Aalto University
School of Science
Degree programme in Engineering Physics and Mathematics

Recognizing Campaigns that Cause Cannibalization

Bachelor's Thesis
11.07.2018

Petra Huttunen

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of bachelor's thesis

Author Petra Huttunen

Title of thesis Recognizing Campaigns that Cause Cannibalization

Degree programme Engineering Physics and Mathematics

Major Mathematics and Systems Science **Code of major** SCI3029

Supervisor Fabricio Oliveira

Thesis advisor(s) Tuomas Viitanen

Date 11.07.2018 **Number of pages** 4+30 **Language** English

Abstract

In today's consumer society, people and companies are growing more and more aware of the waste they produce and ways to minimize it. One way for companies to minimize waste is by predicting and reacting to the demand of their products. Demand forecasting is affected not only by buyers' decisions, but also external factors like weather, holidays and promotions. Even though promotions are managerial decisions that aim to boost sales, increase the revenue of a certain product or product group, or increase customer flow to the store, in some cases, the increased sales of one product can have a negative effect on the sales of another product.

Cannibalization is the phenomenon where a product diverts sales or market share from another product with similar attributes. The other product might seem more appealing to the customer because of a lower price, better marketing or novelty. While most of the research in the field focuses on cannibalization with new product introductions, this thesis sheds light on what promotion attributes cause cannibalization. Promotions have different effects on the demand of products, which means that accurate forecasting of cannibalization requires information about not only the cannibalization relationship, but also the promotion that causes cannibalization.

The analysis was done by fitting a regression model to sales and promotion data of one UK retailer. The data consisted of approximately 1400 product-campaigns from four different perishable product groups and their different parameters extracted from a Relex Solutions environment. In addition to fitting a suitable model to the data, the aim was to gain understanding of the attributes of campaigns that have a significant effect on cannibalization. The dependent variable in the model was the sales deficit of the victim product for the time period when the affecting product was in a promotion.

After fitting a model to the data, it was clear that the dataset contained too much noise to be able to get reasonable results. Information such as marketing plans and product attributes were not included in the dataset, which might have been partly the reason why the model did not explain complex human decisions as well as was expected. In addition, the data included many exceptional data points which were not removed successfully prior to the analysis. Despite the difficulties in finding a suitable model, the analysis gave good insight as to which campaign attributes significantly affect cannibalization. For example the campaign category and the metrics describing the strength of the cannibalization relationship seem to have an effect on the magnitude of cannibalization caused by a campaign.

Keywords cannibalization, regression analysis, retail, promotions, perishable products

Aalto-yliopisto
Perustieteiden
korkeakoulu

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi

Tekniikan kandidaatintyön tiivistelmä

Tekijä Petra Huttunen		
Työn nimi Kannibalisaatiota aiheuttavien kampanjoiden tunnistaminen		
Koulutusohjelma Teknillinen fysiikka ja matematiikka		
Pääaine Matematiikka ja systeemitieteet		**Pääaineen koodi** SCI3029
Vastuuopettaja Fabricio Oliveira		
Työn ohjaaja(t) Tuomas Viitanen		
Päivämäärä 11.07.2018	**Sivumäärä** 4+30	**Kieli** Englanti

Tiivistelmä

Nykypäivän kuluttajayhteiskunnassa ihmiset ja yritykset ovat yhä tietoisempia tuottamastaan jätteestä. Yritykset voivat vähentää jätettä ennustamalla tuotteiden kysyntää ja reagoimalla siihen. Kysynnän ennustaminen on monimutkainen kokonaisuus, johon vaikuttavat inhimillisten päätösten lisäksi ulkoiset tekijät kuten sää, juhlapyhät sekä tarjoukset ja kampanjat. Vaikka kampanjat ovat vähittäiskauppiaalle tapa kasvattaa tuotteen myyntimääriä tai kaupan asiakasmääriä, niillä voi joskus olla negatiivinen vaikutus jonkin toisen tuotteen myyntiin.

Kun tuote kasvattaa myyntiään toisen samankaltaisen tuotteen myyntien kustannukselta, sanotaan että tämä tuote kannibalisoi kyseessä olevaa toista tuotetta. Tuote saattaa olla kuluttajan näkökulmasta houkuttelevampi esimerkiksi kampanjan, uutuudenviehätyksen tai paremman markkinoinnin takia. Uutuuksien aiheuttama kannibalisaatio on suhteellisen yleinen aihe kirjallisuudessa, mutta kampanjoiden aiheuttamaa kannibalisaatiota on tutkittu vähemmän. On kuitenkin totta, että erilaiset kampanjat vaikuttavat tuotteiden myyntiin eri tavoin, joten niiden aiheuttamat kannibalisaatiovaikutukset eroavat. Tämä tutkielma keskittyykin kampanjoiden ominaisuuksiin ja siihen, miten nämä eri ominaisuudet vaikuttavat kampanjan aiheuttaman kannibalisaation voimakkuuteen.

Kannibalisoivia kampanjoita analysoitiin sovittamalla regressiomalli erään vähittäiskauppiaan myynti- ja kampanjadataan. Data koostui neljän eri tuoretuoteryhmän tuotekampanjoista neljän vuoden aikajaksolta. Tuotekampanjoiden parametrit ja kannibalisaatiosuhteet luettiin Relex Solutions -ympäristöstä. Regressiomallin sovittamisen lisäksi tavoitteena oli selvittää, mitkä kampanjan ominaisuuksista vaikuttivat kannibalisaatiovaikutukseen merkittävästi. Mallin selitettävänä muuttujana käytettiin kannibalisaation alaisena olevan tuotteen myynnin alijäämää kannibalisoivan tuotteen kampanjan ajalta.

Mallin sovittamisen jälkeen oli selvää, että data sisälsi liikaa epäpuhtauksia ja poikkeuksia järkevän mallin sovittamista varten. Data ei sisältänyt tietoa esimerkiksi kampanjoiden markkinointisuunnitelmista tai tuotteiden ominaisuuksista, vaikka molemmat vaikuttavat kirjallisuuden mukaan ihmisten ostopäätöksiin. Mallin sovittamisesta koituneista haasteista huolimatta analyysi auttoi selvittämään, mitkä kampanjan ominaisuuksista ovat kannibalisaatioon liittyen merkitseviä. Esimerkiksi kampanjan kategoria sekä kannibalisaatiosuhteen vahvuus näyttäisivät vaikuttavan kannibalisaation voimakkuuteen.

Avainsanat kannibalisaatio, regressiomalli, kampanja, vähittäiskauppa, ennustaminen

Contents

1 Introduction

In today's consumer society, people and companies are growing more and more aware of the waste they produce and ways to minimize it. While there are already some start ups like ResQ Club that use the leftover food from restaurants and grocery stores, the most profitable solution for companies is to minimize waste by predicting and reacting to the demand of their products. For retailers, demand forecasting plays an important role in their supply chain management, affecting everything from supplier costs to customer satisfaction.

Demand forecasting is affected not only by buyers' decisions, but also external factors like weather, holidays and promotions. While weather and holiday effects on demand are involuntary, promotions are managerial decisions that aim to boost sales, increase the revenue of a certain product or product group, or increase customer flow to the store [Dawes, 2012]. In some cases, however, the increased sales of one product can have a negative effect on the sales of another product and even the sales revenue of the whole product group [Blattberg et al., 1995].

Copulsky [1976] defined cannibalization as the phenomenon where a product diverts sales or market share from another product with similar attributes. The other product might seem more appealing to the customer because of a lower price, better marketing or novelty. Even though cannibalization has been known as a concept for over forty years, most of the research in the field focuses on cannibalization with new product introductions. Research that focuses on promotions that cause cannibalization is rare.

