DEGREE PROJECT IN MATHEMATICS,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, SWEDEN 2018

Purchase behaviour analysis in the retail industry using Generalized Linear Models

SOFIA KARLSSON

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ENGINEERING SCIENCES

Purchase behaviour analysis in the retail industry using Generalized Linear Models

SOFIA KARLSSON

Degree Projects in Financial Mathematics (30 ECTS credits)
Degree Programme in Applied and Computational Mathematics
KTH Royal Institute of Technology year 2018
Supervisors at Indiska AB: Karin Lindahl
Supervisor at KTH: Boualem Djehiche
Examiner at KTH: Boualem Djehiche

TRITA-SCI-GRU 2018:358
MAT-E 2018:75

Royal Institute of Technology
School of Engineering Sciences
KTH SCI
SE-100 44 Stockholm, Sweden
URL: www.kth.se/sci

Abstract

This master thesis uses applied mathematical statistics to analyse purchase behaviour based on customer data of the Swedish brand *Indiska*. The aim of the study is to build a model that can help predicting the sales quantities of different product classes and identify which factors are the most significant in the different models and furthermore, to create an algorithm that can provide suggested product combinations in the purchasing process. Generalized linear models with a Negative binomial distribution are applied to retrieve the predicted sales quantity. Moreover, conditional probability is used in the algorithm which results in a product recommendation engine based on the calculated conditional probability that the suggested combinations are purchased.

From the findings, it can be concluded that all variables considered in the models; original price, purchase month, colour, cluster, purchase country and channel are significant for the predicted outcome of the sales quantity for each product class. Furthermore, by using conditional probability and historical sales data, an algorithm can be constructed which creates recommendations of product combinations of either one or two products that can be bought together with an initial product that a customer shows interest in.

Keywords: Generalized linear models, Algorithm, Historical transaction, Retail, Fashion, Recommendation engine

Analys av köpbeteende inom detaljhandeln med hjälp av generaliserade linjära modeller

Sammanfattning

Matematisk statistik tillämpas i denna masteruppsats för att analysera köpbeteende baserat på kunddata från det svenska varumärket *Indiska*. Syftet med studien är att bygga modeller som kan hjälpa till att förutsäga försäljningskvantiteter för olika produktklasser och identifiera vilka faktorer som är mest signifikanta i de olika modellerna och därtill att skapa en algoritm som ger förslag på rekommenderade produktkombinationer i köpprocessen. Generaliserade linjära modeller med en negativ binomialfördelning utvecklades för att beräkna den förutspådda försäljningskvantiteten för de olika produktklasserna. Dessutom används betingad sannolikhet i algoritmen som resulterar i en produktrekommendationsmotor som baseras på den betingade sannolikheten att de föreslagna produktkombinationerna är inköpta.

Från resultaten kan slutsatsen dras att alla variabler som beaktas i modellerna; originalpris, inköpsmånad, produktfärg, kluster, inköpsland och kanal är signifikanta för det predikterade resultatet av försäljningskvantiteten för varje produktklass. Vidare är det möjligt att, med hjälp av betingad sannolikhet och historisk försäljningsdata, konstruera en algoritm som skapar rekommendationer av produktkombinationer av en eller två produkter som kan köpas tillsammans med en produkt som en kund visar intresse för.

Contents

1. Introduction

The digitalisation[1] within the retail industry has implied multifaceted transformations, affecting industry incumbents with both new customer behaviours, new demands on operational structure, global competition and inhouse competence requirements as well as new opportunities to seize competitive advantage. For many traditional fashion retailers, enabling structures to seize these competitive advantages demands large and complex transformations. Instead, the barriers of entry for new fashion retailers and has decreased and allowed for new players, so called digitally native vertical brands[2], with digitally adapted customer experiences and more efficient supply chains (Dart & Lewis, 2018). These new players has fundamentally changed the playing field and the rules of the game and in combination with the decreasing trend in foot traffic within brick-and-mortar, the pressure on brick-and-mortar[3] retailing has increased (Amed et al., 2017).

As a consequence from the effects of increased digitalisation within the industry, customers are becoming more demanding as well as less loyal and unpredictable (Amed et al., 2016). Through the digitisation[4] and the enabling of AI algorithms used to create recommendation engines on websites, the customers have developed a shorter attention span and less time is expected until they find the right product. Consequently, the customers' expectations on personalisation and convenience have been raised. In the Global Fashion survey conducted by McKinsey&Company and Business of Fashion, personalisation proved to be the largest trend in 2018 and findings in a Linkdex survey suggests that more than 70 percent of customers expect some kind of personal touch. Hence, there is shift in consumer behaviour in which they expect retailers to provide curated information and recommendations, and deliver a more tailored shopping experience (Amed et al., 2017). The digital upgrade has also created a price transparency across brands which has allowed for a discount culture among consumers in which they can seek out promotions and discount prices online. Furthermore, the customers search for product information, product reviews and stock balance of wanted products online and are thus becoming increasingly well-informed about the market and less loyal to specific retailer brands (Svensk Handel, 2018). It will likely be required by retail companies to develop and offer their unique selling point to differentiate from other brands and be able to offer a seamless customer experience (Amed et al., 2016).

In order to stay competitive in a market where customers are expecting increased personalisation and their behaviour is simultaneously difficult to predict it is of higher importance that the companies attain advancements within analytics and take benefit from the digitisation. By engaging in customer relationship management, CRM, and utilize the data of which it enables the companies can be able to track the customer behaviour, get to know more about the customer and tailor the customer strategy of the brand (Amed et al., 2017). An increased focus on analysis of large data sets enriched with information that enables the

[1] Digitalisation is the use of digital technologies to change a business model and provide new revenue and value-producing opportunities; it is the process of moving to a digital business (Gartner, 2018)
[2] Digitally native vertical brands are commonly referred to brands started in the digital era with a large focus on digital solutions for the customer experience and operates primarily on the web. (Dunn, Andy, CEO at Bonobos)
[3] Relating to or being a traditional business serving customers in a building as contrasted to an online business (Merriam-Webster, 2018)
[4] Digitisation takes an analog process and changes it to a digital form without any different-in-kind changes to the process itself (Gartner, 2018)

companies to draw conclusions from customer insights is a way to differentiate and create a competitive advantage by better meeting the needs of the consumer. Also, the information can be used as a way for the customer to acknowledge and pay attention to products they might not have realised they needed (Amed et al., 2017). Today, a lot of brands offer a members club which becomes a source for the company to attain more data on the customers and allows the customers to earn points when shopping which can give benefits such as being transferred into reduction vouchers which can be used next time shopping at the store, thereby increasing loyalty (Amed et al., 2017).

Going forward, this shift to digital platforms with increasing number of players will intensify competition and increase market volatility (Amed et al., 2016). E-commerce with goods is growing fast and the fact that retailers lag in digitalisation combined with the importance of the sector for the economy - given its huge store stock and large number of employees - calls for a special focus to understand tomorrow's consumption patterns and economic landscape (Svensk Handel, 2018). The fashion industry in Sweden has over the recent years had pressured margins causing profitability problems. Meanwhile, the operational costs are increasing with higher rents for store facilities and warehouses and there is a stagnating trend in gross profit margins, implicating increasing demands on companies operating in the retail industry. The situation creates an uncertainty that inter alia affects how the fashion companies on OMX Stockholm are valued. Despite the flourishing economy and general optimism at the time of writing, the valuations of the big fashion companies are historically low in relation to their profits, indicating that the market is expecting major challenges in the future (Svensk Handel, 2018).

By taking advantage of the digitisation and storing data such as product interests and purchasing patterns of a consumer, the brand will in the long run have the possibility to predict the next purchase and can address a member with personalized offers which secures that the right product is being offered at the right time to the right customer (Amed et al., 2017). By predicting consumer purchases, the company can also benefit from optimisation of the production levels as well as the supply chain and hence reduce the risk of overstock and markdowns (Amed et al., 2017). Thereby, all the parts of the value chain are affected in the retail industry (Svensk Handel, 2018), and thus, retail brands must pay attention to predicting and understanding the consumer preferences.

This thesis is conducted in association with the Swedish lifestyle brand *Indiska*, founded in 1901 making it the first lifestyle brand, offering both clothing and interior, in Sweden. *Indiska* offers a unique style, *Boho modern*, which is influenced by India and associated with colour, patterns and sustainability targeting independent, expressive and brave women. The products are sold in both stores and online in Sweden, Norway, Finland as main markets and furthermore there are franchise stores in Iceland and Germany. As of today, the members club offered by *Indiska*, *Lovers & Likers*, have over one million members and is continuously growing. The mere part of the members are female Swedish citizens, hence *Lovers & Likers* has a service range to around ten percent of all Swedish women.

Indiska has recently experienced a distressed situation going through a second corporate restructuring in a short period of time. One of the lessons was that, in an attempt to broaden the target audience, parts of the core customer base was overlooked and that the knowledge of the customers was inadequate. Although there exist extensive data of the customer transactions, *Indiska* lacked the possibility to convert the data into the expected customer experience. The corporate restructuring caused a cost improvements focus, including a review of the store network and closing unprofitable stores. The remaining business is turning a corner, focusing on developing data analyses of the customer behaviour and their purchase patterns and to further develop the platform for e-commerce and to offer a more complete omni-channel experience in the long run. There is a fine balance in the transitional phase where part of the strategy is to retain the current consumers but also rejuvenate the brand. This requires that *Indiska* have information regarding the present-day consumer behaviour but also that of the new customer base by learning what they have bought in the past, what they are currently buying and what they will likely buy onwards.

The purpose of this thesis is to provide means for better understanding the behaviour of the customers and their preferences. Desirably, the purchasing behaviour of the customers will be segmented in different channels, purchase months and product specifications. To be able to retrieve that insight, the aim is partly to build a model that can help predicting the sales quantities of different product classes and identify the most significant factors in the results of the model and partly to create an algorithm that can provide product recommendations in the purchasing process.

To be able to fulfil the purpose and aim of this thesis, the following main research question and sub research questions have been formulated:

- How can a midsize, Swedish, fashion brand utilize historical transaction data to predict future sales quantities of the product classes?
    - What parameters should be used to predict consumer purchase behaviour within the fashion industry?
    - Which of the parameters are most significant when predicting the sales quantities for different product classes?

- How can customer purchase behaviour be used to create more personalised product recommendations?

This thesis will be based on historical transaction data from January 2016 to December 2017, provided by *Indiska*. The size of the data will depend on the availability of the transactions which will include all the product lines provided by Indiska but will be limited to the Nordic sales regions.

Between the author and *Indiska*, a Confidential Disclosure Agreement, CDA has been stipulated. Due to this, all the examined products mentioned in the thesis will be given fictional names.

The developments in the retail industry with more shrewd customers who impose higher demands on personalisation as well as the increased competition are universal challenges for most companies in the business. An application of the model built in this thesis is expected to help companies optimize produced quantities as well as product pricing and usage of the algorithm will likely enable ways to progress within customer knowledge. Furthermore, the findings can probably be used for further extensions of analyses in this field of area.

2. Theory and Method

To answer the research questions, this thesis will use a generalized linear model (GLM). Generalized linear models are flexible generalizations of ordinary general linear regression in such a way that the model allows for response variables with distributions other than the normal distribution. Since the distribution is unknown initially and the purchase behaviour not necessarily follows a normal distribution, the flexibility from a GLM is desirable. A GLM allows a relation to the response variable by a link function, and furthermore, the variance of each measurement can be a function of its predicted value (Olsson, 2002).

In this study, the GLM is used to predict future sales quantities of the product classes and identify which factors that are the most significant for the purchased quantity of the product classes, respectively.

