted	$\ln(x_2 y_1) = b_0 + b_1 \ln(x_2 x_1)$
Log carton weighted	$\ln(x_2 y_2) = b_0 + b_1 \ln(x_2 x_1)$
Quadratic volume weighted	$(x_2 y_1)^{1/4} = b_0 + b_1 (x_2 x_1)^{1/4}$
Quadratic carton weighted	$(x_2 y_2)^{1/4} = b_0 + b_1 (x_2 x_1)^{1/4}$

Table 4.6: Six weighted regression models.

Figures 4.6(a) and 4.6(b) are the scatter plots that displays the relationship between variables of the Volume weighted and Carton weighted models respectively. Figures 4.6(c) and 4.6(d) represent the scatter plots for the Log volume weighted and Log carton weighted models. Figures 4.6(e) and 4.6(f) represent the scatter plots for the Quadratic volume weighted and Quadratic carton weighted models respectively.

Analysing the plots in Figure 4.6, the weighted model without transformations, Figures 4.6(a) and 4.6(b), show changes in the variance of the dependent variable, where the variance of the dependent variable increases as the independent variable increases. There are also some outlier observations that can possibly influence regression coefficients. Figures 4.6(e) and 4.6(f) with a fourth root transformation on the weighted model, also show changes in the variance of the dependent variable. Figures 4.6(c) and 4.6(d) that shows the scatter plots for the Log volume weighted and Log carton weighted models does not show much change in the variance of the dependent variable. The heteroscedasticity test results for the weighted models are displayed in Table 4.7.

The log weighted models have smaller χ^2 values than weighted and quadratic weighted models. It is also less than the SLR model shown in Tables 2.6 and 2.9. Figure 4.6(c), predicting volume,

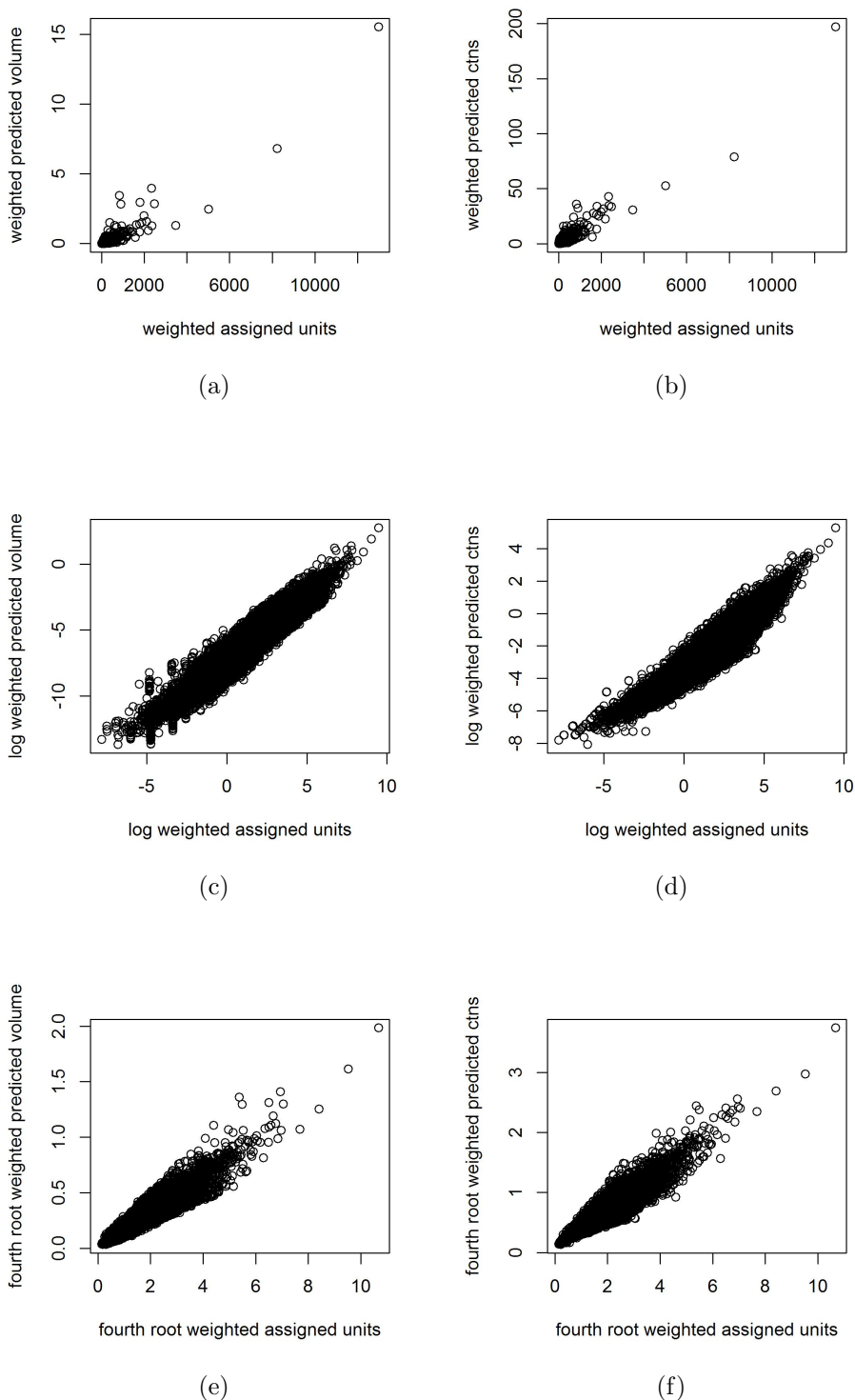


Figure 4.6: Scatter plots for the six weighted models. The top row represent the weighted model without any transformations, plots (c) and (d) are the weighted models with a natural logarithm and plots (e) and (f) is the weighted model with a fourth root transformation.

show some possible outlier observations at the bottom tail, while Figure 4.6(d), predicting cartons, have a slight curve. A straight line regression equation will not fit the plot in Figure 4.6(d) well and

Model	Carton model			Volume model		
	χ^2	DF	P-value	χ^2	DF	P-value
Weighted	29721.38	1	0	27464.55	1	0
Log weighted	405.67	1	0	519.66	1	0
Quadratic weighted	7281.48	1	0	6349.43	1	0

Table 4.7: Heteroscedasticity results from weighted models.

a non-linear model will be needed. The Log weighted models will be used in following chapters.

4.5 Conclusion

The analysis from all the scatter plots and heteroscedasticity test, suggest heteroscedasticity did improve with the transformations compared to the SLR and MLR models in Chapters 2 and 3. It was also found that the transformations is not linear, but the models where both independent and dependent variables are transformed provide relationships between variables that resembles the closest linear relationship.

This chapter indicated that the Log weighted models, where both weighted variables are transformed by a logarithm need to be used in order to alleviate heteroscedasticity. The final transformed variables will be $\ln(y_2x_2)$ for predicting cartons and will be regressed against $\ln(x_1x_2)$. The transformed variable $\ln(y_1x_2)$ will be regressed against $\ln(x_1x_2)$ for volume prediction. A straight regression line will not fit the plot in Figure 4.6(d) well and thus polynomial regression is discussed in the next chapter with regression analysis and conclusions.

Chapter 5

Polynomial regression

The previous chapter provide different transformations in order to find a transformation where the dependent and independent variables have some linear relationship and where heteroscedasticity is alleviated. The best possible transformations were found by the weighted transformed variables, where the predicted volume, predicted cartons and assigned units are multiplied by the assigned volume and where the resulting transformation is transformed further by a logarithm. These transformations are displayed in Figures 4.6(c) and 4.6(d). The model that predicts cartons, (with scatter plot in Figure 4.6(d)), is not perfectly linear. A normal linear regression model would not be sufficient and thus the introduction of polynomial regression.

5.1 Literature review

Polynomial regression can be treated in a similar fashion as multiple linear regression. Polynomial regression, models the relationship between the dependent variable and independent variable with an n^{th} degree polynomial equation [17]. If $n = 1$ then the SLR model holds, if $n > 1$ then the relationship between dependent and independent variables becomes non-linear. OLS can still be used in polynomial regression even if $n > 1$. The standard equation of a polynomial regression model is

$$\hat{y}_j = b_0 + b_1x_j + b_2x_j^2 + \dots + b_nx_j^n. \quad (5.1)$$

There are several problems with using polynomial regression [17],

1. The order of the model must be kept as low as possible. As a general rule the order should not be greater than 2, unless reasons for a higher order is justifiable [17]. A low order model is more preferable than a high order model, as a higher order model will tend to improve the fit. If an order of $n - 1$ is applied to a dataset of n observations then a 100% fit will possibly be achieved. Such a model would unlikely be a good predictor.
2. A strategy must be chosen on determining the appropriate order. Two possible strategies are the forward and backward eliminations. The forward elimination adds an order until the highest order in the model becomes insignificant. Backward elimination start with a high order polynomial and drop the highest order until the t -statistic of highest order in the model are significant. In this thesis the forward elimination technique will be used. At each step a higher order variable model will be implemented starting at order 1. If the higher order is not significantly different from the previous step's lower order model, then

the previous lower order model will be used for prediction. Due to the large data sets (37 000 observations) that is used to do the regression, the backward elimination will not be implemented as high orders will almost always be statistically significant and indicate that higher orders should not be dropped.

3. Polynomial models can in both interpolation and extrapolation change directions inappropriately or unexpectedly.
4. When the order increases the matrix $X^T X$ becomes ill-conditioned, where X is the independent variables in matrix form. The inversion calculations on the matrix can become inaccurate, the parameter estimates might become unreliable and correlation might exist between variables. One attempt to fix this is through centering the variables by subtracting the mean from the independent variables as $(x - \bar{x})$. A centered polynomial regression model is given by

$$\hat{y}_j = b_0 + b_1(x - \bar{x}_j) + b_2(x_j - \bar{x}_j)^2 + \dots + b_n(x_j - \bar{x}_j)^n. \quad (5.2)$$

Centering is not always necessary. According to Pardoe [39] if the highest order is significant then all lower order variables can be kept in model without centering.

5. If the x values (independent variable) are in a limited range then multicollinearity can exist between the order variables.

Other types of polynomial regression models also exist, like piecewise polynomial, orthogonal polynomials and polynomial models with multiple variables [17]. Piecewise polynomial uses splines to model the regression model, where at different points, different model characteristics will be used. Orthogonal polynomials fixes the multicollinearity problem, polynomial regression causes with small independent variables. Polynomial regression can also include multiple variables for example,

$$\hat{y}_j = b_0 + b_1x_{1j} + b_2x_{2j} + b_3x_{1j}^2 + b_4x_{2j}^2 + b_5x_{1j}x_{2j}. \quad (5.3)$$

Sohn *et. al* [45] uses a second order polynomial regression model to predict the number of defective electronic chips in a batch of produced chips, by using the variables that is used to determine whether a chip is a defect as input or independent variables. Loker *et al.* [35] uses piecewise polynomial regression models to determine the amount of milk, protein and fat a pregnant cow produces over time for four types of cow breeds in Canada. Antanasijevic *et al.* [15] predicts green house gas intensity in 26 European countries by using polynomial regression models of order 2 and order 3. These models were compared to an MLR and a neural network model. Pulido-Calvo *et al.* [41] need to predict the flow discharge of water to a pond. The water flow discharge formula looks as, $Q = C_d w \sqrt{2g(h_1 - h_3)}$. The coefficient C_d gets forecasted with polynomial regression of order 2, where h_1 is the upstream water level, h_3 the downstream water level, w the surface area of the filter and b the width of the channel. The results of the order 2 model was compared to an MLR model and a generalised linear model. In this thesis the standard polynomial regression model is used with centering. The next section describe the polynomial models that will be implemented in this thesis.

5.2 Model

The polynomial regression models that are used in this thesis are centered and has the general form as in equation (5.2). This centered polynomial form is implemented on the Log weighted transformed model discussed in the Section 4.4.

Instead of using the current natural logarithm which has base 2.718, the transformation will be changed to a logarithm with base 1000. This decrease the range of the dependent and independent variables' values. The logarithms from here forward will indicate a logarithm with base 1000. The two polynomial regression models then become

$$\log(x_{2j}y_{2j}) = b_0 + b_1(\log(x_{2j}x_{1j}) - m) + b_2(\log(x_{2j}x_{1j}) - m)^2 + \dots + b_n(\log(x_{2j}x_{1j}) - m)^n \quad (5.4)$$

and

$$\log(x_{2j}y_{1j}) = b_0 + b_1(\log(x_{2j}x_{1j}) - m) + b_2(\log(x_{2j}x_{1j}) - m)^2 + \dots + b_n(\log(x_{2j}x_{1j}) - m)^n, \quad (5.5)$$

where b_i is the coefficients for each centered independent variable of the form $(\log(x_2x_1) - m)^n$, m is the average of the observed values $\log(x_2x_1)$. The variables x_2y_2 , x_2y_1 and x_2x_1 represent the assigned volume times predicted cartons, assigned volume times predicted volume and assigned volume times assigned units respectively. The forward elimination approach will be used to determine the best model for prediction on both equations (5.4) and (5.5).

5.3 Results

This section provide results on the polynomial models discussed in the previous section. Equation (5.4) will be referred to as Polynomial regression model 1 (PRM1) where cartons are predicted and equation (5.5) will be referred to as Polynomial regression model 2 (PRM2) where volume are predicted. The first subsection provide results on PRM1 and the second on PRM2. In each subsection, the coefficient statistics, regression statistics, ANOVA tables and BP tests will be provided along with the regression analysis plots for different polynomial orders.

5.3.1 PRM1

Results from polynomial order 1 to 3 are shown and discussed. Table 5.1 contains the coefficients, ANOVA table and regression statistics for PRM1 with order 1, which is also an SLR model. The p -values for the estimates are 0, which indicates that the estimates are significant to the model. The ANOVA table suggest that the model itself is significant with a p -value of 0. The R^2 and adjusted R^2 are also high at 86% as shown in Table 5.1 and the RSE is small at 0.07, which indicate that the fit is very good.

By analysing Table 5.1, the regression equation fit the data well and the model could be used for prediction, but Figure 5.1 shows otherwise. Figure 5.1 displays the regression analysis plots that includes the histogram of the residuals, Q-Q plot of the residuals, fitted plot and residual plot. According to the histogram in Figure 5.1(a) the residuals are normally distributed with fat tails and the residual plot in Figure 5.1(d) shows that variance of the residuals does not change as much compared to previous models, such shown in Figure 2.6(b). The residuals are all within the range of -4 to 5. According to the residual plot there is not a lot of influential observations with very large or very small residuals, but the residual plot does show a triangular shape. This is due to the nonlinear relationship between the log weighted units and log weighted carton variables shown in Figure 5.1(c).

According to the BP test shown in Table 5.1(d) the heteroscedasticity did improve with $\chi^2 = 405.67$, compared to the SLR model in Chapter 2 with $\chi^2 = 3900.69$ for Model 3 and 3074.56 for Model 4 shown in Table 2.6. The p -value for the BP test is 0 but this low value is also influenced

Variable	Estimate	Std. Error	<i>t</i> -value	<i>P</i> -value
Intercept	-0.377	0.00	-1213.07	0
<i>x</i>	0.747	0.00	575.64	0

(a) Coefficients

Source of variation	Df	Sum Sq	Mean Sq	<i>F</i> -value	<i>P</i> -value
Regression	1	1768.01	1768.01	331363.1	0
Error	55184	294.43	0.01		

(b) ANOVA Table

Statistic	Value		Value
R^2	0.86	χ^2	405.67
Adj. R^2	0.86	Degrees of freedom	1
RSE	0.07	<i>p</i> -value	0

(c) Regression statistics

(d) BP test results

Table 5.1: Coefficients, anova table, regression statistics and BP test results for PRM1 of order 1.

by the amount of observations (37 734 observations) as the test statistic for a BP test is equal to nR^2 where R^2 is the coefficient of determination from the auxillary regression. Thus to know whether the heteroscedasticity are acceptable, an improved BP test are required and the residual plots are relied upon. But due to the residual plot having a triangular shape and the regression line does not fit the data very good in Figure 5.1(c), the results with order 2 for PRM1 need to be analysed.

Figure 5.2 provide the regression analysis plots for PRM1 with order 2. The residuals are normally distributed as seen in the histogram, Figure 5.2(a) and Q-Q plot, Figure 5.2(b). The fitted line also fit the data well as seen in Figure 5.2(c). The scatter plot in Figure 5.2(d) displays the relationship between the studensidised residuals and the independent variable which shows a funnel shape to the right. This is caused by the nonlinear relationship between the dependent and independent variables.

The residuals are still within -5 and 5 and according to the residual plot and histogram most of the residuals are centered around a mean of 0 with the minority of points close to -5 and 5 which contribute to the funnel shape. According to the BP test in Table 5.2 this order 2 model has a larger χ^2 value than the order 1 model, which indicate heteroscedasticity worsen but is still an improvement compare to the SLR models and is supported by the residual plot and the fit of the regression line in Figure 5.2(c).

The *F*-statistic prove to be significant with a *p*-value of 0 while the R^2 and Adjusted R^2 are high at 87% and have a small RSE of 0.07. This indicates the fit is very good and it is supported by the fitted plot in Figure 5.2(c). The estimates for the model also prove to be significant to the model with *p*-values of 0 as shown in Table 5.2.

