KTH Royal Institute of Technology

Master Thesis

Marketing Mix Modelling from multiple regression perspective

Ecaterina Mhitarean-Cuvsinov

supervised by
Assoc. Prof. Tatjana Pavlenko
Dr. Daniel Malmquist
Examiner: Tatjana Pavlenko

May 18, 2017

Abstract

The optimal allocation of the marketing budget has become a difficult issue that each company is facing. With the appearance of new marketing techniques, such as online advertising and social media advertising, the complexity of data has increased, making this problem even more challenging. Statistical tools for explanatory and predictive modelling have commonly been used to tackle the problem of budget allocation. Marketing Mix Modelling involves the use of a range of statistical methods which are suitable for modelling the variable of interest (in this thesis it is sales) in terms of advertising strategies and external variables, with the aim to construct an optimal combination of marketing strategies that would maximize the profit.

The purpose of this thesis is to investigate a number of regression-based model building strategies, with the focus on advanced regularization methods of linear regression, with the analysis of advantages and disadvantages of each method. Several crucial problems that modern marketing mix modelling is facing are discussed in the thesis. These include the choice of the most appropriate functional form that describes the relationship between the set of explanatory variables and the response, modelling the dynamical structure of marketing environment by choosing the optimal decays for each marketing advertising strategy, evaluating the seasonality effects and collinearity of marketing instruments.

To efficiently tackle two common challenges when dealing with marketing data, which are multicollinearity and selection of informative variables, regularization methods are exploited. In particular, the performance accuracy of ridge regression, the lasso, the naive elastic net and elastic net is compared using cross-validation approach for the selection of tuning parameters. Specific practical recommendations for modelling and analyzing Nepa marketing data are provided.

Sammanfattning

Att fördela marknadsföringsbudgeten optimalt är en svår uppgift som alla företag ställs inför. Med uppkomsten av nya marknadsföringstekniker, som reklam på nätet och sociala media, har komplexiteten av data ökat, vilket gör detta problem ännu mer utmanande. Statistiska verktyg för förklarande och prediktiv modellering har vanligtvis använts för att hantera problemet med budgetallokering. Marknadsföringsmix Modellering är en term som omfattar klassen av statistiska metoder som är lämpliga för modellering av den intressanta variabeln (i denna uppsats är det försäljning) när det gäller reklamstrategier och externa variaber, med målet att maximera vinsten genom att konstruera en optimal kombination av marknadsstrategier.

Syftet med denna uppsats är att konstruera ett antal modellbyggnadsstrategier, som även inkluderar avancerade regulariseringsmetoder för linjär regression, med en analys av fördelar och nackdelar för varje metod. Flera stora problem som den moderna marknadsföringsmix modellering står inför har beaktats, som till exempel: att välja en passande funktionsformel som bäst beskriver relationen mellan den oberoende variabeln och de beroende variablerna, att hantera marknadsföringens dynamiska omgivningar genom att välja det optimala förfallet hos varje marknadsföringsstrategi, utvärdera säsongsmässiga effekten och marknadsföringsverktygens kollinjäritet.

För att överkomma de två vanligaste problemen inom marknadsföringsekonometri, som är multikollinearitet och val av variabler, har regulariseringsmetoder använts. I synnerhet har prestationsnoggrannheten av ridge regression, lasso, naive elastic net och elastic net jämförts - för att ge specifika rekommendationer för Nepa data. Parametrarna för de regulariserade regressionsmetoderna har valts genom korsvalidering. Modellens resultat visar en hög nivå av förutsägelse noggrannhet. Skillnaden mellan nämnda metoder är inte signifikanta för det givna datasetet.

Acknowledgements

I would like to thank my supervisor Tatjana Pavlenko, Associate Professor at KTH, for her support and guidance during the master thesis. I would also like to thank my supervisor at Nepa AB, Dr. Daniel Malmquist, for his support and feedback throughout the process.

Contents

1 Introduction

This section provides a short introduction into the concept of Marketing Mix Modelling, as well a brief presentation of the company, and the purpose of the study.

1.1 Background

Marketing Mix Modelling is a term that is used to cover statistical methods which are suitable for explanatory and predictive statistical modelling of some variable of interest, for example company's sales or market shares. This thesis is focused on modelling sales as a factor of marketing instruments and environmental variables. In this case, the goal of Marketing Mix Modelling is to explain and predict sales from marketing instruments, while controlling for other factors that influence sales. Its main task is to decompose sales into base volume (which occurs due to such factors as seasonality and brand awareness) and incremental volume (which captures the weekly variation in sales driven by marketing activities). One of the most important Marketing Mix instruments is advertising, thus it is crucial to understand the impact of advertising expenditures on sales.

Model building in marketing started in the middle of twentieth century. Many studies have been conducted since then, which helped managers understand the marketing process. Appropriately constructed market response models helped the managers to determine the instruments that influence sales and take actions that would affect it. Applications show that model benefits include cost savings resulting from improvements in resource allocations. Many studies discuss and describe the model development process, provide a structure for model building and serve as a starting point for this thesis, including: Leeflang (2015), Leeflang (2000), Hanssens (2001), P.M Cain (2010).

This thesis attempts to develop a general model building strategy suitable for a high level of complexity of the data, to establish the most appropriate functional relationships and estimation methods for the Marketing Mix Modelling projects. This strategy will be used by Nepa for systematic analysis of the data collected. All the steps of this model building strategy are implemented in a user-friendly way and will be applied by Nepa for designing a marketing plan for its clients. As an illustration, the thesis analyses the relationship between marketing expenditures and sales on a dataset provided by Nepa. The data comes from a client of Nepa who is one of the largest electronics retailer in Sweden. This dataset contains model-specific weekly sales and marketing activities data, as well as environmental data, for two years. To overcome some of the problems that are commonly encountered when working with marketing data, advanced estimation methods such as ridge regression, the lasso and elastic net were employed to quantify the sales-marketing relationship and identify short and long-run effects of marketing on performance. The thesis describes each method and presents the output for each model introduced. Marketing dynamics were also considered in the model of sales structure, by optimizing the decays for each media variable.

1.2 Nepa

Nepa is an innovative research company founded in 2006 with the ambition to improve the efficiency of the research industry by moving from analog to digital methodologies. It is a company that went beyond phone interviews and mail surveys and pioneered a fully automated and online tracking solution. Today Nepa has more than 350 clients from all over the world and offices in Stockholm, Helsinki, Oslo, Copenhagen, London and Mumbai.

1.3 Purpose

The main purpose of the thesis is to elaborate a methodology that Nepa can use in Marketing Mix Modelling projects. A method is needed to find the optimal parameters to create a model with as good predictability and low multi collinearity as possible, with the following main areas of interest:

Parameter estimation

What type of decay should each media variable have? That is, how much effect does a certain amount invested in a media variable have one week later? This is known as the *carryover effect*, and it appears when some of the marketing strategies have impact not only in the current period, but also in the future periods.

Variable selection

It is important to efficiently tackle the problems of selecting the informative variables and evaluating the seasonality effect. How should be handled season & trend to avoid over- or underestimation of the effects of other variables? Estimating the impact of marketing instruments on sales becomes difficult when advertising activities coincide with seasonal peaks.

Regression modelling

It is often the case that several marketing investments take place at the same time. The collinearity caused makes the parameters estimated with ordinary least squares to be unreliable. The question then arises as to what estimation methods should be used to attain predictability and stability of the models. (Coping with multicollinearity, variable selection, etc.)

2 Theoretical Background

This section presents the mathematical background of the common challenges that marketing mix modellers are facing. It begins with the challenge of choosing the appropriate functional form, continues with the dynamic structure of marketing variables, and finally an approach to account for the effects of seasonality is described.

2.1 Methods of selecting functional forms of the model

An important part of the model building process is deciding upon the functional form that would reflect the most appropriate relationship between the variables. The most commonly applied functional form in Marketing Mix Modelling is the linear model. However, it is often the case that the nonlinear functional forms are used, since they take into account such properties as diminishing/increasing returns to scale and threshold effects. In this section the vector of the model parameters $\boldsymbol{\beta}$ as well as the vector of the disturbance term $\boldsymbol{\varepsilon}$ are named in the same way for different specifications, even though the parameters differ, depending on the functional form.

2.1.1 Linear and Multiplicative Models

Linear models assume constant returns to scale and have the following structure:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_K x_{Kt} + \varepsilon_t, \tag{2.1}$$

where, following the notations in [16]:
y_t = value of the dependent variable in period t ($t = 1, ..., T$, where T is the number of observations),
x_{kt} = value of independent variable k in period t, ($k = 1, ..., K$, where K is the number of covariates), and
$\beta_0, \beta_1, ..., \beta_K$ = model parameters.
ε_t = the (unobserved) value of the disturbance term.

A linear model is often tried first since the estimation of the coefficients and the interpretation of the results are easy. It shows a good predictive performance and a reasonable approximation to an underlying nonlinear function, but only on a limited range.

One drawback of the linearity assumption is that it implies constant returns to scale with respect to each of the covariates, meaning that an increase of one unit in x_{kt} leads to an increase of β_k units in y_t. However, the assumption of constant returns to scale is unrealistic in most real life marketing applications. Usually a sales response curve exhibits a non-constant behavior. One type of non-constant behavior is *diminishing returns to scale*, which happens when the response variable always increase with increases in the covariates, but each additional unit of x_{kt} brings less in y_t than the previous unit did ([15]). One of the functional forms that reflects this phenomenon is the *multiplicative power model* (again, following the notations in [16]):

$$y_t = \beta_0 x_{1t}^{\beta_1} \varepsilon_t, \quad x_{1t} \geq 0, \quad 0 < \beta_1 < 1 \tag{2.2}$$

Model 2.2 can be linearized by taking the logarithms of both sides:

$$\ln y_t = \ln \beta_0 + \beta_1 \ln x_{1t} + \ln \varepsilon_t, \quad x_{1t} \geq 0, \quad 0 < \beta_1 < 1 \tag{2.3}$$

Equation 2.3 is linear in the parameters $\beta_0^\star$, β_1, where $\beta_0^\star = \ln \beta_0$. This model is known as the *double-logarithmic* or the *log-log* model. The version of the multiplicative model that retains the highest-order interaction among the variables for K marketing instruments is:

$$y_t = \beta_0 x_{1t}^{\beta_1} x_{2t}^{\beta_2} \cdots x_{Kt}^{\beta_K} \varepsilon_t \tag{2.4}$$

or more compactly:

$$y_t = \beta_0 (\prod_{k=1}^{K} x_{kt}^{\beta_k}) \varepsilon_t \tag{2.5}$$

In this setting, if some of the variables are "dummies", the corresponding variables are used as exponents. Besides reflecting the non-constant behavior of the sales response function, another advantage of the multiplicative model over the linear model is that it allows for a specific form of interaction between the various instruments. Taking the first-order partial derivative of y_t with respect to any of the independent variables x_{kt}, the impact of a change in x_{kt} on y_t is a function of y_t itself, which means that it depends not only on the value of x_{kt} but on all the other variables as well:

$$\frac{\partial y_t}{\partial x_{kt}} = \beta_0 \beta_k x_{1t}^{\beta_1} x_{2t}^{\beta_2} \cdots x_{kt}^{\beta_{k-1}} \cdots x_{Kt}^{\beta_K} \tag{2.6}$$

When sales response function exhibits *increasing returns to scale*, the *exponential model* can be used:

$$y_t = \beta_0 e^{\beta_1 x_{1t}} \varepsilon_t \tag{2.7}$$

After taking the logarithms of both sides it becomes the *semi-logarithmic* also known as the *log-linear* model:

$$\ln y_t = \ln \beta_0 + \beta_1 x_{1t} + \ln \varepsilon_t \tag{2.8}$$

When the nonlinear model is log-log or log-linear, an adjustment to the forecasts of y_t is required, so that they remain unbiased ([15]). Considering the typical multiplicative specification 2.4, where $\ln \varepsilon_t$ is $N(0, \sigma^2)$, it can be shown that:

$$E[y_t] = \beta_0 x_{1t}^{\beta_1} x_{2t}^{\beta_2} \cdots x_{Kt}^{\beta_K} e^{1/2\sigma^2} \tag{2.9}$$

The forecasts should be calculated from the expression:

$$\hat{y}_t = \hat{\beta}_0 x_{1t}^{\hat{\beta}_1} x_{2t}^{\hat{\beta}_2} \cdots x_{Kt}^{\hat{\beta}_K} e^{1/2\hat{\sigma}^2} \tag{2.10}$$

where hats denote the ordinary least squares (OLS) estimates. A direct re-transformation would under-estimate the forecasts.