2.1 Generalized Linear Models

Consider a general linear model which describes the observation i of the dependent variable y as a linear function of $(p-1)$ independent variables $x_i, x_2, \dots, x_{p-1}$ as follows

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i(p-1)} + e_i$$

which can be written in matrix form in the following way

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

where $\mathbf{y}$ is a vector that consists of the observations of the dependent variable

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

and $\mathbf{X}$ is a matrix with dimension $n \times p$,

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1(p-1)} \\ 1 & x_{21} & \cdots & x_{2(p-1)} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{n(p-1)} \end{pmatrix}$$

where the column of ones corresponds to the intercepts and the other columns contain the values of the independent variables.

$\boldsymbol{\beta}$ is a vector $\beta = (\beta_0, \dots, \beta_{p-1})^T$ containing the parameters p which are to be estimated.

And finally, $\mathbf{e}$ is the residual vector $\mathbf{e} = (e_1, \dots, e_n)^T$.

A generalized linear model, GLM, is a generalization in the assumption regarding the distribution, which can be extended to include all exponential distributions instead of just the normal distribution. Moreover, a GLM models some link function $g(\cdot)$ instead of directly modeling $\boldsymbol{\mu} = E[\mathbf{y}]$ as a function of the linear predictor (Olsson, 2002).

Consider the equation for a general linear model above in matrix form, the linear predictor can be denoted as

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

and hence, the following model is obtained

$$E[\mathbf{y}] = \boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}).$$

To summarize, a generalized linear model involves a specification of the distribution, specification of the link function $g(\cdot)$ and finally specification of the linear predictor $\mathbf{X}\boldsymbol{\beta}$ (Olsson, 2002).

2.1.1 Exponential family of distributions

The probability mass function of the exponential family of distributions is defined in the following form

$$f(y;\theta,\varphi,) = \exp\left(\frac{y\theta - b(\theta)}{a(\varphi)} + c(y,\varphi)\right)$$

where θ is the natural location parameter, also called a canonical parameter. φ is a dispersion parameter and $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are some functions.

Furthermore, for the exponential distribution, the expected value of a random variable Y is expressed as

$$E[Y] = \mu = b'(\theta)$$

and the variance is denoted as follows

$$\mathrm{Var}(Y) = b''(\theta)a(\varphi) = \mu'(\theta)a(\varphi).$$

The above relations of the mean and the variance are derived from the log-likelihood function of the exponential distribution family and its differentials

$$\frac{\partial l(\theta,\varphi;y)}{\partial\theta} = \frac{\partial}{\partial\theta}\left(\frac{y\theta - b(\theta)}{a(\varphi)} + c(y,\varphi)\right) = \frac{y - b'(\theta)}{a(\varphi)}$$

where $l(\theta, \varphi; y) = log\ f(\theta, \varphi; y)$ and consequently

$$\frac{\partial^2 l(\theta, \varphi; y)}{\partial \theta^2} = -\frac{b''(\theta)}{a(\varphi)}.$$

Furthermore, from the theory of Maximum Likelihood Estimation, MLE, the below expressions are known

$$E\left[\frac{\partial l}{\partial \theta}\right] = 0$$

and

$$E\left[\frac{\partial^2 l}{\partial \theta^2}\right] + E\left[\frac{\partial l}{\partial \theta}\right]^2 = 0.$$

Then the following is obtained for the expected value and the variance respectively, $E[Y] = \mu = b'(\theta)$ and $\text{Var}(Y) = b''(\theta)a(\varphi)$. Hence, it has been shown that the relations above hold.

The variance function is often written as $V(\mu) = b''(\theta)$ which indicates the dependence between the mean and the variance (Olsson, 2002).

2.1.2 The link function

The link function must be a monotone, differentiable function $g(\cdot)$ which is to relate the expected value of the dependent variable Y to the predictors $X_1, \dots, X_p$ (Olsson, 2002). If we assume that we can write the expected value of the response variable as

$$E[y_i] = \mu_i$$

and if now considering the linear predictor $\boldsymbol{\eta}$, we obtain the following definition of the link function

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

From this we get the following expression for the expected value

$$E(y_i) = g^{-1}(\eta_i) = g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta}).$$

The inverse function $g^{-1}(\cdot)$ for a monotone function is defined by the relation $g^{-1}\big(g(\mu)\big) = \mu$. Which type of link function to be used depends on the data and the table below includes some link functions and their inverses that are often used in generalized linear models.

Distribution	Link	Function	Inverse function
Normal	Identity	$\eta_i = \mu_i$	$\mu_i = \eta_i$
Binomial	Logit	$\eta_i = ln\left(\frac{\mu_i}{1-\mu_i}\right)$	$\mu_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$, [0,1]
Poisson	Log	$\eta_i = ln(\mu_i)$	$\mu_i = e^{\eta_i}$
Negative Binomial	Log	$\eta_i = ln(\mu_i)$	$\mu_i = e^{\eta_i}$
Exponential	Inverse	$\eta_i = 1/\mu_i$	$\mu_i = 1/\eta_i$
Gamma	Inverse	$\eta_i = 1/\mu_i$	$\mu_i = 1/\eta_i$

Table 2.1. Link functions and inverse functions for some chosen distributions

For the Poisson distribution, the canonical link, which can be seen as a ”natural” link function, is the log link since it permits the mean to be transformed into a canonical location parameter of the exponential dispersion family member (Olsson, 2002).

2.1.3 Poisson distribution

The Poisson distribution is a discrete probability distribution which defines the probability of an event occurring independently and randomly, hence at a constant rate within a given time interval.

For the Poisson distribution, with mean value μ, the probability mass function is expressed as

$$f(y;\mu) = \frac{e^{-\mu}\mu^{y}}{y!}, y = 0,1,2,\dots$$

The above expression can be written as follows

$$f(y;\mu) = \frac{e^{-\mu}\mu^{y}}{y!} = \exp[ylog(\mu) - \mu - \log(y!)]$$

If the latter part of the expression above is compared with the probability mass function of an exponential distribution it can be noted that $\theta = \log(\mu)$which implies that $\mu = \exp(\theta)$.

Inserting this in the probability mass function of the Poisson distribution yields

$$f(y;\mu) = \exp[y\theta - \exp(\theta) - \log(y!)]$$

which can be seen as a special case of the exponential probability function where we have the following expression for the canonical parameter $\theta = \log(\mu)$ (Olsson, 2002). For the arbitrary functions we get $a(\varphi) = 1$, $b(\theta) = \exp(\theta)$ and $c(y, \varphi) = -\log(y!)$.

The derivatives for the Poisson distribution are then as follows

$$b(\theta) = \exp(\theta)$$
$$b'(\theta) = \exp(\theta) = \mu$$
$$b''^{(\theta)} = \exp(\theta) = \mu.$$

2.1.4 Negative binomial distribution

The negative binomial distribution describes the number of trials needed to get a fixed number of successes.

If one is to study the number of trials until r successes has been recorded for a series of Bernoulli trials, and the probability of success is p, then the distribution of y and the probability mass function of the negative binomial distributions can be expressed as

$$P(y) = \binom{y-1}{r-1} p^r (1-p)^{y-r}, \quad \text{for } y = r, r+1, \ldots$$

The Negative binomial distribution is a generalization of the Geometric distribution in such a way that it is the number of trials needed to get the rth success rather than the number of trials needed to get the first success in repeated independent Bernoulli trials (Olsson, 2002).

It can be shown that the mean of the negative binomial distribution is $E\left(\frac{r}{P}\right)$ and the variance can be expressed as $Var\left(\frac{r(1-p)}{p^2}\right)$.

2.2 Parameter Estimation

The Maximum likelihood method is commonly used when estimating the parameters in a GLM. The method aims to find estimates of the parameters β_i that have the highest probability of corresponding to the true values. Hence, when estimating the parameters, the intention is to search for the values that maximize the log likelihood (Olsson, 2002).

The likelihood function, for a single observation, is defined as

$$f(\boldsymbol{y}; \theta, \varphi) = \exp\left(\frac{y\theta - b(\theta)}{a(\varphi)} + c(y, \varphi)\right).$$

Hence, the log-likelihood function can be written as

$$l = L\big(f(\boldsymbol{y};\theta,\varphi)\big) = \frac{y\theta - b(\theta)}{a(\varphi)} + c(y,\varphi).$$

By differentiating l w.r.t. the elements of β, the regression coefficients, using the chain rule the following is obtain

$$\frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \theta}\frac{\partial \theta}{\partial \mu}\frac{\partial \mu}{\partial \eta}\frac{\partial \eta}{\partial \beta_j}.$$

The earlier shown relations of $b' = \mu$, $b'' = V$ and $\eta = \sum_j x_j\beta_j$ implies that $\frac{\partial \mu}{\partial \theta} = V$ and $\frac{\partial \eta}{\partial \beta_j} = x_j$.

By defining

$$W^{-1} = \left(\frac{\partial \eta}{\partial \mu}\right)^2 V$$

and inserting these expressions in the above differential equation yields

$$\frac{\partial l}{\partial \beta_j} = \frac{(y-\mu)}{a(\varphi)}\frac{1}{V}\frac{\partial \mu}{\partial \eta}x_j = \frac{W}{a(\varphi)}(y-\mu)\frac{\partial \eta}{\partial \mu}x_j.$$

The above expressions are valid for one single observation and the extensive notation of the likelihood for one parameter β_j is given by summing over the observations in the following way

$$\sum_i \frac{W_i(y_i-\mu_i)}{a(\varphi)}\frac{\partial \mu_i}{\partial \eta_i}x_{ij}.$$

This can be solved w.r.t β_j since we have that the μ_i:s are functions of the β_j:s. The asymptotic variances and covariance matrix of the estimated parameters is obtained by the inverse of the Fisher information matrix I_θ

$$\begin{pmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) & Cov(\hat{\beta}_0, \hat{\beta}_{p-1}) \\ Cov(\hat{\beta}_1, \hat{\beta}_0) & Var(\hat{\beta}_1) & \\ Cov(\hat{\beta}_{p-1}, \hat{\beta}_0) & & Var(\hat{\beta}_{p-1}) \end{pmatrix} =$$

$$= -E \begin{pmatrix} \frac{\partial^2 l}{\partial \beta_0^2} & \frac{\partial l}{\partial \beta_0}\frac{\partial l}{\partial \beta_1} & \frac{\partial l}{\partial \beta_0}\frac{\partial l}{\partial \beta_{p-1}} \\ \frac{\partial l}{\partial \beta_1}\frac{\partial l}{\partial \beta_0} & \frac{\partial^2 l}{\partial \beta_1^2} & \\ \frac{\partial l}{\partial \beta_{p-1}}\frac{\partial l}{\partial \beta_0} & & \frac{\partial^2 l}{\partial \beta_{p-1}^2} \end{pmatrix}.$$

2.2.1 Numerical procedures

Maximizing the log-likelihood is done by putting the log-likelihood function in extensive notation equal to zero and solve the equation which is done numerically. An approach that is commonly used is iteratively re-weighted least squares (McCullagh & Nelder, 1989) which is described as follows:

1. Linearize the link function $g(\cdot)$ with for example first order Taylor approximation in the following way $g(y) \approx z,\ z = g(\mu) + (y - \mu)g'(\mu)$

2. If we let $\hat{\eta}_0$ be the current estimate of the linear predictor and we let $\hat{\mu}_0$ be the corresponding fitted value given from the link function $\eta = g(\mu)$ then we can form the adjusted dependent variate as $z_0 = \hat{\eta}_0 + (y - \hat{\mu}_0)\left(\frac{\partial \eta}{\partial \mu}\right)^2$ which is evaluated at $\hat{\mu}_0$

3. Derive the weight matrix W from $W^{-1} = \left(\frac{\partial \eta}{\partial \mu}\right)^2 V_0$, V denotes the variance function.

4. Run a weighted regression of z which is the dependent variable on the predictors $x_1, x_2, \dots, x_p$ using the weights W_0. This yields new, updated values of the estimated parameters $\hat{\beta}_1$, from which one can calculate an updated value of the linear predictor estimate $\hat{\eta}_1$.

