

Technische Universität München

CHAIR OF MATHEMATICAL STATISTICS

**Vine Copula based analysis
of online customer demand
under market competition**

Master's Thesis

by

Susanna Elsner

Supervisor: Prof. Claudia Czado, Ph.D.

Advisor: Prof. Claudia Czado, Ph.D. and Daniel Kraus

Submission: 31.07.2015

I hereby declare, that this thesis is my own work and that no other sources have been used except those clearly indicated and referenced.

Garching, 31.07.2015

Abstract

In commerce it is very important to have a good knowledge about the customers' behaviour and about the possibility to influence the sales. Especially in online business, from the customers' side it is very easy to change to another shop and purchase almost everywhere. This results in a very challenging competitive situation.

This master's thesis analyses data from an online shop sampled over a period of two years. Different aspects are covered:

- The product hierarchy is used for several levels of aggregation to support different strategies of the shop from simple pricing to optimisation of the portfolio.
- Several modelling approaches, especially GLMs, GAMs and copula models including different assumptions on the distribution of the response variable are conducted.
- A detailed analysis turns out the relevant covariates like for example prices, advertisement, substitution effects and seasonality.

The results show a good fit and can be applied to the relevant shop activities.

Zusammenfassung

Als Händler ist es besonders wichtig, gut über das Kundenverhalten und die Möglichkeiten, wie man die Verkaufszahlen beeinflussen kann, Bescheid zu wissen. Insbesondere im Online-Handel ist es für die Kunden sehr einfach zwischen verschiedenen Shops zu wechseln und fast überall einzukaufen. Dies führt zu einer herausfordernden Wettbewerbs-Situation.

Diese Masterarbeit beschäftigt sich mit Daten eines Online-Shops, die über einen Zeitraum von zwei Jahren gesammelt wurden. Verschiedene Aspekte werden behandelt:

- Der Produkthierarchie entsprechend werden die Artikel verschieden stark aggregiert, um unterschiedliche Strategien des Shops wie zum Beispiel die Preissetzung oder die Optimierung des Portfolios zu verbessern.
- Mehrere Modellierungsansätze, insbesondere GLMs, GAMs und Copula-Modelle mit verschiedenen Annahmen über die Verteilung der Zielvariablen werden betrachtet.
- Eine detaillierte Analyse zeigt die relevanten Einflussvariablen aus den unterschiedlichen Bereichen wie z.B. Preise, Werbung, Substitution und Saisonalität.

Die Modelle bieten eine gute Möglichkeit, die Daten zu beschreiben. Daraus können im Shop entsprechende Maßnahmen abgeleitet werden.

Contents

| | | |
|----------|----------------------------------------------------------|-----------|
| 1 | Introduction | 1 |
| 2 | Mathematical Background of Modelling | 2 |
| 2.1 | Generalised linear models (GLM) | 2 |
| 2.1.1 | Multiple linear regression models | 2 |
| 2.1.2 | Count regression models | 5 |
| 2.1.3 | Parameter estimation | 11 |
| 2.1.4 | Asymptotic hypothesis tests in generalised linear models | 19 |
| 2.2 | Generalised additive models (GAM) | 20 |
| 2.3 | Exploratory data analysis | 21 |
| 2.3.1 | Dependence measures | 21 |
| 2.3.2 | Scatter plots | 25 |
| 2.3.3 | Interaction effects | 27 |
| 2.3.4 | Seasonal effects | 28 |
| 2.4 | Model selection | 29 |
| 2.4.1 | Information-loss criteria AIC and BIC | 29 |
| 2.4.2 | Deviance | 31 |
| 2.4.3 | Coefficient of determination for linear models | 33 |
| 2.5 | Residual analysis | 34 |
| 2.5.1 | Pearson residuals | 34 |
| 2.5.2 | Deviance residuals | 35 |
| 2.6 | Modelling the economic key figure 'price elasticity' | 36 |
| 2.7 | Copula Modelling | 38 |
| 2.7.1 | Model selection | 39 |
| 2.7.2 | Structure selection | 40 |
| 2.7.3 | Parametric copula families | 43 |
| 2.7.4 | Estimation of Vine copulas | 46 |
| 2.7.5 | Family selection and parameter estimation | 46 |
| 2.7.6 | Comparing copula models | 49 |
| 3 | Study of Online Sales Activities | 51 |
| 3.1 | Online sales activities | 51 |
| 3.1.1 | Data description | 51 |
| 3.1.2 | Data cleaning | 53 |

| | | |
|----------|------------------------------------------------------------------|------------|
| 3.1.3 | Data exploration | 57 |
| 3.2 | Count regression models per article | 71 |
| 3.2.1 | Exploratory data analysis | 72 |
| 3.2.2 | Poisson regression models for article A_6 | 77 |
| 3.2.3 | Negative Binomial regression models for article A_6 | 84 |
| 3.3 | Count regression models per product | 88 |
| 3.3.1 | A Negative Binomial regression model for product P_3 | 88 |
| 3.3.2 | Generalised additive models (GAM) on product basis | 95 |
| 3.4 | Copula modelling on deviance residuals | 107 |
| 3.4.1 | Exploratory data analysis | 107 |
| 3.4.2 | Copula models on product basis | 113 |
| 3.4.3 | Simulating from the copula | 121 |
| 4 | Validation in economical context | 123 |
| 5 | Summary and Outlook | 131 |
| 6 | Indices | 134 |
| 6.1 | List of Figures | 134 |
| 6.2 | List of Tables | 137 |
| 6.3 | Bibliography | 140 |
| 7 | Appendix | 142 |
| 7.1 | A1: Outsourced figures | 142 |
| 7.2 | A2: Outsourced regression results | 149 |
| 7.3 | A3: Outsourced details on copula models | 151 |
| 7.4 | List of the variables of importance | 159 |

Chapter 1

Introduction

In commerce it is very important to have a good knowledge about the customers' behaviour and about the possibility to influence the sales. Especially in online business, from the customers' side it is very easy to change to another shop and purchase almost everywhere. This results in a very challenging competitive situation. The price elasticity thus plays an important role. In addition, the shop owner can carry out various activities to raise the attractiveness of the shop or of the offered items. Further, several strategical aspects have to be covered.

In this thesis we analyse one specific online shop. We investigate the influence of the basic attributes

- prices of the shop
- prices of the competitors
- advertising

on the quantity sold of one specific product group of this online shop. Some mechanisms like substitution effects and seasonality are considered as well. This will be carried out by applying several statistical methods.

At first, in chapter 2, we summarise the mathematical background of the used methods.

In chapter 3 we explain the shop structure and the data available. Following the typical procedure steps

- data cleaning
- exploratory data analysis
- selecting and fitting models

we derive several models for different items and different levels of aggregation.

Finally, in chapter 4 we put the models into the economical context to validate them for application in business.

Chapter 2

Mathematical Background of Modelling

In this chapter we explain some mathematical background of the methods used for modelling and evaluation. Since this thesis does not deal with extending the theoretical concepts, but with applying the appropriate techniques, in the following the relevant common knowledge is described. Thus, referencing each formula in literature is omitted. Nevertheless, the related literature is listed in section 6.3.

2.1 Generalised linear models (GLM)

The typical regression problem looks as follows: We want to explain a variable of interest, the so called response variable, through several independent variables. The most common approach for such a problem are regression models. Since these are an essential component of this thesis, we give a short introduction into the basic concepts. For more details please refer to Czado et al. (2013).

2.1.1 Multiple linear regression models

Setting up a multiple linear regression model, we start from a set of measured data. Furthermore, we have some structural information available about the general framework. This could for example be the order of the measurements or the conditions under which the data was taken. To be able to interpret things correctly, we are looking for structures within the data. We will now introduce the multiple linear regression model. Please note, that this model is only appropriate for normally distributed response variables. Nevertheless, it plays a central role in the theory of statistical regression and can be seen as starting point for generalised linear regression models, which also allow for a non-normally distributed response.

In multiple linear regression we describe our response variable Y as a linear function of the known predictors with a random error variable ϵ . Written as formula, it is

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \epsilon = \beta_0 + \sum_{j=1}^k \beta_j x_j + \epsilon.$$

The parameters to be estimated are the intercept β_0 as well as the regression coefficients β_1, \dots, β_k . This means, we have $p = k + 1$ regression parameters which have to be estimated from the observations $(y_i, \mathbf{x}_i) = (y_i, x_{i1}, \dots, x_{ik})$, where $i = 1, \dots, n$.

Most of the time, we will refer to this using vector notation.

Definition (Linear model in vector notation)

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \text{ with } \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n), \text{ where}$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdot & \cdot & \cdot & x_{1k} \\ \cdot & x_{21} & \cdot & & & & x_{2k} \\ \cdot & & \cdot & & & & \cdot \\ \cdot & & \cdot & & & & \cdot \\ 1 & x_{n1} & x_{n2} & \cdot & \cdot & \cdot & x_{nk} \end{pmatrix} \in \mathbb{R}^{n \times p}, \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix} \in \mathbb{R}^n,$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k) \in \mathbb{R}^p \text{ and } p = k + 1.$$

Estimation and analysis are based on the following assumptions:

Definition (Assumptions on a linear model)

- **Linearity.** The relationship between the covariate vector \mathbf{x}_i and the random response Y_i is of the form

$$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \epsilon_i,$$

where $i = 1, \dots, n$ and ϵ_i is a random variable that satisfies $E(\epsilon_i) = 0$.

- **Independence.** The random variables ϵ_i are independent.
- **Variance homogeneity.** The random variables ϵ_i have a constant variance

$$\text{Var}(\epsilon_i) = \text{Var}(Y_i) = \sigma^2.$$

- **Normality.** The random variables ϵ_i follow a normal distribution.

The assumption of linearity implies, that the expectation of Y_i is a linear function of the unknown regression parameters β_0, \dots, β_k , i.e.

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}.$$

In vector notation, the mean and the variance of the linear model can be written as

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \text{ and } \text{Var}(\mathbf{Y}) = \sigma^2 \mathbf{I}_n.$$

Please note, that in multiple linear regression the interpretation of the regression coefficients is not that straight forward like in case of simple linear regression, because there are no descriptive two dimensional plots giving the relationship between the covariates and the response variable. In multiple linear regression, we do not have a simple regression line, but a hyperplane \mathbf{H} . This does not change anything for β_0 . It can still be seen as the estimated value of the response variable when all of the covariates are equal to zero. But, regarding the remaining β_j , $j = 1, 2, \dots$, we have to be careful. These have to be interpreted as the change in the expectation $E(\mathbf{Y})$ per unit change of the regressor \mathbf{x}_j , when the remaining covariates are held constant. Thus, the β_j correspond to the gradient of the hyperplane with respect to \mathbf{x}_j , i.e. $\beta_j = \frac{\partial \mathbf{H}}{\partial \mathbf{x}_j}$ for $j = 1, 2, \dots$

Estimation of the regression coefficients

We now briefly discuss how to estimate the regression coefficients, i.e. how to obtain

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k) \in \mathbb{R}^p.$$

Generally speaking, there are two most common techniques: the least squares estimation and the maximum likelihood estimation. It is easy to see, that both methods yield the same results, if the model assumptions of independence, homogeneity and normality are fulfilled:

We have

$$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \epsilon_i, \text{ with } \epsilon_i \sim N(0, \sigma^2) \text{ i.i.d. and } i = 1, \dots, n.$$

For reasons of clarity, we use the following abbreviation for the mean term:

$$\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} := \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}.$$

Thus, we have $Y_i \sim N(\mu_i, \sigma^2)$, which are independent for $i = 1, \dots, n$.

To determine our optimal model parameters, let us start with the maximum likelihood method, where we have to maximise

$$\begin{aligned} \log(L(\boldsymbol{\beta}, \sigma^2 | \mathbf{y})) &= \log \left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \mu_i)^2} \right) \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu_i)^2 \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \end{aligned}$$

The first term on the right hand side does not depend on $\boldsymbol{\beta}$. Hence, it is enough to minimise the second term on the right hand side, which is given by $-\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. Due to the negative sign of this term, we get the maximum likelihood estimate from this minimisation.

As $\frac{1}{2\sigma^2}$ is a constant term, we only have to determine

$$\min_{\boldsymbol{\beta}}((\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})), \text{ where } \boldsymbol{\beta} \in \mathbb{R}^p.$$

In case of ordinary linear regression, minimising the above is exactly what we would do when using the least squares estimation. Please note, that this is not the case for generalised linear models.

Assuming that \mathbf{X} has full column rank, the solution to the above is given by

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\mathbf{y}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y},$$

which is a non-random estimate for $\boldsymbol{\beta}$.

Properties of the estimator

$\hat{\boldsymbol{\beta}}$ is an unbiased estimator for $\boldsymbol{\beta}$:

$$E(\hat{\boldsymbol{\beta}}) = E((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E(\mathbf{Y}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \boldsymbol{\beta}.$$

Concerning the variance of $\hat{\boldsymbol{\beta}}$ we define $\mathbf{A} := (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$, which yields

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \text{Var}(\mathbf{A}\mathbf{y}) = \mathbf{A} \text{Var}(\mathbf{y}) \mathbf{A}^T.$$

Since the ϵ_i are uncorrelated with constant variance σ^2 , so are the y_i . Thus, the variance-covariance matrix of \mathbf{y} is given by $\sigma^2 \mathbf{I}_n$, where \mathbf{I}_n is the $n \times n$ identity matrix. Using the symmetry of $\mathbf{X}^T \mathbf{X}$ then gives us the result. Hence,

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \mathbf{A} \sigma^2 \mathbf{I}_n \mathbf{A}^T = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}.$$

2.1.2 Count regression models

In count regression, we usually have to deal with a response variable, which follows a Poisson distribution or a Negative Binomial distribution. An ordinary linear regression model is no appropriate choice for this kind of regression problem, but we have to switch to generalised linear regression. Such being the case, we will now give a short introduction to the corresponding statistical theory.

As we have seen above, the linear model consists of a random component \mathbf{Y} , a systematic component $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ and a link $\boldsymbol{\eta} = \boldsymbol{\mu}$, which connects the random component with the systematic component.

The general ideas of linear modelling are widely applicable, but for generalised linear models we have to extend these definitions to get appropriate equivalents.

For this purpose, we first have to introduce a new class of probability density functions or probability mass functions in case of discrete random variables, respectively. This is the so called exponential family. It describes a class of densities, which comprises a set of distributions ranging both, continuous and discrete random variables. Many well known

probability distributions like for example the Gaussian distribution, the Bernoulli distribution and the Gamma distribution are members of this family. All these distributions follow a general format.

Definition (Exponential family)

A probability density function or a probability mass function, respectively, is member of the exponential family, if it can be written in the form

$$f(\mathbf{y}|\boldsymbol{\theta}, \Phi) = e^{\frac{\boldsymbol{\theta}\mathbf{y}-b(\boldsymbol{\theta})}{a(\Phi)}+c(\mathbf{y},\Phi)},$$

where

- $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions
- $\Phi > 0$ is a dispersion parameter
- $\boldsymbol{\theta}$ is the canonical parameter.

The dispersion parameter Φ is unique only up to a constant.

The generalised linear model thus is a powerful generalisation of the linear regression to the more general exponential family. The observed data enters the model through a linear function $\mathbf{X}\boldsymbol{\beta}$ and the response is drawn from an exponential family distribution with conditional mean $\boldsymbol{\mu}$.

Next, we have a look at the components of a generalised linear model.

Definition (Components of a generalised linear model)

- **Random component.** The responses Y_i , $1 \leq i \leq n$, are independent and follow a probability density function or a probability mass function $f(\mathbf{y}|\boldsymbol{\theta}, \Phi)$ from the exponential family with a canonical parameter $\boldsymbol{\theta}$ and a dispersion parameter $\Phi > 0$.
- **Systematic component.** As linear predictor we define

$$\eta_i(\boldsymbol{\beta}) = \mathbf{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik},$$

where $\boldsymbol{\beta} = (\beta_0, \dots, \beta_k)^T$ consists of the p unknown regression parameters.

- **Parametric link component.** The relationship between the linear predictor η_i and the mean μ_i of Y_i is defined by the link function $g(\mu_i) = \eta_i(\boldsymbol{\beta}) = \mathbf{x}_i^T \boldsymbol{\beta}$.

If \mathbf{Y} follows an exponential family distribution, the expectation and the variance are defined by

$$E(\mathbf{Y}|\mathbf{X}) = b'(\boldsymbol{\theta}) \text{ and } \text{Var}(\mathbf{Y}) = b''(\boldsymbol{\theta})a(\Phi).$$

Both formulas can be derived using the definition of the probability density function or the probability mass function in exponential family form:

We usually can interchange integration and differentiation, and thus the log-likelihood of an exponential family distribution is given by

$$l(\boldsymbol{\theta}, \Phi | \mathbf{y}) := \log(f(\mathbf{y} | \boldsymbol{\theta}, \Phi)).$$

Replacing the concrete observations \mathbf{y} by the random variable \mathbf{Y} allows to calculate the expectation of $\frac{\partial l}{\partial \boldsymbol{\theta}}$, since we then treat l as a random variable. This gives us

$$E\left(\frac{\partial l(\boldsymbol{\theta}, \Phi | \mathbf{Y})}{\partial \boldsymbol{\theta}}\right) = 0$$

as well as

$$E\left(\frac{\partial^2 l(\boldsymbol{\theta}, \Phi | \mathbf{Y})}{\partial \boldsymbol{\theta}^2}\right) + E\left(\left(\frac{\partial l(\boldsymbol{\theta}, \Phi | \mathbf{Y})}{\partial \boldsymbol{\theta}}\right)^2\right) = 0.$$

Setting in the exponential family representation, we get

$$l(\boldsymbol{\theta}, \Phi | \mathbf{y}) = \frac{\boldsymbol{\theta} \mathbf{y} - b(\boldsymbol{\theta})}{a(\Phi)} + c(\mathbf{y}, \Phi).$$

Further, we have

$$\frac{\partial l(\boldsymbol{\theta}, \Phi | \mathbf{y})}{\partial \boldsymbol{\theta}} = \frac{\mathbf{y} - b'(\boldsymbol{\theta})}{a(\Phi)}$$

and

$$\frac{\partial^2 l(\boldsymbol{\theta}, \Phi | \mathbf{y})}{\partial \boldsymbol{\theta}^2} = -\frac{b''(\boldsymbol{\theta})}{a(\Phi)}$$

for the first and the second derivative with respect to $\boldsymbol{\theta}$.

To derive the expectation, we now have to solve

$$0 = E\left(\frac{\partial l}{\partial \boldsymbol{\theta}}\right) = \frac{\boldsymbol{\mu} - b'(\boldsymbol{\theta})}{a(\Phi)},$$

which gives us

$$E(\mathbf{Y}) = \boldsymbol{\mu} = b'(\boldsymbol{\theta}).$$

For the variance, we solve

$$0 = E\left(\frac{\partial^2 l}{\partial \boldsymbol{\theta}^2}\right) + E\left(\frac{\partial l}{\partial \boldsymbol{\theta}}\right)^2 = -\frac{b''(\boldsymbol{\theta})}{a(\Phi)} + \frac{\text{Var}(\mathbf{Y})}{a(\Phi)^2},$$

which yields

$$\text{Var}(\mathbf{Y}) = b''(\boldsymbol{\theta})a(\Phi).$$

Link function

The link function relates the parameters of the distribution of the response y_i and the covariates $\mathbf{x}_i^T = (x_{i1}, \dots, x_{ik})^T$, where k again denotes the total number of regressors. This relation is created by linking together the mean of Y_i and the linear form of the covariates. This means, we start from

$$\mu_i := E(Y_i | \mathbf{X}_i),$$

which, as we have seen above, is a function of $\boldsymbol{\theta}$. Then, we take this as a linear model of the covariates, i.e.

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta},$$

where $g : \mathbb{R} \rightarrow \mathbb{R}$ is called the link function, \mathbf{x}_i is the vector of regressors for the i -th observation and $\boldsymbol{\beta}$ is the parameter vector.

Definition (Canonical link)

In the case that $g(\mu_i) = \theta_i$ for all $i = 1, \dots, n$, $g(\cdot)$ is called a canonical link function.

An example for a generalised linear model with canonical link is the Poisson-GLM with log link, where we have $g : \mathbb{R} \rightarrow \mathbb{R}$, $g(x) = \log(x)$.

One nice property of canonical link functions is, that these ensure $\boldsymbol{\mu}$ to stay within the range of the response variable.

Poisson Regression

A first approach when modelling count data typically is a Poisson regression. Below, we present the general setup of such a Poisson regression model and discuss its most important elements.

Definition (Poisson regression model with exposure units)

In the model setup for Poisson regression, we have n independent random variables Y_i , where

$$Y_i \sim \text{Poisson}(t_i \mu(\mathbf{x}_i, \boldsymbol{\beta})), \quad i = 1, \dots, n,$$

with

$$P(Y_i = y_i | \mathbf{x}_i) = e^{-t_i \mu(\mathbf{x}_i, \boldsymbol{\beta})} \frac{(t_i \mu(\mathbf{x}_i, \boldsymbol{\beta}))^{y_i}}{y_i!}.$$

In the above formulas,

- $t_i > 0$ is the exposure unit
- $\mu(\mathbf{x}_i, \boldsymbol{\beta}) = e^{\mathbf{x}_i^T \boldsymbol{\beta}} > 0$ is the unit Poisson rate.

To simplify calculations, Stirling's formula can numerically be used for large y_i . This asymptotic formula provides a good approximation of the factorial and is given by

$$n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n.$$

Exposure unit

In the Poisson Regression model presented above, the unit Poisson rate $\mu(\mathbf{x}_i, \boldsymbol{\beta})$ is multiplied by an additional parameter t_i , where $i = 1, \dots, n$. Let us now check, what this additional parameter means and how we can specify it correctly.

If we had to analyse data with varying observation period, we would have to bring this inhomogeneity into the model. Thus, we would need the exposure unit: t_i refers to the respective time period of observation for $\mu(\mathbf{x}_i, \boldsymbol{\beta})$ and hence makes it possible to standardise the unit of time.

In this thesis, the time period of observation is equal for the n samples, which yields $t_i = 1 \forall i$. Thus, the model simplifies to

$$Y_i \sim \text{Poisson}(\mu(\mathbf{x}_i, \boldsymbol{\beta})), \text{ where the } Y_i \text{ are independent for } i = 1, \dots, n.$$

Here, $\mu(\mathbf{x}_i, \boldsymbol{\beta})$ yields the expected number of events for observation i .

Let us now bring the Poisson distribution into the exponential family form:

$$\begin{aligned} P(Y_i = y_i) &= \frac{e^{-t_i \mu(\mathbf{x}_i, \boldsymbol{\beta})} \cdot (t_i \mu(\mathbf{x}_i, \boldsymbol{\beta}))^{y_i}}{y_i!} \\ &= e^{y_i \log(t_i \mu(\mathbf{x}_i, \boldsymbol{\beta})) - t_i \mu(\mathbf{x}_i, \boldsymbol{\beta}) - \log(y_i!)} \\ &= e^{\frac{y_i \log(t_i \mu(\mathbf{x}_i, \boldsymbol{\beta})) - (t_i \mu(\mathbf{x}_i, \boldsymbol{\beta}))}{1} - \log(y_i!)} \end{aligned}$$

This yields

$$\theta_i = \log(t_i \mu(\mathbf{x}_i, \boldsymbol{\beta})) \text{ for the canonical parameter,}$$

as well as

$$b(\theta_i) = e^{\theta_i}$$

and

$$c(y_i, \Phi) = -\log(y_i!).$$

The dispersion parameter Φ is equal to 1, which also gives us $a(\Phi) = 1$. Thus, the dispersion is known and so the Poisson distribution is member of the exponential family.

In case of equal time periods of observation, i.e. if $t_i = 1 \forall i$, the link function connects

$$\mu_i = \mu(\mathbf{x}_i^T \boldsymbol{\beta})$$

and

$$\theta_i = \log(e^{\mu(\mathbf{x}_i, \boldsymbol{\beta})})$$

in the way that

$$\mu_i = \theta_i.$$

Hence, our link function $g(\mu_i)$ is of the form $g(\mu_i) = \theta_i$. As mentioned above, this is also called a canonical link.

Please note, that this does not hold when including an exposure unit $t \neq 1 \forall i$, since we then have

$$g(\mu_i) = \log(\mu_i) - \log(t_i) \neq \theta_i.$$

The term $\log(t_i)$ is the so-called 'offset', which is based on known values.

Negative Binomial Regression

Considering the expectation and the variance of the Poisson distribution, we have the rather strong condition

$$E(\mathbf{Y}) = \text{Var}(\mathbf{Y}) = \boldsymbol{\lambda}.$$

This property, which is also called equidispersion, often is not fulfilled by count data. We then speak of overdispersion, if $E(\mathbf{Y}) < \text{Var}(\mathbf{Y})$, and of underdispersion, if $E(\mathbf{Y}) > \text{Var}(\mathbf{Y})$.

Based on experience, underdispersion is rather seldom, whereas overdispersion appears comparatively often. Hence, in many regression problems concerning count data, we have to deal with $E(\mathbf{Y}) < \text{Var}(\mathbf{Y})$. This tells us, that the variance of \mathbf{Y} appears to grow faster than the Poisson model allows by assuming $E(\mathbf{Y}) = \text{Var}(\mathbf{Y})$. In this situation, a Negative Binomial regression model often is a suitable approach.

To bring the overdispersion into the model, we still assume \mathbf{Y} to follow a Poisson distribution, i.e.

$$\mathbf{Y} \sim \text{Poisson}(\boldsymbol{\lambda}),$$

but now, we take $\boldsymbol{\lambda}$ to be a random variable following a Gamma distribution. Thus, we have the probability function of Y_i conditional on a Gamma distributed random variable $Z_i = z_i > 0$, i.e.

$$Y_i | Z_i = z_i \sim \text{Poisson}(z_i),$$

where

$$Z_i \sim \Gamma(\mu_i, \nu_i) \text{ with } \mu_i = t_i e^{\mathbf{x}_i^T \boldsymbol{\beta}} \text{ and } \nu_i = \Phi \mu_i.$$

The density function of Z_i is defined as

$$f_{Z_i}(z_i) = \frac{1}{\Gamma(\Phi \mu_i)} \left(\frac{\Phi \mu_i z_i}{\mu_i} \right)^{\Phi \mu_i} e^{-\frac{\Phi \mu_i z_i}{\mu_i}} \frac{1}{z_i}.$$

Having

$$\text{Var}(Z_i) = \frac{\mu_i^2}{\Phi\mu_i} = \frac{\mu_i}{\Phi},$$

we get the marginal distribution of Y_i through integration of the conditional distribution $Y_i|Z_i = z_i \sim \text{Poisson}(z_i)$ over z_i . With the transformation $s_i = z_i(1 + \Phi)$ it follows

$$\begin{aligned} f(y_i|\mu_i, \Phi) &= \int_0^\infty \frac{e^{-z_i}(z_i)^{y_i}}{y_i!} \frac{1}{\Gamma(\Phi\mu_i)} (\Phi z_i)^{\Phi\mu_i} e^{-\Phi z_i} \frac{1}{z_i} dz_i \\ &= \frac{\Phi^{\Phi\mu_i}}{\Gamma(\Phi\mu_i)y_i!} \int_0^\infty e^{-z_i(1+\Phi)} z_i^{(y_i+\Phi\mu_i-1)} dz_i \\ &= \frac{\Phi^{\Phi\mu_i}}{\Gamma(\Phi\mu_i)y_i!} \int_0^\infty e^{-s_i} \frac{1}{(1+\Phi)^{y_i+\Phi\mu_i-1}} s_i^{y_i+\Phi\mu_i-1} \frac{1}{1+\Phi} ds_i \\ &= \frac{\Phi^{\Phi\mu_i}}{\Gamma(\Phi\mu_i)y_i!(1+\Phi)^{y_i+\Phi\mu_i}} \int_0^\infty e^{-s_i} s_i^{y_i+\Phi\mu_i-1} ds_i \\ &= \frac{\Phi^{\Phi\mu_i}\Gamma(y_i+\Phi\mu_i)}{\Gamma(\Phi\mu_i)y_i!(1+\Phi)^{y_i+\Phi\mu_i}} \\ &= \frac{\Gamma(y_i+\Phi\mu_i)}{\Gamma(\Phi\mu_i)\Gamma(y_i+1)} \left(\frac{1}{1+\frac{1}{\Phi}}\right)^{\Phi\mu_i} \left(\frac{\frac{1}{\Phi}}{1+\frac{1}{\Phi}}\right)^{y_i}. \end{aligned}$$

Thus, the marginal distribution of Y_i is just given by the Negative Binomial distribution

$$Y_i \sim \text{NegBin}(a_i, b_i),$$

where $a_i = \Phi\mu_i$ and $b_i = \frac{1}{1+\frac{1}{\Phi}}$.

Setting $\Phi = \infty$, this converges to the Poisson distribution.

2.1.3 Parameter estimation

In this section, we want to get an idea of how to estimate the regression coefficients for a generalised linear model. Furthermore, we explain how to assess the quality and the accuracy of our gained estimated values. This is the basis for interpreting regression results correctly, which is essential to reach our goal of best possible prediction of future values.

Whereas in ordinary multiple linear regression the coefficients are mainly gained through least squares estimation, this is no appropriate choice in case of generalised linear models.

This is due to the fact, that in least square estimation we aim at minimising the sum of the squared errors. We assume, that the error terms are independent from each other and follow a normal distribution with expectation equal to zero and a constant variance. So we assume that there is no systematic information within the error terms.

Maximum likelihood estimation in turn starts from the assumption, that there is some stochastic dependence present within the measured values. These are taken to follow a certain distribution, but with unknown parameters. The regression coefficients now are estimated by maximising the probability of obtaining the observed values. In case of Gaussian distributed data, i.e. in case of linear regression, this yields the same results as the

least squares estimation, which we have already seen above. But this is not the case for generalised linear models, because we here allow for non-normally distributed data.

Since it is absolutely essential for maximum likelihood estimation to know the distribution characteristics, we restrict to distributions coming from the exponential family when performing generalised linear models.

Summarising the above, maximum likelihood estimation yields a model based estimation of the regression parameters. In the following, we discuss how these estimates are obtained and have a closer look at their properties.

Let $f(\mathbf{y}|\boldsymbol{\theta})$ denote the known probability density or probability mass function, respectively, which comes from the exponential family. Assume, that it is an appropriate choice to describe the distribution of our random vector $\mathbf{Y} = (Y_1, \dots, Y_n)$. Furthermore, assume that we have a full rank design matrix \mathbf{X} . For a fixed realisation of \mathbf{y} , $f(\mathbf{y}|\boldsymbol{\theta})$ is called the likelihood function, which will be denoted by $L(\boldsymbol{\theta})$.

In generalised linear regression we only consider independent random variables Y_1, Y_2, \dots, Y_n with probability density function or probability mass function $f(y_i|\mathbf{x}_i, \boldsymbol{\theta})$ for given covariates $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, which yields

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n f(y_i|\mathbf{x}_i, \boldsymbol{\theta}) := f(\mathbf{x}_1, \dots, \mathbf{x}_n|\boldsymbol{\theta}).$$

The maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ for the unknown parameter $\boldsymbol{\theta}$ now is chosen such that

$$L(\hat{\boldsymbol{\theta}}) \geq L(\boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \in \Theta,$$

where Θ describes the set containing all valid values for $\boldsymbol{\theta}$.

To ease calculations, we take the natural logarithm of our likelihood function. Due to the strict monotonicity, this yields the same maximisation results. In the following, we thus consider

$$l(\boldsymbol{\theta}) = \log(L(\boldsymbol{\theta})) = \sum_{i=1}^n \log(f(y_i|\mathbf{x}_i, \boldsymbol{\theta})) = \sum_{i=1}^n l_i(\boldsymbol{\theta}).$$

To gain our estimated parameter $\hat{\boldsymbol{\beta}}$ out of the above, we first have to bring $\boldsymbol{\beta}$ into the model and then solve

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n \frac{\partial \ln(f(y_i|\mathbf{x}_i, \boldsymbol{\theta}))}{\partial \boldsymbol{\beta}} \stackrel{!}{=} 0.$$

For this purpose, we start from a distribution of the exponential family given in the form

$$f(y_i|\boldsymbol{\theta}, \Phi) = e^{\frac{\theta_i y_i}{a(\Phi)} + c(y_i, \Phi)}$$

and want to get a likelihood function of the form $L(\boldsymbol{\beta})$.

To bring $\boldsymbol{\beta}$ into the model, we use the properties of the components of a generalised linear model: As we have seen above, $\theta_i = \theta_i(\mu_i)$ and the link function $g(\cdot)$ connects the linear predictor η_i and the mean value μ_i in the sense that $g(\eta_i) = \mu_i$, where η_i is given as $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$. This gives

$$L(\boldsymbol{\beta}) = \prod_{i=1}^n e^{\frac{\theta_i(\mathbf{x}_i^T \boldsymbol{\beta}) y_i - b(\theta_i(\mathbf{x}_i^T \boldsymbol{\beta}))}{a(\Phi)} + c(y_i, \Phi)}$$

for the likelihood function, which yields

$$l(\boldsymbol{\beta}) = \sum_{i=1}^n \left(\frac{\theta_i(\mathbf{x}_i^T \boldsymbol{\beta}) y_i - b(\theta_i(\mathbf{x}_i^T \boldsymbol{\beta}))}{a(\Phi)} + c(y_i, \Phi) \right).$$

Taking the partial derivative $\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \in \mathbb{R}^p$ now results in p equations which depend on $\boldsymbol{\beta}$ in a non-linear manner. The vector $\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ of the derivatives of the log-likelihood function with respect to $\boldsymbol{\beta}$ is also called the score vector. This will be denoted by $s(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$. The maximum likelihood equations thus are $s(\boldsymbol{\beta}) = 0$. An analytical solution does not exist. For this reason, the maximum likelihood estimate usually is obtained using the Fisher scoring method. To see how this works, we need a few definitions.

Definition (Weights in generalised linear models)

The weights in a generalised linear model are defined as

$$W_i := W_i(\boldsymbol{\beta}) := \left(\frac{d\mu_i}{d\eta_i} \right)^2 / b''(\theta_i).$$

Definition (Unscaled score equations of a generalised linear model)

The equations defined by

$$s_j(\boldsymbol{\beta}, \mathbf{y}) := \sum_{i=1}^n W_i (y_i - \mu_i) \frac{d\eta_i}{d\mu_i} x_{ij} = 0, \text{ where } j = 1, \dots, p,$$

are called the unscaled score equations.

Definition (Unscaled Hessian matrix and Fisher information in a generalised linear model)

The unscaled Hessian matrix is defined as

$$\mathbf{H} := \mathbf{H}(\boldsymbol{\beta}, \mathbf{y}) := \frac{\partial s(\boldsymbol{\beta}, \mathbf{y})}{\partial \boldsymbol{\beta}},$$

where $s(\boldsymbol{\beta}, \mathbf{y})$ denotes the p -dimensional vector with components $s_j(\boldsymbol{\beta}, \mathbf{y})$ as defined above.

The unscaled Fisher information matrix is given by

$$\mathbf{I} := \mathbf{I}(\boldsymbol{\beta}) := \mathbf{E}(-\mathbf{H}(\boldsymbol{\beta}, \mathbf{Y})).$$

Definition (Unscaled Fisher information of a generalised linear model)

The (r, s) -th element $\mathbf{I}_{r,s}$ of the Fisher information is given by

$$\mathbf{I}_{r,s} = \sum_{i=1}^n \frac{1}{b''(\theta_i)} \left(\frac{d\mu_i}{d\eta_i} \right)^2 x_{is} x_{ir} = \sum_{i=1}^n W_i x_{is} x_{ir}.$$

Defining the diagonal matrix $\mathbf{W} := \mathbf{W}(\boldsymbol{\beta})$ with the i -th diagonal element given by $W_i(\boldsymbol{\beta})$, the Fisher information can be written as

$$\mathbf{I}(\boldsymbol{\beta}) = \mathbf{X}^T \mathbf{W}(\boldsymbol{\beta}) \mathbf{X},$$

where $\mathbf{X}^T := (\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathbb{R}^{p \times n}$.

To understand how the Fisher scoring algorithm works, let us summarise some useful properties:

We have already seen, that

$$\boldsymbol{\mu} := E(\mathbf{Y}) = b'(\boldsymbol{\theta}) \text{ and } \text{Var}(\mathbf{Y}) = a(\Phi) \cdot b''(\boldsymbol{\theta}).$$

Furthermore, we have

$$\frac{\partial \theta_i}{\partial \mu_i} = \left(\frac{\partial \mu_i}{\partial \theta_i} \right)^{-1} = \left(\frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} \right)^{-1} = \frac{1}{b''(\theta_i)}$$

and

$$\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \frac{\partial \mu_i}{\partial g(\mu_i)} \cdot \frac{\partial g(\mu_i)}{\partial \boldsymbol{\beta}} = \left(\frac{\partial g(\mu_i)}{\partial \mu_i} \right)^{-1} \cdot \frac{\partial x_i^T \boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = \left(\frac{\partial g(\mu_i)}{\partial \mu_i} \right)^{-1} \cdot x_i.$$

Having this at hand, we can easily derive the maximum likelihood equations, which calculate as

$$\begin{aligned} \frac{\partial l}{\partial \boldsymbol{\beta}} &= \frac{1}{a(\Phi)} \cdot \sum_{i=1}^n \left(y_i \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} - \frac{\partial b(\theta_i)}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} \right) \\ &= \frac{1}{a(\Phi)} \cdot \sum_{i=1}^n (y_i - \mu_i) \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} \\ &= \frac{1}{a(\Phi)} \cdot \sum_{i=1}^n (y_i - \mu_i) \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \\ &= \frac{1}{a(\Phi)} \cdot \sum_{i=1}^n \frac{(y_i - \mu_i)}{b''(\boldsymbol{\theta}) \cdot \frac{\partial g(\mu_i)}{\partial \mu_i}} \cdot \mathbf{x}_i \\ &= \frac{1}{a(\Phi)} \cdot \sum_{i=1}^n (y_i - \mu_i) \cdot W_i \cdot \frac{\partial g(\mu_i)}{\partial \mu_i} \cdot x_i, \end{aligned}$$

where the W_i are the weights defined above.

In matrix notation, this is

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{1}{a(\Phi)} \cdot \mathbf{X}^T \mathbf{W} \boldsymbol{\Delta} (\mathbf{y} - \boldsymbol{\mu}),$$

where $\boldsymbol{\mu}$ is given by $\boldsymbol{\mu} := (\mu_1, \dots, \mu_n)^T$, \mathbf{X} is the design matrix, and \mathbf{W} is defined as $\mathbf{W} := \text{diag}(\mathbf{W}_i)$, with

$$W_{ij} = \begin{cases} W_i, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}.$$

For $\boldsymbol{\Delta}$, we have

$$\boldsymbol{\Delta} := \text{diag}\left(\frac{\partial g(\mu_i)}{\partial \mu_i}\right).$$

All in all, the maximum likelihood functions now are of the form

$$\mathbf{X}^T \mathbf{W} \Delta \mathbf{y} = \mathbf{X}^T \mathbf{W} \Delta \boldsymbol{\mu}.$$

Unfortunately, these equations are typically highly non-linear in $\boldsymbol{\beta}$. Such non-linear equations are commonly solved using the iterative Newton-Raphson algorithm. In general, this works as follows:

For a system of non-linear equations $f(\mathbf{X}) = (f_1(x_1, \dots, x_n), \dots, f_n(x_1, \dots, x_n)) = 0$ and an initial value $\mathbf{X}^{(0)} \in \mathbb{R}^p$ close to the solution, the set of all solutions is given by

$$\mathbf{x}^{(m+1)} := \mathbf{x}^{(m)} - (Df(\mathbf{x}^{(m)}))^{-1} f(\mathbf{x}^{(m)}),$$

where $Df(\mathbf{x})$ denotes the $p \times p$ Jacobi-matrix of $f(\mathbf{x})$, which is given by

$$Df(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdot & \cdot & \cdot & \frac{\partial f_1}{\partial x_p} \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ \frac{\partial f_n}{\partial x_1} & \cdot & \cdot & \cdot & \frac{\partial f_n}{\partial x_p} \end{pmatrix} \in \mathbb{R}^{p \times p}$$

Coming back to our GLM, we want so solve $\frac{\partial l}{\partial \boldsymbol{\beta}} = 0$. Thus, to be able to apply the Newton-Raphson algorithm, we need the Hessian matrix of the log-likelihood of our model. It is given by

$$\mathbf{H} := \frac{\partial^2 l}{\partial \boldsymbol{\beta}^2} = \begin{pmatrix} \frac{\partial^2 l}{\partial \beta_1 \partial \beta_1} & \cdot & \cdot & \cdot & \frac{\partial^2 l}{\partial \beta_1 \partial \beta_p} \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ \frac{\partial^2 l}{\partial \beta_p \partial \beta_1} & \cdot & \cdot & \cdot & \frac{\partial^2 l}{\partial \beta_p \partial \beta_p} \end{pmatrix} \in \mathbb{R}^{p \times p},$$

where the (r, s) -th element of \mathbf{H} calculates as

$$\begin{aligned} \frac{\partial^2 l}{\partial \beta_r \partial \beta_s} &= \frac{\partial}{\partial \beta_s} \left(\frac{1}{a(\Phi)} \cdot \sum_{i=1}^n \frac{(y_i - \mu_i) \cdot x_{ir}}{b''(\theta_i) \cdot \frac{\partial g(\mu_i)}{\partial \mu_i}} \right) \\ &= \frac{1}{a(\Phi)} \cdot \left(\sum_{i=1}^n (y_i - \mu_i) \cdot \frac{\partial}{\partial \beta_s} \left(\frac{x_{ir}}{b''(\theta_i) \cdot \frac{\partial g(\mu_i)}{\partial \mu_i}} \right) + \sum_{i=1}^n \frac{x_{ir}}{b''(\theta_i) \cdot \frac{\partial g(\mu_i)}{\partial \mu_i}} \cdot \frac{\partial}{\partial \beta_s} (y_i - \mu_i) \right) \\ &= \frac{1}{a(\Phi)} \cdot \left(\sum_{i=1}^n (y_i - \mu_i) \cdot \frac{\partial}{\partial \beta_s} \left(\frac{x_{ir}}{b''(\theta_i) \cdot \frac{\partial g(\mu_i)}{\partial \mu_i}} \right) + \sum_{i=1}^n \frac{x_{ir}}{b''(\theta_i) \cdot \frac{\partial g(\mu_i)}{\partial \mu_i}} \cdot \left(-\frac{1}{\frac{\partial g(\mu_i)}{\partial \mu_i} \cdot x_{is}} \right) \right) \\ &= \frac{1}{a(\Phi)} \cdot \left(\sum_{i=1}^n (y_i - \mu_i) \cdot \frac{\partial}{\partial \beta_s} \left(\frac{x_{ir}}{b''(\theta_i) \cdot \frac{\partial g(\mu_i)}{\partial \mu_i}} \right) - \sum_{i=1}^n W_i x_{ir} x_{is} \right). \end{aligned}$$

This in general depends on y_i . To overcome this problem, we modify the Newton-Raphson method by using the expected unscaled Hessian matrix, which is called the Fisher scoring method.

Taking expectations yields

$$E\left(\frac{\partial^2 l}{\partial \beta_r \partial \beta_s}\right) = -\frac{1}{a(\Phi)^2} \cdot \sum_{i=1}^n W_i x_{ir} x_{is},$$

which in matrix notation is given as

$$E\left(-\frac{\partial^2 l}{\partial \boldsymbol{\beta}^2}\right) = \frac{1}{a(\Phi)^2} \mathbf{X}^T \mathbf{W} \mathbf{X} \in \mathbb{R}^{\mathbf{p} \times \mathbf{p}}.$$

Using the above, the iterative estimating procedure modifies to

$$\boldsymbol{\beta}^{(m+1)} := \boldsymbol{\beta}^{(m)} + (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \Delta(\mathbf{y} - \boldsymbol{\mu}),$$

where \mathbf{W} , Δ and $\boldsymbol{\mu}$ are evaluated at $\boldsymbol{\beta}^{(m)}$ and (m) denotes the m -th iteration step.

Defining $\mathbf{A} := (\mathbf{X}^T \mathbf{W} \mathbf{X})$, this can be written as

$$\mathbf{A} \boldsymbol{\beta}^{(m+1)} = \mathbf{A} \boldsymbol{\beta}^{(m)} + \frac{1}{a(\Phi)^2} \cdot \frac{\partial l}{\partial \boldsymbol{\beta}^{(m)}}.$$

For the j -th element of the left hand side, we now have

$$\begin{aligned} (\mathbf{A} \boldsymbol{\beta}^{(m+1)})_j &= \sum_{s=1}^p A_{js} \beta_s^{(m)} + \sum_{i=1}^n W_i (y_i - \mu_i) \frac{\partial g(\mu_i)}{\partial \mu_i} \cdot x_{ij} \\ &= \sum_{s=1}^p \sum_{i=1}^n W_i x_{is} x_{ij} \beta_s^{(m)} + \sum_{i=1}^n W_i (y_i - \mu_i) \frac{\partial g(\mu_i)}{\partial \mu_i} \cdot x_{ij} \\ &= \sum_{i=1}^n W_i x_{ij} \left(\sum_{s=1}^p x_{is} \beta_s^{(m)} + (y_i - \mu_i) \frac{\partial g(\mu_i)}{\partial \mu_i} \right) \\ &= \sum_{i=1}^n W_i x_{ij} \left(g(\mu_i) + (y_i - \mu_i) \frac{\partial g(\mu_i)}{\partial \mu_i} \right) \\ &= \sum_{i=1}^n W_i x_{ij} z_i, \end{aligned}$$

where $z_i := g(\mu_i) + (y_i - \mu_i) \frac{\partial g(\mu_i)}{\partial \mu_i}$.

$(\mathbf{A} \boldsymbol{\beta}^{(m+1)})$ alternatively can be written as

$$\begin{aligned} (\mathbf{A} \boldsymbol{\beta}^{(m+1)})_j &= \sum_{s=1}^p A_{js} \beta_s^{(m+1)} \\ &= \sum_{s=1}^p \left(\sum_{i=1}^n W_i x_{is} x_{ij} \right) \beta_s^{(m+1)} \\ &= \sum_{i=1}^n W_i x_{ij} g^*(\mu_i), \end{aligned}$$

where $g^*(\mu_i) := g(\mu_i) \big|_{\boldsymbol{\beta} = \boldsymbol{\beta}^{(m+1)}}$.

Thereof, we get

$$\sum_{i=1}^n W_i x_{ij} z_i = \sum_{i=1}^n W_i x_{ij} g^*(\mu_i), \text{ with } j = 1, \dots, p.$$

This now is equivalent to the maximum likelihood.

Summarising the above, the method of the Fisher scoring can be described as follows:

Algorithm (Fisher scoring for generalised linear model estimation of β)

1. Choose starting values $\beta^{(0)}$ and a threshold value $\delta > 0$.
2. For $r \geq 0$ define

$$\beta^{(r+1)} := \beta^{(r)} + \mathbf{I}^{-1}(\beta^{(r)}) \mathbf{s}(\beta^{(r)}, \mathbf{y}),$$

where $\mathbf{s}(\beta, \mathbf{y}) := (s_1(\beta, \mathbf{y}), \dots, s_p(\beta, \mathbf{y}))^T$ is the vector of score functions and \mathbf{I} is the Fisher Information.

3. If $\|\beta^{(r+1)} - \beta^{(r)}\| < \delta$, stop the algorithm. The maximum likelihood estimate of β , which is denoted by $\hat{\beta}$, then is set to $\beta^{(r+1)}$.

Asymptotic normality of the maximum likelihood estimate

Our data sample (y_1, \dots, y_n) is independent, but not identically distributed. Hence, there does not exist an analytical solution for the maximum likelihood estimator $\hat{\theta} = \hat{\beta}$ and we only can deduce asymptotic properties. For this reason, we need appropriate regularity conditions for our maximum likelihood estimators. Thus, we extend our definition of the generalised linear model. For a given sample size n , the maximum likelihood estimate $\hat{\beta}$ will be denoted by $\hat{\beta}_n$ and the corresponding log-likelihood by $l_n(\beta, \Phi|\mathbf{y})$. The regularity conditions then are defined as follows:

Definition (Regularity conditions for asymptotic normality of the maximum likelihood estimators)

- The observation specific dispersion function $a_i(\Phi)$ satisfies $a_i(\Phi) = \frac{\Phi}{W_i}$ for some known and bounded weight W_i , where $i = 1, \dots, n$.
- The set of unknown true regression parameters $\boldsymbol{\beta}$ is an open subset of \mathbb{R}^p .
- For the function $\phi(\boldsymbol{\eta}) := (\psi'(\boldsymbol{\eta}))^2 b''(\psi(\boldsymbol{\eta}))$, where $\psi(\boldsymbol{\eta}) := h(g^{-1}(\boldsymbol{\eta}))$ gives the current canonical parameters, it holds

$$0 < \inf_{i=1, \dots, n} (\phi(\mathbf{x}_i^T \boldsymbol{\beta})) \leq \sup_{i=1, \dots, n} (\phi(\mathbf{x}_i^T \boldsymbol{\beta})) < \infty.$$

- The covariates satisfy the conditions

$$\max_{i=1, \dots, n} (\mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i) \rightarrow 0 \text{ for } n \rightarrow \infty$$

and

$$\lambda_{\min}(\mathbf{X}^T \mathbf{X}) \rightarrow \infty \text{ for } n \rightarrow \infty,$$

where $\lambda_{\min}(\mathbf{A})$ is the minimal eigenvalue of the matrix \mathbf{A} .

Based on this, we define the asymptotic normality of the maximum likelihood estimator in a generalised linear model as follows:

Definition (Asymptotic normality of the maximum likelihood estimator in generalised linear models)

Under the regularity conditions just defined above, it follows that

$$\hat{\boldsymbol{\beta}}_{\mathbf{n}} \rightarrow \boldsymbol{\beta} \text{ in probability for } n \rightarrow \infty$$

and

$$\left(\text{Cov} \left(\frac{\partial l_n(\boldsymbol{\beta}; \Phi | \mathbf{Y})}{\partial \boldsymbol{\beta}} \right) \right)^{\frac{1}{2}} (\hat{\boldsymbol{\beta}}_{\mathbf{n}} - \boldsymbol{\beta}) \rightarrow N_p(\mathbf{0}, \mathbf{I}_p) \text{ in distribution, as } n \rightarrow \infty.$$

The covariance matrix of $\frac{\partial l_n(\boldsymbol{\beta}; \Phi | \mathbf{Y})}{\partial \boldsymbol{\beta}}$ then is defined as

$$\text{Cov} \left(\frac{\partial l_n(\boldsymbol{\beta}; \Phi | \mathbf{Y})}{\partial \boldsymbol{\beta}} \right) = \mathbf{X}^T \mathbf{D}(\boldsymbol{\beta}) \mathbf{X},$$

where $\mathbf{D}(\boldsymbol{\beta})$ is a diagonal matrix with the i -th element given by

$$D_i(\boldsymbol{\beta}) := \frac{W_i(\boldsymbol{\beta})}{a_i(\Phi)} = \frac{w_i}{\Phi} \cdot W_i(\boldsymbol{\beta}).$$

2.1.4 Asymptotic hypothesis tests in generalised linear models

Statistical hypothesis tests help on checking if we have found an appropriate estimator $\hat{\beta}$. In general, we test:

$$H_0 : \mathbf{C}\beta = \psi \text{ versus } H_1 : \mathbf{C}\beta \neq \psi,$$

where \mathbf{C} is taken to be a full rank matrix of rank $q \leq p$.