2.1.2 The Box-Cox transformation

One way to compare between the linear and the multiplicative specifications is the likelihood ratio test, using the Box-Cox transformation. It is based on the following transformation of the dependent variable:

$$\frac{y_t^\lambda - 1}{\lambda} = \beta_0 + \beta_1 x_{1t} + \ldots + \beta_K x_{Kt} + \varepsilon_t. \tag{2.11}$$

To choose the appropriate functional form, the likelihood ratio test of the model above can be used. The idea behind this method is to compute the likelihood ratio for different values of λ and choose the value that maximizes the MLE score. The specification is then chosen according to the value of λ reported. If $\lambda = 1$ then the specification is essentially linear. When λ approaches 0, equation 2.11 approaches the semi-logarithmic form, since:

$$\lim_{\lambda \to 0} \left(\frac{y_t^\lambda - 1}{\lambda} \right) = \ln y_t \tag{2.12}$$

2.2 Marketing Dynamics

Because of the evolving character of markets, the assumption that advertising expenditures have a current and immediate impact on sales rarely happens to be realistic. Most often is happens that parts of the media effects remain noticeable for several future periods. Thus, sales in some period t are affected by advertising expenditures in the same period t, but also by expenditures in previous periods $t-1, t-2, \ldots$. The influence of current marketing expenditures on sales in future periods is called the *carryover effect*. When the effect of a marketing variable is distributed over several time periods, sales in any period are a function of the current and previous marketing expenditures. In the case of just one explanatory variable the equation for sales is:

$$y_t = \beta_0 + \sum_{l=0}^{\infty} \beta_{l+1} x_{t-l} + \varepsilon_t, \tag{2.13}$$

where x_{t-l}, $l = 0, 1, \ldots$, are the lagged terms of the independent variable. The model 2.13 is called the *Infinite Distributed Lag (IDL) Model*. Assuming that all coefficients of the lagged terms of a covariate have the same sign, equation 2.23 can be rewritten as:

$$y_t = \beta_0 + \beta \sum_{l=0}^{\infty} \omega_l x_{t-l} + \varepsilon_t. \tag{2.14}$$

Equation 2.14 is called the *Geometric Lag Model*, where

$$\omega_l \geq 0 \quad \textit{and} \quad \sum_{l=0}^{\infty} \omega_l = 1. \tag{2.15}$$

The omegas can be regarded as probabilities of a discrete-time distribution. As mentioned in [15], the Geometric Distributed Lag (GL) Model is the most commonly used distributed-lag model in marketing. The maximum impact of marketing expenditures on sales is registered instantaneously, then the influence declines geometrically to zero. The impact of any past expenditure in subsequent periods will be a constant fraction of its immediate impact. This constant fraction is called the *retention rate*. If the retention rate is λ the geometric distribution gives:

$$\omega_l = (1 - \lambda)\lambda^l \quad l = 0, 1, 2 \cdots \tag{2.16}$$

where $0 < \lambda < 1$. The specification of the sales response function becomes:

$$y_t = \beta_0 + \beta(1 - \lambda) \sum_{l=0}^{\infty} \lambda^l x_{t-l} + \varepsilon_t, \tag{2.17}$$

or

$$y_t = \beta_0 + \beta_1 x_t + \beta_1 \lambda x_{t-1} + \beta_1 \lambda^2 x_{t-2} + \ldots + \beta_1 \lambda^l x_{t-l} + \ldots + \varepsilon_t, \tag{2.18}$$

where $\beta_1 = \beta(1-\lambda)$. The direct short-term effect of marketing effort is $\beta_1 = \beta(1-\lambda)$, while the *retention rate* λ measures how much of the advertising effect in one period is retained in the next. The implied long-term effect is $\beta = \beta_1/(1-\lambda)$. This model is also approximately equivalent to the *Simple Decay-Effect Model* (Broadbent (1979)):

$$y_t = \beta_0 + \beta_1 a_t + \varepsilon_t, \tag{2.19}$$

where $a_t = f(x_t)$ is the adstock function at time t, x_t is the value of the advertising variable at time t and λ is the decay or lag weight parameter:

$$a_t = f(x_t) = x_t + \lambda a_{t-1}, \quad t = 2, \dots, n \tag{2.20}$$

Recursively substituting and expanding the equation for the adstock function becomes:

$$a_t = x_t + \lambda x_{t-1} + \lambda^2 x_{t-2} + \ldots + \lambda^n x_{t-n}, \tag{2.21}$$

Since $0 < \lambda < 1$, $\lambda \to 0$ as $n \to \infty$. Moving on to the case with K explanatory variables $x_1, \dots, x_K$, each with different retention rates $\lambda_1, \dots, \lambda_K$, the model becomes:

$$y_t = \beta_0 + \beta_1 a_{1t} + \beta_2 a_{2t} + \cdots + \beta_K a_{Kt} + \varepsilon_t \tag{2.22}$$

where

$$a_{it} = f(x_{it}) = x_{it} + \lambda_i a_{it-1}, \quad i = 1, \dots, K \tag{2.23}$$

To estimate the marketing variables coefficients, as well as retention rate values, non-linear least squares can be used. The algorithm is described more detailed in section 3.2. First the adstock at time t is defined for each marketing instrument, as in equation 2.23. The estimated sales are then:

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 a_{1t} + \hat{\beta}_2 a_{2t} + \ldots + \hat{\beta}_K a_{Kt} \tag{2.24}$$

Finally, the optimization problem is:

$$\begin{array}{ll} \text{minimize} & \sum_{t=1}^{T} (y_t - \hat{y}_t)^2 \\ \text{subject to} & 0 \le \lambda_i < 1, \ i = 1, \dots, K. \end{array}$$

For the semi-logarithmic and double-logarithmic models the equation for the predicted sales becomes 2.25 and 2.26 respectively:

$$\ln \hat{y}_t = \hat{\beta}_0^\star + \hat{\beta}_1 a_{1t} + \hat{\beta}_2 a_{2t} + \ldots + \hat{\beta}_K a_{Kt} \tag{2.25}$$

and

$$\ln \hat{y}_t = \hat{\beta}_0^\star + \hat{\beta}_1 \ln a_{1t} + \hat{\beta}_2 \ln a_{2t} + \ldots + \hat{\beta}_K \ln a_{Kt} \tag{2.26}$$

And the optimization problem is:

$$\begin{array}{ll} \text{minimize} & \sum_{t=1}^{T} (\ln y_t - \ln \hat{y}_t)^2 \\ \text{subject to} & 0 \le \lambda_i < 1, \ i = 1, \dots, K. \end{array}$$

2.3 Modelling trend and seasonality

In this section the "classical decomposition" is considered:

$$y_t = m_t + \delta_{it} + \varepsilon_t, \tag{2.27}$$

where: m_t is a slowly changing function (the "trend component");
δ_{it} is a function with known period d (the "seasonal component");
ε_t is a stationary time series.
In trying to explain sales behavior, a linear trend variable ($m_t = 1, 2, \cdots, T$ for $t = 1, 2, \cdots, T$) could be introduced into the sales response function to capture the time-dependent nature of sales growth.

If a variable follows a systematic pattern within the year, it is said to exhibit seasonality. To deal with seasonality, s *dummy variables* could be introduced in the model to express s seasons in the following way:

$$\delta_{it} = \begin{cases} 1, & \text{if } t \text{ is the } i\text{'th period} \\ 0, & \text{otherwise} \end{cases} \qquad i = 1, \cdots, s \quad t = 1, \cdots, T \tag{2.28}$$

These "dummy" variables for seasons and "time" variable m_t for trend could be incorporated into the linear model 2.1:

$$y_t = \beta_0 + m_t + \delta_{1t} + \ldots + \delta_{st} + \beta_1 x_{1t} + \beta_2 x_{2t} + \ldots + \beta_K x_{Kt} + \varepsilon_t, \tag{2.29}$$

and also into the multiplicative models, for example into the log-log model 2.4:

$$y_t = \beta_0 e^{(m_t + \delta_{1t} + \ldots + \delta_{st})} x_{1t}^{\beta_1} x_{2t}^{\beta_2} \ldots x_{Kt}^{\beta_K} \varepsilon_t \tag{2.30}$$

Equation 2.30 is non-linear. For the purposes of estimation, the model is converted into an additive form by taking natural logarithms thus:

$$\ln y_t = \ln \beta_0 + m_t + \delta_{1t} + \ldots + \delta_{st} + \beta_1 \ln x_{1t} + \ldots + \beta_K \ln x_{Kt} + \ln \varepsilon_t \tag{2.31}$$

Taking into account the dynamic structure, equation 2.31 becomes:

$$\ln y_t = \ln \beta_0 + m_t + \delta_{1t} + \ldots + \delta_{st} + \beta_1 \ln a_{1t} + \ldots + \beta_K \ln a_{Kt} + \ln \varepsilon_t \tag{2.32}$$

where $a_{1t}, \ldots, a_{Kt}$ are the adstock variables defined in section 2.2. Equation 2.32 is no longer linear, and was estimated with non-linear least squares, using the Levenberg-Marquardt algorithm, described in section 3.2.

3 Estimation

Once the appropriate functional form is decided, the parameters of the marketing model must be estimated. A description of the estimation methods for the model parameters is provided in this Chapter.

3.1 Ordinary Least Squares

Let us consider the linear model 2.1:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_K x_{Kt} + \varepsilon_t, \quad t = 1, \ldots, T \tag{3.1}$$

where the notations are defined in section 2.1.1. Equation 2.1 can be rewritten in the matrix form:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{K1} \\ 1 & x_{12} & x_{22} & \cdots & x_{K2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1T} & x_{2T} & \cdots & x_{KT} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{pmatrix} \tag{3.2}$$

or:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{3.3}$$

The OLS estimates of the parameters $\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 & \hat{\beta}_1 & \cdots & \hat{\beta}_K \end{pmatrix}^\intercal$ in 3.1 are the values which minimize the *Residual Sum of Squares* (RSS):

$$RSS = \sum_{t=1}^{T} (y_t - \hat{y}_t)^2 = \sum_{t=1}^{T} (y_t - \hat{\beta}_0 - \sum_{k=1}^{K} \hat{\beta}_k x_{kt})^2 \tag{3.4}$$

Following the notations in [8], the *total sum of squares* is defined as: $SS_{tot} = \sum_{t=1}^{T} (y_t - \bar{y}_t)^2$. With RSS and SS_{tot} defined above, the following relationship holds: $SS_{tot} = SS_{reg} + RSS$, where SS_{reg} is the *regression sum of squares*: $SS_{reg} = \sum_{t=1}^{T} (\hat{y}_t - \bar{y}_t)^2$. It is easy to show that the coefficient estimates $\hat{\boldsymbol{\beta}}$ obtained by minimizing the quantity above are:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{y}. \tag{3.5}$$

Assuming that $Cov(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$, the covariance matrix of $\hat{\boldsymbol{\beta}}$ is then $Cov(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\intercal \mathbf{X})^{-1}$, estimated as:

$$\widehat{Cov}(\hat{\boldsymbol{\beta}}) = \frac{RSS}{T - K - 1} (\mathbf{X}^\intercal \mathbf{X})^{-1}$$

An $F_\alpha(1, T - K - 1)$-statistic for the hypothesis $\beta_k = 0$ is calculated as:

$$F = \left(\frac{\hat{\beta}_k}{SE(\hat{\beta}_k)} \right)^2 .$$

where the standard error $SE(\hat{\beta}_k)$ for any $k = 1, \ldots, K$ is the square root of the corresponding diagonal element of $\widehat{Cov}(\hat{\boldsymbol{\beta}})$.

3.2 Non-linear Least Squares

To model the dynamic structure with several explanatory variables, the Levenberg-Marquardt algorithm (LMA) was used. As described in [7], the LMA interpolates between the Gauss-Newton algorithm (GNA) and the method of gradient descent. In the current setting, following the notations defined in the previous sections, the problem is defined in the following way: given a number T of observations of independent and dependent variables, $(\mathbf{x}_t, y_t)$, where $\mathbf{x}_t$ is a vector of length K, containing the K variable measurements corresponding to the observation of the dependent variable y_t, the objective is to optimize K parameters $\boldsymbol{\beta} = \begin{pmatrix}\beta_0 & \beta_1 & \dots & \beta_K\end{pmatrix}^\intercal$ of the model curve $f(\mathbf{X}, \boldsymbol{\beta})$ such that the sum of the squares of the deviations

$$S(\boldsymbol{\beta}) = \sum_{t=1}^{T} [y_t - f(\mathbf{x}_t, \boldsymbol{\beta})]^2 \tag{3.6}$$

is minimized.