5. Repeat step 1-4 until the stop conditions are applied.

2.3 Assessing the fit of the model

When assessing the fit of a statistical model, the model is evaluated on the discrepancy between the observed values and the predicted values, ergo, how well the model results corresponds to the true values. The measures of assessing the fit of the model chosen for this study is a Wald test, Pearson χ^2 statistics, Akaike's Information Criterion, multicollinearity and overdispersion, which will be presented in the following sections.

2.3.1 Variable significance using Wald test

A significance test can be performed to ensure that the chosen variables are significant, and one method of doing this is to perform a Wald test. The outline of the test is to fit a unrestricted model and assess if the results are within the range of sampling variability and agree with the hypothesis (Greene, 2012).

To perform tests on individual coefficients from the model is a simple case of the Wald test.

We can write the null hypothesis as follows

$$H_o: \beta_j = 0, \quad H_1: \beta_j \neq 0,$$

And the following equations yields the test statistics for the null hypothesis

$$Z_j = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

where $se(\hat{\beta}_j) = \frac{1}{\sqrt{I_n(\hat{\beta}_j)}}$, where I_n is the Fisher information, which is the standard error of the maximum likelihood coefficient estimate, $\hat{\beta}_j$, of the j:th independent parameter.

For the significance level α and for the α-quantile of the standard normal distribution depicted as $\lambda_{\alpha/2}$, the null hypothesis can be rejected if $|Z_j| > \lambda_{\alpha/2}$.

2.3.2 Pearson χ^2 statistics

One way of testing the goodness of fit of the model is the generalized Pearson χ^2 test, which has the following definition

$$\chi^2 = \sum_i \frac{(y_i - \hat{\mu})^2}{\hat{V}(\hat{\mu})}$$

where, the estimated variance function is referred as $\hat{V}(\hat{\mu})$ and can be elaborated in the following way $\hat{V}(\hat{\mu}) = b''(b'^{(-1)}(\hat{\mu}))$.

The value of χ^2 is then compared against the critical value of χ_c^2 on a given significance level and with given degrees of freedom. If the critical value is less than the calculated value, then the null hypothesis that the distribution is true can be rejected, hence the model does not fit the data well based on the given level of significance.

When plotting the residuals against the fitted value one should be able to see a "random" pattern with a constant range and zero mean. If the link function used is inaccurate and if non-linear terms are omitted in the linear predictor, this might cause the residual plot to deviate from the desired result (Olsson, 2002).

2.3.3 AIC

A measure of fit that is often used as a means for choosing between models is the following expression

$$D_c = D - \alpha q \varphi$$

where D is the deviance, q is the number of parameters that are included in the model and φ is the dispersion parameter. The general idea is to penalize models that includes extra parameters and by that favour simpler models. When $\alpha = 2$, the above measure is known as the Akaike's Information Criterion, AIC, which can also be written as

$$AIC = -2\ln(L) + 2q$$

where L is referred to the maximum value of the likelihood function for the model and again, q is the number of parameters included in the model. This measure can be used when deciding if or not to include an explanatory variable in a model since the model with the lowest AIC value is preferred over larger values when comparing full and nested model, where one or more variables which are included in the full model have been removed (Montgomery, Peck, & Vining, 2012).

2.3.4 Multicollinearity

In multiple regression, when a number of the explanatory variables are closely dependent on each other, the problem of multicollinearity arises. Multicollinearity can result in larger confidence intervals and larger variances and covariances which might cause skewed results or misinterpreted analyses when predicting the response variable.

One way to detect the presence of multicollinearity is to calculate the VIF, Variance Inflation Factor, which is the ratio of the variance of the values of the coefficients $\hat{\beta}_j$ of the full model divided by the variance of the single value of $\hat{\beta}_j$ if fitted alone.

The VIF is computed in the following way

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

where $R^2_{X_j|X_{-j}}$ is the R^2 from a linear regression of X_j onto all other independent variables.

As a rule of thumb, it is likely that there exists problematic collinearity among the predictor variables within a multiple regression if the value of the VIF exceeds 5 or 10, and such variables should be excluded from the model. In practise though, there will likely be a small amount of multicollinearity between the variables (James, Witten, Hastie, & Tibshirani, 2013).

2.3.5 Overdispersion

The phenomenon of overdispersion occurs when the variance is larger than expected for the choice of distribution. For instance, an attribute of the Poisson distribution is that the variance is equal to the mean which will also be expected when modeling with the Poisson distribution. The foremost cause of presence of overdispersion is lack of homogeneity.

Overdispersion can be detected by measuring the relationship between the residual deviance and the degrees of freedom by dividing the deviance by the *df*. Although, a poor fit of the model can also be a reflection of wrongly made assumptions of the distribution and the link function. Furthermore, the choice of linear predictors might be wrong, causing the bad fit of the model and moreover, the presence of outliers in the data might also cause the bad fit. These reasons should be examined before drawing any conclusions that overdispersion is the reason behind the poor model fit.

Overdisperion is causing underestimations of the standard errors which produces too large test statistics and hence it is easier for the model to produce significant results. If there are signs of overdispersion in a Poisson model, then a Negative binomial distribution can be used as a substitute for the Poisson distribution (Olsson, 2002).

2.4 Conditional probability

When setting up the algorithm the theory of conditional probability will be used in order to set a prioritisation of the suggested product combinations. The combination of which the conditional probability is the highest will be the favoured combination.

Assuming we have two events, A and B, with $P(B) > 0$, then we can define the conditional probability that B occurs given that A occurs as (Gut, 2009)

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

This can be extended to include more than two events and for the specific case of three events we have that the conditional probability of an event is instead conditioned on multiple events as

$$P(C|A \cap B) = \frac{P(A \cap B \cap C)}{P(A \cap B)}.$$

3. Approach

In this thesis, historical transaction data will be used, and the purpose of the study is to provide means for better understanding the behaviour of the customers and their preferences as well as forecasting the purchasing quantities of the product classes. This will be done by modeling the historically purchased quantities of each product class and through analysis of the product combinations in the transactional history. The algorithm will provide suggested product combinations together with the calculated conditional probability of the certain combinations.

The purpose of creating models for each product class is to be able to predict future sales quantities of the classes in favour for the buying department at the company but also for the supply chain in how to distribute the different categories across the channels. Furthermore, the modeling will suggest which variables that will have a significant effect on the outcome of the sales quantity of each product class. This might also come in handy in the processes of designing and pricing the products.

Modeling, instead of analysing the factual sales statistics, allows for more flexibility for the company to do further analyses in the future since it is possible to update the initial transaction data with a more recent one. Hence, this makes the model more reusable and can take new and developed customer behaviours into consideration. Also, modeling makes it easier to consider both existing and prospecting products since if a new product is released, it can be included in the model by categorizing it to its corresponding product class.

3.1 Data processing

This thesis will use a quantitative approach using primary data from the database of the company. The data is procured from the BI analysis tool QlikView. The data spans over the period January 16 to December 2017 and includes transactions from both in-store sales and e-commerce. The transaction data includes transaction ID, store, currency, sales quantity, purchase price, original retail price, currency exchange rate, product ID, purchase date, customer ID if applicable, commodity name and commodity number.

The transaction data is then merged together with the product information data, with product ID as the common denominator, into one complete data file. Apart from product ID, the product data also includes information such as product group and product class.

The following case is presented as a specification of the structure of the products. A specific dress has a certain unique product ID. It is a product which belongs to the product class *Dress* and moreover, *Fashion* is the group to which this dress belongs. Due to the CDA between the author of the thesis and the company, the specific product and its subsequent class and group will not be disclosed in the findings of this thesis in its factual affiliation but instead with a fictional structure consisting of the combination, $n_1.n_2.n_3$, where the first number represents the group of the product, the middle number signifies which class the product belongs to and the last number indicates the specific product. This structure is clarified in *Figure 3.1.* below

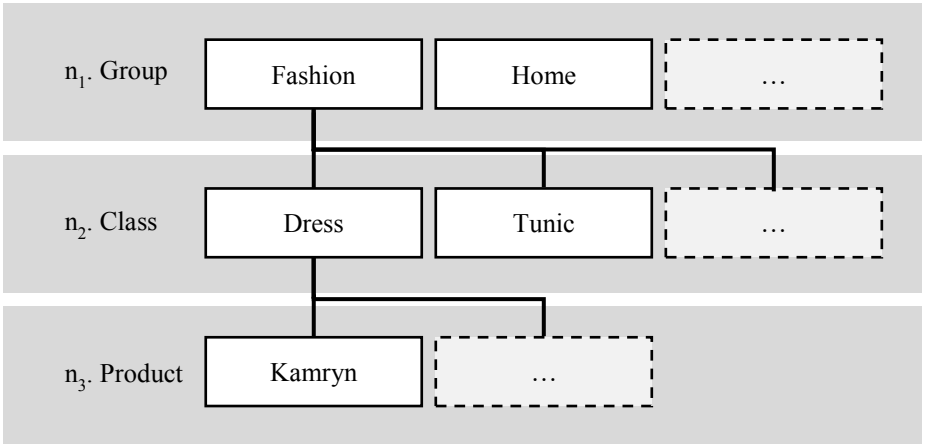

Figure 3.1. Clarification of the product structure

The collected data required refining to only include fully completed transactions and omit the observations with insufficient information regarding product class and original price. Since the data include transactions from not only Sweden but also Norway and Finland the original price of the products needed some configurations to be comparable since the price is given in local currency. For the purchase price of the product the currency exchange rate, which is updated on a monthly basis, is used to yield the equivalent value in SEK for the time of the purchase. For the original price of the product, the value in SEK is applied to all observations.

3.2 Prediction model

The acquired transaction data includes the sold quantity of a certain product class. Hence, this can be referred to as a response variable that counts the occurrences of a specific event, and therefore the data can be considered as count data since the observations of the purchased quantity is restricted to non-negative integer values. Therefore, since modeling the relationship between the response variable as a count, the Poisson distribution have been assumed as the appropriate distribution for the initial trials of modeling the predicted sales quantity (Montgomery et al., 2012).

The collected data needed to be restructured in order to be able to perform a regression analysis. Firstly, the data is distributed into subsets consisting of only the specific product class to be predicted. Secondly, the purchase transactions had to be separated from the refund transactions and the quantity of the latter had to be transformed to absolute values since a Poisson regression is only allowed for positive values. The following notations, x_k and x_r are introduced for the purchased amount and the returned amount, respectively. Assuming that x_k and x_r are independent, then the total predicted sales quantity for the product class will be $x = x_k - x_r$. Thirdly, the independent variables to be included in the model had to be chosen out of the total collected data. Subsequently, each of those explanatory variables had to be clustered together in order for better classifying the characteristics for the variables. Depending on the structure of the variables within each class, different variable groups for different classes had to be constructed for the regression to be stable.

For each of the classes, the actual sales quantity was calculated in order to validate the model by back-tracking the predicted outcome with the historical factual data.

3.2.1 Variable grouping

The regressors are clustered together in order to group the observed values within each variable in such a way that the groups contain sufficiently many observations and are distributed similarly for the regression to converge and to be stable.

The different product groups consist of different products with a wide range of prices, for example some of the classes in the product group fashion, includes much higher price items, for example outerwear, than some of the classes in the product group home, for example candles. As a consequence, the price cluster had to be customized for each class for the cluster to seem sensible. The variables are chosen in a way such that the outcome of the modeling is of importance and of significance for the company operations.

In the table below, a summary of the variables is presented.

Dependent variable	Description
Sales quantity	Sum of sold quantity of the product class during a fixed time period
Independent variables	**Description**
Price	Original sales price in SEK
Month	Month at time of the transaction
Country	Country where the transaction is completed
Cluster	The assortment cluster of which the product belongs to. The cluster also indicates to which stores the product is distributed to
Colour	The specific colour of the product
Channel	The channel of where the transaction is made

Table 3.1. Summary of variables included in the GLM.

The dependent variable can be seen as a response as a count which is a measurement of the number of observations for a specific event. The counts are often summarized in a contingency table where the observations of each variable combination are summed together.