As an example, let us consider the most simple case

$$H_0 : \beta_j = 0 \text{ versus } H_1 : \beta_j \neq 0,$$

which is an important special case of our general test problem.

This is, we test the null hypothesis, which states that the j -th covariate does not have any influence on the response variable, against the alternative hypothesis, that it does have a remarkable effect. Hence, we test for the significance of the effects belonging to β_j . This is done by comparing the model without β_j , i.e. with $\beta_j = 0$, to the full model. The most well known technique that deals with this test problem probably is the likelihood ratio test.

Likelihood ratio test

The likelihood ratio test compares two models, where one of them, the so called 'null model', is a special case of the other one, which is referred to as 'alternative model'. The null model of course represents the model under the null hypothesis, which is a reduced form of the alternative model. The test is based on the likelihood ratio, which indicates how many times more likely the data is to appear under the null model than under the alternative model.

The likelihood ratio statistic is defined as

$$\Lambda = 2(L(\hat{\beta}) - L(\hat{\beta}^*)).$$

Thus, the maximum log-likelihood estimate $L(\hat{\beta})$ of the alternative or 'full' model is compared to the maximum likelihood estimate $L(\hat{\beta}^*)$ of the null model, where the test hypothesis can be formulated as

$$H_0 : \mathbf{C}\hat{\beta}^* = \psi \text{ versus } H_1 : \mathbf{C}\hat{\beta}^* \neq \psi.$$

If the null hypothesis is true, then in the large sample limit it is

$$2(L(\hat{\beta}) - L(\hat{\beta}^*)) \sim \chi_{p_1 - p_0}^2,$$

where p_1 is the number of parameters β_i in the full model and p_0 the number of parameters β_i^* in the null model, respectively.

If the maximum log-likelihood $L(\hat{\beta})$ is significantly larger than $L(\hat{\beta}^*)$, we get a large likelihood ratio statistic. In this case H_0 will be rejected in favour of H_1 .

2.2 Generalised additive models (GAM)

For generalised linear models (GLMs) a linear or at least some other parametric relationship between the covariates and the response variable is required. If the exploratory data analysis already indicates, that this condition is not met by some of the covariates, i.e. if we are not able to cover the non-linearities by appropriate simple transformations or by changing the covariates, we try to improve our models by replacing the linear terms of the covariates with unspecified smooth functions, which are estimated using penalised regression splines.

A generalised additive model (GAM) is a generalised linear model (GLM), where the linear predictor contains a sum of smooth functions of some of the covariates. The model thus look as follows:

$$g(\mu_i) = X_i^* \beta^* + s_1(x_{1i}) + s_2(x_{2i}) + \dots$$

with

$\mu_i \equiv E(Y_i)$ and $Y_i \sim$ a specific member of the exponential family distribution.

In the above notation,

- Y_i is the response variable
- X_i^* is a row of the design matrix for any strictly parametric model components
- β^* is the vector of the regression coefficients corresponding to X_i^*
- s_j are the smooth functions of the covariates x_k .

The smooth functions s_j have to be estimated in a way, that

$$g(\mu_i) = X_i^* \beta^* + s_1(x_{1i}) + s_2(x_{2i}) + \dots$$

becomes a linear model. For this purpose, we define a basis, which spans the space of functions of which the s_j , or at least a close approximation to the s_j , is an element. This is best explained restricting to the univariate case:

Consider a model with one smooth function for one covariate, i.e.

$$\mathbf{y} = s(\mathbf{x}) + \boldsymbol{\epsilon},$$

where

- \mathbf{y} is the response variable
- \mathbf{x} is the covariate
- $\boldsymbol{\epsilon}$ is an i.i.d. $N(0, \sigma^2)$ random variable.

For reasons of understandability, we restrict to $\mathbf{x} \in [0, 1]^n$ in the explanation.

So, let $b_i(x)$ denote the i -th basis function of the above mentioned basis. The smooth function s then can be written as

$$s(x) = \sum_{i=1}^q b_i(x) \cdot \beta_i.$$

The linear predictor thus forecasts some known smooth function of the expectation of the response variable, where the response variable follows an arbitrary distribution belonging to the exponential family. The generalised additive model will then be fitted using penalised likelihood maximisation. This penalisation is needed to avoid complex overfitting, which would result from allowing for any smooth function. So, using this penalised likelihood maximisation, we try to balance between penalising the 'wiggleness' and the 'badness of fit'.

For further details on that, please refer to Wood(2006).

2.3 Exploratory data analysis

Typically, a model will not fit perfectly. Thus, depending on personal preferences and on modelling purposes, we have to select a model, which fits best to our research question. We typically suggest a set of models, which may possibly fit well to the data, and then check them against each other by appropriate measures of goodness of fit. The quality of the final model thus strongly depends on the quality of the set of suggested ones, because the best fit typically just is the best fitting model out of this set. This already explains the importance of the exploratory data analysis: We look for dependency structures within the data and check up to what extent the model assumptions of a linear model are fulfilled. This avoids problems during model fitting and assures a certain quality of the models.

The following explanations are mainly based on Fahrmeir et al. (2001). For a more detailed introduction to the exploratory data analysis please refer to Tukey (1977).

2.3.1 Dependence measures

Correlation coefficients help making out dependencies among the variables. The *Pearson product-moment correlation coefficient*, which is also known as r , R or *Pearson's r* , probably is the most popular one, but, using this correlation coefficient, only linear relationships can be detected.

To bring out also non-linear but monotonous relationships, rank based correlation coefficients can be used. These measure, how well the dependence structure between two variables can be described by an arbitrary monotonous function without knowing the probability distribution of the variables. Furthermore, rank based correlation coefficients are robust against outliers. Two classical examples are the *Kendall's τ* and the *Spearman's ρ* .

The Spearman's ρ is very similar to the Pearson's r . It calculates the difference between the ranks, which makes it a special case of the Pearson's r by converting the data into ranks. The Kendall's τ in contrast does not calculate the differences between the ranks, but it quantifies the difference between the percentage of concordant and discordant pairs among all possible pairwise events. This makes it appear less direct than the Spearman's ρ . For this reason, the Spearman's ρ is mostly used in place of usual linear correlation when working with integer valued scores on a measurement scale, when there is a moderate number of possible scores or when we don't want to make rely on the assumptions about the bivariate relationship. To check whether one has detected a linear or a non-linear relationship using the Spearman's ρ , a nice possibility would be to calculate both, the Spearman's ρ and the Pearson's r , and then look at the difference between them: the bigger it is, the less likely is a linear relationship.

Pearson product-moment correlation

This dimensionless correlation coefficient measures the linear relationship between two variables. It is only suitable for at least approximately normally distributed random variables.

Definition (Pearson product-moment correlation)

Let X and Y be two at least approximately normally distributed random variables with independently sampled observations (x_i, y_i) , $i = 1, \dots, n$. Then, the Pearson product-moment correlation coefficient calculates as

$$r = r_{X,Y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\tilde{s}_{\mathbf{x},\mathbf{y}}}{\tilde{s}_{\mathbf{x}}\tilde{s}_{\mathbf{y}}},$$

where

$$\tilde{s}_{\mathbf{x}} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \text{ and } \tilde{s}_{\mathbf{y}} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}$$

are the standard deviations of x and y , respectively, and

$$\tilde{s}_{\mathbf{x},\mathbf{y}} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

is the empirical covariance.

Let us put this into words: $\tilde{s}_{\mathbf{x},\mathbf{y}}$ accrues from summing up the multiplied deviations of \mathbf{x} and \mathbf{y} from the respective means. The term $\tilde{s}_{\mathbf{x}}\tilde{s}_{\mathbf{y}}$ brings the dispersion into the formula and normalises the correlation coefficient. Due to this normalisation the Pearson's r can only take values between -1 and 1 .

Point clouds concentrated on the first and the third quadrant yield positive correlation, whereas point clouds concentrated on the second and the fourth quadrant yield negative correlation. For randomly scattered point clouds the Pearson's r takes values close to 0 .

Proposition (Properties of the Pearson product-moment correlation)

- r is symmetric
- $-1 \leq r \leq 1$
- if there is no linear correlation it holds $r = 0$
- $r = 1$ in case of a perfect positive correlation
- $r = -1$ in case of a perfect negative correlation.

For the Pearson product-moment correlation, we typically use the following classification scheme:

- $|r| < 0.5$ means weak correlation
- $0.5 < |r| < 0.8$ indicates a medium strong correlation
- $0.8 \leq |r|$ occurs in case of a strong correlation.

Spearman's ρ

In contrast to the Pearson's r , the Spearman's ρ can also be applied if the variables are not normally distributed and even if the distribution is not known at all.

Definition (Spearman's ρ)

Let X and Y be two random variables with n independently sampled observations (x_i, y_i) , $i = 1, \dots, n$, and $(rg(x_i), rg(y_i))$, $i = 1, \dots, n$, the pairs of ranks. The Spearman's ρ then is defined as

$$r_{sp} = \frac{\sum_{i=1}^n (rg(x_i) - \bar{rg}_x)(rg(y_i) - \bar{rg}_y)}{\sqrt{\sum_{i=1}^n (rg(x_i) - \bar{rg}_x)^2 \sum_{i=1}^n (rg(y_i) - \bar{rg}_y)^2}}$$

where the means of rg_x and rg_y are given by

$$\bar{rg}_x = \frac{1}{n} \sum_{i=1}^n rg(x_i) = \frac{1}{n} \sum_{i=1}^n i = \frac{n+1}{2} \text{ and } \bar{rg}_y = \frac{1}{n} \sum_{i=1}^n rg(y_i) = \frac{1}{n} \sum_{i=1}^n i = \frac{n+1}{2}.$$

Obviously, the Spearman's ρ corresponds to the Pearson's r for the pairs of ranks. But, since for the calculation only the rank of the values is required, the Spearman's ρ can also be applied to variables on ordinal scale.

Proposition (Properties of the Spearman's ρ)

- $-1 \leq r_{sp} \leq 1$
- if there is no monotone correlation it holds $r_{sp} = 0$
- $r_{sp} > 0$ means a concordant monotone correlation
- $r_{sp} < 0$ indicates a discordant monotone correlation
- $r = 1$ in case of a perfect concordance
- $r = -1$ in case of a perfect discordance.

Kendall's τ

In this thesis, the Kendall's τ will mainly be used in the context of copulas, when analysing the dependence structure among the model residuals.

Definition (Kendall's τ)

Let (X, Y) and (X^*, Y^*) be independent pairs of continuous random variables, each with a bivariate cumulative distribution function F . The Kendall's τ then is defined as

$$\tau_{X,Y} = P((X - X^*)(Y - Y^*) > 0) - P((X - X^*)(Y - Y^*) < 0) = 2P((X - X^*)(Y - Y^*) > 0) - 1.$$

The Kendall's τ thus calculates as the difference of the probability that the pairs (X, Y) and (X^*, Y^*) are concordant, i.e. that either $X > X^*$ and $Y > Y^*$ or $X < X^*$ and $Y < Y^*$, and the probability, that (X, Y) and (X^*, Y^*) are discordant, i.e. $X > X^*$ and $Y < Y^*$ or $X < X^*$ and $Y > Y^*$.

Besides this, there also exists an empirical version of Kendall's τ .

Definition (Empirical Kendall's τ)

Let (x_i, y_i) , $i = 1, \dots, N$, be independent observations of a pair of continuous random variables (X, Y) with distribution function F . The empirical Kendall's τ then is defined as

$$\tau = \frac{c}{\binom{N}{2}} - \frac{d}{\binom{N}{2}} = \frac{2c}{\binom{N}{2}} - 1.$$

where

- $c = |\{i < j : x_i < x_j, y_i < y_j \wedge x_i > x_j, y_i > y_j\}|$, which gives the number of concordant pairs
- $d = |\{i < j : x_i < x_j, y_i > y_j \wedge x_i > x_j, y_i < y_j\}|$, which gives the number of discordant pairs.

Please note, that we do not have to consider the case $X = Y$ and $X^* = Y^*$ in the above, since the probability of this to appear is zero in case of two continuous random variables.

Due to $\binom{N}{2} = \frac{N(N-1)}{2} = c + d$, which exactly gives the total number of distinct pairs (x_i, y_i) and (x_j, y_j) , $i < j$, the empirical Kendall's τ gives an analogous definition to the theoretical version of the Kendall's τ .

Proposition (Properties of Kendall's τ)

- $-1 \leq \tau_{X,Y} \leq 1$
- $\tau_{X,Y} = 0$ if X and Y are independent

- $\tau_{X,Y} = 1$, if $g : \mathbb{R} \rightarrow \mathbb{R}$ is a strictly increasing function and $Y = g(X)$
- $\tau_{X,Y} = -1$, if $g : \mathbb{R} \rightarrow \mathbb{R}$ is a strictly decreasing function and $Y = g(X)$
- $\tau_{g_1(X),g_2(Y)} = \tau_{X,Y}$, if $g_1 : \mathbb{R} \rightarrow \mathbb{R}$ and $g_2 : \mathbb{R} \rightarrow \mathbb{R}$ are two strictly increasing functions.

The last property mentioned above depicts the difference between the Spearman's ρ and the Kendall's τ by emphasising the fact, that the Kendall's τ does not calculate the difference between the ranks but that it quantifies the difference between the percentage of concordant and discordant pairs of observations.

2.3.2 Scatter plots

In the exploratory data analysis, we typically consider a scatter plot matrix, which shows the pairwise plots of the response and the covariates each against the others.

The upper or lower panel of the scatter plot matrix often is used to display a correlation coefficient, where the scatter plots then help to avoid misinterpretation: If any crucial covariate is missed, spurious correlation, spoofing of correlation or inversion of the sign of correlation may occur. Furthermore, regarding the correlation coefficient, only the strength of correlation can be determined. The effects of correlation cannot be detected. Another problem is, that some dependencies might occur from the data set by random. Investigating the scatter plots, we can identify at least some of these cases.

Furthermore, we typically read out the range of the different covariates from the scatter plots and check the model assumptions of linearity, variance homogeneity and independence:

Range of the covariates

When the mean value of the individual covariates is very large or small, or when the range of a covariate reaches over several orders of magnitude, the numerical optimisation algorithm used to find the regression coefficients may fail. Furthermore, it is always useful to have the covariates on the same order of magnitude in order to prevent misinterpretation of the regression coefficients. Hence, we first have a look at the range of the different covariates and do some scaling if necessary.

Linearity and variance homogeneity

Concerning the linearity assumption, we have a look at the scatter plots of the response variable against the covariates. These should reveal a linear pattern. Otherwise transformations of the covariates are advisable.

In Poisson regression and in Negative Binomial regression, we cannot stay with the identical link like in ordinary linear regression, but we typically use the log-link specification.

Due to this, the linearity assumption of a count model cannot be checked by just looking at the scatter plots of the response variable against the covariates, since we have

$$Y_i \sim \text{Poisson}(t_i e^{\mathbf{x}_i^T \boldsymbol{\beta}})$$

with

$$\log(y_i) = \log(t_i) + \mathbf{x}_i^T \boldsymbol{\beta} \text{ and } \mu_i = E(Y_i).$$

This means, that $\log(y_i) - \log(t_i)$ has to be linear in \mathbf{x}_i , if \mathbf{x}_i should enter the model linearly.

If we furthermore have to deal with discrete covariates with a low number of different levels, i.e. if we have several observations y_i with a common covariate vector \mathbf{x}_i , we can make the plot easier legible by the following simplification:

Having only discrete values v_k of the j -th covariate x_{ij} , $i \in \{1, \dots, n\}$, we calculate the mean values of the corresponding subsets of y_i for each discrete value of \mathbf{x}_j as

$$y_k = \frac{\sum_{i=1}^n 1_{\{x_{ij}=v_k\}} \cdot x_{ij}}{\sum_{i=1}^n 1_{\{x_{ij}=v_k\}}}.$$

Without loss of generality, we can use this mean value y_k instead of plotting all values.

Confidence intervals can be added to these plots, to check, if the assumption of constant variance is reasonable.

Independence

The scatter plots amongst the covariates should preferably show no pattern, since we assume them to be independent. This condition often is not fulfilled by the data. Regression with correlated covariates is possible, but if there are strong correlations this may cause problems when calculating the regression coefficients or lead to model misinterpretation due to a correlation-induced change in the signs of the regression coefficients. Thus, in case of high correlation, it is advisable to remove one of the covariates.

Please note, that this problem cannot be solved including interaction terms: Essentially, correlation between two covariates means the values of the one covariate relate in some way to the values of the other, i.e. the values of one covariate generally co-occur with certain values of the other. But, correlation does not say anything about whether two covariates interact in their effect on a third covariate or not.

Interaction between two covariates however means, that the effect of one of the covariates on the response variable is not constant, but that it differs at different values of the other. Thus, two interacting covariates may be correlated or not. How to investigate interactions correctly, will be our next issue concerning the exploratory data analysis.

2.3.3 Interaction effects

Interaction terms allow to bring non-additive simultaneous effects between two or more covariates into the model. This commonly is used, if we assume, that the relationship between each of the interacting covariates and the response variable depends on the value of the other interacting covariates. Whenever there is an interaction effect present in the data, which is not included into the model, this means, that the interpretation of the individual covariates may be incomplete or misleading.

The lowest level of interaction is the two-way interaction. In the corresponding interaction plot, the vertical axis represents the response variable. One of the two covariates, which are investigated, is drawn on the horizontal axis and the other one is included by plotting multiple lines on the graph. It is always advisable to put the covariate with the higher number of levels on the horizontal axis in order to reduce the resulting number of lines in the plot. If both have the same number of levels, we should choose the covariate that has numerical values, if there is one. The correct interpretation of these plots can roughly be summarised differentiating between two cases: Parallel lines in the plot mean no interaction, whereas any crossing of the lines indicates, that there may be some interaction.

To investigate three-way interactions, there can be drawn several plots for each unique pair of the three covariates: We pick the covariate with the smallest number of levels and perform the same plots as above for the remaining two ones, where we now draw a separate plot for each level of the covariate that was picked. If the resulting plots look all the same, there is no interaction present. Whenever we find some differences among them, it might be useful to check for this three-way interaction.

These plots of course are easy to understand, but they tell nothing about the significance of the interaction effects. To investigate that, we could either reduce the full interaction model until only significant terms remain, run a stepwise AIC approach or perform a partial F test. This is part of the model selection and will be explained later on.

Please note, that including interaction terms into a regression model changes the interpretation of the coefficients of the covariates from unconditional, i.e. from the case that there is no interaction included, to conditional. Without the interaction terms, the regression coefficients show the relationship between the covariate, to which they belong, and the response variable, assuming all the other covariates to be on average value. Please make sure, that you do not mix that up with the interpretation of β_0 , which shows the value of the response variable \mathbf{Y} when all the covariates are equal to zero. Including interactions, the situation changes: We now have to read the regression coefficient of a covariate in the way, that it shows the effect of that covariate when the other covariates contained in this interaction are equal to zero and averaged over the remaining covariates.

Let us first consider a two-way interaction model. The interaction term is constructed out of the interacting covariates and usually comes into the model as multiplicative effect.

This looks as follows:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_1 \cdot X_2.$$

In such a model, the regression coefficients can be interpreted as follows:

- β_1 shows the relationship between X_1 and Y , when X_2 is equal to zero
- β_2 shows the relationship between X_2 and Y , when X_1 is equal to zero
- β_3 gives the strength of the interaction between X_1 and X_2 .

We now consider three-way interactions. A model with a three-way interaction (and all two-way interactions) looks as follows:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot X_1 \cdot X_2 + \beta_5 \cdot X_1 \cdot X_3 + \beta_6 \cdot X_2 \cdot X_3 + \beta_7 \cdot X_1 \cdot X_2 \cdot X_3.$$

When the interaction $X_1 \cdot X_2 \cdot X_3$ is included

- β_4 shows the interaction between X_1 and X_2 , when X_3 is equal to zero
- β_5 shows the interaction between X_1 and X_3 , when X_2 is equal to zero
- β_6 shows the interaction between X_2 and X_3 , when X_1 is equal to zero.

This principle can directly be extended to any order of interaction.

Including interactions, the interpretation of the regression coefficients may cause problems, if the covariates cannot attain the value zero. Consider for example the covariate 'price'. Interpreting the effect of any other covariate when the price is equal to zero does not make sense, since this case will never occur. Here, centring may help on getting a reasonable interpretation. Centring means, that we subtract the mean value for the sample of the covariate from each individual observation, which yields a mean equal to zero for the covariate. This allows to interpret the regression coefficient assuming the covariate to be on average value instead of being equal to zero. The resulting regression coefficients then usually are much more similar to the unconditional ones.

2.3.4 Seasonal effects

If the data was taken over several months or years, it is always worth to look for time effects. Covering seasonal periodicity in appropriate dummy variables can explain a lot of the variation in the model and thus may improve the model fit significantly.

To detect seasonal effects, we can cluster the data by weekday, month or quarter year and check if we can make out any pattern. We can either just sum up the variable values within the formed groups or normalise by broader defined grouping. If we for example

look for monthly effects, we can divide the values for the individual months each by the total sum over the corresponding year. Such kind of normalising is advisable especially when there is an overall increase or decrease of the values over time, since this gives us values, which all lie in the same range of the vertical axis. This eases the comparison of the groups a lot.

2.4 Model selection

Assessing the goodness of fit, we want to check, if we are able to explain our data by the chosen model in a reasonable manner. We usually measure the goodness of fit by means of statistical scores or by appropriate statistical tests. In the following, we present the two most common information-loss criteria AIC and BIC, and give a short introduction to hypothesis testing and residual analysis.

2.4.1 Information-loss criteria AIC and BIC

The idea of information-loss criteria, or just 'information criteria' in short notation, is to select an appropriate but preferably simple model, where the complexity is measured based on the number of parameters. The models are compared on the log-likelihood, penalising additional parameters that come into the model in a manner depending on the information criterion we use. This penalisation is applied, because if we add covariates to our model, which actually are not needed, the likelihood almost always increases, since the extra parameters usually let the model get closer to the data.

Especially, if our models are nested, the AIC and the BIC are appropriate methods for assessing the goodness of fit, because the quality of the individual models is measured relative to the others. Unfortunately, the goodness of fit cannot be measured in an absolute sense.

AIC

The probably most well known information criterion is the Akaike information criterion (AIC). Using AIC, we aim at minimising the information loss we have to take when representing the true data by a model, i.e. we want to choose the model with the highest value for $l(\hat{\beta}) - p$. This is equivalent to choosing the model with the lowest AIC, which is defined as

$$\text{AIC} = -2l(\hat{\beta}) + 2p,$$

where

- $\hat{\beta}$ is the parameter gained from the maximum likelihood estimation
- p is the degree of freedom.

Using the AIC, the penalisation is done by adding twice the number of parameters to be estimated.

Given a set of candidate models, the value of the AIC gives us the relative probability that the individual models minimise the estimated information loss:

Let AIC_m denote the minimal AIC among our set of candidate models, then

$$e^{\frac{AIC_{\min} - AIC_i}{2}}$$

gives us the probability, that model i minimises information loss compared to the model with the minimal AIC. This probability obviously decreases with increasing AIC. For this reason, we always choose the model minimising the AIC as the 'best model'.

Please note, that the AIC is valid only asymptotically. For small data sets, correction by including correlation might be necessary.

BIC

The BIC calculates as

$$BIC = -2l(\hat{\beta}) + \log(n)p,$$

where

- $\hat{\beta}$ is the parameter gained from the maximum likelihood estimation
- p is the degree of freedom
- n is the number of observations contained in our data set.

In contrast to the AIC, we here multiply p by the logarithm of the sample size n when penalising the free parameters. This is done, because large sample sizes often help on improving the maximum likelihood estimate and thus models containing comparatively many parameters appear more attractive than they actually are. Hence, using the BIC we clearly penalise harder compared to the AIC for a number of observations greater than 8, since we then have $\log(n)|_{n \geq 8} > 2$.

To answer the question which one of these two information criteria we should use when comparing our models, let us investigate them in more detail.

The AIC and the BIC both are maximum likelihood estimate driven information criteria and penalise free parameters in an effort to avoid overfitting. In both cases, the best model is the one that minimises the score. Due to the different ways of penalising the free parameters, AIC presents the danger that it might overfit the model, whereas BIC tends to underfitting. The AIC is aimed at finding the best approximating model to the unknown data generating process. As such it fails to converge in probability to the model. For the BIC by contrast convergence in probability is present when tending to infinity.

We conclude, that the AIC and the BIC both are mathematically convenient approximations, which we can use in order to efficiently compare our models. Most of the time they will more or less agree on the preferred model anyhow. If they would give significantly different results for the 'best' model, this indicates, that we probably have a high model uncertainty.

2.4.2 Deviance

Deviance is a measure of goodness of fit for generalised linear models, which can be interpreted in a similar way like the residual sum of squares in ordinary linear regression. It is defined as

$$2(L(\hat{\boldsymbol{\beta}}_{\max}) - L(\hat{\boldsymbol{\beta}})) \cdot \Phi,$$

where $\hat{\boldsymbol{\beta}}_{\max}$ represents the maximised likelihood of the saturated model containing one parameter per data point. Hence, $\hat{\boldsymbol{\beta}}_{\max}$ is evaluated by setting $\hat{\boldsymbol{\mu}} = \mathbf{y}$, which results in the highest value that the likelihood can achieve for the given data.

In practice, we mostly use the scaled deviance, which calculates as

$$D^* = \frac{D}{\Phi}.$$

In case of Poisson regression and Negative Binomial regression, the deviance and the scaled deviance are the same due to $\Phi = 1$.

From the results of the likelihood ratio test, we conclude, that, in case of correct model specification, we approximately have

$$D^* \sim \chi_{n-p}^2.$$

In general, we distinguish between two forms of deviance: The null deviance and the residual deviance. The null deviance indicates, how well the response variable is explained by the null model, that only contains the intercept. The residual deviance gives us the deviance for the model including the selected covariates. In case of properly chosen covariates, the residual deviance of course is significantly smaller than the null deviance, since the covariates should explain most of the variation in the response variable.

Deviance is often used for model selection in generalised linear models. We usually consider an analysis of deviance table, which can be seen as the equivalent to ANOVA tables in case of ordinary linear regression. Below we discuss two statistical hypothesis test, which are often used in this context.

Residual deviance test

The residual deviance test is used to check the model assumptions of a specified generalised linear model. This comprises verifying the correct specification of the response distribution, the link function and the linear predictors.

Definition (Residual deviance test for a generalised linear model)

Reject the null hypothesis

H_0 : the model assumptions of the specified generalised linear model are satisfied against the alternative hypothesis

$$H_1: \text{not } H_0$$

at level α , if and only if

$$\frac{D(\hat{\mu}, y)}{\hat{\Phi}} > \chi_{n-q, 1-\alpha}^2,$$

where

- $\chi_{r, 1-\alpha}^2$ denotes the $100(1-\alpha)\%$ quantile of a χ^2 -distribution with r degrees of freedom
- $\hat{\Phi}$ is an estimate for the dispersion parameter Φ .

Partial deviance test

If we want to compare the fit of two nested generalised linear models, we most often use the partial deviance test.

Definition (Partial deviance test for nested generalised linear models)

Reject the null hypothesis

$$H_0 : \beta_2 = 0$$

versus the alternative hypothesis

$$H_1 : \beta_2 \neq 0,$$

if and only if

$$\frac{D_R - D_F}{\hat{\Phi}_F} > \chi_{p_2, 1-\alpha}^2,$$

where

- $\beta_1 \in \mathbb{R}^{p_1}$ and $\beta_2 \in \mathbb{R}^{p_2}$ with $p_1 + p_2 = p$
- D_R gives the deviance of the reduced model
- D_F gives the deviance of the full model
- $\hat{\Phi}_F$ is an estimate for the dispersion parameter Φ based on the full model.

2.4.3 Coefficient of determination for linear models

The *coefficient of determination* (R^2) is a measure for the overall fit of a linear model and gives the percentage of variability in the response variable, which can be explained by the model. It is defined as

$$R^2 := \frac{\text{SSR}}{\text{SST}},$$

where

- $\text{SSR} := \sum_{i=1}^n (\hat{Y}_i - \bar{\mathbf{Y}})^2$ is the regression sum of squares
- $\text{SST} := \sum_{i=1}^n (Y_i - \bar{\mathbf{Y}})^2$ is the total sum of squares
- $\bar{\mathbf{Y}} := \frac{1}{n} \sum_{i=1}^n Y_i$.

Defining the sum of squared error as

$$\text{SSE} := \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

and using the relation

$$\sum_{i=1}^n (\hat{Y}_i - \bar{\mathbf{Y}})(Y_i - \hat{Y}_i) = 0,$$

we can write

$$\text{SST} = \text{SSR} + \text{SSE}.$$

This yields

$$R^2 := \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}.$$

The R^2 lies in the unit interval $[0, 1]$ and attains the value $R^2 = 0$ in the case, that there is no linear relationship. The closer it is to 1, the better our model accounts for the variability in the response. That means, in case of $R^2 = 0$, the best model would be $\mathbf{Y} = \beta_0$. This occurs, if $\hat{Y}_i = \bar{\mathbf{Y}} \forall i$. In case of $R^2 = 1$, \mathbf{Y} can fully be explained by the model, i.e. $Y_i = \hat{Y}_i \forall i$.

Please note, that the R^2 only measures the quality of the linear approximation, but does not check the model specification. This results in a higher R^2 for the models, that were estimated using least squares. A high value for R^2 thus does not necessarily indicate a good fit of the model.

If we want to compare several multiple linear regression models using R^2 , we have to be careful: Adding a new covariate to an existing model will always increase the model's R^2 , or at least will not decrease it. As a consequence, model selection according to R^2 often results in overfitting.

To overcome this problem, we better use the *adjusted* R^2 . This statistic, which is very similar to the R^2 , is defined as

$$R_{\text{adj}}^2 := 1 - \frac{\text{SSE}/(n-p)}{\text{SST}/(n-1)}.$$

In contrast to the R^2 , the adjusted R^2 may also attain negative values.

The adjusted R^2 is always lesser than or equal to the R^2 . When adding an additional covariate to the model, the adjusted R^2 only increases, if the R^2 attains a greater value than we would expect if the new covariate had no explanatory value at all.

Although the adjusted R^2 takes the number of estimated regression parameters into account, the penalisation of unnecessary complex models is not very strong. This again ends up in models containing too many covariates.

2.5 Residual analysis

The examination of the residuals is one of the most important techniques when checking the model fit. The residuals are the deviation of the observations from the sample mean. Having a data set consisting of n observations, the i -th raw residual is defined as

$$r_i = y_i - \hat{y}_i,$$

where

- y_i is the observed value
- \hat{y}_i represents the fitted value.

Regression analysis always tries to minimise these residuals e.g. by least squares estimation or by the maximum likelihood method.

In case of generalised linear models we cannot simply examine these raw residuals, but the residuals have to be standardised. The main reason for this need of standardisation is the problem of checking the validity of the assumed mean variance relation ship.

In case of Poisson regression for example, the variance of the residuals should increase in direct proportion to the size of the fitted values μ_i . Plotting the raw residuals against the fitted values, it would cost a lot of effort to judge on the correct model specification, since we would have to check whether the residual variability increases in proportion to the mean. Examining standardised residuals instead, these approximately behave like residuals from an ordinary linear regression and we just have to check for equal variance among them.

2.5.1 Pearson residuals

The Pearson residuals calculate as the raw residuals divided by the square root of the variance function $V(\boldsymbol{\mu})$.

Remember from above, that the variance function of a generalised linear model is given by $V(\boldsymbol{\theta}) = b''(\boldsymbol{\theta})$, which in mean parametrisation can be written as $V(\boldsymbol{\theta}) = V(h(\boldsymbol{\mu})) := V(\boldsymbol{\mu})$. Thus,

$$r_i^P := \frac{r_i}{\sqrt{V(\hat{\mu}_i)}},$$

where

- r_i is the i -th raw residual
- $V(\hat{\mu}_i)$ is the estimated variance function for the i -th observation.

These residuals have approximately zero mean and variance Φ in case of correct model specification. Further, they should not show any pattern when plotted against the fitted values or the covariates.

Unfortunately, the distribution of the Pearson residuals often is quite asymmetric around zero, which means that they do not behave like the residuals from an ordinary linear regression. For this reason, we prefer the deviance residuals when checking the goodness of fit.

2.5.2 Deviance residuals

The deviance residuals are best comparable to the residuals from an ordinary linear regression. This is due to the fact, that the deviance is the equivalent to the residual sum of squares in ordinary linear regression:

In the ordinary linear regression, the deviance calculates as the sum of the squared residuals. The residuals thus are the square roots of the components of the deviance with the appropriate sign.

By analogy to ordinary linear regression, we hence define the i -th deviance residual in a generalised linear model as

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$$

where

- y_i is the i -th observation of our data set
- $\hat{\mu}_i$ are the fitted means
- d_i is the deviance contribution of observation i .

The deviance contribution is defined as

$$d_i = -2[l((\hat{\mu})_i, \Phi|y_i) - l(y_i, \Phi|y_i)] \cdot \Phi.$$

The sum of squares of these deviance residuals then gives the deviance itself. Thus, if all parameters were known, we would have $D^* \sim \chi_n^2$, which implies $d_i \sim \chi_1^2$. This yields $r_i^D \sim N(0, 1)$, which proves the similarity between the residuals from an ordinary linear regression and the deviance residuals.

Of course, it is not very meaningful to identify the distribution of one single d_i , but this nevertheless suggests that we can expect the deviance residuals to behave very similar like $N(0,1)$ random variables, at least in the case of a well-fitting model.

Please note, that the residuals for non-normal generalised linear models are skewed.

2.6 Modelling the economic key figure 'price elasticity'

Concerning price setting, the basic mechanism connecting the supply and the demand has to be considered. Producers like to supply more goods at higher prices, because selling at high prices brings along a good revenue and profit. Mathematically spoken, this means a positive correlation between the price and the quantity sold.

The demand of the clients in contrast has a negative correlation. This means, that less people are willing to purchase, if prices increase (except the product is essential). The market is stabilised automatically, so that supply and demand are balanced. If prices exceed this equilibrium price, demand will decrease below supply. Lower prices by contrast will push sales accordingly.

An important index in this context is the price elasticity.

Definition (Price elasticity)

The price elasticity is defined as

$$\eta = \frac{\Delta \text{quantity sold}\%}{\Delta \text{price}\%}.$$

Expressing this in mathematical terms, we get

$$\eta = \frac{\partial E(Y_i|X_i)}{\partial x_{ji}} \cdot \frac{x_{ji}}{E(Y_i|X_i)},$$

where

- $E(Y_i|X_i)$ is the conditional expectation of the quantity sold Y_i for given prices x_{ji}
- x_{ji} is a concrete price.

Let us now check, how we can retrieve this parameter directly from the regression coefficients.

In Poisson regression we do not have the simple linear interpretation of the regression coefficients like in case of multivariate linear regression, but we have multiplicative exponentiated coefficients. It is

$$E(Y_i|X_i) = e^{\mathbf{x}_i^T \boldsymbol{\beta}},$$

where $\mathbf{x}_i^T = (x_{1i}, \dots, x_{ki})^T$ is the i -th column vector of the design matrix \mathbf{X} .

The regression coefficients thus indicate the effect of a one unit change in x_{ji} on rate, i.e. changing x_{ji} by one unit induces a multiplication of $E(Y_i|X_i)$ with e^{β_j} .

Differentiation yields

$$\begin{aligned} \frac{\partial E(Y_i|X_i)}{\partial x_{ji}} &= \frac{\partial e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{\partial x_{ji}} \\ &= \frac{\partial e^{\beta_1 x_{1i} + \dots + \beta_k x_{ki}}}{\partial x_{ji}} \\ &= \beta_j \cdot e^{\beta_1 x_{1i} + \dots + \beta_k x_{ki}} \\ &= \beta_j \cdot E(Y_i|X_i). \end{aligned}$$

Using this result in the formula for the price elasticity gives us

$$\begin{aligned} \eta &= \frac{\partial E(Y_i|X_i)}{\partial x_{ji}} \cdot \frac{x_{ji}}{E(Y_i|X_i)} \\ &= \beta_j \cdot E(Y_i|X_i) \cdot \frac{x_{ji}}{E(Y_i|X_i)} \\ &= \beta_j \cdot x_{ji}. \end{aligned}$$

Hence, our regression coefficient β_j gives us the proportionate change of the expected quantity sold $E(Y_i|X_i)$ induced by a one unit change in the price x_{ji} . Thus, if β_j is positive, the mean value grows with increasing x_{ji} , whereas for negative β_j it decreases with increasing x_{ji} . Thus, in this type of model, we have η depending on x_{ji} .

Let us now inspect a model with the covariate for price on log-scale.

This gives us

$$E(Y_i|X_i) = e^{\log(\mathbf{x}_i^T \boldsymbol{\beta})},$$

where \mathbf{x}_i^T again is the i -th column vector of the design matrix \mathbf{X} .

Differentiation yields

$$\begin{aligned} \frac{\partial E(Y_i|X_i)}{\partial x_{ji}} &= \frac{\partial e^{\boldsymbol{\beta} \log(\mathbf{x}_i^T \boldsymbol{\beta})}}{\partial x_{ji}} \\ &= \frac{\partial e^{\beta_1 \log(x_{1i}) + \dots + \beta_k \log(x_{ki})}}{\partial x_{ji}} \\ &= \frac{\beta_j}{x_{ji}} e^{\beta_1 \log(x_{1i}) + \dots + \beta_k \log(x_{ki})} \\ &= \frac{\beta_j}{x_{ji}} E(Y_i|X_i). \end{aligned}$$

In this case we have

$$\beta_j = \frac{\partial E(Y_i|X_i) \cdot x_{ji}}{\partial x_{ji} \cdot E(Y_i|X_i)},$$

which exactly is the definition of the price elasticity. Thus, we have

$$\eta = \beta_j,$$

i.e. with prices on log scale, β_j now measures the percentage change in $E(Y_i|X_i)$ due to a percentage change in x_{ji} . Thus, η does not depend on x_{ji} .

For further details on that, please refer to Yamamura (2012).

In economical context, the preferred approach is to have a constant value for η for a small range of prices x_{ji} . So, we preferably take all covariates concerning prices on log scale. This way, the model directly gives us the well-known price elasticity.

2.7 Copula Modelling

Generally speaking, a copula represents the distribution function of a multidimensional random vector with uniform margins, where the marginal and the common dependencies are modelled separately. A detailed introduction to that is for example given in Czado (2015).

In this thesis, we use copula models to model the dependency structure of the products using the deviance residuals of the derived GAMs.

Definition (Copula)

A function $C : [0, 1]^d \rightarrow [0, 1]$ is called a d-dimensional copula, if there exists a random vector (U_1, \dots, U_d) with $U_i \sim U[0, 1]$, $i = 1, \dots, d$, such that

$$P(U_1 \leq u_1, \dots, U_d \leq u_d) = C(u_1, \dots, u_d).$$

Copula models are widely applicable, since Abe Sklar (1959) proved, that any multivariate distribution can be split into its margins and a copula. This became the core theorem of the copula theory.

Theorem (Sklar's Theorem)

Let \mathbf{X} be a d-dimensional random vector with joint distribution function F and marginal distribution functions F_i , $i = 1, \dots, d$, then

$$F(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d))$$

and

$$f(x_1, \dots, x_d) = c(F_1(x_1), \dots, F_d(x_d))f_1(x_1)\dots f_d(x_d)$$

for some d -dimensional copula C with density c . For absolutely continuous distributions the copula is unique.

The inverse also holds, i.e. the copula corresponding to a multivariate distribution is given by

$$C(u_1, \dots, u_d) = F(F_1^{-1}(u_1), \dots, F_d^{-1}(u_d))$$

and

$$c(u_1, \dots, u_d) = \frac{f(F_1^{-1}(u_1), \dots, F_d^{-1}(u_d))}{f_1(F_1^{-1}(u_1)) \dots f_d(F_d^{-1}(u_d))}.$$

From this theorem, we can easily conclude, why we define a copula as the distribution of uniformly distributed random variables: For a random vector (X_1, \dots, X_d) with marginal distributions F_1, \dots, F_d and joint distribution F , it holds that $F_i(X_i) \sim U[0, 1]$ for $i = 1, \dots, d$. The corresponding copula C then is defined as the distribution function of $(F_1(X_1), \dots, F_d(X_d))$.

One possibility to model copulas is by defining a Vine structure. We decompose the multivariate distribution into a set of bivariate distributions, which are represented as bivariate copulas. Since this decomposition is not unique, we need to organise the bivariate copulas in an appropriate tree structure. This allows to describe the whole multivariate distribution as a copula, which calculates out of the bivariate copula densities.

A copula contains all information on the dependence structure in a multivariate random vector. Due to Sklar's Theorem, we can select bivariate copulas from a wide range of parametric families. This allows a lot of flexibility regarding the shape of the distribution, which results in much more precise approaches than classical multivariate models do.

2.7.1 Model selection

A copula model basically consists of

- a tree structure
- a set of copula families
- a set of copula parameters.

To specify an appropriate model for our data set we thus have to

- select a Vine structure specifying which unconditional and conditional pairs to use
- choose a bivariate copula family for each pair contained in the Vine structure
- estimate the corresponding parameters for each copula.

2.7.2 Structure selection

Vine Copulas

Vine copulas are also called pair-copula constructions. This is due to the fact, that any d -dimensional copula density can be decomposed into a product of $\frac{d(d-1)}{2}$ bivariate (conditional) copula densities, or, looking at it the other way round, an arbitrary d -dimensional copula density can be constructed by using only bivariate building blocks.

Regular vine trees (R-vine trees)

As already mentioned above, the decomposition of the density is not unique. Especially in higher dimensions there is a countless number of different possibilities. To really find the 'best' model, we would have to check all the possible R-Vine constructions. But, as the number of possible R-Vines on n variables increases with $\frac{n!}{2} \cdot 2^{\binom{n-2}{2}}$ (for details see Morales-Napolean et al. (2010)), we usually apply an heuristic algorithm to specify the structure of our n -dimensional model.

The structure selection typically is based on Kendall's τ . This allows to measure dependence independent from the assumed distributions, which turns out to be very useful when working with different copula families.

Having this at hand, we investigate the corresponding algorithm in more detail: Starting from the first tree, we define $n - 2$ conditional trees by a sequential method in such a way, that the chosen pairs model the strongest pairwise dependencies, which are present in the data. This iterative procedure of constructing the trees does not ensure to find the 'global optimum', but, nevertheless, it returns very good results. This is justified by the fact, that the copula families specified in the first tree most often have the greatest influence on the model fit. Furthermore, concentrating on the strongest dependencies, we are very likely to attain a good fit since the copula distribution functions are very similar when coming closer to independence.

To describe the structure of a Vine copula, usually a sequence of linked trees $T_k = (V_k, E_k)$, $k = 1, \dots, d - 1$, is used. V_k denotes the set of nodes and E_k the set of edges in the tree T_k . The tree structure then is constructed as follows:

Definition (Regular Vine (R-vine))

$\nu = (T_1, \dots, T_{d-1})$ is an R-vine tree sequence on d elements, if

- (1) Each tree T_j is connected, i.e. for all nodes $a, b \in T_j$, $j = 1, \dots, d - 1$, there exists a path $(n_1, \dots, n_k) \in N^k$ with $a = n_1, b = n_k$.
- (2) T_1 is a tree with nodes $N_1 = 1, \dots, d$ and a set of edges E_1 .
- (3) For $j \geq 2$, T_j is a tree with nodes $N_j = E_{j-1}$ and edges E_j .
- (4) For $j = 2, \dots, d - 1$ and $\{a, b\} \in E_j$ it must hold that $|a \cup b| = 1$.

Condition (4) is the so-called proximity condition, which ensures that if a and b are connected by e in tree T_j , $j \geq 2$, then a and b must share a common node in tree T_{j-1} .

Please note, that these trees are not directed, i.e. the set notation $e = \{a, b\}$ does not induce any ordering of its elements.

Regular vine copulas (R-vine copulas)

To be able to use the R-vine tree for copula modelling, we need to include a stochastic component. For this purpose, we need an R-vine distribution.

R-vine copula

A copula C corresponding to a random vector (U_1, \dots, U_d) , with $U_i \sim U[0, 1]$, $i = 1, \dots, d$, is called an R-vine copula, if there is a tuple (V, C) such that

- (1) V is an R-vine tree sequence on d elements.
- (2) $C = \{C_e | e \in E_k, k = 1, \dots, d-1\}$, where C_e is a bivariate copula.
- (3) Each $e \in E_k$, $k = 1, \dots, d-1$, can be identified as $\{a_e, b_e; D_e\}$, and C_e is the copula corresponding to $(U_{a_e}, U_{b_e}) | (U_k)_{k \in D_e} = (u_k)_{k \in D_e}$.

For an R-vine copula C corresponding to the tuple (V, C) , where all included copulas allow for a density, we can write the overall density as

$$c(\mathbf{u}) = \prod_{k=1}^{d-1} \prod_{e \in E_k} c_{a_e, b_e; D_e}(C_{a_e | D_e}(u_{a_e} | \mathbf{u}_{D_e}), C_{b_e | D_e}(u_{b_e} | \mathbf{u}_{D_e})),$$

where $\mathbf{u}_{D_e} = (u_j)_{j \in D_e}$ is a subvector of $\mathbf{u} = (u_1, \dots, u_d)$ and $C_{j_e | D_e}$ is the conditional distribution of $U_{j_e} | \mathbf{U}_{D_e} = \mathbf{u}_{D_e}$, $j_e \in \{1, \dots, d\}$.

In the following, this will be referred to in short notation as

$$u_{j_e | D_e} := C_{j_e | D_e}(u_{j_e} | \mathbf{u}_{D_e}).$$

We thus need $n-1$ unconditional copulas for the first tree, $n-2$ conditional copulas for the second tree, etc. The corresponding copula families and parameters are chosen arbitrarily and are independent of each other. But, due to the conditional variables in the tree structure, the choice of the different copulas will of course influence each other.

The above defined density of an R-vine copula involves conditional distributions of the form $C_{j_e | D_e}$, where $j_e \in \{a_e, b_e\}$. These conditional distributions can be expressed as a recursive application of conditional distributions corresponding to bivariate copulas contained in C . This is covered by h-functions.

Definition (h-function)

Let $U_1, U_2 \sim U[0, 1]$ and C be the copula of (U_1, U_2) , then the corresponding h-functions are defined as

$$h_{1|2}(u_1|u_2) = C_{1|2}(u_1|u_2) = \frac{\partial C(u_1, u_2)}{\partial u_2} = P(U_1 \leq u_1 | U_2 = u_2),$$

and

$$h_{2|1}(u_2|u_1) = C_{2|1}(u_2|u_1) = \frac{\partial C(u_2, u_1)}{\partial u_1} = P(U_2 \leq u_2 | U_1 = u_1).$$

Representing R-vines using R-vine matrices

The idea of the R-vine matrix is to store the indices $\{i(e), j(e) | D(e), e \in T_j, j = 1, \dots, n\}$ in a lower triangular matrix, which then describes the tree structure.

Definition (R-Vine Matrix)

Let \mathbf{M} be a lower triangular matrix with entries $m_{i,j}$, $i \leq j$. Each entry $m_{i,j}$ is allowed to take integer values from 1 to n . \mathbf{M} is called an R-vine matrix if it satisfies the following conditions:

- (1) $\{m_{i,i}, \dots, m_{n,i}\} \subset \{m_{j,j}, \dots, m_{n,j}\}$ for $1 \leq i \leq j \leq n$.
- (2) $m_{i,i} \notin \{m_{i,i+1}, \dots, m_{n,i+1}\}$ for $i = 1, \dots, n-1$.
- (3) For all $i = 1, \dots, n$ and $k = i+1, \dots, n-1$ there exists a j in $\{i+1, \dots, n-1\}$ such that

$$\{m_{k,i}, \{m_{k+1,i}, \dots, m_{n,i}\}\} = \{m_{j,j}, \{m_{k,j}, \dots, m_{n,j}\}\}$$

or

$$\{m_{k,i}, \{m_{k+1,i}, \dots, m_{n,i}\}\} = \{m_{k,j}, \{m_{k+1,j}, \dots, m_{n,j}, m_{j,j}\}\}.$$

From (1) it follows, that all the entries of the columns on the right of a selected column are contained in this column. From (2) we conclude, that the diagonal entry of a column is not contained in any column further to the right. Condition (3) exactly is the counterpart of the proximity condition for R-vine trees.

The most important properties of an R-Vine matrix are, that all elements in a column are different and that deleting the first row and column form a n -dimensional R-Vine matrix always gives a $(n-1)$ -dimensional R-Vine matrix.

Since each entry below the diagonal of this matrix describes one edge of the corresponding tree, this type of lower triangular matrix is also used to describe the copula family and parameters by entering the values at the matrix field for the corresponding edge.

Special cases of R-vine copulas

Recall, that having n parameters, the number of possible R-vines is $\frac{n!}{2} \cdot 2^{\binom{n-2}{2}}$ and that taking into account, that numerous copula families do exist with their unknown parameters, it is a quite expensive calculation to determine the adequate tree structure with the best fitting copulas. As we have already seen, there exist some heuristic algorithms to receive well-fitting R-vines.

Another possibility of simplification is to restrict the tree structure. There are two special types of R-vines, which have been applied successfully in various fields like for example in the context of stock market or health. These two tree types are the canonical vines (C-vines) and the drawable vines (D-vines).

A C-Vine is an R-vine which contains one node with the maximal degree in each tree, i.e. each tree of a C-vine has a star structure. The ordering of the root nodes in the first tree thus completely determines the C-vine structure.

A D-vine is an R-vine for which the first tree has nodes with degree two or less, which gives a path structure. Subsequently, there is no other possibility to derive conditional trees than following this path structure. Hence, the vine structure is completely defined by the first tree.