According to a study about food waste in the supplier-retailer interface in the UK and Spain, inaccurate forecasting and cannibalization are some of the main causes of food waste in the supply chain [Mena et al., 2011]. Food waste can be reduced by accurate forecasting, which takes into account weather, promotions and cannibalization. Forecasting the demand of perishable products can be especially challenging because of short shelf life and the attractiveness of products which have been on shelf for a shorter time period [Hvolby and Steger-Jensen, 2015].

This thesis aims to shed light on what promotion attributes cause cannibalization by fitting a regression model to sales and promotion data of one UK retailer. The data consists of approximately 1400 product-campaigns from four different perishable product groups and their different parameters extracted from a Relex Solutions environment. First this paper will go over

previous literature on cannibalization. After that, the method for conducting the analysis is explained followed by the results of the analysis. In the end the results and areas of further research are assessed.

2 Background

Cannibalization as a phenomenon can appear in several different forms. In all of the cases there are at least two subjects that are competing for the same market share or customers — product A, the affecting subject which diverts sales from product B, the victim subject [Raghavan Srinivasan et al., 2005a]. When the products are offered by competing companies, the situation is ideal and the companies can compete with each other. However, if the products are offered by the same company, cannibalization can decrease the total volume of sales and revenue if not taken into account while planning the assortment and promotions. [Mason and Milne, 1994]

The subjects of cannibalization can be products, product groups, brands and even channels [Raghavan Srinivasan et al., 2005a]. The most interesting situation in terms of this thesis is product cannibalization, where one product's sales grow at the expense of another product. The issue is relevant especially for multi-brand companies, which offer a great selection of similar products to satisfy customer's needs [Mason and Milne, 1994].

Product cannibalization can occur in different scenarios. Raghavan Srinivasan et al. [2005a] summarized the scenarios into four different types: In multi-product pack cannibalization, products that have been packed together (e.g. toothbrush and toothpaste) cannibalize individual products. Combo-product cannibalization is similar to multi-product pack cannibalization, except that the products cannot be bought separately. Inter-product cannibalization covers the situation where different products with similar functionalities cannibalize each other - a good example here could be chicken and minced meat. Intra-product cannibalization occurs within the same product group, like two brands of minced meat. [Raghavan Srinivasan et al., 2005a]

A subcategory of intra-product cannibalization is brand-pack size cannibalization, where for example a 12-pack of Coca Cola cannibalizes a single Coca Cola can [Dawes, 2012]. Within the same group product, the remaining shelf life of a product can also be a determining factor for a customer, who would rather buy fresh products instead of products that have been in the store for a longer period of time [Hvolby and Steger-Jensen, 2015].

In this thesis, cannibalization is divided into two categories: cannibalization caused by the product assortment and new product introductions, and cannibalization caused by promotions. In addition to investigating these two areas, different models that have been derived to take cannibalization into account will be examined.

2.1 Product assortment and cannibalization

Offering multiple different choices of products to the customer is seen to be better than offering only one. It's better that the customer chooses between two of the products within that company, than between two products from competing companies. When product managers decide their assortment, it's important that they consider the whole product portfolio and assess the risks of cannibalization within that portfolio. Here branding is especially important, as Copulsky [1976] noticed with coffee brands. If a company brings out a product which is branded similarly as an existing product, even though the product in itself is different, the customer segment or "niche" is the same and the existing product is cannibalized.

Marketing has been compared to ecological phenomena in literature several times. Mason and Milne [1994] used niches in their article to investigate cannibalization within the cigarette market. The niches are demonstrated in Figure 1, where A, B and C are different brands and a, b and c are consumers who bought the corresponding brands. Core consumers are considered to be within the niche of the brand, and fringe customers are outside the niche but might still purchase the brand in question.

Niches were decided based on consumer attributes, like age and gender. If the consumers were within multiple niches, they were subject to cannibalization, which can be seen from the figure where brands A and B overlap. Some consumers bought product a and others product b in that overlapping area. If the niches of two different companies overlapped like A and C, it was considered competition and not cannibalization. The cigarette industry is quite different to other industries in terms of management, as smokers often tend to prefer only one product from the vast variety of options, but the study pointed out that the more products there were in the product portfolio, the more cannibalization occurred. [Mason and Milne, 1994]

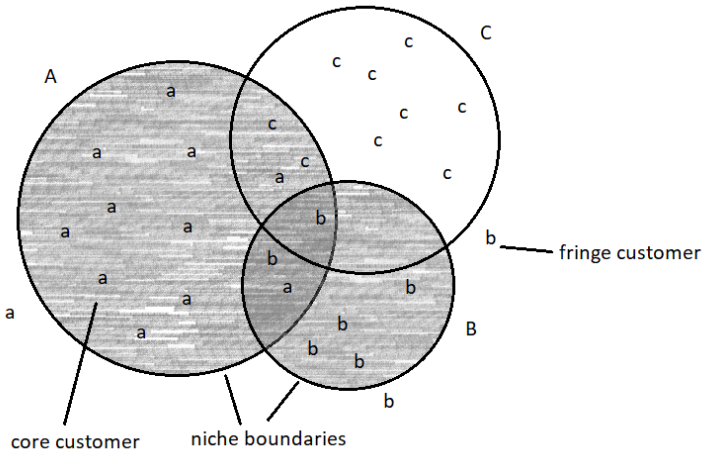

Figure 1: Niche space of brands A, B and C according to Mason and Milne [1994]

Together with branding the product and appealing to distinct customer segments, product attributes are a factor closely knit with cannibalization. Raghavan Srinivasan et al. [2005a] identified different product attributes like package size, package type, family and brand, and researched how much each attribute contributed to cannibalization by comparing sales data trends. In another article, Raghavan Srinivasan et al. [2005b] incorporated cannibalization into forecasting of a new product by using a system where victims were ranked based on the the similarity of product attributes. Based on the rank, the level of cannibalization was determined and then included in the original forecast model.

Raghavan Srinivasan et al. [2005a] noticed in their study that the loss of sales was directly proportional to attribute similarity, and inversely proportional to the number of products with the same level of attribute similarity as the new product. The more attributes the new and old product had in common, the more loss could be expected. As a conclusion, Raghavan Srinivasan et al. stated that attributes are the drivers of product cannibalization, and combining the model in question with the original forecast model reduced over-prediction of sales by 28-46%.[Raghavan Srinivasan et al., 2005b]

However, similarity of products doesn't always lead to higher cannibalization — some brands might have a stronger image and therefore attract more customers, even though the products are very similar [Blattberg et al., 1995]. With perishable products, the freshness of a product can make one product more appealing than the other, so even though the attributes are similar, the

customer can prefer one product over the other [Hvolby and Steger-Jensen, 2015]. Another factor is the functionality of the product — if the product has some added health advantages, it might attract more customers [Yuan et al., 2009].

New product introductions need to be planned carefully to take into account the possible cannibalization. Copulsky [1976] gave Ford and Chevrolet as examples of well and poorly handled introductions: when Ford introduced the Falcon car in 1959, the market share of the standard-sized Ford actually decreased, because the new car was marketed as the smaller version of the old one and stole market share from it. Chevrolet introduced the Corvair at the same time, and it was marketed as a completely new car with different attributes compared to the standard Chevrolet. As a result, the market share of the standard Chevrolet remained the same. Forecasting the demand of a new product is a complex but interesting topic, since there are so many variables that are challenging to predict like timing, marketing and customer characteristics [Raghavan Srinivasan et al., 2005a].

