GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Master Thesis

Multivariate modelling of the dependency structure between article sales of a sportswear manufacturer

Author
Petros Christanas
from Nuremberg

Matriculation Number
11604278

Applied Statistics M.Sc.
Chair of Statistics and Econometrics

Supervisors
Prof. Dr. Thomas Kneib
Dipl.-Vw. Quant. Fabian H. C. Raters

Submitted February 20, 2020
Processing time of 20 weeks

Contents

1 Introduction

Write introduction here and "upper" subsections here (adidas, Motivation, etc...)

1.1 Data Sources

Throughout each season, transactional data are collected from online purchases of the sports brand's eCommerce website. Specifically, we are provided with weekly sales data for western European countries depicted in table 1.1.

Column	Description	Values
week_id	Calender week of a specific year (YYYYWW)	Factors: *201648, ..., 201852*
article_number	Unique article identification number (article ID)	Factors: *10669, 10, ...*
min_date_of_week	Minimum date of the respective week; always a Monday (YYYY-MM-DD)	Dates: *2016-11-28, ..., 2018-12-24*
art_min_price	Minimal recorded price of the article	Non-negative (integer) value
month_id	Calender month of a specific year (YYYYMM)	Factors: *201612, ..., 201812*
season	Season of year (format: SSYY) (Spring-Summer [SS]: December - May) Fall-Winter [FW]: June - November)	Factors: *SS17, FW17,* *SS18, FW18, SS19*
bf_w	Weekly "Black Friday" promotion intensity of the article	Between 0 and 1
ff_w	Weekly "Friends & Family" promotion intensity of the article	Between 0 and 1
ot_w	Weekly article promotion intensity of "Other" type	Between 0 and 1
gross_demand_quantity	Weekly amount of added articles to shopping cart	Non-negative (integer) value
base_price_locf	Retail price of the article without any discounts	Non-negative (integer) value
total_markdown_pct		
day_of_month	Day of the month	Integers: *1 - 31*
month_of_year	Month of the year	Factors: *January, ..., December*
year	Year	Integers: *2016, 2017, 2018*
week_of_year	Week of the year	Integers: *1 - 52*

Table 1.1: Transactional raw data description from online purchases of western European countries

Due to legal regulations of the company, some columns had to undergo anonymization in order for the data to be released. To ensure data protection and confidentiality, numeric variables (with exception of time-indicating columns) were transformed. As a consequence for the analysis part, most integer values were converted to float numbers. This fact should be kept in mind by the reader, since the above table serves as a reminder and reference point for the data documentation.

Another peculiarity of this setup is to be considered, too. We will often refer to the variable *gross demand quantity* as *sales*, even though it is obviously not exactly the same. In the eCommerce environment, there are several stages before the purchase is complete, e.g. addition to cart, removal from cart, proceeding to checkout & even the return of bought articles. Targeting the articles added to cart, i.e. the (gross) demand quantity, provides the optimal data extraction for analytical purposes and is the closest to adequately model the dependency structure between net sales of articles.

Besides the transactional data, *article master data*, i.e. attributes of the articles, are provided and depicted in table 1.2.

Column	**Description**	**Values**
week_id	Calender week of a specific year (YYYYWW)	Factors: *201648, ..., 201852*
article_number	Unique article identification number (article ID)	Factors: *10669, 10, ...*
min_date_of_week	Minimum date of the respective week; always a Monday (YYYY-MM-DD)	Dates: *2016-11-28, ..., 2018-12-24*
art_min_price	Minimal recorded price of the article	Non-negative (integer) value
month_id	Calender month of a specific year (YYYYMM)	Factors: *201612, ..., 201812*
season	Season of year (format: SSYY) (Spring-Summer [SS]: December - May) Fall-Winter [FW]: June - November)	Factors: *SS17, FW17,* *SS18, FW18, SS19*
bf_w	Weekly "Black Friday" promotion intensity of the article	Between 0 and 1
ff_w	Weekly "Friends & Family" promotion intensity of the article	Between 0 and 1
ot_w	Weekly article promotion intensity of "Other" type	Between 0 and 1
gross_demand_quantity	Weekly amount of added articles to shopping cart	Non-negative (integer) value
base_price_locf	Retail price of the article without any discounts	Non-negative (integer) value
total_markdown_pct		
day_of_month	Day of the month	Integers: *1 - 31*
month_of_year	Month of the year	Factors: *January, ..., December*
year	Year	Integers: *2016, 2017, 2018*
week_of_year	Week of the year	Integers: *1 - 52*

Table 1.2: Article master data

2 Theory & Methods

This chapter introduces various statistical methods used during the conduction of this thesis. It is assumed that basic understanding and knowledge of the reader regarding mathematical foundations of statistics (like linear algebra, probability theory, etc) already exists.

2.1 Generalized Linear Models

Generalized Linear Models (GLMs) are an extension of the *Classical Linear Regression Model (LM)*

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n$$

which in matrix notation can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where the response variable y_i can take values from several probability distributions (e.g. Poisson, Binomial, Gamma and others), which are members of the exponential family [Fahrmeir et al., 2003]. The linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i = \boldsymbol{x}_i'\boldsymbol{\beta}$$

is passed through a *response function h* (a one-to-one, twice differentiable transformation), such that

$$E(y_i) = h(\eta_i)$$

i.e. the expected value of the response variable belongs to the value range of *h*. The inverse of the response function,

$$g = h^{-1}$$

is called the *link function*.

2.2 Generalized Additive Models

2.3 Mixed Effects Models

Linear Mixed Models (LMMs) are powerful tools when dealing with clustered data or data with a longitudinal structure (repeated measurements of individuals). As in the classical LM, there are population-specific effects, namely the parameter vector of *fixed effects* $\boldsymbol{\beta}$, as well as the cluster- or individual-specific effects of such models called *random effects* [Fahrmeir et al., 2003]. In the following, we will refer to our clusters or individuals as "groups" for briefness. Mathematically speaking, the linear predictor $\eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta}$ is extended to

$$\eta_{ij} = \boldsymbol{x'}_{ij}\boldsymbol{\beta} + \boldsymbol{u}'_{ij}\boldsymbol{\gamma}_i, \quad j = 1, \ldots, m, \quad i = 1, \ldots, n_i, \quad \text{where}$$

- i is the number of groups
- j is the number of observations per group
- $\boldsymbol{\beta}$ is the vector of fixed effects
- γ_i is the vector of random effects
- $\boldsymbol{x}'_{ij}$ is the vector of covariates and
- $\boldsymbol{u}'_{ij}$ is a subvector of $\boldsymbol{x}'_{ij}$.

$\boldsymbol{x}'_{ij} = (1, x_{ij1}, \ldots, x_{ijk})$ & $\boldsymbol{u}'_{ij} = (1, u_{ij1}, \ldots, u_{ijk})$ are therefore the design vectors and ϵ_{ij} are the error terms of the *measurement model*

$$y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{u}'_{ij}\boldsymbol{\gamma}_i + \epsilon_{ij}, \quad \varepsilon_{ij} \overset{i.i.d.}{\sim} N\left(0, \sigma^2\right) \tag{2.1}$$

or in matrix notation
WRITE THE MATRIX NOTATION HERE

List of Figures

List of Tables

List of Abbreviations

BIC Bayesian Information Criterion

GLM Generalized Linear Model

LM Linear Model

LMM Linear Mixed Model

References

[Fahrmeir et al., 2003] Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2003). Regression; Models, Methods and Applications. 2013.

[Lütkepohl, 2005] Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.