C-vines and D-vines both have $\frac{n!}{2}$ possible tree structures for n parameters, which is remarkably less than for an R-vine. For $n = 7$ this reduces the number of possibilities by the factor of 1/1024. Due to the limits in the structure, C-vines and D-vines are most suitable for data fulfilling some criteria: C-vine copulas for example can especially be applied to data, of which pivotal variables can be identified and D-vine copulas may be useful in particular for time series data. Nevertheless, both types can be used with arbitrary data. But, of course, some statistical tests like for example the Vuong-test should be carried out to examine the reasonableness of the selected models.

In the following, some methods and descriptions will be introduced referring to R-vines. These in principle are also valid for C-vines and D-vines as special cases of R-vines.

2.7.3 Parametric copula families

We have already seen, that Sklar's theorem directly gives us the copula density as

$$c(u_1, \dots, u_d) = \frac{f(F_1^{-1}(u_1), \dots, F_d^{-1}(u_d))}{f_1(F_1^{-1}(u_1)) \dots f_d(F_d^{-1}(u_d))} = \frac{\partial^d C(u_1, \dots, u_d)}{\partial u_1 \dots \partial u_d},$$

where f, f_1, \dots, f_d denote the corresponding density functions of F, F_1, \dots, F_d .

This allows us to use arbitrary parametric distribution functions F to construct parametric copula families. A short overview over the most well-known ones is given below: We differentiate between elliptical and Archimedean copula families and introduce the multivariate Gaussian copula and the t-copula, which both are elliptical copulas, as well

as four Archimedean copula families, namely the Clayton copula, the Gumbel copula, the Frank copula and the Joe copula.

Elliptical copula families

Multivariate Gaussian copula

Applying the above to the multivariate Gaussian distribution with zero mean and correlation matrix R yields the multivariate Gaussian copula, which is defined as

$$C(\mathbf{u}; R) = \Phi_R(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)),$$

where Φ^{-1} denotes the inverse of the standard normal cumulative distribution function Φ and $\Phi_d(\cdot \cdot \cdot; R)$ the multivariate standard normal distribution function with zero mean and symmetric positive definite correlation matrix $R \in [-1; 1]^d$.

The copula is given by

$$c(\mathbf{u}; R) = |R|^{-0.5} e^{\frac{1}{2} \mathbf{x}^T (\mathbf{I}_d - R^{-1}) \mathbf{x}},$$

where $\mathbf{x} = (x_1, \dots, x_d)^T \in \mathbb{R}^d$ with $x_i = \Phi^{-1}(u_i)$ and $i = 1, \dots, d$.

Multivariate t-copula

Similar to the multivariate Gaussian copula we can derive the multivariate t-copula from the multivariate t-distribution as

$$C(\mathbf{u}; R, \nu) = t_{R, \nu}(t_\nu^{-1}(u_1), \dots, t_\nu^{-1}(u_d)),$$

where $t_{R, \nu}$ denotes the distribution function of the multivariate standard t-distribution with correlation matrix $R \in [-1, 1]^d$ and $\nu > 0$ degrees of freedom. Further, t_ν^{-1} represents the inverse of the distribution function t_ν of the univariate standard t-distribution with ν degrees of freedom.

The multivariate Gaussian copula and the multivariate t-copula both range among the most widely used copulas and are quite similar, which especially is true for high numbers of degrees of freedom. For this reason it is common to abstain from using the t-copula for degrees of freedom higher than 30, but to use Gaussian copulas instead.

Archimedean copulas

Besides this, there is a second important class of copulas, the one-parametric Archimedean copulas. Some well-known examples for this class of copulas are the Clayton copula, the Gumbel copula, the Frank copula and the Joe copula.

Definition (Archimedean copula)

Let $\Phi : [0, 1] \rightarrow [0, \infty]$ be continuous, strictly monotonous decreasing and convex with $\Phi(1) = 0$. Define Φ^{-} as the generalised inverse of Φ , i.e.

$$\Phi^{-}(y) := \inf\{x \in [0, \infty] : \Phi(x) \geq y\}, y \in [0, 1].$$

Then, the Archimedean copula function is given by

$$C(u_1, u_2) = \Phi^{-}(\Phi(u_1) + \Phi(u_2)).$$

Restricting to the bivariate case, this yields the following:

Clayton copula

For $\Phi_{\theta}(x) = \frac{x^{-\theta}-1}{\theta}$, $\theta \in (0, \infty)$, the Clayton copula is defined as

$$C(u_1, u_2) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-\frac{1}{\theta}}.$$

Gumbel copula

With $\Phi_{\theta}(x) = (-\log(x))^{\theta}$, $\theta \in [1, \infty]$, the Gumbel copula is given by

$$C(u_1, u_2) = e^{-[(-\log(u_1))^{\theta} + (-\log(u_2))^{\theta}]^{\frac{1}{\theta}}}.$$

Frank copula

Having $\Phi_{\theta}(x) = -\log(\frac{e^{-\theta x}-1}{e^{-\theta}-1})$, with $\theta \in \mathbb{R} \setminus \{0\}$, the bivariate Frank copula is defined as

$$C(u_1, u_2) = -\frac{1}{\theta} \log\left(1 + \frac{(e^{-\theta u_1}-1)(e^{-\theta u_2}-1)}{e^{-\theta}-1}\right).$$

Joe copula

With $\Phi_{\theta}(x) = -\log(1 - (1-x)^{\theta})$, $\theta \in [1, \infty]$, the Joe copula is derived as

$$C(u_1, u_2) = 1 - ((1-u_1)^{\theta} + (1-u_2)^{\theta} - (1-u_1)^{\theta}(1-u_2)^{\theta})^{\frac{1}{\theta}}.$$

For further details on that please refer to Nelsen (2006).

Independence copula

The independence copula is given by

$$C(\mathbf{u}) = \prod_{i=1}^n u_i.$$

With Skar's theorem, we conclude, that a set of random variables is independent if and only if their copula is the independence copula.

The related copula density simply is constant.

Truncated regular copulas

If all pair copulas at levels above a distinct level M are set to bivariate independence copulas, the corresponding R-vine copula is said to be *truncated at level M* .

2.7.4 Estimation of Vine copulas

In parametric models, the estimates of R-vine copulas usually are gained by maximisation of the full likelihood. Due to the huge number of parameters that comes with increasing dimension, there often is used a sequential estimation approach to find good starting values for this optimisation, which is based on bivariate estimation only.

Definition (Sequential estimate of an R-vine copula density)

Let C be an R-vine copula corresponding to the tuple (V, C) and $(u_1^{(i)}, \dots, u_d^{(i)})_{i=1, \dots, n}$ given i.i.d. samples from C . Then, a sequential estimate of an R-vine copula density is obtained as follows:

- (1) For all $e \in E_1$ obtain estimates for c_{a_e, b_e} .
- (2) For $k = 2, \dots, d - 1$:
For all $e \in E_k$ and $j = a_k, b_k$:
 - (i) Let $j' \in D_e$ be another index such that $C_{j, j'; D_e \setminus j'} \in C$ and define $D'_e := D_e \setminus j'$.
 - (ii) Based on the sample $(u_{j|D'_e}^{(i)}, u_{j'|D'_e}^{(i)})_{i=1, \dots, n}$, obtain an estimate of the h-function $h_{j|j'; D'_e}$.
 - (iii) Define $u_{j|D_e}^{(i)} := \hat{h}_{j|j'; D'_e}(u_{j|D'_e}^{(i)} | u_{j'|D'_e}^{(i)})$, $i = 1, \dots, n$.
 - (iv) Based on $(u_{a_e|D_e}^{(i)}, u_{b_e|D_e}^{(i)})_{i=1, \dots, n}$ obtain an estimate of the copula density $c_{a_e, b_e; D_e}$.

This algorithm thus works as follows: In the first tree, each node is assigned to one random variable. From the samples of these random variables we then obtain estimates for all pair-copulas that correspond to the edges of the tree. To get samples from the second tree, we estimate the h-functions and apply them to obtain pseudo-samples, which can be used to estimate the copulas for the edges in the second tree. Doing this also for the remaining trees, we obtain estimates for all copula densities and all h-functions that are required to determine the density of the full R-vine copula.

2.7.5 Family selection and parameter estimation

Coming to the family selection and the estimation of the corresponding parameters, the importance of the tree structure finds even more expression, since the pair copula families and the corresponding parameters both depend on the tree structure.

We have already seen, that the pair copula families are independent of each other and can be chosen arbitrarily. So the question arising is, how to choose the appropriate parametric families.

A first hint on that is given by the pair plots. Since pair copulas are defined for two variables in $[0, 1]^2$, the pair plots resulting from any specified margins may not be processed.

To overcome this problem, we always consider the transformation to a joint distribution with standard normal margins instead. This allows a direct comparison to multivariate normal shapes and brings out some characteristics like for example sharp corners in case of tail dependence. Further, due to the characteristic shape of the pair plots, we can read out well-fitting parametric families from this graphical examination.

But, to get more precise results, we usually want to compare the modelling results based on some measure of goodness of fit. Of course, it is not viable to check all possible R-vine copulas to find out the globally best fitting model. But, the parameters usually are estimated for several different parametric families among which the best model is selected.

For a given parametric family, the corresponding family parameters usually are gained via maximum likelihood estimation.

Definition (Maximum likelihood estimator MLE)

Let $(U_1, \dots, U_d) \sim C_{\boldsymbol{\theta}}^{(\cdot)}$, where $\boldsymbol{\theta} \in \Theta$ and $\Theta \in \mathbb{R}^p$ is the family's parameter space. Further, denote $c_{\boldsymbol{\theta}}^{(\cdot)}$ as the density of $C_{\boldsymbol{\theta}}^{(\cdot)}$. The maximum likelihood estimator of the parameter vector $\boldsymbol{\theta}$ then is defined as

$$\hat{\boldsymbol{\theta}}_n^{\text{MLE}} := \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} \prod_{i=1}^n c_{\boldsymbol{\theta}}^{(\cdot)}(u_1^{(i)}, \dots, u_d^{(i)}).$$

This value has to be maximised by varying the model parameters. As already described, some penalising can be implemented to identify good and preferably simple models, which usually is done using the information criteria AIC and BIC.

Definition (Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC))

The two information criteria AIC and BIC are defined as

$$\text{AIC} := -2 \sum_{i=1}^n \log(c_{\hat{\boldsymbol{\theta}}_n}^{(\cdot)}(u_1^{(i)}, \dots, u_d^{(i)})) + 2p,$$

and

$$\text{BIC} := -2 \sum_{i=1}^n \log(c_{\hat{\boldsymbol{\theta}}_n}^{(\cdot)}(u_1^{(i)}, \dots, u_d^{(i)})) + \log(n) \cdot p,$$

where p is the number of parameters of the respective family and $\hat{\boldsymbol{\theta}}_n$ the parameter estimate.

As usual, the best fitting model minimises the score.

In case of bivariate estimation of one-parametric families, an alternative estimation is provided by the inversion of the empirical Kendall's τ . This is based on the one-to-one relationship of the Kendall's τ and the parameter of some families.

Inversion of the empirical Kendall's τ

Let $(U_1, U_2) \sim C_{\boldsymbol{\theta}}^{(\cdot)}$, where $\boldsymbol{\theta} \in \Theta$ and $\Theta \subset \mathbb{R}$ is the family's parameter space. Let further $\Psi : \Theta \rightarrow [-1; 1]$ be a bijective function, such that $\Psi(\boldsymbol{\theta}) = \tau(U_1, U_2)$. The inversion of the empirical Kendall's τ then yields

$$\hat{\boldsymbol{\theta}}_n = \Psi^{-1}(\hat{\tau}_n(U_1, U_2))$$

as an estimator for $\boldsymbol{\theta}$.

In both cases consistency is lost when the model is misspecified.

Relationship between the copula parameters and the Kendall's τ

For bivariate Archimedean and elliptical copulas the parameters of the copula family and the Kendall's τ can be related as follows:

Theorem (Kendall's τ for bivariate Archimedean and elliptical copulas)

If ϕ is a generator of a bivariate Archimedean copula, then the corresponding Kendall's τ satisfies

$$\tau = 1 + 4 \int_0^1 \frac{\phi(t)}{\phi'(t)} dt.$$

For elliptical copulas we have the following relationship between the association parameter δ and Kendall's τ :

$$\delta = \sin\left(\frac{\pi}{2}\tau\right).$$

For the different bivariate copula families introduced above, the Kendall's τ is given as follows:

| Family | Kendall's τ | $\tau \in$ |
|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
| Gaussian | $\tau = \frac{2}{\pi} \arcsin(\rho)$ | $[-1, 1]$ |
| t | $\tau = \frac{2}{\pi} \arcsin(\rho)$ | $[-1, 1]$ |
| Gumbel | $\tau = 1 - \frac{1}{\delta}$ | $[0, 1]$ |
| Clayton | $\tau = \frac{\delta}{\delta+2}$ | $[0, 1]$ |
| Frank | $\tau = 1 - \frac{4}{\delta} + 4 \frac{D_1(\delta)}{\delta}$ with $D_1(\delta) = \int_0^\delta \frac{x}{e^x - 1} dx$ (Debye-Function) | $[-1, 1]$ |
| Joe | $\tau = 1 + \left(\frac{-2+2\gamma+2\ln(2)+\psi(\frac{1}{\delta})+\psi(\frac{1}{2}\frac{2+\delta}{\delta})+\delta}{-2+\delta} \right)$ with Euler's constant $\gamma = \lim_{n \rightarrow \infty} (\sum_{i=1}^n \frac{1}{i} - \ln(n)) \approx 0.57721$ and Digamma-function $\psi(x) = \frac{d}{dx} \ln(\Gamma(x)) = \frac{d}{dx} \frac{\Gamma(x)}{\Gamma(x)}$ | $[0, 1]$ |

Table 2.1: Relationship between the Kendall's τ and the copula parameters for selected elliptical and Archimedean copula families.

2.7.6 Comparing copula models

The information criteria AIC and BIC, which are used to evaluate the goodness of fit in a model, normally are only applicable to nested models. Since truncated copulas and copulas containing independence copulas are nested within the corresponding copulas without independence copulas, these can be compared properly using the above.

But, for a precise comparison of different vine copula models, we need a suitable criterion for non-nested models. For this purpose, we now introduce the Vuong test. This statistical test is based on the Kullback-Leibler criterion, which measures the distance between the true but unknown distribution and a specified, approximative model with parameter estimate $\hat{\boldsymbol{\theta}}$. Please note, that $\hat{\boldsymbol{\theta}}$ is an estimate for the so called 'pseudo true value' $\boldsymbol{\theta}$, but not for the parameter of the true density.

Definition (Kullback-Leibler criterion (KLIC))

The Kullback-Leibler criterion for the true density $h_0(\cdot)$ and the estimated density $f(\cdot|\hat{\boldsymbol{\theta}})$ of a random vector \mathbf{X} is defined as

$$\text{KLIC}(h_0, f, \hat{\boldsymbol{\theta}}) := \int h_0(x) \log \left(\frac{h_0(x)}{f(x|\hat{\boldsymbol{\theta}})} \right) dx = E_0(\log(h_0(\mathbf{X}))) - E_0(\log(f(\mathbf{X}|\hat{\boldsymbol{\theta}}))),$$

where E_0 denotes the expectation with respect to the true density h_0 .

Of course, we strongly prefer the model with the minimal KLIC. As the true density h_0 usually is unknown, the equivalent method is to choose the model with maximal $E_0(\log(f(\mathbf{X}|\hat{\boldsymbol{\theta}}))$.

Vuong test

Let us consider two competing models

- **Model 1** $X \sim f_1(\cdot|\boldsymbol{\theta}_1)$
- **Model 2** $X \sim f_2(\cdot|\boldsymbol{\theta}_2)$.

To decide, which of the models is more appropriate, Vuong's closeness test - in the following shortly Vuong test - is used. The Vuong test investigates the null hypothesis that both models are an equally close approach to the data, i.e. that

$$H_0 : \text{KLIC}(h_0, f_1, \boldsymbol{\theta}_1) = \text{KLIC}(h_0, f_2, \boldsymbol{\theta}_2).$$

This can equivalently be expressed as

$$H_0 : E_0(\log(f_1(\mathbf{X}|\boldsymbol{\theta}_1))) = E_0(\log(f_2(\mathbf{X}|\boldsymbol{\theta}_2))).$$

Naturally, in case of

$$E_0(\log(f_1(\mathbf{X}|\boldsymbol{\theta}_1))) > E_0(\log(f_2(\mathbf{X}|\boldsymbol{\theta}_2))),$$

we decide in favour of model 1, and vice versa.

The Vuong test also assigns a significance level to its decision.

Theorem (Asymptotic normality of the likelihood ratio statistics for two approximating non-nested models)

With $LR_n(\hat{\theta}_1, \hat{\theta}_2)$ as the likelihood ratio for the two models and ω^2 as the variance of the random variable $\log\left(\frac{f_1(\mathbf{X}|\hat{\theta}_1)}{f_2(\mathbf{X}|\hat{\theta}_2)}\right)$, the term ν_n asymptotically follows the Normal distribution $N(0, 1)$ and is defined as

$$\nu_n := \frac{LR_n(\hat{\theta}_1, \hat{\theta}_2)}{\sqrt{n\omega^2}} \xrightarrow[n \rightarrow \infty]{\mathcal{D}} N(0, 1).$$

Thus, ν_n can be used to determine the significance level α , which is given by:

$$\alpha = 2\Phi(-|\nu_n|).$$

So far, we did not take the number of model parameters into account. Whenever we suppose the problem of overfitting models, we can use the adjusted version of the Vuong test.

Definition (Adjusted Vuong test statistics)

The adjusted Vuong test statistics is defined as

$$\tilde{LR}_n(\hat{\theta}_1, \hat{\theta}_2) := LR_n(\hat{\theta}_1, \hat{\theta}_2) - K_n(f_1, f_2).$$

In the above formula, $K_n(f_1, f_2)$ denotes a correction term, for which the following two versions are suggested:

- *Akaike correction*: $K_n^A(f_1, f_2) := k_1 - k_2$
- *Schwarz correction*: $K_n^S(f_1, f_2) := \binom{k_1}{2} \log(n) - \binom{k_2}{2} \log(n)$

In the above, k_1 and k_2 denote the number of parameters in model 1 and in model 2, respectively.

All in all, the Vuong statistic gives us a simple decision rule on hand, where for a given significance level α the decisions are made as follows:

$$\text{prefer ... } \begin{cases} \text{model 1,} & \text{if } \Phi^{-1}(1 - \frac{\alpha}{2}) \leq \nu_n \\ \text{no model,} & \text{if } \Phi^{-1}(\frac{\alpha}{2}) < \nu_n < \Phi^{-1}(1 - \frac{\alpha}{2}) \\ \text{model 2,} & \text{if } \nu_n \leq \Phi^{-1}(\frac{\alpha}{2}) \end{cases} .$$

Chapter 3

Study of Online Sales Activities

3.1 Online sales activities

3.1.1 Data description

In the online shop, the hierarchy for organising the items is as follows:

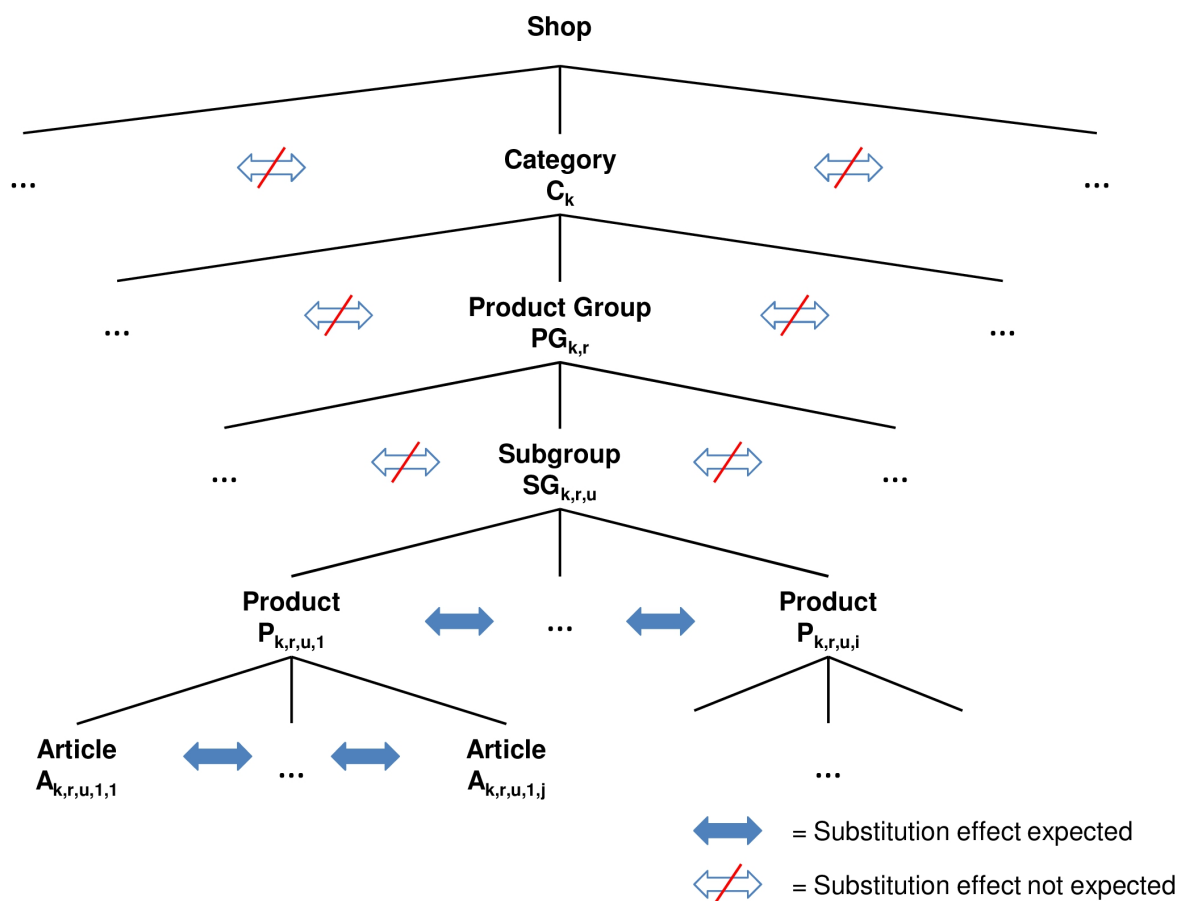


Figure 3.1: Shop structure consisting of several levels: Category, product group, subgroup, product and article.

There are different categories. For each of the categories, there are several product groups, which again consist of subgroups. As next level, there are products. These consist of several product variants, which we will refer to as 'articles' in the following.

Let us explain this shop structure using a supermarket example:

A **category** could be *vegetables* or *beauty care*.

The **product groups** are a rather rough grouping of the products, classifying them by purpose of use. Thus, they are assumed to be completely independent. In our example of the category *beauty care*, two possible product groups could be *deodorants* and *shampoo*.

The **subgroups**, which will be referred to as **types** in the following, pool the products by common characteristics. Again, illustrated at the supermarket's product range, possible subgroups of *deodorants* may be *roll-on deodorant* and *deodorant spray*, as well as a split by brands. The definition of a subgroup aims at bundling the products, which theoretically can be substituted with each other. Usually, customers do not switch between the subgroups due to habit or personal preferences. Hence, also the subgroups are assumed to be independent.

Internally, more than one characteristic are assigned to the products. So evaluations can be handled flexible by picking a specific characteristic as 'subgroup specifier'. Defining subgroups using more than one characteristic, e.g. by multi-level subgroups, would lead to a low number of members, which is not very helpful when doing statistics.

To have the shop as customer-friendly as possible, the client can select products by specifying any of the characteristic to find his preferred product. On the one hand, we have subgroups classifying the products by type and on the other hand, we have an individual subgroup for each brand. This helps the customers to increase their searching efficiency and allows for different search strategies. Since all the different characteristics are maintained for the articles, the customer will find the desired article with any search strategy.

As a consequence of the definition of subgroups, the **products** within one subgroup are comparatively similar. So we cannot assume independence here. Sticking to the deodorant example, a product would be a specific type from a specific manufacturer like a *deodorant for sensitive female skin with rose scent* from one brand. Thus, 'product' is still abstract.

When speaking of **articles**, these are 'concrete' variants of the products, which form the assortment of the shop. Depending on the product, articles have different packaging sizes, colours or are multipacks of other articles. The strongest substitution effects are expected here.

Some articles are special offers and consist of one main item and a so called freebie, which has a relative low value compared to the main item. To have unbiased prices regarding these special offers, the value of the freebie is deducted from the sales price of the main item in the analysis.

In this thesis we analyse one specific product group. The data set retrieved from the online shop covers the daily sales data collected over a time period of two years; more precisely from 01/11/2012 - 31/10/2014, which adds up to a total of 730 days. It covers 287 articles, where we have several data fields available for each article and time point. These are shown in the table below.

| data field | description |
|-------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| DATE | Date of observation. All further field contents refer to this specific date. The data set contains sales data over a time period of 730 days, from 01/11/2012 to 31/10/2014 on daily basis. |
| BRAND | We will consider articles of seven different brands A, B, C, D, E, F and G. |
| TYPE | Independent of brand, the articles are grouped in four types a, b, c and d. |
| PRODUCT | Product to which the article belongs. |
| AS_ID | Unique ID for the articles in the shop. |
| ASH_PRICE | Price at which the articles are sold in the shop. |
| BLACK_PRICE | Recommended retail price. |
| COMP_1_PRICE | Article price from the competitor assumed to be most important (rank 1). |
| COMP_2_PRICE | Article price from the competitor (rank 2). |
| ⋮ | Article price from the competitor (rank ...). |
| COMP_10_PRICE | Article price from the competitor assumed to be least important (rank 10). |
| SHOP_ABSATZ | Quantity sold of the different articles. |
| SALES_BMF | SHOP_ABSATZ times ASH_PRICE. |
| STOCK | Boolean flag: 'TRUE' indicates, that this article is in stock. |
| NL_STATUS | Boolean flag: 'TRUE' indicates that this article was in a newsletter. |
| GENERAL_NL | Boolean flag: 'TRUE' indicates a general newsletter advertisement independent of the articles. |
| BONUS | Boolean flag: 'TRUE' indicates, that a freebie is packed to this article additionally. |
| WEIGHT | Content in UoM kg (of a single pack). |
| BUNDLE_SIZE | Amount of packs, of which the article consists. |

Table 3.1: Available data fields in the data source.

The raw data has to be examined and prepared for the regression analysis.

3.1.2 Data cleaning

The assortment is quite dynamic. Several articles are not available in the shop during the whole period of observation: Some were dropped from the product range or listed as new

products, others run out of stock. Articles which are strongly loaded by these problems would normally be dropped from the data set in order to have more or less complete data for each article.

But, among these articles, there are several special offers which only are available from time to time. These articles cannot be dropped, since they influence the sales significantly and are needed to explain slumps in the sales of the regular articles. Thus, we use a more sophisticated selection method for the articles based on the product affiliation:

As described above, the individual articles are different packaging or bundle sizes of the products. Consider for example an arbitrary product, which is available in three different packaging sizes and from time to time as a special offer, too. Thus, for this product, we have four individual articles in the data set. We then select as follows:

If at each time point under consideration at least one of these four articles is available, all the four of them will be included when modelling. Under this condition, we can create complete data for example by taking average values for the different products, brands or other groups of articles instead of only regarding one article at a time. Using this method, the effect of special offers can easily be covered when modelling.

Otherwise, neither the product nor any of the corresponding articles will be considered. This selection method reduces the amount from 287 articles to 29 articles. In the following all explanations and figures are based on this subset of the data.

Structure of the articles selected for modelling

Having removed the inappropriate products, there remains a total amount of 29 articles selected for modelling. Shown in the hierarchical structure of the shop, this looks as follows:

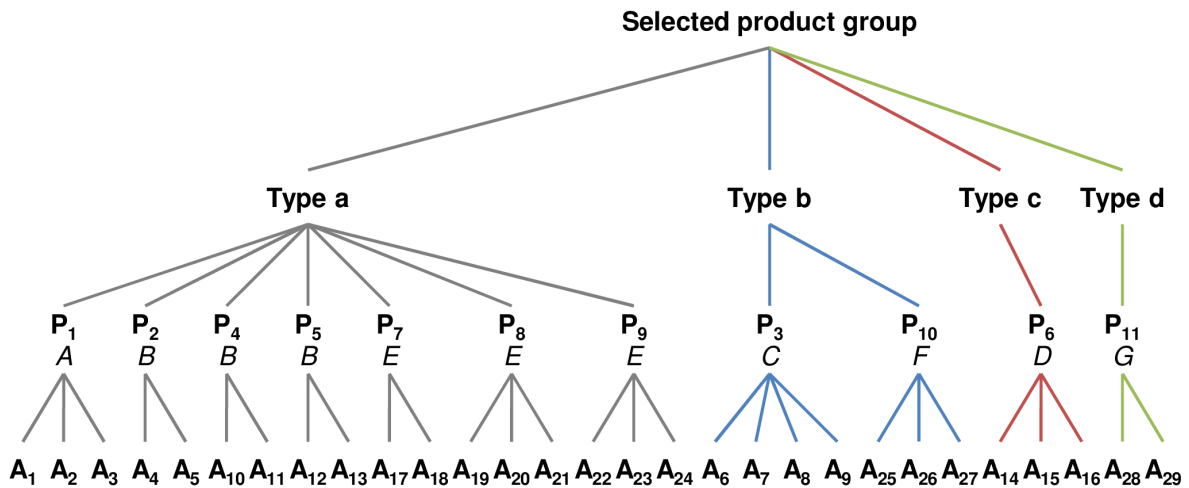


Figure 3.2: Hierarchical structure of the considered product group showing the selected 29 articles. For the ease of use, the brand of the product is always displayed below the product in italics.

We consider the articles A_1, \dots, A_{29} , which come from 11 different products P_1, \dots, P_{11} . The

products come from 7 brands, which are named from A to G and belong to four different types named a, b, c and d.

The table below summarises the articles grouped by type, brand and product, respectively.

| type | brands no. | products no. | articles no. | brand | products no. | articles no. | product | articles no. |
|----------|------------|--------------|--------------|----------|--------------|--------------|----------|--------------|
| a | 3 | 7 | 17 | A | 1 | 3 | P_1 | 3 |
| | | | | B | 3 | 6 | P_2 | 2 |
| | | | | | | | P_4 | 2 |
| | | | | | | | P_5 | 2 |
| | | | | E | 3 | 8 | P_7 | 2 |
| | | | | | | | P_8 | 3 |
| P_9 | 3 | | | | | | | |
| b | 2 | 2 | 7 | C | 1 | 4 | P_3 | 4 |
| | | | | F | 1 | 3 | P_{10} | 3 |
| c | 1 | 1 | 3 | D | 1 | 3 | P_6 | 3 |
| d | 1 | 1 | 2 | G | 1 | 2 | P_{11} | 2 |
| Σ | 7 | 11 | 29 | Σ | 11 | 29 | Σ | 29 |

Table 3.2: Grouping of articles, products and brands in absolute numbers.

| type | brands no. | products no. | articles no. | brand | products no. | articles no. | product | articles no. |
|----------|------------|--------------|--------------|----------|--------------|--------------|----------|--------------|
| a | 43% | 64% | 59% | A | 9 % | 10% | P_1 | 10% |
| | | | | B | 27% | 21% | P_2 | 7% |
| | | | | | | | P_4 | 7% |
| | | | | | | | P_5 | 7% |
| | | | | E | 27% | 28% | P_7 | 7% |
| | | | | | | | P_8 | 10% |
| P_9 | 10% | | | | | | | |
| b | 29% | 18% | 24% | C | 9% | 14% | P_3 | 14% |
| | | | | F | 9% | 10% | P_{10} | 10% |
| c | 14% | 9% | 10% | D | 9% | 10% | P_6 | 10% |
| d | 14% | 9% | 7% | G | 9% | 7% | P_{11} | 7% |
| Σ | 100 % | 100% | 100% | Σ | 100% | 100% | Σ | 100% |

Table 3.3: Grouping of articles, products and brands in percentage.

Sales data of the articles selected for modelling

To get an overview over the sales relevance of the products, the table below contains the average quantity sold for each of the articles. Weight and bundle size are necessary to calculate the basic price for the articles later on. For some articles a freebie is added as a bonus, which can be recognised at the boolean flag *bonus*.

| article | product | type | brand | weight | bundle size | bonus | avg. daily quantity sold |
|---------|----------|------|-------|--------|-------------|-------|--------------------------|
| A1 | P_1 | a | A | 15 kg | 1 | 0 | 45 |
| A2 | P_1 | a | A | 15 kg | 2 | 0 | 72 |
| A3 | P_1 | a | A | 7 kg | 1 | 0 | 30 |
| A4 | P_2 | a | B | 14 kg | 1 | 0 | 33 |
| A5 | P_2 | a | B | 14 kg | 2 | 0 | 155 |
| A6 | P_3 | b | C | 4.5 kg | 1 | 0 | 53 |
| A7 | P_3 | b | C | 9 kg | 1 | 0 | 101 |
| A8 | P_3 | b | C | 18 kg | 1 | 0 | 162 |
| A9 | P_3 | b | C | 18 kg | 1 | 1 | 67 |
| A10 | P_4 | a | B | 14 kg | 2 | 0 | 33 |
| A11 | P_4 | a | B | 14 kg | 1 | 0 | 7 |
| A12 | P_5 | a | B | 14 kg | 2 | 0 | 14 |
| A13 | P_5 | a | B | 14 kg | 1 | 0 | 3 |
| A14 | P_6 | c | D | 5 kg | 3 | 0 | 12 |
| A15 | P_6 | c | D | 5 kg | 1 | 0 | 5 |
| A16 | P_6 | c | D | 5 kg | 2 | 0 | 39 |
| A17 | P_7 | a | E | 15 kg | 1 | 0 | 2 |
| A18 | P_7 | a | E | 15 kg | 2 | 0 | 15 |
| A19 | P_8 | a | E | 15 kg | 2 | 0 | 45 |
| A20 | P_8 | a | E | 15 kg | 1 | 0 | 8 |
| A21 | P_8 | a | E | 15 kg | 2 | 1 | 42 |
| A22 | P_9 | a | E | 14 kg | 1 | 0 | 8 |
| A23 | P_9 | a | E | 14 kg | 2 | 0 | 58 |
| A24 | P_9 | a | E | 14 kg | 2 | 1 | 55 |
| A25 | P_{10} | b | F | 8 kg | 3 | 0 | 2 |
| A26 | P_{10} | b | F | 8 kg | 1 | 0 | 2 |
| A27 | P_{10} | b | F | 8 kg | 2 | 0 | 1 |
| A28 | P_{11} | d | G | 2 kg | 3 | 0 | 8 |
| A29 | P_{11} | d | G | 2 kg | 1 | 0 | 1 |

Table 3.4: This table matches the articles with the corresponding product, type, brand, weight and bundle size. Furthermore, it is displayed, if there is given away a bonus article for free. To get an overview over the importance of the different articles, there is also given the average quantity sold per day.

Stock availability of the articles selected for modelling

The table below gives a short overview over the availability of the articles, where most of them have a good availability. The ones being less than half of the time in stock probably are these special offers of limited duration. During analysis, the days with no availability of an article are ignored when analysing this article. But, when calculating an averaged price for the product containing this article, we always use the data of the available articles per time point. As already mentioned, the selection of articles was done in a way, that on each day at least one of the articles within a product was in stock. Hence, sufficient data is available.

| availability av | products | |
|------------------------|----------|------------|
| | number | percentage |
| $0\% \leq av < 50\%$ | 5 | 17% |
| $50\% \leq av < 80\%$ | 2 | 7% |
| $80\% \leq av < 90\%$ | 1 | 3% |
| $90\% \leq av < 100\%$ | 4 | 14% |
| $av = 100\%$ | 17 | 59% |
| Total: | 29 | 100% |

Table 3.5: Availability of the articles given in absolute numbers and in percentage share.

Competitor prices of the articles selected for modelling

An important question is how to handle competitor prices. There are single days or short periods where the prices for particular articles and competitors are not available for diverse reasons. Furthermore, there might be slight deviations from the actual competitor price for example because of time lags when taking the prices. To have complete data, the lastly reported prices will be taken if no data is available. This compensates the lack of data without removing observations from the data set. Furthermore, competitor prices will be ignored if the deviation from the usual price range of the individual products is too high, because then the data may be erroneous. In this case, also the lastly reported valid price will be taken.

3.1.3 Data exploration

Any shop is heading for high sales, but, in the end, the clients' behaviour determines the sales. Thus, the quantity sold is set as response variable. It 'responds' to the changes made by the shop on directly affectable parameters like for example the prices, which are the so called covariates. In this subsection we will explore the data. Suitable and necessary conversions will be examined.

There are four topics to be worked through for the covariates: the influence of the shop price itself, the effect of the competitor prices, the impact of advertising and the substitution effects.

Quantity sold

The quantity sold will be modelled as response variable. It will be denoted as

$$n_{A_m,t} := \text{quantity sold of article } A_m \text{ at time } t.$$

No matter which kind of shop article is considered, the quantity sold does not count the number of items the considered article consists of, but the number of sales. Thus, if we for example consider a multipack consisting of three identical items, which was sold seven times, then we have a quantity sold of seven, although we actually have sold twenty-one items.

If we are interested in coarsening our models for example by clustering the articles by type or brand, the response variable has to be adjusted. Clustering will merge articles of different packaging sizes, as well as multipacks and special offers. Switching from quantity sold to units sold, which in this context will be kilograms sold, allows for this kind of consolidation. The figure 'kilograms sold' is calculated as 'quantity sold' times 'weight' times 'bundle size'.

To get a first impression of our response variable, the box plots with the standard boundaries 25% and 75%, which are given below, visualise the dispersion of the quantity sold per day and article.

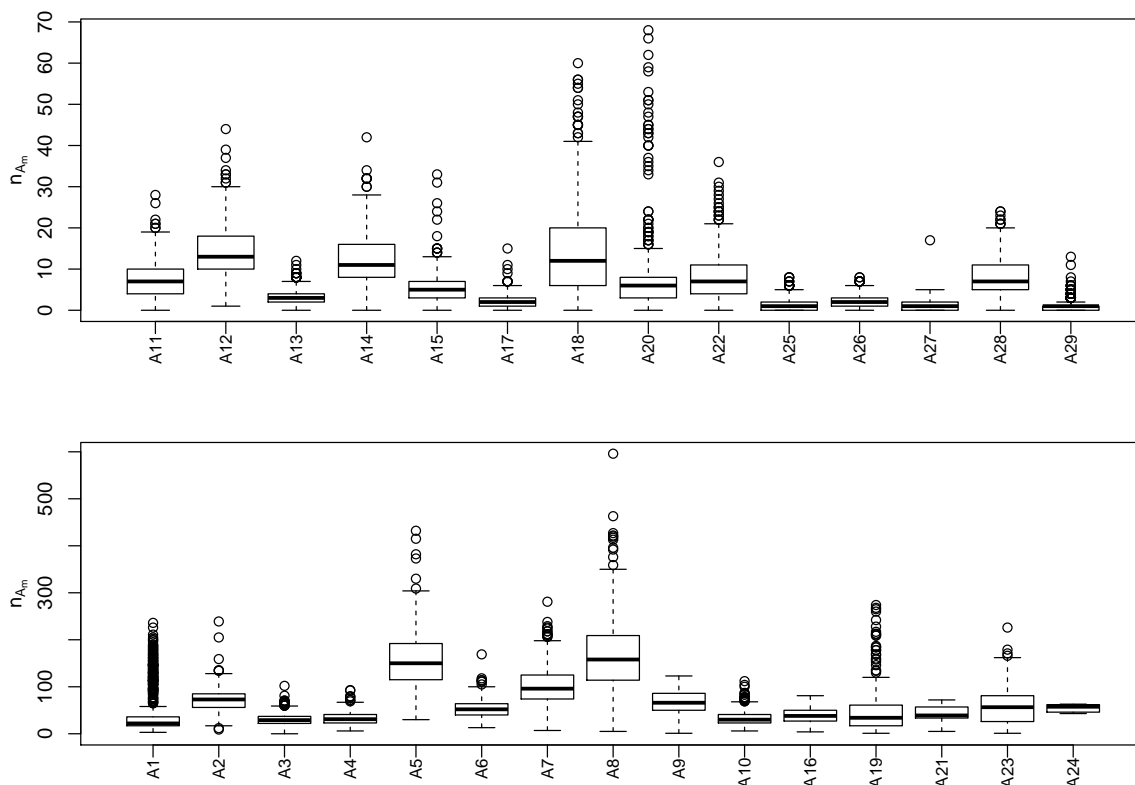


Figure 3.3: Box plots of the quantity sold on daily basis. Each article is displayed separately.

The box plots are slightly skewed (as expected for count data) and show notable outliers in both directions. The outliers showing extremely high quantity sold suggest the assumption, that this is not just a random behaviour, but that this probably is induced by temporary price reductions or advertisement.

Prices

Concerning prices, we have to be careful when measuring the influence of possible substitutes. Consider for example a single article of arbitrary packaging size. Possible substitute products are other packaging sizes or multipacks of the same product and similar articles from other brands and potentially other packaging sizes.

Since we are doing pricing policy, the influence of the substitute products will be measured by comparing prices. To keep things comparable, we will always consider the basic prices, which in this context means the prices per kilogram. This way we avoid the problem concerning different packaging sizes and multipacks.

Nevertheless, using the current article price or the article's basic price will both have the same relative increase or decrease and the result of the regression will be identical.

The freebies coming with special offers are not comparable with the articles considered, because normally these do not come from a related product group. For this reason, the regular selling price of these freebies will be determined and subtracted from the article price before calculating the basic price (per kilogram) of the special offers. Thus, the attractiveness gained by freebies will be included in our models by lower basic prices for the concerned articles.

The variable used for the current price of the articles A_m , where $m = 1, \dots, 29$, will be

$$p_{A_m,t} := \text{shop price of article } A_m \text{ per kilogram at time } t.$$

Typically, for each article there is a set of predefined price values. Of course, these predefined values can also be modified, but this is done rather seldom. The number of predefined prices does vary across the articles. For this reason, we will differentiate between the number of different prices and the number of price changes for each article.

The dynamics in prices are shown in the following table. The number of prices means the number of different prices used per article during the time period of observation. The number of price changes indicates, how many times the price of the article was changed during the time period of observation, where changing a price means to switch from one of the predefined values to another.

| article | number of prices | number of price changes | article | number of prices | number of price changes |
|----------|------------------|-------------------------|---------|------------------|-------------------------|
| A1 | 3 | 5 | A16 | 3 | 2 |
| A2 | 3 | 5 | A17 | 2 | 14 |
| A3 | 3 | 2 | A18 | 1 | 0 |
| A4 | 3 | 4 | A19 | 2 | 3 |
| A5 | 5 | 8 | A20 | 3 | 34 |
| A6 | 3 | 17 | A21 | 1 | 0 |
| A7 | 5 | 26 | A22 | 3 | 47 |
| A8 | 6 | 45 | A23 | 5 | 32 |
| A9 | 2 | 1 | A24 | 1 | 0 |
| A10 | 4 | 5 | A25 | 4 | 5 |
| A11 | 2 | 1 | A26 | 1 | 0 |
| A12 | 3 | 3 | A27 | 1 | 0 |
| A13 | 3 | 2 | A28 | 5 | 9 |
| A14 | 5 | 12 | A29 | 2 | 1 |
| A15 | 3 | 2 | | | |
| Average: | | | | 3 | 9.8 |

Table 3.6: Dynamics of the prices per article: The number of prices gives the cardinality of the set of different prices, whereas a price change means switching between the prices available.

This evaluation shows sufficient dynamics on prices to do regression analysis.

The box plots below give a first insight into the interdependencies between the shop price and the quantity sold. The relative frequencies of the prices are displayed above the plots. Please note, that these relative frequencies always refer to the whole time period of observation. Consequently, they will only add up to 1 for the 17 articles, that were in stock the whole time. Three articles were selected to show the different effect types. The whole set of the plots can be found in the appendix 7.1 to appendix 7.3.

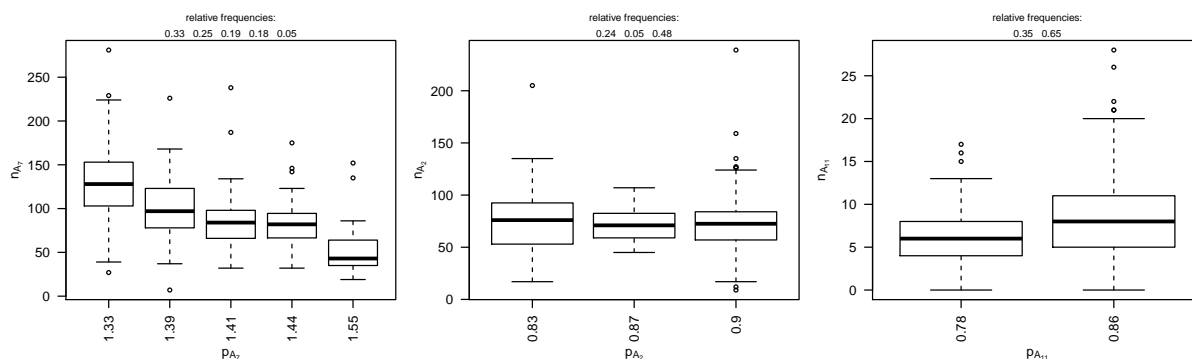


Figure 3.4: Box plots of the quantity sold over the shop price for the individual articles. The data is given on daily basis. The box plots for all of the articles can be found in the appendix 7.1 to appendix 7.3.

The first effect characteristic is decreasing sales at increasing prices. This would be commonly expected. The behaviour was found at 17 articles.

The second characteristic is almost no effect on sales by price changes. This was found ten times. Investigating the concerned articles in more detail, we find, that for 6 out of them a pricing effect cannot be detected, because they did only have one price each.

The third effect is increasing sales at increasing prices. This occurs two times. These articles did only use two prices each and had very low quantities sold. This thus is assumed to be statistical noise.

We furthermore notice, that there are several outliers for most of the articles. These seem to be concentrated at the prices with high relative frequencies. Thus, the more days a price is active in the shop, the more likely it is, that outliers appear at this price.

Price cut

In the shop, there are always displayed two prices: the current shop price and the recommended retail price. One could assume, that customers react on the percentage difference between these two prices. Thus, this price cut could be defined as an independent variable. But models including this as covariate nevertheless show comparatively bad fits. One reason could be, that customers rather compare our current shop price for example to the current prices of other online shops than to the recommended retail price.

Similarly, a price cut between the current shop price and the competitor prices could be used.

But, no matter which price cut is used, since the price cut is calculated as the difference between the current sales price and another price, there is a linear correlation between the price cut and any price used in the subtraction. Since the current price and the competitor price will surely be used as covariates, it is not reasonable to introduce the price cut as an additional covariate.

Competitor prices

Competitor prices are taken for the most important competitors. The number of competitors to be tracked varies from article to article. Prices can only be taken for common brands. How many competitor prices are available, depends on the popularity of both, the brand and the product variant. For private labels of course no competitor prices are available. These thus cannot be compared to other online shops, which exactly is the intention of such products.

The variable for the competitor prices will be defined as

$$c_{A_m,t}^y := \text{price of article } A_m \text{ per kilogram at time } t \text{ at competitor } y.$$

The bar plot below gives an overview over the number of competitors to be tracked for each article.

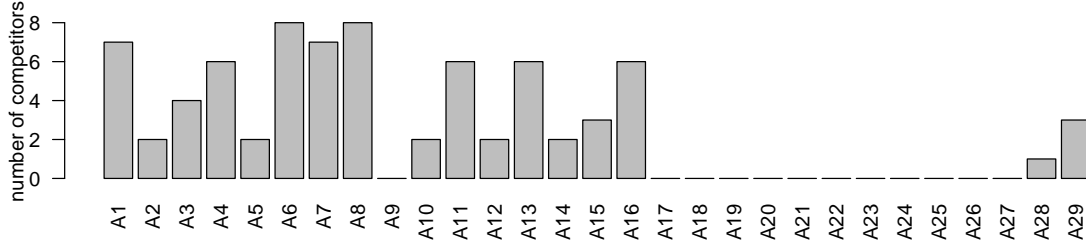


Figure 3.5: Bar plot showing the number of competitors to be tracked for each article.

The number of competitor prices available obviously varies a lot. Even if it was the same for any two articles, the competitors could be different ones. The influence of the competitor prices might furthermore depend on the market share and on the prominence of the competitor.

Concerning competitors, different influences are conceivable: The minimal competitor price, the average competitor price or the competitors' price range could be considered. As these are highly correlated, it is not reasonable to include all of them in regression analysis. For this reason, we will restrict to the minimum competitor price, because this seems to be the most promising one.

Minimum competitor price

The minimum competitor price for article A_m at time t is defined as

$$c_{A_m,t}^m := \min_y(c_{A_m,t}^y),$$

where we do not consider the raw article price, but the price per kilogram. Thus, it is the minimum out of the competitor prices per time point, i.e. it is the currently lowest price amongst all competitors. It does not represent the price of one specific competitor like the cheapest in general, the cheapest averaged over time or across products.

The three dimensional scatter plots below visualise the dependence between the shop price, the cheapest competitor price and the sales. For each unique pair $(p_{A_m}^*, c_{A_m,t}^{m*})$ of shop price and cheapest competitor price the average sales are taken, i.e. we consider

$$\bar{n}_{A_m, p_{A_m}^*, c_{A_m,t}^{m*}} := \frac{\sum_t 1_{\{p_{A_m,t} = p_{A_m}^*, c_{A_m,t}^m = c_{A_m,t}^{m*}\}} \cdot n_{A_m,t}}{\sum_t 1_{\{p_{A_m,t} = p_{A_m}^*, c_{A_m,t}^m = c_{A_m,t}^{m*}\}}}.$$

Of course, only articles with competitor prices available will be considered. The plots below confirm the uplift of the average sales for a decreasing shop price and an increasing minimum competitor price.