2.2 Promotions and cannibalization

Promotions are a managerial decision, where the aim is to boost sales or attract customer traffic to the store. Promotions have a significant effect the demand and revenue of products, which are important factors for both the retailers and manufacturers. Some manufacturers actually spend more on promotions than they do on advertisement. The reason why promotions are so popular is that they are very effective in increasing sales. Promotions can also be used to affect the brand or image of the store or product. Because promotions have such a significant effect on the retailer and manufacturer, it's extremely important to take cannibalization into account when planning promotions in order to minimize the risk of a negative impact on sales. [Blattberg et al., 1995]

Dawes [2012] researched the relationship between temporary price changes and cannibalization between different pack sizes within a brand for different product groups like cereals, soft drinks and toothpaste. In the study, the ScanPro model was used to investigate the relationship between unit sales effects and short-term changes in price or other promotional variables. The study was focused on different pack sizes within the same brand in multiple different product groups. The ScanPro model is a regression model with unit sales as the dependent variable [Andrews et al., 2008]. Independent variables were price indices of the promoted product and other non-promoted pack

sizes of the same brand, with appropriate lead and lag times. The model also included information about the price of the products, bonus and coupon offers as well as competitor's price and coupon offers. In addition, unit sales of non-promotional pack sizes and holidays were taken into account.

According to the research, price promotions caused cannibalization within the same product group. In 74% of the cases cannibalization occurred, and in those cases 22% of the sales of the promoted product came from a non-promoted product. Cannibalization decreased the revenue more than it did sales quantity. In the study, Dawes [2012] also assessed the significance of the duration of the promotion. During the first week of the promotion the sales peak is usually the highest, and after that the additional sales dampen. The effect of the promotion can also stretch before or after the promotion, if consumers anticipate the promotion or stock up their supplies during promotions - this is called the pre- or post-promotion dip. In most cases, however, the cannibalization effect was limited to the week of the promotion.

Based on further research, Dawes observed that products with a higher price-per-unit are less likely to cannibalize other products but the pack size in itself and the magnitude of the price cut don't make a difference. If the package type itself is different, cannibalization is less likely, and popular products cannibalize other products less. These support the earlier statements that similar product attributes are in the core of cannibalization. However, it was surprising that the magnitude of the price cut didn't make a difference. The relationship between frequency of promotions and cannibalization was also questionable - on one hand frequent promotions train customers to switch brands easily, and on the other hand frequent promotions might make customers immune for promotions. [Dawes, 2012]

However, price promotions are not the only promotions that retailers and manufacturers use. Different types of coupons, bundles and other offers are just as common, and different studies have pointed out that there is a psychological difference between these and price promotions: with price promotions, customers can often see the situation as spending less, while non-price promotions give customers the chance to gain more. Based on some studies, non-price promotions can actually cause more cannibalization than price promotions. [González-Benito et al., 2010]

Hailu et al. [2014] studied the effect of different types of promotions on the demand of pork products in Canada. The Almost Ideal Demand System (AIDS) was used to recognize cannibalization between 11 pork products in Canada. The AIDS model is commonly utilized in studying consumer behaviour [Deaton and Muellbauer, 1980]. Both promotional and demographi-

cal variables were incorporated in the analysis, and the research was done on product-customer level. In the study, it was found out that consumers might be more responsive to coupon offers than to price discounts. The cannibalization effect differed between different pork categories - for some, coupon offers caused cannibalization, and conversely for others the price discounts seemed to cause more negative effects.

Brand loyalty is also a big factor in cannibalization and promotions. A good example is a study conducted within the soft drink industry in Mexico. The goal of the study was to determine how much of the promotion sales of a product came from within the same brand, how much from different brands, and what was the difference between non-price and price promotions. This was achieved by using the attraction model to analyze promotional effects on market share. Most of the promotional sales actually came from the same brand but different sized products (cannibalization), which would mean that in the soft drink industry, customers are reluctant to change brands. This makes sense, since it's common for consumers to prefer either Pepsi or Coca Cola and not switch between the brands. [González-Benito et al., 2010]

3 Research question and methods

The goal of the analysis was to fit a regression model to the campaign and cannibalization data of four perishable product groups. The aim was to identify relationships between the cannibalization effect and different variables, in order to predict whether a campaign will cause cannibalization or not. In addition, the attributes of campaigns and how they affect cannibalization were investigated.

3.1 Data

The data set used in the analysis comprised of four different product groups: chicken, coca cola, beef and frozen potatoes (referred to as chips). The product groups were chosen based on their sales data, promotion history and varying attributes. The sales of the product groups were relatively constant throughout the three-year period without seasonal fluctuations. Each product group had different types of promotions with different durations throughout the time period. Sales data was used from years 2013-2016, of which the last year was used as a test period. The level of investigation was product campaigns, and the parameters and metrics used in the analysis can be seen in Appendix A.

The data set included three different classifications for campaigns: the campaign category, campaign type and campaign subtype. The campaign categories "onshelf", "offshelf" and "FOS" or Front-of-store defined where in the store the campaign was visible to consumers. The campaign type gave information on the magnitude of the price-cut. The campaign subtype divided campaigns into "multibuy" and "other", where campaigns of subtype "other" were campaigns with a price discount.

Cannibalization relationships were calculated from the scope on the lowest level of product hierarchy, because the amount of common attributes is closely related to the strength of the cannibalization effect [Raghavan Srinivasan et al., 2005a]. Based on the calculated relationships, all of the product campaigns that cannibalized other products where chosen for investigation. The cannibalization occurrences were tagged by creating cannibalization events for all of the victim products for the time periods when the cannibalizing products were in a campaign. The baseline sales were calculated for the promotions and cannibalization events using the baseline sales before and after the promotion or event, to get an idea of the sales increases

and decreases.

The data was combined so that for each cannibalizing product campaign there was information about the promotion itself, the cannibalization event and the cannibalization relationship. This way the impact of the relationship strength could be taken into account, as well as the effect of the promotion in explaining the sales difference from the baseline sales of the cannibalization event. The sales difference compared to the baseline was referred to as the additional sales. The additional sales of a cannibalizing campaign were positive, while the additional sales of a cannibalization event were either negative (cannibalization) or positive (no cannibalization).

3.2 Preparing the data

Prior to conducting any analysis, the data was cleaned and normalized. Out of 1400 data points, there were many exceptions that would have skewed the results. These data points were inspected individually and excluded if there was a reasonable explanation for the exception.

In order to find differences in attributes between true cannibalization and false positives, the cannibalization events with both negative and positive additional sales were of interest. However, very positive additional sales of cannibalization events were skewing the data. These data points turned out to be a result of a promotion of the victim product overlapping with the cannibalization event. Distinguishing which part of the sales quantity was an effect of the victim's own promotion and which of it was a result of a cannibalizing product's promotion was not possible in our case. These data points with the high positive additional sales of the victim product were therefore excluded from the data.

The data set included promotions with a duration between some days up to a year. Long promotions were irrelevant in the study because they overlapped with shorter promotions and the cannibalization effect was again harder to measure accurately. Therefore campaigns longer than 50 days were also eliminated from the data set. In addition, only promotions with sales quantity and baseline sales of the cannibalizing promotion being larger than zero were included.

After cleaning the data, the quantity metrics were normalized. The campaigns were of different lengths and for different products, which meant that the deviation of additional sales was significantly high. To normalize the additional sales metrics, the additional sales were divided by the baseline

sales of that product campaign. This seemed to be a reasonable way of making the product campaigns more comparable with each other. The same normalization was performed to the sales quantity metric.

3.3 Analysis

The analysis was conducted by fitting a regression model to the data. As the dependent variable we used the normalized additional sales of the cannibalization event. The independent variables used can be seen in Appendix A. The aim was to test different models to see which variables had a significant effect on the additional sales of the victim, and to build a model which would give us the answer to the question: will the promotion cannibalize the sales of the victim product?