3.2.1 The Gradient Descent Method

The idea behind the steepest descent method is that it updates parameter estimates in the direction opposite to the gradient of the objective function. The gradient of S with respect to $\boldsymbol{\beta}$ is

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 2(\mathbf{y} - f(\mathbf{X}, \boldsymbol{\beta}))^\intercal \frac{\partial}{\partial \boldsymbol{\beta}}(\mathbf{y} - f(\mathbf{X}, \boldsymbol{\beta})) = -2(\mathbf{y} - f(\mathbf{X}, \boldsymbol{\beta}))^\intercal \frac{\partial f(\mathbf{X}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2(\mathbf{y} - f(\mathbf{X}, \boldsymbol{\beta}))^\intercal J \tag{3.7}$$

where the Jacobian matrix

$$J = \frac{\partial f(\mathbf{X}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

represents the the change of $f(\mathbf{X}, \boldsymbol{\beta})$ to variation in the parameters $\boldsymbol{\beta}$. In each iteration step, the parameter increment δ that moves the parameters $\boldsymbol{\beta}$ in the direction of steepest descent is given by

$$\delta_{gd} = \alpha J^\intercal (\mathbf{y} - f(\mathbf{X}, \boldsymbol{\beta})) \tag{3.8}$$

The positive scalar α determines the length of the step in the steepest-descent direction.

3.2.2 The Gauss-Newton Method

The Gauss-Newton method assumes that the objective function is approximately quadratic in the parameters near the optimal solution ([7]). The parameter increment δ is found by approximating the functions $f(\mathbf{x}_t, \boldsymbol{\beta} + \delta)$ by their linearizations

$$f(\mathbf{x}_t, \boldsymbol{\beta} + \delta) \approx f(\mathbf{x}_t, \boldsymbol{\beta}) + J_t \delta \tag{3.9}$$

where

$$J_t = \frac{\partial f(\mathbf{x}_t, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

The above first-order approximation of $f(\mathbf{x}_t, \boldsymbol{\beta} + \delta)$ gives

$$S(\boldsymbol{\beta} + \delta) \approx (\mathbf{y} - f(\mathbf{X}, \boldsymbol{\beta}))^\intercal (\mathbf{y} - f(\mathbf{X}, \boldsymbol{\beta})) - 2(\mathbf{y} - f(\mathbf{X}, \boldsymbol{\beta}))^\intercal J \delta + \delta^\intercal J^\intercal J \delta \tag{3.10}$$

Taking the derivative of $S(\boldsymbol{\beta} + \delta)$ with respect to δ and setting the result to zero gives:

$$(J^\intercal J)\delta_{gn} = J^\intercal [\mathbf{y} - f(\mathbf{X}, \boldsymbol{\beta})] \tag{3.11}$$

3.2.3 The Levenberg-Marquardt Method

The Levenberg-Marquardt algorithm interpolates between the Gauss-Newton method and the method of gradient descent.

$$(J^\intercal J + \lambda \mathbf{I})\delta_{lm} = J^\intercal[\mathbf{y} - f(\mathbf{X}, \boldsymbol{\beta})] \tag{3.12}$$

Small values of the damping parameter λ result in a Gauss-Newton update and large values of λ result in a gradient descent update. In each step the parameter λ is iteratively adjusted, that is λ is increased $S(\boldsymbol{\beta} + \delta) > S(\boldsymbol{\beta})$, and is decreased otherwise. To avoid slow convergence in the direction of small gradient, Marquardt provided the insight that the values of λ should be scaled to the values of $J^\intercal J$ ([7]):

$$[J^\intercal J + \lambda \ \mathrm{diag}((J^\intercal J)]\delta_{lm} = J^\intercal[\mathbf{y} - f(\mathbf{X}, \boldsymbol{\beta})]. \tag{3.13}$$

4 Validation and Testing

The process of validation and testing of the model begins with testing model's statistical assumptions. This part is called specification error analysis (section 4.2). The next step is to test the regression results. This involves tests of significance described in section 4.1.

4.1 Methods of Model Assessment

In this section it is assumed that there are no specification errors. Linear regression assumes that the disturbances are normally distributed: $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I})$, thus $\hat{\boldsymbol{\beta}} \sim N(\beta, \sigma^2 (\mathbf{X}^\intercal \mathbf{X})^{-1})$. A test statistic for the hypothesis that all of the β's are all equal to zero:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_K = 0 \quad \text{vs} \quad H_1 : \text{at least one } \beta_i \neq 0$$

is:

$$F = \frac{SS_{reg}/K}{RSS/(T-K-1)}$$

which has an approximate $F(K, T-K-1)$ distribution under the null. To determine the amount of variation "explained" by the covariates, one looks at a descriptive statistic R^2, called the *coefficient of determination* or *goodness of fit*.

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{RSS}{SS_{tot}} \tag{4.1}$$

There is also an adjusted R^2 that considers an adjustment for degrees of freedom:

$$\bar{R}^2 = 1 - \frac{T-1}{T-K-1} \frac{RSS}{SS_{tot}} \tag{4.2}$$

To determine which covariates are contributing to the fit, one has to examine each covariate separately. The test statistic for the null hypothesis that a coefficient is zero must be calculated as explained in section 3.1. A common test to determine which covariate should enter the regression is the Akaike Information Criterion test:

$$AIC = T \ln(RSS) + 2K. \tag{4.3}$$

The model with the lowest AIC is prefered, since it minimizes the information loss ([13]).

4.2 Specification Error Analysis

To obtain point estimates of the coefficients and perform statistical inferences based on those point estimates (for example: tests of significance, confidence intervals) the following assumptions must be satisfied:

- $\mathrm{E}[\varepsilon_t] = 0$ for all t;
- $\mathrm{Var}[\varepsilon_t] = \sigma^2$ for all t;
- $\mathrm{Cov}[\varepsilon_t, \varepsilon_{t'}] = 0$ for $t \neq t'$;

- ε_t is normally distributed.
- The matrix $\mathbf{X}$ has full rank, thus $\mathbf{X}^\intercal\mathbf{X}$ is non-singular.

Table 1 based on [16] is a part of model building strategy from the perspective of violation of assumptions. It presents a short summary of reasons, remedies, and ways to detect possible violations of each assumption. The Table is adapted to the given problem, and the methods applied in this thesis.

4.2.1 Nonzero expectation of the residuals

The violation of the assumption that the residuals are normally distributed could be a sign of incorrect functional form, or omitted variables. If the assumed functional form is incorrect, a plot of the residuals $e_t = y_t - \hat{y}_t$, $t = 1, \ldots, T$ against each predictor should show a systematic pattern in the residual values. However, this plot will not show that a variable has been omitted ([16]).

One way to test the possibility of an omitted variable is to add additional variables in the original regression. Ramsey recommends to add powers of the fitted response as additional terms. The test is based on the estimation of the following model:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_K x_{Kt} + \gamma_1 \hat{y}_t^2 + \gamma_2 \hat{y}_t^3 + \cdots + \gamma_m \hat{y}_t^{m+1} + \varepsilon_t^\star, \tag{4.4}$$

The null hypothesis is that the tested model is the true model, meaning that the additional variables should not have an impact on the dependent variable in the model 4.4. Since the regression equation 4.4 is useful to detect both omitted variables and nonlinearities, it is difficult to determine the exact cause of the test failure ([12]).

4.2.2 Heteroscedasticity

The second assumption is that all residuals ε_t have the same standard deviation. In this case standard errors and F-statistics will be computed from the estimated covariance matrix. However, if the model has heteroscedastic residuals and is misspecified as homoscedastic, then the estimators of the standard errors of the coefficient estimates will be wrong, and therefore the F-tests will be invalid ([13]). Thus, OLS estimates of the coefficients of the model will still be unbiased but not efficient. One solution is to use another estimation method, like generalized least squares or the method of maximum likelihood ([4]). In many cases the critical remedy is to use an appropriately adjusted formula for the variances and covariances of the parameter estimates.

Heteroscedascticity can be detected using the Breusch-Pagan test ([21]). The idea of this test to run a regression of the squared residuals on the covariates from the original equation:

$$\hat{\varepsilon}_t^2 = \delta_0 + \delta_1 x_{1t} + \delta_2 x_{2t} + \ldots + \delta_K x_{Kt} + \nu_t \tag{4.5}$$

where ν_t is a disturbance term with mean zero given the x_{kt} , $k = 1, \ldots, K$. The null hypothesis of homoscedasticity is:

$$H_0 : \delta_1 = \delta_2 = \ldots = \delta_K x_{Kt} = 0 \tag{4.6}$$

The F-statistic of the test is calculated in the following way:

$$F = \frac{R^2_{\hat{\varepsilon}^2}/K}{(1 - R^2_{\hat{\varepsilon}^2})/(T - K - 1)} \tag{4.7}$$

where $R^2_{\hat{\varepsilon}^2}$ is the R-squared from the regression 4.5. This F-statistic has (approximately) an $F_{K,T-K-1}$ distribution under the null.

Table 1: Violations of the assumptions about the disturbance term: reasons, consequences, tests and remedies (based on [16])

Violated Assumption	**Possible Reasons**	**Consequence**	**Detection**	**Remedy**
1. $\mathrm{E}[\varepsilon_t] \neq 0$	• Incorrect functional form(s) • Omitted variable(s)	• Biased parameter estimate	• Plot residual against each predictor variable • RESET test • Box-Cox transformation	• Modify the model specification in terms of functional form • Add relevant predictors
2. $\mathrm{Var}[\varepsilon_t] \neq \sigma^2$	• Error proportional to variance of the predictor	• Inefficient parameter estimate	• Plot residual against each predictor variable • Breusch-Pagan test	• Modify the specification • Use heteroscedasticity consistent estimation (e.g. GLS)
3. $\mathrm{Cov}[\varepsilon_t, \varepsilon_{t'}] \neq 0$	• See 1.	• See 2.	• Plot residuals against time • Durbin Watson test	• See 1.
4. Nonnormal errors	• See 1.	• p-values cannot be trusted	• Inspect the distribution of residuals • Normality tests	• See 1. • Box-Cox transformation
5. Multicollinearity	• Relations between predictor variables	• Unreliable parameters	• Inspect the correlation matrix of the predictor variables • Some VIF ≥ 5 • Condition number of the matrix $(\mathbf{X}^\intercal\mathbf{X})^{-1}$ is greater than 30	• Apply other estimation methods • Eliminate predictor variable(s)

4.2.3 Correlated Disturbances

Instead of assuming a model where the disturbance term is 0, let us consider the following simple linear additive relation for T time-series observations:

$$y_t = \beta_0 + \beta_1 x_t + u_t, \quad t = 1, \ldots, T \tag{4.8}$$

where the disturbances are correlated in the following way:

$$u_t = \rho u_{t-1} + \varepsilon_t, \quad |\rho| < 1 \tag{4.9}$$

and:

$$\mathrm{E}[\varepsilon_t] = 0, \quad Cov(\varepsilon_t, \varepsilon_{t'}) = 0, \quad t \neq t'$$

In 4.8 the error terms $u_1, u_2, \ldots, u_T$ follow a first-order AutoRegressive (AR) process with autocorrelation parameter ρ. In this case, the parameter estimates are no longer efficient, although still unbiased, and the usual F-statistic cannot be trusted.