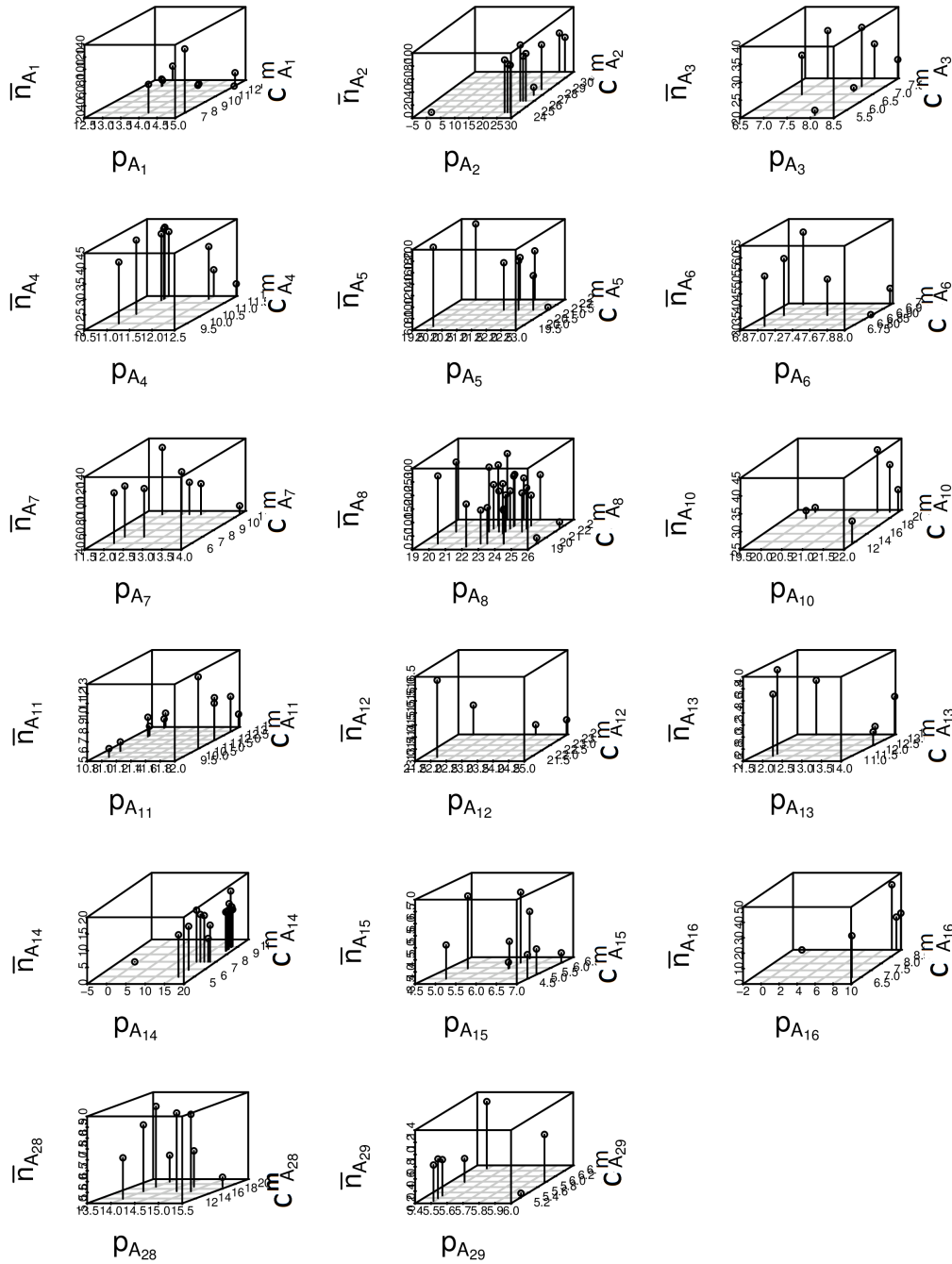


Figure 3.6: Three dimensional scatter plots of the average daily sales for every pair of shop price and currently cheapest competitor price. Of course, only the articles with competitor prices available are considered.

Article-specific newsletter advertisement

Article-specific newsletters inform about currently reduced prices and motivate to buy by emphasising the limited validity of these special offers. The variable contained in the data set is a boolean flag, indicating whether the product was in the newsletter that day or

not, i.e.

$$\text{ad}_{A_m,t} = \begin{cases} 1, & \text{if article } A_m \text{ is in newsletter at time } t \\ 0, & \text{otherwise} \end{cases}$$

Since this variable is on nominal scale, a scatter plot would not be helpful. Instead of that, we have a look at the number of days having the article in the newsletter and the average daily sales with and without article-specific newsletter advertisement.

To get the bar plot below, we sum up over the above defined variable, i.e. we consider

$$\text{ad}_{A_m}^* = \sum_{t \in [1;730]} \text{ad}_{A_m,t}.$$

Thus, for each of the articles, $\text{ad}_{A_m}^*$ gives the overall number of days having this article in the newsletter advertisement.

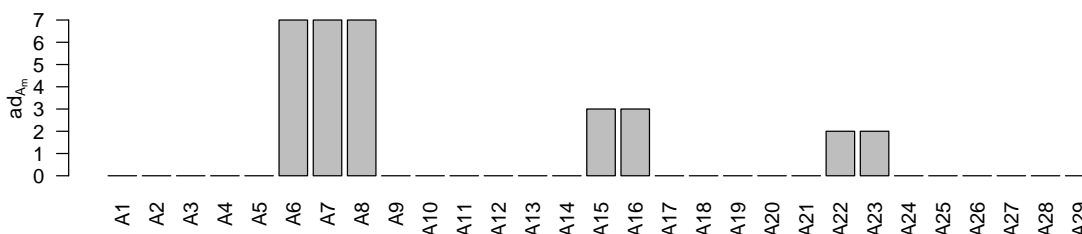


Figure 3.7: Bar plot showing the number of days on which the individual articles appeared in the article-specific newsletter.

Most of the considered articles do never appear in the article-specific newsletter during the whole time period of consideration, because normally only articles, which are classified as 'top sellers' are promoted with this kind of advertisement as a specific shop strategy.

Nevertheless, compared to the time period of observation of 730 days, having an overall number of at the longest 7 days of article-specific newsletter advertisement for an article is a comparatively low number, since this is less than 1% of the time.

General advertisement

Besides this, there are also special offers, which affect the whole assortment. These are for example price coupons or bonus articles, that are given away for free, when achieving a certain minimum order value independent of the products you buy.

In the following, the general advertisement will be denoted by

$$\text{ad}_t^g = \begin{cases} 1, & \text{if there is any general advertisement at time } t \\ 0, & \text{otherwise} \end{cases}$$

Within the time period of consideration of 730 days, there are 157 days with general advertisement. This is a share of approximately 21,5% of the time.

Relative uplift of sales induced by advertisement

In the following, we investigate the relative uplift of sales induced by newsletter advertisement. We distinguish between days with article-specific newsletter advertisement, with general newsletter advertisement, with both kinds of newsletter advertisement at the same time and with no newsletter advertisement.

Let

$$k_{A_m} := \text{number of days with article } A_m \text{ being in advertisement,}$$

$$k^* := \text{number of days with general advertisement,}$$

and

$$k^{**} := \text{number of days with both article-specific and general advertisement.}$$

Then, the relative uplift of sales induced by article-specific newsletter advertisement is defined as

$$up_{A_m} = \frac{N_{1,A_m}}{N_{0,A_m}},$$

where N_{1,A_m} represents the average number of sales for article A_m when appearing in newsletter, i.e.

$$N_{1,A_m} = \frac{1}{k_{A_m}} \sum_t 1_{\{\text{ad}_{A_m,t}=1, \text{ad}_t^g=0\}} \cdot n_{A_m,t},$$

and N_{0,A_m} represents the average number of sales for article A_m without any newsletter advertisement at all, i.e.

$$N_{0,A_m} = \frac{1}{730 - k_{A_m} - k^* - k^{**}} \sum_t 1_{\{\text{ad}_{A_m,t}=0, \text{ad}_t^g=0\}} \cdot n_{A_m,t}.$$

Similarly, for the relative uplift induced by general advertisement we have

$$up_{A_m}^g = \frac{N_{1,A_m}^*}{N_{0,A_m}^*},$$

where N_{1,A_m}^* stands for the average sales of article A_m when general advertisement is present, i.e.

$$N_{1,A_m}^* = \frac{1}{k^*} \sum_t 1_{\{\text{ad}_{A_m,t}=0, \text{ad}_t^g=1\}} \cdot n_{A_m,t}.$$

In the case, that both kinds of newsletter advertisement are active at the same time, we calculate

$$up^{both} = \frac{N_{1,A_m}^{**}}{N_{0,A_m}^{**}},$$

where

$$N_{1,A_m}^{**} = \frac{1}{k^{**}} \sum_t 1_{\{ad_{A_m,t}=1, ad_t^g=1\}} \cdot n_{A_m,t}.$$

The three box plots below summarise the results. The first box plot shows the relative uplift of sales induced by article-specific newsletters. Of course only articles appearing at least once in the article-specific newsletter are considered. The second one summarises the effect of general advertisement, where all of the 29 articles are taken into consideration. The third one visualises the effect, which is brought about when article-specific advertisement and general newsletter are active at the same time, where again of course only articles that were at least once in the article-specific newsletter advertisement are considered.

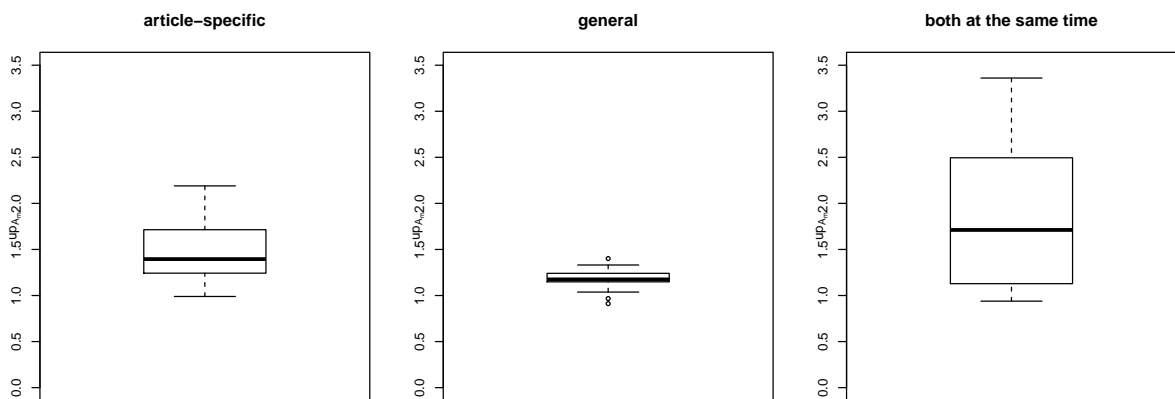


Figure 3.8: Box plots showing the relative uplift of sales induced by the article-specific advertisement, by general newsletters and by both kinds of newsletter at the same time, respectively. On the left and on the right, only articles, which appear at least once in the article-specific newsletter are included. In the middle, all the 29 articles are considered.

Obviously, general advertisement also influences sales significantly, but not as much as the article-specific advertisement does. Combining both kinds of newsletter advertisement on average yields the highest number of sales. When comparing the box plots, it has to be kept in mind, that an article-specific advertisement only lifts up the quantity sold for *one* article on average by 40%. General advertisement however lifts up sales for *all* articles on average by 18%. Just as a very rough estimate: If the shop only offered these 29 articles, the total shop sales would raise by 18% on general advertisement and by $40\% / 29 = 1.4\%$ on article-specific newsletter. So from a global perspective, general advertisement has a much greater lever than the article-specific advertisement.

Below, the effect of the different kinds of advertisement is compared for the articles appearing at least once in the article-specific newsletter.

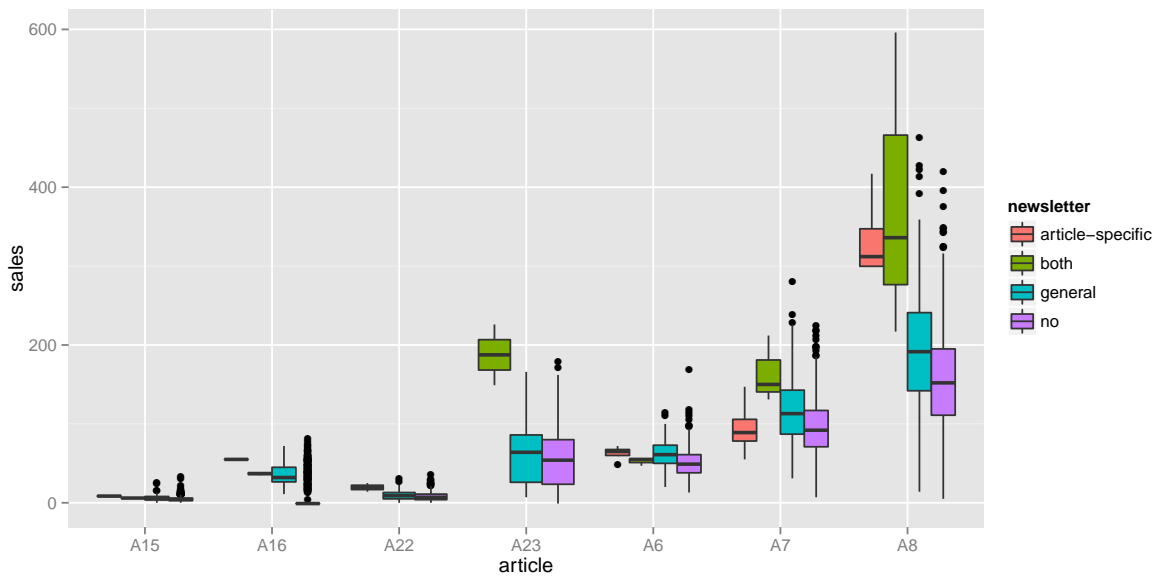


Figure 3.9: Box plots visualising the effect of the different kinds of advertisement on the articles appearing at least once in the article-specific newsletter.

The plot above highlights the surpassing effect of the combination of both kinds of newsletter advertisement. Furthermore, we read out, that the general advertisement seems to have a similar effect on all of the articles. It makes its presence felt in the quantity sold, but not as much as the combination of both kinds of advertisement does. Concerning the article-specific advertisement, we cannot make out a clear pattern of effect across the articles.

We furthermore notice, that there are almost no outliers present in case of article-specific newsletter advertisement and in case of the combination of both kinds of advertisement. On the one hand we could of course argue, that article-specific newsletter advertisement always causes a peak in sales, which makes the data points lie closely together, but we also have to keep in mind, that due to the sparse presence of the article-specific advertisement, the corresponding box plots are drawn from a very small number of data points, which makes outliers rather unlikely to appear.

The two plots below visualise the effect of advertisement on the articles without article-specific newsletter advertisement. On these articles, too, general advertisement has a similar effect across all the articles considered. Like in the plot above, there can be read out a clear uplift of the quantity sold by this kind of advertisement.

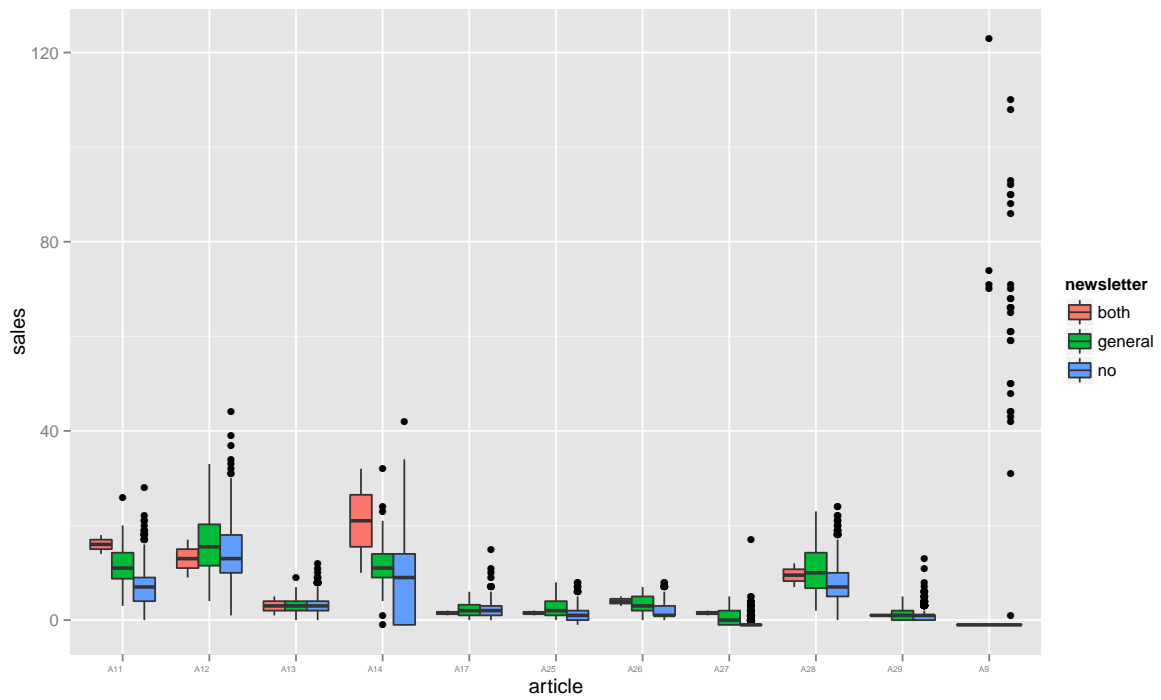


Figure 3.10: Part 1: Box plots visualising the effect of the different kinds of advertisement for the articles without article-specific advertisement.

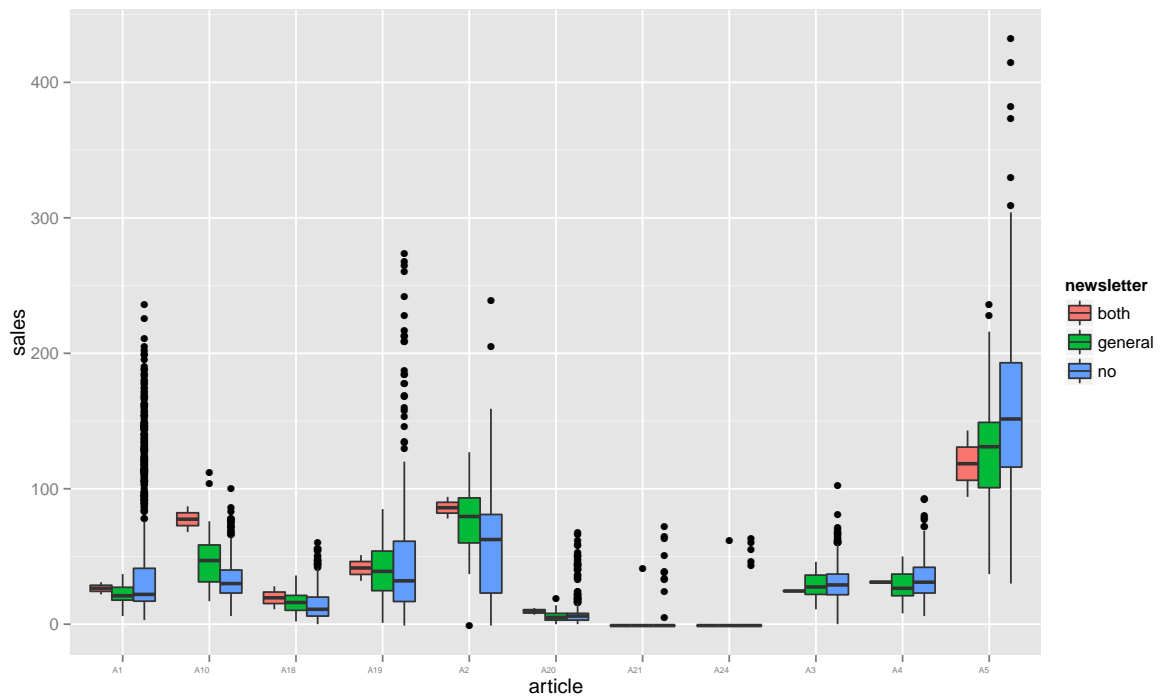


Figure 3.11: Part 2: Box plots visualising the effect of the different kinds of advertisement for the articles without article-specific advertisement.

Concerning the combination of both kinds of advertisement, we typically would expect a lower uplift of the quantity sold by this kind of advertisement than by general newsletters, since the considered articles are not included in the article-specific newsletter advertisement. So, in case of the combination of both kinds of advertisement, there is some general purchasing incentive, but, at the same time, there is article-specific advertisement for closely related articles. Thus, the financially most worthwhile buying behaviour would be to benefit from the general offer via buying the low-priced articles, which appeared in the article-specific newsletter advertisement. This obviously is not the case. For some of the articles, the combination of both kinds of newsletter advertisement, where the considered articles are not included in the article-specific advertisement, induces an even greater uplift of sales than the general advertisement does. A plausible explanation would be, that customers are tempted to the online shop by the low-priced articles in the article-specific newsletter advertisement, but then decide to buy another related article.

Refining the box plots above, we could also consider the newsletter induced uplift of sales multiplied by the depth of discount. This could be justified arguing that the effect of advertisement strongly depend on the attractiveness of the current price. But, this would result in strong correlations between the covariates 'depth of discount' and 'shop price'. Thus, weighting advertisement by the depth of discount does not bring additional information into the model compared to just taking the raw prices and a boolean flag for advertisement.

Substitution effects

An important issue is to investigate the substitution effects between the different articles. It has to be taken into account, that changing the prices of the individual articles will probably make a part of the customers switch to alternative products with more attractive prices. Thus, for optimal pricing, it is important to know about these dynamics to be able to control the so-called substitution effects. The aim here is to get a kind of substitution effect matrix, i.e. a matrix, which represents the strength of influence of the articles each on another. This matrix does not necessarily have to be symmetric: It is easy conceivable, that price induced switching from lower quality to higher quality is much more natural than switching from high quality to cheaper products: Imagine we usually buy a rather cheap product. If then there is a special offer for some premium brand, temporarily sold at a price below the regular price of the cheaper product, we are very likely to change to the premium brand and then back to our cheap product, when the special offer has ended. Vice versa, if we usually buy the premium brand, we are much more resistant to price levels, since we then obviously make decisions based on quality characteristics and not on prices. Only if the prices get extremely unattractive or if our product is temporarily not available we are about to look for alternative products.

To get more detailed information about substitution, we consider the following substitution possibility matrix. Depending on how strongly the products are related to each other, some clustering can be done. Considering one specific article, it is not far to seek, that the other articles belonging to this product may have the most significant influence. If for example the article we are intended to buy is available as a single pack and as a

multipack at different basic prices we will reasonably take the article with the lower basic price. Only in case that even the cheapest version of our desired product is not cheap enough, we tend to look for product alternatives. Subject to brand awareness, we probably first look for alternatives within our usual brand and then expand our searching to products of the same type.

Hence, if two articles are of same type, brand or product, these are considered as possible substitutes. This will be indicated by '1' if two articles are of the same type only, by '2' if they are also of the same brand and by '3' if they are even of the same product. Thus, for the substitution possibility matrix S , we define

$$s_{ij} = \begin{cases} 0, & \text{for unrelated articles} \\ 1, & \text{for articles of the same type} \\ 2, & \text{for articles of the same type and brand} \\ 3, & \text{for articles of the same type, brand and product} \end{cases} .$$

In contrast to the substitution effect matrix, which we want to quantify via regression analysis, the substitution possibility matrix of course has to be symmetric.

$$S = \begin{pmatrix} 0 & 3 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 3 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 2 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 2 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 3 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 3 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 3 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 3 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 3 & 0 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 2 & 2 & 3 & 0 & 3 & 2 & 2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 2 & 2 & 3 & 3 & 0 & 2 & 2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 2 & 2 & 2 & 2 & 2 & 2 & 0 & 3 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 3 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 3 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 \end{pmatrix} \in \mathbb{R}^{29 \times 29}$$

To keep the number of covariates on a reasonably low level, we pool the remaining articles of the same product within one variable and capture possible substitutes of other products per brand and not per product or article. Above all, if we have to deal with special offers, this seems to be a promising approach: Special offers usually are only available for a short period within the time period of consideration. Due to this rare data, it would be difficult to examine the substitution effect of one single special offer article. Nevertheless, these articles normally affect sales significantly, which can easily be brought into the model when considering the whole brand instead of the individual articles. For this purpose, either the minimum basic price or the average basic price of the considered brand could be used. Like in case of the competitor prices, we will consider the minimum basic price of the substitutes in the following.

3.2 Count regression models for the daily number sold per article

In this subsection, we build some models, which predict the sales for article A_6 . Our response variable thus will be the quantity sold $n_{A_6,t}$ for $t = 1, \dots, 730$. We have to fit a model to count data, which will be done using Poisson regression and Negative Binomial regression.

As we understand from the graphic below, the considered article A_6 is one out of the four articles belonging to product P_3 . Product P_3 together with product P_{10} make up the subgroup 'type b'. Concerning substitution effects, we have the three remaining articles of the 'original' product P_3 , and the three articles of the substitute product P_{10} , where P_{10} was identified as a substitute as shown in the substitution possibility matrix introduced above. To keep the number of covariates on a desirably low level, we will not consider the substitution effect of each individual article, but the influence of the 'original' product P_3 , and of the substitute product P_{10} , respectively, by one covariate each. So, we stay with two covariates to cover possible substitution effects.

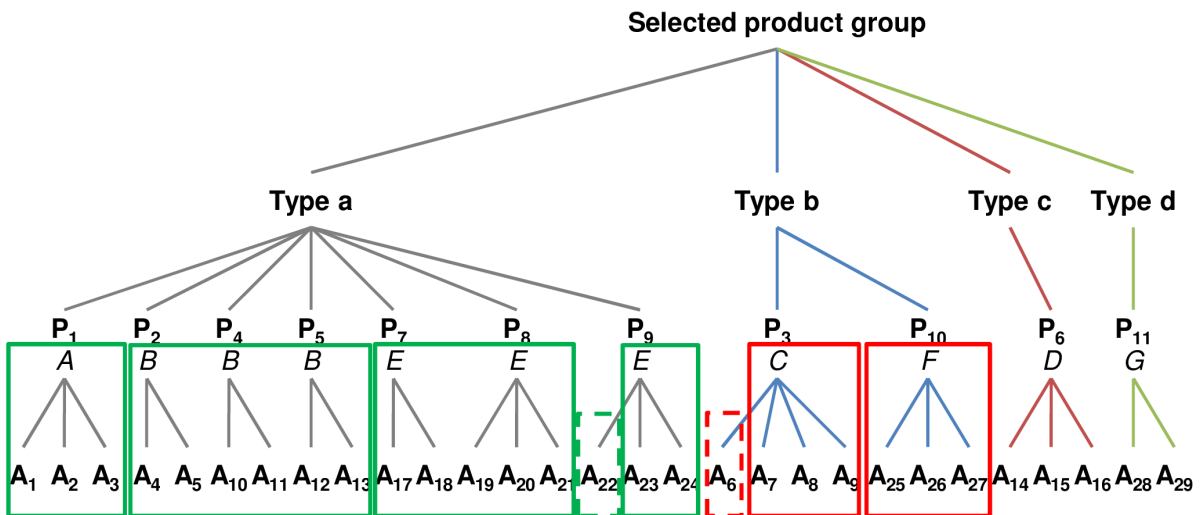


Figure 3.12: This graphic exemplary illustrates the grouping of the articles to form suitable covariates covering possible substitution effects. Each of the green boxes forms one covariate, when analysing article A_{22} ; each of the red boxes for article A_6 , respectively.

3.2.1 Exploratory data analysis

In order to set up a model properly, certain preconditions and dependencies have to be explored. In the following, we inspect the covariates as well as interactions and possible time effects.

Choosing the covariates

As described in chapter 2, we have to check, if for $i \in \{1, \dots, n\}$ the plots of x_{ij} versus $\log(y_i) - \log(t_i)$ show a linear dependency. Due to $t_i = 1$ for all $i = 1, \dots, n$ we always have $\log(t_i) = 0$ and hence can skip this term.

Since one of our major goals is to find out the price elasticity of the different articles, we aim at bringing all covariates concerning prices into the model as logarithmised values. So, for the covariates 'article price', 'minimum competitor price', 'minimum price of the other articles belonging to the same product' and 'minimum price of the substitute product P_{10} ', we check the plots of $\log(x_{ij})$ versus $\log(y_i)$ instead of x_{ij} versus $\log(y_i)$. If these plots indicate a linear relationship, we will not check the plots for the 'raw' variables or other transformations.

Since we only have discrete values v_k for the j -th covariate x_{ij} , $i \in \{1, \dots, n\}$, we calculate the mean values of the corresponding subsets of y_i for each discrete value of \mathbf{x}_i as

$$y_k = \frac{\sum_{i=1}^n 1_{\{x_{ij}=v_k\}} \cdot x_{ij}}{\sum_{i=1}^n 1_{\{x_{ij}=v_k\}}}.$$

Without loss of generality, we can use these mean values y_k together with the corresponding confidence bands instead of plotting all the values. This makes the plots easier legible.

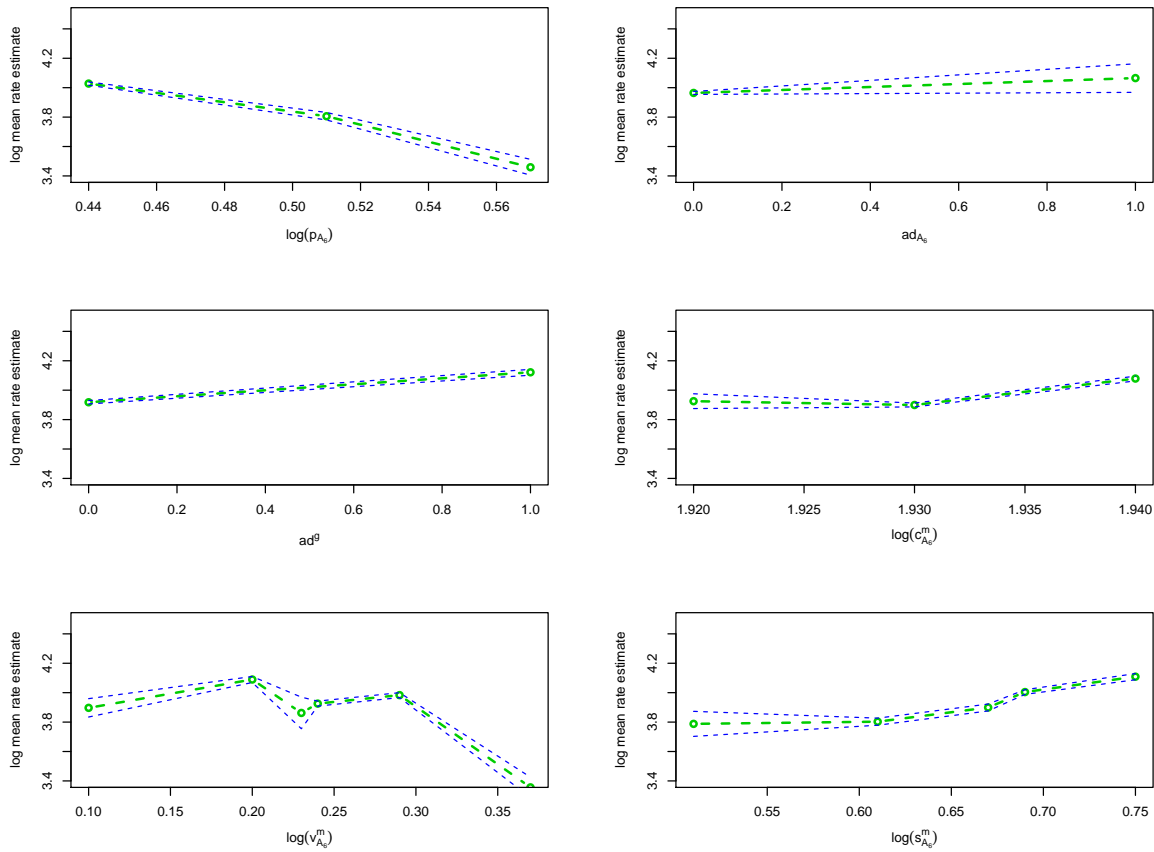


Figure 3.13: EDA-plots of the covariates versus the log mean rate estimate.

For the article price, we do only have three different values. Such being the case, we abstain from bringing the article price into the model on logarithmic scale. Although this way we do not get the price elasticity out of the regression coefficient, we treat the article price as factor variable instead.

The variables concerning advertisement do only have two levels each. This makes it impossible to decide about the linearity. Being convinced, that advertisement does effect our response variable, these two variables nevertheless will be used for modelling.

Regarding the minimum competitor price, the minimum price of the remaining articles belonging to product P_3 and the minimum price out of the articles belonging to the substitute product P_{10} , we detect several problems: At first, we notice, that for the minimum competitor price, we do only have three unique values, too. Thus, it is not far to seek, to treat it as factor variable. But, since the plots of the minimum price out of the remaining articles belonging to product P_3 as well as the minimum price among the articles belonging to the substitute product P_{10} neither are that convincing, we check, if we can improve the plots when modifying these covariates.

For the minimum price of the remaining articles belonging to product P_3 , we could try

to find a suitable variable transformation. But, due to the small range of the prices, the non-linearity of the logarithm and square root does not have a visible effect when transforming the variable. Hence, using square root or the non-transformed prices instead of the logarithmised values will not help to improve the plots.

Instead, we check, if there are modified forms of the three covariates, that are more suitable for modelling.

For each time point t , we define the average competitor price for article A_6 as

$$c_{A_6,t}^a := \frac{1}{y_{A_6}^*} \sum_y c_{A_6,t}^y,$$

the average price of the remaining articles belonging to the same product, to which article A_6 belongs, as

$$v_{A_6,t}^a := \frac{1}{\sum_{m=7}^9 1_{\{p_{A_m,t} > 0\}}} \sum_{A_m=7}^9 p_{A_m,t},$$

and the average price for the substitute product as

$$s_{A_6,t}^a := \frac{1}{\sum_{m=25}^{27} 1_{\{p_{A_m,t} > 0\}}} \sum_{A_m=25}^{27} p_{A_m,t}.$$

Using these average prices instead of the minimum values yields the following plots:

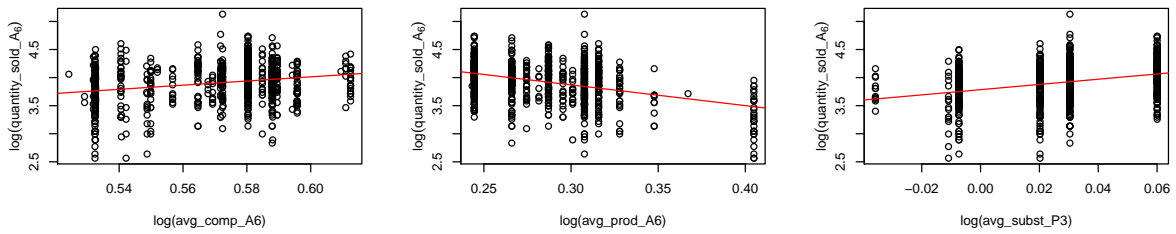


Figure 3.14: Scatter plots of the response variable against the average prices of comparable articles.

Taking the average prices instead of the minimum prices obviously gives us more different values for the covariates. Hence, there is no need any more to treat one of them as factor variable. Furthermore, checking the Pearson product-moment correlation coefficient, we find, that the correlation improves slightly. This tells us, that the relationship between the logarithm of the response variable and the average prices seems to be more linear than in case of the minimum values. Such being the case, we will always consider the average prices in the following.

Time effects

Since our data was collected over a time period of two years, we possibly have to deal with seasonality.

As a first approach, we cluster the data by month, and check, if we detect any pattern in the plot.

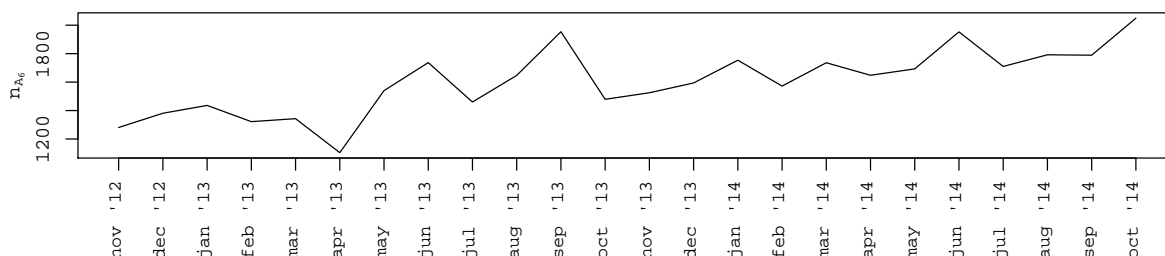


Figure 3.15: Quantity sold of article A_6 clustered by month.

This does not yield a very clear pattern and seems to be random. Monthly seasonality thus will not be included into the models. But, sales obviously increase over time. This of course has to be covered by an appropriate covariate.

Another approach regarding time effects is, to look for weekday seasonality. For this purpose, we cluster the data by year and weekday.

We have to keep in mind, that in 2012 the data was collected for eight weeks only, so this does not reflect a whole year. In 2013 the data has been collected for the full year and in 2014 for ten months. To have evenly spread data, we will not cluster by the year in the classical sense, but we will split the data into two parts, one from the beginning of November 2012 to the end of October 2013 and one from the beginning of November 2013 to the end of October 2014. This way we end up with two 'full' years.

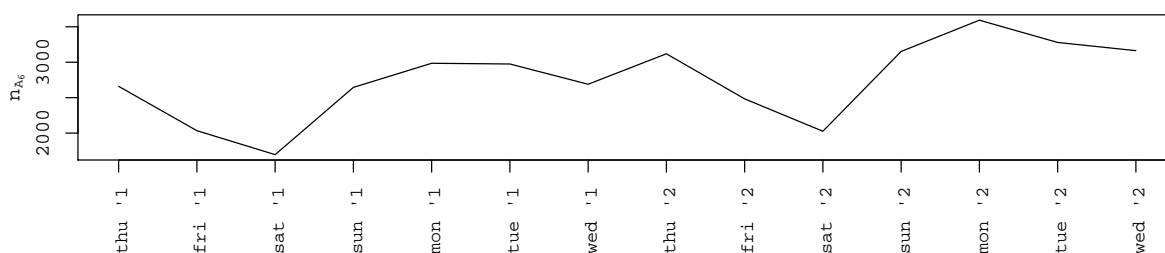


Figure 3.16: Quantity sold of article A_6 clustered by weekday and year 1 and year 2, respectively.

We read out from the plot, that sales are rather poor on Friday and Saturday. We thus will cluster these two days of the week together as 'weekend' and use this as a covariate to cover the weekday seasonality when modelling.

Interaction effects

Next, we check, if it is advisable to include some interaction terms when modelling. For reasons of interpretation, we always consider centred variables in this context.

The matrix below shows all possible pairwise interaction plots.

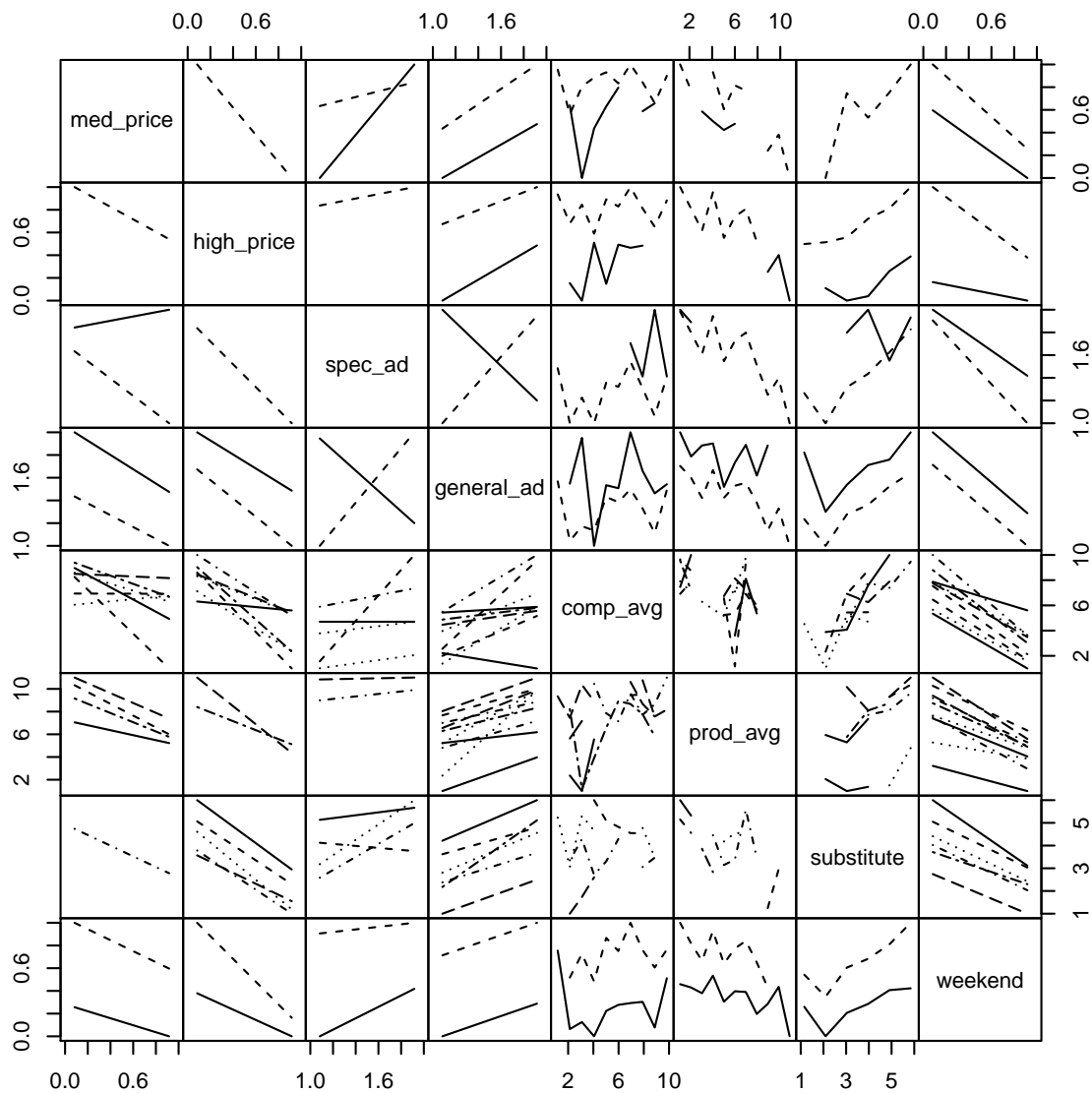


Figure 3.17: Interaction plot matrix for the set of covariates.

From this plot, we can roughly deduce, between which of our covariates there is some interaction present.

The question arising is, which of these interaction effects are statistically significant and

consequently should be included into the model. This will be examined when doing the regression.

3.2.2 Poisson regression models for article A_6

In this subsection, we compare different Poisson regression models for article A_6 and decide, which one fits best to our data. At first, we perform a simple Poisson generalised linear model. After that, we check, if we can improve the fit by including time effects and interactions.

The simple Poisson regression model, which we want to start with, looks as follows:

$$n_{A_6,t} \sim \text{Poisson}(\mu(\mathbf{x}_t, \boldsymbol{\beta})), t = 1, \dots, 730,$$

where

$$\mu(\mathbf{x}_t, \boldsymbol{\beta}) = e^{\{\mathbf{x}_t \boldsymbol{\beta}\}},$$

and

$$\mathbf{x}_t \boldsymbol{\beta} = \beta_0 + \beta_1 \cdot p_{A_m,t} + \beta_2 \cdot \text{ad}_{A_6,t} + \beta_3 \cdot \text{ad}_t^g + \beta_4 \cdot \log(c_{A_6,t}^a) + \beta_5 \cdot \log(v_{A_6,t}^a) + \beta_6 \cdot \log(s_{A_6,t}^a).$$

Before we start analysing this model, let us have a closer look at the covariates. We differ between quantitative and qualitative covariates and match them with the classical levels of measurement. For each time point t , we have:

| covariate | classification | scale of measurement | description |
|---------------------|----------------|----------------------|----------------------------------------------------------------------------------------------------------|
| $n_{A_6,t}$ | quantitative | ratio scale | discrete number representing the quantity sold |
| $p_{A_m,t}$ | qualitative | ordinal scale | factor variable covering the current shop price per kilogram with three different 'levels' |
| $\text{ad}_{A_6,t}$ | qualitative | nominal scale | boolean flag representing the presence respectively absence of article-specific newsletter advertisement |
| ad_t^g | qualitative | nominal scale | boolean flag indicating the presence respectively absence of general advertisement |
| $\log(c_{A_6,t}^a)$ | quantitative | ratio scale | logarithm of the average competitor price per kilogram |
| $\log(v_{A_6,t}^a)$ | quantitative | ratio scale | logarithm of the average price of the remaining product variants of product P_3 per kilogram |
| $\log(s_{A_6,t}^a)$ | quantitative | ratio scale | logarithm of the average price of the substitute product P_{10} per kilogram |

Table 3.7: Statistical classification and description of the variables used for modelling.

The simple Poisson regression model with the variables as described in the table above yields the following regression results:

| | Estimate | Std. Error | z value | Pr(> z) |
|--------------------|-----------------------------------|------------|---------|----------|
| (Intercept) | 2.50 | 0.33 | 7.46 | 0.00 |
| medium price | -0.09 | 0.02 | -4.25 | 0.00 |
| high price | -0.43 | 0.03 | -12.48 | 0.00 |
| spec_nl_A6 | 0.03 | 0.05 | 0.61 | 0.54 |
| general_nl | 0.18 | 0.01 | 15.09 | 0.00 |
| log(avg_comp_A6) | -0.38 | 0.28 | -1.33 | 0.18 |
| log(avg_prod_A6) | -0.91 | 0.24 | -3.86 | 0.00 |
| log(avg_subst_P3) | 2.69 | 0.38 | 7.14 | 0.00 |
| Null deviance: | 4520.32 on 729 degrees of freedom | | | |
| Residual deviance: | 3481.87 on 722 degrees of freedom | | | |
| AIC: | 7691.418 | | | |

Table 3.8: Regression results for article A_6 in the simple Poisson regression model.

We read out from the table, that the p-values are on a very low level for most of the covariates. Hence, performing the test $H_j : \beta_j = 0$ against $K_j : \beta_j \neq 0$, the hypothesis can be rejected at a significance level of $\alpha = 0.001$ and smaller for $j \in \{0, 1, 2, 4, 6, 7\}$.

The article price comes into the model as factor variable. Thus, the low p-values of the different price levels tell us, that the means of the medium and the high price level are significantly different from the mean of the low price, which is the reference level. But, these two significances can only tell us about the pairwise differences between the levels. To find out whether the factor variable 'article price' as a whole is significant, we have to test whether there is any heterogeneity in the means of the levels of the article price. This can be done using a χ^2 -test, which we use to compare our current Poisson model including the factor variable 'article price' and the model without the factor variable 'article price'.

| | Resid. Df | Resid. Dev | Df | Deviance | Rao | Pr(>Chi) |
|---|-----------|------------|-------|----------|---------|----------|
| 1 | 722.00 | 3481.87 | | | | |
| 2 | 724.00 | 3665.03 | -2.00 | -183.17 | -173.86 | 0.00 |

Table 3.9: χ^2 -test of the models with and without the factor variable 'article price', respectively.

As the p-value is clearly smaller than the significance level of 0.05, we do reject the null hypothesis, and conclude, that the article price as a whole is highly significant.

The standard errors in our regression results are a first hint, that the Poisson model is no good fit. In case of good fit, we would expect them to be at least one order of magnitude below the regression coefficients. This is not fulfilled for the medium price, for the article-specific newsletter advertisement, for the average competitor price and for the

average price of the remaining articles belonging to product P_3 .

The summary further yields the residual deviance, which here is rather high compared to the number of degrees of freedom. If the residual deviance divided by the number of degrees of freedom was close to one, this would indicate a good fit of the Poisson model. Having a value of approximately 4.8, we are far away from that. This also tells us, that we should try to find a more appropriate model. Hence, we will now check, if we can improve the fit by including time effects.

The exploratory data analysis already hinted at a weekend seasonality and at an overall trend. These effects will come into the model by including a trend variable and a dummy variable to cover the weekend seasonality.

| | Estimate | Std. Error | z value | Pr(> z) |
|--------------------|-----------------------------------|------------|---------|----------|
| (Intercept) | 3.24 | 0.47 | 6.82 | 0.00 |
| medium price | -0.07 | 0.02 | -3.35 | 0.00 |
| high price | -0.43 | 0.04 | -12.02 | 0.00 |
| spec_nl_A6 | 0.01 | 0.05 | 0.19 | 0.85 |
| general_nl | 0.09 | 0.01 | 7.80 | 0.00 |
| log(avg_comp_A6) | -0.41 | 0.29 | -1.39 | 0.16 |
| log(avg_prod_A6) | -0.47 | 0.27 | -1.74 | 0.08 |
| log(avg_subst_P3) | 1.56 | 0.63 | 2.47 | 0.01 |
| weekend | -0.35 | 0.01 | -27.76 | 0.00 |
| trend | 0.00 | 0.00 | 2.47 | 0.01 |
| Null deviance: | 4520.32 on 729 degrees of freedom | | | |
| Residual deviance: | 2661.35 on 720 degrees of freedom | | | |
| AIC: | 6874.905 | | | |

Table 3.10: Regression results for article A_6 in the Poisson regression model including time effects.

The influence of the weekend seasonality on the quantity sold obviously is strongly supported by the model. Regarding the significance of the covariates, which were already included in the simple Poisson model, we do not detect remarkable changes for the article price, for the variables covering advertisement and for the average competitor price. But, the average price of the remaining product variants of product P_3 loses significance, and the average price of the substitute product P_{10} becomes less significant compared to the results from the simple Poisson regression model without any time effects.

For the standard errors, we stay with the problem of the comparatively high orders of magnitude for some of the covariates, but dividing the residual deviance by the number of degrees of freedom, we get a much smaller value, which now is about 3.70. Further, the AIC has decreased distinctly.

We conclude, that this model clearly is a better choice than the simple Poisson regression model.

Nevertheless, we are still far away from good fit. We now check, if we can reach further improvement by including interaction effects. To ease the model interpretation, we consider centred covariates.

One possibility to investigate the two-way interactions is to start with the full two-way interaction model and then delete the non-significant interaction terms. This approach yields the following model:

| | Estimate | Std. Error | z value | Pr(> z) |
|----------------------------|-----------------------------------|------------|---------|----------|
| (Intercept) | 3.93 | 0.03 | 124.14 | 0.00 |
| medium price | -0.05 | 0.02 | -2.29 | 0.02 |
| high price | -0.36 | 0.04 | -8.30 | 0.00 |
| spec_nl_centr | 0.19 | 0.06 | 2.86 | 0.00 |
| gen_nl_centr | 0.09 | 0.01 | 7.59 | 0.00 |
| avg_comp_centr | 0.17 | 0.32 | 0.53 | 0.60 |
| Prod_centr | 4.37 | 0.88 | 4.97 | 0.00 |
| Sub_centr | 0.31 | 0.65 | 0.48 | 0.63 |
| weekend | -0.36 | 0.01 | -28.09 | 0.00 |
| trend | 0.00 | 0.00 | 3.83 | 0.00 |
| med:Prod_centr | -5.05 | 0.87 | -5.80 | 0.00 |
| high:avg_comp_centr | 14.76 | 2.01 | 7.33 | 0.00 |
| spec_nl_centr:gen_nl_centr | -0.39 | 0.10 | -3.75 | 0.00 |
| Prod_centr:trend | -0.01 | 0.00 | -5.03 | 0.00 |
| Null deviance: | 4520.32 on 729 degrees of freedom | | | |
| Residual deviance: | 2583.51 on 716 degrees of freedom | | | |
| AIC: | 6805.061 | | | |

Table 3.11: Regression results for article A_6 in the Poisson regression model including time effects and interactions.