3.3.1 Dependent variables

Based on literature, the following factors were especially interesting:

1. **Relative sales increase of the promotion**
   Based on the definition of cannibalization, the promoted item diverts sales from the cannibalized product [Copulsky, 1976]. If this definition is taken further, it can be stated that the higher the additional sales of the promotion are, the more sales need to be diverted from another product, leading to an increased cannibalization.

2. **Duration of promotion**
   Longer promotions have less additional sales units per week, because the effect of the promotion dampens during time [Dawes, 2012].

3. **Discount percentage and other price metrics**
   Previous studies have given conflicting results regarding the effect of price on a product's sales. In multiple studies, it has been stated that the lower price does cause customers to switch products, but only if the new price is low enough compared to the other products within that group [Blattberg et al., 1995].

4. **Campaign type (price or non-price)**
   According to the research done by González-Benito et al. [2010] and Hailu et al. [2014], the type of promotion affects the magnitude of cannibalization.

5. **Product group**
   Different product groups have different types of cannibalization. This can be seen from different studies, with for example the cigarette industry [Mason and Milne, 1994], soft drinks [González-Benito et al., 2010] and pork products [Hailu et al., 2014].

6. **Strength of the cannibalization relationship**
   The cannibalization relationships have in our case been calculated beforehand. The strength of these relationships should by common sense have an effect on cannibalization.

After identifying potentially interesting factors, the correlations between each independent variable and the dependent variable were investigated.

The data contained the following categorical variables: campaign category, campaign type, campaign subtype and product group. These were analyzed separately to check if there were any differences in the subsets of data, and whether a separate regression model should be fitted to the subsets.

3.3.2 Fitting the regression model

A regression model is a combination of three different parts: the dependent variable y, the systematic part of the model with the independent variables x and the non-systematic part of the model, the residuals ε. The idea is to find coefficients β so that the model,

$$y = f(x; \beta) + \varepsilon \tag{1}$$

best fits to the data. The fit of the model can be maximized for example by using the ordinary least squares method

$$\min_{\beta_0 \ldots \beta_k} \sum_{i=1}^{n} (\varepsilon_i)^2 = \min_{\beta_0 \ldots \beta_k} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})^2 , \tag{2}$$

where the square of the sum of residuals ε is minimized. [Viitasaari, 2017b]

In RStudio the lm-function can be used to find the best fitting model [RDocumentation]. The function chooses best estimates of β for the given variables. In this analysis, a step-by-step process was used. This means that a model was built with all of the dependent variables in the data set from 2013 to 2015, and then insignificant variables were eliminated. Models were evaluated based on the following methods:

1. **The coefficient of determination R-squared and adjusted R-squared**
   These values tell how much of the variance of the data is explained by the model. The variance not explained by the model are the residuals of the model. The adjusted R-squared should be as close to 100% as possible.

2. **P-values of the estimates**
   The p-value of an estimate indicates whether the estimate is significant or not. Significance of estimates was evaluated on the 5%-level.

3. **VIF values of variables**
   A large Variance Inflation Factor (VIF value) $VIF = \frac{1}{1-R^2}$ of the variable means that the risk of multicollinearity is high. Here R^2 is the coefficient of determination of the linear regression model where the variable in question is the dependent variable, and the other variables of the original model are the independent variables. Multicollinearity means that independent variables are strongly correlated. If the VIF value was high for a give variable, the correlation matrix (Appendix B) was checked to confirm cross-correlation and decide which variable should be left out of the model. By eliminating the variable in question, the model's coefficient of determination should not decrease dramatically. [Viitasaari, 2017a]

4. **Cook's distance**
   Cook's distance demonstrates whether there are any significant outliers in the data which affect the fit of the model. The limit for Cook's distances can be decided in multiple ways, but the rule of thumb is that if the value is significantly larger than the other Cook's distances, the data point in question should be examined. [StatisticsHowTo]

5. **Graphs and histogram of residuals**
   When fitting the regression model, assumptions are made regarding the residuals and their distribution. The residuals should be homoscedastic and normally distributed. The variance of homoscedastic residuals is not dependent on the fitted values. Viitasaari [2017a]

The lm-function takes categorical variables into account as dummy variables. This means that the model adds n-1 variables, where n is the number of categories, and addresses those as binary variables. One of the categories is used as the baseline. [RDocumentation]

In addition to testing linear models, we tested a log-transformed model where the dependent and independent variables were logarithmic. Some variables

were also tested as nonlinear variables using their squares and cubes.

3.3.3 Predicting with the model

The model was built based on data from 2013 to 2015. After finding a suitable model, the predictive power of the model was tested on data from 2016. After predicting the test data, the results were validated by comparing the predicted values to the actual values of the test data, and evaluating the standard errors of the prediction.

The R-code for fitting the regression model, validating the model and predicting with the model can be found in Appendix C.

4 Results

Investigating the individual relationships between the dependent variable and independent variables showed that it's difficult to see any pattern in how they correlate with each other. As Figure 2 shows, the relations between variables seem random. The calculated correlations helped in finding patterns in the graphs, and they can be found in Appendix B. For example, the correlation between the normalized additional sales and the absolute price during campaign is slightly positive, and the correlation between normalized additional sales and campaign duration seems to be slightly negative. Based on the plots, it is difficult to make assumptions regarding the linearity or nonlinearity of variables.

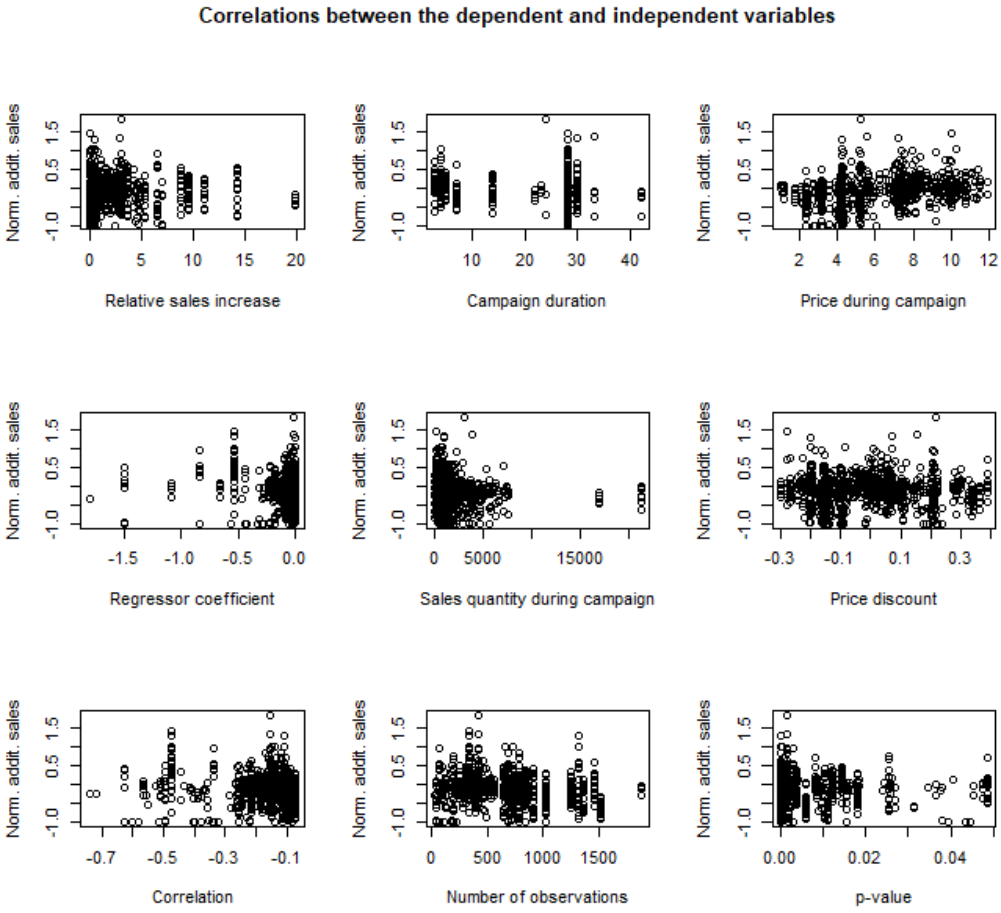

Figure 2: Correlation between the dependent and independent variables. Based on the plots, the relationships between variables seemed very random.