In terms of fit and the regression plots in Figure 5.2, this order 2 model is a significant improvement on the order 1 model and is a possible suggested model for prediction. Figure E.1 shows

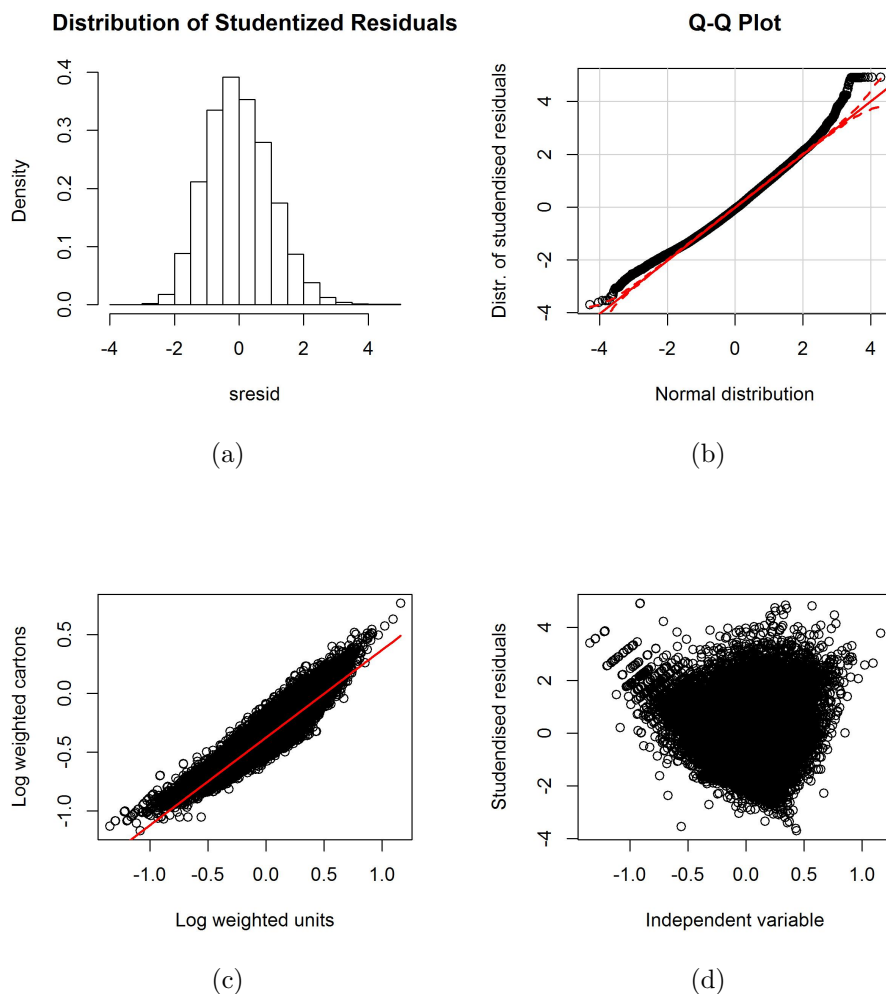


Figure 5.1: Regression analysis graphs for PRM1 of order 1 (SLR model)

the regression analysis plots for PRM1 with order 3. The residuals are normally distributed as indicated by the histogram of residuals and Q-Q plot. The model fit the data very good as seen in the fitted plot of Figure E.1(c) and the adding of the third variable does decrease the significance of the lower variables as seen in Table E.1. The F -statistic in Table E.1 suggest that the order 3 model is significant with a p -value of 0, but the R^2 , Adjusted R^2 and RSE in Table E.1 did not improve from the order 2 model. Because PRM1 order 3 does not improve or differ significantly from the order 2 model, the order 2 will be used for PRM1.

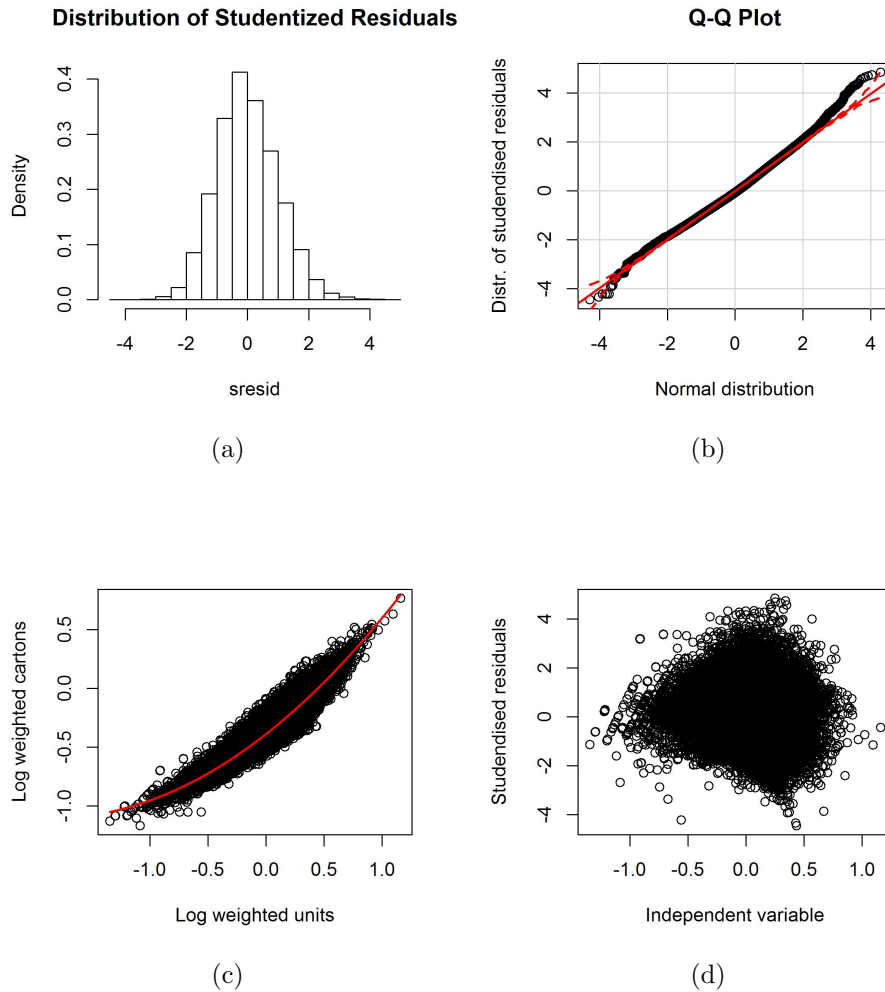


Figure 5.2: Regression analysis graphs for PRM1 of order 2

5.3.2 PRM2

Table 5.3 shows the coefficient statistics, ANOVA table, regression statistics and BP test results for PRM2 of order 1. The p -value of the coefficients in the Table 5.3(a) is 0 which indicate the coefficient is significant to the model. The ANOVA table shows that the model is significant with p -value equal of 0, while the R^2 and Adjusted R^2 values are high with a small RSE.

The data in Table 5.3 suggest that the regression equation fit the data well and the plots in Figure 5.3 confirm this. The histogram of residuals is slightly skewed and the Q-Q plot shows some outlier observations on the higher tail. The fit plot suggest there are a linear relationship between the independent and dependent variables and the order 1 model which is an SLR model fit the data well. The p -value of the BP test is 0 but the χ^2 value of 519.67 is an improvement from the SLR models in Chapter 2. The residual plot also does not show any significant change in the variance of the residuals but also displays some of the outlier observations.

The significance of the first coefficient of variable x in PRM2 of order 2 decreased, with a smaller t -value from the order 1 model as indicated in Table E.2(a). All three coefficients are significant to the model, but the third coefficient has a small t -value and the coefficient value itself is very

Variable	Estimate	Std. Error	t -value	P -value
Intercept	-0.39	0.00	-1138.12	0
x	0.77	0.00	592.71	0
x^2	0.21	0.00	71.88	0

(a) Coefficients

Source of variation	Df	Sum Sq	Mean Sq	F value	P -value
Regression	2	1793.22	896.61	183776.5	0
Error	55183	269.23	0.01		

(b) ANOVA Table

Statistic	Value		Value
R^2	0.87	χ^2	824.41
Adj R^2	0.87	Degrees of freedom	2
RSE	0.07	p -value	0

(c) Regression statistics

(d) BP test results

Table 5.2: Coefficients, anova table, regression statistics and BP test results for PRM1 of order 2.

Variable	Estimate	Std. Error	t -value	P -value
Intercept	-0.85	0.00	-2801.62	0
x	0.97	0.00	761.97	0

(a) Coefficients

Source of variation	Df	Sum Sq	Mean Sq	F value	P -value
Regression	1	2986.7	2986.7	580593.2	0
Error	55184	283.87	0.01		

(b) ANOVA table

Statistic	Value		Value
R^2	0.91	χ^2	519.67
Adj R^2	0.91	Degrees of freedom	1
RSE	0.07	p -value	0

(c) Regression statistics

(d) BP test results

Table 5.3: Coefficients, anova table, regression statistics and BP test results for PRM2 of order 1.

small at 0.04. This suggest that the x^2 variable does not have a major impact on the value of the dependent variable. The ANOVA table in Table E.2(b) has a large F -statistic with a p -value of 0 and the R^2 values are high with a small RSE. The plots in Figure E.2 also shows that the

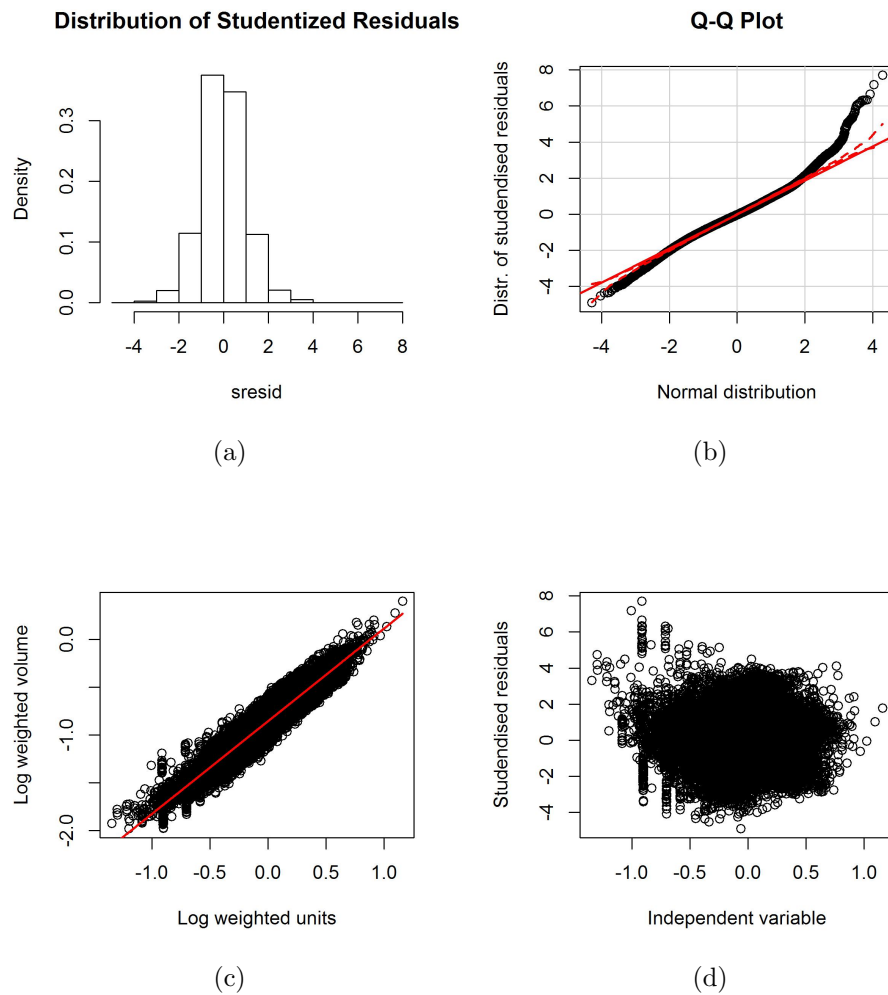


Figure 5.3: Regression analysis graphs for PRM2 of order 1

regression fit the data well, with the residuals distributed normally with mean 0. The results from the order 2 model is not significantly differently from the order 1 model and the coefficient value of x^2 does not add much to the model regarding the dependent variable. Thus the order 1 model will be used for PRM2.

5.4 Conclusion

The polynomial model PRM1 showed with a fit plot that the relationship between the transformed variables is not linear, and thus the introduction of polynomial regression where higher order variables can be used. It was then shown that an order of 2 is sufficient for PRM1. Model PRM2 does have a linear relationship between the transformed variables and it was shown for PRM2 an order 1 polynomial are sufficient. It was shown that the residuals under these polynomial models are normally distributed with PRM2 having some outlier observations. The BP tests has smaller χ^2 values than the SLR models in Chapter 2 indicating heteroscedasticity improved compared to the SLR models. But from the fit plots of these suggested models the regression lines fit the data well, and even though the residual plots, particularly for PRM1, does show some

sign of heteroscedasticity, it is all caused by some outlier observations. Due to the significant improvements in terms of the BLUE assumptions compared to the SLR models in Chapter 2, these polynomial models will be used for carton and volume forecasting.

Chapter 6

Carton and volume forecasting

In this chapter the polynomial transformed models described in the previous chapter are applied to the test data set. This data contains the units and volume per store on picking lines that has been set up for the Kuilsrivier PEP DC during the period of 16 July 2014 until 31 July 2014. The carton forecasting model use a polynomial regression model of order 2 while the volume prediction model uses an SLR model where the dependent variables (predicted volume and predicted cartons) and independent variable (assigned units) are weighted by the assigned volume and further transformed with a logarithm with base 1000. The carton and volume forecasting for the test set data are obtained and will be compared to the actual values. Five measures of accuracy are used for each model. These measures are, root mean square error (RMSE) where

$$\text{RMSE} = \sqrt{\frac{\sum_{j=1}^n (a_j - y_j)^2}{n}}, \quad (6.1)$$

normalised root mean square error (NRMSE),

$$\text{NRMSE} = \frac{\text{RMSE}}{\bar{a}}, \quad (6.2)$$

mean absolute error (MAE),

$$\text{MAE} = \frac{\sum_{j=1}^n |a_j - y_j|}{n}, \quad (6.3)$$

mean absolute percentage error (MAPE),

$$\text{MAPE} = \sum_{j=1}^n \frac{|a_j - y_j|}{na_j} 100, \quad (6.4)$$

and percentage bias (PBIAS),

$$\text{PBIAS} = \frac{\sum_{j=1}^n (a_j - y_j)}{\sum_{j=1}^n a_j} 100. \quad (6.5)$$

The a_j values where $j = 1, \dots, n$, is the actual values, y_j is the forecasted values and n is the number of observations. The RMSE is a unit measure that give one value that measure the overall accuracy of the forecasted values against the actuals. This measure is good for comparing models with the same unit measures against each other for forecast accuracy. The NRMSE is a percentage

measure also called the coefficient of variation of the RMSE. The NRMSE are non-dimensional and can be used to compare models with different measures. The SLR and MLR models will also be applied to the test set data to determine a forecast. The NRMSE can then be used to compare the overall forecast accuracy between the models. The MAE is an overall unit measure of forecast accuracy, which represents the average difference between the actual and predicted values. This will determine on average with how much does a model over or under forecast and can only be used to compare models with the same measures. The MAPE represent the average percentage difference between the actual and predicted values and can be used to compare the average error of models with different measures.

The RMSE and NRMSE give more weight to larger errors due to the power of two and when comparing two models, the RMSE or NRMSE could indicate that one model is less accurate than another due to a few outlier errors compared to the average error. The RMSE, NRMSE, MAE and MAPE will be used together to determine whether one model predicts better than the other. The PBIAS measures the degree to which a model over or under predict [15]. The RMSE, MAE and MAPE must be as small as possible, where a MAPE of less than 10% are considered as highly accurate. A perfect value for PBIAS is 0, where a negative PBIAS indicate the model over predict while a positive value indicates under prediction. This chapter are divided into two sections, one for forecasting cartons and one for forecasting volume. The first section provide the results on the carton forecasting.

6.1 Carton forecasting

The carton forecasting makes use of the polynomial model in equation (5.4) with 2 degrees. Table 5.2 provide the coefficients for PRM1 with 2 degrees for the training dataset. Given these results the model for carton forecasting is given as

$$\log(y_2x_2) = -0.39 + 0.77(\log(x_1x_2) - 0.213) + 0.21(\log(x_1x_2) - 0.213)^2. \quad (6.6)$$

The variables y_2x_2 represents the forecasted cartons times assigned volume and x_1x_2 the assigned units times assigned volume. The values 0.77 and 0.21 are the coefficients for the two terms from Table 5.2 with -0.39 the intercept term and 0.213 are the average of the values, $\log(x_1x_2)$ in the data that was used to calculate the coefficients in Table 5.2 that centers the polynomial model. The forecasted cartons (y_2) can be calculated as

$$y_2 = \frac{1000(-0.39+0.77(\log(x_1x_2)-0.213)+0.21(\log(x_1x_2)-0.213)^2)}{x_2}. \quad (6.7)$$

The assigned volume, assigned units and forecasted carton values from the test set picking line data were used in the equation (6.7). The resulting actual versus forecasted values per day that were obtained are shown in Table 6.1 and displayed in the bar chart in Figure 6.1. The values in Table 6.1 and Figure 6.1 are aggregated values from the store predictions and actuals. The total actual cartons over the period 16 July 2014 until 25 July 2014 that were sent to the stores were 21315 cartons while the model predicted 23941 cartons, thus 2626 more.

Table 6.2 provide the forecast accuracy measures for the polynomial models which predicted cartons and volume, and the SLR model with and without influential observations that was detected by the Cook's distances in, Section 2.5.2. The table shows the model in the first column, the

Date	Actual cartons	Predicted cartons	Error	Percentage error
16-Jul-2014	2911	3129	218	8%
17-Jul-2014	3311	3236	-75	-2%
18-Jul-2014	1674	2144	470	28%
21-Jul-2014	2588	3698	1110	43%
22-Jul-2014	2723	3792	1069	39%
23-Jul-2014	3467	3370	-97	-3%
24-Jul-2014	2641	2532	-109	-4%
25-Jul-2014	2000	2039	39	2%
Total	21315	23941	2626	12%

Table 6.1: Actual versus predicted cartons for model PRM1 of order 2 per day from 16 July 2014 until 25 July 2014.