In this model, we have four interaction terms. We notice, that compared to the model without interactions, there are remarkable changes regarding the significance of the article-specific newsletter and the average price of the remaining product variants of product P_3 , which now are classified as significant, whereas the average price of the substitute product loses significance.

The residual deviance divided by the number of degrees of freedom for this model is about 3.61. Compared to the value from the Poisson model with time effects only, which was about 3.70, we do not notice a distinct improvement. Also the AIC has only decreased slightly. This indicates, that although we have included four additional covariates, our model still is not able to describe our data adequately.

One reason might be, that the interaction terms cannot explain much of the variation in the model, although they reach the significance level of 0.05.

We notice, that the interaction between the article-specific newsletter advertisement and the general advertisement is a very sparse covariate: The multiplication of these two covariates yields an even more sparse variable, which in the concrete case of article A_6 results in a covariate where only 3 out of 730 observations are non-zero. This is a share of about 0.4% of the time period of observation. Thus, this interaction term is equal to zero for more than 99% of the time period of observation, which makes it appear little convenient.

To find out, if this interaction model nevertheless is a better choice than the time effect Poisson model without interactions, we again perform a χ^2 -test:

| | Resid. Df | Resid. Dev | Df | Deviance | Rao | Pr(>Chi) |
|---|-----------|------------|------|----------|-------|----------|
| 1 | 720.00 | 2661.35 | | | | |
| 2 | 716.00 | 2583.51 | 4.00 | 77.84 | 78.95 | 0.00 |

Table 3.12: χ^2 -test for the Poisson model including time effects and the Poisson model with interactions.

As the p-value is clearly below the significance level of 0.05, we do reject the null hypothesis, that the Poisson model including time effects is good enough, and decide in favour of the Poisson interaction model.

Not being able to further improve the model, we now check the residuals and the link specification of our best fit. This helps to find out, if we should look for alternative modelling approaches.

Model validation

The model validation will be done based on the best model we have set up so far, i.e. on the model including weekend seasonality and interaction terms.

Investigating the residual plots, we consider the following graphs: A plot of the deviance residuals against time, a normal Q-Q plot, a plot of the fitted values versus the deviance residuals and the plots of the covariates against the deviance residuals. As the latter look quite similar for all of our covariates, we only present one of them. The full set of plots can be found in the appendix 7.4 and 7.5.

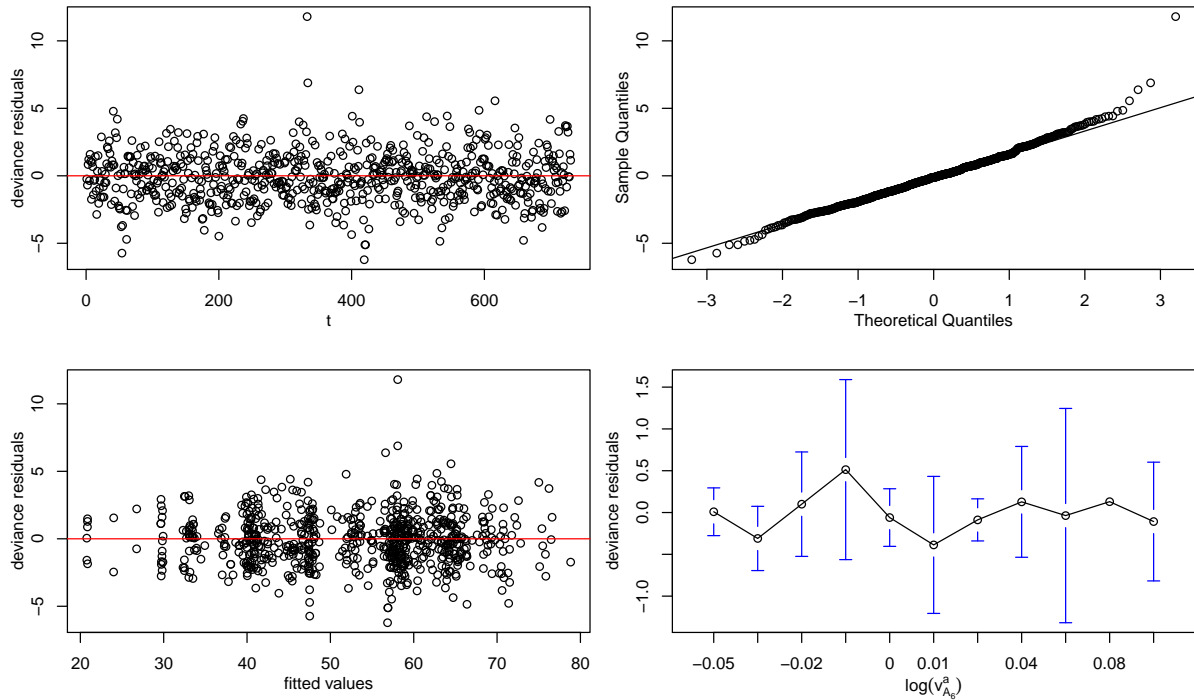


Figure 3.18: Residual analysis for the Poisson model including time effects and interactions: The left upper plot shows the deviance residuals for the 730 observations and the right upper plot the normal Q-Q plot of the deviance residuals. The lower plots show the expected response and exemplary one covariate against the deviance residuals.

The left upper plot just shows the deviance residuals for all 730 observations. As these are randomly scattered around zero, this plot does not indicate any lack of fit. The Q-Q plot of the deviance residuals helps on checking the desired normal distribution. When the residuals follow a normal distribution, they completely lie on the bisection line, which here is represented by a solid black line. Whereas for the residuals lying between -4 and 4, a normal distribution seems to be a good fit, we detect stronger deviations from the straight line for small and above all for large residuals. Thus, our distribution is slightly positively skewed and heavy tailed.

Furthermore, we should check, if the residuals are independent of the variables. We plot the estimated expectation of the response and the covariates against the deviance residuals, where no pattern, i.e. randomly scattering around zero, indicates a good fit. As all of our covariates have a rather low number of different levels, we use mean plots for the graphical representation. These plots show the group means and the corresponding confidence intervals. Concerning the response variable, the condition of independence from the residuals seems to be fulfilled, which is not the case for the covariates. This is exemplary shown in the plot for the average product price.

Let us now move on to the link function. When performing a Poisson regression model, typically the log-link is chosen. To check, if this fits well to the data, we plot the linear

predictors versus the response variable, where we expect randomly scattering around the exponential function. As this is rather hard to see from the plot, we consider the plot of the fitted means versus the response variable instead. This means, we do not use the linear predictor $\mathbf{x}_i^t \hat{\boldsymbol{\beta}}$, but we consider the exponential values $\exp(\mathbf{x}_i^t \hat{\boldsymbol{\beta}})$. Here, in case of a suitable link specification, we expect the values to lie on the bisection line.

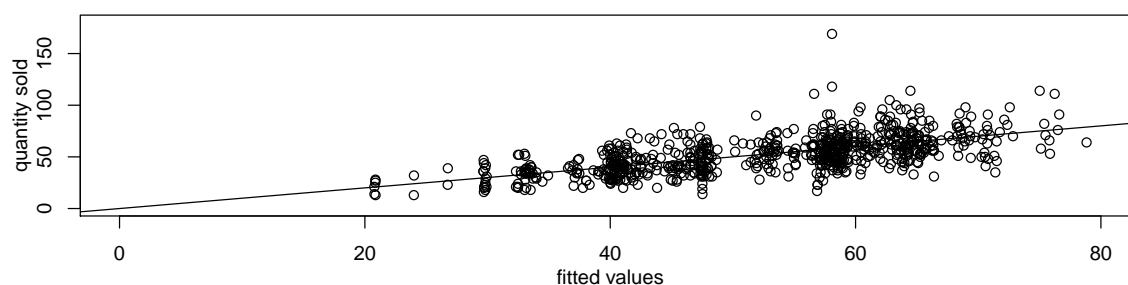


Figure 3.19: Check of the log-link specification: Plot of the linear predictor versus the response variable.

We conclude, that the log-link seems to be a good choice for our model. Nevertheless, we have seen, that the overall fit of the Poisson models is not sufficient.

An apparent reason for the bad model fit could be, that the strict model assumption of equal mean and variance is not met by the data. The way larger residual deviance than the number of degrees of freedom is a first hint at possible overdispersion. Of course, this must not be the case, but since we have already investigated on interaction effects and verified the link specification, overdispersion is not far to seek. Thereof, we conclude, that Negative Binomial regression models probably are more appropriate to describe our data.

3.2.3 Negative Binomial regression models for article A_6

To keep the Negative Binomial models comparable to the Poisson regression performed before, we stay with the same covariates. Here, too, we consider three different regression models: a simple Negative Binomial model, a Negative Binomial model covering time effects and one including interaction terms.

The regression results of the simple Negative Binomial regression model look as follows:

| | Estimate | Std. Error | z value | Pr(> z) |
|--------------------|----------------------------------|------------|---------|----------|
| (Intercept) | 2.53 | 0.73 | 3.48 | 0.00 |
| medium price | -0.09 | 0.04 | -2.06 | 0.04 |
| high price | -0.43 | 0.07 | -6.38 | 0.00 |
| spec_nl_A6 | 0.05 | 0.11 | 0.46 | 0.64 |
| general_nl | 0.18 | 0.03 | 6.68 | 0.00 |
| log(avg_comp_A6) | -0.30 | 0.61 | -0.49 | 0.62 |
| log(avg_prod_A6) | -0.94 | 0.51 | -1.83 | 0.07 |
| log(avg_subst_P3) | 2.60 | 0.83 | 3.12 | 0.00 |
| Null deviance: | 968.94 on 729 degrees of freedom | | | |
| Residual deviance: | 741.02 on 722 degrees of freedom | | | |
| AIC: | 6080.61 | | | |

Table 3.13: Regression results for article A_6 in the simple Negative Binomial regression model.

Concerning the regression coefficients, the Poisson model and the Negative Binomial model yield very similar results. Further, neglecting the level of significance for a moment, the two models more or less yield agreeing results regarding significance.

Concerning the deviance, we recognise strong improvements. It decreased from 3481.87 to 741.02, which gives a dispersion parameter very close to 1. This supposes, that we have a much better fit in the Negative Binomial model. At the same time, the standard errors double when using Negative Binomial regression. This does not mean that the estimates are less precise. The larger standard errors are appropriate, reflecting the fact, that there is more uncertainty than the Poisson model allows due to $E(\mathbf{Y}) = \text{Var}(\mathbf{Y})$.

We now check, if we can improve our model including weekend seasonality.

| | Estimate | Std. Error | z value | Pr(> z) |
|--------------------|----------------------------------|------------|---------|----------|
| (Intercept) | 3.22 | 0.90 | 3.58 | 0.00 |
| medium price | -0.07 | 0.04 | -1.76 | 0.08 |
| high price | -0.42 | 0.06 | -6.77 | 0.00 |
| spec_nl_A6 | 0.04 | 0.10 | 0.37 | 0.71 |
| general_nl | 0.10 | 0.02 | 4.11 | 0.00 |
| log(avg_comp_A6) | -0.28 | 0.56 | -0.50 | 0.62 |
| log(avg_prod_A6) | -0.51 | 0.52 | -0.99 | 0.32 |
| log(avg_subst_P3) | 1.49 | 1.20 | 1.25 | 0.21 |
| weekend | -0.35 | 0.02 | -15.35 | 0.00 |
| trend | 0.00 | 0.00 | 1.33 | 0.18 |
| Null deviance: | 1271.2 on 729 degrees of freedom | | | |
| Residual deviance: | 738.98 on 720 degrees of freedom | | | |
| AIC: | 5881.874 | | | |

Table 3.14: Regression results for article A_6 in the Negative Binomial regression model including time effects.

Again, the dummy variable covering the weekend seasonality is classified as significant. But, in contrast to the Poisson model, the effect of the trend variable is not supported by the Negative Binomial model. Apart from this, we get similar regression results like before.

Next, we investigate the interaction effects. This time, the model will be fitted using the stepwise AIC approach, which can be calculated automatically using 'R'. Therewith, we get the following model:

| | Estimate | Std. Error | z value | Pr(> z) |
|--------------------------|-----------------------------------|------------|---------|----------|
| (Intercept) | 3.94 | 0.03 | 112.83 | 0.00 |
| medium price | -0.10 | 0.04 | -2.65 | 0.01 |
| high price | -0.32 | 0.06 | -5.04 | 0.00 |
| spec_nl_cent | 0.24 | 0.13 | 1.86 | 0.06 |
| gen_nl_cent | 0.10 | 0.02 | 4.28 | 0.00 |
| avg_comp_cent | -0.41 | 0.54 | -0.76 | 0.45 |
| weekend | -0.36 | 0.02 | -15.71 | 0.00 |
| trend | 0.00 | 0.00 | 4.62 | 0.00 |
| high:avg_comp_cent | 7.31 | 2.25 | 3.24 | 0.00 |
| spec_nl_cent:gen_nl_cent | -0.47 | 0.20 | -2.37 | 0.02 |
| Null deviance: | 1292.84 on 729 degrees of freedom | | | |
| Residual deviance: | 737.39 on 720 degrees of freedom | | | |
| AIC: | 5867.834 | | | |

Table 3.15: Regression results for article A_6 in the Negative Binomial regression model including time effects and interactions.

This model suggests to include two interaction terms, namely the one between the high article price and the average competitor price, as well as the one between the two differ-

ent kinds of advertisement. Further, the main effect concerning the average price of the remaining articles belonging to product P_3 , as well as the one covering the average price for the substitute product do not appear any more in this model.

Due to this 'reduction' of the main effects, we are not able to compare this interaction model to the models we have fitted so far. We thus have limited the number of steps allowed in the stepwise AIC approach to a number, which only just avoids, that any main effect is deleted. This yields the following regression results:

| | Estimate | Std. Error | z value | Pr(> z) |
|----------------------------|-----------------------------------|------------|---------|----------|
| (Intercept) | 3.98 | 0.06 | 70.30 | 0.00 |
| medium price | -0.09 | 0.04 | -2.28 | 0.02 |
| high price | -0.31 | 0.07 | -4.23 | 0.00 |
| spec_nl_centr | 0.23 | 0.13 | 1.74 | 0.08 |
| gen_nl_centr | 0.10 | 0.02 | 4.30 | 0.00 |
| avg_comp_centr | -0.40 | 0.55 | -0.72 | 0.47 |
| Prod_centr | -0.24 | 0.51 | -0.46 | 0.64 |
| Sub_centr | 1.04 | 1.19 | 0.87 | 0.38 |
| weekend | -0.36 | 0.02 | -15.70 | 0.00 |
| trend | 0.00 | 0.00 | 1.52 | 0.13 |
| high:avg_comp_centr | 7.09 | 2.27 | 3.12 | 0.00 |
| spec_nl_centr:gen_nl_centr | -0.46 | 0.20 | -2.29 | 0.02 |
| Null deviance: | 1294.49 on 729 degrees of freedom | | | |
| Residual deviance: | 737.48 on 718 degrees of freedom | | | |
| AIC: | 5870.987 | | | |

Table 3.16: Regression results for article A_6 in the 'refitted' Negative Binomial regression model including time effects and interactions.

This model suggests the same interaction terms like the one performed above and the AIC only worsens slightly. Hence, the two interaction models are very close. Due to reasons of comparability we will refer to the just performed one when speaking of the Negative Binomial interaction model.

All in all, the Negative Binomial regression definitely is more appropriate to describe our data. So, the next step is, to find the best fitting one among the three of them performed above.

Regarding the residual deviance divided by the number of degrees of freedom, we get a value of about 1.03 for all the three models. The AIC is minimised in the model including weekend seasonality and interactions. This is a first indicator, that this model is our best fit. Also the χ^2 -test strongly suggests this model. The details on this test can be found in the appendix 7.3. Thus, again, we choose the model including weekend seasonality and interaction effects as our best fit.

To assess the goodness of fit of our 'best model', we use the residual deviance test. This test is based on the fact, that the deviance is given by the χ^2 -value at a certain degree of freedom df . In order to test for significance, we can find out the associated p-value, which calculates as $1 - \chi^2_{\text{deviance},df}$. We of course aim at accepting the null hypothesis that our model is 'good enough'. Thus, we need high p-values. In our example, we get a p-value of 0.30 and do not reject the null hypothesis. Thus, all in all, our Negative Binomial regression model including weekend seasonality and interaction terms seems to be an acceptable fit to our data.

Performing Negative Binomial regression models also for the remaining articles belonging to product P_3 , we get similar results. For more details on that, please refer to the model summaries, which are presented in table 7.1 and in table 7.2 in the appendix.

Checking the residuals

Poisson regression and Negative Binomial regression are closely related in the sense, that the Poisson regression is a special case of the Negative Binomial regression. Checking for the goodness of fit, we thus can proceed just as before, when investigating the residual plots.

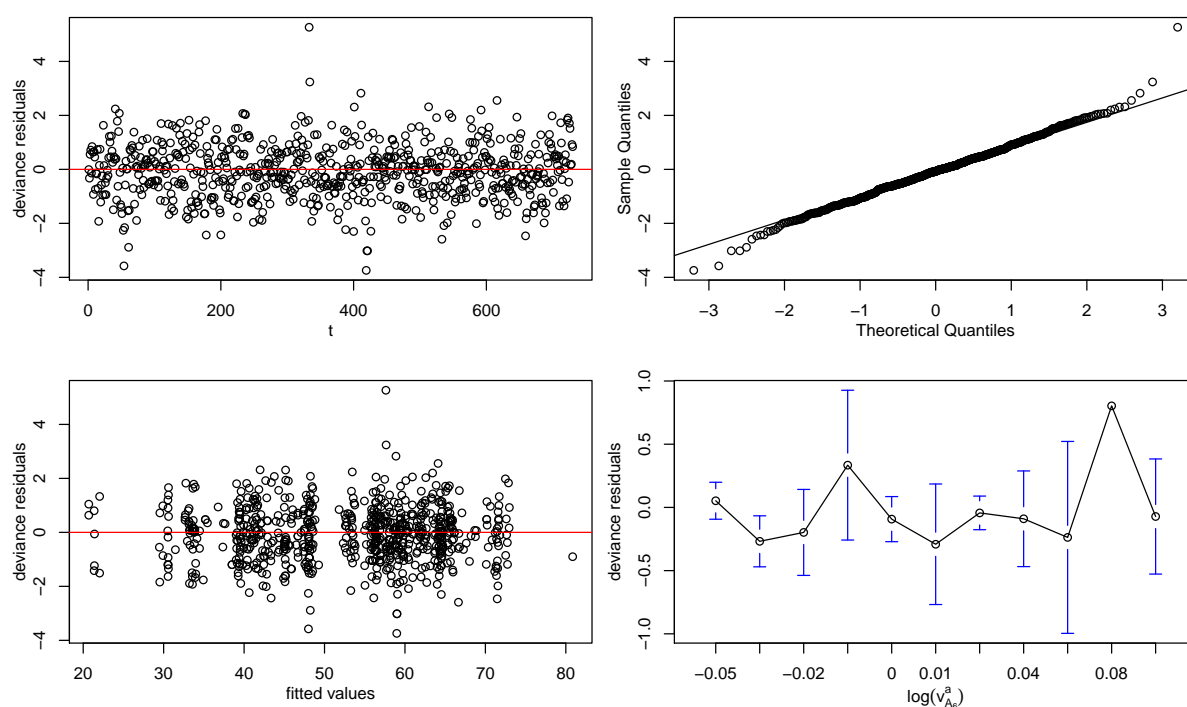


Figure 3.20: Residual analysis for Negative Binomial regression model including time effects and interactions: The left upper plot shows the deviance residuals of the 730 observations and the right upper plot the Q-Q plot of the deviance residuals. The lower plots show the expected response and exemplary one covariate against the deviance residuals. The full set of plots can be found in the appendix 7.6 and ??.

Again the residuals randomly scatter around zero. Furthermore, we notice, that the deviance is much smaller now. Whereas in case of Poisson regression, we get values lying between -5 and 10, we now are between -4 and 4.

All in all, we definitely improved the model fit using Negative Binomial regression models.

3.3 Count regression models for the daily amount sold per product

3.3.1 A Negative Binomial regression model for product P_3

Considering product P_3 , we have already built appropriate generalised linear models for the individual articles A_6 , A_7 and A_8 . For article A_9 we did not set up a separate model, because A_9 is a special offer, which was available only for a short period of time. This of course comes with raw data for the article.

We now pool all of the articles belonging to P_3 and compare the results. This way we check, how detailed we have to proceed to get an acceptable sales prediction.

Our best fit for the individual articles always was a Negative Binomial model. We thus assume the whole product to follow a Negative Binomial distribution, too.

Since we now have to deal with different packaging sizes, we can no longer stay with the quantity sold as response variable. Instead, we consider the units sold, i.e. we take $w_{A_m} \cdot b_{A_m} \cdot n_{A_m,t}$, where w_{A_m} is the weight of article A_m , b_{A_m} the bundle size and $n_{A_m,t}$ the quantity sold of article A_m at time point t . For each time point t we then sum up over all articles belonging to product P_3 . Since there are some packaging sizes of for example 4.5 kilogram, we furthermore multiply the units sold by 2. This ensures that we stay with an integer valued response variable, which is a necessary condition to be able to perform the desired Negative Binomial regression model.

Our response variable thus is defined as

$$N_{P_3,t} := 2 \cdot \sum_{m=6}^9 w_{A_m} \cdot b_{A_m} \cdot n_{A_m,t}.$$

To get an appropriate price for our product, we take the weighted mean over the basic prices of the individual articles. The basic price accordingly is measured per one half kilogram and thus calculates as

$$P_{P_3,t} := \frac{\sum_{m=6}^9 n_{A_m,t} \cdot p_{A_m,t} \cdot w_{A_m}}{2 \cdot \sum_{m=6}^9 n_{A_m} \cdot w_{A_m} \cdot b_{A_m}}.$$

Concerning the article-specific advertisement, for each time point t , we take the sum over the boolean flags of all the articles belonging to product P_3 . This gives us an integer value lying between 0 and 4 for the newsletter advertisement, which indicates, how many of the considered product variants were in advertisement at which time point. We define

$$\text{ad}_{P_3} := \sum_{m=6}^9 \text{ad}_{A_m}.$$

For the competitor prices, we of course take the average basic price over all product variants per one half kilogram. This results in

$$C_{P_3,t}^a := \frac{1}{2 \cdot \sum_{m=6}^9 1_{\{c_{A_m,t}^a > 0\}}} \sum_{m=6}^9 c_{A_m,t}^a.$$

The variable covering the influence of the other product variants of course has to be dropped, since these are already pooled and modelled as a whole now.

Concerning substitution effects, too, the average price of the substitute product P_{10} per one half kilogram is considered.

Exploratory data analysis

Since we want to build a Negative Binomial model with log-link, we at first check for the linear dependence between the logarithm of the response variable and the covariates.

To be able to interpret the regression coefficients concerning prices as elasticities, we need to measure them on logarithmic scale. This gives us the following plots:

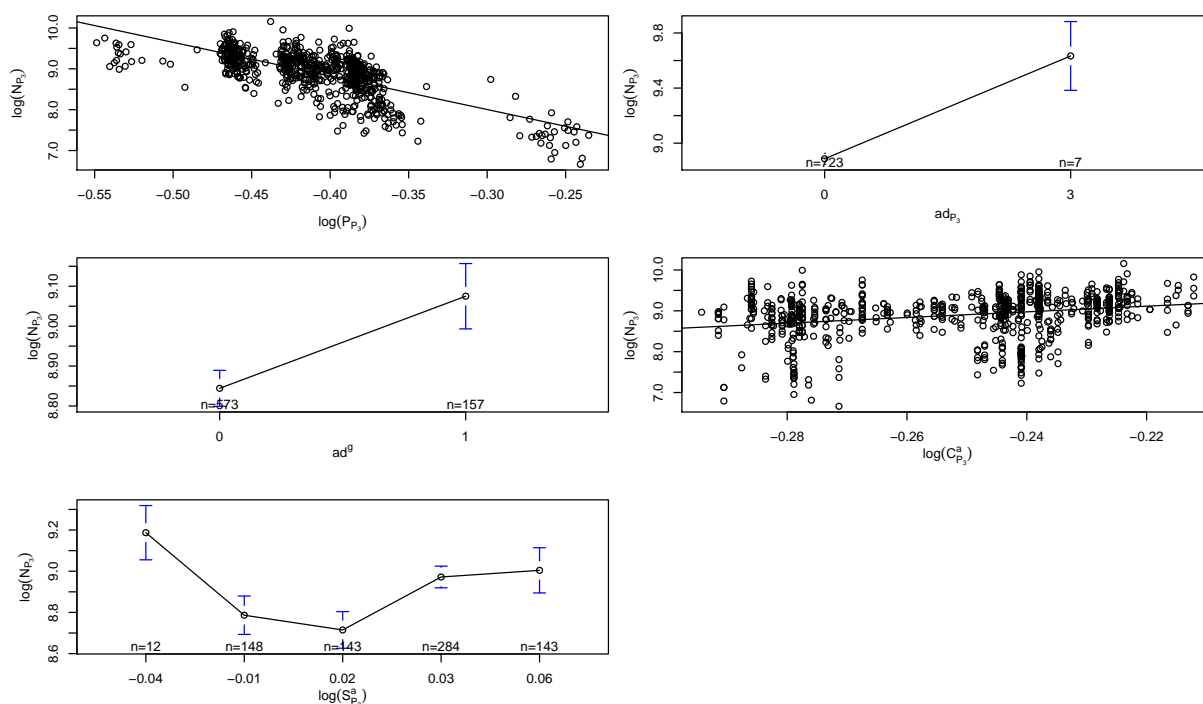


Figure 3.21: Checking for linearity: Plots of the logarithm of the response variable versus the covariates.

From the plots, we conclude, that the assumption of linearity is more or less fulfilled for most of the covariates.

Seasonal effects

As we have seen above, there is some seasonality present in the data for article A_6 . We thus expect a similar seasonal pattern for the whole product P_3 . We again check both, seasonal effects on monthly basis and weekday seasonality.

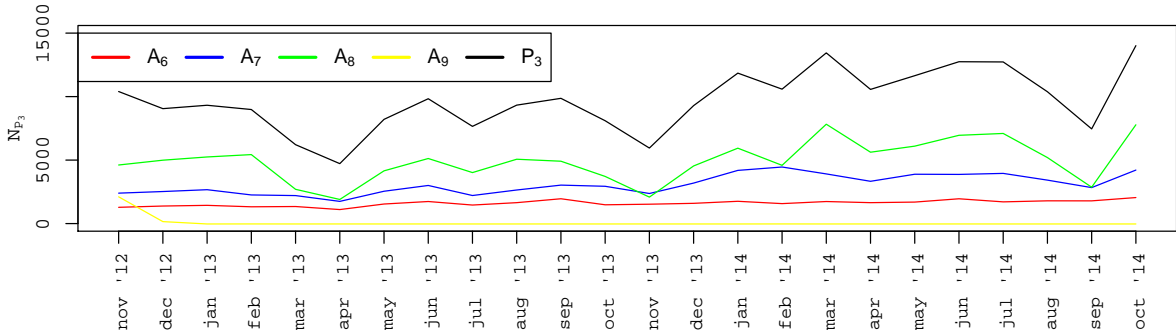


Figure 3.22: Seasonal effects on monthly basis for product P_3 . The curves for the individual articles belonging to P_3 are also shown.

The behaviour of the different articles belonging to product P_3 is very similar. We notice immediately, that apart from the beginning of the period of observation, the whole product behaves exactly like the article A_8 , which, due to the comparatively high number of sales, has a lot of influence on the product's behaviour. The deviations at the beginning result from the special offer A_9 , which was available in the shop at that time and influenced the sales significantly.

Due to the lack of a clear pattern, we do not take into account this type of seasonality. But, we notice, that the sales seem to grow over time, which of course will be regarded when modelling.

Concerning weekday seasonality we get very similar results for all of the articles belonging to product P_3 . Thus, at least in connection with this product, this seems to be the typical customer behaviour.

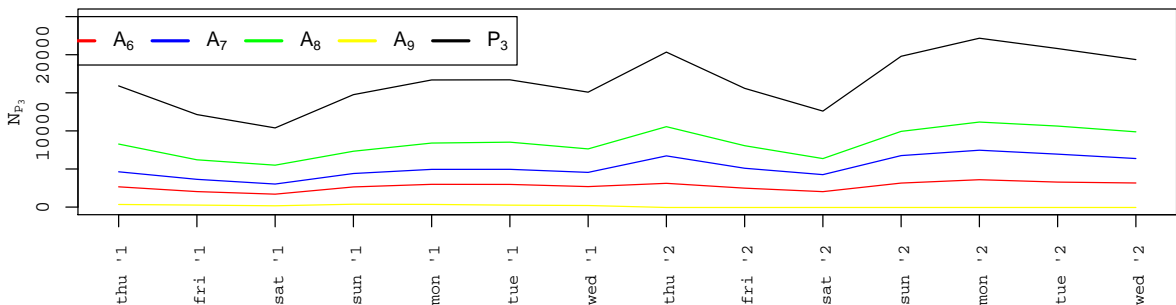


Figure 3.23: Checking for weekday seasonality of product P_3 . The curves for the individual articles, which belong to product P_3 , are also shown.

Model selection

Of course, for the product P_3 , too, we could perform different Poisson regression models and Negative Binomial regression models and then choose a model based on appropriate statistical tests or coefficients. But, we have already seen in the exploratory data analysis, that, concerning time effects, the whole product behaves exactly like the individual articles and should further also follow a Negative Binomial distribution.

Above, we always decided in favour of the Negative Binomial regression model including time effects and interaction terms. So this is very likely to be an appropriate model for the product as a whole, too. But, since the Poisson model and the Negative Binomial model did not agree on which interaction terms to include into the model and we further could only improve the model fit very slightly when using interactions, for the product as a whole, we only perform a Negative Binomial model including time effects, but without interaction terms.

This model yields the following regression results:

| | Estimate | Std. Error | z value | Pr(> z) |
|--------------------|-----------------------------------|------------|---------|----------|
| (Intercept) | 6.61 | 0.30 | 22.01 | 0.00 |
| log(price_P3) | -8.01 | 0.29 | -27.76 | 0.00 |
| spec_nl_P3 | 0.03 | 0.04 | 0.79 | 0.43 |
| general_nl | 0.13 | 0.03 | 4.36 | 0.00 |
| log(avg_comp_P3) | 2.93 | 0.95 | 3.10 | 0.00 |
| log(avg_subst_P3) | 1.90 | 1.40 | 1.36 | 0.18 |
| weekend | -0.33 | 0.03 | -11.70 | 0.00 |
| trend | -0.00 | 0.00 | -2.82 | 0.00 |
| Null deviance: | 1651.09 on 729 degrees of freedom | | | |
| Residual deviance: | 743.43 on 722 degrees of freedom | | | |
| AIC: | 13487.22 | | | |

Table 3.17: Regression results for product P_3 in the Negative Binomial regression model including time effects.

Concerning the scale, the sign and the significance level of the regression coefficients, we get very similar results like in case of the Negative Binomial regression models for the individual articles, and the residual deviance divided by the degree of freedom again yields a value of 1.03. We suppose, that this model is an appropriate choice for modelling the product as a whole, too. But, of course, we have to investigate the residuals to get a more detailed insight into the goodness of fit of this model.

Residual analysis

Like for the other generalised linear models, we now perform a residual analysis to check the goodness of fit.

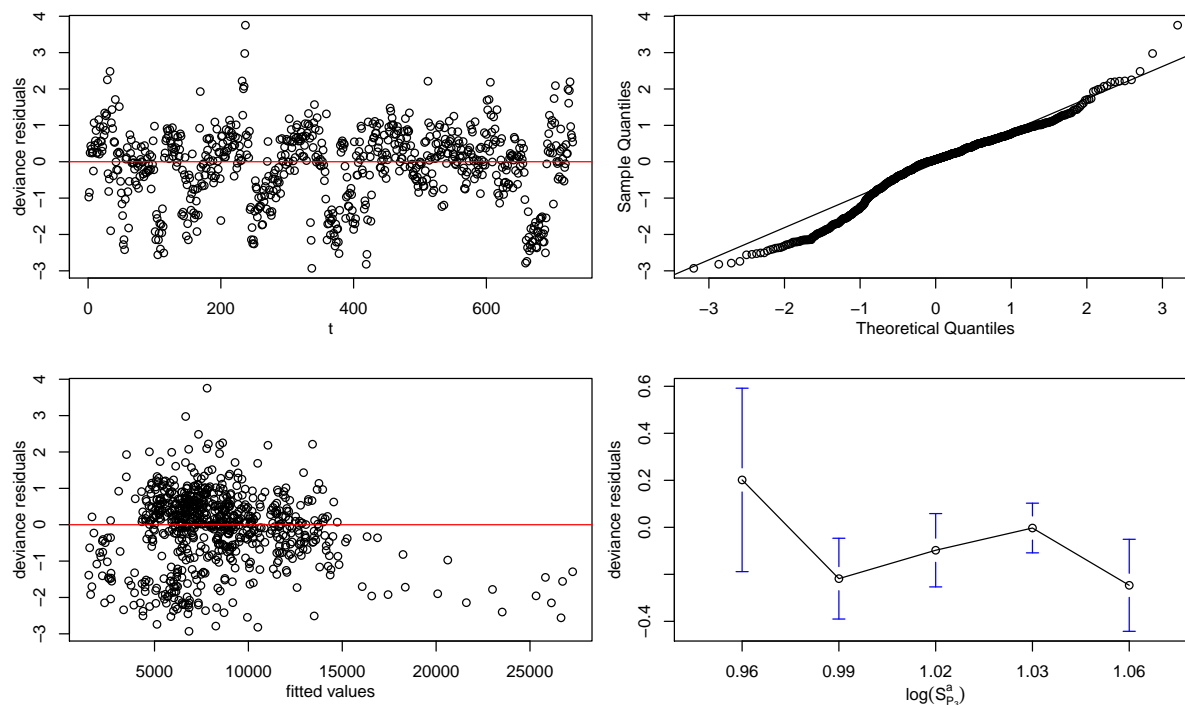


Figure 3.24: Residual analysis for the Negative Binomial regression model including time effects for product P_3 : The left upper plot shows the deviance residuals for the 730 observations and the right upper plot the normal Q-Q plot of the deviance residuals. The lower plots show the expected response and exemplary one of the covariates against the deviance residuals. The full set of plots can be found in the appendix 7.7.

This model obviously does not really fit to our data. We suppose, that this is due to the non-linear relationship between the response variable and some of the covariates. Using variable transformations, we were not able to compensate these non-linearities. Thus, in the next chapter, we try to improve the model fit using generalised additive models.

Model interpretation

Up to now, we have set up different count models on article basis and for the product as a whole. We now want to check, if the regression coefficients are economically interpretable. On that point, we will not only concentrate on one model at a time, but we want to compare the model outcomes for the different articles belonging to product P_3 , and the model for the product as a whole.

In the following, we will always refer to the Negative Binomial models including time effects, since these provide a quite good fit across all the articles and for the product as a whole.

Covering the **shop price** in a factor variable for the individual articles, we unfortunately do not get the price elasticity out of the corresponding models. Nevertheless, we read out from the regression coefficients, that the customers react on price changes in a reasonable manner: Our reference level for the shop price always was the lowest price available. The regression coefficients of the remaining prices, which are higher than the reference price, all have a negative sign. This means, that in comparison to the lowest price, there were less articles sold at all of the higher prices. Furthermore, the regression coefficient for the most part becomes smaller with growing prices, i.e. the effect strengthens with growing prices. This exactly is what we have expected: The quantity sold decreases with increasing prices.

For the product as a whole, the shop price was brought into the model on logarithmic scale. This exactly yields the price elasticity. Having a regression coefficient of -8.01 hence tells us, that sales would decrease by about 8%, if the product price was increased by 1%. Investigating the product P_3 in more detail, this is quite convincing: The product appeared in the newsletter advertisement for several times and there is a special offer article among its product variants. Thus, the customer is used to special offers and reduced prices for this product, which possibly makes him 'price sensitive' in the sense, that some get into the habit of panic buying and store up with this product to overcome the time periods with less attractive prices.

Concerning **advertisement**, we get positive signs for all of the regression coefficients. This approves the uplift of sales generated by newsletter advertisement.

Regarding article A_6 , we find, that the quantity sold increases by about 4% if the article appears in the newsletter. The general advertisement however seems to have more than twice this effect on that article. Concerning article A_7 , we draw similar conclusions. But, for article A_8 we observe a clearly larger regression coefficient for the article-specific newsletters than for general advertisement.

Comparing these regression coefficients to the box plots in figure 3.9, 3.10 and 3.11, we can verify our regression results: All of the regression coefficients attribute a relative uplift of sales lying between 4% and 42% to article-specific newsletter advertisement and of about 10% to 14% to general advertisement. Regarding the box plots in figure 3.8, we conclude, that these seem to be typical values.

Also regarding the product P_3 as a whole, the regression results are very convincing, which can be verified by summarising the effect of advertisement across the individual articles belonging to product P_3 and taking account of the respective numbers of sales.

Concerning the **average competitor prices**, we do not find agreeing results concerning the sign of the regression coefficient. For article A_7 and A_8 , as well as for the product P_3 as a whole, we get a positive regression coefficient, which tells us, that sales in general increase with increasing competitor prices. This is absolutely what we have expected. Nevertheless, things are different for article A_6 , where we get a negative regression coefficient.

Let us investigate that in more detail:

For the whole product P_3 , we get a regression coefficient of 2.93, which, due to the logarithmised covariate, can be interpreted in the way, that sales increase by 2.93% if the average competitor price raises by 1%. This is absolutely convincing: When the average competitor price raises, some of the customers will search for cheaper offers and thus probably prefer our multipacks or special offers rather than their 'usual' article at the competitor's online shop.

This also gives a good explanation for the negative sign of the regression coefficient for article A_6 : This is the smallest packaging size of this product, which is available at our shop and which thus is the most expensive one referring to the basic prices. If customers visit our online shop because of too expensive prices at the competitors, they will rather take the large packages with the lower basic prices because it is very likely, that the basic price of the small package is still higher than the price for the large package at the competitors. So customers are rather unlikely to change the online shop, if, neglecting packaging size for a moment, the product is available at a lower basic price at their habitual online shop. This dependence between the competitor prices and the packaging size become even more convincing, if we compare the regression coefficients of article A_6 , A_7 and A_8 : Whereas we make out an average decrease in sales of about -0.28% per 1% uplift in the competitor prices concerning article A_6 , we find an uplift of about 1.19% when considering article A_7 , and of even 4.40% concerning article A_8 .

The regression coefficients of the other **product variants** show similar characteristics like the ones concerning the competitor prices: Price adjustments usually strike the whole product and not only individual product variants. Thus, if the prices for one of the articles raise, so will do the prices of the remaining articles of this product, too. Hence, having again a negative regression coefficient for article A_6 , but positive ones for article A_7 and A_8 , we conclude, that customers are more likely to buy larger packaging sizes with growing product prices. This again can easily be explained by the fact, that the basic prices decrease with growing packaging size.

Although sales increase for the medium and large packaging size with growing product price, we suppose a decrease in sales when considering the product P_3 as a whole. Since all articles belonging to product P_3 are clustered together in the respective product model, we of course do not have a covariate covering the influence of 'the other product variants'. But, we have the price sensitivity, which, due to the negative sign tells us, that sales decrease with growing product prices.

Further, if we do not have a general price adjustment, but just set the prices from reduced price to 'normal price', the customers probably have stored up with this product during the period of price reduction. This makes buying the product at normal price temporarily quite unattractive.

We conclude, that price appreciations on the one hand make customers switch to larger packaging sizes, but bring about a decrease in sales regarding the product as a whole.

Concerning the **substitute product** P_{10} , we get positive regression coefficients for the article A_6 and the product P_3 as a whole, whereas the models for article A_7 and A_8 suggest a negative sign for the substitution effect. This could at best be interpreted in the sense,

that customers temporarily switch from product P_{10} to small packages of product P_3 in case of price appreciation for product P_{10} . But, we cannot explain why the number of sales should decrease for the larger packaging sizes at the same time. Hence, we suppose these effects to be random. This assumption is confirmed by the fact, that substitution does not reach the significance level of 0.05 in any of the models.

3.3.2 Generalised additive models (GAM) on product basis

For each of the eleven products, we now fit an appropriate Negative Binomial generalised additive model including time effects.

Using this model type, we are able to cover non-linear dependencies of the response variable on the covariates, which cannot be resolved using a simple transformation. For this purpose, we fit a spline to the concerned covariates. Since it is rather difficult to identify these covariates before setting up the model, our first approach for each of the products will be to fit a spline to each of the covariates concerning prices as well as to the one covering the overall trend.

Having set up these models, we check the splines that were fitted to the different covariates. If these look linear, we conclude, that it is not worth fitting a spline, because this would increase the model complexity without improving the fit. The concerned covariates then will be included as simple main effect into the model.

The covariates for the weekend seasonality and for the different kinds of advertisement will always come into the model as simple main effects, since it is not reasonable to fit a spline to dummy variables with two possible values only.

For each of the products contained in our data set, we present the set of plots resulting from the first modelling approach. The 95% confidence intervals are also drawn. Further, the model summary of the final GAM is displayed.

Product P_1

Product P_1 consists of the three articles A_1 , A_2 and A_3 . For these articles, there is no article-specific newsletter advertisement, but we have competitor prices available as well as the prices of six possible substitute products of the same product type. This results in a model including the article price, the general advertisement, the influence of the competitor prices, the influence of the substitute products and the time effects.

To check for non-linear dependencies, we investigate the spline plots of the different covariates.

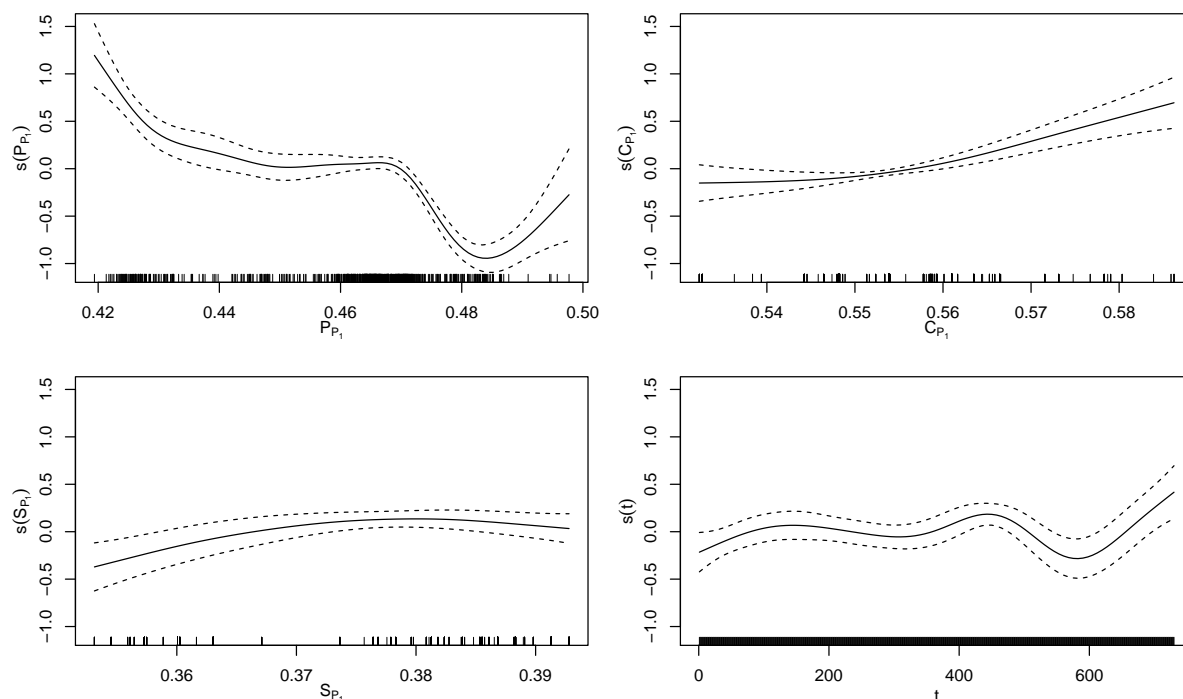


Figure 3.25: Plots of the splines which were fitted to the covariates of the Negative Binomial GAM for product P_1 . Confidence intervals are also drawn.

| A. parametric coefficients | Estimate | Std. Error | t-value | p-value |
|----------------------------|----------|------------|----------|----------|
| (Intercept) | 8.5048 | 0.0217 | 391.7097 | < 0.0001 |
| general_nl | 0.1029 | 0.0397 | 2.5905 | 0.0096 |
| weekend | -0.2912 | 0.0367 | -7.9308 | < 0.0001 |
| B. smooth terms | edf | Ref.df | F-value | p-value |
| s(price_P1) | 7.6619 | 8.5510 | 296.2195 | < 0.0001 |
| s(avg_comp_P1) | 2.3069 | 2.8937 | 29.7761 | < 0.0001 |
| s(avg_subst_P1) | 2.4123 | 2.9746 | 15.1414 | 0.0011 |
| s(trend) | 7.0344 | 8.0473 | 57.9398 | < 0.0001 |

Table 3.18: Regression results of the final Negative Binomial GAM for product P_1 .

Product P_2

Product P_2 consists of the two articles A_4 and A_5 . Here, too, we do not have article-specific newsletter advertisement but six substitute products. Furthermore, we have some competitor prices available.

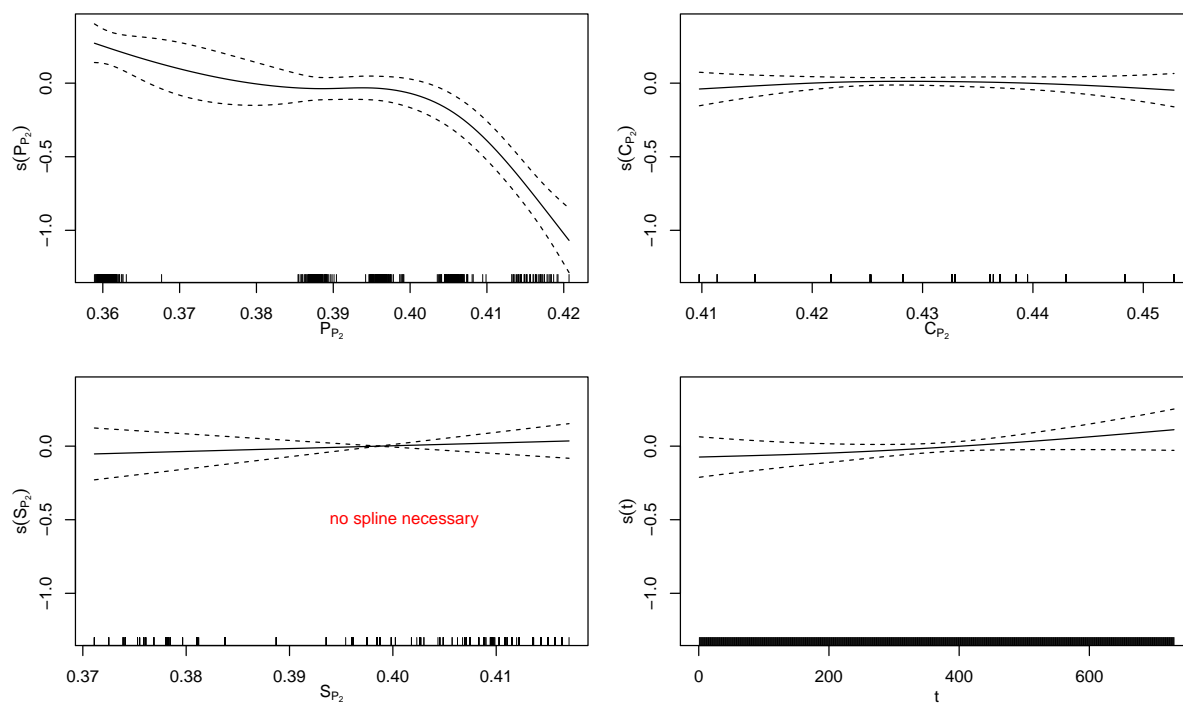


Figure 3.26: Plots of the splines which were fitted to the covariates of the Negative Binomial GAM for product P_2 . Confidence intervals are also drawn.

The plots indicate, that there is no need for a spline concerning the prices of the substitute products. Thus, we include this covariate as simple main effect.

| A. parametric coefficients | Estimate | Std. Error | t-value | p-value |
|----------------------------|----------|------------|----------|----------|
| (Intercept) | 8.4248 | 1.2745 | 6.6101 | < 0.0001 |
| general_nl | 0.1245 | 0.0351 | 3.5468 | 0.0004 |
| avg_subst_P2 | 1.9194 | 3.1974 | 0.6003 | 0.5483 |
| weekend | -0.3498 | 0.0317 | -11.0210 | < 0.0001 |
| B. smooth terms | edf | Ref.df | F-value | p-value |
| s(price_P2) | 3.6784 | 4.3917 | 130.6987 | < 0.0001 |
| s(avg_comp_P2) | 1.7366 | 2.1428 | 1.4504 | 0.5093 |
| s(trend) | 1.3115 | 1.5270 | 2.9612 | 0.2305 |

Table 3.19: Regression results of the final Negative Binomial GAM for product P_2 .

Product P_3

Concerning product P_3 , we have already seen a lot of details on the individual articles A_6 to A_9 , as well as on the product itself. For this product, too, we now fit a generalised additive model and see if we can improve the fit compared to the generalised linear model from above.