The independent variable price is a quantitative variable in its original form. Although, when grouping the prices together in levels corresponding to prices that are related as higher or lower

in the same product class, the variable is transformed to an ordinal variable since they are not necessarily equally spaced.

Since the variable purchase month is grouped together randomly over the year depending on the number of observations in each month, this variable is transformed from a quantitative variable that only includes the single month to a categorical variable with different levels without any specific ordering.

All of the variables purchase country, cluster, colour and channel are referred to as categorical variables without an intrinsic order of the observed values.

A sample of the variable groups for the two product classes *1.25* and *3.51* are shown respectively in the tables below:

Independent variables	Group 1	Group 2	Group 3
Price	0-350	>350	
Month	1,6-7,9-12	2-5,8	
Country	SE	NO,FI	
Cluster	B	C,F	
Colour	BLACK,WHITE	COLOR,MULTI,METAL	
Channel	E-commerce	Store	

Table 3.2. Variable grouping of the class 1.25.

Independent variables	Group 1	Group 2	Group 3
Price	0-100	>100	
Month	5-8	1-4,9-10	11-12
Country	SE	NO,FI	
Cluster	B	F	
Colour	BLACK,WHITE,CLEAR	COLOR,MULTI,METAL	
Channel	E-commerce	Store	

Table 3.3. Variable grouping of the class 3.51

Due to the CDA, a more intuitively visual description of the distribution of the data cannot be provided.

3.2.2 Prediction of sales quantity

Out of the generated coefficients from the GLM a prediction of the future sales quantity of each product class can be computed. Depending on the characteristics of the product in the product class, different sales amounts will be predicted. For this thesis, the characteristics of the product classes that maximizes the predicted outcome of the sales quantity will be used.

3.3 Product combinations

The product combinations were constructed in such a way that the input product that is chosen by the user of the algorithm was be paired together with other purchased products that occurs on the same transaction ID as the input product. To be able to generate the product combinations, the commodity number of each product was combined with the colour to separate the same type of products with different colours. The reason behind this was that it seems highly realistic that a customer buys the same product, but in different colours. Hence, the combinations will be allowed to include the same product classes but in different colours. Although, the combinations will be restricted and exclude products in the same colour but in different sizes when it comes to products of the group Fashion. This will not be the case for products that belong to the group Home since the same product in different sizes have different commodity numbers. Moreover, this is desirable since when observing the original data set, it appears common that customers buy different sizes of the same Home product in one transaction. The original sales data set was then used in the algorithm to calculate the underlying quantities in order to compute the conditional probabilities of each identified product combination.

For the combinations including three products, a number of conditions had to be formulated in the algorithm for the results to be reliable. Firstly, the counted quantity of transactions including all three products is set to a minimum of 5 counts because of the extensiveness of the product assortment which allows for a wider range of product combinations. Furthermore, the probability of the two input products being purchased together is required to be high enough in order for the probability, when including the combination product being bought, to be significant. Hence, both P(*input product 2* ∩ *input product 1*) and P(*input product 1* ∩ *input product 2*) is required to exceed a certain numeric criteria, which in this thesis was set to 0.05, again due to the breadth of the product assortment.

For simplicity, the following notations, A, B and C, are introduced for the components *input product 1*, *input product 2* and *combination product*, respectively. The probabilities of the components included in the formulas for the conditional probability are defined as stated below.

$$\mathrm{P}(A) = \frac{Number\ of\ unique\ transactions\ including\ input\ product}{Total\ number\ of\ unique\ transactions}$$

$$\mathrm{P}(C \cap A) = \frac{Number\ of\ unique\ transactions\ including\ input\ product\ and\ combining\ product}{Total\ number\ of\ unique\ transactions}$$

$$\mathrm{P}(C \cap B \cap = \frac{Number\ of\ unique\ transactions\ including\ both\ input\ products\ and\ combining\ product}{Total\ number\ of\ unique\ transactions}$$

3.4 Algorithm

The algorithm takes the original data set as input and combines the commodity number with the colour of the specific product which will be used to identify the product or products entered by the user. Firstly, the user of the algorithm enters the input product(s) and the class of the product is identified to enable the possibility to compute the predicted sales quantity. Since the aim is to analyse which type of products are being bought together, the original data set is then reduced by excluding the returned products. Furthermore, since some products are basic products which exist in the assortment throughout a whole year, the possibility of recommending typically seasonal products that are bought together which such running product styles had to be restricted. This is done by letting the user decide if all data is to be used or if to limit the data set to include not only the current month at the time of algorithm usage but also the two enclosing months because of an assumption of seasonal resemblances.

Secondly, all the transactions made that includes the input product(s) are identified. Subsequently, the number of transactions made with other products which are bought together with the input product(s) are aggregated. Thirdly, the algorithm counts the total number of made transactions and the number of transactions made of the input product(s). Then the probability of the input product(s) being bought alone and together with each combining product is calculated, respectively. Finally, the conditional probability of each of the combining products are calculated and the result is printed.

If two products are entered by the user as input products, the algorithm first calculates the conditional probability of only the two input products being purchased together and then checks for the number of transactions made of the input products together with each combining product. If the requirements are fulfilled, the algorithm provides a list of the input products together with the top ten, or if less than so, the available number of suggested product combinations along with the related conditional probability.

Although, since the physical stores charge for bags, they end up as a line on the receipt. Hence, in the printed result, combining products including bags are removed since it is not of interest. The same is applied for combinations including shipping costs from e-commerce transactions.

The resulting combinations are inter-product combinations, although the algorithm does not account for product combinations that includes different sizes of the same products unless the commodity number separates products of different sizes.

The specific requirements applied in the algorithm are specified in the table below.

Applied when:	Usability	Requirement value
Counting quantity of transactions made including all three products	In order for the mathematical result to be reliable when presenting the conditional probability of three products	5
When computing the conditional probability of the two input products being bought together, respectively	Have to be high enough to be able to calculate the conditional probability and securing that the probability of the third product being bought is reliable	0.05

Table 3.3. Applied requirements when building the algorithm

4. Prediction model

This section will present the result of each of the steps involved when building the generalized linear models for each of the product classes.

4.1 Multicollinearity

For each of the product classes, a generalized linear model has been performed from which it is possible to compute the variance inflation factor, VIF, to get an indication of the existence of multicollinearity among the regressors.

For the product classes *3.50, 3.85* and *2.40,* the results are presented in Table *4.1.* below for each of the included variables in the regression.

Product class	3.50		3.85		2.40	
Variables	VIF x_k	VIF x_r	VIF x_k	VIF x_r	VIF x_k	VIF x_r
Price	1.277505	1.267734	1.000683	1.005519	1.001399	1.011192
Month	1.005459	1.044083	1.000525	1.031591	1.011159	1.068831
Country	1.573982	1.435308	1.001739	1.060412	1.006108	1.045646
Cluster	1.007153	1.058774	1.001058	1.035924	1.004830	1.052696
2.Colour	1.270949	1.056019	1.000522	1.015384	1.004117	1.004135
Channel	1.023286	1.113116	1.001700	1.064823	1.015572	1.119837

Table 4.1. Variance inflation factors for an extraction of the product classes

As is shown, there is not one of the variables in the table above, for each of the classes, that exceeds the critical level of 5 for any of the two data sets x_k or x_r. In fact, for all variables, the variance inflation factor is less than two and therefore, the conclusion is drawn that there is no problem with multicollinearity among the regressor in the data.

A similar tendency is apparent when observing the variance inflation factors for the other product classes as well.

4.2 Selection of variables

When running a GLM, a summary output of the overall performance of the model, including variable significance, AIC values as well as the estimated coefficients are presented. The results from an extraction of the modelled product classes will now be presented.

The summary output from the regression models for x_k and x_r for class *3.70* using Poisson distribution and are presented in Tables *5.2.* and *5.3.*, respectively. The group (Intercept) corresponds to the first level of every variable and is set as a reference group, in this case for the class *3.70*, containing prices between 0-180, months 1,11-12, cluster B, country SE, the colours BLACK, WHITE, CLEAR and the channel E-com. For every coefficient, the significance is computed using the Wald where '***','**','*' is shown if the p-value is less than or equal to 0.001, 0.01, 0.05, respectively.

Poisson regression, x_k

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	6.520594	0.010208	638.79	<2e-16	***
price.group>180	0.199406	0.003742	53.28	<2e-16	***
month.group6-7,10	-0.587985	0.004433	-132.65	<2e-16	***
month.group2-5,8-9	-0.746235	0.004669	-159.81	<2e-16	***
cluster.groupC,F	-3.364808	0.011163	-301.42	<2e-16	***
country.groupNO,FI	-1.322345	0.004563	-289.79	<2e-16	***
color.groupCOLOR,MULTI,METAL	0.600449	0.003886	154.53	<2e-16	***
channelStore	3.205377	0.009626	333.00	<2e-16	***

Null deviance:	768584	on	87	degrees of freedom
Residual deviance:	22894	on	80	degrees of freedom
AIC:	23574			

*Table 4.2. Summary output from Poisson regression of x_k of class 3.70 including Wald test with significance levels 0.001 '***', 0.01 '**', 0.05 '* '*

Poisson regression, x_r

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	3.07153	0.05401	56.87	<2e-16	***
price.group>180	0.51918	0.01972	26.32	<2e-16	***
month.group6-7,10	-0.55783	0.02359	-23.64	<2e-16	***
month.group2-5,8-9	-0.38187	0.02239	-17.05	<2e-16	***
cluster.groupC&F	-2.80070	0.04452	-62.91	<2e-16	***
country.groupNOFI	-1.89756	0.02831	-67.03	<2e-16	***
color.groupCOLOR&METAL&MULTI	0.48250	0.01953	24.70	<2e-16	***
channelStore	3.25479	0.05081	64.06	<2e-16	***

Null deviance:	26576	on	71	degrees of freedom
Residual deviance:	824	on	64	degrees of freedom
AIC:	1208			

*Table 4.3. Summary output from Poisson regression of x_r of class 3.70 including Wald test with significance levels 0.001 '***', 0.01 '**', 0.05 '* '*

Judging from the summary outputs presented in Tables *5.2.* and *5.3.*, the standard errors are fairly small while the significance is very high for all variables for both x_k and x_r. Furthermore, the ratio between the residual deviance and the degrees of freedom are significantly larger than one, which indicates that there might be overdispersion in the data. Moreover, the value of AIC for the model of x_k is rather large, and when removing the explanatory variables one by one, the value only becomes larger, suggesting that all the chosen linear predictors should be included in the model.

The result from the models for x_k and x_r for class *2.44* using Poisson distribution and are presented in Tables *5.4.* and *5.5.*, respectively. For this product class, the group (Intercept) is instead a reference group consisting of prices between 0-100, purchase months 1,7,12, cluster B, country SE, the colour METAL and the channel E-com.