A plot of the residuals against time could help to detect a violation of the assumption of uncorrelated disturbances. Another way is to use the test developed by Durbin and Watson ([9], [10]), based on the variance of the difference between two successive disturbances:

$$\mathrm{E}[(u_t - u_{t-1})^2] = \mathrm{E}[u_t^2] + \mathrm{E}[u_{t-1}^2] - 2\mathrm{E}[u_t u_{t-1}] \tag{4.10}$$

The Durbin-Watson test statistic varies between zero and four and is calculated in the following way:

$$\mathrm{DW} = \frac{\sum_{t=2}^{T} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{T} \hat{u}^2} \tag{4.11}$$

Values of the DW test below (above) 2 are associated with positive (negative) autocorrelation. The test statistic is used as described in ([16]):

1. Tests for positive autocorrelation:
   (a) If $\mathrm{DW} < d_L$, there is positive autocorrelation;
   (b) If $d_L < \mathrm{DW} < d_U$, the result is inconclusive;
   (c) If $\mathrm{DW} > d_U$, there is no positive autocorrelation.
2. Tests for negative autocorrelation:
   (a) If $\mathrm{DW} > 4 - d_L$, there is negative autocorrelation;
   (b) If $4 - d_U < \mathrm{DW} < 4 - d_L$, the result is inconclusive;
   (c) If $\mathrm{DW} < 4 - d_U$, there is no negative autocorrelation.

where the lower and upper bounds d_L and d_U depend on significance level and sample size.

When first-order autocorrelation is detected, a two-step estimation procedure is required. The first step involves obtaining an estimate of ρ by means of OLS estimation. The second step requires this estimate of ρ to be used in an estimated generalized least squares (GLS) regression ([15]). However, according to [16], this remedy should only be a last resort option.

4.2.4 Nonnormal Errors

The assumption of normally distributes is required for hypothesis testing and confidence intervals to be applicable. When this assumption is violated the standard statistical tests cannot be performed although the least squares estimates of the parameters remain unbiased as well as consistent.

The normality of the errors can be examined through the residuals. For this, an inspection of the distribution function of the residuals as well as normality tests might be used. In this thesis, the Lilliefors test was employed to assess the normality assumption of the residuals.

4.2.5 Multicollinearity

In linear model, matrix of observations $\mathbf{X}$ is assumed to have full rank, otherwise $\mathbf{X}^\intercal\mathbf{X}$ will be singular, and the OLS estimates cannot be uniquely determined. When the number of covariates is smaller than the number of observations, $\mathbf{X}^\intercal\mathbf{X}$ will be singular when some of the columns of $\mathbf{X}$ are collinear. In practice however, more often the problem that arrises is imperfect multicollinearity, when a column of $\mathbf{X}$ is nearly a linear combination of the other columns. In this case, $(\mathbf{X}^\intercal\mathbf{X})^{-1}$ exists, but its elements will be large, thus the standard errors of one or more of the regression coefficients become very large, and the point estimates of the those coefficients will be imprecise. This problem is encountered in marketing area, since data often show high degrees of correlation between media variables. Some methods of diagnosing multicollinearity in a given dataset include:

1. Examining the correlation matrix of the predictor variables. A correlation coefficient close to 1 or -1 is considered as an indicator of positive or negative collinearity.

2. Looking at the *Variance Inflation Factor* (VIF). This measure is based on the regression of each individual predictor variable on all the other predictor variables. VIF is computed as $1/(1 - R_k^2)$, where R_k^2 values result from the regressions above. There is no exact value of VIF that would be considered as a sign of multicollinearity. Some analysts argue that a VIF value greater than 5 is a signal that collinearity is a problem.

3. Comparing results for F-test and t-tests. Multicollinearity may be regarded as acute if the F-statistic shows significance and none of the t statistics for the slope coefficients is significant.

4. Looking at the condition number of the matrix $(\mathbf{X}^\intercal\mathbf{X})^{-1}$, which is the ratio of its largest eigenvalue to its smallest eigenvalue $\lambda_{Max}/\lambda_{Min}$. The data matrix should first be normalized so that each column has equal length - usually unit length. A rule of thumb is that a condition index of 15 indicates some degree of collinearity, and a condition index above 30 is an indicator of severe multicollinearity.

The main solution proposed to Nepa for solving multicollinearity was to apply regularization methods specifically developed for the cases with severe multicollinearity.

5 Linear Model Selection and Regularization

In this section, there are discussed distinct ways that might improve the linear (or linearizable) models, using variable selection or alternative estimation methods. To avoid confusion with the systematic literature, for this Chapter n states for the number of observations and p is the number covariates.

5.1 Subset selection

There are several methods for selecting subsets of predictors. These include best subset and stepwise model selection procedures.

Best subset selection is performed by fitting a least squared regression for each possible combination of the total number of p predictors. Then all the resulting 2^p models are examined, with the goal of identifying the one that is *best*. Since there are 2^p models to be examined, the number of all possible models that must be considered grows rapidly as p increases. For computational reasons, ***stepwise methods*** come as alternatives for best subset selection, which include: forward stepwise, backward stepwise selection as well as hybrid approaches.

Forward stepwise selection begins with no predictors, and then gradually adds predictors to the model by adding at each step the variable that gives the greatest *additional* improvement to the fit (in terms of RSS or R^2). On the other hand, backward stepwise selection starts with the full model containing all p predictors, and iteratively removes the least useful one. Finally, it is possible to combine both forward and backward stepwise selection, in which variables are added to the model sequentially, but at each step the method may also remove any variables that no longer provide an improvement in the model fit. The *best* model can be selected according to various criteria, such as: C_p, AIC, BIC or Adjusted R^2, where AIC and Adjusted R^2 are defined in the previous chapters, and C_p and BIC are computed using the following equations:

$$C_p = \frac{1}{n}(RSS + 2p\hat{\sigma}^2) \tag{5.1}$$

and

$$BIC = \frac{1}{n}(RSS + \ln(n)p\hat{\sigma}^2). \tag{5.2}$$

Both C_p and BIC will return a small value for the models with a low test error, so similar to AIC, the models with lowest C_p and BIC are prefered.

Validation and Cross-Validation

The validation set approach involves randomly dividing the observations into a training set and a test set. The coefficients are estimated with the training set, and then used to predict the dependent values in the test set. In k-fold cross-validation approach the data is divided into k groups. Each of these groups is subsequently used as a test set, while the rest of the observations is used as a training set. Obtaining k estimates of the test error, the k-fold cross-validation estimate is computed by averaging these values. The advantage of cross-validation relative to the methods mentioned above is that it provides a direct estimate of the test error, and helps to avoid overfitting. To obtain accurate estimates of the test error, only the training observations should be used.

5.2 Shrinkage Methods

As an alternative to the subset selection methods described in section 5.1 above, techniques that shrink coefficient estimates towards zero can be used. These include *ridge regression*, the *lasso*, and the *elastic net*. To understand better these techniques, first the concept of bias-variance trade off is introduced.

5.2.1 The Bias-Variance Trade-Off

Recall that *mean squared error* (MSE) is estimated as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}[(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\intercal(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})]$$

where $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1})$. Note that the covariates are fixed, only the responses are random. The expected test MSE, for a given vector $\mathbf{x}_0$ of length K that contains new measurements, can be decomposed in the following way:

$$\text{MSE}_0 = \text{Var}[\mathbf{x}_0^\intercal\hat{\boldsymbol{\beta}}] + \text{Bias}^2(\mathbf{x}_0^\intercal\hat{\boldsymbol{\beta}}) + \text{Var}[\boldsymbol{\varepsilon}]$$

where $\text{Bias}(\mathbf{x}_0^\intercal\hat{\boldsymbol{\beta}}) = \text{E}[\mathbf{x}_0^\intercal\hat{\boldsymbol{\beta}}] - \mathbf{x}_0^\intercal\boldsymbol{\beta}$. In practice, some bias might be accepted for a reduction in variance of the coefficient estimates. This can be achieved by employing regularized regression methods described in the following sections.

5.2.2 Ridge Regression

The balance between bias and variance may be achieved by placing constraints on the estimated coefficients $\boldsymbol{\beta}$. Instead of minimizing the sum of the squared residuals RSS defined above, the ridge regression coefficient estimates are found by minimizing the following value:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ji})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{5.3}$$

where $\lambda \geq 0$ is a *tuning parameter* to be determined, and the term $\lambda\sum_{j=1}^{p}\beta_j^2$ is called a *shrinkage penalty*. The result is the ridge regression estimator:

$$\hat{\boldsymbol{\beta}}_{ridge}(\lambda) = ((\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbf{I}_r)^{-1}\mathbf{X}^\intercal\mathbf{y} = \mathbf{W}(\lambda)\hat{\boldsymbol{\beta}}_{OLS},$$

where $\mathbf{W}(\lambda) = ((\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbf{I}_r)^{-1}\mathbf{X}^\intercal\mathbf{X}$. For each value of λ ridge regression will produce a set of coefficient estimates. When $\lambda = 0$ ridge estimates will be equal to the least squares estimates. As $\lambda \to \infty$ the ridge regression coefficient estimates will approach zero. The intercept remains simply the mean value of the response.

Because ridge coefficients change substantially when multiplying a covariate by a constant, ridge regression should be applied using standardized predictors ([6]):

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}} \tag{5.4}$$

Standardizing the predictors also makes it possible to compare estimated coefficients with each other.

5.2.3 The Lasso

Ridge regression will shrink all the coefficient estimates towards zero, but it will not set any of them exactly equal to zero, which might be a drawback if the purpose of the model is also variable selection. An alternative method called the *lasso* overcomes this disadvantage. The lasso coefficients are the values that minimize the quantity:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ji})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j| \tag{5.5}$$

Like ridge regression, the lasso shrinks the coefficient estimates towards zero. However, in the case of the lasso penalty, some of the coefficient estimates will be exactly zero when the tuning parameter λ is sufficiently large. Hence the lasso performs also variable selection, which makes the interpretation of the model much easier.

5.2.4 Comparing Ridge regression and The Lasso

One can show that the lasso and ridge regression coefficient estimates solve the problems:

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \left\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2\right\} \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j| \leq s \tag{5.6}$$

and

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \left\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2\right\} \quad \text{subject to} \quad \sum_{j=1}^{p}\beta_j^2 \leq s, \tag{5.7}$$

respectively.

Note that the restriction $\sum_{j=1}^{p}\beta_j^2 \leq s$ on $\boldsymbol{\beta}$ is a hypersphere centered at the origin with bounded squared radius s, where the value of s determines the value of k. Figure 1 (taken from [6]) shows the restrictions for the lasso and ridge regression for the two-parameter case.

Choosing among the regularization methods is not trivial. Which model produces better prediction accuracy depends on the dataset used. Since lasso assumes that several coefficients are in fact equal to zero, it will perform better when some of the predictors are not related to the response. In the case when all coefficients substantially differ from zero, ridge regression is expected to outperform the lasso. Since the number of coefficient related to the response is never known, cross-validation approach can be used to determine the best method for each dataset. For a deeper discussion see ([6]) on how to select the regularization approach.

5.2.5 Selecting the Tuning Parameter

Cross-validation approach tackles the problem of selecting the appropriate tuning parameter λ in the following way: the cross-validation error is computed over a grid of λ values, then the tuning parameter value is selected for which the cross-validation error is smallest. Finally, the model is refitted using all the available observations and the selected value of the tuning parameter.

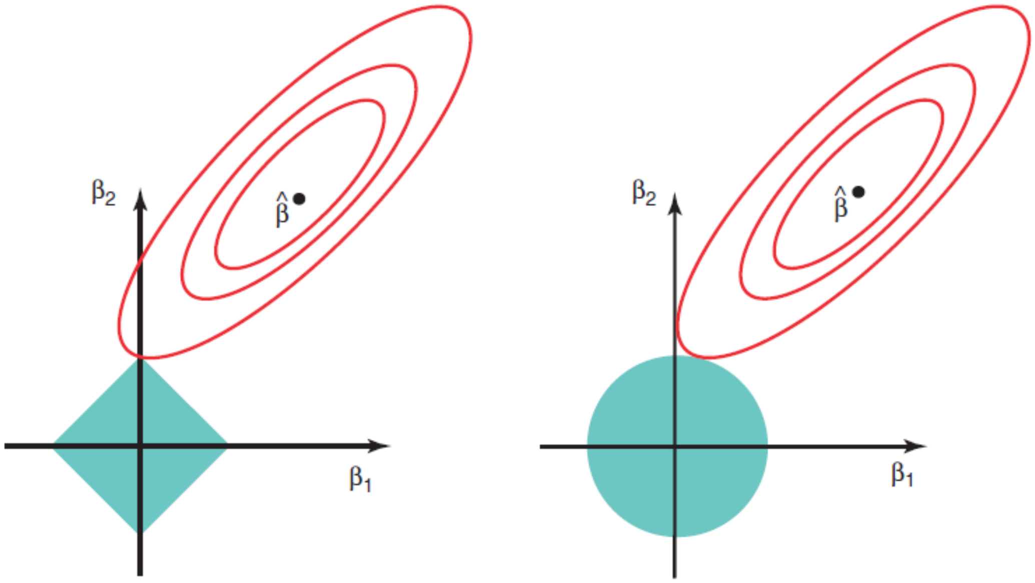

Figure 1: (taken from [1]). Contours of the error and constraint functions for the lasso (left) and ridge regression (right). The solid blue areas are the regions $|\beta_1|+|\beta_2|\leq s$ and $\beta_1^2+\beta_2^2\leq s$, respectively, while the red ellipses are the countors of the RSS.