Normalized additional sales of different campaign categories, types and subtypes can be seen in Figure 3. Front-of-store or FOS campaigns seemed to cause the most cannibalization, as well as NO-CUT campaigns. Over 50% discounts seemed to cause the least cannibalization, which was against common sense. However, the amount of campaigns with a discount of over 50% was also very small. Campaign subtypes didn't seem to make a significant difference in the effect of cannibalization.

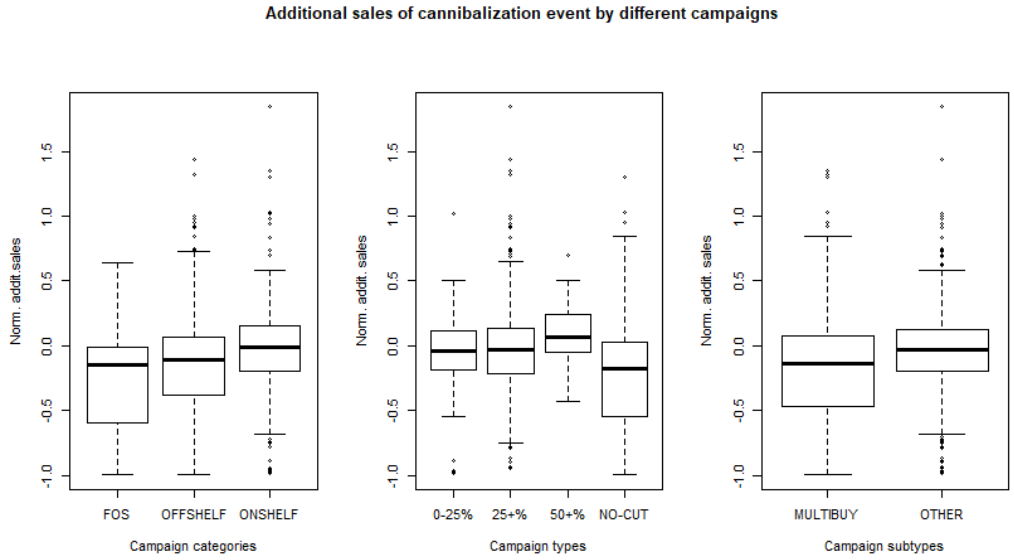

Figure 3: Normalized additional sales by different campaigns. As can be seen from the plots, front-of-store (FOS) campaigns and NO-CUT campaigns seemed to cause the most cannibalization.

Product groups used in the study were quite different compared to each other based on, for example, the number of products in each group. As Figure 4 shows, the normalized additional sales of the cannibalization events within the product group were quite similar, but price and sales quantity differed. The price during the campaign was highest for beef products, and lowest for chips and coca cola. Chips sold significantly higher quantities during campaigns than the other product groups.

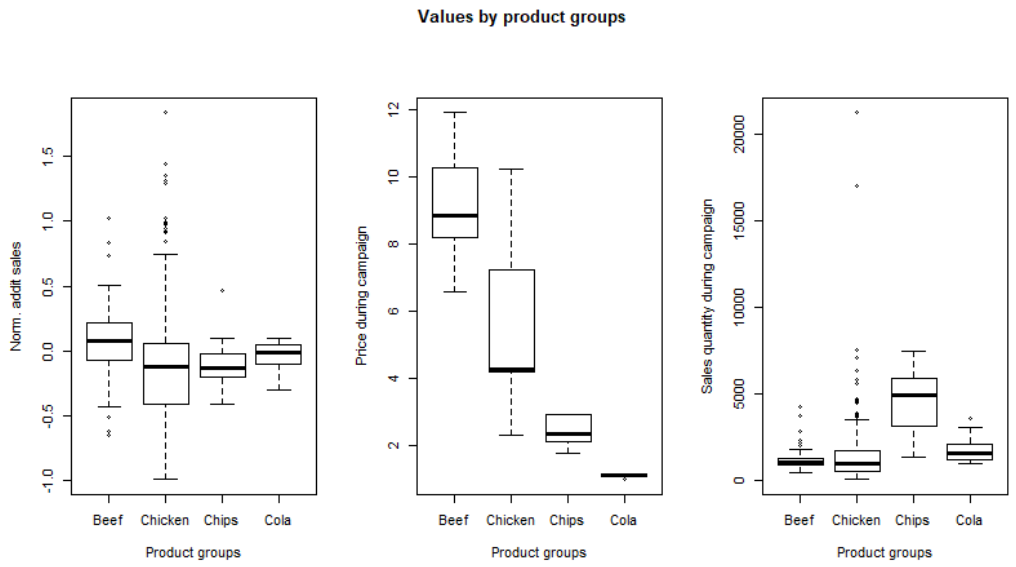

Figure 4: Normalized additional sales of cannibalization event, price during campaign and sales quantity by product groups. Beef seems to be the most expensive product group, while Chips seem to sell the highest quantities during campaigns.

The final regression model which was fitted to the data can be seen in Table 1. All of the coefficients pass the cutoff point of $p < 0.05$, except for the offshelf variable, which is one of the three campaign categories. The high p-value in this case refers to the fact that the offshelf campaigns are not different from the baseline category, i.e. FOS campaigns. Eliminating just one category of a categorical variable from the model was not possible, so the offshelf variable was included in the model regardless of the high p-value.

Table 1: Estimates and p-values of the variables in the final regression model.

Coefficient	**Estimate**	**p-value**
Intercept	0.25	0.0034
Category: offshelf	0.065	0.13
Category: onshelf	0.20	$3.13 \cdot 10^{-5}$
Relative sales increase	0.020	$1.80 \cdot 10^{-6}$
Campaign duration	-0.0058	0.0010
Correlation	1.1	$9.72 \cdot 10^{-7}$
Number of observations	-0.00029	$2.44 \cdot 10^{-9}$
p-value of relationship	-3.4	0.036

The coefficient of determination (R-squared) of the model was $R^2 = 16.45\%$, and the adjusted R-squared was 15.63%. The VIF values were under 2 for all variables, meaning that the risk of multicollinearity was small. However, when examining the correlations between the different variables in the model in Appendix B in Figure 11, it can be seen that there is high cross-correlation between the p-value, number of observations and the correlation of the two products.

From the Cook's distances in Figure 5 it can be seen that there seem to be no significant outliers that have higher Cook's distances than the other observations. However, the number of observations is in our case over 700, and the mean of Cook's distances is $\mu \approx 0.001$. All observations with a Cook's distance of over $3\mu \approx 0.003$ could possibly be significant outliers.

Figure 5: Cook's distances of the model. Even though Cook's distances are relatively small, compared to the mean $\mu \approx 0.001$ there are many potential outliers.

The distribution of the residuals can be seen in Figure 6. Figure 8 also shows that the residuals appear to be random and homoscedastic. The distribution of residuals was analyzed in more detail with the quantile-quantile (Q-Q) plot in Figure 7, which shows the correlation between the residuals and the normal distribution. From the figure, we can see that with higher residuals, the values are not normally distributed.