Poisson regression, x_k

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	6.712725	0.009793	685.50	<2e-16	***
price.group>100	0.099784	0.003449	28.93	<2e-16	***
month.group2-6,8-11	0.550169	0.003575	153.89	<2e-16	***
cluster.groupC&F	-3.438420	0.010139	-339.13	<2e-16	***
country.groupNOFI	-1.056165	0.003935	-268.38	<2e-16	***
color.groupBLACK&WHITE&CLEAR	-1.767819	0.006106	-289.50	<2e-16	***
color.groupCOLOR&MULTI	-0.444288	0.003706	-119.90	<2e-16	***
channelStore	3.306963	0.009329	354.48	<2e-16	***

Null deviance:	927116	on	86	degrees of freedom
Residual deviance:	19891	on	79	degrees of freedom
AIC:	20570			

*Table 4.4. Summary output from Poisson regression of x_k of class 2.44 including Wald test with significance levels 0.001 '***', 0.01 '**', 0.05 '* '*

Poisson regression, x_r

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	3.37405	0.05103	66.12	<2e-16	***
price.group>100	0.77418	0.02629	29.45	<2e-16	***
month.group2-6,8-11	0.55916	0.02534	22.07	<2e-16	***
cluster.groupC&F	-2.99016	0.05930	-50.43	<2e-16	***
country.groupNOFI	-1.83075	0.03542	-51.69	<2e-16	***
color.groupBLACK&WHITE&CLEAR	-1.61621	0.03981	-40.60	<2e-16	***
color.groupCOLOR&MULTI	-0.59558	0.02703	-22.03	<2e-16	***
channelStore	2.43823	0.04515	54.01	<2e-16	***

Null deviance:	17638	on	75	degrees of freedom
Residual deviance:	544	on	68	degrees of freedom
AIC:	912			

*Table 4.6. Summary output from Poisson regression of x_r of class 2.44 including Wald test with significance levels 0.001 '***', 0.01 '**', 0.05 '* '*

When examining the summary outputs for product class *2.44*, yet again, small standard errors are apparent and high significance occurs for all variables for both x_k and x_r. The result show also for this product class, that there might be a problem with overdispersion in the data.

The above reasoning appeared to be the case for all the product classes, which indicates that another distribution than the Poisson distribution might be better suited to model the data. An alternative distribution that can be used when modeling overdispersed data is the Negative binomial distribution.

Therefore, new regressions were performed for all the classes, but instead with a Negative binomial distribution applied in the models, of which the results for the two classes presented above are displayed below.

Negative binomial regression, x_k

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	6.1318	0.2128	28.812	<2e-16	***
price.group>180	1.1598	0.1508	7.689	1.48e-14	***
month.group6-7,10	-0.5136	0.1853	-2.772	0.00557	**
month.group2-5,8-9	-0.7533	0.1856	-4.059	4.93e-05	***
cluster.groupC,F	-3.2479	0.1517	-21.412	<2e-16	***
country.groupNO,FI	-1.6561	0.1487	-11.138	<2e-16	***
color.groupCOLOR,MULTI,METAL	0.6665	0.1485	4.487	7.23e-06	***
channelStore	3.0701	0.1498	20.493	<2e-16	***

Null deviance:	891	on	87	degrees of freedom
Residual deviance:	95	on	80	degrees of freedom
AIC:	1225			

*Table 4.7. Summary output from Negative binomial regression of x_k of class 3.70 including Wald test with significance levels 0.001 '***', 0.01 '**', 0.05 '* '*

Negative binomial regression, x_r

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	2.7949	0.1614	17.319	<2e-16	***
price.group>180	0.8196	0.1180	6.943	3.83e-12	***
month.group6-7,10	-0.6498	0.1391	-4.672	2.99e-06	***
month.group2-5,8-9	-0.4284	0.1381	-3.102	0.00193	**
cluster.groupC&F	-2.8040	0.1342	-20.901	<2e-16	***
country.groupNOFI	-1.8461	0.1195	-15.446	<2e-16	***
color.group1COLOR&METAL&MULTI	0.6654	0.1140	5.836	5.34e-09	***
channelStore	3.2577	0.1308	24.901	<2e-16	***

Null deviance:	1327	on	71	degrees of freedom
Residual deviance:	78	on	64	degrees of freedom
AIC:	605			

*Table 4.8. Summary output from Negative binomial regression of x_r of class 3.70 including Wald test with significance level 0.001 '***', 0.01 '**', 0.05 '* '*

When observing the new summary outputs of x_k and x_r in Tables *5.7.* and *5.8.*, with a Negative binomial distribution instead, it is evident that the standard errors are higher while the significance of the variables are still very high, in fact all variables are significant on a level of 0.05 or less.

Moreover, the value of AIC for x_k is considerably lower for the model with Negative binomial distribution than for the Poisson regression, suggesting that the former model is the desirable one out of the two.

Furthermore, the relationship of the residual deviance and the degrees of freedom are now much closer to one than in the previous model which indicates that overdispersed data was in fact present and have now been handled correctly.

Below, the results from the regression using Negative binomial distribution for the product class *2.44* are shown for both x_k and x_r.

Negative binomial regression, x_k

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	6.2457	0.1624	38.456	<2e-16	***
price.group>100	0.6933	0.1219	5.686	1.30e-08	***
month.group2-6,8-11	0.8780	0.1196	7.341	2.12e-13	***
cluster.groupC&F	-3.4346	0.1244	-27.606	<2e-16	***
country.groupNOFI	-1.3266	0.1197	-11.079	<2e-16	***
color.groupBLACK&WHITE&CLEAR	-1.5947	0.1540	-10.355	<2e-16	***
color.groupCOLOR&MULTI	-0.6770	0.1399	-4.839	1.31e-06	***
channelStore	3.4643	0.1203	28.791	<2e-16	***

Null deviance:	1456	on	86	degrees of freedom
Residual deviance:	93	on	79	degrees of freedom
AIC:	1182			

*Table 4.9. Summary output from Negative binomial regression of x_k of class 2.44 including Wald test with significance levels 0.001 '***', 0.01 '**', 0.05 '* '*

Negative binomial regression, x_r

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	3.4064	0.1676	20.330	<2e-16	***
price.group>100	0.9287	0.1269	7.318	2.52e-13	***
month.group2-6,8-11	0.6060	0.1229	4.933	8.12e-07	***
cluster.groupC&F	-2.9603	0.1501	-19.727	<2e-16	***
country.groupNOFI	-1.6723	0.1275	-13.116	<2e-16	***
color.groupBLACK&WHITE&CLEAR	-1.6657	0.1581	-10.534	<2e-16	***
color.groupCOLOR&MULTI	-0.6258	0.1422	-4.402	1.07e-05	***
channelStore	2.1818	0.1325	16.464	<2e-16	***

Null deviance:	1104	on	75	degrees of freedom
Residual deviance:	78	on	68	degrees of freedom
AIC:	579			

*Table 4.10. Summary output from Negative binomial regression of x_r of class 2.44 including Wald test with significance levels 0.001 '***', 0.01 '**', 0.05 '* '*

Again, the conclusions can be drawn that the variable significance is high, although the standard errors are higher than for the Poisson regression and the AIC value of 1182 is substantially lower than the value 20570 of the corresponding model using a Poisson distribution. Moreover, the relationship is again closer to one between the residual deviance and the degrees of freedom.

In fact, the similar result was evident when examining the models of the remaining product classes. Based on the differences in the results from the models using Poisson distribution compared to the models with a Negative binomial distribution, the conclusion is drawn that the Negative binomial distribution is suitable and that the model fits the data well.

In order to evaluate if all the chosen variables should be included in the Negative binomial models, a comparison was made between the AIC values of the model containing all the chosen variables and a model where each one of the variables is omitted. The result is presented in Table *4.11.* below for each of the product classes *2.47* and *1.24.*

Product	2.47		1.24	
	AIC x_k	AIC x_r	AIC x_k	AIC x_r
Full model	1045	576	1081	747
Without price	1056	594	1162	818
Without month	1128	648	1172	807
Without country	1099	638	1157	826
Without cluster	1108	610	1239	870
Without colour	1046	578	1097	769
Without channel	1137	634	1200	821

Table 4.11. AIC values for the classes 2.47 and 1.24

For each of the classes, the lowest AIC value is obtained from the full model, suggesting that all the variables should be included when modeling the data. Actually, this was the case for all the evaluated classes.

4.3 Assessing the fit of the model

4.3.1 Pearson χ^2 statistics

One of the measure chosen for this thesis to evaluate the goodness of fit of the model is the Pearson χ^2 statistics, which is calculated for both the model with Poisson distribution and for the one with Negative binomial distribution, of each product class.

The resulting values for the Pearson χ^2 statistics is presented in the Table *5.12.*

Product class		3.81		1.19	
Distribution		χ^2	χ_c^2	χ^2	χ_c^2
Poisson regression	x_k	49299.47	100.75	31955.37	72.15
Negative binomial regression		75.73	100.75	46.67	72.15
Poisson regression	x_r	883.60	79.08	3984.88	67.50
Negative binomial regression		68.58	79.08	49.87	67.50

Table 4.12. Pearson χ^2 statistics of the two different models for the data sets x_k and x_r

For both of the presented product classes, the computed value, χ^2, is extensively larger than the critical value when the Poisson distribution is applied while the computed value is within bounds of the critical value for the Negative binomial distribution. The findings of these tests are further proof that the Negative binomial distribution is the more appropriate distribution to apply in the model out of the two distributions.

The similar findings were evident across all analysed product classes.

4.3.2 Residual plots

Residuals plots are another useful tool in assessing the fit of the model since they can disclose patterns which are unexplained in the data. For a GLM the plotted residuals are the deviance residuals.

Poisson distribution

The initial GLM with the Poisson distribution of the data set x_k and x_r respectively of the class *1.14* resulted in the plots below.

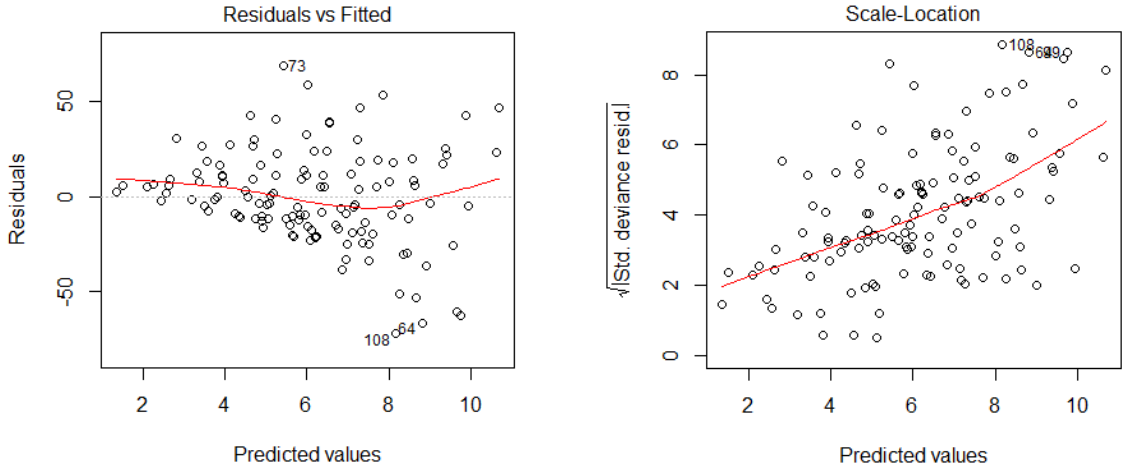

Figure 4.1. Left: Residuals vs Fitted for GLM with Poisson distribution for x_k of class 1.14
Right: Scale-Location for GLM with Poisson distribution for x_k of class 1.14

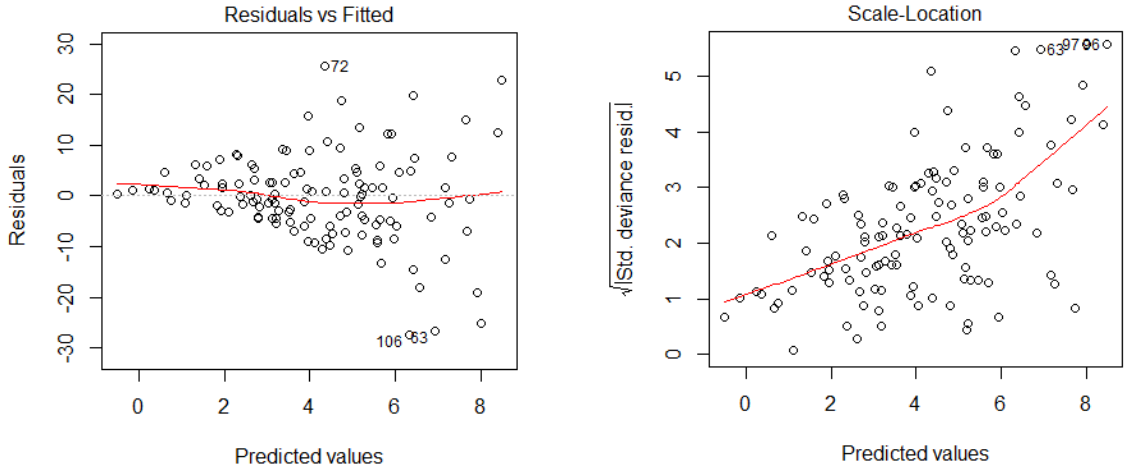

Figure 4.2. Left: Residuals vs Fitted for GLM with Poisson distribution for x_r of class 1.14
Right: Scale-Location for GLM with Poisson distribution for x_r of class 1.14

Both x_k, displayed in *Figure 5.1.* and x_r, in *Figure 5.2.*, show similar tendencies. In the left figures, which represent the relationship between the predicted values and the residuals, it is apparent that the residuals seem to deviate gradually from the constant dashed line around zero as the predicted values increases. Hence there is no random pattern of the observations in the plots which suggests that there might exist a non-linear relationship between the response variable and the explanatory variables that is not captured correctly by the model and instead was left out in the residuals.