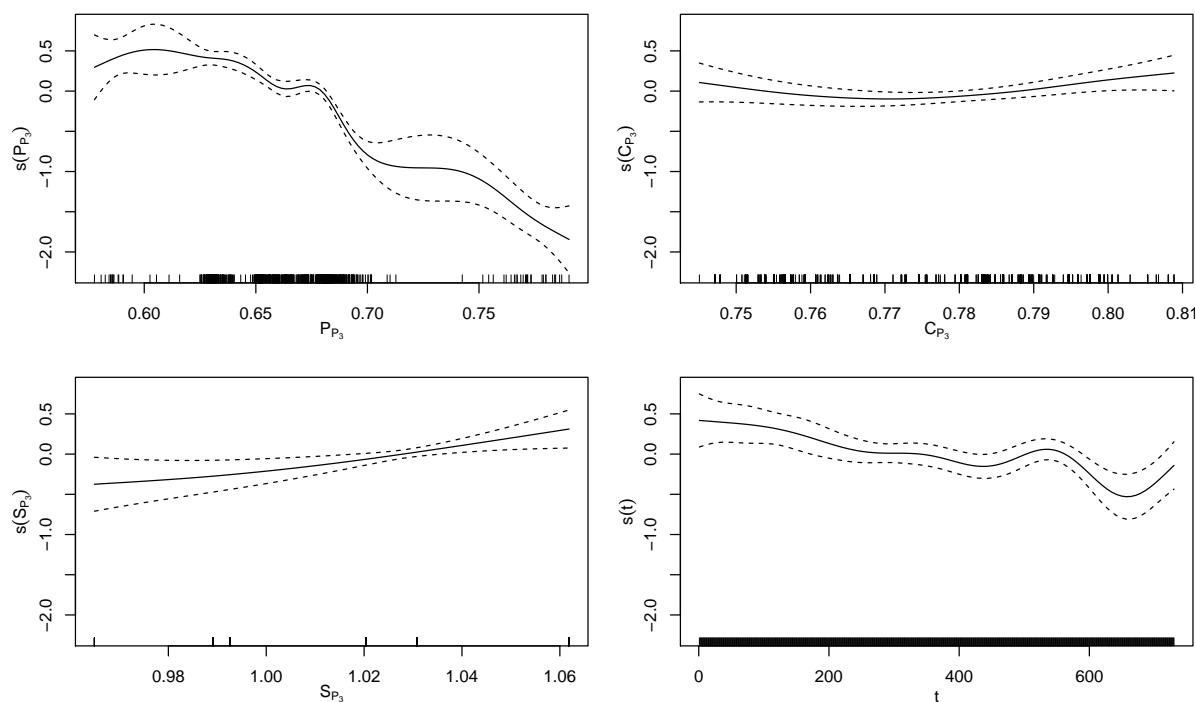


Figure 3.27: Plots of the splines which were fitted to the covariates of the Negative Binomial GAM for product P_3 . Confidence intervals are also drawn.

| A. parametric coefficients | Estimate | Std. Error | t-value | p-value |
|----------------------------|----------|------------|----------|----------|
| (Intercept) | 8.9958 | 0.0226 | 398.7704 | < 0.0001 |
| spec_nl_P3 | 0.0954 | 0.0578 | 1.6503 | 0.0989 |
| general_nl | 0.1171 | 0.0415 | 2.8171 | 0.0048 |
| weekend | -0.3223 | 0.0374 | -8.6257 | < 0.0001 |
| B. smooth terms | edf | Ref.df | F-value | p-value |
| s(price_P3) | 8.3488 | 8.8683 | 358.7439 | < 0.0001 |
| s(avg_comp_P3) | 2.5892 | 3.2840 | 9.5921 | 0.0252 |
| s(avg_subst_P3) | 1.6992 | 2.0688 | 8.2220 | 0.0178 |
| s(trend) | 8.3524 | 8.8362 | 26.8904 | 0.0011 |

Table 3.20: Regression results of the final Negative Binomial GAM for product P_3 .

Product P_4

Belonging to product type 'a', product P_4 has six possible substitute products. The two articles A_{10} and A_{11} , of which product P_4 consists, do not appear in the article-specific newsletter advertisement, but we have some competitor prices available.

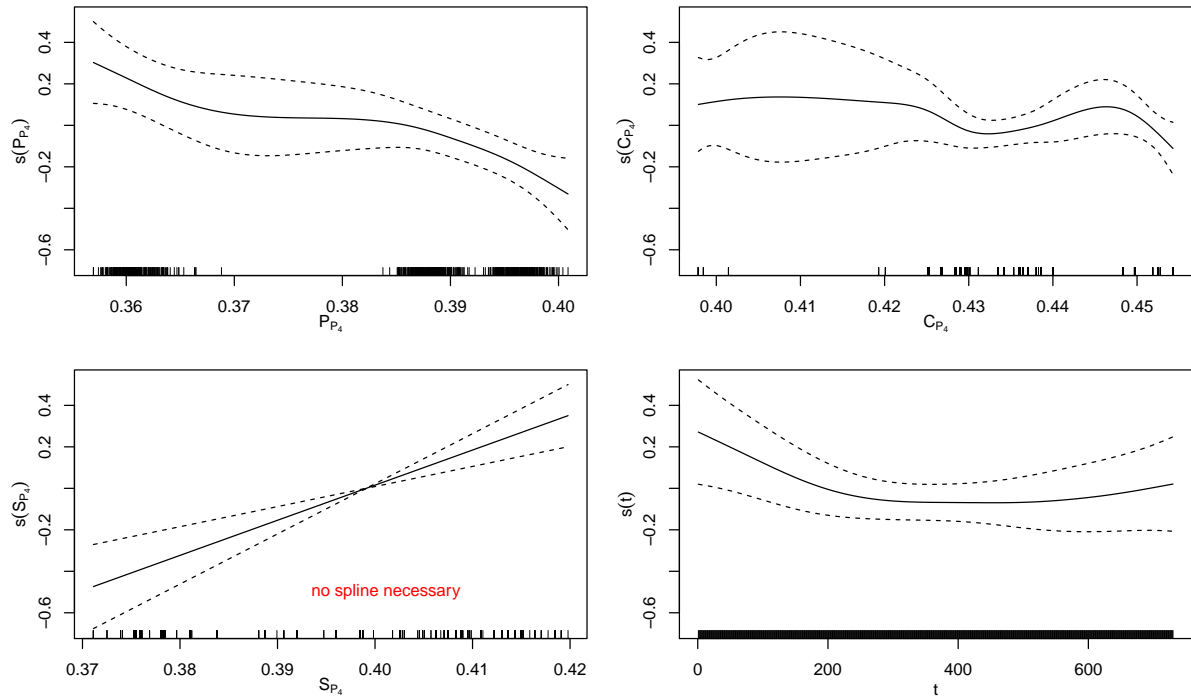


Figure 3.28: Plots of the splines which were fitted to the covariates of the Negative Binomial GAM for product P_4 . Confidence intervals are also drawn.

The spline for the price of the substitute products is not necessary. Thus, we refit the model with this covariate as simple main effect.

| A. parametric coefficients | Estimate | Std. Error | t-value | p-value |
|----------------------------|----------|------------|---------|----------|
| (Intercept) | 0.9075 | 1.4465 | 0.6274 | 0.5304 |
| general_nl | 0.1156 | 0.0405 | 2.8537 | 0.0043 |
| avg_subst_P4 | 16.9280 | 3.6224 | 4.6732 | < 0.0001 |
| weekend | -0.3473 | 0.0366 | -9.4806 | < 0.0001 |
| B. smooth terms | edf | Ref.df | F-value | p-value |
| s(price_P4) | 3.0215 | 3.7172 | 21.3864 | 0.0003 |
| s(avg_comp_P4) | 4.9076 | 5.9010 | 7.3762 | 0.2768 |
| s(trend) | 2.6185 | 3.2480 | 9.6663 | 0.0582 |

Table 3.21: Regression results of the final Negative Binomial GAM for product P_4 .

Product P_5

Product P_5 also belongs to product type 'a' and consequently has six possible substitute products. It consists of article A_{12} and A_{13} , which do not appear in the article-specific advertisement but for which there are some competitor prices available.

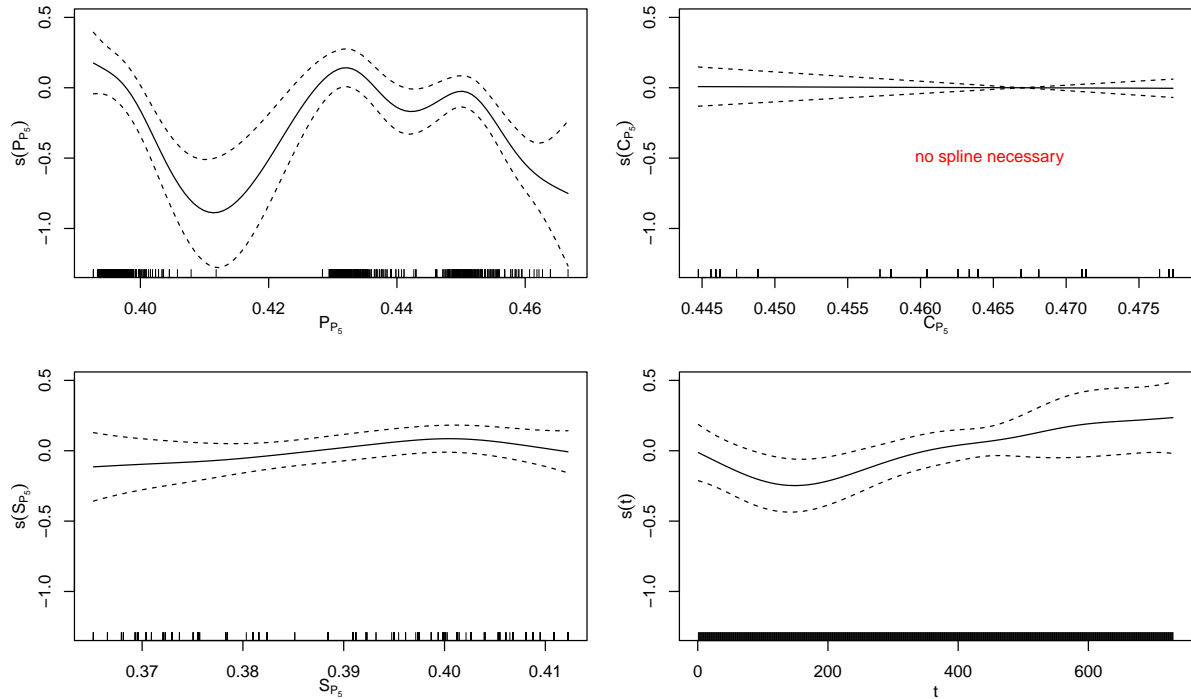


Figure 3.29: Plots of the splines which were fitted to the covariates of the Negative Binomial GAM for product P_5 . Confidence intervals are also drawn.

We drop the spline for the competitor price.

| A. parametric coefficients | Estimate | Std. Error | t-value | p-value |
|----------------------------|----------|------------|---------|----------|
| (Intercept) | 7.0015 | 1.4654 | 4.7780 | < 0.0001 |
| general_nl | 0.1196 | 0.0385 | 3.1053 | 0.0019 |
| avg_comp_P5 | -0.3666 | 3.1387 | -0.1168 | 0.9070 |
| weekend | -0.3374 | 0.0350 | -9.6295 | < 0.0001 |
| B. smooth terms | edf | Ref.df | F-value | p-value |
| s(price_P5) | 7.8792 | 8.6590 | 90.9503 | < 0.0001 |
| s(avg_subst_P5) | 2.8276 | 3.5251 | 4.9763 | 0.3151 |
| s(trend) | 4.3870 | 5.3989 | 12.0368 | 0.0473 |

Table 3.22: Regression results of the final Negative Binomial GAM for product P_5 .

Product P_6

Product P_6 is the only product of product type 'c' and consists of the three articles A_{14} , A_{15} and A_{16} . There are some competitor prices available and two out of the three articles appeared in the article-specific newsletter advertisement.

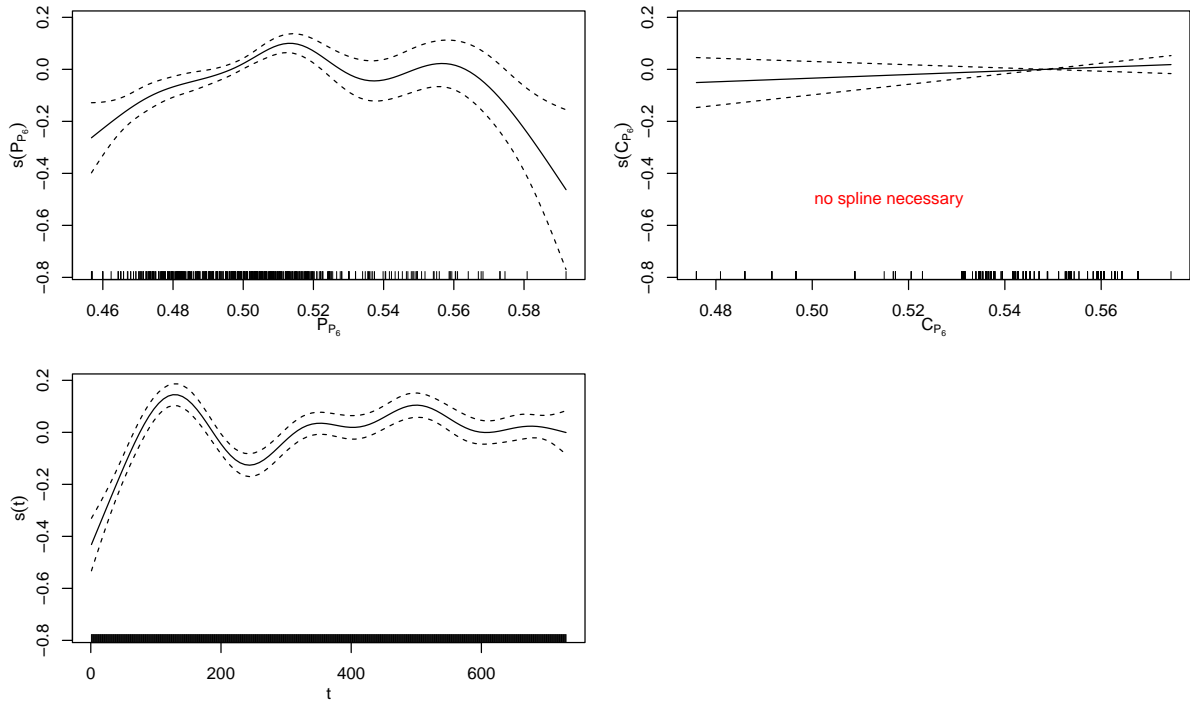


Figure 3.30: Plots of the splines which were fitted to the covariates of the Negative Binomial GAM for product P_6 . Confidence intervals are also drawn.

We drop the spline for the average competitor price.

| A. parametric coefficients | Estimate | Std. Error | t-value | p-value |
|----------------------------|----------|------------|----------|----------|
| (Intercept) | 6.7041 | 0.3636 | 18.4402 | < 0.0001 |
| spec_nl_P6 | 0.1054 | 0.0789 | 1.3367 | 0.1813 |
| general_nl | -0.0066 | 0.0178 | -0.3722 | 0.7097 |
| avg_comp_P6 | 0.7005 | 0.6625 | 1.0574 | 0.2903 |
| weekend | -0.0333 | 0.0167 | -1.9910 | 0.0465 |
| B. smooth terms | edf | Ref.df | F-value | p-value |
| s(price_P6) | 7.0778 | 8.1617 | 77.6089 | < 0.0001 |
| s(trend) | 8.5731 | 8.9465 | 157.4948 | < 0.0001 |

Table 3.23: Regression results of the final Negative Binomial GAM for product P_6 .

Product P_7

Product P_7 consist of the two articles A_{17} and A_{18} and is of product type 'a'. It has six possible substitute products and does not appear in the article-specific newsletter advertisement. There are no competitor prices available, because product P_7 is a private label product.

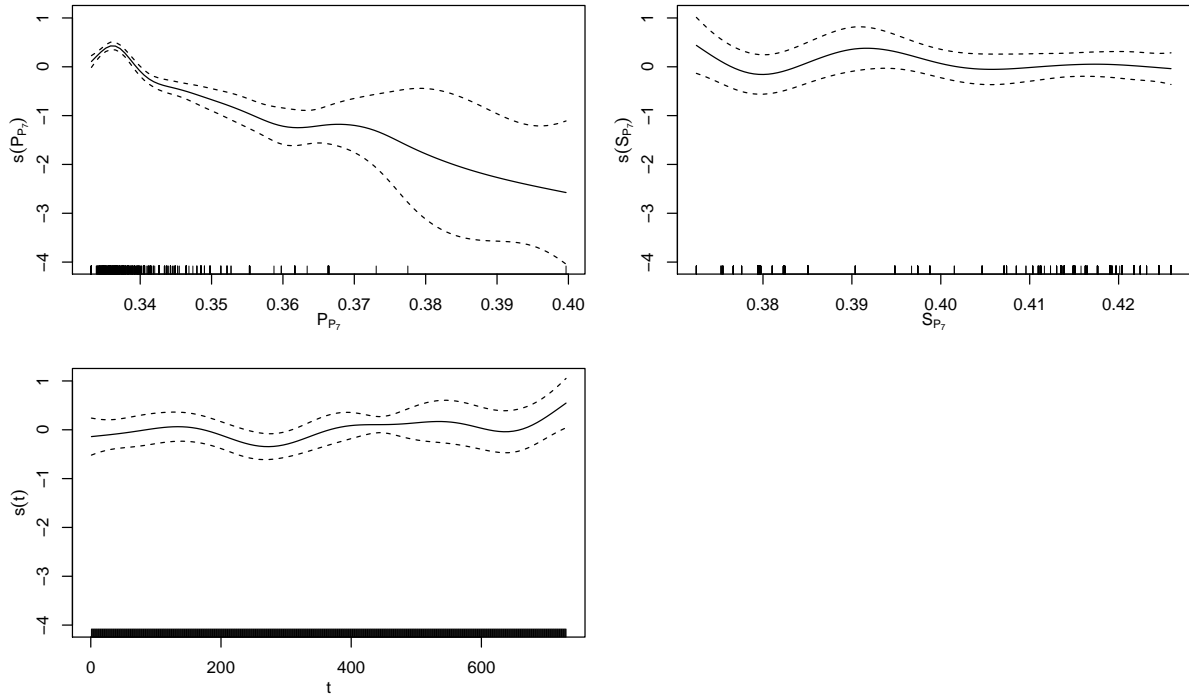


Figure 3.31: Plots of the splines which were fitted to the covariates of the Negative Binomial GAM for product P_7 . Confidence intervals are also drawn.

| A. parametric coefficients | Estimate | Std. Error | t-value | p-value |
|----------------------------|----------|------------|----------|----------|
| (Intercept) | 6.7760 | 0.0369 | 183.8541 | < 0.0001 |
| general_nl | 0.0741 | 0.0685 | 1.0823 | 0.2791 |
| weekend | -0.2746 | 0.0613 | -4.4825 | < 0.0001 |
| B. smooth terms | edf | Ref.df | F-value | p-value |
| s(price_P7) | 7.7170 | 8.5568 | 247.5148 | < 0.0001 |
| s(avg_subst_P7) | 6.3688 | 7.4823 | 13.0744 | 0.0759 |
| s(trend) | 6.9784 | 7.9955 | 20.5493 | 0.0083 |

Table 3.24: Regression results of the final Negative Binomial GAM for product P_7 .

Product P_8

Product P_8 consists of the three articles A_{19} , A_{20} and A_{21} . It is of product type 'a' and has six possible substitute products. Since it is a private label product, there are no competitor prices available. Further, the product did not appear in the article-specific advertisement.

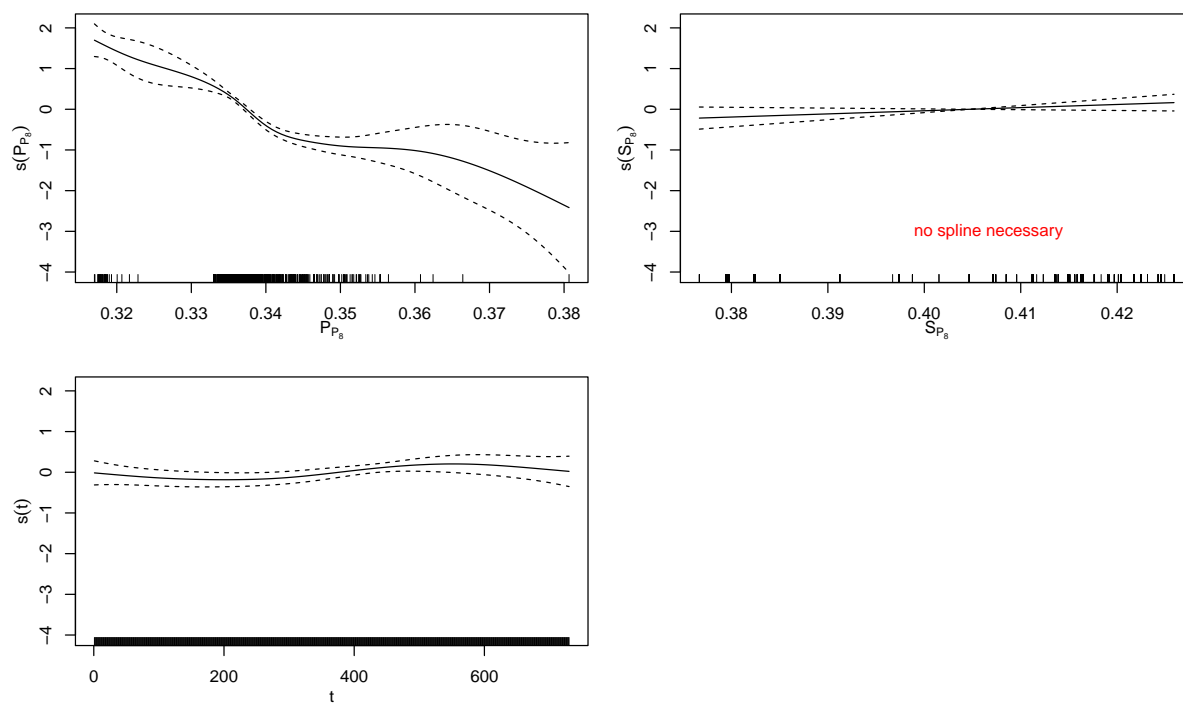


Figure 3.32: Plots of the splines which were fitted to the covariates of the Negative Binomial GAM for product P_8 . Confidence intervals are also drawn.

We do not need the spline for the price of the substitute products.

| A. parametric coefficients | Estimate | Std. Error | t-value | p-value |
|----------------------------|----------|------------|----------|----------|
| (Intercept) | 4.6825 | 1.9594 | 2.3898 | 0.0169 |
| general_nl | 0.1072 | 0.0793 | 1.3519 | 0.1764 |
| avg_subst_P8 | 7.7213 | 4.8380 | 1.5960 | 0.1105 |
| weekend | -0.0841 | 0.0741 | -1.1346 | 0.2565 |
| B. smooth terms | edf | Ref.df | F-value | p-value |
| s(price_P8) | 5.5814 | 6.6781 | 295.5534 | < 0.0001 |
| s(trend) | 3.4164 | 4.2197 | 8.5145 | 0.0893 |

Table 3.25: Regression results of the final Negative Binomial GAM for product P_8 .

Product P_9

Product P_9 , too, is of product type 'a' and thus there are six possible substitute products. It consists of the three articles A_{22} , A_{23} and A_{24} , where two out of them did appear in the article-specific newsletter advertisement. Since P_9 is a private label product, there are no competitor prices available.

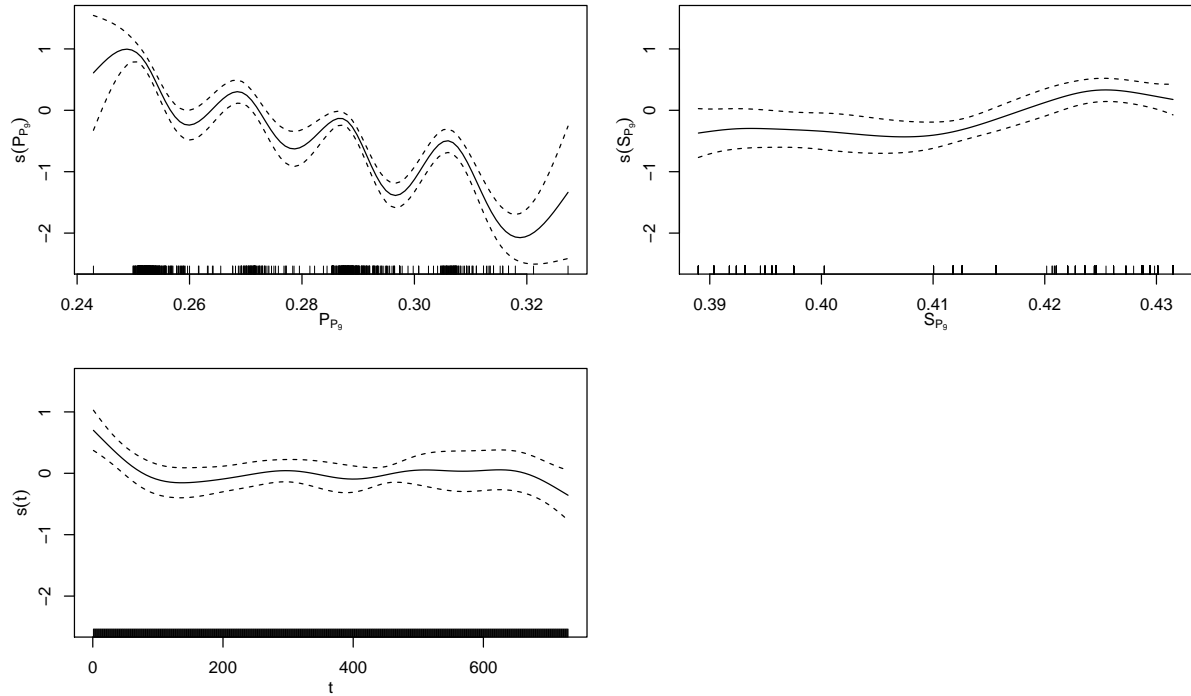


Figure 3.33: Plots of the splines which were fitted to the covariates of the Negative Binomial GAM for article A_9 . Confidence intervals are also drawn.

| A. parametric coefficients | Estimate | Std. Error | t-value | p-value |
|----------------------------|----------|------------|----------|----------|
| (Intercept) | 8.0646 | 0.0305 | 264.2631 | < 0.0001 |
| spec_nl_P9 | 0.5387 | 0.2488 | 2.1648 | 0.0304 |
| general_nl | 0.0684 | 0.0566 | 1.2094 | 0.2265 |
| weekend | -0.2438 | 0.0508 | -4.7958 | < 0.0001 |
| B. smooth terms | edf | Ref.df | F-value | p-value |
| s(price_P9) | 8.8971 | 8.9963 | 400.2889 | < 0.0001 |
| s(avg_subst_P9) | 4.6018 | 5.5389 | 33.1451 | < 0.0001 |
| s(trend) | 6.9910 | 8.0276 | 40.6523 | < 0.0001 |

Table 3.26: Regression results of the final Negative Binomial GAM for product P_9 .

Product P_{10}

Product P_{10} consists of the three articles A_{25} , A_{26} and A_{27} . It is of product type 'b' and there is one possible substitute product. The product did not appear in the article-specific advertisement and there are no competitor prices available.

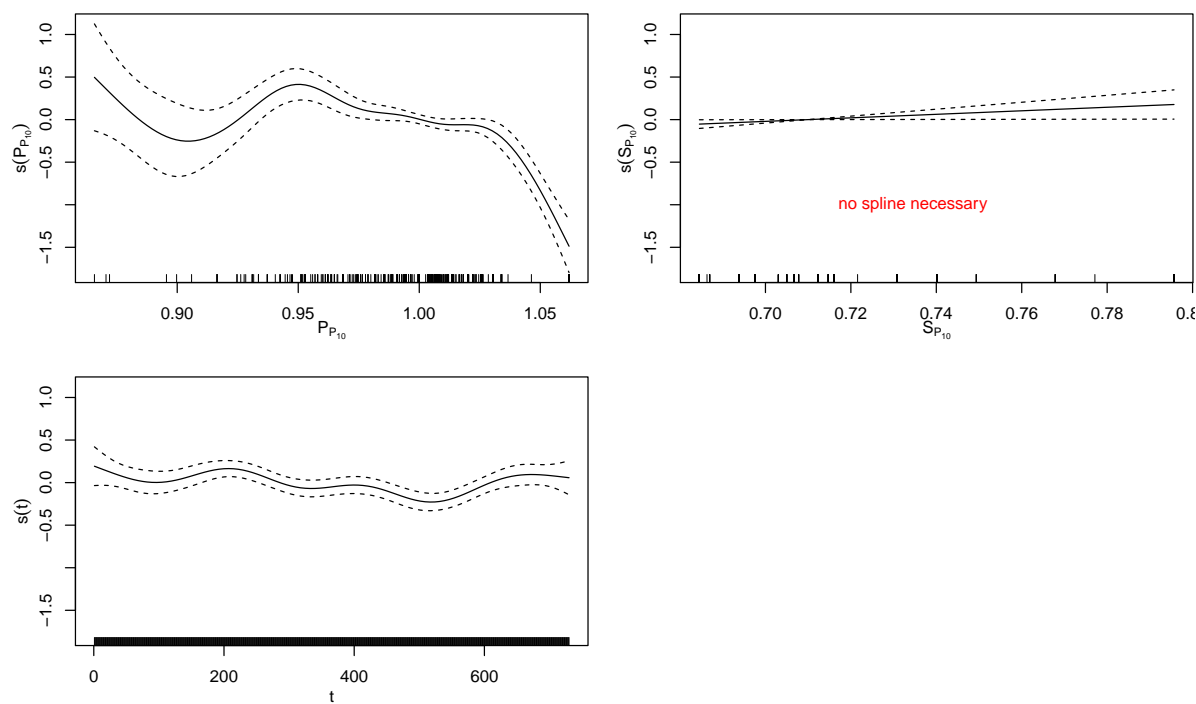


Figure 3.34: Plots of the splines which were fitted to the covariates of the Negative Binomial GAM for product P_{10} . Confidence intervals are also drawn.

We drop the spline for the average price of the substitute product.

| A. parametric coefficients | Estimate | Std. Error | t-value | p-value |
|----------------------------|----------|------------|----------|----------|
| (Intercept) | 3.6576 | 0.7123 | 5.1347 | < 0.0001 |
| general_nl | 0.0418 | 0.0437 | 0.9549 | 0.3396 |
| avg_subst_P10 | 2.0718 | 1.0032 | 2.0651 | 0.0389 |
| weekend | -0.1297 | 0.0400 | -3.2455 | 0.0012 |
| B. smooth terms | edf | Ref.df | F-value | p-value |
| s(price_P10) | 7.4661 | 8.3805 | 149.1011 | < 0.0001 |
| s(trend) | 7.5693 | 8.4921 | 31.6039 | 0.0002 |

Table 3.27: Regression results of the final Negative Binomial GAM for product P_{10} .

Product P_{11}

Product P_{11} consists of the two articles A_{28} and A_{29} and is the only product of product type 'd'. The product did not appear in the article-specific newsletter advertisement, but there are some competitor prices available.

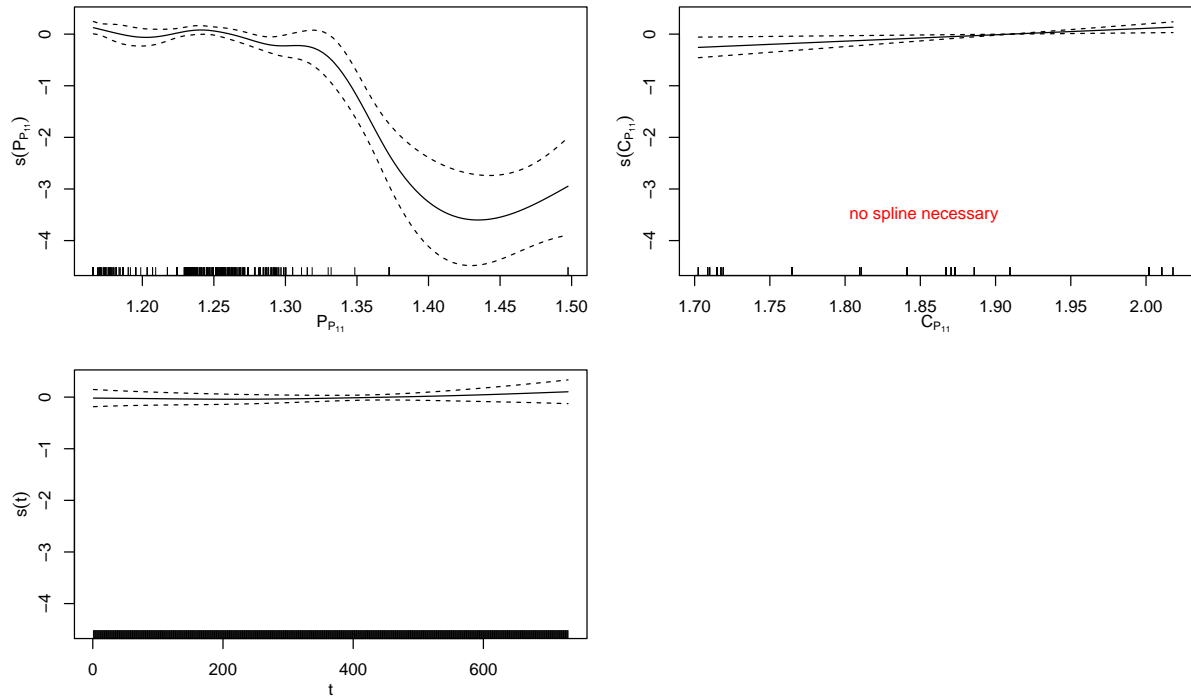


Figure 3.35: Plots of the splines which were fitted to the covariates of the Negative Binomial GAM for product P_{11} . Confidence intervals are also drawn.

We skip the spline for the average competitor price.

| A. parametric coefficients | Estimate | Std. Error | t-value | p-value |
|----------------------------|----------|------------|----------|----------|
| (Intercept) | 2.6689 | 1.2205 | 2.1867 | 0.0288 |
| general_nl | 0.0696 | 0.0500 | 1.3913 | 0.1641 |
| avg_comp_P11 | 1.0261 | 0.6389 | 1.6060 | 0.1083 |
| weekend | -0.2348 | 0.0455 | -5.1556 | < 0.0001 |
| B. smooth terms | edf | Ref.df | F-value | p-value |
| s(price_P11) | 7.7966 | 8.5645 | 119.7222 | < 0.0001 |
| s(trend) | 5.5153 | 6.7194 | 8.3696 | 0.3168 |

Table 3.28: Regression results of the final Negative Binomial GAM for product P_{11} .

Model interpretation

Inspecting the models we find, that the covariate for the product price needed a spline in every model. The expected global tendency of a negative price elasticity is clearly reflected. Some splines are rather 'wiggly', e.g. the one for the prices of product P_5 and P_9 or for the time trend of product P_6 . It is doubtful, whether this is meaningful in reality.

The adjusted R^2 as a measure for the goodness of fit looks as follows:

| Product | adjusted R^2 |
|----------|----------------|
| P_1 | 0.52 |
| P_2 | 0.66 |
| P_3 | 0.64 |
| P_4 | 0.55 |
| P_5 | 0.32 |
| P_6 | 0.33 |
| P_7 | 0.46 |
| P_8 | 0.64 |
| P_9 | 0.52 |
| P_{10} | 0.29 |
| P_{11} | 0.13 |

Table 3.29: Summary of the goodness of fit of the different models.

All in all, the model fit is satisfying.

3.4 Copula modelling on deviance residuals

In this section, we investigate the dependence structure between the deviance residuals of the Negative Binomial GAMs including time effects for the different products.

3.4.1 Exploratory data analysis

Before we start with copula modelling, we apply a probability integral transform to the deviance residuals, which are assumed to be asymptotically normally distributed. If the normal distribution however is no good approach, we check, if using a skew-t distribution yields better results.

This analysis will be done graphically. That means, we investigate the normal Q-Q plots of the deviance residuals to check the appropriateness of a normal distribution. If these plots indicate any lack of fit, we try to improve the approach fitting a skew-t distribution.

To be able to decide which of the distributions is more appropriate, we have a look at the histograms of the probability integral transform. Due to

$$F_Y(y) = P(Y \leq y) = P(F_X(X) \leq y) = P(X \leq F_X(y)^{-1}) = F_X(F_X(y)^{-1}) = y$$

we have $Y \sim U[0, 1]$ for any random variable Y . Hence, for our deviance residuals, we will always choose the distribution with the histogram which is closer to the one of a uniformly distributed random variable.

We use the following notation:

- r_i represents the deviance residuals for the Negative Binomial GAM including time effects for product P_i
- F_N represents the distribution function of any normal distribution
- F_t represents the distribution function of any skew-t distribution.

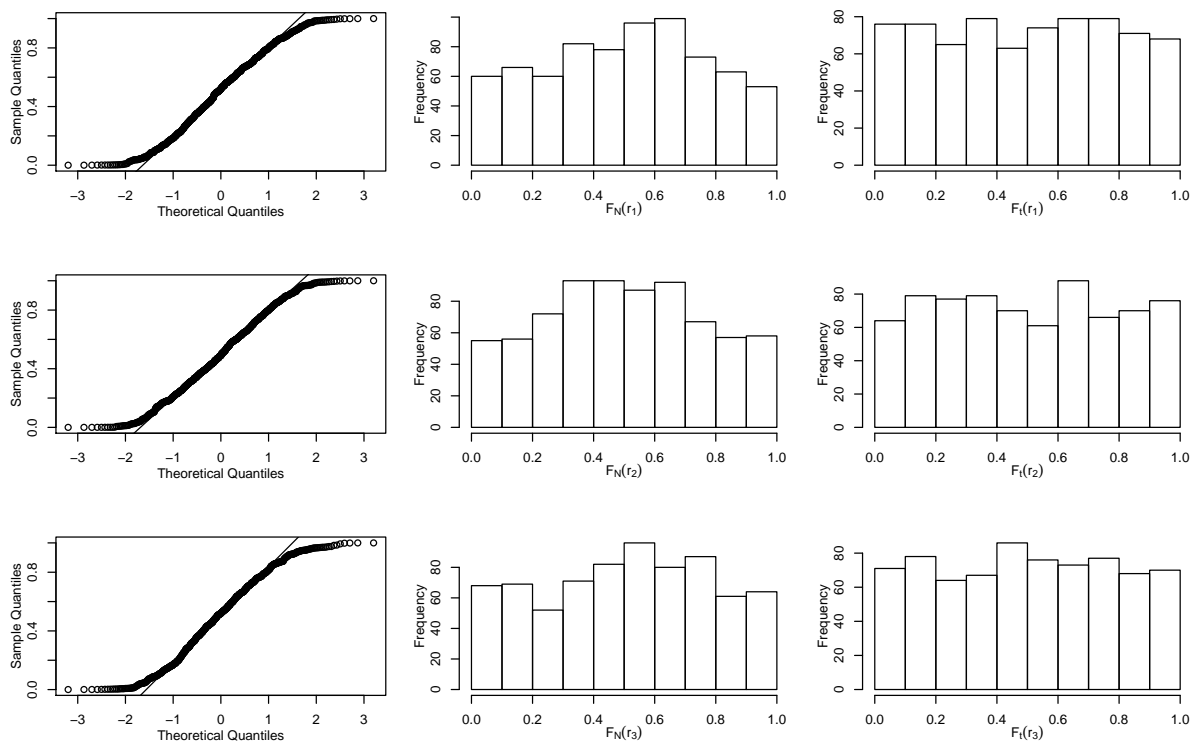


Figure 3.36: Residual analysis for the Negative Binomial GAMs of product P_1 to P_3 : For each of the products, there is drawn a normal Q-Q plot of the standardised deviance residuals and the corresponding histograms after having applied a normal probability integral transform and a skew-t probability integral transform, respectively.

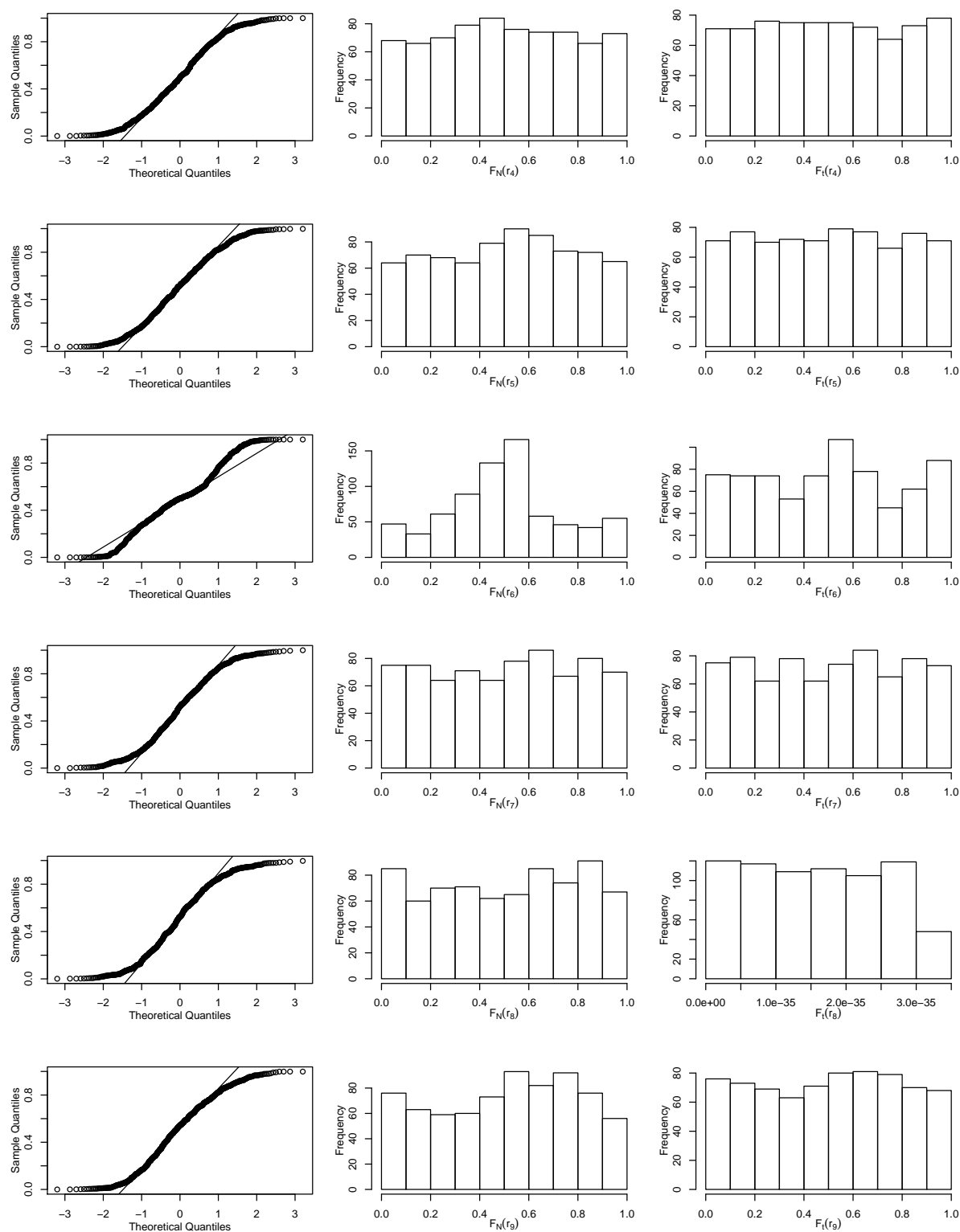


Figure 3.37: Residual analysis for the Negative Binomial GAMs of product P_4 to P_9 : For each of the products, there is drawn a normal Q-Q plot of the standardised deviance residuals and the corresponding histograms after having applied a normal probability integral transform and a skew-t probability integral transform, respectively.

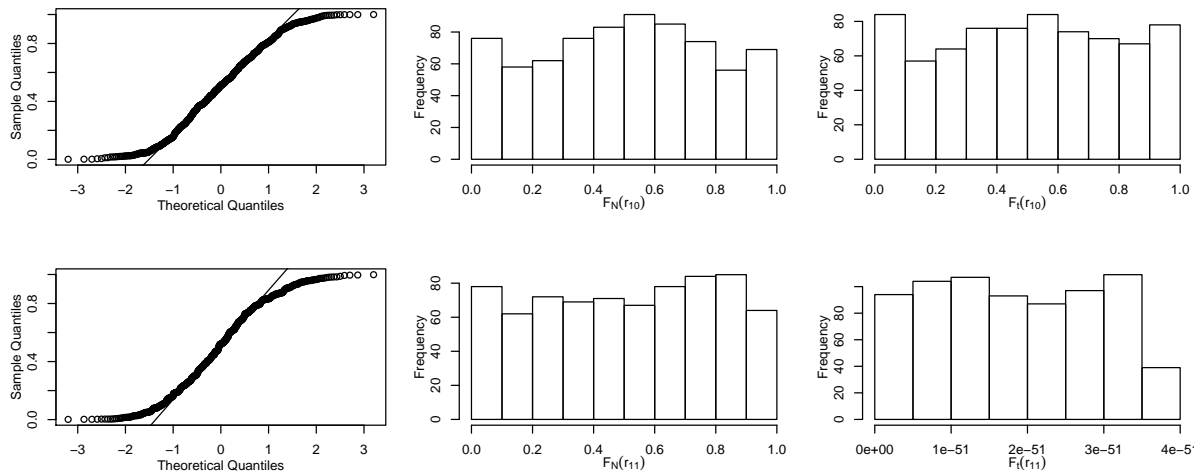


Figure 3.38: Residual analysis for the Negative Binomial GAMs of product P_{10} to P_{11} : For each of the products, there is drawn a normal Q-Q plot of the standardised deviance residuals and the corresponding histograms after having applied a normal probability integral transform and a skew-t probability integral transform, respectively.

From these plots we conclude, that for our data the skew-t distribution always fits better. However, for some of the products, none of these two distributions provides an appropriate approach. This especially is the case for product P_6 , P_8 , P_{10} and P_{11} . Without further refinement, these products thus are useless for further modelling. For this reason, they will be excluded in the following and we confine ourselves to the reduced data set consisting of the products P_1 , P_2 , P_3 , P_4 , P_5 , P_7 and P_9 .

The plot matrix below shows the histograms of the probability integral transform of the deviance residuals on the diagonal. Since we need a measure for the strength of dependence between the different products for further modelling, we furthermore compute the pairwise Kendall's τ , where higher absolute values indicate stronger dependencies. These are displayed in the upper panel. For better visualisation of the strength of dependencies we use a heat map, where a darker colouring means a stronger dependence. The lower panel shows the pairwise contour plots.

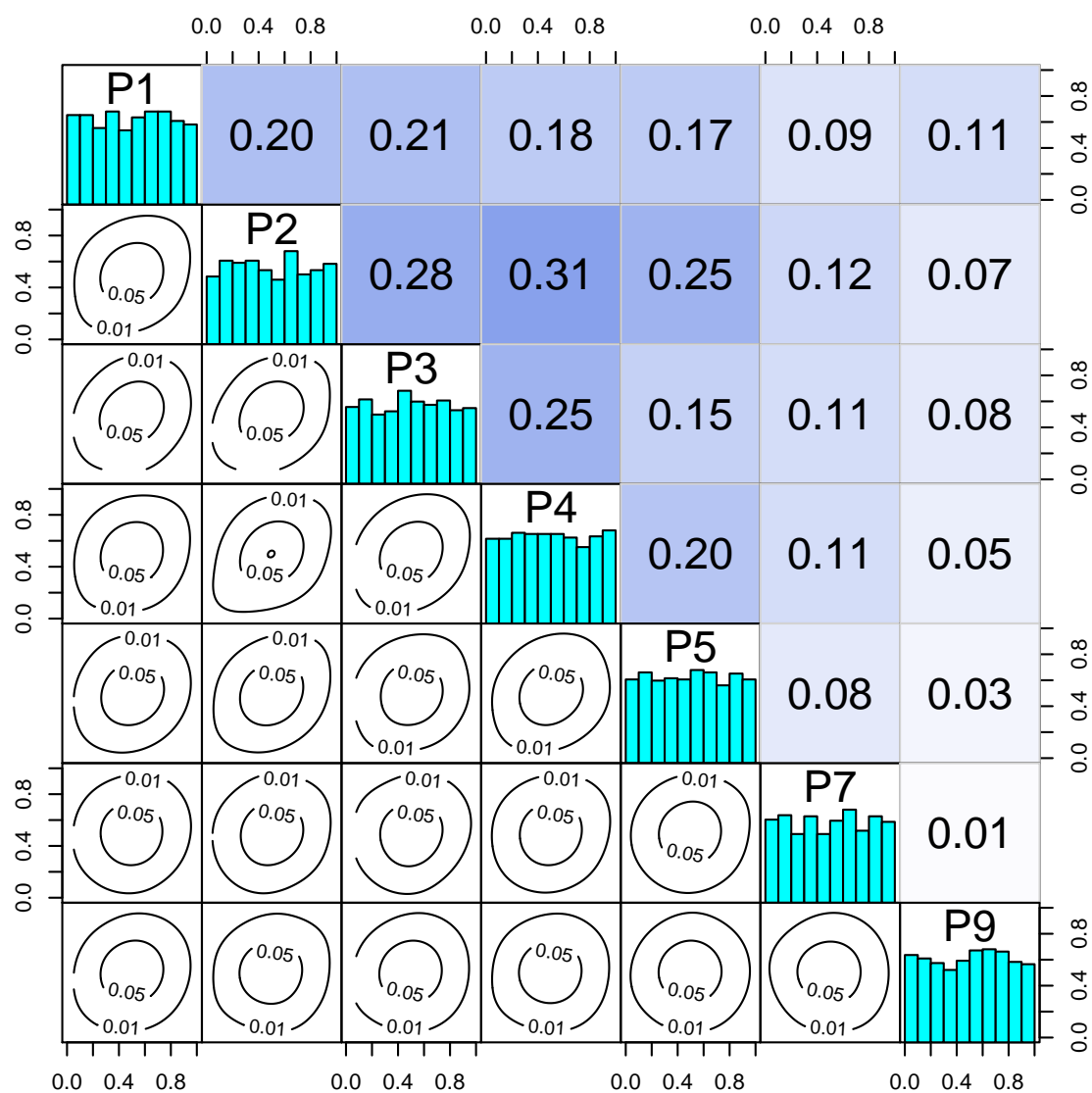


Figure 3.39: On the diagonal there are drawn the histograms of the probability integral transform of the deviance residuals of the Negative Binomial GAMs for the selected subset of products. The upper panel shows a heat map of the absolute values of the pairwise Kendall's τ and the lower panel the pairwise contour plots.

The absolute values of Kendall's τ all lie between 0.01 and 0.31. We conclude, that there only are weak dependencies present in our data. We further read out from the plot, that dependencies can mostly be detected between the deviance residuals of the models for product P_1 to P_5 , whereas dependencies are weaker regarding product P_7 and product P_9 . Putting that into the context of the structure of the product group as shown in figure 3.2, some results can be read out:

- The dependencies between the products P_1 , P_2 , P_4 and P_5 are rather strong. This is just what we have expected, because these four products all are of the same type. Since the absolute values of Kendall's τ are highest among the products belonging to brand B, we conclude, that dependencies within a brand are stronger than within a type.
- The private label products P_7 and P_9 apparently do not depend on each other and show weak dependencies to the remaining products, too. This is probably due to the fact, that the private label products are mostly cheaper than the products of the well-known brands, which makes price comparisons more or less unnecessary. So, price sensitive customers probably just buy one these products regardless of brand and without further price comparisons. This is reflected by the small values of the Kendall's τ .
- There are comparatively strong dependencies between the products P_1 to P_5 , which all come from well-known manufacturers and are popular brands. This nevertheless is surprising insofar, that product P_3 belongs to product type 'b', whereas the other products belong to product type 'a'. We conclude, that the brand awareness for these products is even stronger than the preference of a product type.