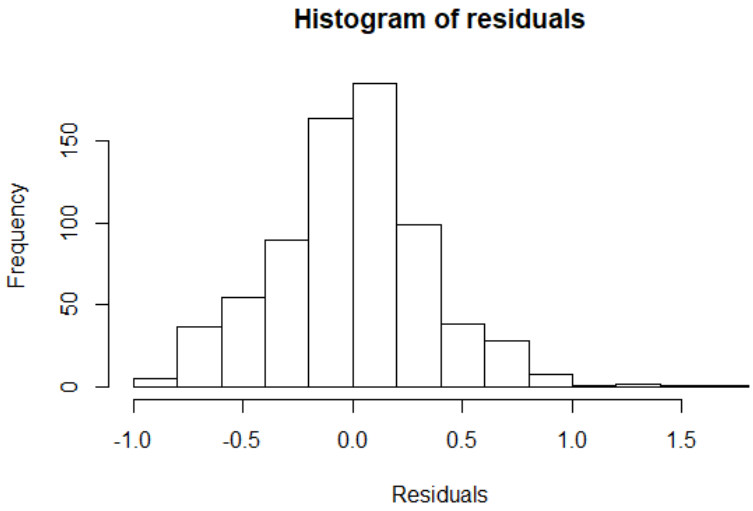

Figure 6: Based on the histogram of the residuals of the model, the residuals seem to be normally distributed.

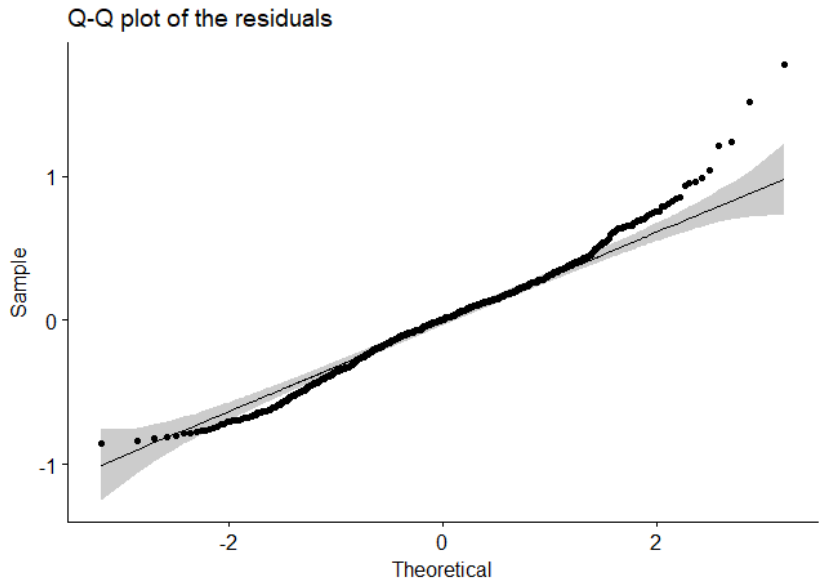

Figure 7: The Q-Q plot of the residuals shows that the residuals with higher values do not exactly follow a normal distribution.

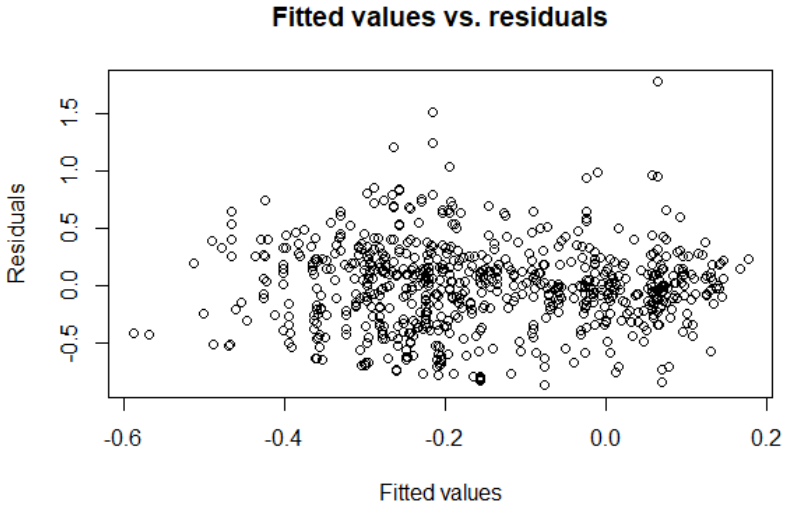

Figure 8: Fitted values and the residuals of the model are quite randomly spread and therefore, the residuals seem homoscedastic.

The prediction of the test data calculated by the model can be seen in Figure 9. The red line demonstrates the optimal prediction where the prediction is the same as the real data. As we can see from the figure, the prediction is random but somewhat in line with the test data, as the data points are spread quite evenly around the optimal line. When we look at the standard errors of the prediction in Figure 10, we can see that the largest standard errors occur with the smallest additional sales. The predictive power of the model seems to be in accordance with fitting results.

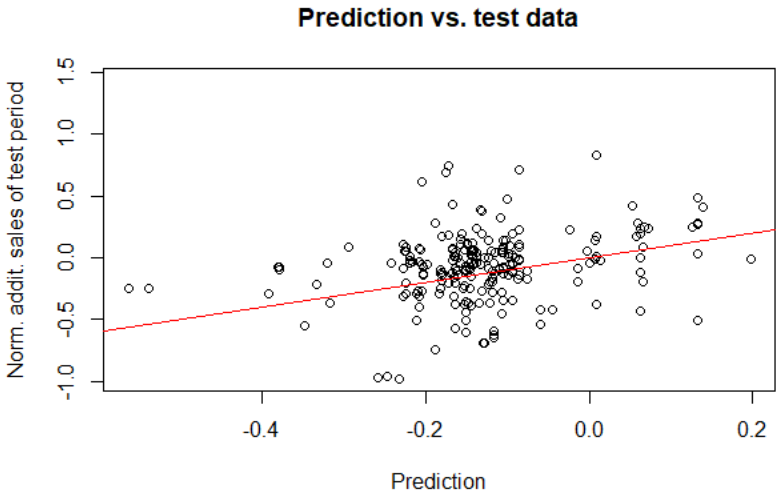

Figure 9: Prediction of the test data by using the constructed regression model. As can be seen, the prediction seems quite random.

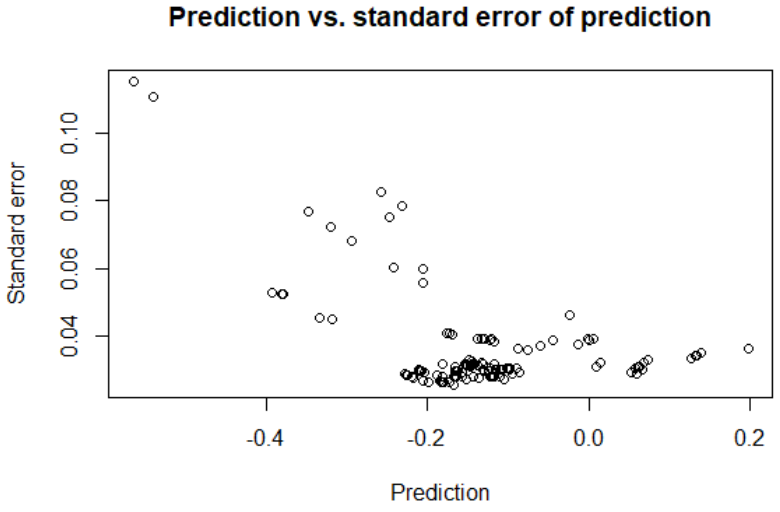

Figure 10: Prediction of the test data vs. the standard error of the prediction. The standard error increases when the prediction decreases.