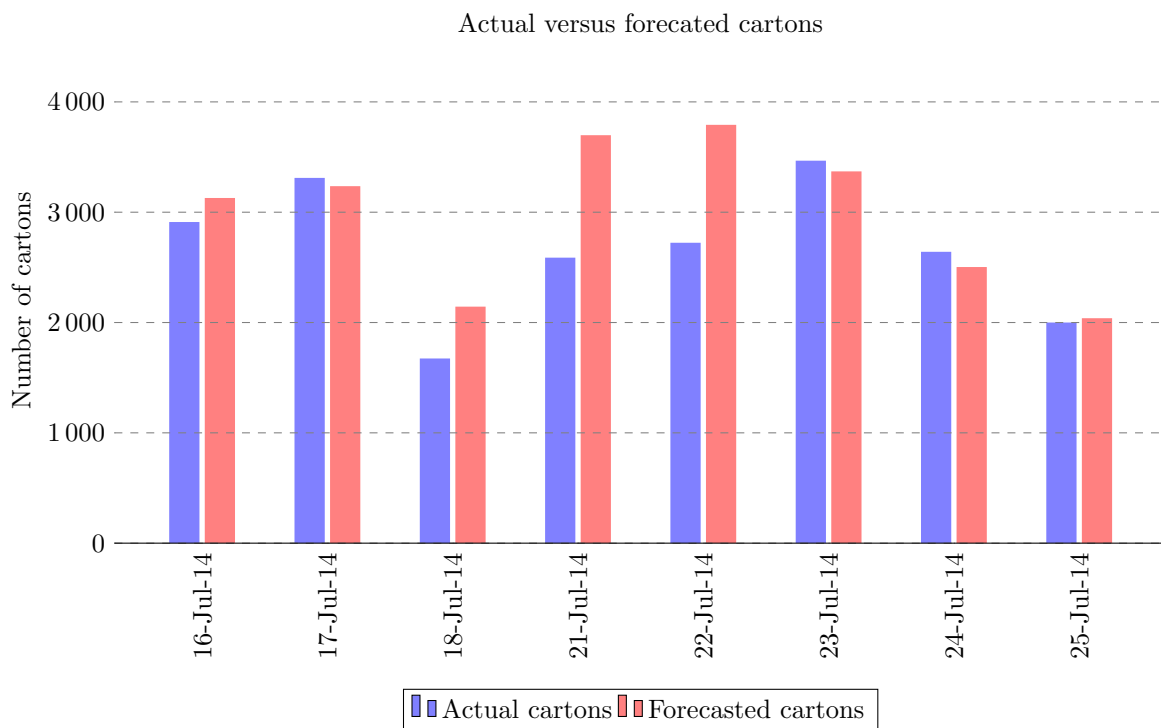


Figure 6.1: Bar chart of the actual versus predicted cartons for model PRM1 of order 2 per day from 16 July 2014 until 25 July 2014.

dependent variable in the second column and independent variable in the third column, like the assigned volume or assigned units. The next 5 columns shows the accuracy measures, RMSE, NRMSE, MAE, MAPE and PBIAS. The last column shows the total forecast for the dependent variable the model provide.

In Table 6.2 the models with dependent variable as cartons has similar RMSE and NRMSEs between each other. The polynomial model for cartons has the same RMSE (1.54) and NRMSE (0.77) values as for the SLR model without influential observations and volume as independent variable. According to the MAE and MAPE the average error that is made on the models with cartons as dependent variable and volume as independent variable is greater than those with units as independent variable and thus predicting cartons with units as independent variable will give

smaller errors. The polynomial model for predicting cartons also has a smaller MAE (0.83) and MAPE (46.90%) than the SLR models that has volume as independent variable, but are larger than the SLR models with units as independent variable.

The SLR models tend to under forecast cartons except for the model where volume are an independent variable on the SLR model without influential observations. The PBIAS on that model is -11.92%. The polynomial model tend to over predict with a PBIAS of -12.21%. The total actual cartons as displayed in Table 6.1 are 21315 cartons and the polynomial model and SLR models with volume as independent variable results into a prediction close to the actual with a difference of about 2000 cartons. Even though the MAE and MAPE for the models with units as independent variable are smaller than the models with volume as independent variable and the polynomial model, the total prediction value is smaller than the other models.

The SLR model with influential observations and units as independent variable has a forecast that are about 5000 cartons less than the actual value while the SLR model without influential observations and units as independent variable has a forecast that are about 2000 cartons less than the actual value. Thus on cartons, more stores are over predicted with the SLR models with volume as independent variable and polynomial model, compared with the SLR models with units as independent variable.

Model	Dependent variable	Independent variable	RMSE	NRMSE	MAE	MAPE	PBIAS	Total forecast
Polynomial model	Cartons	Weighted units	1.54	0.77	0.83	46.90%	-12.32%	23941
	Volume	Weighted units	0.08	1.08	0.04	70.58%	-31.78%	1097
SLR model	Cartons	Volume	1.29	0.65	0.88	48.48%	3.70%	20526
	Cartons	Units	1.12	0.56	0.73	38.91%	25.39%	15903
	Volume	Volume	0.03	0.43	0	4.33%	4.15%	798
	Volume	Units	0.08	0.98	0.04	54.07%	-15.10%	958
SLR model without influential observations	Cartons	Volume	1.54	0.77	1	54.37%	-11.92%	23855
	Cartons	Units	1.18	0.59	0.69	37.48%	11.26%	18914
	Volume	Volume	0.03	0.39	0.01	5.56%	-1.97%	849
	Volume	Units	0.08	1.02	0.04	60.72%	-24.66%	1037

Table 6.2: Accuracy performance metrics of the actual versus forecasted cartons for the PRM1 polynomial model of order 2 per day from 16 July 2014 until 25 July 2014.

The SLR models as discussed in Chapter 2 does not adhere to the BLUE assumptions, but regardless of this, the SLR models for carton prediction in particular the models with volume as independent variable, provide similar accuracy results as the polynomial model PRM1.

6.2 Volume forecasting

Volume forecasting uses the PRM2 model described in Chapter 5, sections 5.2 and 5.3. The model is a polynomial regression model as in Equation 5.5 with 1 degree, which is an SLR model. Given the coefficients in Table 5.3 the regression equation to be used for volume forecasting is

$$\log(y_1x_2) = -0.85 + 0.97(\log(x_1x_2) - 0.213). \quad (6.8)$$

The variables y_1x_2 represents the forecasted volume times assigned volume, while x_1x_2 represents the assigned units times assigned volume. The values 0.97 and -0.85 are the coefficients from Table 5.3 and the value 0.213 are the average of the values of $\log(x_1x_2)$ from each observation in the data used to determine the coefficients that centers the polynomial model. The forecasted volume, y_1 can be calculated by the formula

$$y_1 = \frac{1000^{(-0.85+0.97(\log(x_1x_2)-0.213))}}{x_2}. \quad (6.9)$$

The picking line data for 16 July 2014 until 25 July 2014 was used in Equation 6.9 and the resulting forecasted volume are compared to the actual volume sent to the stores. The actual versus the forecasted volume are shown in Table 6.3 and corresponding bar chart in Figure 6.2. The total actual volume are 832.76 m³ and the total forecasted volume are 1079.04 m³. Analysing Table 6.3 and Figure 6.2 the model over predict the volume on all the days except for one day, 24 July 2014.

Date	Actual volume	Fcst volume	Error	Percentage error
16-Jul-2014	107.63	132.28	24.65	23%
17-Jul-2014	109.23	136.8	27.57	25%
18-Jul-2014	75.52	102.5	26.98	36%
21-Jul-2014	114.58	180.02	65.44	57%
22-Jul-2014	114.51	183.34	68.83	60%
23-Jul-2014	130.08	151.03	20.95	16%
24-Jul-2014	111.85	104.94	-6.91	-6%
25-Jul-2014	69.36	88.12	18.76	27%
Total	832.76	1079.04	246.28	30%

Table 6.3: Actual versus forecasted volume measured in cubic meters (m³) for the M2 polynomial model of order 1 (SLR) per day from 16 July 2014 until 25 July 2014.

By studying the models with volume as dependent variable in Table 6.2, the RMSE and NRMSE for the polynomial model are similar to both SLR models where the independent variable is units. The polynomial model have an RMSE and NRMSE of 0.08 and 1.08, the SLR models with units as independent variable has an RMSE and NRMSE of 0.08 and 0.98 and for the model without influential observations an RMSE of 0.08 and NRMSE of 1.02. The MAE is equal to 0.04 for all the models with unit as independent variable and volume as dependent variable, which shows the average error that is made is very small. The models with volume as independent variable have a smaller MAPE than the models with units as independent variable which indicate smaller errors. The PBIAS suggest that for the volume predictions, the models tend to over predict with negative PBIAS values except for the SLR model with influential observations and volume as independent variable.

Table F.1 provide the accuracy measures for all the MLR models. The RMSE and NRMSE values for the different MLR models does not differ significantly from each other except for the BU clothing indicator model and BU combined dummy variable model where units are the independent variable. The RMSE and NRMSE for all the models including the polynomial and SLR models tend to be about 1 for carton forecasting, but for the BU clothing indicator model the RMSEs is 5.13, NRMSE is 2.56 and for the BU combined dummy variable model the RMSE is 3.78 and NRMSE 1.89. The RMSE and NRMSE for volume forecasting is in the range of 0.03 to 0.08 and

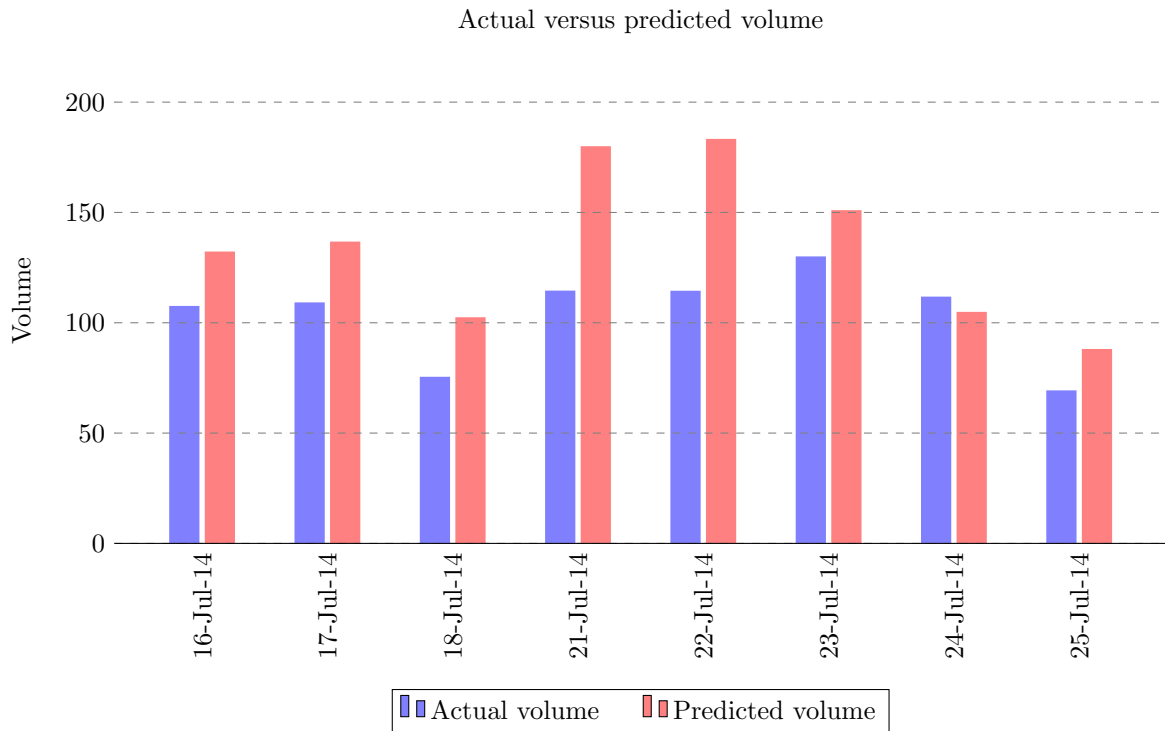


Figure 6.2: Bar chart of the actual versus forecasted volume measured in cubic meters (m^3) for the M2 polynomial model of order 1 (SLR) per day from 16 July 2014 until 25 July 2014.

0.57 to 0.96, but for the BU clothing indicator model and BU combined dummy variable model the RMSE are 0.46 and 0.43 while the NRMSE are 5.88 and 5.49 for the two models respectively. The MAE and MAPE for these two models are also higher and shows that the BU model with clothing indicator and combined dummy variables should not be considered as a forecasting model.

6.3 Conclusion

The polynomial regression models for carton and volume produced total predictions that are close to the actuals for the days 16 July 2014 until 25 July 2014. In Table 6.1 that shows the actual cartons against forecasted cartons, the errors between actuals and predictions are relatively small with an average percentage error of 12%. Two of the days has a forecasted value more than a 1000 cartons that is large compared to the other days.

The volume predictions are consistently more than the actual values with an average percentage error of 30%, displayed in Table 6.3. The PBIAS also indicated that the polynomial model tend to over forecast both cartons and volume. The small MAE (0.83 for carton prediction and 0.04 for volume prediction) value shows that the average error produced by the model are small which indicate although the models over predict the errors for each store per day are small.

The errors produced by the various models tend to be small with the MAE value of about 1. The relationship between the RMSE, NRMSE, MAE and MAPE for all the models (polynomial, SLR, MLR), are consistent with an increase in RMSE and NRMSE leads to an increase in MAE and MAPE as well. This shows that there is not a substantial number of observations with very large errors. The RMSE and NRMSE also does not differ significantly between the models, but the

BU MLR models with added dummy variables give large error performance metrics compared to the other models. Given these results and results from previous chapters the polynomial models could be considered for predicting cartons and volume.

Chapter 7

Case study

In this chapter the polynomial models for carton and volume prediction are implemented on a new set of data for picking lines that was set up during 2 November 2015 until 13 November 2015. The predictions will be determined by using the coefficients found from the training set in Chapter 5. This predictions from these coefficients will be compared to the predicted cartons and volumes that was sent from the Kuilsrivier DC to the hub and the actual carton and volume values.

7.1 Carton and volume forecast against actuals

The Kuilsrivier DC generated a total carton and volume prediction for each day from 7 November 2015 until 13 November 2015 that was sent to the hub. This prediction is based on the number of cartons a picker can pick per hour, the number of hours a picker works, the number of pickers and the average volume per carton. The pickers can on average fill 22 cartons per hour and each carton on average is 0.05 m^3 . If a picker has 8 hours to pick, then one picker will be able to fill 176 cartons and pick 8.8 m^3 of stock. Multiply this by the number of pickers in a day give the carton and volume predictions per day that the DC send to the hub. The regression method discussed in this thesis is different to the current method in the DC as the regression method does not take the pickers into account.

The coefficients calculated in Chapter 5 are used on the picking line data from 2 November 2015 until 13 November 2015. The predicted results are compared to actual carton and volume for the period and also the prediction from the current DC method. Table 7.1 shows the carton prediction from the model PRM1 with polynomial order 2 with the coefficients from Chapter 5 in column two, the prediction from the DC in the third column, the actual cartons in column four and the difference between the predicted and the actual cartons is in the last two columns. The Prediction diff. column displays the difference between the prediction and the actual cartons and the DC diff. column shows the difference between the prediction from the DC and the actual cartons. The predictions from the model under predict by 8481 cartons, while the DC predictions are over by a 1092 cartons. This shows that the DC forecast for cartons is more accurate than the forecast from the model PRM1.

Table 7.2 shows the volume predictions in the second column from the model PRM2 with polynomial order 1 with the coefficients in Chapter 5, the DC volume prediction, actual volume and the predicted versus actual differences as in Table 7.1. The predictions with model PRM2 under predict by 87 m^3 , while the DC predictions are over by 592 m^3 . Thus the DC predictions for

Day	Predicted Cartons	DC carton prediction	Actual cartons	Prediction diff.	DC diff.
02-Nov-2015	2005	1650	1654	352	-4
03-Nov-2015	3995	4460	4460	-465	0
04-Nov-2015	4181	3960	3880	302	80
05-Nov-2015	3183	3960	3960	-777	0
06-Nov-2015	3603	5060	4862	-1260	198
09-Nov-2015	4873	6600	6410	-1537	190
10-Nov-2015	5581	6600	5945	-364	655
11-Nov-2015	6039	7100	7089	-1050	11
12-Nov-2015	4793	7100	7114	-2322	-14
13-Nov-2015	2990	4327	4350	-1361	-23
Total	41244	50817	49725	-8481	1092

Table 7.1: Forecasted cartons compared to the actual carton and volume for the period 2 November 2015 until 13 November 2015

volume are less accurate over this period than the regression model. The regression model under predict more in the second week from 9 November 2015 until 13 November 2015. The actual volume increased in the second week compared to the first week and the same for the cartons as shown in Table 7.1. The number of pickers was also increased in the second week from 24 pickers to 40 in order to accommodate for the increase in volume to pick. During the second week some Christmas stock and back to school stock, that is normally high volume, was sent to stores. This shows that for higher volumes, the regression model tend to under predict. The coefficients were calculated during non-peak sales weeks and the under prediction of volume and cartons indicate that the coefficients should be reviewed during the different periods of the year to accommodate for the change in volume. The higher volume lead to more units per carton and/or more cartons that are sent to stores and in turn affect the ratio between the independent and dependent variables during peak sales periods.

Day	Predicted volume	DC volume forecast	Actual volume	Prediction diff.	DC diff.
02-Nov-2015	77	83	67	10	16
03-Nov-2015	175	223	116	59	107
04-Nov-2015	190	198	120	70	78
05-Nov-2015	145	198	112	33	86
06-Nov-2015	164	253	164	0	89
09-Nov-2015	228	330	275	-47	55
10-Nov-2015	259	330	324	-65	6
11-Nov-2015	275	355	343	-68	12
12-Nov-2015	223	355	296	-73	59
13-Nov-2015	126	216	132	-6	84
Total	1862	2541	1949	-87	592

Table 7.2: Forecasted volume compared to the actual carton and volume for the period 2 November 2015 until 13 November 2015

7.2 Conclusion

The coefficient from model PRM1 tend to under predict cartons and the DC over predict the cartons and it is shown that the DC also predict the cartons more accurately. The model PRM2 predict the volume more accurately than the DC but the model under predict more in the second week when the pick volume increased. This under prediction give the impression that the coefficients should be reviewed during different periods of the year.

Chapter 8

Conclusion

The error terms of the regression models for the picking line input (assigned volume and assigned units) and output (predicted cartons and predicted volume) data from PEP's Kuilsrivier DC is not normally distributed and the variance of the data is not the same at every data point. The larger the input data, the larger the variance of the output data and the more unpredictable the output data becomes. Given this information, the SLR model did not work very well and the BLUE assumptions of regression modelling did not hold, because heteroscedasticity were present. Various MLR models was also introduced but heteroscedasticity was still present and different transformations on the SLR model was considered.

Of the transformed actions considered, the logarithmic weighted transformation displays the best results. In this model the input and output data have a positive linear relationship and the heteroscedasticity is minimised. The logarithmic weighted transformation is applied to both carton and volume forecasting but for the carton forecasting, the relationship between the input and output data displays a curve while for the volume forecast it is a straight line relationship. An SLR model is used for volume forecasting but for the carton forecasting a polynomial model with order 2 is a better option, because polynomial regression accommodate for nonlinear relationships.

The regression analysis on the transformed data displays an alleviation in heteroscedasticity with the polynomial regression of order 2 for carton forecasting and the SLR model for volume forecasting. According to the accuracy tests on the test set data, the accuracy of these models are also very good. With carton forecasting, the MAE or average error is 0.83 and for volume forecasting, the MAE were 0.04 which is very small compared to actual values and the average error on each picking line and store combination. About 25% of the observations, has one carton as actual value and a lot of the volume values is less than one and close to zero, where a minimum volume value were recorded at 0.06 m^3 .