5.2.6 Naive Elastic Net

Considering the model above with $\mathbf{y}$ being the vector of response variable and $\mathbf{X} = \begin{pmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_p\end{pmatrix}$ being the matrix of predictors, the naive elastic net estimator $\hat{\boldsymbol{\beta}}$ is the one that minimizes the quantity:

$$L(\lambda_1,\lambda_2,\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ji})^2+\lambda_1\sum_{j=1}^{p}|\beta_j|+\lambda_2\sum_{j=1}^{p}\beta_j^2 \tag{5.8}$$

for any λ_1 and λ_2. Similar to ridge regression and the lasso, this procedure can be viewed as a penalized least squares. If α is defined as $\alpha = \lambda_2/(\lambda_1+\lambda_2)$, then solving for $\boldsymbol{\beta}$ in equation 5.8 is equivalent to the optimization problem:

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \left\{\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij})^2\right\} \quad \text{subject to} \quad (1-\alpha)\sum_{j=1}^{p}|\beta_j|+\alpha\sum_{j=1}^{p}\beta_j^2\leq s, \tag{5.9}$$

The function $(1-\alpha)\sum_{j=1}^{p}|\beta_j|+\alpha\sum_{j=1}^{p}\beta_j^2\leq s$ is called the elastic net penalty. When $\alpha=1$, the naive elastic net becomes simple ridge regression. For all $\alpha\in[0,1)$, the elastic net penalty function is singular (without first derivative) at 0 and it is strictly convex for all $\alpha>0$. Note that the lasso penalty ($\alpha=0$) is convex but not strictly convex. The two-dimensional contours of the penalty function for ridge, lasso and naive elastic net are given in Figure 2 (taken from [26]). In the article [27] Hui Zou and Trevor Hastie develope a method to solve the naive elastic net problem efficiently.

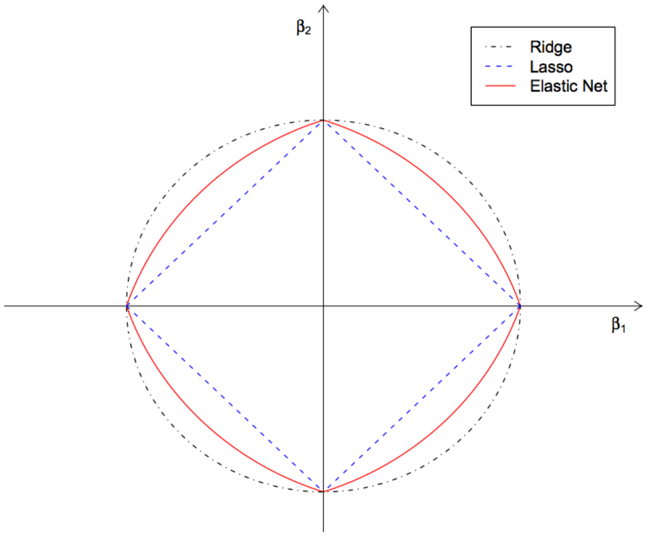

Figure 2: (taken from [26]). Two-dimensional contour plots of the ridge, the lasso, and $\alpha = 0.5$ elastic net penalties.

It turns out that minimizing equation 5.8 is equivalent to a lasso-type optimization problem. This fact implies that the naive elastic net also enjoys the computational advantage of the lasso. The next Lemma is a result from paper [27].

Lemma 1. *Given a dataset* $(\mathbf{y}, \mathbf{X})$ *and* (λ_1, λ_2)*, an artificial dataset* $(\mathbf{y}^\star, \mathbf{X}^\star)$ *is defined by:*

$$\mathbf{X}^\star_{(n+p)*p} = (1+\lambda_2)^{-1/2} \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I} \end{pmatrix}, \quad \mathbf{y}^\star_{(n+p)} = \begin{pmatrix} \mathbf{y} \\ 0 \end{pmatrix} \tag{5.10}$$

Let $\gamma = \lambda_1/\sqrt{1+\lambda_2}$ *and* $\boldsymbol{\beta}^\star = \sqrt{1+\lambda_2}\boldsymbol{\beta}$*. Then the naive elastic net criterion can be given as:*

$$L(\gamma, \boldsymbol{\beta}) = L(\gamma, \boldsymbol{\beta}^\star) = \sum_{t=1}^{n} (y_i^\star - \beta_0^\star - \sum_{j=1}^{p} \beta_j^\star x_{ji}^\star)^2 + \gamma \sum_{j=1}^{p} |\beta_j^\star|. \tag{5.11}$$

Let $\hat{\boldsymbol{\beta}}^\star = (\hat{\beta}_1^\star, \ldots, \hat{\beta}_p^\star)^\intercal$ *be the vector that minimizes the quantity above. Then*

$$\hat{\boldsymbol{\beta}} = \frac{1}{\sqrt{1+\lambda_2}} \hat{\boldsymbol{\beta}}^\star \tag{5.12}$$

Note that the sample size in the augmented problem is $n + p$ and $\mathbf{X}^\star$ has rank p, which means that the naive elastic net can potentially select all p predictors in all situations. Lemma 1 also shows that the naive elastic net can perform an automatic variable selection in a fashion similar to the lasso.

5.2.7 Elastic Net

The studies in [27] show that one drawback of the naive elastic net is that it performs best when it is very close to either ridge regression or the lasso. The estimation of the coefficients implies a double shrinkage procedure, which causes an increase in bias. In their paper, Hui Zou and Trevor Hastie propose a scaling of the naive elastic net coefficients which keeps the advantage of variable selection property, avoiding the undesirable double shrinkage. Following the notations in section 5.2.6, the naive elastic net solves a lasso-type problem:

$$\hat{\boldsymbol{\beta}}^{\star} = \arg\min_{\boldsymbol{\beta}^{\star}} |\mathbf{y}^{\star} - \mathbf{X}^{\star}\boldsymbol{\beta}^{\star}|^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}}|\boldsymbol{\beta}^{\star}|_1 \tag{5.13}$$

where $(\mathbf{y}^{\star}, \mathbf{X}^{\star})$ is the augmented data defined in 5.10, and (λ_1, λ_2) is the penalty parameter. The elastic net corrected estimates are defined by:

$$\hat{\boldsymbol{\beta}}_{enet} = \sqrt{1+\lambda_2}\hat{\boldsymbol{\beta}}^{\star} \tag{5.14}$$

Recall that $\hat{\boldsymbol{\beta}}_{naive\ enet} = (1/\sqrt{1+\lambda_2})\hat{\boldsymbol{\beta}}^{\star}$, thus:

$$\hat{\boldsymbol{\beta}}_{enet} = (1+\lambda_2)\hat{\boldsymbol{\beta}}_{naive\ enet} \tag{5.15}$$

In elastic net one could choose the type of the tuning parameters as (λ_2, s) where $s \in [0;1]$ is the fraction of the l_1-norm. The tuning parameters were chosen using a two-dimensional tenfold cross-validation method, following the procedure suggested in [27]: first a (relatively small) grid of values for λ_2 is picked, then the other tuning parameter is selected by cross-validation. The value of λ_2 is chosen such as to give the smallest CV error.

6 Results

In this section, the results for the models specified in the previous sections are presented. It starts from the linear and multiplicative models, described in section 2.1.1. The linear, log-linear and log-log functional forms are estimated. Next, the retention rates for the functional form chosen are estimated using non-linear least squares. Finally, the results for the modern approaches are illustrated and compared.

6.1 Choosing among functional forms for Marketing Mix Modelling

To estimate the parameters for the equations formulated in the above sections, the following data provided by Nepa was used:

y_t	= value of sales in week t,
TV_t	= Advertising expenditures for Television in week t,
DR_t	= Advertising expenditures for ads coming with the mail in week t,
$\mathsf{DR.POSTEN}_t$	= Investments in ads coming with the mail, but in a slightly different format,
$\mathsf{OUTDOOR}_t$	= Investments in ads put up outdoor, i.e. at bus stops in week t,
RADIO_t	= RADIO advertising expenditures in week t,
PRINT_t	= PRINT advertising expenditures in week t,
$\mathsf{SOCIALMEDIA}_t$	= SOCIALMEDIA advertising expenditures in week t,
Rain_t	= Rain quantity in week t,
sal_t	= Dummy variable, indicating whether it was salary week,
$\mathsf{HOLIDAY}_t$	= Dummy variable, indicating whether it was a Holiday in week t,

With the variables defined above the linear model 2.1 becomes:

$$\begin{aligned} y_t = {} & \alpha_0 + \alpha_1 \mathsf{TV}_t + \alpha_2 \mathsf{DR}_t + \alpha_3 \mathsf{DR.POSTEN}_t + \alpha_4 \mathsf{OUTDOOR}_t + \\ & + \alpha_5 \mathsf{RADIO}_t + \alpha_6 \mathsf{PRINT}_t + \alpha_7 \mathsf{SOCIALMEDIA}_t + \\ & + \alpha_8 \mathsf{Rain}_t + \alpha_9 \mathsf{sal}_t + \alpha_{10} \mathsf{HOLIDAY}_t + \varepsilon_t^{(1)}, \end{aligned} \tag{6.1}$$

The exponential model 2.7 model takes the form:

$$\begin{aligned} y_t = {} & \beta_0 e^{\beta_1 \mathsf{TV}_t + \beta_2 \mathsf{DR}_t + \beta_3 \mathsf{DR.POSTEN}_t + \beta_4 \mathsf{OUTDOOR}_t + \beta_5 \mathsf{RADIO}_t} \cdot \\ & \cdot e^{\beta_6 \mathsf{PRINT}_t + \beta_7 \mathsf{SOCIALMEDIA}_t} e^{\beta_8 \mathsf{Rain}_t + \beta_9 \mathsf{sal}_t + \beta_{10} \mathsf{HOLIDAY}_t} \varepsilon_t^{(2)} \end{aligned} \tag{6.2}$$

which after taking the natural logarithm on both sides becomes:

$$\begin{aligned} \ln(y_t) = {} & \ln(\beta_0) + \beta_1 \mathsf{TV}_t + \beta_2 \mathsf{DR}_t + \beta_3 \mathsf{DR.POSTEN}_t + \beta_4 \mathsf{OUTDOOR}_t + \\ & + \beta_5 \mathsf{RADIO}_t + \beta_6 \mathsf{PRINT}_t + \beta_7 \mathsf{SOCIALMEDIA}_t + \\ & + \beta_8 \mathsf{Rain}_t + \beta_9 \mathsf{sal}_t + \beta_{10} \mathsf{HOLIDAY}_t + \ln(\varepsilon_t^{(2)}), \end{aligned} \tag{6.3}$$

Finally, the multiplicative model 2.4 becomes:

$$\begin{aligned} y_t = {} & \gamma_0 \mathsf{TV}_t^{\beta_1} \mathsf{DR}_t^{\gamma_2} \mathsf{DR.POSTEN}_t^{\gamma_3} \mathsf{OUTDOOR}_t^{\gamma_4} \mathsf{RADIO}_t^{\gamma_5} \cdot \\ & \cdot \mathsf{PRINT}_t^{\gamma_6} \mathsf{SOCIALMEDIA}_t^{\gamma_7} \gamma_8^{\mathsf{Rain}_t} \gamma_9^{\mathsf{sal}_t} \gamma_{10}^{\mathsf{HOLIDAY}_t} \varepsilon_t^{(3)} \end{aligned} \tag{6.4}$$

and after the transformation:

$$\begin{aligned}\ln(y_t) = \ln(\gamma_0) + \gamma_1 \ln(\mathsf{TV}_t) + \gamma_2 \ln(\mathsf{DR}_t) + \gamma_3 \ln(\mathsf{DR.POSTEN}_t) + \\ +\gamma_4 \ln(\mathsf{OUTDOOR}_t) + \gamma_5 \ln(\mathsf{RADIO}_t) + \gamma_6 \ln(\mathsf{PRINT}_t) + \\ +\gamma_7 \ln(\mathsf{SOCIALMEDIA}_t) + \gamma_8^\star \mathsf{Rain}_t + \gamma_9^\star \mathsf{sal}_t + \\ +\gamma_{10}^\star \mathsf{HOLIDAY}_t + \ln(\varepsilon_t^{(3)}),\end{aligned} \tag{6.5}$$

where:

$$\gamma_8^\star = \ln(\gamma_8)$$
$$\gamma_9^\star = \ln(\gamma_9)$$
$$\gamma_{10}^\star = \ln(\gamma_{10})$$

The following tables illustrate the estimation results for 6.1, 6.3 and 6.5, respectively.