5 Conclusions

5.1 Discussion of the model and results

The regression model's adjusted coefficient of determination was $R^2 = 15.63\%$, which means that the model only explains 15.63% of the variance of the data. This is a very low percentage, as the ideal percentage would be as close to 100% as possible. This is no surprise when we look at the graphs in Figure 2 — the relationships seemed quite random already in the beginning and fitting the model was challenging.

The validity of the model is questionable. The Cook's distances (Figure 5) seem relevantly small and therefore according to some authors, the outliers are insignificant. However, when the outliers are compared to the mean of the Cook's distances, there are tens of significant outliers. The most significant outliers were data points with generally normal values, and there was no reason to clean those from the dataset. The residuals were homoscedastic according to Figure 8, which could also be a result of randomness of the data. Figure 7 showed that the distribution of residuals is close to a normal distribution, but the higher values of residuals do not quite fit that description.

When we look at the coefficients of the model in Table 1, we can see the effects the significant variables have on the dependent variable. The campaign category "onshelf" increases the normalised additional sales compared to the other two categories. This makes sense when we look at Figure 3: The FOS and offshelf campaigns have very similar additional sales, but the onshelf campaigns seem to cause less cannibalization.

The relative sales increase and campaign duration affect cannibalization in the opposite way as expected. Campaign duration is inversely proportional to the normalized additional sales, which means that longer campaigns cannibalize more than shorter campaigns. However, the estimate is only slightly negative, and the reason behind this could be the fact that more consumers have time to hear about the campaign if the duration is longer. Even more questionable is the fact the bigger the relative sales increase of a campaign is, the less it will cannibalize. This finding is completely against the assumptions where the campaign sales are diverted from the cannibalized product. One explanation, however, could be that in campaigns with very high additional sales, the sales are diverted from competing companies, not from within the same company.

The estimates of variables representing the strength of the cannibalization relationship were logical. The correlation was directly proportional to the additional sales. In our case, correlation gets only negative values and the closer to zero (i.e. the bigger it is), the less there appears cannibalization (i.e. the larger the additional sales). Both the number of observations and p-value of the relationship are inversely proportional to the additional sales, therefore the less observations of cannibalization there have been and the smaller the p-value is, the more cannibalization will occur. As all of these variables represent the strength of the cannibalization relationship, it's no surprise that the cross-correlation is high, based on the table in Appendix B. The validity of the model is therefore questionable. Using only one of these three variables decreased the R-squared of the model and resulted in that specific variable to be insignificant in the model.

In the process of finding the most suitable model, we tried log- and nonlinear models. For example price and relationship values were tested as nonlinear variables, but the R-squared value decreased with each try. Logarithmic functions didn't improve the model either. These are logical outcomes based on Figure 2 — the relations are linear if anything. These findings are also in line with the cannibalization models derived by Raghavan Srinivasan et al. [2005b] and Dawes [2012], who also used linear bases for their models.

The data set in use contained both absolute values and relative values (Appendix A). Using absolute values to model relative values or using absolute values as the dependent variable was in our case questionable, because we had four very different product groups in the data set. On the other hand, absolute variables like price during campaign and sales quantity of a campaign were significant in some models that were tested out. However, none of these kinds of variables were in the final model.

In addition to sales quantity and price during campaign being significant variables in some models, subgroups as a categorical variable was also an interesting factor. As mentioned before, the attributes of products are key drivers of cannibalization according to Raghavan Srinivasan et al. [2005a], so the product group having an effect on cannibalization would be logical. Modeling subgroups separately was also tested during the analysis, but it didn't provide any improvement on results.

Price discount was not a significant variable in any of the models reviewed. This is quite an interesting find and is against common sense, but is in line with the findings of Dawes [2012] and partly in line with the findings of Hailu et al. [2014].

5.2 Future prospects

Even though the data was cleaned before the analysis to see that each half-year period had sales, shorter zero-sales periods remained in the data. Because the data was comprised on product campaign level, we were only able to exclude whole campaigns and cannibalization events with zero sales quantity. If for one week of a two-week campaign the product was out of stock, there is no way that we could have excluded that specific data point. These periods can affect the additional sales of cannibalization events quite dramatically, because they cause the cannibalization effect to be stronger than it is in reality. This should be taken into account in further research.

The baseline sales are calculated based on the average sales of the weeks before and after the campaign. If those weeks are also promotion weeks, the calculation logic has to find regular weeks from further away. This means that long back-to-back campaign or cannibalization event periods might not have the most reliable baseline sales. Therefore, the additional sales of campaign and cannibalization events might also be inaccurate in some cases. Identifying and excluding these cases could possibly improve the results.

Another factor that made the analysis challenging was holiday events that overlapped with the cannibalization events or with campaigns, which usually increase the additional sales during that time period. Excluding all product-campaigns overlapping with for example Christmas and Easter would have decreased the number of data points radically, but the effect of the holidays could be researched more.

The data set used in the analysis consisted mostly of numeric variables in addition to four category variables. However, the human decisions leading to buying a certain product can be affected by many more factors than just the campaign type and product group [Blattberg et al., 1995]. The analysis conducted in this thesis didn't take into account the marketing plans of the campaign, or product attributes such as quality, brand, functional attributes, package type and package size of the product, because that information was not available. According to the literature, these factors have an affect on cannibalization as well, so gathering the information and taking it into account could improve results.

In this analysis, the cannibalization relationships were provided by Relex Solutions. The research done in the beginning of this thesis did provide ideas as to how the relationships could be calculated more accurately. The relationship calculation was restricted to the level of product groups, even

though there can be cannibalization between product groups as well like Raghavan Srinivasan et al. [2005a] suggested. In addition, product attributes could be taken into account even better in the relationship calculation.

References

R.L. Andrews, I.S. Currim, P. Leeflang, and J. Lim. Estimating the scanpro model of store sales: Hb, fm or just ols? *International Journal of research in marketing*, 25(1):22–33, 2008.

R.C. Blattberg, R. Briesch, and E.J. Fox. How promotions work. *Marketing science*, 14(3):G122–G132, 1995.

W. Copulsky. Cannibalism in the marketplace. *The Journal of Marketing*, pages 103–105, 1976.

J.G. Dawes. Brand-pack size cannibalization arising from temporary price promotions. *Journal of Retailing*, 88(3):343–355, 2012.

A. Deaton and J. Muellbauer. An almost ideal demand system. *The American economic review*, 70(3):312–326, 1980.

Ó. González-Benito, Z.I. Loyola-Galván, and P.A. Muñoz-Gallego. Inter-size and inter-brand competition analysis within a product category: Scope of cannibalization effects. *Journal of Brand Management*, 17(4):254–265, 2010.

G. Hailu, R.J. Vyn, and Y. Ma. The demand for pork products in canada: Discount promotions and cannibalization. *Agribusiness*, 30(4):470–492, 2014.

H.H. Hvolby and K. Steger-Jensen. Managing cannibalization of perishable food products in the retail sector. *Procedia Computer Science*, 64:1051–1056, 2015.

C.H. Mason and G.R. Milne. An approach for identifying cannibalization within product line extensions and multi-brand strategies. *Journal of Business Research*, 31(2-3):163–170, 1994.

C. Mena, B. Adenso-Diaz, and O. Yurt. The causes of food waste in the supplier–retailer interface: Evidences from the uk and spain. *Resources, Conservation and Recycling*, 55(6):648–658, 2011.