The coefficients from the regression models were used in a case study on a new set of data and was compared to actual data the Kuilsrivier DC sent to the hub during 02 November 2015 until 13 November 2015. It was found that the models can be used by PEP as small forecast errors were obtained and coefficients will have to be reviewed periodically by doing regression analysis on corresponding data in a previous year in order to accommodate for changes in volume over peak sales periods.

8.1 Recommendations to PEP

It would benefit PEP to be able to predict carton and volume output days before the actual carton and volume are sent to the hub. This will help hubs to schedule trucks and routes more accurately and in turn save costs. The models described in this thesis can help PEP significantly with the forecasting and be able to provide the hubs with a detail forecast on store level where the current forecast from PEP to hubs are normally on total volume and cartons. The functionality to predict the cartons and volume by using the models described in this thesis could be implemented into PEP's warehouse management system, such that if a picking line is set up the carton and volume forecast can be generated and send to the hubs.

8.2 Objectives achieved

In this section all the objectives achieved are discussed as well as where it has been achieved.

1. In Chapter 2, SLR models were discussed and the picking line data from PEP's Kuilsrivier DC was used to predict cartons and volume. It was found that there exist heteroscedasticity in the residuals and the BLUE assumptions were not adhere to. Chapter 3 follows with a literature review on MLR models and eight MLR models were considered where BUs and categories were independent variables and a SKU count and a clothing indicator were added as dummy variables.
2. Different transformations were analysed in Chapter 4 to find a transformation where heteroscedasticity are alleviated. A literature review were provided on different types of transformations and a logarithmic weighted transformation was chosen as the best possible transformation that will eliminate heteroscedasticity.
3. Polynomial regression models are discussed in Chapter 5. A literature review are provided and regression analysis on the picking line data to predict cartons and volume were implemented on the training data set. The resulting coefficients were used in Chapter 6 and forecast accuracy where measured.
4. In Chapter 7 provide a case study. The coefficients from the results in Chapter 5 where used on picking line data from 2 November 2015 until 13 November 2015. These forecasts were compared to the Kuilsrivier DC forecast that was sent to the hub in that period and also the actual carton and volume for the period.
5. In Chapter 6 and 7 it was found that the models generate small errors and can be used for forecasting cartons and volume from a picking line. But it was also found that the coefficients need to be reviewed periodically. In Chapter 8 it is recommended that PEP use the models in this thesis and implement the models into their system.

8.3 Further studies and recommendations

The assumption is made that the picking lines are already planned. This gives PEP only a few days to generate the forecast and send it to the hub. As a further study, one could investigate methods to forecast the assigned units and volume that will be sent from the PEP central office

to the DC and use methods to schedule picking lines virtually before it is physically build. This will give PEP the ability to send the hubs a forecast earlier.

The units and volume predictions from the central office can possibly be calculated by using historical data and some smoothing forecasting algorithm like the Holt-Winter smoothing algorithm [51]. The virtual scheduling could be implemented into the warehouse management system that extract the forecast that will be used for the scheduling of picking lines. The predictions from the central office will contain many products, but each picking line can only take a limited amount of products. These products can be divided into the various picking lines by using a bin packing algorithm [32] where the objective could be to maximise the number of time products need to be picked together.

If PEP can determine in advance what units and volume will need to be scheduled, then stock can be retrieved from the large storage racks sooner, alleviate work on cranes and give them the ability to send cartons and volume predictions earlier to the hubs. This will help PEP with better visibility in future and will also help with supply chain planning during peak sales periods like December and January. During these peak sales period, it is important for the DC to optimally schedule workload between workers and get the stock out of the DC when needed. It is also important for the hubs so that they do not get overstocked with the large volume and have problems getting the stock to stores on time.

Another investigation is to use the methods in this thesis and analyse the effect it have on the scheduling of trucks and routes at the hub. The forecast that are currently sent to the hub is not broken down into store level and the hubs can only start scheduling trucks and routes once they have the break down per store. Currently only once the actual cartons and volume are received then the scheduling of trucks can commence. The models in this thesis can provide a carton and volume forecast per store before the actual cartons and volume are received by the hub and scheduling can start earlier. The polynomial regression models in this thesis can be implemented on picking lines for a certain period like a month. Once the picking lines are setup, the regression models can be implemented and the resulting carton and volume predictions can be used to schedule the trucks and routes.

Nallusamy [27] *et. al* describes the multiple vehicle routing problem where multiple vehicles need to serve a number of customers and the travelling distance of serving the customers must be minimised and a generic algorithm is also introduced. This problem resembles how the PKL hub schedule trucks on routes to serve stores with the local transport. Ropke [43] introduces an algorithm that can be applied to different variants of the vehicle routing problem. These and similar heuristics or algorithms can be used to schedule routes and trucks from the Kuilsrivier hub where the travelling distances can be minimised and also the number of trucks used. The resulting costs (truck hiring, petrol, wages) from these schedules can be compared to the costs resulting from schedules that is done with the current method at the hub.

As an extra dimension, both actual and forecasted data can be implemented into these heuristics or algorithms. That way we will be able to analyse whether the current scheduling method are sufficient and there is not any other better ways of scheduling. We will also be able to analyse whether the predicted data can lead to more cost efficient schedules. Carton and volume predictions from other retailers PKL serves and also other hubs can also be included into the data.

In Chapter 7 it is shown that the models under predict when the volume that need to be picked increased due to peak season sales. Different periods of the year probably need different coefficients for the regression models. During peak sales season more stock will be send to stores than any other time of the year and this could lead to pickers having the ability to utilise cartons more efficiently.

A study can follow that calculate different coefficients at different times of the year for instance per month or per quarter or a specific time of the year like December and January. Along with forecast accuracy tests, predictions can be found with these different coefficients and hypothesis testing can be implemented on the different prediction sets with the different coefficients. The hypothesis tests can test the null hypothesis that all the prediction sets does not differ from each other against the null hypothesis that at least one differ from the rest. The Diebold-Mariano test [16] can also be used to compare the predictions resulting from the different coefficients.

Bibliography

- [1] N.N. ABDELMALEK & N. OTSU, 1985, *Restoration of images with missing high-frequency components by minimizing the L1 norm of the solution vector*, Applied Optics, **24**(10), pp. 1415-1420
- [2] S. ABDINNOUR-HELM, *Network design in supply chain management*, International Journal of Agile Management Systems, **1**(2), pp. 99-106
- [3] M.A. ADELMAN & G.C. WATKINS, 1995, *Reserve Asset Values and the Hotelling Valuation Principle: Further Evidence*, Southern Economic Journal, **61**(3), 664-673
- [4] C. AGOSTINELLI, 2002, *Robust stepwise regression*, Journal of Applied Statistics, **29**(6), pp. 825-840
- [5] Y. AN, Y.ZHANG & B.ZENG, *The Reliable Hub-and-spoke Design Problem: Models and Algorithms*, [ONLINE], Available from <http://www.optimization-online.org>
- [6] C. ARMERO & J. FERRÁNDIZ, 2002, *Simulation in the simple linear regression model*, Teaching Statistics, **24**(1), pp. 12-16
- [7] T.W. ARNOLD, 2010, *Uninformative parameters and model selection using Akaike's information criterion*, Journal of wildlife magement, **74**(6), pp. 1175-1178
- [8] M. BARRETO & DAVID MAHARRY, 2006, *Least median squares and regression through the origin*, Computational statistics and data analysis, **50**(2006), pp. 1394-1397
- [9] J.J. BARTHOLDI & S.T. HACKMAN, 2014, *Warehouse and distribution science*, [ONLINE], Available from <http://www.warehouse-science.com>
- [10] R.L. CHAMBERS & R. DUNSTAN, 1986, *Estimating function from survey data*, Biometrika, **73**(3), pp. 597-604
- [11] S. CHATTERJEE & A.S. HADI, 2006, *Regression by example*(4 ed), Hoboken, N.J. : Wiley-Interscience
- [12] J. CHEN, 1997, *Mental chronometry with simple linear regression*, Perceptual and Motor Skills, **85**(2), 499-513
- [13] S. CHOPRA & P. MENDEIL, 2004, *Supply Chain Management*, 2nd Edition, New Jersey, Upper saddle river: Pearson Prentice Hall
- [14] R.D. COOK, 2002, *Detection of Influential Observation in Linear Regression*, Technometrics, **19**(1), pp. 65-68

- [15] D. ANTANASIJEVIC, V. POCAJT, M. RISTIC & A. PERIC-GRUJIC, 2015, *Modeling of energy consumption and related GHG (greenhouse gas) intensity and emissions in Europe using general regression neural networks*, Energy, **84**, pp. 816-824
- [16] F.X. DIEBOLD AND R. MARIANO, 1995, *Comparing Predictive Accuracy*, Journal of Business and Economic Statistics, **13**(3), pp. 253-265.
- [17] D.C. MONTGOMERY, E.A. PECK, 1992, *Introduction to linear regression analysis 2nd Edition*, Wiley
- [18] J.G. EISENHAUER, 2003, *Regression through the origin*, Teaching statistics, **25**(3), pp. 76-80
- [19] W. FRANK, G. SELENA, A. LANCE, M. JASON & H. JAMES, 2005, *USING A LOGARITHMIC REGRESSION TO IDENTIFY THE HEART-RATE THRESHOLD IN CYCLISTS.*, Journal of Strength and Conditioning Research **19**(4), pp. 838 - 841
- [20] I. M. GHANDI & S. AHMAD, 2010, *Stepwise multiple regression method to forecast fish landing*, Procedia Social and Behavioral Sciences, **8**, pp. 549-554
- [21] A. GILONI & M. PADBERG, 2002, *Alternative methods of linear regression*, Mathematical and Computer Modeling, **35**(3), pp. 361-374
- [22] D. GUJARATI, 1992, *Essentials of econometrics*, 1st Edition, New York, McGraw-Hill
- [23] S.L. HAKIMI, 1964, *Optimum loactions of switching centers and the absolute centers and medians of a graph*, Operations Research, **12**(3), pp. 450-459
- [24] Microsoft. (2010). Microsoft Excel [Computer software]. Redmond, Washington: Microsoft.
- [25] S.M. MILLER, J.R. GILES & C.P. OERTEL, 2009, *Weighted exponential regression for characterizing radionuclide concentrations in soil depth profiles*, J Radioanal Nucl Chem, **282**(2), pp. 487 - 491
- [26] B. MOLNÁR & G. LIROVSZKI, *Multi-objective routing and scheduling of order pickers in a warehouse*, International Journal of Simulation, **6**(5), pp. 23-32
- [27] R. NALLUSAMY, K. DURAISWAMY, R. DHANALAKSMI, P. PARTHIBAN, 2009, *Optimization multiple vehicl routing problems using approximation algorithms*, International Journal of Engineering Science and Technology, **1**(3), pp. 129-135
- [28] Y. KAYHAN & S. GUNAY, 2008, *A new approach to least median of squares and regression through the origin*, Communications in Statistics - Theory and Methods, **37**(5), pp. 773-781
- [29] K. MARIL, 2004, *Advanced Statistics: Linear Regression, Part 1: Simple Linear Regression*, ACAD EMERG MED, **11**(1), pp. 87-93
- [30] MASSEY UNIVERSITY, SIMPLE LINEAR REGRESSION,[ONLINE], Retrieved from <http://www.massey.ac.nz/~mbjones/Book/Chap9.pdf>
- [31] C. KIM, 1996, *Cook's distance in spline smoothing*, Statistics and Probability Letters, **31**(2), pp. 139-144
- [32] R. KORF, 2002, *A new algorithm for optimal bin packing*, AAI, pp. 731-736

- [33] H.C. LAU, M. SIM & K.M. TEO, 2003, *Vehicle routing problem with time windows and a limited number of vehicles*, European Journal of Operational Research, **148**(3), pp. 559-569
- [34] LEX JANSEN, REGRESSION ANALYSIS IN THE REAL WORLD, [ONLINE], Retrieved from <http://www.lexjansen.com/nesug/nesug90/NESUG9009.PDF>
- [35] S. LOKER, F. MIGLIOR, J. BOHMANOVA, J. JAMROZIK & L.R. SCHAEFFER, 2009, *Phenotypic analysis of pregnancy effect on milk, fat, and protein yields of Canadian Ayrshire, Jersey, Brown Swiss, and Guernsey breeds*, Journal of Dairy Science, **92**(3), pp. 1300-1312
- [36] P. R. MURPHY & D. F. WOOD, 2008, *Contemporary logistics*, 9th Edition, Boston, Upper saddle river: Pearson Prentice Hall
- [37] M.E. O'KELLY, 1987, *A quadratic integer program for the location of interacting hub facilities*, European Journal of Operational Research, **32**(3), pp. 393-404
- [38] B. PAPADOPOULOS, K.P. TSAGARAKIS & A. YANNOPOULOS, 2007, *Cost and Land Functions for Wastewater Treatment Projects: Typical Simple Linear Regression versus Fuzzy Linear Regression*, Journal of environmental engineering, **133**(6), pp. 581-586
- [39] I. PARDOE, 2006, *Applied Regression modelling a business approach*, Wiley
- [40] M.L. PEARL, C.M. YASHAR, C.M. JOHNSON, R.K. REYNOLDS & J.A. ROBERTS, 1994, *Exponential regression of CA 125 during salvage treatment of ovarian cancer with taxol.*, Gynecologic oncology **53**(3), pp. 339 - 343
- [41] I. PULIDO-CALVO, J.C. GUTIERREZ-ESTRADS, E. DIAZ-RUBIO & I. DE LA ROSA, 2014, *Assisted management of water exchange in traditional semi-intensive aquaculture ponds*, Computers and electronics in Agriculture, **101**, pp. 128-134
- [42] R CORE TEAM, R-3.2.0 FOR WINDOWS, [Computer program], Available at <http://cran.r-project.org/bin/windows/base/>, (Accessed 10 Jun 2015)
- [43] S. ROPKE, 2005, *Heuristics and exact algorithms for vehicle routing problems*, Ph.D Thesis, University of Copenhagen, Copenhagen
- [44] P.J. ROUSSEUW, 1984, *Least median of squares regression*, Journal of the american statistical association, **79**(388), pp. 871-880
- [45] S.Y. SOHN & S.G. LEE, 2012, *Probe test yield optimization based on canonical correlation analysis between process control monitoring variables and probe bin variables*, Expert Systems with Applications, **39**(4), pp. 4377-4382
- [46] S.J. STEEL & D.W. UYS, 2007, *Variable selection in multiple linear regression: The influence of individual cases*, ORION, **23**(2), pp. 123-136
- [47] B.C. TANSEL, R.L. FRANCIS & T.J. LOWE, 1983, *State of the art - location on Networks: A Survey. Part 1: The p-Center and p-Median Problems*, Management science, **29**(4), pp. 482-497
- [48] G. TREMBLAY, P. LEGENDRE, J.F. DOYON, R. VERDON & R. SCHETAGNE, 1998, *The use of polynomial regression analysis with indicator variables for interpretation of mercury in fish data*, Biogeochemistry, **40**(2), pp. 189-201

-
- [49] G.K. UYANIK & N. GULER, 2013, *A study on multiple linear regression analysis*, Procedia Social and Behavioral Sciences, **106**, pp. 234-240
- [50] Z. WANG, Z. HE & J.D.Z. CHEN, 2005, *Robust Time Delay Estimation of Bioelectric Signals Using Least Absolute Deviation Neural Network*, IEEE Transactions on biomedical engineering, **52**(3), pp. 454-462
- [51] J.H. WILSON & B. KEATING, 2009, *Business Forecasting with ForecastX*, 6th Edition, Boston, McGraw-Hill/Irwin
- [52] X. XIAO, E.P. WHITE, M.B. HOOTEN & S.L. DURHAM, 2011, *On the use of log-transformation vs. nonlinear regression for analyzing biological power laws.*, Ecology, **92**(10), pp. 1887-1894

Appendix A

SLR analysis plots

A.0.1 Results part 1

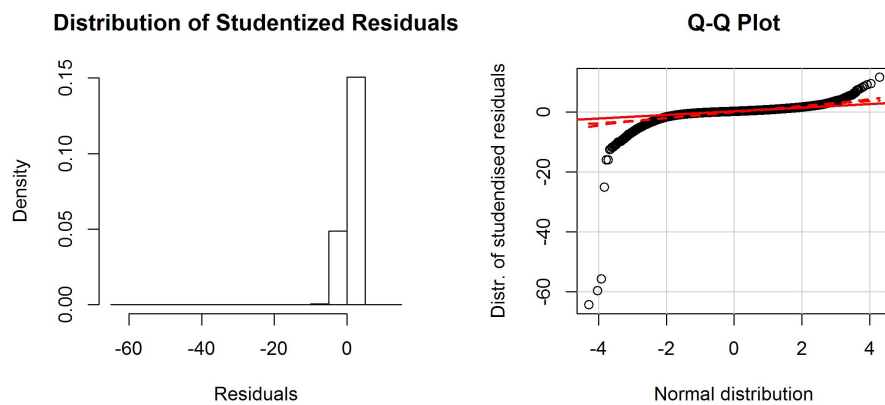


Figure A.1: Residual histogram and Q-Q plots for the model with volume independent variables and volume as dependent variable.

A.0.2 Results part 2

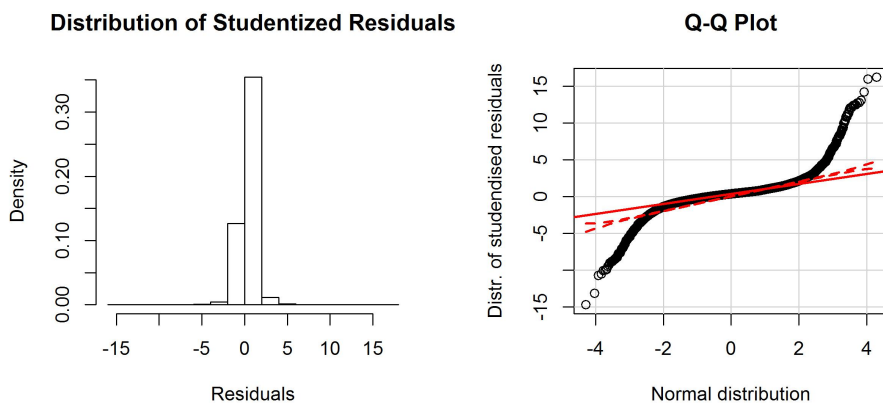


Figure A.2: Residual histogram and Q-Q plots for the model with units independent variables and cartons as dependent variable.