Table 2: Estimation results of the linear model (OLS)

```
##
## Call:
## lm(formula = "SALES_TOT ~ TV+DR+DR.POSTEN+OUTDOOR+RADIO+PRINT+SOCIALMEDIA+Rain..
    mm.+sal+HOLIDAY",
##     data = regdata)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -23290744  -7598418  -1641095   7793738  82440104
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.226e+07  3.988e+06  10.598  < 2e-16 ***
## TV           2.239e+01  2.567e+00   8.721 9.59e-14 ***
## DR           1.453e+01  8.034e+00   1.809  0.07369 .
## DR.POSTEN    1.742e+01  5.417e+00   3.216  0.00178 **
## OUTDOOR      3.756e+01  1.035e+01   3.631  0.00046 ***
## RADIO       -6.633e+01  2.957e+01  -2.243  0.02724 *
## PRINT        6.304e+00  5.876e+00   1.073  0.28610
## SOCIALMEDIA  1.832e+02  2.400e+01   7.636 1.84e-11 ***
## Rain..mm.    2.796e+05  1.461e+05   1.913  0.05877 .
## sal          6.885e+06  3.734e+06   1.844  0.06834 .
## HOLIDAY      2.906e+06  6.383e+06   0.455  0.64999
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14380000 on 94 degrees of freedom
## Multiple R-squared:  0.8203, Adjusted R-squared:  0.8012
## F-statistic: 42.92 on 10 and 94 DF,  p-value: < 2.2e-16
```

For the linear functional form, relatively high value for R^2 was expected, given the fact that time series data was used. Note that R^2 can be compared among different models, only if the models have exactly the same LHS and exactly the same observations. The values of F-statistics in all

Table 3: Estimation results of the log-linear model (OLS)

```
##
## Call:
## lm(formula = "log(SALES_TOT) ~ TV+DR+DR.POSTEN+OUTDOOR+RADIO+PRINT+SOCIALMEDIA+
    Rain..mm.+sal+HOLIDAY",
##     data = regdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20887 -0.07726 -0.00648  0.06248  0.32894
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.787e+01  3.063e-02 583.581  < 2e-16 ***
## TV           1.661e-07  1.972e-08   8.425 4.07e-13 ***
## DR           1.273e-07  6.171e-08   2.062 0.041929 *
## DR.POSTEN    1.423e-07  4.161e-08   3.420 0.000929 ***
## OUTDOOR      3.334e-07  7.946e-08   4.196 6.15e-05 ***
## RADIO       -6.494e-07  2.271e-07  -2.860 0.005224 **
## PRINT        7.119e-08  4.513e-08   1.577 0.118097
## SOCIALMEDIA  1.547e-06  1.843e-07   8.393 4.76e-13 ***
## Rain..mm.    2.251e-03  1.122e-03   2.006 0.047757 *
## sal          5.934e-02  2.868e-02   2.069 0.041258 *
## HOLIDAY     -4.279e-02  4.903e-02  -0.873 0.385054
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1105 on 94 degrees of freedom
## Multiple R-squared:  0.8332, Adjusted R-squared:  0.8154
## F-statistic: 46.95 on 10 and 94 DF,  p-value: < 2.2e-16
```

cases indicate that all three models (especially the first two) are highly significant. The number of significant parameters slightly vary in each model. The significant parameters that all three models have in common are the intercept, TV, DR.POSTEN, OUTDOOR, RADIO, SOCIALMEDIA.For there parameters the corresponding p-values are smaller than 0.05. In the multiplicative log-linear model, also the parameters for DR, Rain, and sal are significant. For the log-log model, the parameters log(PRINT) and Rain along with the common ones mentioned above are significant.

It is important to mention that each specification has its own unique economic interpretation. That is, the choice of a log versus linear specification should be made largely based on the underlying economics. Table 5 (taken from [25]) summarizes the interpretation of the estimates for each case.

Table 4: Estimation results of the log-log model (OLS)

```
## 
## Call:
## lm(formula = "log(SALES_TOT) ~ log(TV)+log(DR)+log(DR.POSTEN)+log(OUTDOOR)+log(
    RADIO)+log(PRINT)+log(SOCIALMEDIA)+Rain..mm.+sal+HOLIDAY", 
##     data = regdata)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.36417 -0.10845  0.01072  0.06971  0.79316 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      14.953425   0.501883  29.795  < 2e-16 ***
## log(TV)           0.012518   0.003359   3.727 0.000331 ***
## log(DR)           0.005367   0.007390   0.726 0.469484    
## log(DR.POSTEN)    0.013922   0.006767   2.057 0.042410 *  
## log(OUTDOOR)      0.052534   0.014639   3.589 0.000530 ***
## log(RADIO)       -0.012448   0.004007  -3.107 0.002500 ** 
## log(PRINT)        0.181852   0.036741   4.950 3.26e-06 ***
## log(SOCIALMEDIA)  0.012168   0.003768   3.229 0.001709 ** 
## Rain..mm.         0.004394   0.001918   2.292 0.024164 *  
## sal               0.049464   0.047164   1.049 0.296974    
## HOLIDAY          -0.050739   0.086438  -0.587 0.558615    
## ---
## Signif. codes:  
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1858 on 94 degrees of freedom
## Multiple R-squared:  0.5279, Adjusted R-squared:  0.4777 
## F-statistic: 10.51 on 10 and 94 DF,  p-value: 1.001e-11
```

Table 5: (taken from [25]). Summary of the interpretation of Marketing Mix Modelling functional forms

	Dependent Variable	**Independent Variable**	**Interpretation of β**	**Marginal Effect of Δx**
linear model	y	x	$\beta = \Delta y/\Delta x$	β
log-linear model	$\ln(y)$	x	$100 \cdot \beta = \ \%\Delta y/\Delta x$	$y \cdot \beta$
log-log model	$\ln(y)$	$\ln(x)$	$\beta = \ \%\Delta y/ \ \%\Delta x$	$y \cdot \beta/x$

RSS, ESS, and $\hat{\sigma}$ are not comparable in size across the models. This is due to the fact that several variables, including the dependent variable, were transformed to be able to estimate the multiplicative models using least square. Therefore, it is not possible to compare the estimated values for the parameters across the models.

Note that the numbers in Table 4 are estimates for the parameters in 6.3, the linearized version of the log-log model (6.2). To find the estimates for the independent variables that had been logged, an 'anti-ln' transformation must be applied. Instead of just taking the exponential of the estimates from Table 4 to obtain proper estimates for the parameters in the log-log model, the following

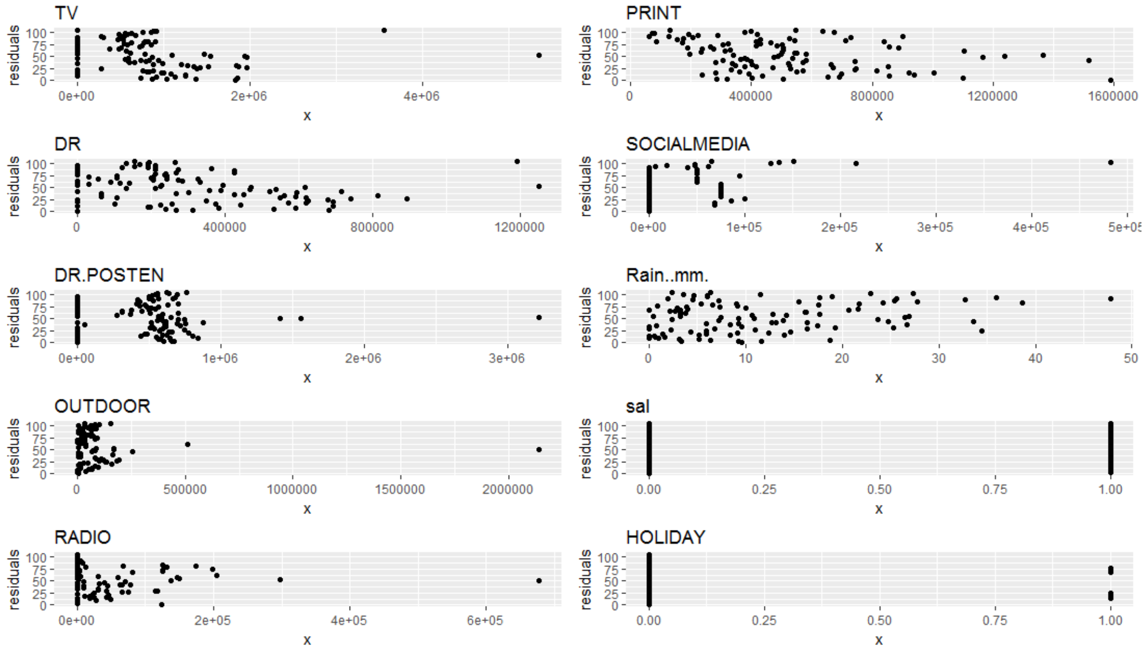

Figure 3: Plot of residuals against each predictor variable for the linear model

correction must be employed ([16]):

$$\hat{\gamma} = \exp(\hat{\gamma}^{\star}) \cdot \exp(-\frac{1}{2}\sigma^2_{\hat{\gamma}^{\star}}) \tag{6.6}$$

The estimates $\hat{\gamma}_8$, $\hat{\gamma}_9$, $\hat{\gamma}_{10}$ from equation 6.4 become:

$$\hat{\gamma}_8 = e^{0.004394} \cdot e^{-\frac{1}{2}0.001918^2} = 1.0044$$

$$\hat{\gamma}_9 = e^{0.049464} \cdot e^{-\frac{1}{2}0.047164^2} = 1.0495$$

$$\hat{\gamma}_{10} = e^{-0.050739} \cdot e^{-\frac{1}{2}0.086438^2} = 0.9470$$

In order to test the first assumption of nonzero expectation of the residuals for each of the models presented above the plots of the residuals against each predictor variable for the models 6.1, 6.3, 6.5 must be examined. The plots of the residuals against each predictor for the models above are presented in Figures 3 and 4 and 5, respectively. These plots should be examined by inspecting each independent variable, to asses if for certain values the residuals differ systematically from zero. To assess this assumption even more carefully, the RESET test was employed for each of the models. Tables 6 and 7 display the results for the RESET test with power 2 and power 3 of the fitted response $\hat{y}_t$, respectively. Table 8 shows the results of the RESET test with both power 2 and 3 of the fitted response. The most appropriate functional form is the log-linear functional form, but one cannot reject the hypothesis that both the second and the third powers of the fitted response are insignificant explanatory variables in the model.