The relationship between the predicted values and the square root of the standardized deviance residuals is expressed in the scale-location plots to the right. For the Poisson distribution, the resulting plots for x_k and x_r, are presented to the right in *Figures 5.1.* and *5.2.*, respectively. In the plots, there are an apparent curve upwards indicating that the residuals are not equally spread, suggesting inequality in variance.

The corresponding plots but instead for the product class *2.41* are presented below, with the similar findings as for the conducted analysis of the product class *1.14* using Poisson distribution.

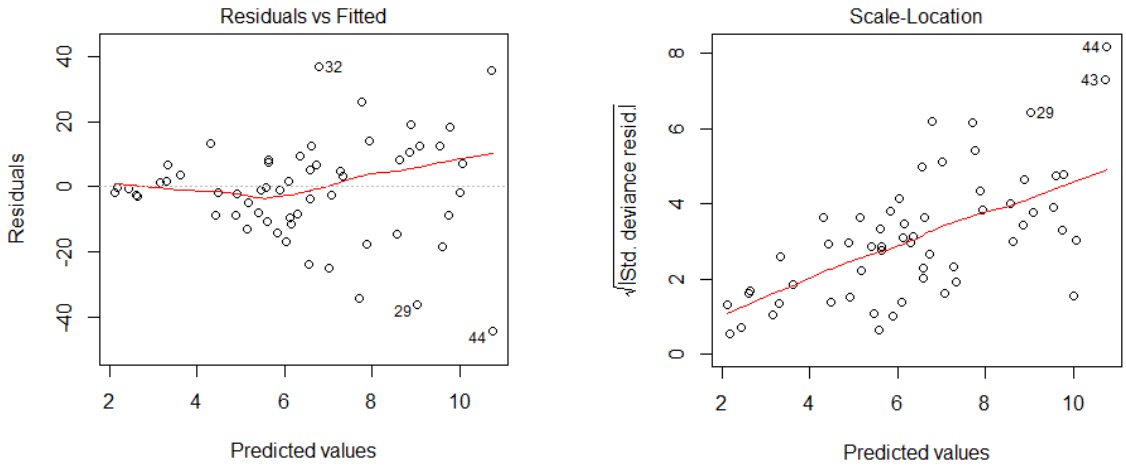

Figure 4.3. *Left: Residuals vs Fitted for GLM with Poisson distribution for x_k of class 2.41*
Right: Scale-Location for GLM with Poisson distribution for x_k of class 2.41

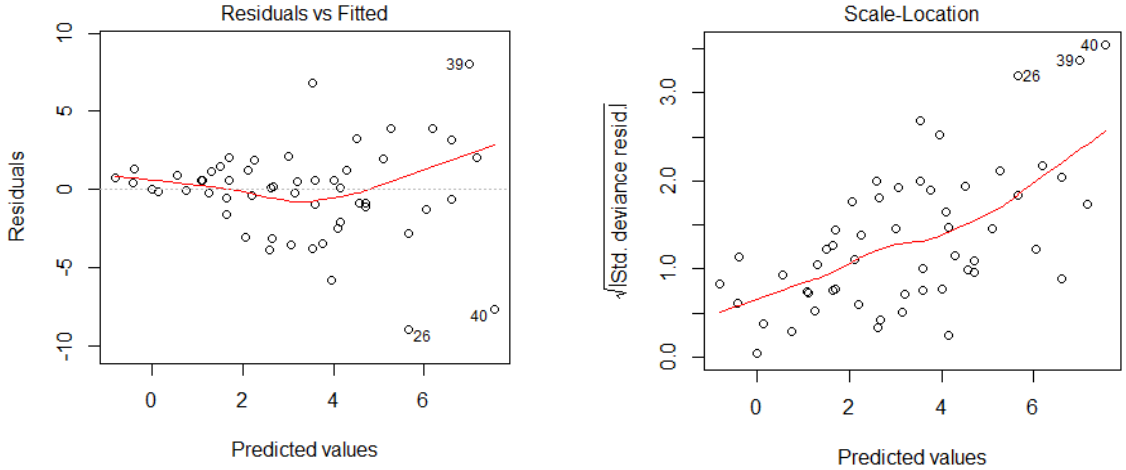

Figure 4.4. *Left: Residuals vs Fitted for GLM with Poisson distribution for x_r of class 2.41*
Right: Scale-Location for GLM with Poisson distribution for x_r of class 2.41

Negative binomial distribution

The second GLM with the Negative binomial distribution instead produced the following plots.

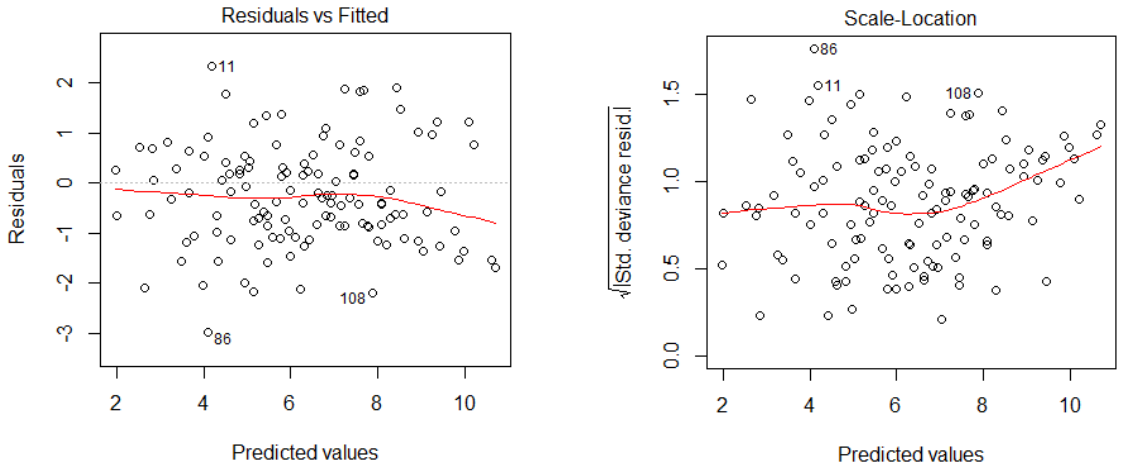

Figure 4.5. *Left: Residuals vs Fitted for GLM with Negative binomial distribution for x_k of class 1.14*
Right: Scale-Location for GLM with Negative binomial distribution for x_k of class 1.14

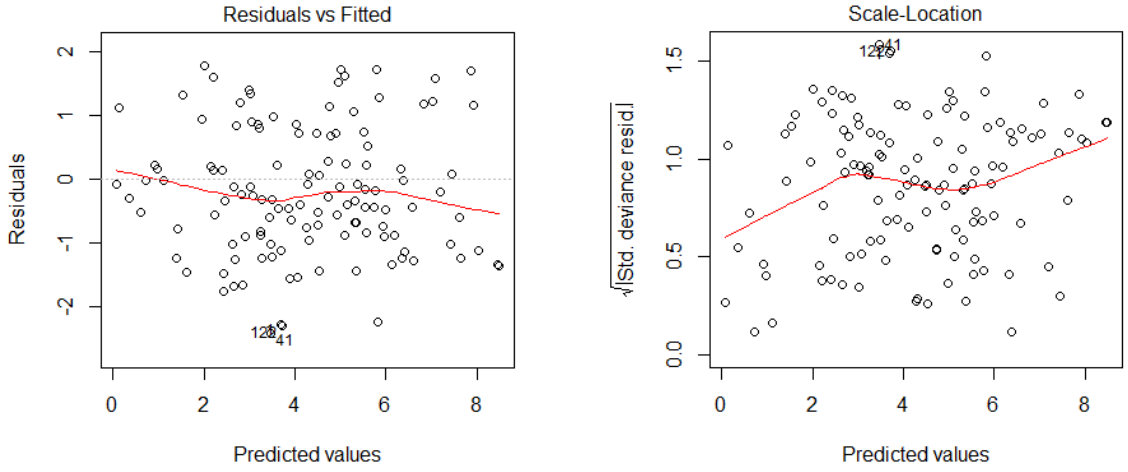

Figure 4.6. *Left: Residuals vs Fitted for GLM with Negative binomial distribution for x_r of class 1.14*
Right: Scale-Location for GLM with Negative binomial distribution for x_r of class 1.14

By examining the left plots in *Figures 5.5.* and *5.6.* for the GLM with the Negative binomial distribution, it is obvious that there is no apparent pattern. Instead, the residuals are equally spread out around the horizontal dashed line which is the desired result since it indicates that the relationship between the response variable and the explanatory variables are captured by the model.

Moreover, the Negative binomial regression represented to the right in *Figures 5.5.* and *5.6.*, shows a lot smaller curvature from the horizontal line, suggesting homogeneity of variance. Especially notice the difference in scale on the y-axis between the plots resulted from Poisson regression compared to the Negative binomial regression since it at first glance can appear that the curvature of the red line resembles one another in the Residuals vs Fitted plots, although the distances in the Poisson plot are greater.

The corresponding plots but instead for the product class *2.41* are presented below, with the similar findings as for the conducted analysis of the product class *1.14* using Poisson distribution.

The residual plots generated from the model using Negative binomial model, for the class *2.41*, are shown below. As for the product class *1.14*, the findings from the Negative binomial regression is favourable compared to the Poisson regression.