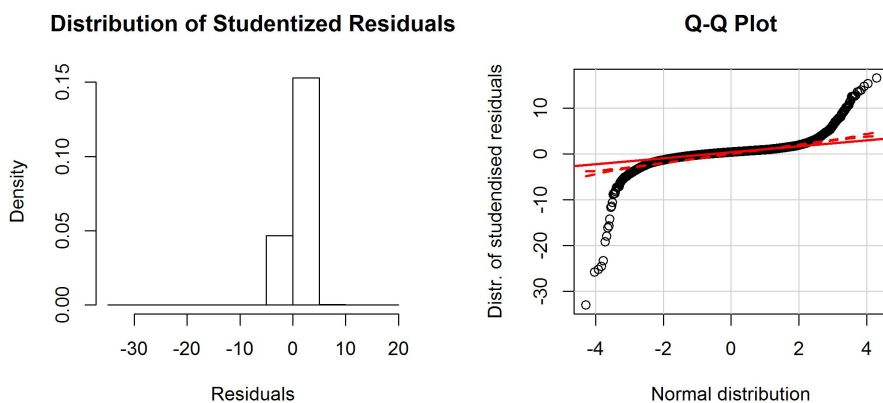


Figure A.3: Residual histogram and Q-Q plots for the model with volumes independent variables and cartons as dependent variable.

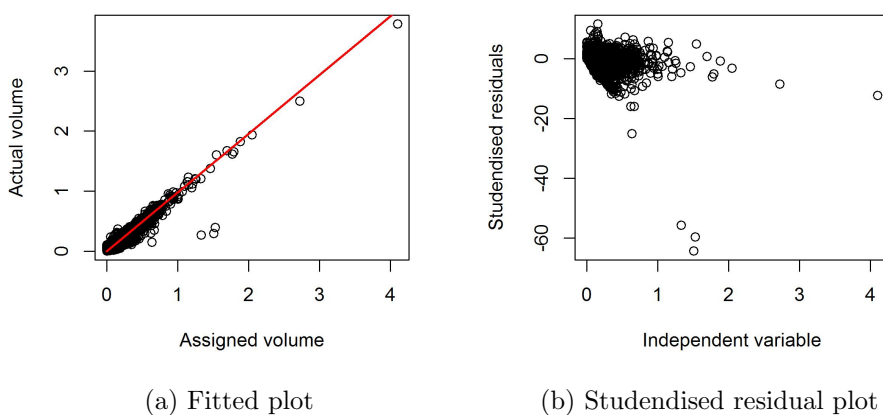


Figure A.4: Fitted and studentised residual plots for the SLR model 2.

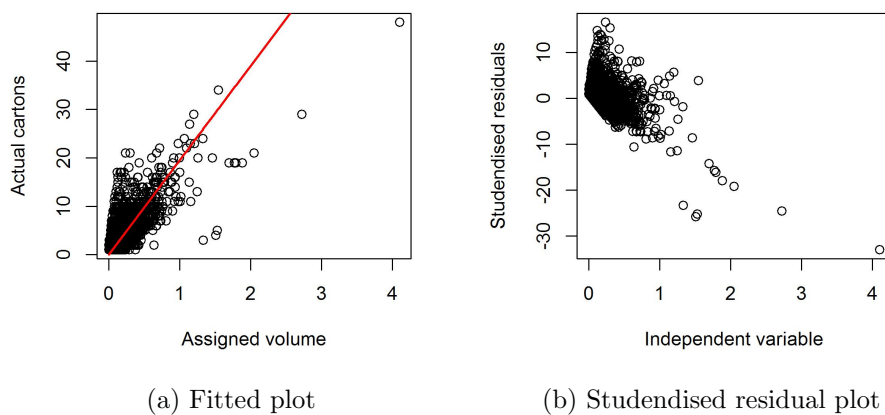


Figure A.5: Fitted and studentised residual plots for the SLR model 3.

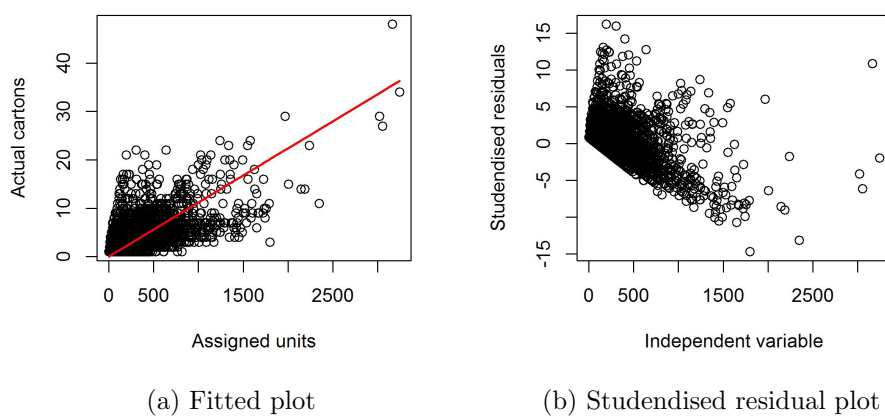


Figure A.6: Fitted and studentised residual plots for the SLR model 4.

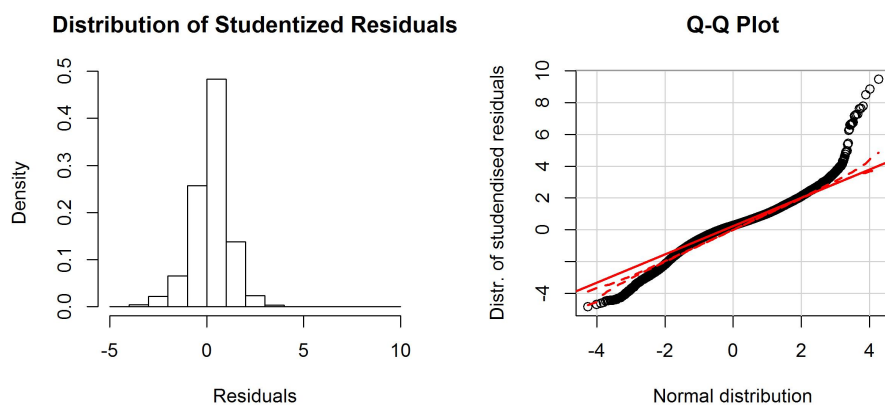


Figure A.7: Residual histogram and Q-Q plots for model 2 after influential observations.

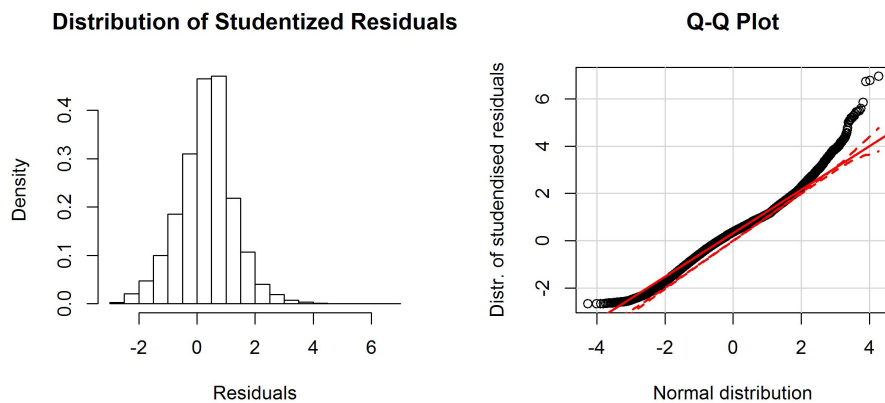


Figure A.8: Residual histogram and Q-Q plots for model 3 after influential observations.

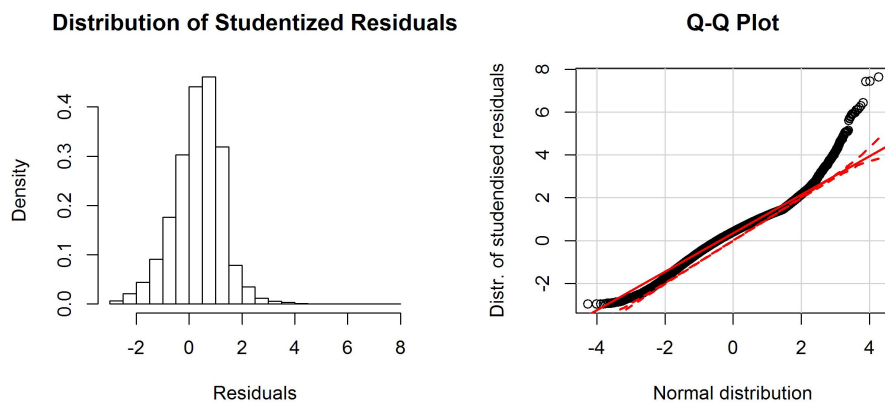


Figure A.9: Residual histogram and Q-Q plots for model 4 after influential observations.

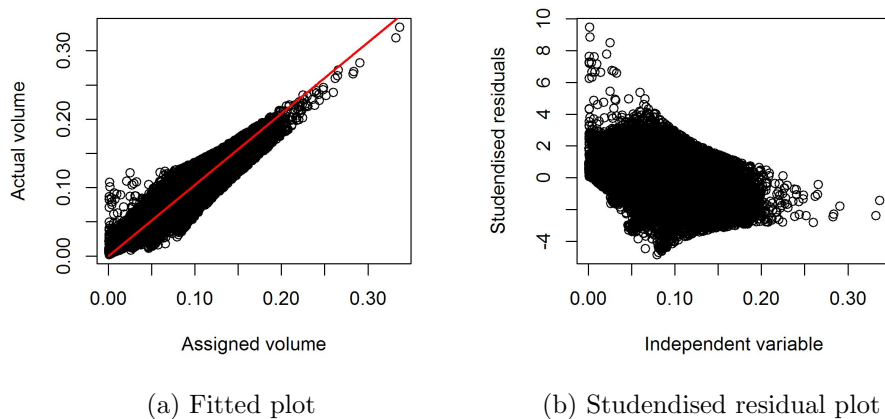


Figure A.10: Fitted and studentised residual plots for the SLR model 2 after influential observations were removed.

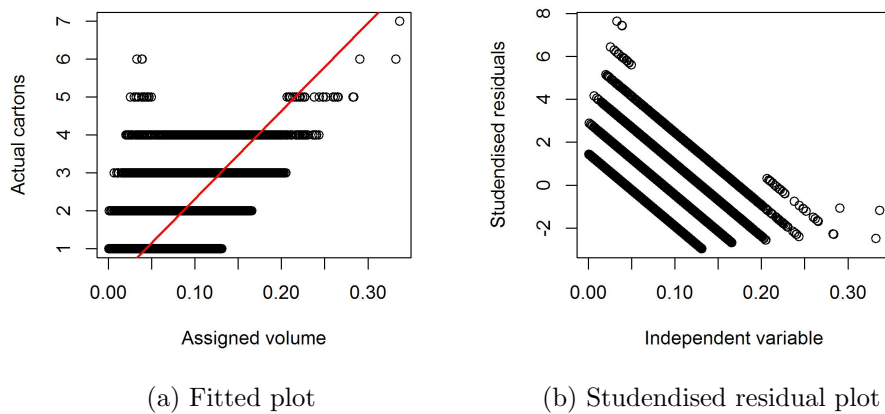


Figure A.11: Fitted and studentised residual plots for the SLR model 3 after influential observations were removed.

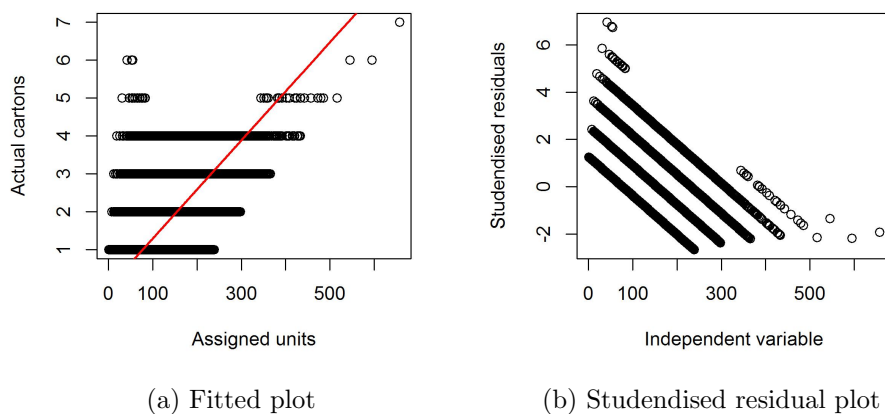


Figure A.12: Fitted and studentised residual plots for the SLR model 4 after influential observations were removed.

Appendix B

MLR analysis

B.1 MLR coefficient statistics

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
BU1 - Babies	0.00024	0	98.686	0
BU2 - Kids	0.00063	0	60.353	0
BU3 - Adults	0.00035	0	58.068	0
BU4 - Home	0.00003	0	3.384	0.001
BU5 - FMCG	0.00065	0	189.679	0
BU10 - Ess	0.00081	0	239.339	0

Table B.1: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the BU MLR model for sub model 2, BVU.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
BU1 - Babies	14.41503	0.121	118.676	0
BU2 - Kids	18.188	0.341	53.29	0
BU3 - Adults	14.64683	0.233	62.9	0
BU4 - Home	23.43314	0.481	48.74	0
BU5 - FMCG	18.50445	0.123	150.651	0
BU10 - Ess	22.84125	0.075	303.726	0

Table B.2: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values on the BU MLR model for sub model 3, BCV.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
BU1 - Babies	0.00517	0	91.221	0
BU2 - Kids	0.01074	0	44.246	0
BU3 - Adults	0.00797	0	57.868	0
BU4 - Home	0.00522	0	22.459	0
BU5 - FMCG	0.01343	0	168.163	0
BU10 - Ess	0.01742	0	223.588	0

Table B.3: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values on the BU MLR model for sub model 4, BCU.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
CAT1 - Accessories	0.0004	0	93.497	0
CAT2 - Clothing	0.00028	0	125.119	0
CAT3 - FMCG	0.00058	0	175.374	0
CAT4 - Underwear	0.00072	0	179.876	0
CAT5 - Home	0.00112	0	40.255	0
CAT6 - Luggage	0.00303	0	163.667	0
CAT7 - Stationary	-0.00011	0	-5.21	0

Table B.4: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the Category MLR model for sub model 2, CVU.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
CAT1 - Accessories	14.70252	0.15	98.217	0
CAT2 - Clothing	17.63179	0.106	166.622	0
CAT3 - FMCG	18.32229	0.124	148.068	0
CAT4 - Underwear	29.74873	0.174	171.396	0
CAT5 - Home	21.10866	0.54	39.088	0
CAT6 - Luggage	18.54273	0.166	111.536	0
CAT7 - Stationary	28.99212	1.546	18.754	0

Table B.5: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the Category MLR model for sub model 3, CCV.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
CAT1 - Accessories	0.00865	0	82.162	0
CAT2 - Clothing	0.00633	0	113.954	0
CAT3 - FMCG	0.01242	0	153.569	0
CAT4 - Underwear	0.01583	0	161.396	0
CAT5 - Home	0.03099	0.001	45.381	0
CAT6 - Luggage	0.05215	0	114.448	0
CAT7 - Stationary	0.00124	0.001	2.334	0.01958

Table B.6: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the Category MLR model for sub model 4, CCU.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
BU1 - Babies	0.00023	0	88.73	0
BU2 - Kids	0.00059	0	55.921	0
BU3 - Adults	0.00028	0	42.174	0
BU4 - Home	0	0	-0.468	0.63964
BU6 - FMCG	0.00064	0	184.291	0
BU10 - Ess	0.00068	0	98.5	0
SKU count	0.00052	0	21.911	0

Table B.7: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the BU SKU count MLR model for sub model 2, BSVU.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
BU1 - Babies	11.23372	0.114	98.432	0
BU2 - Kids	14.53535	0.312	46.629	0
BU3 - Adults	7.73869	0.221	35.044	0
BU4 - Home	17.23343	0.44	39.143	0
BU6 - FMCG	17.54913	0.112	156.863	0
BU10 - Ess	14.43479	0.103	139.701	0
SKU count	0.03214	0	108.404	0

Table B.8: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the BU SKU count MLR model for sub model 3, BSCV.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
BU1 - Babies	0.00429	0	74.309	0
BU2 - Kids	0.0085	0	35.412	0
BU3 - Adults	0.00432	0	28.7	0
BU4 - Home	0.00306	0	13.31	0
BU6 - FMCG	0.01273	0	161.299	0
BU10 - Ess	0.01013	0	65.251	0
SKU count	0.02872	0.001	53.873	0

Table B.9: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the BU Sku count MLR model for sub model 4, BSCU.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
CAT1 - Accessories	0.00024	0	51.283	0
CAT2 - Clothing	0.00024	0	104.462	0
CAT3 - FMCG	0.00054	0	167.508	0
CAT4 - Underwear	0.00028	0	37.329	0
CAT5 - Home	0.00089	0	33.034	0
CAT6 - Luggage	0.00278	0	152.936	0
CAT7 - Stationary	-0.00019	0	-9.264	0
SKU count	0.0014	0	66.341	0

Table B.10: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the Category Sku count MLR model for sub model 2, CSVU.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
CAT1 - Accessories	8.17445	0.146	55.897	0
CAT2 - Clothing	12.76709	0.104	122.463	0
CAT3 - FMCG	16.91739	0.112	151.106	0
CAT4 - Underwear	4.16561	0.273	15.261	0
CAT5 - Home	11.91799	0.492	24.211	0
CAT6 - Luggage	16.391	0.151	108.775	0
CAT7 - Stationary	15.13024	1.396	10.842	0
SKU count	0.0443	0	114.25	0

Table B.11: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the Category Sku count MLR model for sub model 3, CSCV.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
CAT1 - Accessories	0.00297	0	26.735	0
CAT2 - Clothing	0.00471	0	88.288	0
CAT3 - FMCG	0.01105	0	146.631	0
CAT4 - Underwear	0.0001	0	0.577	0.56383
CAT5 - Home	0.02274	0.001	36.041	0
CAT6 - Luggage	0.04321	0	101.304	0
CAT7 - Stationary	-0.00167	0	-3.427	0.00061
SKU count	0.05092	0	102.621	0