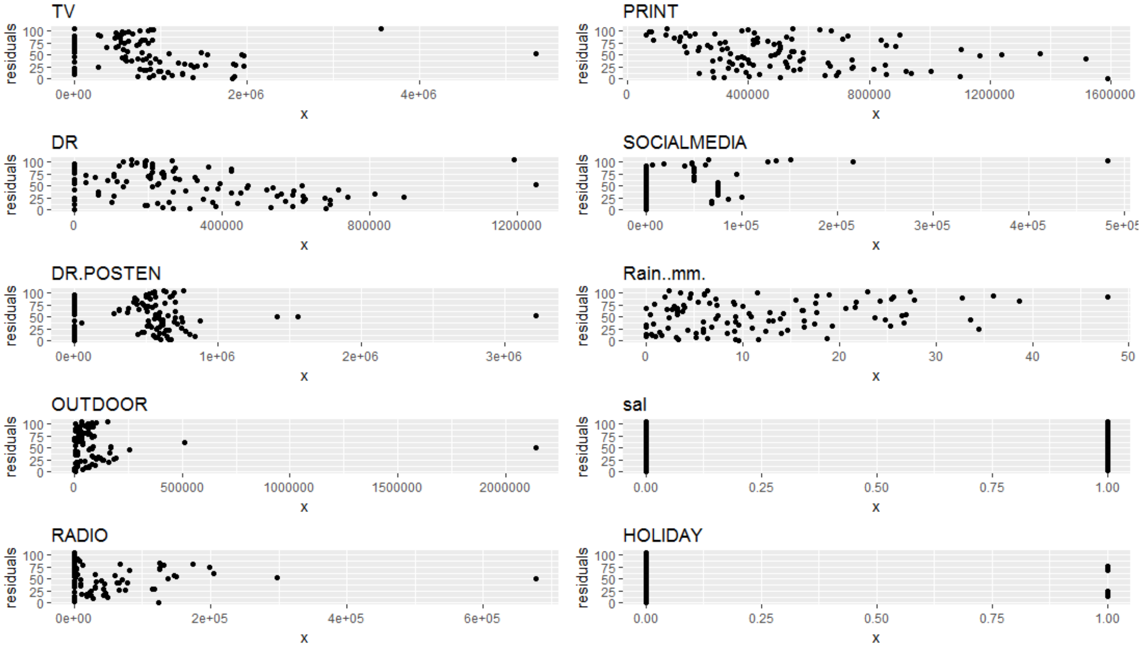

Figure 4: Plot of residuals against each predictor variable for the log-linear model

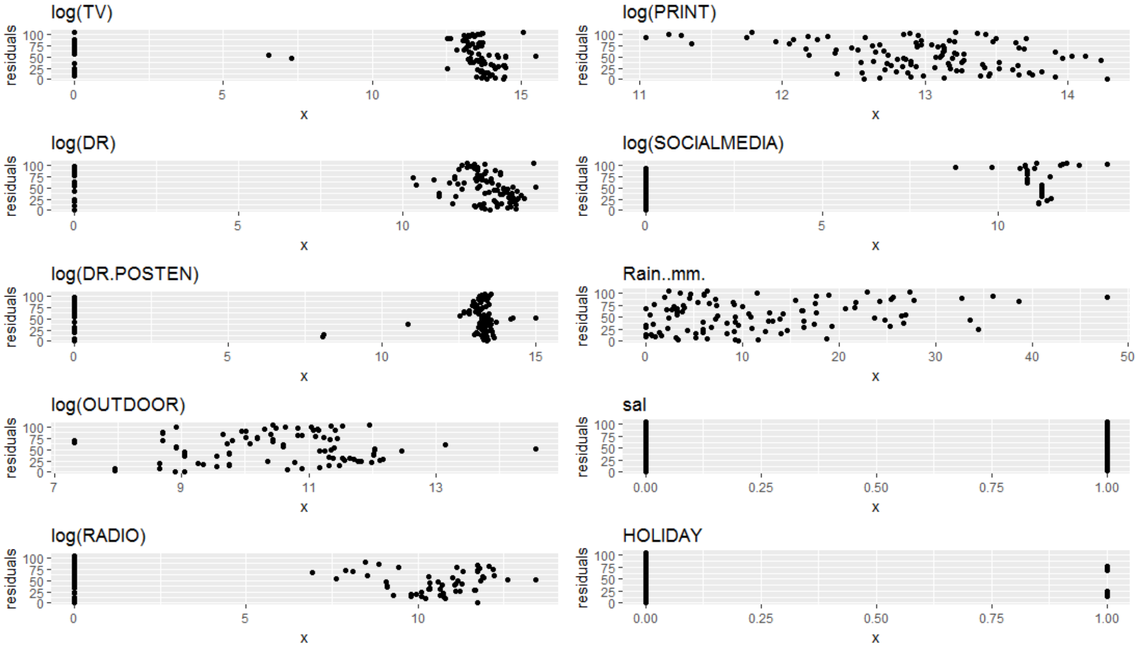

Figure 5: Plot of residuals against each predictor variable for the log-log model

Table 8: RESET test. Power 2 and 3 of the fitted response

linear model	RESET = 38.269, df1 = 2, df2 = 92, p-value = 8.055e-13
log-linear model	RESET = 4.5081, df1 = 2, df2 = 92, p-value = 0.01356
log-log model	RESET = 26.782, df1 = 2, df2 = 92, p-value = 6.818e-10

The Box-Cox test can also be used to determine whether transformations of variables are required. Figure 6 shows the log of the likelihood ratio test for different values of λ. The best fitting transformation is $\lambda = -0.4242424$, which is closest to the log-linear specification. Note that if the task was to fit historical data, the value of λ above would have been chosen. However, this has no economic meaning, thus the log-linear specification is preferred.

Table 6: RESET test. Power 2 of the fitted response

linear model	RESET = 52.878, df1 = 1, df2 = 93, p-value = 1.092e-10
log-linear model	RESET = 1.7934, df1 = 1, df2 = 93, p-value = 0.1838
log-log model	RESET = 35.013, df1 = 1, df2 = 93, p-value = 5.417e-08

Table 7: RESET test. Power 3 of the fitted response

linear model	RESET = 31.869, df1 = 1, df2 = 93, p-value = 1.778e-07
log-linear model	RESET = 0.3821, df1 = 1, df2 = 93, p-value = 0.538
log-log model	RESET = 38.029, df1 = 1, df2 = 93, p-value = 1.785e-08

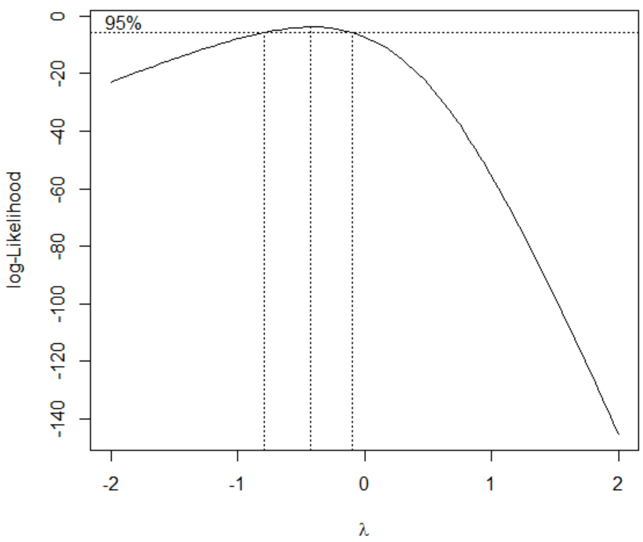

Figure 6: Box-Cox transformation of the response variable, with 95% confidence interval of the parameter λ

6.2 Marketing dynamics

Considering the specification chosen above, the next step is to find the appropriate retention rate for each marketing variable, by solving the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & \sum_{t=1}^{T} (\ln y_t - \ln \hat{y}_t)^2 \\ \text{subject to} \quad & 0 \leq \lambda_i < 1, \; i = 1, \ldots, K. \end{aligned}$$

where:

$$\begin{aligned} \ln \hat{y}_t = \beta_0^\star &+ \beta_1 f_1(\mathsf{TV}_t, \lambda_1) + \beta_2 f_2(\mathsf{DR}_t, \lambda_2) + \beta_3 f_3(\mathsf{DR.POSTEN}_t, \lambda_3) + \beta_4 f_4(\mathsf{OUTDOOR}_t, \lambda_4) + \\ &+ \beta_5 f_5(\mathsf{RADIO}_t, \lambda_5) + \beta_6 f_6(\mathsf{PRINT}_t, \lambda_6) + \beta_7 f_7(\mathsf{SOCIALMEDIA}_t, \lambda_7) + \\ &+ \beta_8 \mathsf{Rain}_t + \beta_9 \mathsf{sal}_t + \beta_{10} \mathsf{HOLIDAY}_t, \end{aligned} \tag{6.7}$$

and $f_1(\mathsf{TV}_t), \ldots, f_7(\mathsf{SOCIALMEDIA}_t)$ are the *adstock functions*, defined as:

$$\begin{aligned} f_1(\mathsf{TV}_t, \lambda_1) &= \mathsf{TV}_t + \lambda_1 f(\mathsf{TV}_{t-1}) \\ &\vdots \\ f_7(\mathsf{SOCIALMEDIA}_t, \lambda_7) &= \mathsf{SOCIALMEDIA}_t + \lambda_7 f(\mathsf{SOCIALMEDIA}_{t-1}) \end{aligned} \tag{6.8}$$

The starting values for all the decays is 0, and as starting values for the parameter coefficients the estimates of the equation 6.3 were used, shown in Table 3. The results of the Non-Linear Least squares regression are shown in Table 9. The estimates are quite close to the ones provided in 3, although their significance changes. To avoid over-fitting, one might pick the significant decays and run a linear regression again. Observe that from all the decays, the `SOCIALMEDIA_adstock` has the p-value 0.055435, so the null hypothesis that the decay for SOCIALMEDIA equals 0 is rejected at significance level $\alpha = 0.1$. The adstock function for SOCIALMEDIA expenditures looks in the following way:

$$f_7(\mathsf{SOCIALMEDIA}_t, 0.2644602) = \mathsf{SOCIALMEDIA}_t + 0.2644602 \cdot f(\mathsf{SOCIALMEDIA}_{t-1})$$

$$t = 2, \ldots, T$$

Table 9: Estimation results of the log-linear model, dynamic structure using Levenberg-Marquardt method

```
##
## Formula: log(SALES_TOT) ~ Intercept + TV_coefficient * adstock(TV, TV_adstock) +
##     DR_coefficient * adstock(DR, DR_adstock) + DR.POSTEN_coefficient *
##     adstock(DR.POSTEN, DR.POSTEN_adstock) + OUTDOOR_coefficient *
##     adstock(OUTDOOR, OUTDOOR_adstock) + RADIO_coefficient * adstock(RADIO,
##     RADIO_adstock) + PRINT_coefficient * adstock(PRINT, PRINT_adstock) +
##     SOCIALMEDIA_coefficient * adstock(SOCIALMEDIA, SOCIALMEDIA_adstock) +
##     Rain..mm._coefficient * Rain..mm. + sal_coefficient * sal +
##     HOLIDAY_coefficient * HOLIDAY
##
## Parameters:
##                           Estimate Std. Error t value Pr(>|t|)
## Intercept                1.787e+01  4.094e-02 436.519  < 2e-16 ***
## TV_coefficient           1.593e-07  2.313e-08   6.885 8.54e-10 ***
## TV_adstock               1.488e-02  1.286e-01   0.116 0.908188
## DR_coefficient           1.118e-07  6.615e-08   1.690 0.094613 .
## DR_adstock               0.000e+00  5.868e-01   0.000 1.000000
## DR.POSTEN_coefficient    1.445e-07  5.444e-08   2.655 0.009436 **
## DR.POSTEN_adstock        7.176e-04  3.168e-01   0.002 0.998198
## OUTDOOR_coefficient      3.229e-07  8.639e-08   3.737 0.000332 ***
## OUTDOOR_adstock          0.000e+00  3.257e-01   0.000 1.000000
## RADIO_coefficient       -5.799e-07  2.428e-07  -2.389 0.019075 *
## RADIO_adstock            3.523e-02  4.130e-01   0.085 0.932215
## PRINT_coefficient        7.731e-08  4.689e-08   1.649 0.102793
## PRINT_adstock            0.000e+00  5.808e-01   0.000 1.000000
## SOCIALMEDIA_coefficient  1.409e-06  2.141e-07   6.581 3.38e-09 ***
## SOCIALMEDIA_adstock      2.645e-01  1.362e-01   1.941 0.055435 .
## Rain..mm._coefficient    2.141e-03  1.173e-03   1.826 0.071339 .
## sal_coefficient          5.876e-02  3.178e-02   1.849 0.067882 .
## HOLIDAY_coefficient     -3.850e-02  5.050e-02  -0.762 0.447846
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1105 on 87 degrees of freedom
##
## Number of iterations to convergence: 20
## Achieved convergence tolerance: 1.49e-08
```

6.3 Re-estimation and testing the OLS assumptions

In this section the lag weight parameters found using the Levenberg-Marquardt algorithm are being used, and the re-estimated model with OLS is tested. Since the only significant decay found in the previous section is the one for SOCIALMEDIA, the following equation is estimated with OLS:

$$\begin{aligned}\ln(y_t) = {} & \beta_0^\star + \beta_1 \mathsf{TV}_t + \beta_2 \mathsf{DR}_t + \beta_3 \mathsf{DR.POSTEN}_t + \beta_4 \mathsf{OUTDOOR}_t + \\ & + \beta_5 \mathsf{RADIO}_t + \beta_6 \mathsf{PRINT}_t + \beta_7 f_7(\mathsf{SOCIALMEDIA}_t, 0.2644602) + \\ & + \beta_8 \mathsf{Rain}_t + \beta_9 \mathsf{sal}_t + \beta_{10} \mathsf{HOLIDAY}_t + \varepsilon_t^\star,\end{aligned} \tag{6.9}$$

The results are shown in Table 10.