3.4.2 Copula modelling on the deviance residuals of the GAMs for a selected subset of products

To find an appropriate copula model for our data, we compare different approaches using R-Vines, Gauss-Vines, C-Vines and D-Vines. A Gauss-Vine is an R-Vine, where the only family permitted is the Gaussian copula. Each time we fit one vine object excluding the independence copula and another including the independence copula. This gives us a set of eight different models.

For each of these eight models, we start with a sequential estimation using a forward selection of trees. Based on this tree structure and the selected pair-copula families, we optimise the models via maximum likelihood estimation.

In total, we thus fitted 16 different models. The results are summarised in the table below. For each of the models, there are given the number of parameters and the log-likelihood, as well as the two information criteria AIC and BIC. The results for the sequentially estimated models are always denoted by '-seq' and for the maximum likelihood optimised model by '-mle', respectively. Further, '-ind' means, that we permitted for the independence copula.

| | par | loglik_seq | loglik_mle | AIC_seq | AIC_mle | BIC_seq | BIC_mle |
|----------------|-----|------------|------------|---------|---------|---------|---------|
| R-vine | 26 | 462.12 | 462.70 | -872.24 | -873.39 | -752.82 | -753.97 |
| R-vine-ind | 18 | 451.79 | 452.45 | -867.59 | -868.90 | -784.91 | -786.22 |
| Gauss-vine | 21 | 417.01 | 417.01 | -792.03 | -792.03 | -695.57 | -695.57 |
| Gauss-vine-ind | 14 | 410.83 | 410.83 | -793.66 | -793.66 | -729.36 | -729.36 |
| C-vine | 25 | 473.83 | 474.09 | -897.66 | -898.19 | -782.84 | -783.36 |
| C-vine-ind | 20 | 468.31 | 468.62 | -896.63 | -897.23 | -804.77 | -805.37 |
| D-vine | 26 | 458.04 | 458.72 | -864.08 | -865.44 | -744.66 | -746.02 |
| D-vine-ind | 18 | 446.51 | 447.13 | -857.02 | -858.27 | -774.35 | -775.59 |

Table 3.30: Sequential and maximum likelihood optimised approach of the fitted vine models. Each time the number of parameters and the log-likelihood, as well as the AIC and the BIC are displayed. The results for the sequentially estimated models are denoted by '-seq' and for the maximum likelihood optimised models by '-mle', respectively. If the independence copula is permitted, this is denoted by '-ind'.

We read out from the table, that for each of the different vine types the log-likelihood as well as the AIC and the BIC improve slightly in case of the maximum likelihood optimised model. We thus restrict our further analysis to these eight models.

If two models are of the same vine type, the model including the independence copula and the one excluding the independence copula of course are nested. This justifies the above evaluation using the AIC and the BIC. Concerning the full set of the eight maximum likelihood optimised models, these are not nested, but closely related to each other. For this reason, the AIC and the BIC cannot be used. Instead, we perform a Vuong test for the models each against the others, which is a correct method to compare non-nested models.

In the table below, 'stat' denotes the classical Vuong statistic whereas 'stat_A' and 'stat_S' denote the penalised versions thereof according to Akaike and Schwarz, respectively. The corresponding p-values are denoted by 'p_val', 'p_val_A' and 'p_val_S'.

| | stat | p_val | stat_A | p_val_A | stat_S | p_val_S |
|-------------------------------|-------|-------|--------|---------|--------|---------|
| R-vine vs. R-vine-ind | 2.21 | 0.03 | 0.48 | 0.63 | -3.48 | 0.00 |
| R-vine vs. Gauss-vine | 4.33 | 0.00 | 3.86 | 0.00 | 2.77 | 0.01 |
| R-vine vs. Gauss-vine-ind | 4.52 | 0.00 | 3.47 | 0.00 | 1.07 | 0.28 |
| R-vine vs. C-vine | -2.51 | 0.01 | -2.73 | 0.01 | -3.23 | 0.00 |
| R-vine vs. C-vine-ind | -1.10 | 0.27 | -2.21 | 0.03 | -4.77 | 0.00 |
| R-vine vs. D-vine | 1.08 | 0.28 | 1.08 | 0.28 | 1.08 | 0.28 |
| R-vine vs. D-vine-ind | 2.64 | 0.01 | 1.28 | 0.20 | -1.84 | 0.07 |
| R-vine-ind vs. Gauss-vine | 3.33 | 0.00 | 3.62 | 0.00 | 4.26 | 0.00 |
| R-vine-ind vs. Gauss-vine-ind | 3.99 | 0.00 | 3.61 | 0.00 | 2.73 | 0.01 |
| R-vine-ind vs. C-vine | -3.16 | 0.00 | -2.14 | 0.03 | 0.21 | 0.83 |
| R-vine-ind vs. C-vine-ind | -2.73 | 0.01 | -2.39 | 0.02 | -1.62 | 0.11 |
| R-vine-ind vs. D-vine | -1.05 | 0.29 | 0.29 | 0.77 | 3.37 | 0.00 |
| R-vine-ind vs. D-vine-ind | 1.28 | 0.20 | 1.28 | 0.20 | 1.28 | 0.20 |
| Gauss-vine vs. Gauss-vine-ind | 1.74 | 0.08 | -0.23 | 0.82 | -4.74 | 0.00 |
| Gauss-vine vs. C-vine | -5.22 | 0.00 | -4.85 | 0.00 | -4.01 | 0.00 |
| Gauss-vine vs. C-vine-ind | -4.68 | 0.00 | -4.77 | 0.00 | -4.98 | 0.00 |
| Gauss-vine vs. D-vine | -3.93 | 0.00 | -3.46 | 0.00 | -2.38 | 0.02 |
| Gauss-vine vs. D-vine-ind | -2.83 | 0.00 | -3.12 | 0.00 | -3.77 | 0.00 |
| Gauss-vine-ind vs. C-vine | -5.32 | 0.00 | -4.39 | 0.00 | -2.27 | 0.02 |
| Gauss-vine-ind vs. C-vine-ind | -5.08 | 0.00 | -4.55 | 0.00 | -3.34 | 0.00 |
| Gauss-vine-ind vs. D-vine | -4.19 | 0.00 | -3.14 | 0.00 | -0.73 | 0.47 |
| Gauss-vine-ind vs. D-vine-ind | -3.47 | 0.00 | -3.09 | 0.00 | -2.21 | 0.03 |
| C-vine vs. C-vine-ind | 1.72 | 0.08 | 0.15 | 0.88 | -3.46 | 0.00 |
| C-vine vs. D-vine | 2.65 | 0.01 | 2.82 | 0.00 | 3.22 | 0.00 |
| C-vine vs. D-vine-ind | 3.52 | 0.00 | 2.61 | 0.01 | 0.51 | 0.61 |
| C-vine-ind vs. D-vine | 1.52 | 0.13 | 2.44 | 0.01 | 4.56 | 0.00 |
| C-vine-ind vs. D-vine-ind | 3.01 | 0.00 | 2.73 | 0.01 | 2.09 | 0.04 |
| D-vine vs. D-vine-ind | 2.46 | 0.01 | 0.76 | 0.45 | -3.14 | 0.00 |

Table 3.31: Vuong tests for all possible pairs of maximum likelihood optimised vine models. The classical Vuong statistic as well as the two corrected versions according to Akaike and Schwarz are displayed with the respective p-values. The abbreviations 'stat', 'stat_A' and 'stat_S' denote the classical Vuong statistic, and the penalised Vuong statistic according to Akaike and Schwarz, respectively. The corresponding p-values are denoted by 'p_val', 'p_val_A' and 'p_val_S'.

Immediately the question arises, which of the above given statistics to use for the model selection. To decide on this, let us first compare the results from the Vuong test for all the three of them. Comparing the models each to the others, the table below summarises, how many times the individual models were preferred against another according to the different variants of the Vuong test. We distinguish between significant preference and non-significant preference at a significance level of 0.05.

| | sig | Nsig | tot | A_sig | A_Nsig | A_tot | S_sig | S_Nsig | S_tot |
|----------------|-----|------|-----|-------|--------|-------|-------|--------|-------|
| R-vine | 4 | 1 | 5 | 2 | 3 | 5 | 1 | 2 | 3 |
| R-vine-ind | 2 | 1 | 3 | 2 | 2 | 4 | 4 | 2 | 6 |
| Gauss-vine | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Gauss-vine-ind | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 |
| C-vine | 6 | 1 | 7 | 6 | 1 | 7 | 4 | 1 | 5 |
| C-vine-ind | 4 | 2 | 6 | 6 | 0 | 6 | 6 | 1 | 7 |
| D-vine | 3 | 1 | 4 | 2 | 1 | 3 | 1 | 1 | 2 |
| D-vine-ind | 2 | 0 | 2 | 2 | 0 | 2 | 3 | 1 | 4 |
| total | 21 | 7 | 28 | 20 | 8 | 28 | 20 | 8 | 28 |

Table 3.32: Summarising the results from the different versions of the Vuong test. The abbreviations 'sig', 'A_sig' and 'S_sig' denote a significant preference for the model, and 'Nsig', 'A_Nsig' and 'S_Nsig' a non-significant preference, respectively. The columns 'tot', 'A_tot' and 'S_tot' sum up the significant and the non-significant preferences for each of the models, whereas the row 'total' gives the overall numbers for the different versions of the Vuong test.

This table yields the following results:

- The non-penalised version of the Vuong test clearly prefers the C-vine.
- The penalised version according to Akaike yields a similar result, but here, we cannot decide between the C-vine and the C-vine-ind based on the significant preferences. Only when we include the non-significant preferences into decision making, we again come to the C-vine, which shows, that there only is a weak penalisation, which still prefers more complicated models.
- The penalisation of large number of parameters makes its presence felt, when correcting the Vuong statistic according to Schwarz, since we here clearly decide in favour of the C-vine-ind and not in favour of the C-vine without the independence copula.

Summarising the table, we notice, that the C-vine and the D-vine obviously yield different model fits. This is not very surprising, since both are special cases of R-vines but with strict boundary conditions, which are more or less suitable depending on the data.

Further, the Gauss-vine seems not to be an appropriate choice for our data. Also this is understandable regarding the families which were chosen in the different models, because the Gaussian family appears rather seldom. For more details on that, please refer to appendix 7.3.

But we wonder, why we always decide in favour of a C-vine: At first glance, we would of course expect the R-vine to yield the best results, since C-vines and D-vines are just special cases of R-vines. So, despite both these structures are permitted in case of an R-vine, the heuristic algorithm suggests another, apparently less fitting tree structure. But we notice, that the R-vine and the C-vine are very close in the Vuong test. So, let us investigate them in more detail.

We have already seen, that the values for Kendall's τ most often are rather small, and thus including the independence copula into the model absolutely makes sense. Further, we always prefer simple models according to 'KISS'. Thus, we base our decisions on the Vuong test statistic corrected according to Schwarz and compare the first tree of the two best fitting models, i.e. of the R-vine-ind and of the C-vine-ind, respectively:

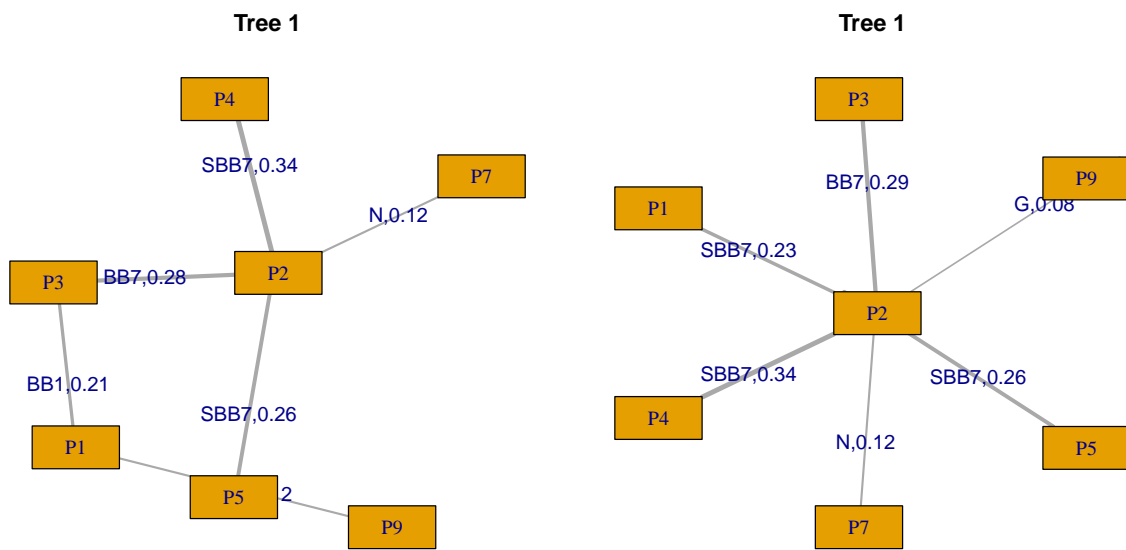


Figure 3.40: Comparing the R-vine-ind and the C-vine-ind: The first three of the R-vine-ind on the left, and the first tree of the C-vine-ind on the right.

The first tree of the R-vine-ind and the C-vine-in are very similar concerning the structure and the pair copula families. That the R-vine does not exactly correspond to the C-vine is probably due to the kind of 'soft facts' based decisions of the heuristic algorithm, since it obviously is not reasonable to try all possible vine structures. So, we conclude, that it absolutely makes sense to examine the special cases of C-vines and D-vines to achieve the best possible model fit.

In the following, we investigate our best fitting model, i.e. the C-vine-ind, in more detail. At first, we have a closer look at the values of the conditional pairwise Kendall's τ . These are summarised in the table below:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|-------|------|------|------|------|------|------|------|
| | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| tree6 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| tree5 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| tree4 | 0.07 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| tree3 | 0.07 | 0.00 | 0.07 | 0.08 | 0.00 | 0.00 | 0.00 |
| tree2 | 0.07 | 0.06 | 0.15 | 0.16 | 0.06 | 0.00 | 0.00 |
| tree1 | 0.26 | 0.12 | 0.34 | 0.23 | 0.29 | 0.08 | 0.00 |

Table 3.33: Values of the fitted Kendall's τ associated to the conditional pair copulas in the selected C-vine model (C-vine-ind).

Calculating the corresponding ranges of the fitted Kendall's τ for each of the trees in the vine tree sequence, we get the following:

| | min | max |
|--------|------|------|
| tree 1 | 0.08 | 0.34 |
| tree 2 | 0.06 | 0.16 |
| tree 3 | 0.00 | 0.08 |
| tree 4 | 0.00 | 0.07 |
| tree 5 | 0.00 | 0.00 |
| tree 6 | 0.00 | 0.00 |

Table 3.34: Ranges of the fitted Kendall's τ for each of the trees in the vine tree sequence of the selected C-vine model (C-vine-ind).

We check, if it is reasonable to consider the corresponding truncated C-vine at level 3 instead:

| | par | loglik | AIC | BIC |
|------------------|-----|--------|---------|---------|
| C-vine-ind | 20 | 468.62 | -897.23 | -805.37 |
| C-vine-ind-trunc | 18 | 461.25 | -886.51 | -803.83 |
| (C-vine) | 25 | 474.09 | -898.19 | -783.36 |
| (C-vine-trunc) | 19 | 464.18 | -890.36 | -803.09 |

Table 3.35: Comparing the selected C-vine model (C-vine-ind) and the corresponding model truncated at level 3. For information, there is also given the C-vine and the truncated version thereof at level 3. For each of the models, there are given the number of parameters, the log-likelihood, the AIC and the BIC.

Due to the model simplification, we normally expect an improvement of the AIC and the BIC by truncating the model. This is, what we get when comparing the C-vine without the independence copula with the truncated C-vine at level 3: There is a decline in the log-likelihood, but, since the truncated model has 6 parameters less than the non-truncated one, we notice a slight improvement in the AIC and in the BIC.

Comparing the C-vine-ind and the truncated version thereof at level 3, we make out, that the truncation does not improve the model. This is quite understandable, since we already allowed for the independence copula in the C-vine-ind. As we have seen above, this resulted in a model, which is truncated at level 4 and which additionally includes two further independence copulas at lower levels. So, the independence test reveals, that some pair copula family is more appropriate than the independence copula at level 4 and thus, the non-truncated model yields the better fit, which can be read out comparing the log-likelihoods. Despite the reduction of the parameters from 20 to 18, the AIC and the BIC did not improve, because the model fit however worsened. So, we stay with the non-truncated C-vine including the independence copula as our best fit.

Let us consider this best fitting model in more detail.

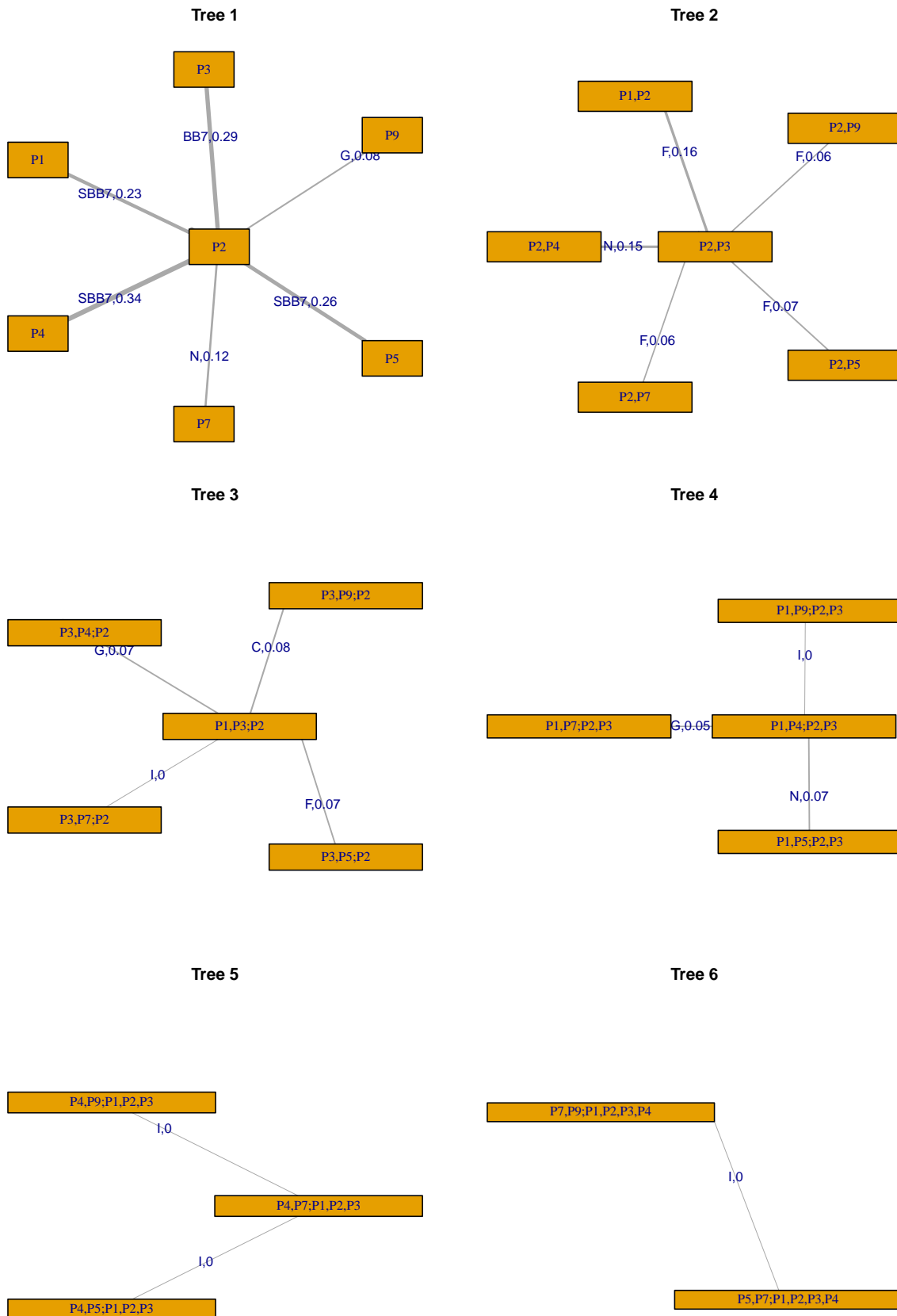


Figure 3.41: Vine tree sequence for the selected C-vine model (C-vine-ind).

The corresponding C-Vine Matrix is given by

$$\text{RVM} = \begin{bmatrix} 5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 7 & 6 & 0 & 0 & 0 & 0 & 0 \\ 6 & 7 & 4 & 0 & 0 & 0 & 0 \\ 4 & 4 & 7 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 7 & 3 & 0 & 0 \\ 3 & 3 & 3 & 3 & 7 & 2 & 0 \\ 2 & 2 & 2 & 2 & 2 & 7 & 7 \end{bmatrix}$$

This matrix describes the constraints within the trees, where each row corresponds to one tree.

The numbers in the matrix represent the index of the product in the data frame parsed to the function calculating the C-Vine.

| Vine Index | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|------------|-------|-------|-------|-------|-------|-------|-------|
| Product | P_1 | P_2 | P_3 | P_4 | P_5 | P_7 | P_9 |

Table 3.36: Mapping of the C-Vine matrix indices to the products.

The bottom row refers to the unconditional tree (tree 1). Going up the rows, you get the trees for copulas with preconditions. The third tree for example contains one edge described in column one, which is the connection between product P_1 and product P_5 under condition of product P_2 and product P_3 .

Next, we have a look at the chosen pair copula families and the corresponding family parameters:

$$\text{families} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 14 & 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 4 & 3 & 0 & 0 & 0 \\ 5 & 5 & 1 & 5 & 5 & 0 & 0 \\ 19 & 1 & 19 & 19 & 9 & 4 & 0 \end{bmatrix}$$

$$\text{par} = \begin{bmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.1052 & 1.0506 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.6757 & 0.0000 & 1.0732 & 0.1666 & 0.0000 & 0.0000 & 0.0000 \\ 0.5993 & 0.5739 & 0.2258 & 1.4292 & 0.5332 & 0.0000 & 0.0000 \\ 1.2526 & 0.1911 & 1.5664 & 1.1833 & 1.3090 & 1.0924 & 0.0000 \end{bmatrix}$$

$$\text{par2} = \begin{bmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.4341 & 0.0000 & 0.4266 & 0.3849 & 0.4639 & 0.0000 & 0.0000 \end{bmatrix}$$

Summarising the chosen pair copula families yields the following table:

| index | familyName | count |
|-------|----------------------|-------|
| 0 | Independence | 5 |
| 1 | Gaussian | 3 |
| 3 | Clayton | 1 |
| 4 | Gumbel | 2 |
| 5 | Frank | 5 |
| 9 | Joe-Clayton | 1 |
| 14 | Survival Gumbel | 1 |
| 19 | Survival Joe-Clayton | 3 |

From the table we read out, that there appear seven different pair copula families in the above tree structure. Furthermore, due to the small values of the Kendall's τ , the independence copula often is the most appropriate approach.

3.4.3 Simulating from the copula

Next, we simulate from our chosen copula model to check if we were able to describe our data properly. Based on this simulated data, we create a plot matrix with the corresponding histograms on the diagonal, with the absolute values of Kendall's τ in the upper panel and with the corresponding contour plots in the lower panel. We expect a plot similar to the one showing the observed data in figure 3.39.

Since the simulated values disperse in each simulation, we furthermore run another simulation with ten times as many simulated points. The following figure shows the plot matrix obtained from the observed data as well as the one based on a simulation of length 730 and on one of length 7300, respectively.

All the plots show a similar heat map for the Kendall's τ . The simulation from the 7300 simulated points has almost unified histograms and the absolute values of the Kendall's τ are very close to the ones obtained from the observed data. Hence, our model provides an acceptable fit.

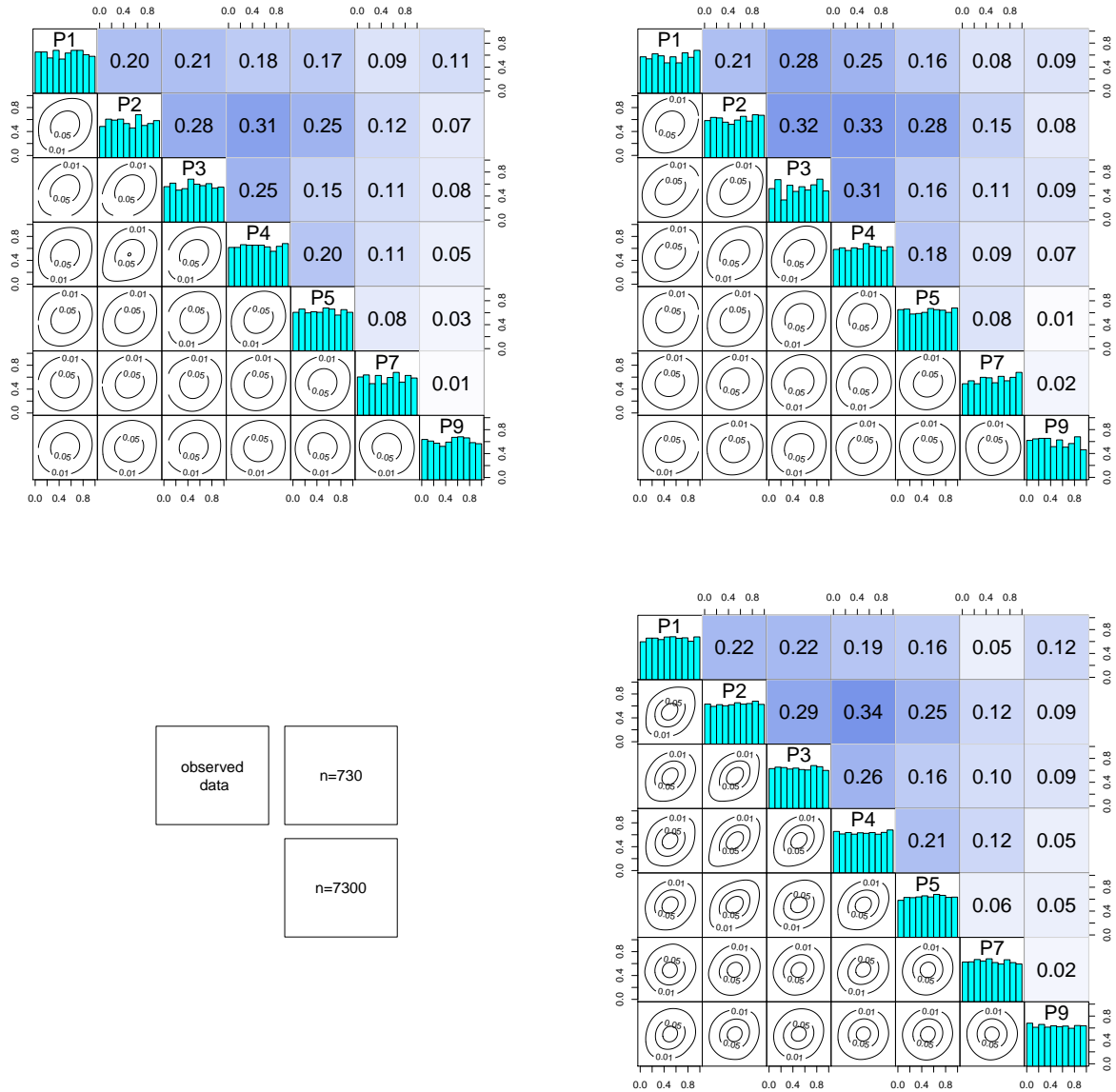


Figure 3.42: Observed data and simulation values of the selected C-vine model (C-vine-ind) with 730 and 7300 simulated values, respectively: On the diagonal there are drawn the histograms. The upper panel shows the pairwise Kendall's τ and the lower panel the pairwise contour plots.

Chapter 4

Validation of the results in the economical context

Having fitted a lot of different models to our data, we want to transfer the gained insights to our research question: *'Which influences do affect the sales to what extent?'*

For this purpose, we assess the goodness of fit within the different model types, compare the quality of prediction across the model types and check the explicability of the modelling results in the economical context. It is important to factor in these three aspects simultaneously, because considering only one single aspect could be misleading.

Furthermore, we put the models into the context of the questions arising from the shop business.

Goodness of fit

The goodness of fit has already been carefully optimized with the relevant techniques and examined in detail during the process of model fitting.

The typical measures for goodness of fit only allow to compare models of the same type, but nevertheless, they already give some absolute indication of suitability.

Regarding the models on article basis, we conclude, that all of the Poisson models showed an unacceptably bad fit, whereas the Negative Binomial models seem to be an appropriate choice for our data. Also regarding the generalized additive models, we reached an acceptable level of fit. The table below gives a short overview. 'Rd. of xx on yy df.' denotes 'Residual deviance of xx on yy degrees of freedom'.

| item | model | Goodness of fit | rating |
|-------|-----------------------------|-----------------------------------|--------|
| A_6 | simple Poisson | 4520.32 on 729 degrees of freedom | - |
| A_6 | + time effect | 4520.32 on 729 degrees of freedom | - |
| A_6 | + interaction | 4520.32 on 729 degrees of freedom | - |
| A_6 | simple NegBin | 968.94 on 729 degrees of freedom | ++ |
| A_6 | + time effect | 1271.2 on 729 degrees of freedom | ++ |
| A_6 | + interaction | 1294.49 on 729 degrees of freedom | ++ |
| P_3 | NegBin + time effects | 1651.09 on 729 degrees of freedom | ++ |
| P_3 | NegBin + time effects (GAM) | 0.64 | ++ |

Table 4.1: Summary of the goodness of fit of the different models.

Quality of prediction

Besides the goodness of fit, we need a common criterion to measure the quality of prediction independent of the type of the model. The probably most important method to use is the root of the mean squared error (RMSE), which is also known as the standard error of the regression. This measure is particularly suitable for comparison, since it decides on the width of the confidence intervals for the predictions in the sense that the 95% confidence interval is approximately equal to the prediction plus and minus twice this standard error. Please note, that the RMSE does not give an absolute indication of 'good' or 'bad', but that it is very suitable when comparing predictions relatively to each other.

Due to the fact that our regression models use log-link, we have to interpret the RMSE and the confidence intervals for the quantity sold. We did the following calculations:

The RMSE for the i -th value of the logarithmised quantity sold is derived as

$$RMSE_i = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_{i,s} - y_{i,o})^2},$$

where

- n is the number of observations
- $y_{i,s}$ is the i -th value of the simulated response
- $y_{i,o}$ is the i -th value of the observed response.

With the above definition of the RMSE, we use the following formula for the 95% confidence interval:

$$\log(\text{quantity sold}) := \text{prediction} \pm 1.96 \cdot \text{RMSE}.$$

This yields

$$\text{quantity sold} := e^{\text{prediction}} \cdot e^{\pm 1.96 \cdot \text{RMSE}}.$$

Thus, due to the log-link, the confidence interval for the quantity sold is defined by multiplying with the factor $e^{\pm 1.96 \cdot \text{RMSE}}$ instead of just adding $\pm 1.96 \cdot \text{RMSE}$.

This leads to the inequality

$$e^{-1.96 \cdot \text{RMSE}} \leq \frac{\text{quantity sold}}{e^{\text{prediction}}} \leq e^{+1.96 \cdot \text{RMSE}}.$$

Defining

$$A := \frac{\text{quantity sold}}{e^{\text{prediction}}}$$

as the accuracy of prediction, we can interpret our results: $A = 1$ means a perfect prediction. Thus, the smaller the confidence interval, the better the accuracy of prediction.

The following tables show the RMSE for the log-link and the 95% confidence interval for the accuracy of prediction. Let us first inspect article A_6 .

| article | model | RMSE | 95% confidence interval |
|---------|----------------|------|-------------------------|
| A_6 | simple Poisson | 0.32 | $0.54 \leq A \leq 1.87$ |
| A_6 | + time effect | 0.28 | $0.58 \leq A \leq 1.72$ |
| A_6 | + interaction | 0.27 | $0.59 \leq A \leq 1.7$ |
| A_6 | simple NegBin | 0.32 | $0.54 \leq A \leq 1.87$ |
| A_6 | + time effect | 0.28 | $0.58 \leq A \leq 1.72$ |
| A_6 | + interaction | 0.27 | $0.59 \leq A \leq 1.71$ |

Table 4.2: Overall comparison of the different models for article A_6 based on the root of the mean squared error, which is given for the logarithm of the quantity sold.

Comparing these models, the following points come up:

- Model type: Due to the strong relation between the Poisson type models and the Negative Binomial type models, it is not very surprising, that the RMSE behaves similarly. But this also shows, that decisions should not only be based on the RMSE, since the goodness of fit is not covered sufficiently. Consequently, due to the overall bad performance of the Poisson type models, it is not reasonable to pick one of these for application or for further analysis.
- Time effect: Enhancing the model by time effects, the RMSE drops noticeably.
- Interaction effects: Including interaction effects on top of the time effects only improves the RMSE slightly. It thus is not worth including the additional parameters that come into the model using this approach.

The table below compares the RMSE of the models for product P_3 and the ones of the Negative Binomial models including time effects for the articles belonging to product P_3 .

| item | model | RMSE | 95% confidence interval |
|-------|-----------------------------|------|-------------------------|
| A_6 | NegBin + time effects | 0.28 | $0.58 \leq A \leq 1.72$ |
| A_7 | NegBin + time effects | 0.29 | $0.56 \leq A \leq 1.78$ |
| A_8 | NegBin + time effects | 0.57 | $0.32 \leq A \leq 3.08$ |
| P_3 | NegBin + time effects | 0.35 | $0.5 \leq A \leq 2$ |
| P_3 | NegBin + time effects (GAM) | 0.28 | $0.57 \leq A \leq 1.75$ |

Table 4.3: Overall comparison of the different Negative Binomial models including time effects for product P_3 and for the articles belonging to P_3 , i.e. for article A_6 , A_7 and A_8 . Again, the comparison is based on the root of the mean squared error, which is given for the logarithm of the quantity sold.

Explainability

Explainability does not mean to explain, how the model parameters are derived. On the contrary, it means to explain, if the derived model can be aligned with the reality. To judge this, we need an appropriate indicator.

When setting up the models, we already spent some efforts on defining reasonable covariates. Now we evaluate the models and determine the lever of each of the covariates. We concentrate on the question: *'How big is the change of the predicted quantity sold when varying one covariate within the range of the observed values?'*

This lever does not reflect the statistical significance of the covariates in the model, but it describes the economical importance on the sales.

When calculating these levers, we assume the range of the individual covariates to be a set of 'typical values', out of which we calculate the maximal impact by taking the minimum and the maximum value of the observed data. We have to differentiate between several cases:

- the untransformed variables,
- the covariates that come into the model on logarithmic scale,
- the factor variables,
- the covariates to which a spline was fitted (in case of GAM).

For our purpose we define the lever of a covariate as the variance of the response on varying this covariate. The lever is characterised by a factor, which is the ratio of the

response at the maximal value of the covariate and the response at the minimal value of the covariate. This factor can be derived from the basic equation, which describes our models:

$$\widehat{\text{quantity sold}}_i = e^{\beta_0 + \beta_{\mathbf{x}_1} \cdot x_{1i} + \beta_{\mathbf{x}_2} \cdot \log(x_{2i}) + \dots}.$$

To calculate the factor for the untransformed covariate \mathbf{x}_1 we set

- x_{1i} as the minimal value of the covariate \mathbf{x}_1 ,
- x_{1j} as the maximal value of the covariate \mathbf{x}_1 and
- x_{2k} as an arbitrary value of \mathbf{x}_2 (not to be varied),

and receive:

$$\text{factor}_{\mathbf{x}_1} = \frac{e^{\beta_0 + \beta_{\mathbf{x}_1} \cdot x_{1i} + \beta_{\mathbf{x}_2} \cdot \log(x_{2k}) + \dots}}{e^{\beta_0 + \beta_{\mathbf{x}_1} \cdot x_{1j} + \beta_{\mathbf{x}_2} \cdot \log(x_{2k}) + \dots}} = \frac{e^{\beta_{\mathbf{x}_1} \cdot x_{1i}}}{e^{\beta_{\mathbf{x}_1} \cdot x_{1j}}} = e^{\beta_{\mathbf{x}_1} \cdot (x_{1i} - x_{1j})},$$

where all not-varied terms can be cancelled.

For the variable \mathbf{x}_2 , which is assumed to come into the model on logarithmic scale, the factor correspondingly calculates as:

$$\text{factor}_{\mathbf{x}_2} = \frac{e^{\beta_0 + \beta_{\mathbf{x}_1} \cdot x_{1k} + \beta_{\mathbf{x}_2} \cdot \log(x_{2i}) + \dots}}{e^{\beta_0 + \beta_{\mathbf{x}_1} \cdot x_{1k} + \beta_{\mathbf{x}_2} \cdot \log(x_{2j}) + \dots}} = \frac{e^{\beta_{\mathbf{x}_2} \cdot \log(x_{2i})}}{e^{\beta_{\mathbf{x}_2} \cdot \log(x_{2j})}} = \left(\frac{x_{2i}}{x_{2j}} \right)^{\beta_{\mathbf{x}_2}}.$$

Due to the strict monotonicity of $\beta_{\mathbf{x}_a} \cdot \mathbf{x}_a$ or $\beta_{\mathbf{x}_a} \cdot \log(\mathbf{x}_a)$, respectively, we can use the formulas above and get the minimum and maximum response on inspecting the minimum and maximum value of the covariate x_a . For a positive sign of $\beta_{\mathbf{x}_a}$ we get a factor greater than 1 meaning an *increasing* response on increasing covariate values. For a negative sign of $\beta_{\mathbf{x}_a}$ we get a factor less than 1 meaning a *decreasing* response on increasing covariate values.

For factor variables and GAM-splines we have to adjust our approach. Factor variables as well as GAM-splines do not have a strict monotonicity. Equally to before, we determine the variance of the response on varying one covariate in its observed range. To get a similar information like above, whether increasing covariate values on the whole lead to increasing response or not, we determine the all-over tendency by inspecting the slope of the simple linear regression of the covariate \mathbf{x}_i on the predicted response. Thus, we use the following formula, where

- x_{3i} denotes the value of the covariate \mathbf{x}_3 with the maximal response,
- x_{3j} denotes the value of the covariate \mathbf{x}_3 with the minimal response,
- x_{2k} denotes an arbitrary value of the covariate \mathbf{x}_2 (not to be varied) and

- m_r denotes the regression coefficient (slope) of the linear regression of the predicted response,

and receive:

$$\begin{aligned} \text{factor}_{x_3} &= \left(\frac{e^{\beta_0 + s(x_{3i}) + s(x_{2k}) + \dots}}{e^{\beta_0 + s(x_{3j}) + s(x_{2k}) + \dots}} \right)^{\text{sign}(m_r)} \\ &= \left(\frac{e^{s(x_{3i})}}{e^{s(x_{3j})}} \right)^{\text{sign}(m_r)} \\ &= \left(e^{s(x_{3i}) - s(x_{3j})} \right)^{\text{sign}(m_r)}. \end{aligned}$$

To handle factor variables, we have to prepare some data. The original covariate \mathbf{x}_i with its n levels is split up into $n - 1$ new covariates \mathbf{z}_{ik} with their corresponding β_k , where the first level of the original covariate is factored into the intercept. The latter can also be expressed as a new covariate with the corresponding $\beta = 0$. We shortly denote this in the equation:

$$\widehat{\text{quantity sold}}_i = e^{\beta_0 + \sum_k \beta_k \cdot z_{4ki} + \beta_{\mathbf{x}_2} \cdot x_{2i} + \dots},$$

where $i = 1$ denotes the first level of the covariate and subsequently $\beta_1 = 0$. The used level of the original covariate is indicated by setting the k -th new covariate \mathbf{z}_{ik} equal to 1 and setting the others equal to 0. This yields to n predictions for the n levels of the covariate. We do a linear regression of the covariate \mathbf{x}_i on the predicted response. Now we can use a similar formula like for GAM-splines.

To analyse the factor for the factored covariate \mathbf{x}_4 we use

- r as the level of the covariate \mathbf{x}_4 with the maximal response and thus we set $z_{4ri} = 1$,
- s as the level of the covariate \mathbf{x}_4 with the minimal response and thus we set $z_{4si} = 1$,
- x_{2k} as an arbitrary value of \mathbf{x}_2 (not to be varied) and
- m_r as the regression coefficient (slope) of the linear regression of the predicted response,

and receive:

$$\begin{aligned} \text{factor}_{\mathbf{x}_4} &= \left(\frac{e^{\beta_0 + \beta_r \cdot z_{4ri} + \beta_{\mathbf{x}_2} \cdot x_{2i} + \dots}}{e^{\beta_0 + \beta_s \cdot z_{4si} + \beta_{\mathbf{x}_2} \cdot x_{2i} + \dots}} \right)^{\text{sign}(m_r)} \\ &= \left(\frac{e^{\beta_r}}{e^{\beta_s}} \right)^{\text{sign}(m_r)} \\ &= \left(e^{\beta_r - \beta_s} \right)^{\text{sign}(m_r)}. \end{aligned}$$

For easier comprehension, we transform this factor into a percentage increase or decrease compared to the base value, which is our desired lever:

$$\text{lever} = (\text{factor} - 1) \cdot 100\%.$$

The following two tables show these levers exemplary for our favourite models for article A_6 and product P_3 , respectively.

| NegBin with time effects A_6 | | | |
|--------------------------------|-----------------|-------|--------|
| covariate | range | lever | impact |
| price_A6 | [1.55 , 1.78] | -34 % | +++ |
| spec_nl_A6 | {0 ; 1} | 4 % | |
| general_nl | {0 ; 1} | 10 % | + |
| avg_comp_A6 | [1.69 , 1.84] | -2 % | |
| avg_prod_A6 | [1.28 , 1.5] | -8 % | + |
| avg_subst_P3 | [0.96 , 1.06] | 15 % | ++ |
| weekend | {0 ; 1} | -30 % | +++ |
| trend | {0 ; ... ; 730} | 16 % | ++ |

Table 4.4: Levers of the different covariates on the predicted quantity sold in the Negative Binomial model including time effects for article A_6 .

In general, all the levers are quite plausible regarding their sign and amount. It is not very surprising, that the article price has the largest influence on the sales. Also the high lever of the weekend can easily be explained: Since the customers are not bound to opening hours, they probably spend their free time on the weekend on leisure activities. The small lever of the article-specific newsletter advertisement may result from the sparseness of this variable and the general unattractiveness of this article compared to the remaining articles belonging to product P_3 , where the latter may also explain the low influence of the competitor prices. The time trend reflects the general growth of the shop.

Next, we investigate the whole product P_3 in more detail. Please remember, that the covariates concerning prices are measured per half kilogram in the corresponding models. This is the reason for the large differences in the price ranges compared to article A_6 .

| NegBin with time effects (GAM) P_3 | | | |
|--------------------------------------|-----------------|-------|--------|
| covariate | range | lever | impact |
| price_P3 | [0.58 , 0.79] | -91 % | +++ |
| spec_nl_P3 | {0 ; 1} | 10 % | + |
| general_nl | {0 ; 1} | 12 % | + |
| avg_comp_P3 | [0.75 , 0.81] | -28 % | ++ |
| avg_subst_P3 | [0.96 , 1.06] | 99 % | +++ |
| weekend | {0 ; 1} | -28 % | ++ |
| trend | {0 ; ... ; 730} | -61 % | +++ |

Table 4.5: Levers of the different covariates on the predicted quantity sold in the Negative Binomial model including time effects (GAM) for product P_3 .

Also for the whole product, most of the levers are quite plausible regarding their sign and amount. The high lever of the article-specific newsletter advertisement results from the large packaging sizes and multipacks, which are included when modelling the whole product. This also explains the high lever of the price. The negative time trend may result from the fact, that the special offer article with very high quantity sold was removed from the shop's assortment during the time period of observation. Although the splines fitted to the covariates lead to a good model fit, the negative time trend is hardly to be reasonable for prediction. If this trend was true, the demand for this product would shrink to a negligible amount. This product thus needs some attention in order to mitigate this.

Chapter 5

Summary and Outlook

We started with a large data set from the online shop. After inspecting and cleaning the data, we restricted to the 29 articles with sufficient data quality.

Having analysed this data, we used different modelling approaches regarding model type and aggregation level of the articles. We assumed Poisson and Negative Binomial distributions for the response variable and used GLMs, GAMs and copula models on deviance residuals to describe our data.

Three levels of aggregation were examined: the articles, the products and the product group as a whole.

In the end, the following covariates turned out to be reasonable:

- Influences controlled by the online shop:
The sales price, the article-specific newsletter and the general newsletter are important control variables for the online shop to stimulate sales.
- Market influence:
Of course, the competitor prices are very important. So they are consequently monitored and taken into account in the pricing strategy.
- Customer behaviour:
Sales depend on the customers' behaviour and preferences. The decision for purchasing a single article among other things depends on the availability of attractive substitute items of the online shop. These substitution effects have been considered and modelled on each level of aggregation.
- Seasonal effects:
The influence of the weekend and of the all-over time trend were remarkable.

In general, good model fits were achieved. Further, the modelling results were explainable and well balanced considering simplicity and preciseness. Interactions between the chosen covariates seem not to be relevant.

Validating the models, we see a good support to optimise our pricing strategy and the product assortment. Of course, the expected tendencies regarding e.g. the price elasticity and the effect of advertising were confirmed. At the starting point we were already

aware of the dependencies between the response variable and the covariates in a qualitative manner. During this thesis we quantified the effects by retrieving the corresponding parameters in the simple case of GLMs or have the regression splines in case of GAMs. Thus, we are now able to predict the customer demand and can support the optimisation of the strategy of the online shop.

Article-based models

The models for the individual articles can be used for the day-to-day business of pricing. For example, if our supply is short we could adjust the prices according to the model coefficients. This reduces sales in a manner to avoid running out of stock, which otherwise might annoy some of the customers. This way, despite the poor availability, the shop is able to satisfy the customers' demand sufficiently.

Product-based models

The product model however is useful when analysing the shop's strategy. The influences of rather general covariates like for example advertisement or competitor prices can be interpreted more easily when regarding the whole product, since there might just be some shifts among the different articles belonging to the product, but no overall effect on the product itself, which cannot be seen when just regarding one single article at a time. Furthermore, the product model can help to make out the 'position' of the product in the shop's assortment, which may be helpful if we are interested in how to compose the assortment.

Copula-based models for the product group

A first assumption was, that we do not expect a substitution effect between different product types. A second assumption was, that grouping the products by brand for evaluation of substitution effects generally is a sufficient approach for modelling. Doing copula modelling, some more elaborate dependencies were found. Within brand B, P_2 and P_4 have a stronger dependency than both of them with P_5 . Surprisingly, a comparatively strong dependency between P_3 from 'type b' and P_2 and P_4 from 'type a' is derived. For the future shop strategy, it will be helpful to understand the underlying mechanism of customer behaviour and subsequently optimise the portfolio using these insights. Of course, one has to be careful, because the underlying mechanisms may not be basic rules and hence may vary due to 'temporary fashion'.

Outlook

Besides using the models for now, some further activities are left. The current models predict the quantity sold. The gross profit, which is most important in the end, has to be calculated separately. If the price is set below the break-even-point,

then the predicted quantity sold may be incredible high, but you receive losses. Doing the maths, you can optimise the gross profit in absolute figures.

All in all, we received relevant models for the different levels of the shop's strategy. Since the parameters used in the models are comparatively volatile e.g. due to new products, eco-trends or health-trends, it is recommended to implement a continuous monitoring process to keep the models up to date.