Table B.12: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the Category Sku count MLR model for sub model 4, CSCU.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
BU1 - Babies	0.00484	0	78.031	0
BU2 - Kids	0.00064	0	57.799	0
BU3 - Adults	0.0003	0	47.536	0
BU4 - Home	0.0001	0	9.704	0
BU6 - FMCG	0.00066	0	179.58	0
BU10 - Ess	0.00067	0	134.974	0
Clothing Dummy	0.02466	0	74.949	0

Table B.13: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the BU Clothing indicator MLR model for sub model 2, BCIVU.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
BU1 - Babies	13.68702	0.316	43.248	0
BU2 - Kids	17.78056	0.364	48.786	0
BU3 - Adults	11.02698	0.253	43.642	0
BU4 - Home	23.01951	0.512	44.922	0
BU6 - FMCG	18.86247	0.132	143.38	0
BU10 - Ess	21.4013	0.129	165.327	0
Clothing Dummy	0.95447	0.007	145.047	0

Table B.14: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the BU Clothing indicator MLR model for sub model 3, BCICV.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
BU1 - Babies	0.0539	0.001	37.861	0
BU2 - Kids	0.01058	0	41.868	0
BU3 - Adults	0.00665	0	45.795	0
BU4 - Home	0.00577	0	23.787	0
BU6 - FMCG	0.01344	0	160.488	0
BU10 - Ess	0.01269	0	111.638	0
Clothing Dummy	0.79601	0.008	105.399	0

Table B.15: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the BU Clothing indicator MLR model for sub model 4, BCICU.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
CAT1 - Accessories	0.00036	0	77.621	0
CAT3 - FMCG	0.00058	0	163.9	0
CAT4 - Underwear	0.00063	0	128.711	0
CAT5 - Home	0.00111	0	37.233	0
CAT6 - Luggage	0.00288	0	144.417	0
CAT7 - Stationary	-0.00024	0	-10.276	0
Clothing Dummy	0.02555	0	82.31	0

Table B.16: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the Category Clothing indicator MLR model for sub model 2, CCIVU.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
CAT1 - Accessories	13.89649	0.162	85.748	0
CAT3 - FMCG	19.1294	0.132	144.755	0
CAT4 - Underwear	24.11268	0.208	115.71	0
CAT5 - Home	18.66719	0.579	32.236	0
CAT6 - Luggage	19.33582	0.178	108.855	0
CAT7 - Stationary	15.75821	1.654	9.527	0
Clothing Dummy	0.91564	0.007	131.413	0

Table B.17: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the Category Clothing indicator MLR model for sub model 3, CCICU.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
CAT1 - Accessories	0.00718	0	66.906	0
CAT3 - FMCG	0.0123	0	151.192	0
CAT4 - Underwear	0.01198	0	105.709	0
CAT5 - Home	0.02886	0.001	41.954	0
CAT6 - Luggage	0.04715	0	102.317	0
CAT7 - Stationary	-0.00164	0.001	-3.092	0.00199
Clothing Dummy	0.79521	0.007	110.684	0

Table B.18: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the Category Clothing indicator MLR model for sub model 4, CCICU.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
BU1 - Babies	0.00451	0	71.872	0
BU2 - Kids	0.00058	0	52.031	0
BU3 - Adults	0.00022	0	31.463	0
BU4 - Home	0.00004	0	4.081	0.00005
BU6 - FMCG	0.00064	0	175.257	0
BU10 - Ess	0.00049	0	59.183	0
SKU count	0.01775	0	27.006	0
Clothing dummy	0.00079	0	42.788	0

Table B.19: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the BU Combined variable MLR model for sub model 2, BCVU.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
BU1 - Babies	9.12562	0.304	30.045	0
BU2 - Kids	14.60895	0.346	42.214	0
BU3 - Adults	6.33983	0.245	25.864	0
BU4 - Home	17.8137	0.488	36.528	0
BU6 - FMCG	18.253	0.124	146.772	0
BU10 - Ess	14.16145	0.15	94.111	0
SKU count	0.48289	0	82.39	0
Clothing dummy	0.03324	0.008	57.18	0

Table B.20: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the BU Combined variable MLR model for sub model 3, BCCV.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
BU1 - Babies	0.03911	0.001	27.667	0
BU2 - Kids	0.00801	0	31.935	0
BU3 - Adults	0.00296	0	18.878	0
BU4 - Home	0.0031	0	12.871	0
BU6 - FMCG	0.01282	0	155.58	0
BU10 - Ess	0.00457	0	24.548	0
SKU count	0.03571	0.009	51.733	0
Clothing dummy	0.48325	0.001	54.347	0

Table B.21: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the BU Combined variable MLR model for sub model 4, BCCU.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
CAT1 - Accessories	0.00022	0	42.8	0
CAT3 - FMCG	0.00054	0	154.701	0
CAT4 - Underwear	0.00026	0	30.527	0
CAT5 - Home	0.00094	0	32.233	0
CAT6 - Luggage	0.0027	0	136.996	0
CAT7 - Stationary	-0.0003	0	-13.2	0
SKU count	0.00149	0	33.18	0
Clothing dummy	0.01281	0	53.181	0

Table B.22: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the Category Combined dummy variable MLR model for sub model 2, CCVU.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
CAT1 - Accessories	8.25422	0.162	50.848	0
CAT3 - FMCG	17.55831	0.124	141.623	0
CAT4 - Underwear	1.91292	0.307	6.233	0
CAT5 - Home	10.95256	0.545	20.113	0
CAT6 - Luggage	17.28404	0.167	103.771	0
CAT7 - Stationary	6.27473	1.541	4.072	0.00005
SKU count	0.048	0.009	46.932	0
Clothing dummy	0.39976	0.001	93.241	0

Table B.23: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the Category Combined dummy variable MLR model for sub model 3, CCCV.

Variable	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
CAT1 - Accessories	0.00294	0	24.952	0
CAT3 - FMCG	0.0112	0	141.355	0
CAT4 - Underwear	0.0005	0	2.592	0.00954
CAT5 - Home	0.02371	0.001	35.857	0
CAT6 - Luggage	0.0417	0	93.339	0
CAT7 - Stationary	-0.0035	0.001	-6.877	0
SKU count	0.04588	0.009	45.887	0
Clothing dummy	0.40174	0.001	72.427	0

Table B.24: This table contains the coefficients and corresponding estimates, standard errors, *t*-value and *p*-values for the Category Combined dummy variable MLR model for sub model 4, CCCU.

B.2 MLR regression statistics

Sub model nr.	Sub model desc	R^2	Adj. R^2	RSE	Breusch-Pagan test		
					χ^2	DF	p -value
2	BVU	0.82	0.82	0.04	3945.54	5	0
3	BCV	0.79	0.79	1.02	3900.69	5	0
4	BCU	0.81	0.81	0.96	5348.55	5	0

Table B.25: This table contains the R^2 , adjusted R^2 , RSE and Breusch-Pagan heteroscedasticity test results for the BU MLR model for sub models 2, 3 and 4 (CVU, CCV, CCU).

Sub model nr.	Sub model desc	R^2	Adj. R^2	RSE	Breusch-Pagan test		
					χ^2	DF	p -value
2	CVU	0.84	0.84	0.04	3499.89	5	0
3	CCV	0.80	0.80	1.01	1085.33	5	0
4	CCU	0.81	0.81	0.96	9008.55	5	0

Table B.26: This table contains the R^2 , adjusted R^2 , RSE and Breusch-Pagan heteroscedasticity test results for the Category MLR model for sub models 2, 3 and 4 (CVU, CCV, CCU).

Sub model nr.	Sub model desc	R^2	Adj. R^2	RSE	Breusch-Pagan test		
					χ^2	DF	p -value
2	BSVU	0.82	0.82	0.04	4478.58	5	0
3	BSCV	0.80	0.80	0.99	4555.57	5	0
4	BSCU	0.85	0.85	0.87	5105.49	5	0

Table B.27: This table contains the R^2 , adjusted R^2 , RSE and Breusch-Pagan heteroscedasticity test results for the BU Sku count MLR model for sub models 2, 3 and 4 (BSVU, BSCV, BSCU).

Sub model nr.	Sub model desc	R^2	Adj. R^2	RSE	Breusch-Pagan test		
					χ^2	DF	p -value
2	CSVU	0.85	0.85	0.04	3815.65	6	0
3	CSCV	0.83	0.83	0.92	1355.47	6	0
4	CSVU	0.85	0.85	0.86	6629.81	6	0

Table B.28: This table contains the R^2 , adjusted R^2 , RSE and Breusch-Pagan heteroscedasticity test results for the Category Sku count MLR model for sub models 2, 3 and 4 (CSVU, CSCV, CSCU).

Sub model nr.	Sub model desc	R^2	Adj. R^2	RSE	Breusch-Pagan test		
					χ^2	DF	p -value
2	BCIVU	0.80	0.80	0.05	4043.42	6	0
3	BCICV	0.77	0.77	1.06	4938.65	6	0
4	BCICU	0.89	0.89	1.02	7665.43	6	0

Table B.29: This table contains the R^2 , adjusted R^2 , RSE and Breusch-Pagan heteroscedasticity test results for the BU Clothing indicator MLR model for sub models 2, 3 and 4 (BCIVU, BCICV, BCICU).

Sub model nr.	Sub model desc	R^2	Adj. R^2	RSE	Breusch-Pagan test		
					χ^2	DF	p -value
2	CCIVU	0.82	0.82	0.04	3565.76	6	0
3	CCICV	0.80	0.80	1.01	1247.62	6	0
4	CCICU	0.80	0.80	1.03	9204.05	6	0

Table B.30: This table contains the R^2 , adjusted R^2 , RSE and Breusch-Pagan heteroscedasticity test results for the Category Clothing indicator MLR model for sub models 2, 3 and 4 (CCIVU, CCICV, CCICU).

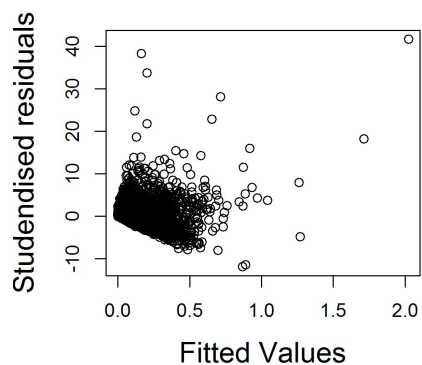
Sub model nr.	Sub model desc	R^2	Adj. R^2	RSE	Breusch-Pagan test		
					χ^2	DF	p -value
2	BCDVU	0.80	0.80	0.05	4470.30	7	0
3	BCDCV	0.78	0.78	1.03	4910.66	7	0
4	BCDCU	0.81	0.81	0.97	5557.73	7	0

Table B.31: This table contains the R^2 , adjusted R^2 , RSE and Breusch-Pagan heteroscedasticity test results for the BU Combined variable MLR model for sub models 2, 3 and 4 (BCVU, BCCV, BCCU).

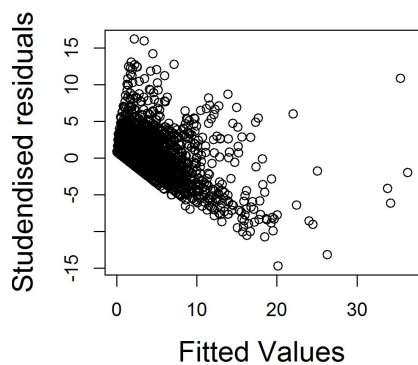
Sub model nr.	Sub model desc	R^2	Adj. R^2	RSE	Breusch-Pagan test		
					χ^2	DF	p -value
2	CCDVU	0.82	0.82	0.04	3767.04	7	0
3	CCDCV	0.81	0.81	0.97	1472.91	7	0
4	CCDCU	0.82	0.82	0.95	7171.44	7	0

Table B.32: This table contains the R^2 , adjusted R^2 , RSE and Breusch-Pagan heteroscedasticity test results for the Category Combined variable MLR model for sub models 2, 3 and 4 (CCVU, CCCV, CCCU).

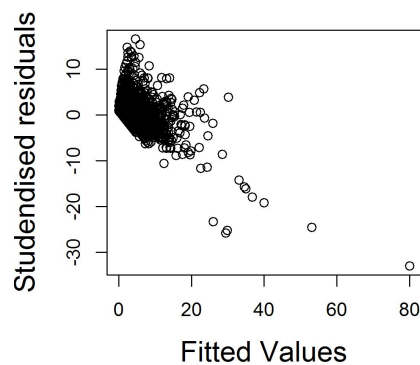
B.3 MLR residual plots



(a) Predicted volumes and assigned units (BVU)

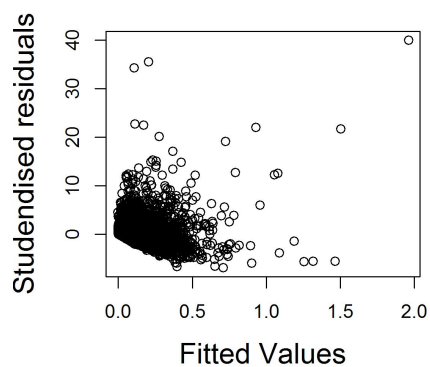


(b) Predicted cartons and assigned volumes (BCV)

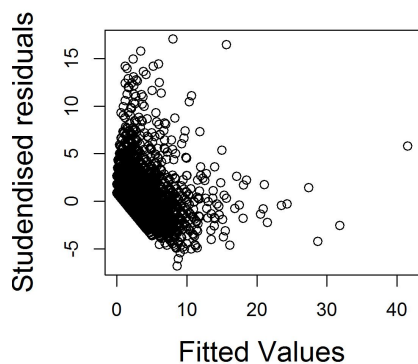


(c) Predicted cartons and assigned units (BCU)

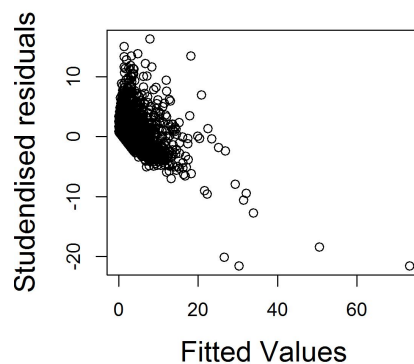
Figure B.1: The studentised residual plots for the BU MLR model for sub models 2, 3 and 4 (BVU, BCV, BCU), that is indicating heteroscedasticity in the residuals. The variance of the studentised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.



(a) Predicted volumes and assigned units (CVU)

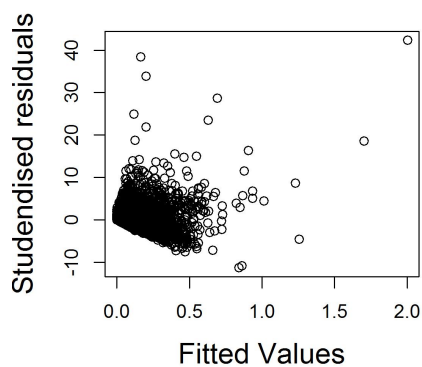


(b) Predicted cartons and assigned volumes (CCV)

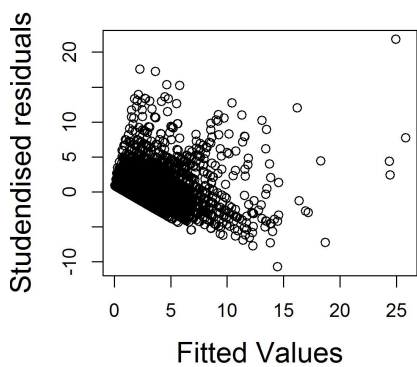


(c) Predicted cartons and assigned units (CCU)

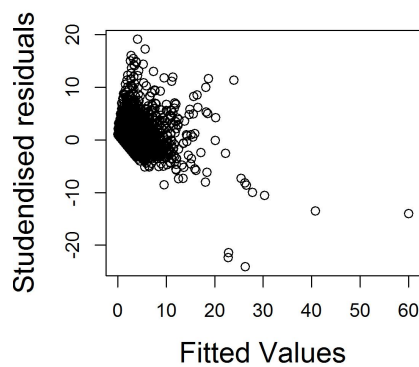
Figure B.2: The studensised residual plots for the Category MLR model for sub models 2, 3 and 4 (CVU, CCV, CCU), that is indicating heteroscedasticity in the residuals. The variance of the studensised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.



(a) Predicted volumes and assigned units (BSVU)

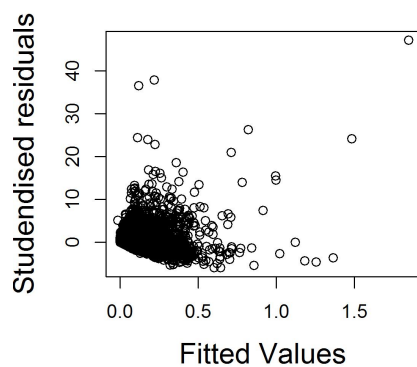


(b) Predicted cartons and assigned volumes (BSCV)

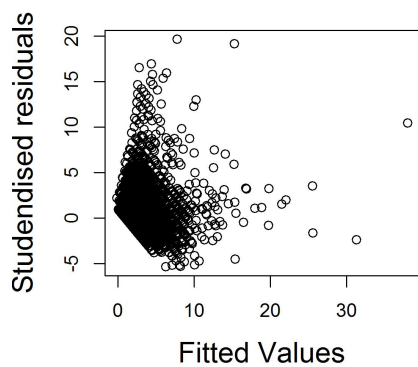


(c) Predicted cartons and assigned units (BSCU)

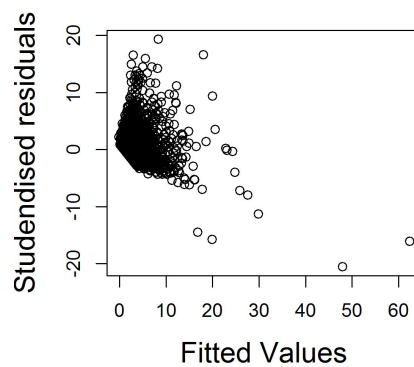
Figure B.3: The studensised residual plots for the BU Sku count MLR model for sub models 2, 3 and 4 (BSVU, BSCV, BSCU), that is indicating heteroscedasticity in the residuals. The variance of the studensised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.