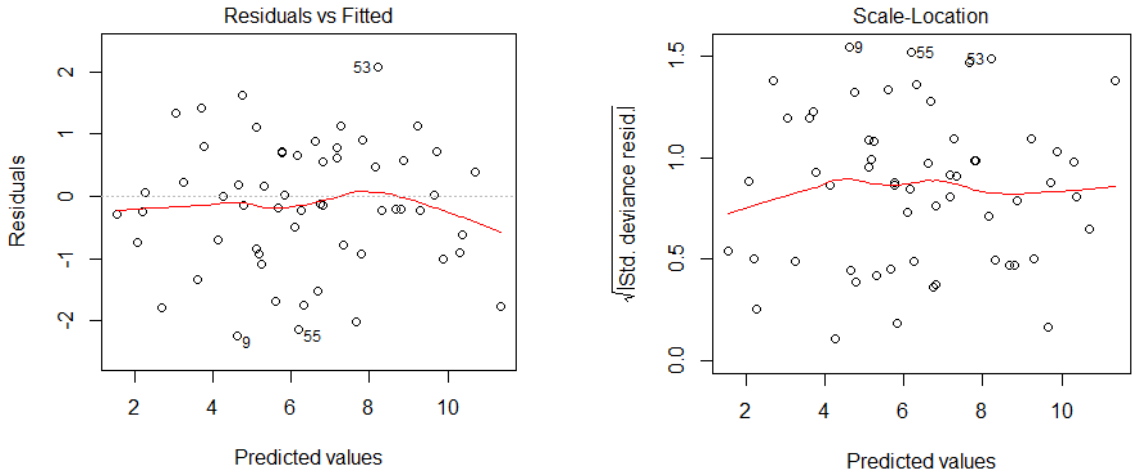

Figure 5.7. *Left: Residuals vs Fitted for GLM with Negative binomial distribution for x_k of class 2.41*
Right: Scale-Location for GLM with Negative binomial distribution for x_k of class 2.41

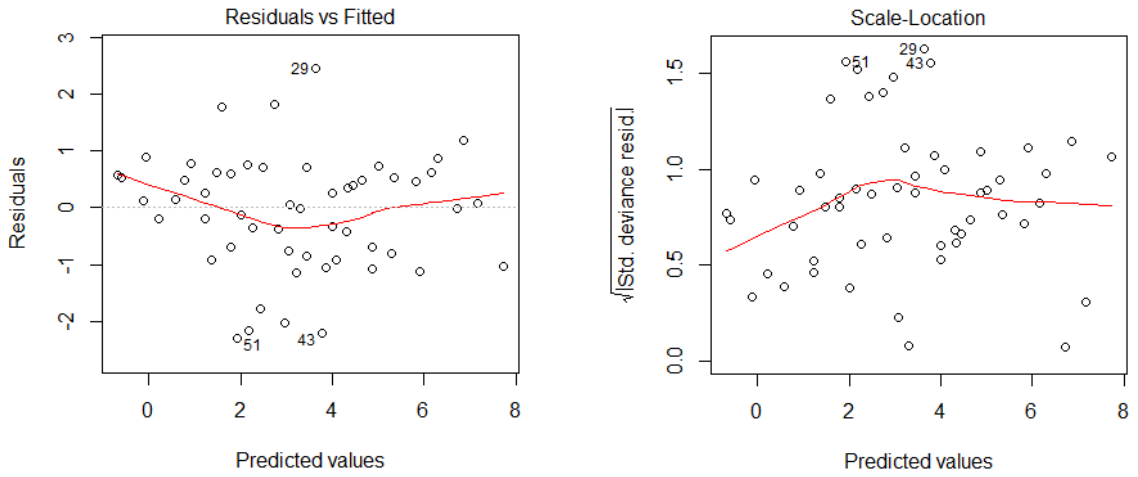

Figure 5.8. *Left: Residuals vs Fitted for GLM with Negative binomial distribution for x_r of class 2.41*
Right: Scale-Location for GLM with Negative binomial distribution for x_r of class 2.41

The similar procedure is applied when examining all the plots resulted from the Poisson regression and the GLM with Negative binomial distribution for each product class. Since the results show the similar tendencies for every product class, there is further evidence that the Negative binomial distribution is preferable to the Poisson distribution.

In the plots, there are certain observations that are situated outside the clusters which could be marked as potential outliers. Although, once these points were deleted from the data set there were no obvious changes in the plots and therefore the decision was made to keep those points in the data set to not diminish the amount of data.

5. Results

This section will illustrate the results from the Generalized linear model and present the resulting product combinations suggested from the algorithm and the corresponding conditional probabilities.

5.1 Generalized Linear Model

5.1.1 Model Prediction

The concluding model for the predicted sales quantity, x_k, for all the product classes is presented below

$$\log(x_k) = \beta_{0_{x_k}} + \sum_{i=1}^{n} x_i \beta_i$$

and the predicted returned quantity, x_r, for all the products classes are as follows

$$\log(x_r) = \beta_{0_{x_r}} + \sum_{j=1}^{n} x_j \beta_j.$$

Since the assumption is made that x_k and x_r are independent, the finalized total sales quantity, x, can be expressed as

$$\log(x) = \log\left(\frac{x_k}{x_r}\right) = \beta_{0_{x_k}} - \beta_{0_{x_r}} + \sum_{i=1}^{n} x_i \beta_i - \sum_{j=1}^{n} x_j \beta_j$$

where n is the number of parameters, β_i and β_j, included in x_k and x_r, respectively, for each product class.

For the class *1.14*, the final model can be elaborated into the following expression for x_k.

$$\log(x_k) = \beta_{0_{x_k}} + \begin{cases} 0 & \text{if original price 0-400} \\ \beta_1 & \text{if original price >400} \end{cases} + \begin{cases} 0 & \text{if month 2-4} \\ \beta_2 & \text{if month 11-12} \\ \beta_3 & \text{if month } 1{,}6-7 \\ \beta_4 & \text{if month } 5{,}8-10 \end{cases}$$

$$+ \begin{cases} 0 & \text{if cluster B} \\ \beta_5 & \text{if cluster C, F} \end{cases} + \begin{cases} 0 & \text{if country SE} \\ \beta_6 & \text{if country NO,FI} \end{cases}$$

$$+ \begin{cases} 0 & \text{if colour BLACK, WHITE} \\ \beta_7 & \text{if colour COLOR, MULTI} \end{cases} + \begin{cases} 0 & \text{if channel E-commerce} \\ \beta_8 & \text{if channel Store} \end{cases}$$

and correspondently for $\log(x_r)$.

The parameter values for x_k and x_r for class *1.14* are summarized in the next table.

Reference group	Parameter	Parameter value	
		x_k	x_r
Intercept	β_0	5.2503	3.4127
Original price >400	β_1	0.3772	0.3956
Month 11-12	β_2	0.4370	0.6761
Month 1,6-7	β_3	1.3322	1.2036
Month 5,8-10	β_4	1.4182	1.2463
Cluster C,F	β_5	-2.3572	-2.0487
Country NO,FI	β_6	-1.6991	-2.1531
Colour COLOR,MULTI	β_7	0.9959	0.9264
Channel Store	β_8	2.6179	2.3623

Table 5.1. Parameter values for class 1.14

By observing the sign of the parameter values, it is apparent that the cluster C,F as well as country NO,FI have a negative effect on both the predicted sales quantity and on the predicted returned quantity. When interpreting the parameter values, it is important to take into account that there are fewer products categorized in the cluster C,F compared to the cluster B, which might be an explanation for the outcome. Furthermore, since Sweden is the biggest market, this might also be an explanation for the fact that the group country NO,FI have a negative parameter value.

An original price above 400 SEK increases the predicted sales quantity by a factor of 1.46 compared to the intercept values, ceteris paribus. This might be explained by most of the products within the product class having a higher price or by the fact that the customers are willing, or even want, to pay a higher price for the items of this product class, perhaps for perceived quality reasons.

The intercept value includes the months 2-4, and compared to this, the predicted sales amount is increased for every one of the other month groupings. It is evident that this product class has a high predicted sales quantity during the months May and August to October, also the months January and June to July have a higher predicted quantity than for the intercept. The latter months, are often periods with high reductions in the stores and online which might indicate that the product class have a high attraction when on sale.

The colour group COLOR,MULTI, have a positive factor of 2.69 compared with the intercept, ceteris paribus, indicating that colourful products of this category are highly preferred by the customers.

The predicted sales quantity depends highly on the channel of which the purchase is made. The predicted outcome is considerably higher, ceteris paribus, for the group Store than for the group E-com which is included in the reference group. An explanation for this it that the turnover share for the channel E-com is relatively small in comparison.

For the class *3.51*, the final model for x_k can be elaborated into the following expression

$$\log(x_k) = \beta_{0_{x_k}} + \begin{cases} 0 & \text{if original price 0-100} \\ \beta_1 & \text{if original price >100} \end{cases} + \begin{cases} 0 & \text{if month 5-8} \\ \beta_2 & \text{if month 1-4,9-10} \\ \beta_3 & \text{if month 11-12} \end{cases}$$

$$+ \begin{cases} 0 & \text{if cluster B} \\ \beta_4 & \text{if cluster F} \end{cases} + \begin{cases} 0 & \text{if country SE} \\ \beta_5 & \text{if country NO,FI} \end{cases}$$

$$+ \begin{cases} 0 & \text{if colour BLACK, WHITE, CLEAR} \\ \beta_7 & \text{if colour COLOR, MULTI, METAL} \end{cases} + \begin{cases} 0 & \text{if channel E-commerce} \\ \beta_8 & \text{if channel Store} \end{cases}$$

where again, the similar applies for x_r.

For the class *3.51*, the parameter values for x_k and x_r are summarized in the next table.

Reference group	Parameter	Parameter value	
		x_k	x_r
Intercept	β_0	6.3647	2.46655
Original price >100	β_1	-1.2282	-0.66742
Month 1-4,9-10	β_2	0.4515	-0.08147
Month 11-12	β_3	0.5944	-3.37089
Cluster F	β_5	-3.8909	-1.53802
Country NO,FI	β_6	-1.5441	-0.37414
Colour COLOR,MULTI,METAL	β_7	0.4137	0.83742
Channel Store	β_8	3.7081	2.77497

Table 5.2. Parameter values for class 3.51

On the contrary of product class *1.14*, the class *3.51* have a sold quantity enhancing sales price below or equal to 100 SEK. Again, an explanation for this could be that most of the products within the product class have a lower price or that the customers expect a lower price for the products of this class. Furthermore, also for this product class, the cluster group F and the country group NO,FI affects the predicted sales volume negatively. Although, this could again be explained by the fact that there are fewer products categorized in the cluster F compared to the cluster B and since the major market is Sweden.

Observing the month groups, it becomes evident that May to August, which is included intercept as the reference group, have the lowest predicted sales quantity. If instead examining the returned quantity, the same months also have the highest predicted returned quantity. Looking at parameter value for x_r for the month group 11-12, it is much lower than compared to the base line but also to the month group 1-4,9-10. An explanation for this outcome may be that in November and December, it is common to shop for Christmas gifts and that during that period there is an extended return policy until early January. Further evidence for this reasoning is that the predicted sales quantity is the highest for this period, with all other variables unchanged.

Comparing the two presented classes above, based on their intercept values for both x_k and x_r, show that the predicted sales quantity is higher for the product class *3.51* while also the predicted returned quantity is lower than for the product class *1.14*. Hence, in total, the predicted sales volume of the class *3.51* will be higher when referring to the reference groups, respectively.

5.1.2 Outcome of predicted sales quantities

The above model is then used to calculate the predicted total sales quantity for different characteristics within every product classes. Since the data is based on two years and since the month group in some cases contain several months, an average per month is achieved by dividing the outcome in two and the number of month included in the month group. Because of the similarity of the distributed quantity in each month group, i.e. that the including months in each group are as homogenous in sales quantity as possible, the generated average is assumed to be a good valuation.

For each month, the predicted total sales quantity of each product class with the characteristics that are the most sales enhancing for each class has been calculated for each month and are presented in the *Table 5.3.* below. For example, for the class *1.14* presented above, the sales enhancing factors are original price>400, Cluster B, Country SE, Colour COLOR, MULTI and channel Store and are applied for each month

Class	1	2	3	4	5	6	7	8	9	10	11	12
3.50	7903	7903	7903	7903	7903	7903	7903	7903	7903	7903	7903	22231
1.24	9253	3013	3013	3013	9253	9253	9253	9253	9253	9253	9253	9253
2.41	5273	2942	5273	5273	2942	2942	5273	2942	5273	5273	2942	5273
3.81	16739	16739	16739	16739	16739	16739	16739	16739	16739	16739	16739	38694
1.19	12006	6397	6397	6397	12006	12006	12006	12006	12006	12006	12006	12006

Table 5.3. Predicted sales quantities per month for the product classes 3.50, 1.24, 2.41, 3.81 and 1.19

As is presented in table above, the class *3.50* have a predicted sales quantity in total that amounts to 22231 for December and 7903 for the remaining months. While for the class *1.24*, the predicted quantity instead equals to 9253 for the each of the months 1,5-12 and 3013 for the months February to April.

For the class *3.81* is it clear that the predicted total sales quantity really high during the whole year while it increases significantly during December, possibly due do it being a popular Christmas gift.