Chapter 6

Indices

6.1 List of Figures

| | | |
|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 3.1 | Shop structure consisting of several levels: Category, product group, sub-group, product and article. | 51 |
| 3.2 | Hierarchical structure of the considered product group showing the selected 29 articles. For the ease of use, the brand of the product is always displayed below the product in italics. | 54 |
| 3.3 | Box plots of the quantity sold on daily basis. Each article is displayed separately. | 58 |
| 3.4 | Box plots of the quantity sold over the shop price for the individual articles. The data is given on daily basis. The box plots for all of the articles can be found in the appendix 7.1 to appendix 7.3. | 60 |
| 3.5 | Bar plot showing the number of competitors to be tracked for each article. | 62 |
| 3.6 | Three dimensional scatter plots of the average daily sales for every pair of shop price and currently cheapest competitor price. Of course, only the articles with competitor prices available are considered. | 63 |
| 3.7 | Bar plot showing the number of days on which the individual articles appeared in the article-specific newsletter. | 64 |
| 3.8 | Box plots showing the relative uplift of sales induced by the article-specific advertisement, by general newsletters and by both kinds of newsletter at the same time, respectively. On the left and on the right, only articles, which appear at least once in the article-specific newsletter are included. In the middle, all the 29 articles are considered. | 66 |
| 3.9 | Box plots visualising the effect of the different kinds of advertisement on the articles appearing at least once in the article-specific newsletter. | 67 |
| 3.10 | Part 1: Box plots visualising the effect of the different kinds of advertisement for the articles without article-specific advertisement. | 68 |
| 3.11 | Part 2: Box plots visualising the effect of the different kinds of advertisement for the articles without article-specific advertisement. | 68 |

| | | |
|------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 3.12 | This graphic exemplary illustrates the grouping of the articles to form suitable covariates covering possible substitution effects. Each of the green boxes forms one covariate, when analysing article A_{22} ; each of the red boxes for article A_6 , respectively. | 71 |
| 3.13 | EDA-plots of the covariates versus the log mean rate estimate. | 73 |
| 3.14 | Scatter plots of the response variable against the average prices of comparable articles. | 74 |
| 3.15 | Quantity sold of article A_6 clustered by month. | 75 |
| 3.16 | Quantity sold of article A_6 clustered by weekday and year 1 and year 2, respectively. | 75 |
| 3.17 | Interaction plot matrix for the set of covariates. | 76 |
| 3.18 | Residual analysis for the Poisson model including time effects and interactions: The left upper plot shows the deviance residuals for the 730 observations and the right upper plot the normal Q-Q plot of the deviance residuals. The lower plots show the expected response and exemplary one covariate against the deviance residuals. | 82 |
| 3.19 | Check of the log-link specification: Plot of the linear predictor versus the response variable. | 83 |
| 3.20 | Residual analysis for Negative Binomial regression model including time effects and interactions: The left upper plot shows the deviance residuals of the 730 observations and the right upper plot the Q-Q plot of the deviance residuals. The lower plots show the expected response and exemplary one covariate against the deviance residuals. The full set of plots can be found in the appendix 7.6 and ?? | 87 |
| 3.21 | Checking for linearity: Plots of the logarithm of the response variable versus the covariates. | 89 |
| 3.22 | Seasonal effects on monthly basis for product P_3 . The curves for the individual articles belonging to P_3 are also shown. | 90 |
| 3.23 | Checking for weekday seasonality of product P_3 . The curves for the individual articles, which belong to product P_3 , are also shown. | 90 |
| 3.24 | Residual analysis for the Negative Binomial regression model including time effects for product P_3 : The left upper plot shows the deviance residuals for the 730 observations and the right upper plot the normal Q-Q plot of the deviance residuals. The lower plots show the expected response and exemplary one of the covariates against the deviance residuals. The full set of plots can be found in the appendix 7.7. | 92 |
| 3.25 | Plots of the splines which were fitted to the covariates of the Negative Binomial GAM for product P_1 . Confidence intervals are also drawn. | 96 |
| 3.26 | Plots of the splines which were fitted to the covariates of the Negative Binomial GAM for product P_2 . Confidence intervals are also drawn. | 97 |
| 3.27 | Plots of the splines which were fitted to the covariates of the Negative Binomial GAM for product P_3 . Confidence intervals are also drawn. | 98 |
| 3.28 | Plots of the splines which were fitted to the covariates of the Negative Binomial GAM for product P_4 . Confidence intervals are also drawn. | 99 |

| | | |
|------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 3.29 | Plots of the splines which were fitted to the covariates of the Negative Binomial GAM for product P_5 . Confidence intervals are also drawn. | 100 |
| 3.30 | Plots of the splines which were fitted to the covariates of the Negative Binomial GAM for product P_6 . Confidence intervals are also drawn. | 101 |
| 3.31 | Plots of the splines which were fitted to the covariates of the Negative Binomial GAM for product P_7 . Confidence intervals are also drawn. | 102 |
| 3.32 | Plots of the splines which were fitted to the covariates of the Negative Binomial GAM for product P_8 . Confidence intervals are also drawn. | 103 |
| 3.33 | Plots of the splines which were fitted to the covariates of the Negative Binomial GAM for article A_9 . Confidence intervals are also drawn. | 104 |
| 3.34 | Plots of the splines which were fitted to the covariates of the Negative Binomial GAM for product P_{10} . Confidence intervals are also drawn. | 105 |
| 3.35 | Plots of the splines which were fitted to the covariates of the Negative Binomial GAM for product P_{11} . Confidence intervals are also drawn. | 106 |
| 3.36 | Residual analysis for the Negative Binomial GAMs of product P_1 to P_3 : For each of the products, there is drawn a normal Q-Q plot of the standardised deviance residuals and the corresponding histograms after having applied a normal probability integral transform and a skew-t probability integral transform, respectively. | 108 |
| 3.37 | Residual analysis for the Negative Binomial GAMs of product P_4 to P_9 : For each of the products, there is drawn a normal Q-Q plot of the standardised deviance residuals and the corresponding histograms after having applied a normal probability integral transform and a skew-t probability integral transform, respectively. | 109 |
| 3.38 | Residual analysis for the Negative Binomial GAMs of product P_{10} to P_{11} : For each of the products, there is drawn a normal Q-Q plot of the standardised deviance residuals and the corresponding histograms after having applied a normal probability integral transform and a skew-t probability integral transform, respectively. | 110 |
| 3.39 | On the diagonal there are drawn the histograms of the probability integral transform of the deviance residuals of the Negative Binomial GAMs for the selected subset of products. The upper panel shows a heat map of the absolute values of the pairwise Kendall's τ and the lower panel the pairwise contour plots. | 111 |
| 3.40 | Comparing the R-vine-ind and the C-vine-ind: The first three of the R-vine-ind on the left, and the first tree of the C-vine-ind on the right. | 116 |
| 3.41 | Vine tree sequence for the selected C-vine model (C-vine-ind). | 119 |
| 3.42 | Observed data and simulation values of the selected C-vine model (C-vine-ind) with 730 and 7300 simulated values, respectively: On the diagonal there are drawn the histograms. The upper panel shows the pairwise Kendall's τ and the lower panel the pairwise contour plots. | 122 |
| 7.1 | Article A_1 to article A_6 : Box plots of the quantity sold over the shop price for the individual articles. The data is given on daily basis. | 142 |

| | | |
|-----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 7.2 | Article A_7 to article A_{18} : Box plots of the quantity sold over the shop price for the individual articles. The data is given on daily basis. | 143 |
| 7.3 | Article A_{19} to article A_{29} : Box plots of the quantity sold over the shop price for the individual articles. The data is given on daily basis. | 144 |
| 7.4 | Residual analysis for the Poisson model for article A_6 including time effects and interactions (part 1): Plots showing the covariates against the deviance residuals. | 145 |
| 7.5 | Residual analysis for the Poisson model for article A_6 including time effects and interactions (part 2): Plots showing the covariates against the deviance residuals. | 146 |
| 7.6 | Residual analysis for the Negative Binomial model for article A_6 including time effects and interactions: Plots showing the covariates against the deviance residuals. | 147 |
| 7.7 | Residual analysis for the Negative Binomial model including time effects for product P_3 : Plots showing the covariates against the deviance residuals. | 148 |

6.2 List of Tables

| | | |
|------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 2.1 | Relationship between the Kendall's τ and the copula parameters for selected elliptical and Archimedean copula families. | 48 |
| 3.1 | Available data fields in the data source. | 53 |
| 3.2 | Grouping of articles, products and brands in absolute numbers. | 55 |
| 3.3 | Grouping of articles, products and brands in percentage. | 55 |
| 3.4 | This table matches the articles with the corresponding product, type, brand, weight and bundle size. Furthermore, it is displayed, if there is given away a bonus article for free. To get an overview over the importance of the different articles, there is also given the average quantity sold per day. | 56 |
| 3.5 | Availability of the articles given in absolute numbers and in percentage share. | 57 |
| 3.6 | Dynamics of the prices per article: The number of prices gives the cardinality of the set of different prices, whereas a price change means switching between the prices available. | 60 |
| 3.7 | Statistical classification and description of the variables used for modelling. | 77 |
| 3.8 | Regression results for article A_6 in the simple Poisson regression model. | 78 |
| 3.9 | χ^2 -test of the models with and without the factor variable 'article price', respectively. | 78 |
| 3.10 | Regression results for article A_6 in the Poisson regression model including time effects. | 79 |
| 3.11 | Regression results for article A_6 in the Poisson regression model including time effects and interactions. | 80 |
| 3.12 | χ^2 -test for the Poisson model including time effects and the Poisson model with interactions. | 81 |
| 3.13 | Regression results for article A_6 in the simple Negative Binomial regression model. | 84 |

| | | |
|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 3.14 | Regression results for article A_6 in the Negative Binomial regression model including time effects. | 85 |
| 3.15 | Regression results for article A_6 in the Negative Binomial regression model including time effects and interactions. | 85 |
| 3.16 | Regression results for article A_6 in the 'refitted' Negative Binomial regression model including time effects and interactions. | 86 |
| 3.17 | Regression results for product P_3 in the Negative Binomial regression model including time effects. | 91 |
| 3.18 | Regression results of the final Negative Binomial GAM for product P_1 . . . | 96 |
| 3.19 | Regression results of the final Negative Binomial GAM for product P_2 . . . | 97 |
| 3.20 | Regression results of the final Negative Binomial GAM for product P_3 . . . | 98 |
| 3.21 | Regression results of the final Negative Binomial GAM for product P_4 . . . | 99 |
| 3.22 | Regression results of the final Negative Binomial GAM for product P_5 . . . | 100 |
| 3.23 | Regression results of the final Negative Binomial GAM for product P_6 . . . | 101 |
| 3.24 | Regression results of the final Negative Binomial GAM for product P_7 . . . | 102 |
| 3.25 | Regression results of the final Negative Binomial GAM for product P_8 . . . | 103 |
| 3.26 | Regression results of the final Negative Binomial GAM for product P_9 . . . | 104 |
| 3.27 | Regression results of the final Negative Binomial GAM for product P_{10} . . . | 105 |
| 3.28 | Regression results of the final Negative Binomial GAM for product P_{11} . . . | 106 |
| 3.29 | Summary of the goodness of fit of the different models. | 107 |
| 3.30 | Sequential and maximum likelihood optimised approach of the fitted vine models. Each time the number of parameters and the log-likelihood, as well as the AIC and the BIC are displayed. The results for the sequentially estimated models are denoted by 'seq' and for the maximum likelihood optimised models by 'mle', respectively. If the independence copula is permitted, this is denoted by 'ind'. | 113 |
| 3.31 | Vuong tests for all possible pairs of maximum likelihood optimised vine models. The classical Vuong statistic as well as the two corrected versions according to Akaike and Schwarz are displayed with the respective p-values. The abbreviations 'stat', 'stat_A' and 'stat_S' denote the classical Vuong statistic, and the penalised Vuong statistic according to Akaike and Schwarz, respectively. The corresponding p-values are denoted by 'p_val', 'p_val_A' and 'p_val_S'. | 114 |
| 3.32 | Summarising the results from the different versions of the Vuong test. The abbreviations 'sig', 'A_sig' and 'S_sig' denote a significant preference for the model, and 'Nsig', 'A_Nsig' and 'S_Nsig' a non-significant preference, respectively. The columns 'tot', 'A_tot' and 'S_tot' sum up the significant and the non-significant preferences for each of the models, whereas the row 'total' gives the overall numbers for the different versions of the Vuong test. | 115 |
| 3.33 | Values of the fitted Kendall's τ associated to the conditional pair copulas in the selected C-vine model (C-vine-ind). | 117 |
| 3.34 | Ranges of the fitted Kendall's τ for each of the trees in the vine tree sequence of the selected C-vine model (C-vine-ind). | 117 |

| | | |
|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 3.35 | Comparing the selected C-vine model (C-vine-ind) and the corresponding model truncated at level 3. For information, there is also given the C-vine and the truncated version thereof at level 3. For each of the models, there are given the number of parameters, the log-likelihood, the AIC and the BIC. | 117 |
| 3.36 | Mapping of the C-Vine matrix indices to the products. | 120 |
| 4.1 | Summary of the goodness of fit of the different models. | 124 |
| 4.2 | Overall comparison of the different models for article A_6 based on the root of the mean squared error, which is given for the logarithm of the quantity sold. | 125 |
| 4.3 | Overall comparison of the different Negative Binomial models including time effects for product P_3 and for the articles belonging to P_3 , i.e. for article A_6 , A_7 and A_8 . Again, the comparison is based on the root of the mean squared error, which is given for the logarithm of the quantity sold. . | 126 |
| 4.4 | Levers of the different covariates on the predicted quantity sold in the Negative Binomial model including time effects for article A_6 | 129 |
| 4.5 | Levers of the different covariates on the predicted quantity sold in the Negative Binomial model including time effects (GAM) for product P_3 . . . | 130 |
| 7.1 | Regression results for article A_7 in the Negative Binomial regression model including time effects. | 149 |
| 7.2 | Regression results for article A_8 in the Negative Binomial regression model including time effects. | 149 |
| 7.3 | χ^2 -test for article A_6 : The Negative Binomial model including time effects with and without interactions. | 150 |

6.3 Bibliography

Fahrmeir, L.; Künstler, R.; Pigeot, I. & Tutz, G. (2001): Statistik: Der Weg zur Datenanalyse (3. Auflage). Berlin/Heidelberg/New York: Springer.

Tukey, J.W. (1977): Exploratory Data Analysis. Reading, Massachusetts/Menlo Park, California/London/Amsterdam/Don Mills, Ontario/Sydney: ADDISON-WESLEY PUBLISHING COMPANY.

Yamamura, M. (2012): A Report on Seminar Titled 'Interpretation of the estimation results by logistic regression model'. Hiroshima University, Graduate School of Education, Department of Mathematics Education.

Czado, C. et al. (2013): Generalized linear models with applications. Technische Universität München, Lehrstuhl für Mathematische Statistik, Boltzmannstr. 3, 85747 Garching, Deutschland.

Czado, C. & Sikora, I. (2002): Quantifying overdispersion effects in count regression data. Working paper. Ludwig-Maximilians-Universität München, Institut für Statistik, Sonderforschungsbereich 386, paper 289.

Zhou, M.; Li, L.; Dunson, D. & Carin, L. (2012): Lognormal and Gamma Mixed Negative Binomial Regression. Working paper. Duke University, Durham NC 27708, USA.

Cameron, A.C. & Trivedi, P.K. (1999): Essentials of Count Data Regression. Working paper.

Engelhardt, M.E. (1994): Events in Time: Basic Analysis of Poisson Data. Working paper, Idaho National Engineering Laboratory.

Karaca-Mandic, P.; Norton, E. C. & Dowd, B. (2011): Interaction Terms in Nonlinear Models. Health Services Research, Methods Corner.

Wood, S. (2006): Generalized Additive Models: An introduction with R.

Hastie, T. & Tibshirani, R. (1986): Generalized Additive Models. Statistical Science, Vol.1, No. 3, p. 297-318.

Czado, C. (2015): Statistical Modelling with Copulas. Lecture notes.

Czado, C. (2013): Vine copulas with applications. Lecture notes.

Dorey, M. & Joubert, P. (2015): Modelling Copulas: An Overview. The Staple Inn Actuarial Society.

Karlis, D. & Nikoloulopoulos, A. K. (2009): Modelling multivariate count data using copulas. *Communication in Statistics - Simulation and Computation*, Taylor & Francis: STM, Behavioural Science and Public Health Titles, pp. 172-187.

Parsa, R. A. & Klugman S. A. (2011): Copula regression. *Casualty actuarial society*, volume 5, issue 1.

Krämer, N. & Schepsmeier, U. (2011): Introduction to vine copulas. NIPS Workshop, Granada.

Gijbels, I.; Veraverbeke, N. & Omelka, M. (2009): Conditional copulas, association measures and their applications.

Dißmann, J.; Brechmann, E.C.; Czado, C. & Kurowicka, D. (2012): Selecting and estimating regular vine copulae and application to financial returns.

Nelsen, R. B. (2006): *An Introduction to Copulas* (Springer Series in Statistics). Springer-Verlag New York, Inc.

Morales-Nápoles, O.; Cooke, R.; Kurowicka, D. (2010): About the number of vines and regular vines on n nodes. Submitted for publication.

Chapter 7

Appendix

7.1 A1: Outsourced figures

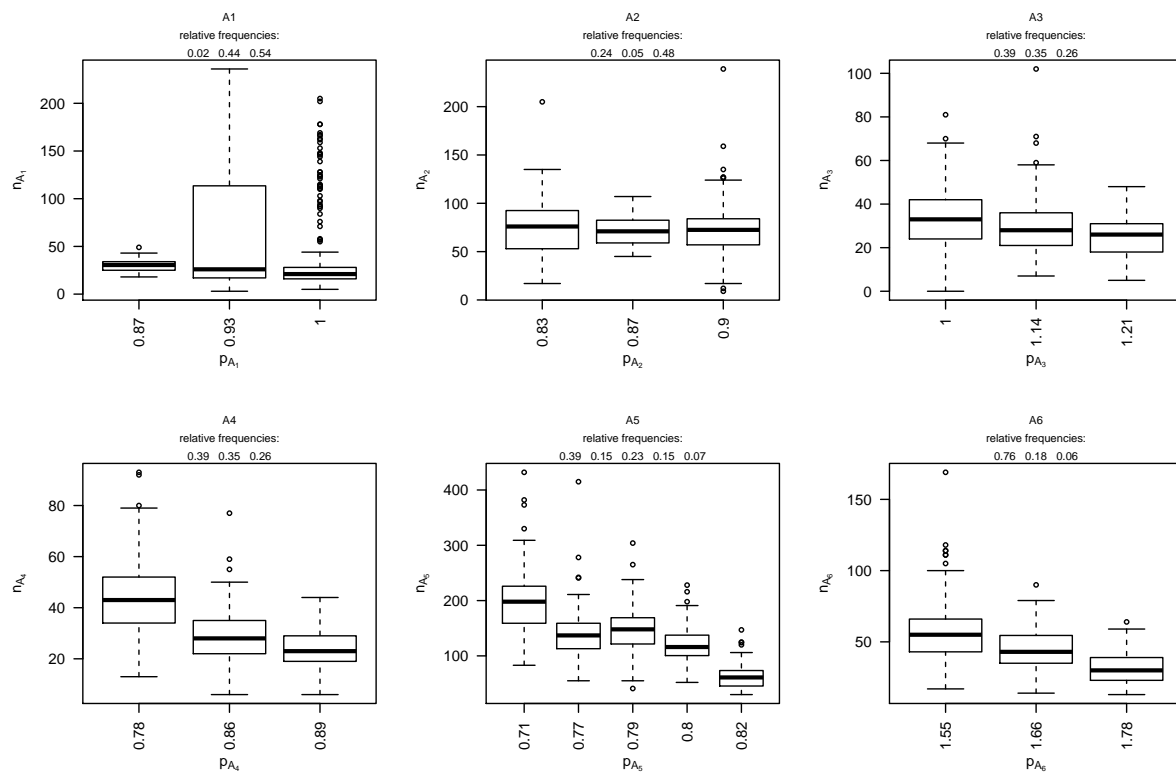


Figure 7.1: Article A_1 to article A_6 : Box plots of the quantity sold over the shop price for the individual articles. The data is given on daily basis.

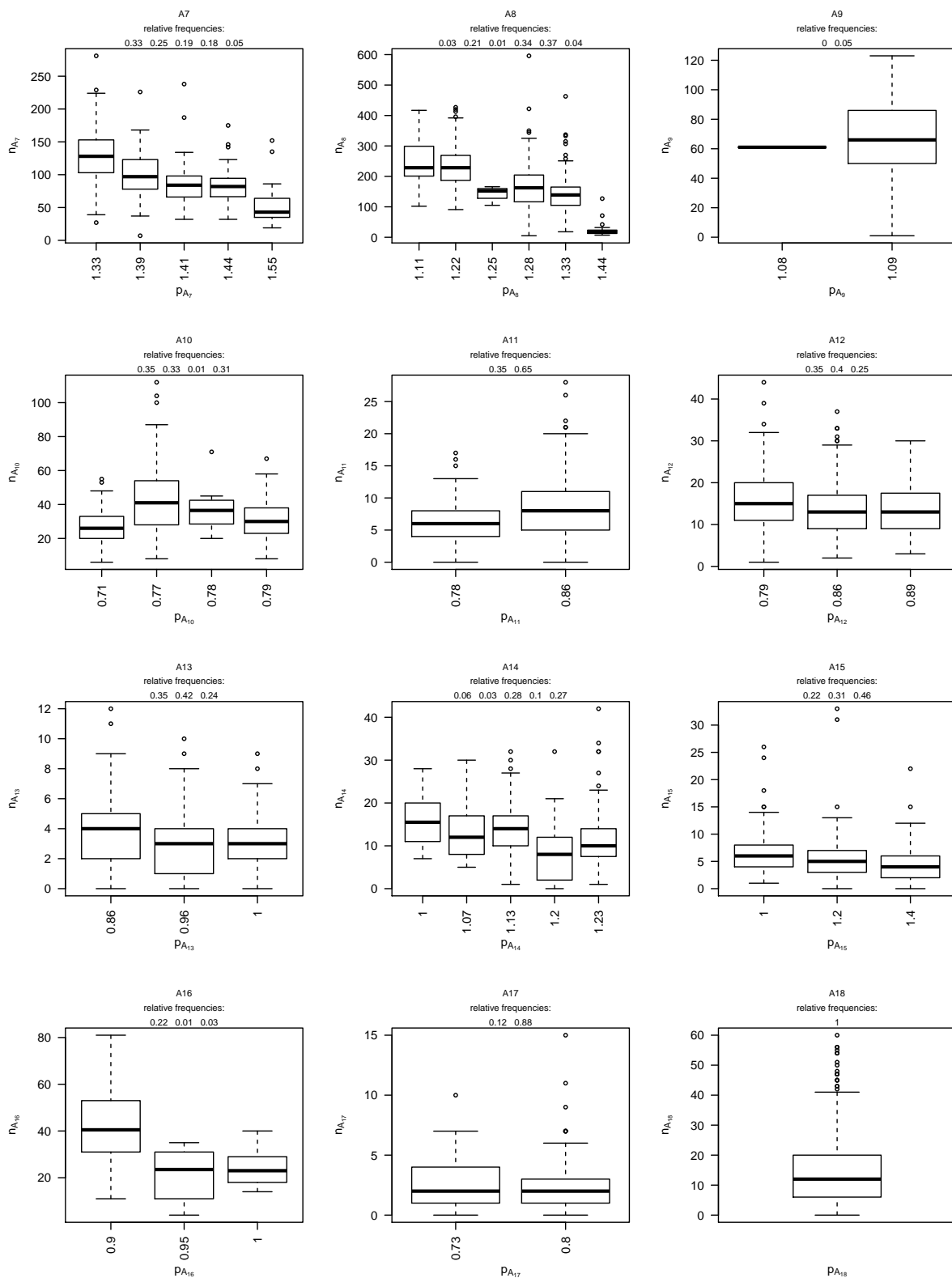


Figure 7.2: Article A_7 to article A_{18} : Box plots of the quantity sold over the shop price for the individual articles. The data is given on daily basis.

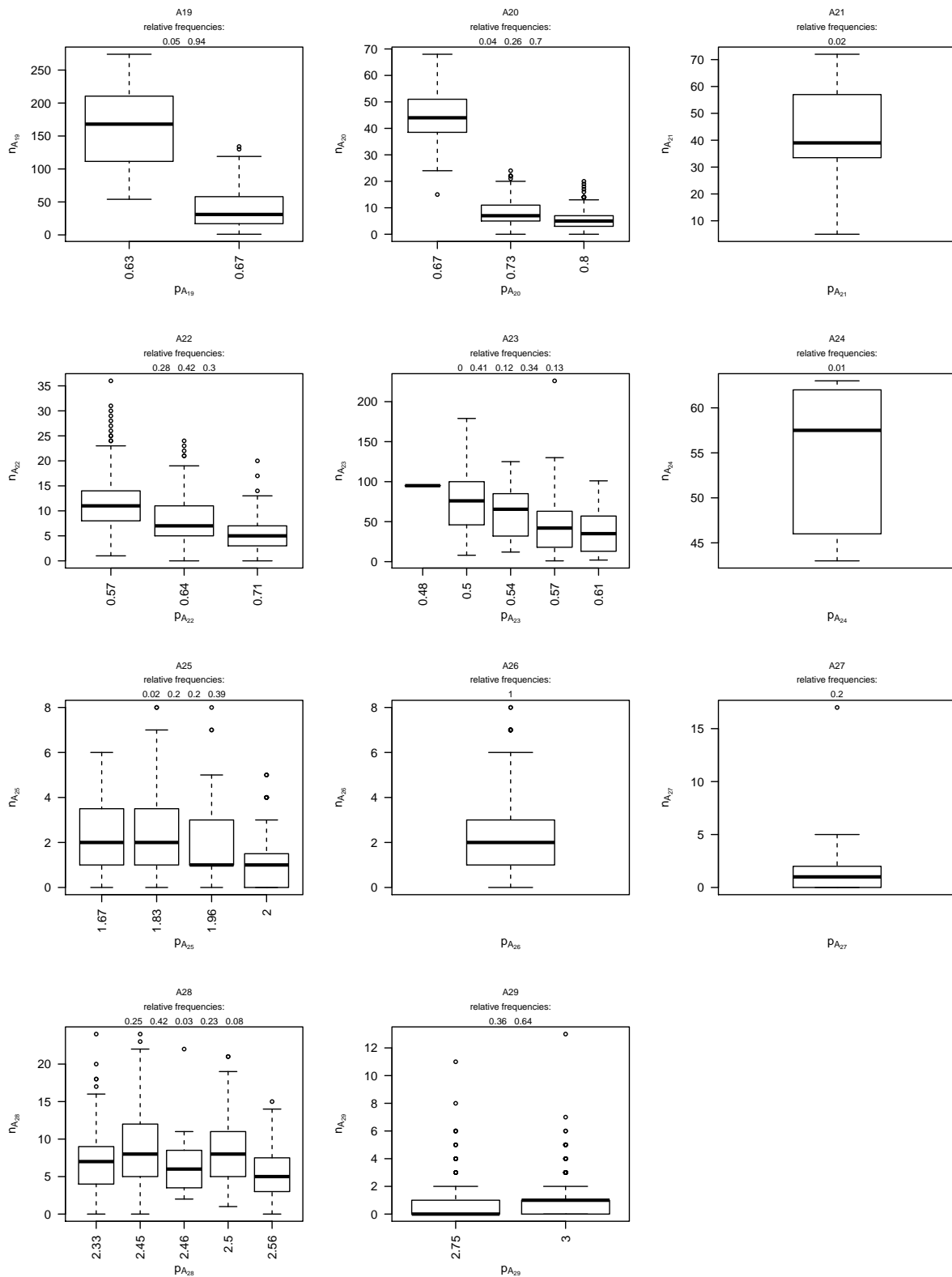


Figure 7.3: Article A_{19} to article A_{29} : Box plots of the quantity sold over the shop price for the individual articles. The data is given on daily basis.

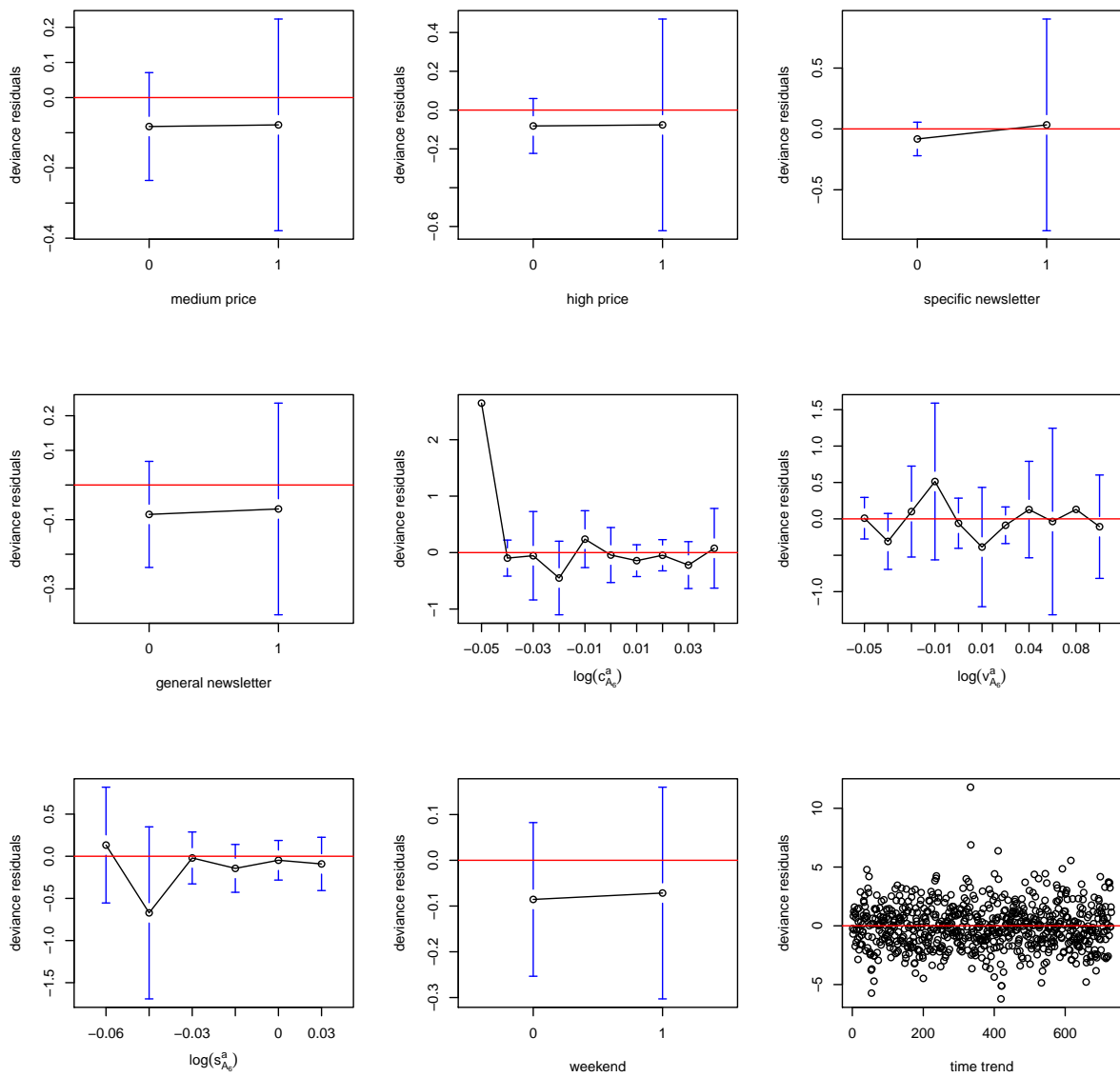


Figure 7.4: Residual analysis for the Poisson model for article A_6 including time effects and interactions (part 1): Plots showing the covariates against the deviance residuals.

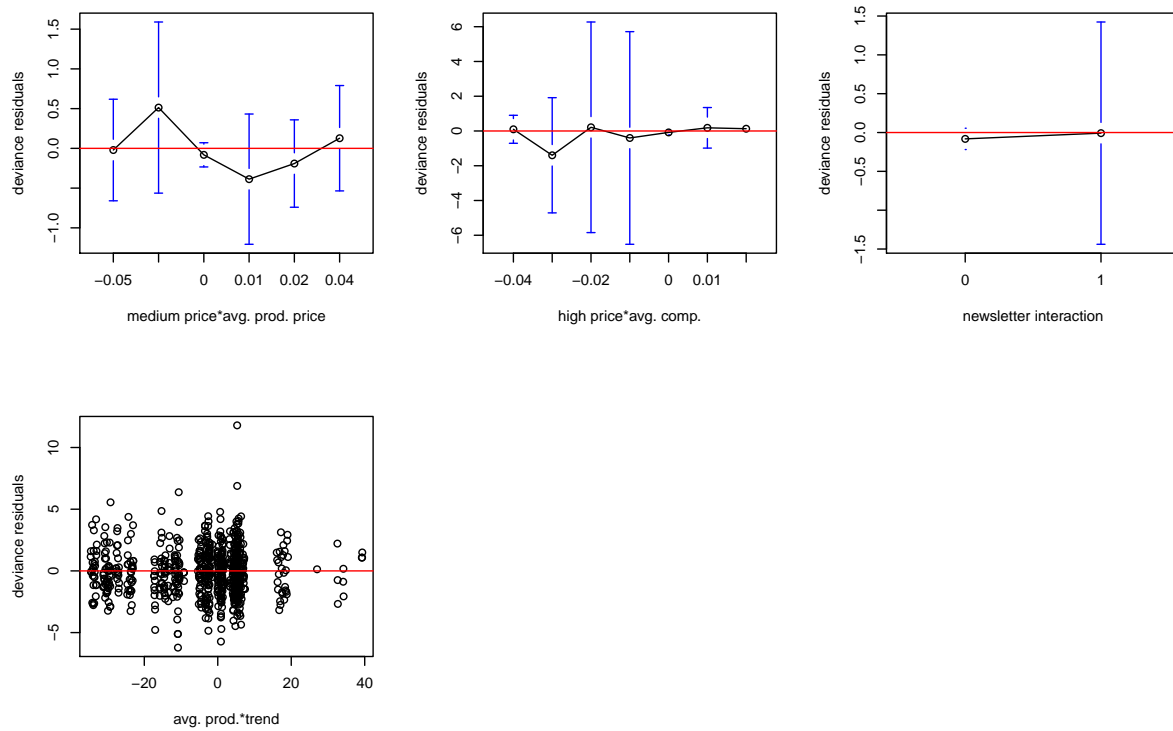


Figure 7.5: Residual analysis for the Poisson model for article A_6 including time effects and interactions (part 2): Plots showing the covariates against the deviance residuals.

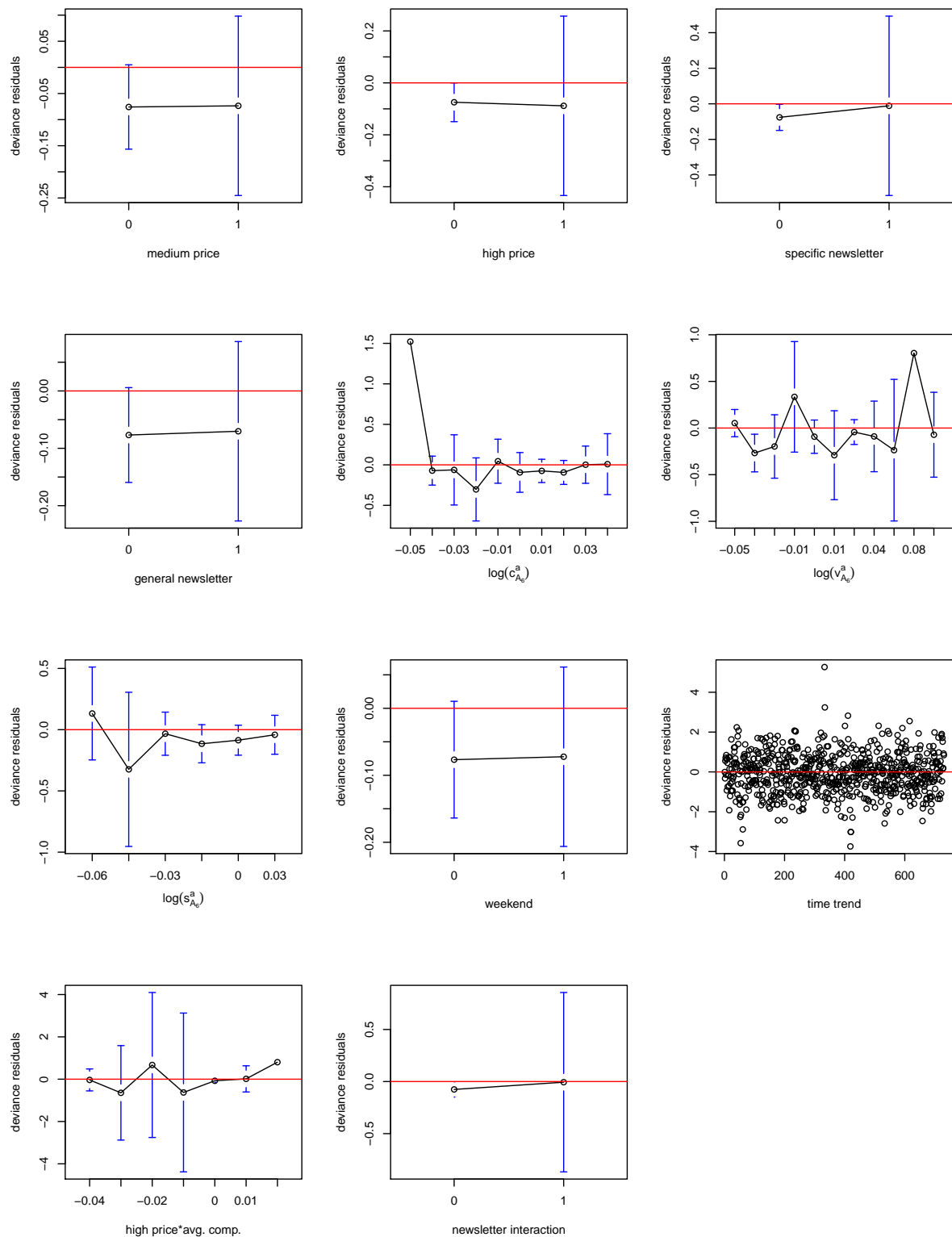


Figure 7.6: Residual analysis for the Negative Binomial model for article A_6 including time effects and interactions: Plots showing the covariates against the deviance residuals.

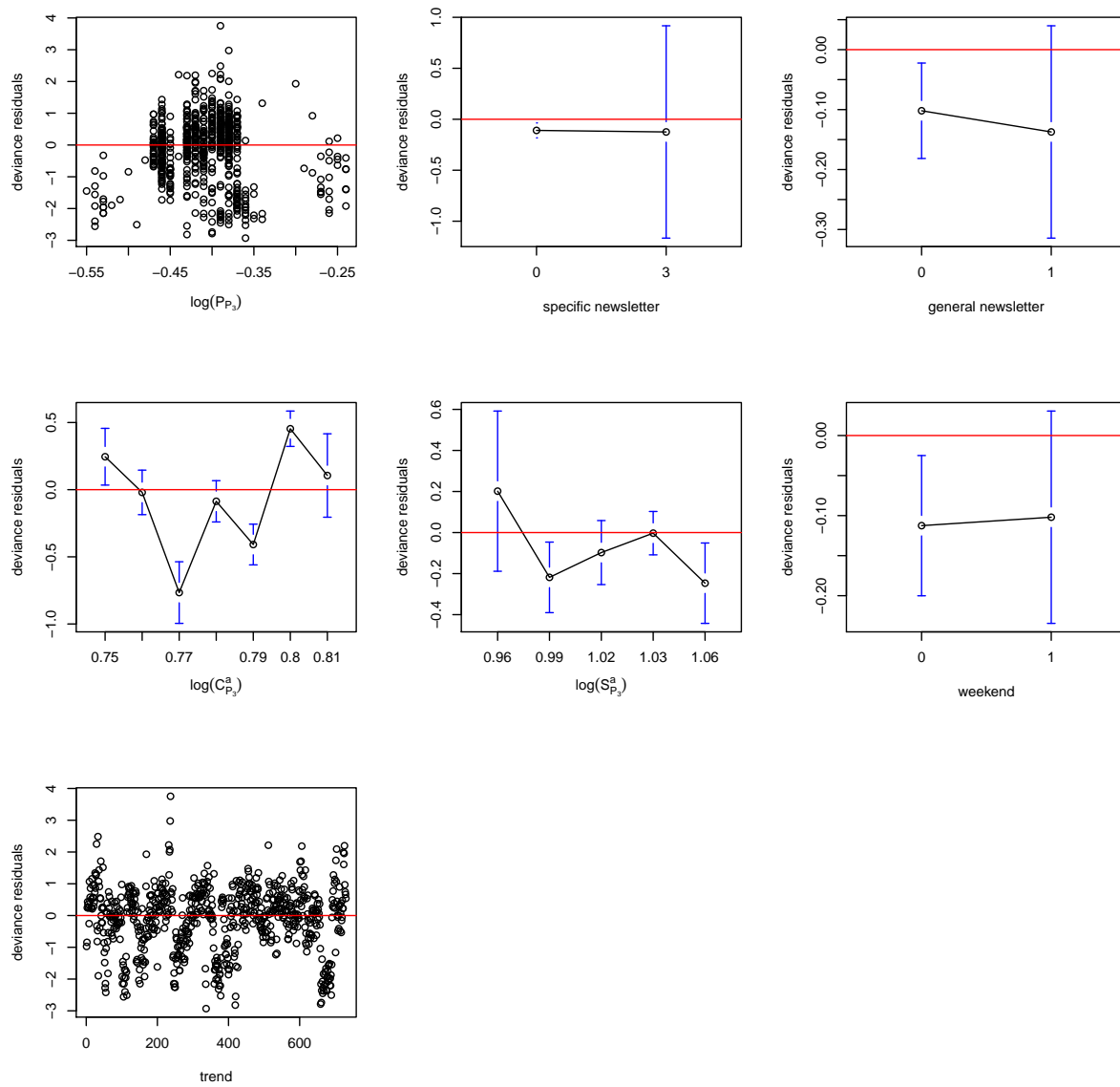


Figure 7.7: Residual analysis for the Negative Binomial model including time effects for product P_3 : Plots showing the covariates against the deviance residuals.

7.2 A2: Outsourced regression results

| | Estimate | Std. Error | t value | Pr(> t) |
|--------------------|-----------------------------------|------------|---------|----------|
| (Intercept) | 4.04 | 0.84 | 4.80 | 0.00 |
| price level 2 | -0.24 | 0.03 | -8.32 | 0.00 |
| price level 3 | -0.24 | 0.05 | -5.40 | 0.00 |
| price level 4 | -0.20 | 0.06 | -3.52 | 0.00 |
| price level 5 | -0.93 | 0.08 | -11.41 | 0.00 |
| spec_nl_A7 | 0.08 | 0.10 | 0.74 | 0.46 |
| general_nl | 0.12 | 0.02 | 4.80 | 0.00 |
| log(avg_comp_A7) | 1.19 | 0.51 | 2.33 | 0.02 |
| log(avg_prod_A7) | 1.84 | 0.60 | 3.09 | 0.00 |
| log(avg_subst_P3) | -0.82 | 1.23 | -0.66 | 0.51 |
| weekend | -0.32 | 0.02 | -13.96 | 0.00 |
| trend | 0.00 | 0.00 | 3.54 | 0.00 |
| Null deviance: | 1596.02 on 729 degrees of freedom | | | |
| Residual deviance: | 746.97 on 718 degrees of freedom | | | |
| AIC: | 6839.889 | | | |

Table 7.1: Regression results for article A_7 in the Negative Binomial regression model including time effects.

| | Estimate | Std. Error | t value | Pr(> t) |
|--------------------|-----------------------------------|------------|---------|----------|
| (Intercept) | 4.58 | 1.33 | 3.44 | 0.00 |
| price level 2 | -0.13 | 0.11 | -1.21 | 0.23 |
| price level 3 | -0.69 | 0.20 | -3.53 | 0.00 |
| price level 4 | -0.57 | 0.11 | -5.36 | 0.00 |
| price level 5 | -0.73 | 0.11 | -6.87 | 0.00 |
| price level 6 | -2.43 | 0.13 | -18.54 | 0.00 |
| spec_nl_A8 | 0.42 | 0.16 | 2.63 | 0.01 |
| general_nl | 0.14 | 0.04 | 3.63 | 0.00 |
| log(avg_comp_A8) | 4.40 | 0.82 | 5.34 | 0.00 |
| log(avg_prod_A8) | 0.05 | 0.09 | 0.57 | 0.57 |
| log(avg_subst_P3) | -0.28 | 1.92 | -0.14 | 0.89 |
| weekend | -0.33 | 0.04 | -9.26 | 0.00 |
| trend | -0.00 | 0.00 | -1.49 | 0.14 |
| Null deviance: | 1331.71 on 729 degrees of freedom | | | |
| Residual deviance: | 768.4 on 717 degrees of freedom | | | |
| AIC: | 8192.525 | | | |

Table 7.2: Regression results for article A_8 in the Negative Binomial regression model including time effects.

χ^2 -test for article A_6 in the Negative Binomial models including time effect with and without interactions:

| | Resid. Df | Resid. Dev | Df | Deviance | Rao | Pr(>Chi) |
|---|-----------|------------|------|----------|-------|----------|
| 1 | 720.00 | 738.98 | | | | |
| 2 | 718.00 | 737.48 | 2.00 | 1.49 | 15.01 | 0.00 |

Table 7.3: χ^2 -test for article A_6 : The Negative Binomial model including time effects with and without interactions.

7.3 A3: Outsourced details on copula models

R_mle_seq:

$$\begin{aligned}
 & \text{families} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 13 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 23 & 0 & 0 & 0 & 0 & 0 \\ 5 & 16 & 13 & 0 & 0 & 0 & 0 \\ 5 & 13 & 4 & 13 & 0 & 0 & 0 \\ 1 & 5 & 4 & 1 & 1 & 0 & 0 \\ 19 & 2 & 7 & 9 & 19 & 1 & 0 \end{bmatrix} \\
 \\
 & \text{par} = \begin{bmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0383 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ -0.0412 & -0.0629 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.6346 & 1.0177 & 0.0845 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.4049 & 0.0745 & 1.0708 & 0.0917 & 0.0000 & 0.0000 & 0.0000 \\ 0.1332 & 0.3976 & 1.1640 & 0.2258 & 0.1166 & 0.0000 & 0.0000 \\ 1.2515 & 0.1888 & 0.2441 & 1.2807 & 1.5653 & 0.1910 & 0.0000 \end{bmatrix} \\
 \\
 & \text{par2} = \begin{bmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.4287 & 12.3066 & 1.1331 & 0.4867 & 0.4243 & 0.0000 & 0.0000 \end{bmatrix}
 \end{aligned}$$

R_mle_ind:

$$\text{families} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 13 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & 1 & 0 & 0 \\ 2 & 19 & 7 & 9 & 19 & 1 & 0 \end{bmatrix}$$

$$\text{par} = \begin{bmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.6572 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 1.0740 & 0.0920 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.1332 & 1.1669 & 0.2258 & 0.1166 & 0.0000 & 0.0000 \\ 0.1887 & 1.2456 & 0.2398 & 1.2801 & 1.5639 & 0.1911 & 0.0000 \end{bmatrix}$$

$$\text{par2} = \begin{bmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 12.3048 & 0.4319 & 1.1276 & 0.4900 & 0.4213 & 0.0000 & 0.0000 \end{bmatrix}$$

C_mle_seq:

$$\text{families} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 13 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 33 & 0 & 0 & 0 & 0 & 0 \\ 1 & 14 & 16 & 0 & 0 & 0 & 0 \\ 5 & 13 & 4 & 3 & 0 & 0 & 0 \\ 5 & 5 & 1 & 5 & 5 & 0 & 0 \\ 19 & 1 & 19 & 19 & 9 & 4 & 0 \end{bmatrix}$$

$$\text{par} = \begin{bmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0426 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ -0.0333 & -0.0635 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.1052 & 1.0496 & 1.0149 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.6791 & 0.0941 & 1.0761 & 0.1622 & 0.0000 & 0.0000 & 0.0000 \\ 0.5913 & 0.5800 & 0.2258 & 1.4306 & 0.5512 & 0.0000 & 0.0000 \\ 1.2566 & 0.1910 & 1.5677 & 1.1877 & 1.3108 & 1.0933 & 0.0000 \end{bmatrix}$$

$$\text{par2} = \begin{bmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.4303 & 0.0000 & 0.4259 & 0.3795 & 0.4632 & 0.0000 & 0.0000 \end{bmatrix}$$

C_mle_ind:

$$\text{families} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 14 & 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 4 & 3 & 0 & 0 & 0 \\ 5 & 5 & 1 & 5 & 5 & 0 & 0 \\ 19 & 1 & 19 & 19 & 9 & 4 & 0 \end{bmatrix}$$

$$\text{par} = \begin{bmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.1052 & 1.0506 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.6757 & 0.0000 & 1.0732 & 0.1666 & 0.0000 & 0.0000 & 0.0000 \\ 0.5993 & 0.5739 & 0.2258 & 1.4292 & 0.5332 & 0.0000 & 0.0000 \\ 1.2526 & 0.1911 & 1.5664 & 1.1833 & 1.3090 & 1.0924 & 0.0000 \end{bmatrix}$$

$$\text{par2} = \begin{bmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.4341 & 0.0000 & 0.4266 & 0.3849 & 0.4639 & 0.0000 & 0.0000 \end{bmatrix}$$

D_mle_seq:

$$\text{families} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 13 & 0 & 5 & 0 & 0 & 0 & 0 \\ 13 & 0 & 13 & 5 & 0 & 0 & 0 \\ 13 & 0 & 9 & 1 & 1 & 0 & 0 \\ 0 & 2 & 7 & 19 & 19 & 19 & 0 \end{bmatrix}$$

$$\text{par} = \begin{bmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0359 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.4677 & -0.0393 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.1263 & 1.0380 & 0.6403 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.1542 & 0.0300 & 0.2182 & 0.4102 & 0.0000 & 0.0000 & 0.0000 \\ 0.1753 & 0.4003 & 1.1294 & 0.2881 & 0.1332 & 0.0000 & 0.0000 \\ -0.0067 & 0.1888 & 0.2121 & 1.2679 & 1.5715 & 1.2510 & 0.0000 \end{bmatrix}$$

$$\text{par2} = \begin{bmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.1239 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 12.3072 & 1.1422 & 0.4017 & 0.4109 & 0.4278 & 0.0000 \end{bmatrix}$$

D_mle_ind:

$$\text{families} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 13 & 0 & 5 & 0 & 0 & 0 & 0 \\ 13 & 0 & 13 & 5 & 0 & 0 & 0 \\ 13 & 0 & 9 & 1 & 1 & 0 & 0 \\ 0 & 2 & 7 & 19 & 19 & 19 & 0 \end{bmatrix}$$

$$\text{par} = \begin{bmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.4825 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.1293 & 0.0000 & 0.6490 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.1579 & 0.0000 & 0.2169 & 0.4120 & 0.0000 & 0.0000 & 0.0000 \\ 0.1763 & 0.0000 & 1.1294 & 0.2926 & 0.1376 & 0.0000 & 0.0000 \\ 0.0000 & 0.1939 & 0.2153 & 1.2684 & 1.5741 & 1.2498 & 0.0000 \end{bmatrix}$$

$$\text{par2} = \begin{bmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.1231 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 12.5760 & 1.1404 & 0.4022 & 0.4120 & 0.4281 & 0.0000 \end{bmatrix}$$

7.4 List of the variables of importance

| variable | description |
|---------------|--------------------------------------------------------------------------------------------------------------|
| w_{A_m} | Weight of article A_m . |
| b_{A_m} | Bundle size of article A_m . |
| $n_{A_m,t}$ | Quantity sold of article A_m at time t . |
| $N_{P_m,t}$ | Quantity sold of product P_m at time t . |
| $p_{A_m,t}$ | Shop price of article A_m per kilogram at time t . |
| $P_{P_m,t}$ | Avg. price of product P_m per half kilogram at time t . |
| $c_{A_m,t}^y$ | Price of article A_m per kilogram at time t at competitor y . |
| $c_{A_m,t}^m$ | Minimum competitors price for article A_m at time t . |
| $c_{A_m,t}^a$ | Average competitors basic price for article A_m at time t . |
| $C_{P_m,t}^a$ | Average competitors basic price for product P_m at time t . |
| $v_{s,t}^m$ | Minimal price for the remaining articles belonging to the same product per kilogram at time t . |
| $v_{s,t}^a$ | Average basic price for the remaining articles belonging to the same product per kilogram at time t . |
| $s_{s,t}^m$ | Minimal price for the substitute product per kilogram at time t . |
| $s_{s,t}^a$ | Average basic price for the substitute product per kilogram at time t . |
| $S_{s,t}^a$ | Average basic price for the substitute product per half kilogram at time t . |
| $ad_{A_m,t}$ | Boolean flag: TRUE if article A_m appeared in the newsletter at time t . |
| $ad_{P_m,t}$ | Integer value indicating how many product variants of product P_m appeared in the newsletter at time t . |
| ad^g | Boolean flag: TRUE if there was general advertisement active at time t . |