(a) Predicted volumes and assigned units (CSVU)

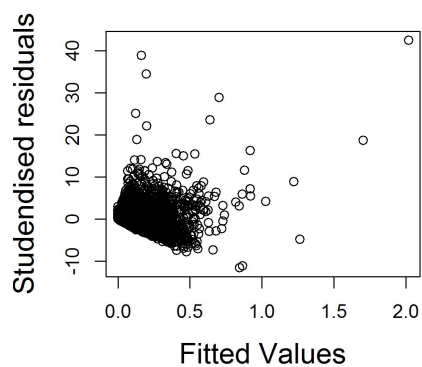


(b) Predicted cartons and assigned volumes (CSCV)

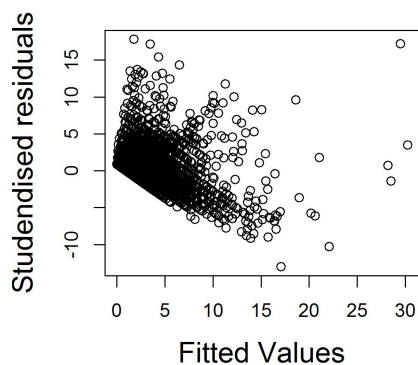


(c) Predicted cartons and assigned units (CSCU)

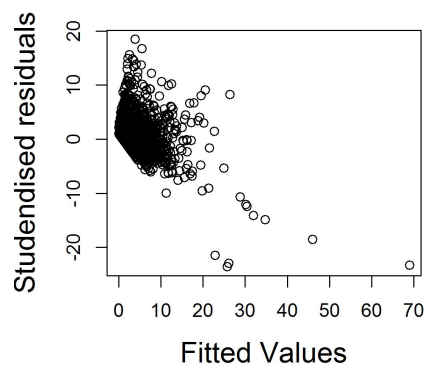
Figure B.4: The studensised residual plots for the Categories Sku count MLR model for sub models 2, 3 and 4 (CSVU, CSCV, CSCU), that is indicating heteroscedasticity in the residuals. The variance of the studensised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.



(a) Predicted volumes and assigned units (BCIVU)

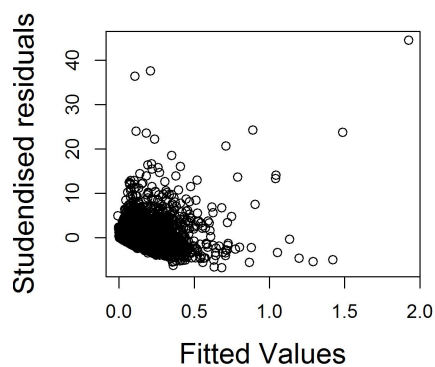


(b) Predicted cartons and assigned volumes (BCICV)

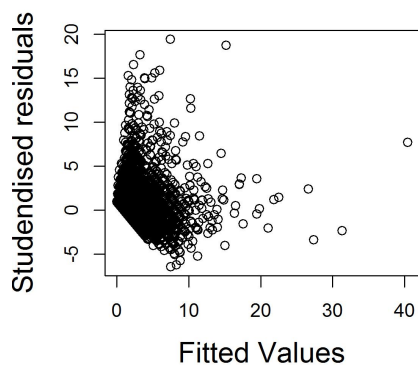


(c) Predicted cartons and assigned units (BCICU)

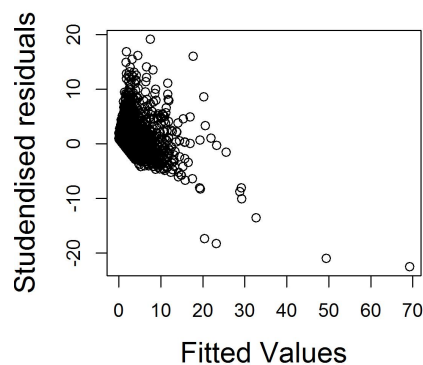
Figure B.5: The studentised residual plots for the BU Clothing indicator MLR model for sub models 2, 3 and 4 (BCIVU, BCICV, BCICU), that is indicating heteroscedasticity in the residuals. The variance of the studentised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.



(a) Predicted volumes and assigned units (CCIVU)

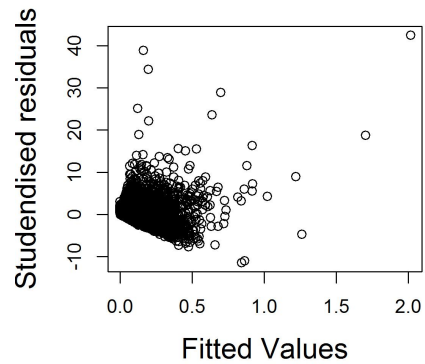


(b) Predicted cartons and assigned volumes (CCICV)

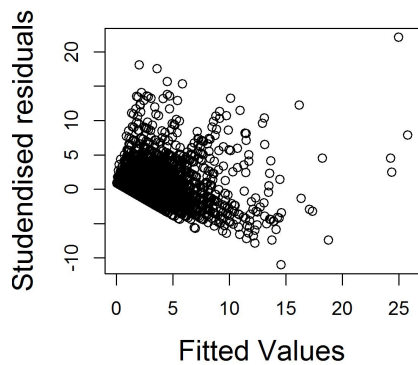


(c) Predicted cartons and assigned units (CCICU)

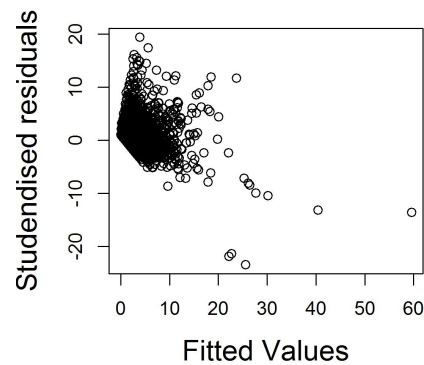
Figure B.6: The studensised residual plots for the Categories Clothing indicator MLR model for sub models 2, 3 and 4 (CCIVU, CCICV, CCICU), that is indicating heteroscedasticity in the residuals. The variance of the studensised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.



(a) Predicted volumes and assigned units (BCDVU)

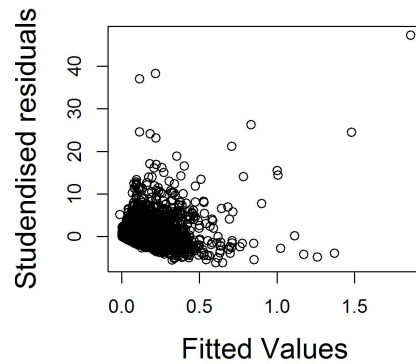


(b) Predicted cartons and assigned volumes (BCDCV)

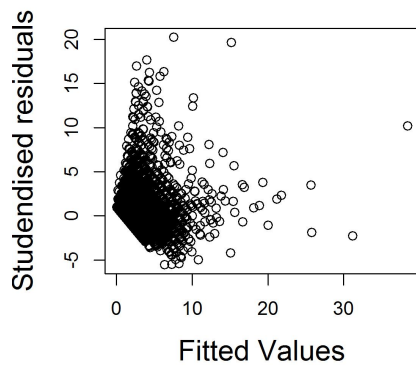


(c) Predicted cartons and assigned units (BCDCU)

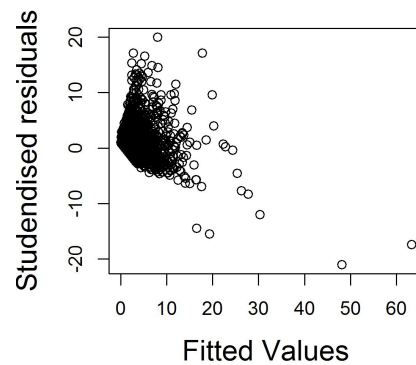
Figure B.7: The studentised residual plots for the BU Combined variable MLR model for sub models 2, 3 and 4 (BCVU, BCCV, BCCU), that is indicating heteroscedasticity in the residuals. The variance of the studentised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.



(a) Predicted volumes and assigned units (CCDVU)



(b) Predicted cartons and assigned volumes (CCDCV)



(c) Predicted cartons and assigned units (CCDCU)

Figure B.8: The studensised residual plots for the Categories Combined variable MLR model for sub models 2, 3 and 4 (CCVU, CCCV, CCCU), that is indicating heteroscedasticity in the residuals. The variance of the studensised residuals increase as the fitted values increase. There are also signs of influential observations that can be considered as outliers.

Appendix C

Heteroscedasticity test results

Model	Predicted carton			Predicted volume		
	χ^2	DF	P-value	χ^2	DF	P-value
Linear-linear	3074.56	0	0	1080.22	0	0
Log-Linear	1048.96	1	0	874.68	1	0
Linear-Log	10087.92	1	0	8976.69	1	0
Log-Log	2753.70	1	0.01	347.06	1	0
Squared-linear	9228.38	1	0	7922.58	1	0
Linear-squared	3504.94	1	0	1540.83	1	0
Squared-squared	3110.93	1	0	1413.45	1	0
Fourthroot-linear	8932.92	1	0	8185.06	1	0
Linear-fourthroot	1845.99	1	0	686.26	1	0
Fourthroot-fourthroot	2123.82	1	0	1413.45	1	0

Table C.1: Heteroscedasticity results from different models with assigned volume as independent variable.

Model	Predicted carton			Predicted volume		
	χ^2	DF	P-value	χ^2	DF	P-value
Linear-linear	3900.69	0	0	3074.56	0	0
Log-Linear	353.89	1	0	285.28	1	0
Linear-Log	7107.09	1	0	8236.49	1	0
Log-Log	500.52	1	0.01	3223.06	1	0
Squared-linear	6693.94	1	0	11408.40	1	0
Linear-squared	2777.40	1	0	1162.63	1	0
Squared-squared	4164.70	1	0	4491.87	1	0
Fourthroot-linear	8891.25	1	0	10601.35	1	0
Linear-fourthroot	1758.13	1	0	647.35	1	0
Fourthroot-fourthroot	3829.43	1	0	1273.36	1	0

Table C.2: Heteroscedasticity results from models with assigned units as independent variable.

Model	Predicted carton			Predicted volume		
	χ^2	DF	P-value	χ^2	DF	P-value
Weighted	2340537	1	0	3268642	1	0
Log weighted	205	1	0	788	1	0
Quadratic weighted	11988	1	0	18478	1	0

Table C.3: Heteroscedasticity results from weighted models.

Appendix D

Quadratic transformation plots and BP test results

D.1 BP test results

Model	Dependent variable	Independent variable	Breusch-Pagan test		
			χ^2	Degrees of freedom	p -value
1	predicted volume	(assigned units) ^{$\frac{1}{2}$}	1162.63	1	0
2	predicted volume	(assigned volume) ^{$\frac{1}{2}$}	1540.84	1	0
3	predicted cartons	(assigned units) ^{$\frac{1}{2}$}	2777.41	1	0
4	predicted cartons	(assigned volume) ^{$\frac{1}{2}$}	3504.94	1	0

Table D.1: Breusch-Pagan heteroscedasticity test results for the linear-squared transformations.

Model	Dependent variable	Independent variable	Breusch-Pagan test		
			χ^2	Degrees of freedom	p -value
1	(predicted volume) ^{$\frac{1}{2}$}	(assigned units) ^{$\frac{1}{2}$}	4491.88	1	0
2	(predicted volume) ^{$\frac{1}{2}$}	(assigned volume) ^{$\frac{1}{2}$}	1413.45	1	0
3	(predicted cartons) ^{$\frac{1}{2}$}	(assigned units) ^{$\frac{1}{2}$}	4164.70	1	0
4	(predicted cartons) ^{$\frac{1}{2}$}	(assigned volume) ^{$\frac{1}{2}$}	3110.93	1	0

Table D.2: Breusch-Pagan heteroscedasticity test results for the squared-squared transformations.

Model	Dependent variable	Independent variable	Breusch-Pagan test		
			χ^2	Degrees of freedom	p -value
1	(predicted volume) ^{$\frac{1}{4}$}	assigned units	10601.35	1	0
2	(predicted volume) ^{$\frac{1}{4}$}	assigned volume	8185.07	1	0
3	(predicted cartons) ^{$\frac{1}{4}$}	assigned units	8891.26	1	0
4	(predicted cartons) ^{$\frac{1}{4}$}	assigned volume	8932.92	1	0

Table D.3: Breusch-Pagan heteroscedasticity test results for the fourth-linear transformations.

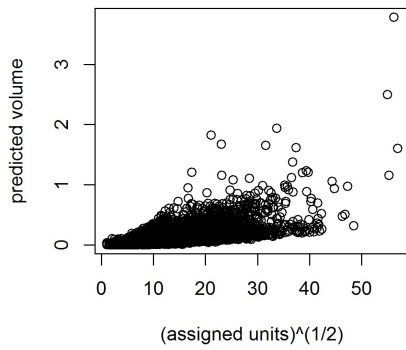
Model	Dependent variable	Independent variable	Breusch-Pagan test		
			χ^2	Degrees of freedom	p -value
1	predicted volume	(assigned units) ^{$\frac{1}{4}$}	647.35	1	0
2	predicted volume	(assigned volume) ^{$\frac{1}{4}$}	686.26	1	0
3	predicted cartons	(assigned units) ^{$\frac{1}{4}$}	1758.13	1	0
4	predicted cartons	(assigned volume) ^{$\frac{1}{4}$}	1845.99	1	0

Table D.4: Breusch-Pagan heteroscedasticity test results for the linear-fourth transformations.

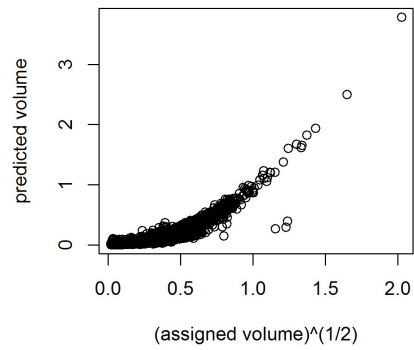
Model	Dependent variable	Independent variable	Breusch-Pagan test		
			χ^2	Degrees of freedom	p -value
1	(predicted volume) ^{$\frac{1}{4}$}	(assigned units) ^{$\frac{1}{4}$}	1273.36	1	0
2	(predicted volume) ^{$\frac{1}{4}$}	(assigned volume) ^{$\frac{1}{4}$}	1413.45	1	0
3	(predicted cartons) ^{$\frac{1}{4}$}	(assigned units) ^{$\frac{1}{4}$}	3829.43	1	0
4	(predicted cartons) ^{$\frac{1}{4}$}	(assigned volume) ^{$\frac{1}{4}$}	2123.82	1	0

Table D.5: Breusch-Pagan heteroscedasticity test results for the fourth-fourth transformations.

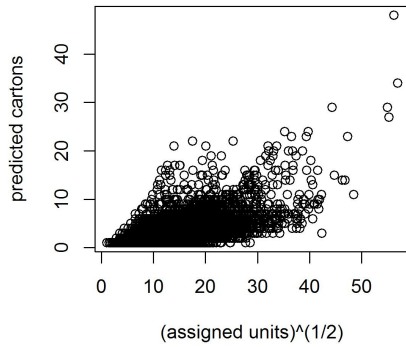
D.2 Transformation plots



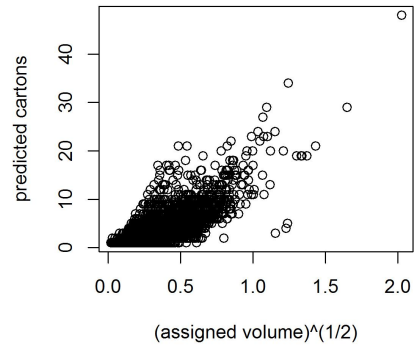
(a)



(b)



(c)



(d)

Figure D.1: Scatter plots between assigned volume/units and predicted volume/cartons with a linear-squared transformation.

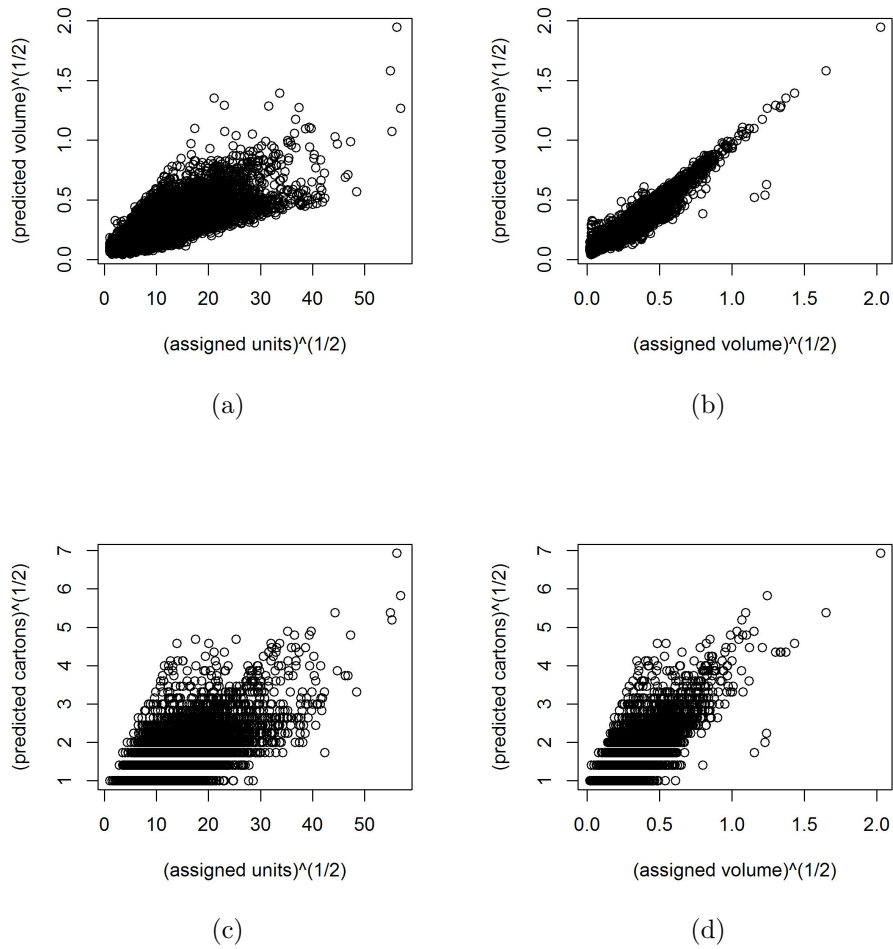


Figure D.2: Scatter plots between assigned volume/units and predicted volume/cartons with a squared-squared transformation.