S. Raghavan Srinivasan, S. Ramakrishnan, and S.E. Grasman. Identifying the effects of cannibalization on the product portfolio. *Marketing intelligence and planning*, 23(4):359–371, 2005a.

S. Raghavan Srinivasan, S. Ramakrishnan, and S.E. Grasman. Incorporating cannibalization models into demand forecasting. *Marketing intelligence and planning*, 23(5):470–485, 2005b.

RDocumentation. Documentation regarding the lm-function. https://www.rdocumentation.org/packages/stats/versions/3.5.0/topics/lm. Online; accessed 1.6.2018.

ResQ Club. Waste food restaurant. https://www.resq-club.com. Online; accessed 01.06.2018.

StatisticsHowTo. Cook's distance. http://www.statisticshowto.com/cooks-distance/. Online; accessed 7.6.2018.

L. Viitasaari. Regressiodiagnostiikka ja regressiomallin valinta, lecture materials. https://mycourses.aalto.fi/mod/resource/view.php?id=260615, 2017a. Online; accessed 1.6.2018.

L. Viitasaari. Yleinen lienaarinen malli, lecture materials. https://mycourses.aalto.fi/mod/resource/view.php?id=260614, 2017b. Online; accessed 1.6.2018.

Y. Yuan, O. Capps, and R.M. Nayga. Assessing the demand for a functional food product: is there cannibalization in the orange juice category? *Agricultural and Resource Economics Review*, 38(2):153–165, 2009.

A Variables used in the analysis

Table 2: Table of the different variables used in the analysis.

Name of variable	**Type**	**Category**
Sales quantity	Quantity	Cannibalizing campaign
Baseline sales	Quantity	Cannibalizing campaign
Additional sales	Quantity	Cannibalizing campaign
Campaign duration	Days	Cannibalizing campaign
Price during campaign	Value	Cannibalizing campaign
Price before campaign	Value	Cannibalizing campaign
Price discount	Percentage	Cannibalizing campaign
Relative sales increase	Percentage	Cannibalizing campaign
Sales quantity of victim	Quantity	Cannibalization event
Baseline sales of victim	Quantity	Cannibalization event
Additional sales of victim	Quantity	Cannibalization event
Normalized additional sales of victim	Percentage	Cannibalization event
Regressor coefficient	Quantity	Cannibalization relationship
Number of observations	Quantity	Cannibalization relationship
Correlation	Quantity	Cannibalization relationship
T-statistic	Quantity	Cannibalization relationship
P-value	Quantity	Cannibalization relationship

B Correlation matrix of the dataset

	Sales quantity	Baseline sales	Additional sales	Campaign duration	Price during campaign	Price before campaign	Price discount	Relative sales increase	Sales quantity of victim	Baseline sales of victim	Additional sales of victim	Normalized additional sales	Regressor coefficient	Number of observations	Correlation	T-statistic	P-value
Sales quantity	1.00	0.77	0.94	0.15	-0.28	-0.17	0.38	0.25	0.27	0.33	-0.26	-0.11	0.10	-0.01	-0.07	0.00	0.08
Baseline sales	0.77	1.00	0.49	0.24	-0.41	-0.37	0.21	0.23	0.28	0.36	-0.33	-0.18	0.05	0.13	-0.07	-0.05	0.03
Additional sales	0.94	0.49	1.00	0.08	-0.16	-0.02	0.40	0.21	0.21	0.25	-0.17	-0.06	0.10	-0.09	-0.06	0.03	0.10
Campaign duration	0.15	0.24	0.08	1.00	-0.55	-0.53	0.00	0.19	0.10	0.17	-0.22	-0.22	-0.10	0.10	-0.08	-0.10	0.09
Price during campaign	-0.28	-0.41	-0.16	-0.55	1.00	0.92	-0.17	-0.23	-0.17	-0.24	0.24	0.28	0.11	-0.31	0.12	0.18	-0.05
Price before campaign	-0.17	-0.37	-0.02	-0.53	0.92	1.00	0.20	-0.13	-0.15	-0.22	0.25	0.31	0.13	-0.41	0.10	0.23	0.02
Price discount	0.38	0.21	0.40	0.00	-0.17	0.20	1.00	0.15	0.08	0.08	-0.04	0.04	0.08	-0.27	-0.08	0.11	0.14
Relative sales increase	0.25	0.23	0.21	0.19	-0.23	-0.13	0.15	1.00	-0.05	-0.06	0.06	0.04	0.01	0.09	0.00	-0.05	0.03
Sales quantity of victim	0.27	0.28	0.21	0.10	-0.17	-0.15	0.08	-0.05	1.00	0.94	-0.07	0.19	-0.28	-0.10	-0.21	-0.22	-0.03
Baseline sales of victim	0.33	0.36	0.25	0.17	-0.24	-0.22	0.08	-0.06	0.94	1.00	-0.40	-0.05	-0.28	-0.08	-0.23	-0.23	-0.04
Additional sales of victim	-0.26	-0.33	-0.17	-0.22	0.24	0.25	-0.04	0.06	-0.07	-0.40	1.00	0.65	0.06	-0.04	0.10	0.07	0.05
Normalized additional sales	-0.11	-0.18	-0.06	-0.22	0.28	0.31	0.04	0.04	0.19	-0.05	0.65	1.00	-0.08	-0.18	-0.07	-0.05	-0.01
Regressor coefficient	0.10	0.05	0.10	-0.10	0.11	0.13	0.08	0.01	-0.28	-0.28	0.06	-0.08	1.00	0.32	0.68	0.55	0.11
Number of observations	-0.01	0.13	-0.09	0.10	-0.31	-0.41	-0.27	0.09	-0.10	-0.08	-0.04	-0.18	0.32	1.00	0.50	0.01	-0.25
Correlation	-0.07	-0.07	-0.06	-0.08	0.12	0.10	-0.08	0.00	-0.21	-0.23	0.10	-0.07	0.68	0.50	1.00	0.77	0.21
T-statistic	0.00	-0.05	0.03	-0.10	0.18	0.23	0.11	-0.05	-0.22	-0.23	0.07	-0.05	0.55	0.01	0.77	1.00	0.57
P-value	0.08	0.03	0.10	0.09	-0.05	0.02	0.14	0.03	-0.03	-0.04	0.05	-0.01	0.11	-0.25	0.21	0.57	1.00

Figure 11: Correlation matrix of the dataset. Correlations of above 0.20 are marked in green, and correlations below -0.20 are marked in red.

C R-code of the analysis

```
#Fitting the model

model1 <- lm(Norm.addit.sales ~
           factor(Category) +
              Stat..increase.combo +
             Campaign.duration +
             Correlation +
              Number.of.observations +
              P.value
                , data = analysis)

#Validating the model
summary(model1)
vif(model1)
hist(model1$residuals, xlab = "Residuals",
main = "Histogram of residuals")
AIC(model1)
plot(cooks.distance(model1), xlab = "Index",
ylab = "Cook's distance")
title("Cook's distances of the regression model")
plot(model1$fitted.values, model1$residuals,
xlab = "Fitted values", ylab = "Residuals")
title("Fitted values vs. residuals")
par(mfrow=c(2,2))
plot(model1)

#Predicting with the model

prediction1 <- predict.lm(model1, training,
se.fit = TRUE,interval = "confidence",
           level = 0.95, na.action = na.pass)
plot(prediction1$fit[,1], training$Norm.addit.sales,
xlab = "Prediction", ylab = "Norm. addit. sales of training period")
abline(a=0, b=1,col="red"  )
title("Prediction vs. training data")
plot(prediction1$fit[,1], prediction1$se.fit,
xlab = "Prediction", ylab = "Standard error")
title("Prediction vs. standard error of prediction")
```