Both for the class *1.24* as well as for *1.19*, there is a decline in predicted sales quantity during February to April. Hence, for these two classes, which also belong to the same product group, there is a lower demand during these spring months.

By developing a similar table as above but expanding it to include all product lines provided by *Indiska*, guiding principle are given for the predicted total sales quantity for every month.

Hence, fulfilling the intention of making it easier to make decisions, based on quantitative data, on how many products to order place in the buying process.

5.2 Product combinations and conditional probability

The algorithm is used to create the product combinations depending on the conditional probability of the combination being purchased.

The user can decide if all the data should be considered or if the algorithm shall only process specific month(s) when suggesting the combining products together with the analysed input product.

The resulting combinations for an extraction of some products are presented in the following tables.

1.23.2030		
Priority	Product	P(C | A)
1	1.23.2040	0.10209033
2	1.23.2031	0.09251936
3	1.23.2018	0.05081254
4	1.23.2044	0.03987903
5	1.23.2043	0.02126882
6	1.25.2617	0.01767970
7	1.25.2733	0.01345917
8	1.25.2681	0.01262836
9	1.20.1552	0.01259513
10	1.19.1321	0.01086704

Table 5.4. Suggested product combinations for product 1.23.2030

As is shown in *Table 5.4.*, given that the product *1.23.2030* has been purchased, the conditional probability for the first combining product *1.23.240* is 0.10209033. Especially noticeable for this product is that the combinations are solely intra-group combinations, e.g. all the suggested combinations include products from the same group and the first five suggested combination are combinations of products within the same product class.

The following table presents the suggested product combinations for the product *3.50.5760*.

3.50.5760		
Priority	Product	P(C | A)
1	3.50.5762	0.27578913
2	1.23.2030	0.01025056
3	3.50.5794	0.00846078
4	3.50.5769	0.00813537
5	1.23.2018	0.00797266
6	3.60.6085	0.00780995
7	3.50.5777	0.00764725
8	3.51.6007	0.00748454
9	3.50.5772	0.00699642
10	3.50.5771	0.00699642

Table 5.5. Suggested product combinations for product 3.50.5760

All of the product combinations, except for number two and five, consists of products from the same group. Although, the conditional probability for the first combinations is high, the conditional probability for the other combinations decreases rapidly when descending in priority.

Below, the results are presented for the product *2.40.3443*.

2.40.3443		
Order of priority	Product	P(C | A)
1	1.25.2733	0.02226891
2	1.23.2018	0.02100840
3	1.24.2398	0.01848739
4	1.23.2070	0.01722689
5	1.19.1321	0.01701681
6	1.23.2030	0.01659664
7	1.14.852	0.01596639
8	1.24.2397	0.01575630
9	1.23.2044	0.01470588
10	1.14.858	0.01470588

Table 5.6. Suggested product combinations for product 2.40.3443

The combinations for this product are, as for the product *3.50.5760*, intra-group combinations. Although, for the product in *Table* 5.6*.,* the result show that when buying this product from the group *2,* all combinations include products from the group *1.* Nonetheless, for this product, the product classes of the combinations vary to a greater extent.

To present the findings slightly more intuitively, the algorithm calculates for instance the conditional probability of purchasing ‘Tiles Mini Tray’ in the colourway Blue given that ‘Mini Tray’ in the Nival print also is purchased, to 0.10428101 for the months June, July and August. This case is presented as a clarification and is independent from the extracted products presented in the tables above.

If instead entering the following two input products, *3.81.8703* and *3.81.8748*, the below suggested product combinations are generated from the algorithm.

3.81.8703 and 3.81.8748		
Order of priority	Product	P(C | B ∩ A)
1	3.81.8715	0.30645161
2	3.81.8721	0.22903226
3	3.81.8681	0.13548387
4	3.81.8749	0.12903226
5	3.81.8704	0.10967742
6	3.81.8710	0.04838710
7	3.81.8747	0.04677419
8	3.81.8716	0.03870968
9	3.81.8722	0.03709677
10	3.81.8900	0.02741935

Table 5.7. Suggested product combinations for products 3.81.8703 and 3.81.8748

In *Table* 5.*7.,* it is evident that if the customers buy two products from the same product class, it is highly likely that they also buy a third product from that class. In fact, the conditional probabilities of buying the two input products as well as one of the five first suggested combinations are all above 10 percent.

For example, the conditional probability of purchasing the dress ‘Bardot’ given that the top ‘Birgitte’ and the shorts ‘Zola’ are also purchased is 0.1386138 for the months June, July and August. Again, the presented case is only for clarification and is independent from the extracted products presented in the table above.

6. Conclusion

This section will enlighten the findings of the research questions stated in the initial part of the thesis.

6.1 Research questions

The purpose with the study was to provide means for better understanding the behaviour of the customers and their preferences by using applied mathematics and statistical frameworks. Altogether, prediction models using a generalized linear model with a Negative binomial distribution have been developed for all analysed product classes. Furthermore, an algorithm has been formulated to suggest product recommendations based on historical transaction patterns.

6.1.1 How can a midsize, Swedish, fashion brand, utilize historical transaction data to predict future sales quantities of the product classes?

It can be established that a generalized linear model can be used to predict future sales quantities of the different product classes provided by *Indiska*. For the data used, it can be confirmed that a Negative binomial distribution is suitable to use in the modeling. The findings are presented in in its entirety in Section 5.

6.1.2 What parameters should be used to predict consumer purchase behaviour within the fashion industry?

From the regression analysis, it can be concluded that all the factors included in the model, original price, purchase month, colour, cluster, purchase country and channel are significant for the predicted outcome of the sales quantity for each product class.

6.1.3 Which of the parameters are most significant when predicting the sales quantities for different product classes?

Judging from the results from the generalized linear model with Negative binomial regression, all explanatory variables have been found significant for the predicted outcome of the sales quantity. There are however different characteristics of each factor which have the most significant influence on the total sales quantity, although these depends and varies between the different product classes. See section 7.2 for a more elaborated analysis.

6.1.4 How can customer purchase behaviour be used to create more personalised product recommendations?

The historical transaction data can be used as a way of analysing customer purchases behaviour. It has been shown that an algorithm can be constructed that can, by using conditional probability and historical sales data, create recommendations of suggested product combinations of one or two products that can be bought together with an initial product that a customer shows interest

in. Although, these findings are a theoretical personalisation and actual tests by implementing the results would be needed to prove the actual effect.

6.2 Elaborated findings

Stated below is an interpretation of similarities and differences of each of the factors included in the model for the different product classes.

Original price
The predicted parameters of the groupings of the factor original price depends a great deal on the product class. For example, for product classes of which the customers might expect a lower price, grouping which includes a higher price have been found to have a negative effect on the predicted outcome. Although, this might also be an effect from the fact the most items of the class have a lower price.

Cluster
For most product classes, B is the cluster that tended to have the highest sales quantity. The quantity of products with article classification B is much higher in relation to products with article classification C or F which might be an explanation for this outcome.

Purchase country
The variable purchase country was divided in three different groups, Sweden, Norway and Finland and it is evident that Sweden is the most selling country which is also the biggest market by far and thus this result was expected.

Purchase month
The grouping of purchase month turned out very individual depending on the product class due to very high seasonal variations. It is though evident that for product classes which are highly demanded on sale, there are a peaking tendency of the month groups including periods with sale.

Colour
The colour group of which includes COLOR and MULTI have generally an enhancing factor of the predicted sales quantity. This is proof of that the customers prefer colourful products and that unique business idea of *Indiska* should be preserved and even enhanced.

Channel
For all analysed product classes, the channel group Store tends to have been found to be most significant in the prediction of the sales quantity. What might be an explanation for this outcome is that the Stores largely account for the total sales.

7. Discussion

For the modeling using GLM and Negative binomial distribution the significance of the variable grouping turned out be high for most of the product classes. The residual analysis showed promising results and turned out much more desirable for the Negative binomial distribution than for the Poisson distribution. This desirable result is probably an effect of the large amount of data and observations implemented in the regression analysis.

Although, a Negative binominal distribution fitted the data set well, there were a few product classes of which the model did not fit the distribution perfectly. However, the grouping of the variables turned out to be contributing factor to the significance of the models and the estimated parameters.

The structure of the initial data contained many errors which did not originate from the data sampling. Some of the transactions included in the original data set lacked information regarding original price and article classification and was therefore deleted from the data set with a smaller amount of data consequently. Moreover, the structure of entering product colours in the product system is inadequate since a repeating product, but with a different print, cannot have the same colour as one of the products in a previous season. This sometimes calls for sub optimate measures, entering a colour of the product which do not correspond to the true colour. Considering the huge size of the initial data, it is not certain that all these errors were discovered and properly taken care of.

The requirements that were implemented in the algorithm to secure mathematically stable results might have disregarded probable product combinations. Furthermore, requirements that are lacking in reasoning and explanation can have a negative effect on the resulting combinations generated by the algorithm. Since the algorithm do not handle changed preferences of the customers, a way to adjust for the shifts of interests is by updating the initial sales data on a regularly basis.

Since a Negative binomial distribution proved to be a good fit for the data, the results can be perceived as reliable. Although, by implementing requirements in the algorithm when generating the product combinations, the reliability of the concluding results might have been reduced.

Depending on the structure and the extent of the initial data, a similar approach with a Negative binomial distribution as used in this thesis might yield different outcomes and results and thereby affecting the validity of the thesis negatively.

For some classes, the significance was however still low after the grouping of the variables which might have had a negative effect on the reliability of the regression analysis. It is possible that a different grouping might have yielded a higher significance of the variables, nut since time was limited for this thesis, such groupings were not conceivable. The fact that a wrong

colour might be registered for some products might as well affect the reliability in a negative way.

The retail industry and especially the fashion sector is volatile by nature. Since the preferences and demands of a customer are undergoing constant changes, the presented results have a limited time of reliability.

8. Further research

One difficulty that reoccurred on several occasions during the pre-phase of the study was to handle the flaws of the structure of the initial data. By constructing a better way of sorting and cleaning the initial data not only will time consuming actions be reduced but also the building of the model will benefit.

Since shifts of customer interests is disregarded in this study and considering the rapid movements in the sector, an interesting development for further research would be to add more flexibility regarding behaviour changing factors such as existing and future trends, sale, campaigns and marketing when modeling and building the algorithm.

To immerse even more in personalisation, a way for the algorithm to take into account more personalised data such as age, residence and preferences when creating the product recommendations would be of interest.

9. References

Amed, I., Berg, A., Brantberg, L., Hedrich, S., Leon, J., & Young, R. (2016). *The State of Fashion 2017*. Business of Fashion, McKinsey&Company.

Amed, I., Berg, A., Kappelmark, S., Hedrich, S., Andersson, J., Drageset, M., & Young, R. (2017). *The State of Fashion 2018*. Business of Fashion, McKinsey&Company.

Dart, M., & Lewis, R. (2018). Digitally Native Vertical Brands & Guideshops. Retrieved from https://www.atkearney.com/web/consumers-250/article?/a/digitally-native-vertical-brands-guideshops

Greene, W. H. (2012). *Econometric Analysis*. *Prentice Hall* (7th ed.). Pearson Education Limited.

Gut, A. (2009). *An Intermediate Course in Probability* (2nd ed.). New York: Springer Series+Business Media.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. *Springer Texts in Statistics*. New York: Springer Series+Business Media.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed.). Chapman and Hall.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (Fifth Edit). Hoboken, New Jersey: John Wley & Sons, Inc.

Olsson, U. (2002). *Generalized Linear Models*. *Studentlitteratur*. Lund: Studentlitteratur.

Svensk Handel. (2018). *Det Stora Detaljhandelsskiftet*. Stockholm. Retrieved from http://www.svenskhandel.se/globalassets/dokument/aktuellt-och-opinion/rapporter-och-foldrar/e-handelsrapporter/det-stora-detaljhandelsskiftet_2018-digital-version-08052018.pdf

TRITA -SCI-GRU 2018:358