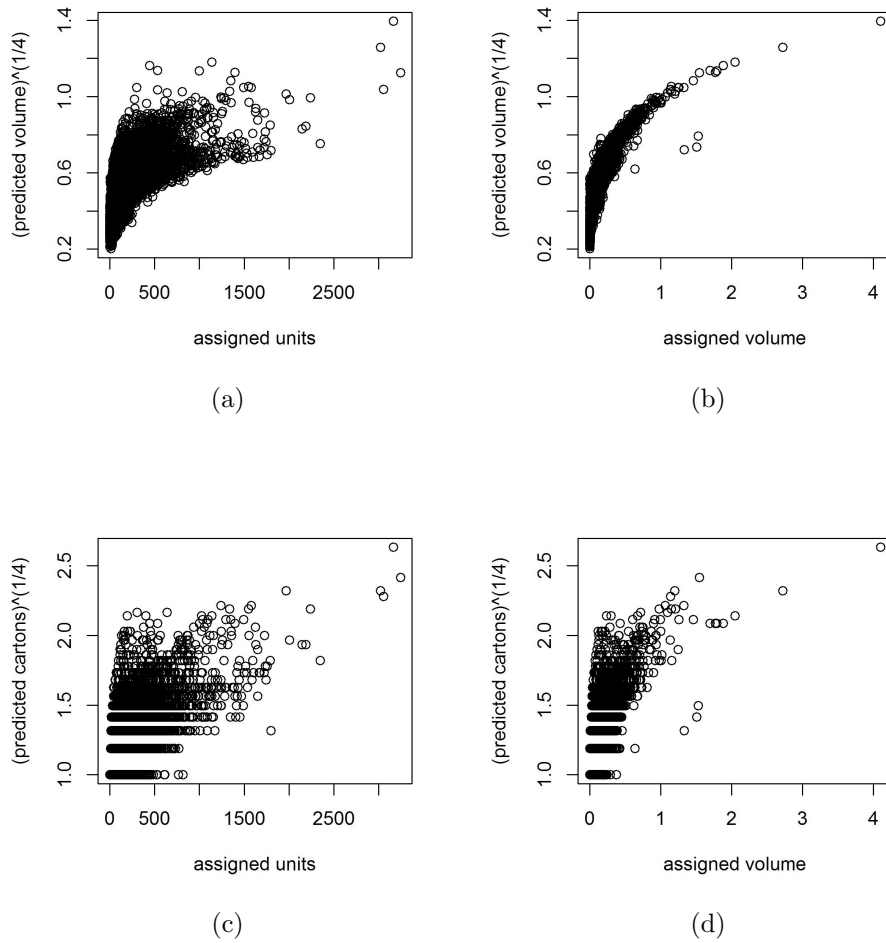


Figure D.3: Scatter plots between assigned volume/units and predicted volume/cartons with a fourth-linear transformation.

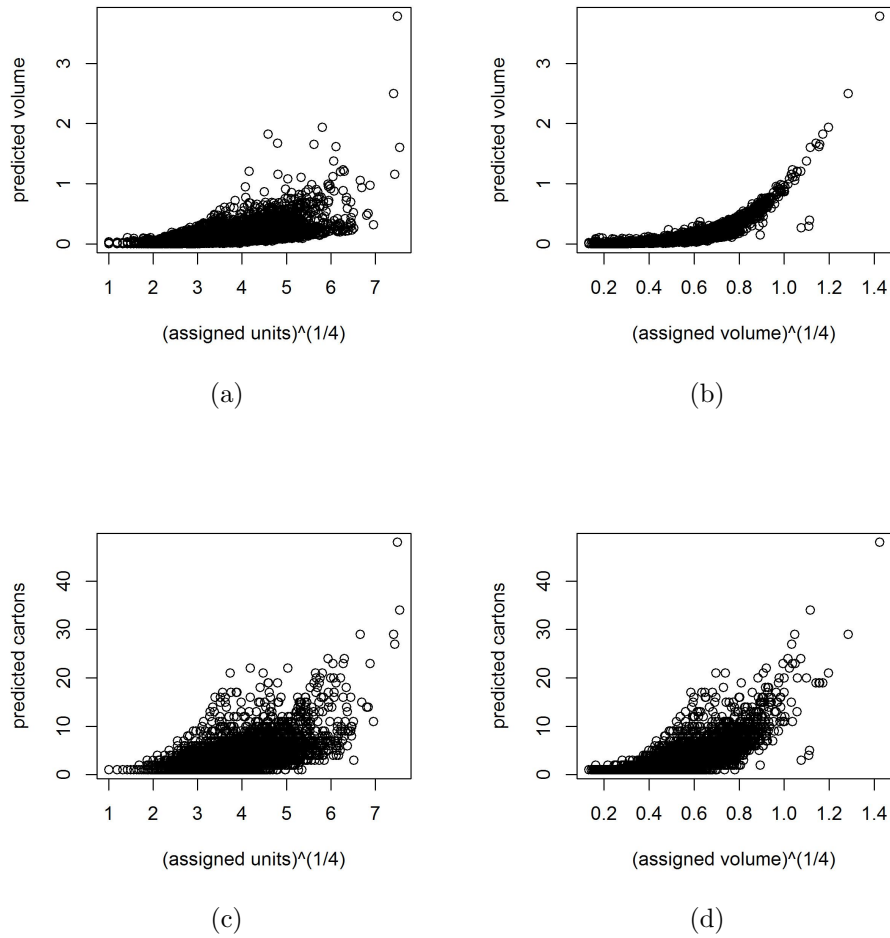


Figure D.4: Scatter plots between assigned volume/units and predicted volume/cartons with a linear-fourth transformation.

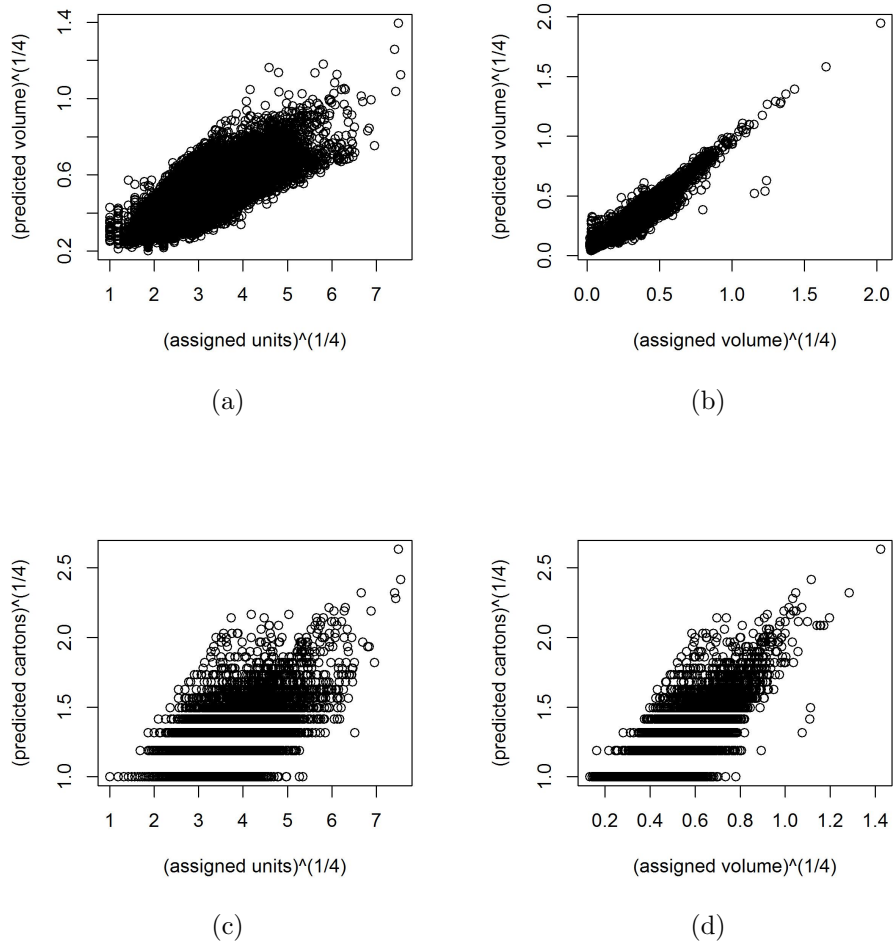


Figure D.5: Scatter plots between assigned volume/units and predicted volume/cartons with a fourth-fourth transformation.

Appendix E

Polynomial plots

Variable	Estimate	Std. Error	<i>t</i> -value	<i>P</i> -value
Intercept	-0.39	0.00	-1119.67	0
<i>x</i>	0.76	0.00	436.98	0
<i>x</i> ²	0.24	0.00	67.27	0
<i>x</i> ³	0.08	0.01	14.28	0

(a) Coefficients

Source of variation	Df	Sum Sq	Mean Sq	F value	<i>P</i> -value
Regression	3	1794.21	598.07	123036.4	0
Error	55182	268.24	0.01		

(b) ANOVA table

Statistic	Value		Value
<i>R</i> ²	0.87	χ^2	1350.87
Adj <i>R</i> ²	0.87	Degrees of freedom	3
RSE	0.07	<i>p</i> -value	0

(c) Regression statistics

(d) BP test results

Table E.1: Coefficients, anova table, regression statistics and BP test results for M1 of order 3.

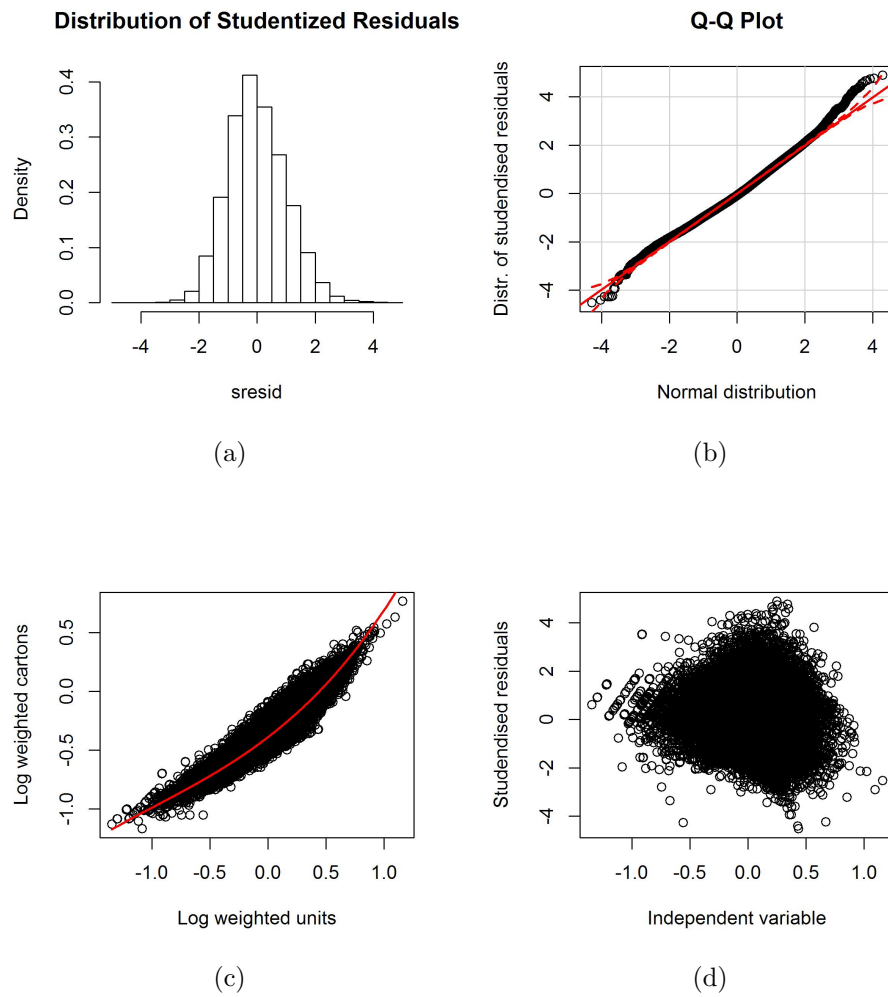


Figure E.1: Regression analysis graphs for M1 of order 3

Variable	Estimate	Std. Error	<i>t</i> -value	<i>P</i> -value
Intercept	-0.857	0.00	-2443.5	0
<i>x</i>	0.976	0.00	725.25	0
<i>x</i> ²	0.035	0.00	11.44	0

(a) Coefficients

Source of variation	Df	Sum Sq	Mean Sq	F value	<i>P</i> -value
Regression	2	2987.37	1493.68	291046.0	0
Error	55183	283.21	0.01		

(b) ANOVA Table

Statistic	Value		Value
<i>R</i> ²	0.91	χ^2	1719.01
Adj <i>R</i> ²	0.91	Degrees of freedom	2
RSE	0.07	<i>p</i> -value	0

(c) Regression statistics

(d) BP test results

Table E.2: Coefficients and anova table for M2 of order 2.

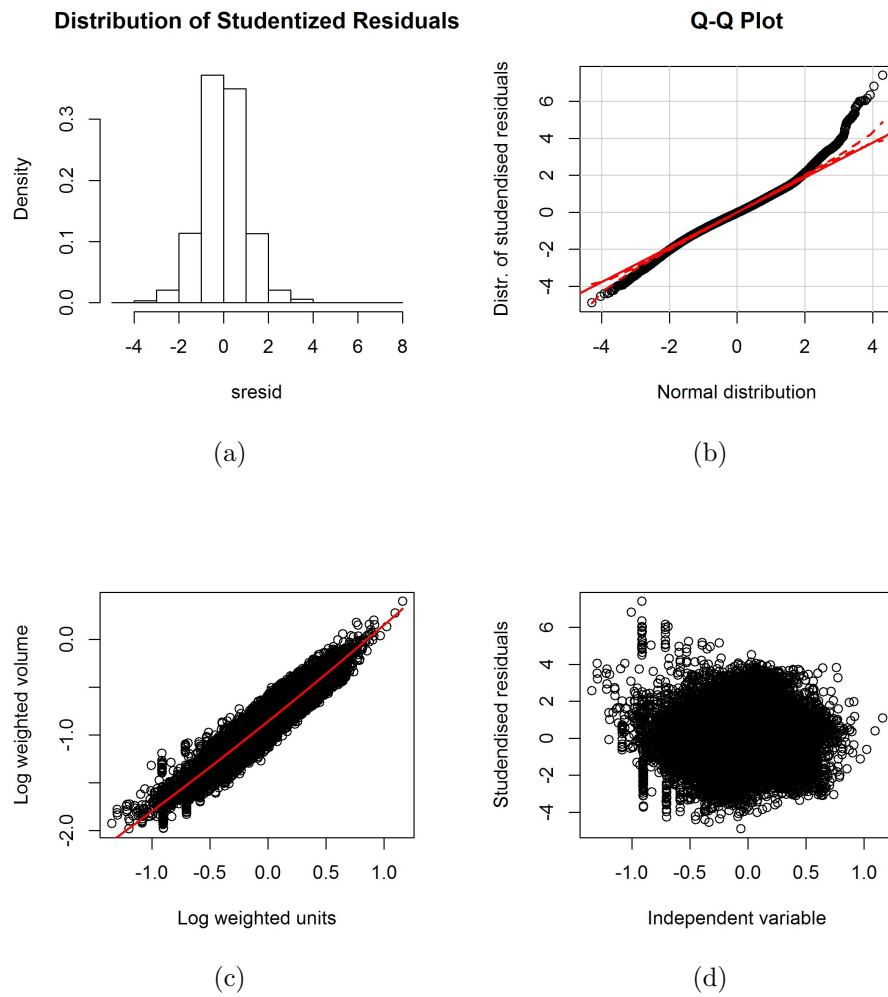


Figure E.2: Regression analysis graphs for M2 of order 2

Appendix F

Forecasting accuracy

Model	Dependent variable	Independent variable	RMSE	NRMSE	MAE	MAPE	PBIAS
BU MLR	Cartons	Volume	1.1	0.55	0.72	38.23%	23.28%
	Cartons	Units	1.23	0.62	0.87	49.61%	-1.71%
	Volume	Volume	0.03	0.4	0	5.72%	3.79%
	Volume	Units	0.07	0.92	0.04	59.47%	-22.54%
Category MLR	Cartons	Volume	1.08	0.54	0.72	38.84%	24.29%
	Cartons	Units	1.31	0.66	0.89	50.21%	0.73%
	Volume	Volume	0.03	0.41	0.01	9.02%	4.57%
	Volume	Units	0.07	0.96	0.04	56.19%	-16.82%
BU SKU count model	Cartons	Volume	1.03	0.52	0.68	38.26%	8.35%
	Cartons	Units	1.23	0.61	0.88	51.93%	-6.14%
	Volume	Volume	0.03	0.42	0.01	13.56%	-3.67%
	Volume	Units	0.07	0.93	0.04	62.33%	-24.58%
Category SKU count model	Cartons	Volume	1.03	0.51	0.68	38.68%	6.94%
	Cartons	Units	1.18	0.59	0.81	47.53%	-5.86%
	Volume	Volume	0.03	0.43	0.01	13.12%	-1.44%
	Volume	Units	0.07	0.95	0.04	57.69%	-21.47%
BU Clothing indicator model	Cartons	Volume	1.12	0.56	0.76	47.65%	-9.62%
	Cartons	Units	5.13	2.56	2.34	116.51%	-99.84%
	Volume	Volume	0.04	0.57	0.03	58.46%	-36.59%
	Volume	Units	0.46	5.88	0.19	253.27%	-233.49%
Category Clothing indicator model	Cartons	Volume	1.15	0.57	0.7	36.53%	12.78%
	Cartons	Units	1.25	0.63	0.82	45.02%	-4.25%
	Volume	Volume	0.04	0.58	0.02	36.30%	-1.41%
	Volume	Units	0.08	0.99	0.04	63.62%	-19.30%
BU combined dummy variable model	Cartons	Volume	1.09	0.54	0.72	44.00%	-10.10%
	Cartons	Units	3.78	1.89	1.88	97.08%	-74.44%
	Volume	Volume	0.04	0.57	0.03	56.53%	-36.92%
	Volume	Units	0.43	5.49	0.18	240.31%	-219.13%
Category combined dummy variable model	Cartons	Volume	1.13	0.57	0.74	41.77%	2.15%
	Cartons	Units	1.22	0.61	0.81	46.30%	-6.39%
	Volume	Volume	0.05	0.59	0.02	41.81%	-8.61%
	Volume	Units	0.08	0.99	0.04	63.12%	-21.08%

Table F.1: Accuracy performance metrics of the actual versus forecasted cartons for the MLR models per day from 16 July 2014 until 25 July 2014.