Table 10: Estimation results of the log-linear functional form with adstock model considered for SOCIALMEDIA variable

```
##
## Call:
## lm(formula = "log(SALES_TOT) ~ TV+DR+DR.POSTEN+OUTDOOR+RADIO+PRINT+SOCIALMEDIA+
    Rain..mm.+sal+HOLIDAY",
##     data = regdata)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.231972 -0.073825 -0.009095  0.061918  0.291512
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.787e+01  2.955e-02 604.603  < 2e-16 ***
## TV           1.540e-07  1.914e-08   8.048 2.53e-12 ***
## DR           1.124e-07  5.925e-08   1.897 0.060940 .
## DR.POSTEN    1.471e-07  3.999e-08   3.678 0.000391 ***
## OUTDOOR      3.153e-07  7.667e-08   4.113 8.37e-05 ***
## RADIO       -5.581e-07  2.197e-07  -2.540 0.012712 *
## PRINT        8.798e-08  4.351e-08   2.022 0.045990 *
## SOCIALMEDIA  1.901e-06  2.078e-07   9.146 1.20e-14 ***
## Rain..mm.    2.176e-03  1.080e-03   2.015 0.046764 *
## sal          5.672e-02  2.757e-02   2.057 0.042443 *
## HOLIDAY     -4.193e-02  4.717e-02  -0.889 0.376266
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1063 on 94 degrees of freedom
## Multiple R-squared:  0.8456, Adjusted R-squared:  0.8292
## F-statistic: 51.48 on 10 and 94 DF,  p-value: < 2.2e-16
```

The following results indicate that most of the parameters are significant, except HOLIDAY. The goodness of fit has increased, and the p-value for the F-statistics indicates that the model is highly significant. The short-term effect of marketing effort for SOCIALMEDIA is $1.901 * 10^{-6}$. The implied long-term effect is $1.901 * 10^{-6}/(1 - 0.2644602) = 2.584496 * 10^{-6}$.

In order to test the first assumption $\mathrm{E}[\varepsilon_t] = 0$, the residuals against each predictor variable are plotted again (Figure 7). The graphs do not show any systematic pattern in the residuals. Next

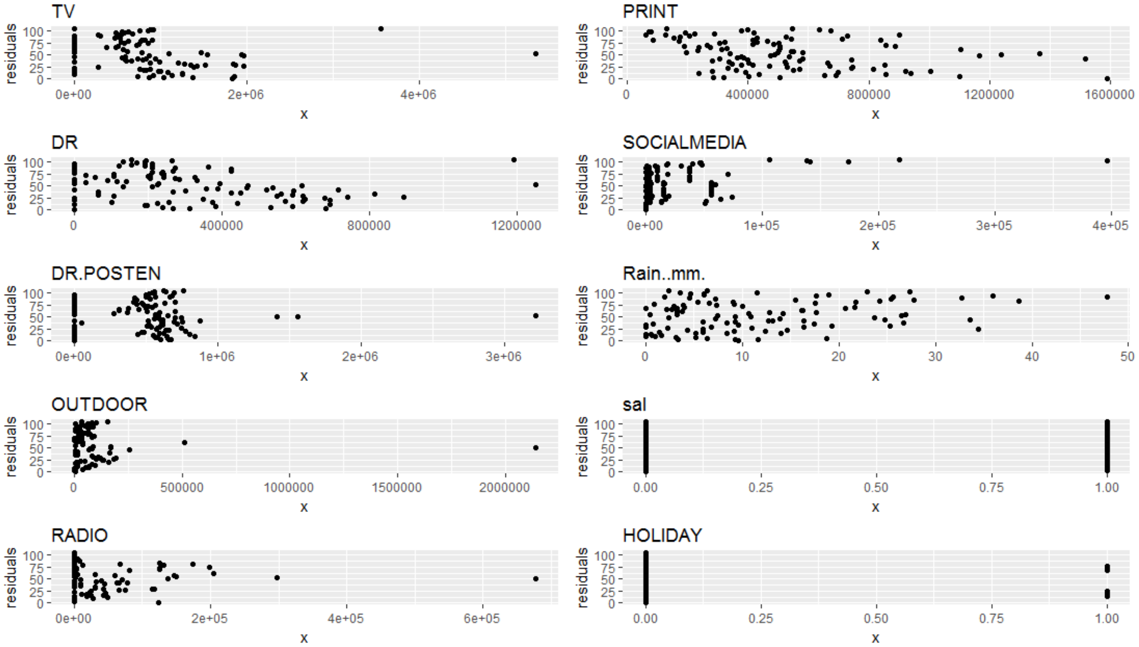

Figure 7: Plot of residuals against each predictor variable for the log-linear functional form with adstock model considered for SOCIALMEDIA variable (equation 6.9)

the RESET test was employed with powers of the fitted response. The results shown in Table 11 indicate that there is no strong evidence of misspecification.

Table 11: RESET test for log-linear adstock model, using powers of the fitted response

power 2	`RESET = 1.1517, df1 = 1, df2 = 93, p-value = 0.286`
power 3	`RESET = 0.3628, df1 = 1, df2 = 93, p-value = 0.5484`
power 2 and 3	`RESET = 2.6459, df1 = 2, df2 = 92, p-value = 0.07634`

In order to test the second assumption $\mathrm{Var}[\varepsilon_t] = 0$ for all t, Figure 7 must be examined again, but now with the purpose to detect changes in the variability of the residuals. To test for heteroscedasticity more formally, the Breusch-Pagan test is employed, by running a regression of the squared residuals on the explanatory variables that appear in equation 6.9. The p-value associated with Breusch-Pagan test is 0.9635, indicating that no significant heteroscedasticity is detected. To test the normality of the residuals, normality tests together with visual assessment were employed. Figure 9 shows the empirical cumulative distribution function together with normal cumulative distribution function. The normal probability plot is shown in Figure 8 *Right*. The plots don't indicate evidence of non-normality of the residuals. To asses nonnormality of the residuals more carefully, Lilliefors test was employed, which returned a p-value equal to 0.7248. From the results above it can be concluded that nonnormality is not an issue. To check the presence of multicollinearity first the correlation matrix of the explanatory variables must be inspected. The correlation matrix

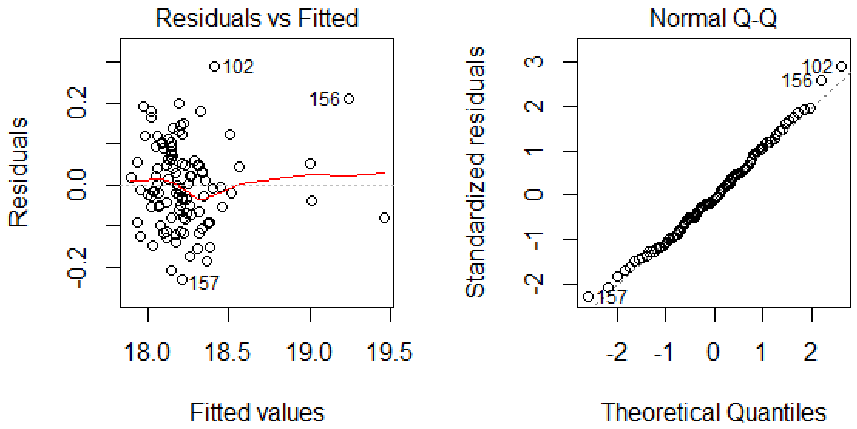

Figure 8: Diagnostics of the regression model considered in equation 6.9. *Left*: scatterplot of the residuals against fitted values. *Right*: Normal probability plot of the residuals

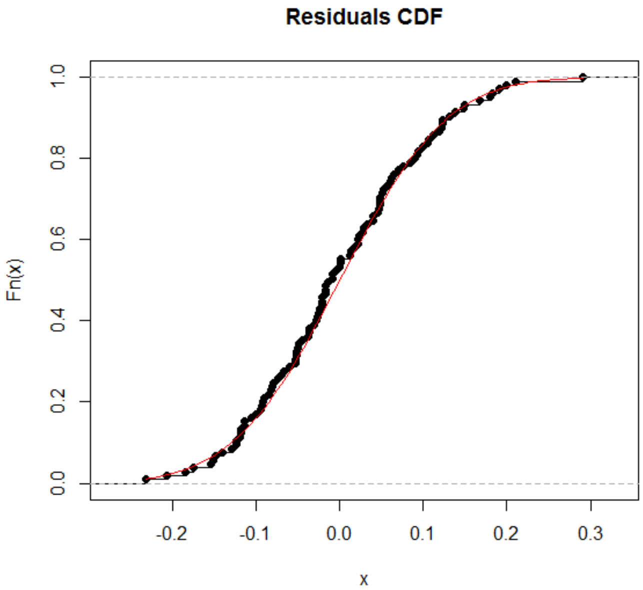

Figure 9: Empirical cumulative distribution function of the residuals for the log-linear functional form with adstock model considered for SOCIALMEDIA variable (equation 6.9)

Table 12: Correlation matrix of the explanatory variables

```
##             (Intercept)    TV     DR DR.POSTEN OUTDOOR   RADIO  PRINT
## (Intercept)        1.00  0.14 -0.240    -0.254  -0.020  0.1763 -0.610
## TV                 0.14  1.00 -0.346    -0.186   0.154 -0.2015 -0.310
## DR                -0.24 -0.35  1.000    -0.490  -0.096  0.1570  0.071
## DR.POSTEN         -0.25 -0.19 -0.490     1.000  -0.030 -0.2383  0.186
## OUTDOOR           -0.02  0.15 -0.096    -0.030   1.000 -0.7079  0.017
## RADIO              0.18 -0.20  0.157    -0.238  -0.708  1.0000 -0.266
## PRINT             -0.61 -0.31  0.071     0.186   0.017 -0.2660  1.000
## SOCIALMEDIA       -0.15 -0.17  0.050    -0.055  -0.158  0.1543  0.067
## Rain..mm.         -0.47 -0.11  0.148    -0.054   0.010  0.0287 -0.015
## sal                0.13  0.14 -0.110    -0.263   0.178 -0.0183 -0.362
## HOLIDAY           -0.16  0.11  0.018     0.050   0.028 -0.0039 -0.176
##             SOCIALMEDIA Rain..mm.    sal HOLIDAY
## (Intercept)     -0.1485    -0.471  0.128 -0.1558
## TV              -0.1671    -0.106  0.138  0.1058
## DR               0.0496     0.148 -0.110  0.0183
## DR.POSTEN       -0.0551    -0.054 -0.263  0.0495
## OUTDOOR         -0.1580     0.010  0.178  0.0281
## RADIO            0.1543     0.029 -0.018 -0.0039
## PRINT            0.0670    -0.015 -0.362 -0.1759
## SOCIALMEDIA      1.0000    -0.036  0.037 -0.0063
## Rain..mm.       -0.0360     1.000  0.129  0.1921
## sal              0.0370     0.129  1.000  0.0745
## HOLIDAY         -0.0063     0.192  0.074  1.0000
```

indicates that RADIO and OUTDOOR are negatively correlated (-0.708). Also, there is evidence of negative correlation between DR and DR.POSTEN (-0.490). Multicollinearity issue might be the cause of the non-significance of some of the coefficients. In Table 13 there are presented the VIF values of the explanatory variables.

Table 13: VIF values of the explanatory variables

	VIF
TV	1.919631
DR	2.122950
DR.POSTEN	2.389954
OUTDOOR	2.487447
RADIO	3.113011
PRINT	1.584613
SOCIALMEDIA	1.077286
Rain	1.099352
sal	1.416964
HOLIDAY	1.114468

The condition number of the matrix $(\mathbf{X}^\intercal\mathbf{X})^{-1}$, after normalizing the data matrix is 16.0652, also indicating moderate degree of multicollinearity. Although in this case the multicollinearity detected is not severe, to provide Nepa with a strategy for the cases of severe multicollinearity, this issue will be addressed in section 6.5.

To assess autocorrelation, first the plot of the residuals over time (Figure 10) were examined.

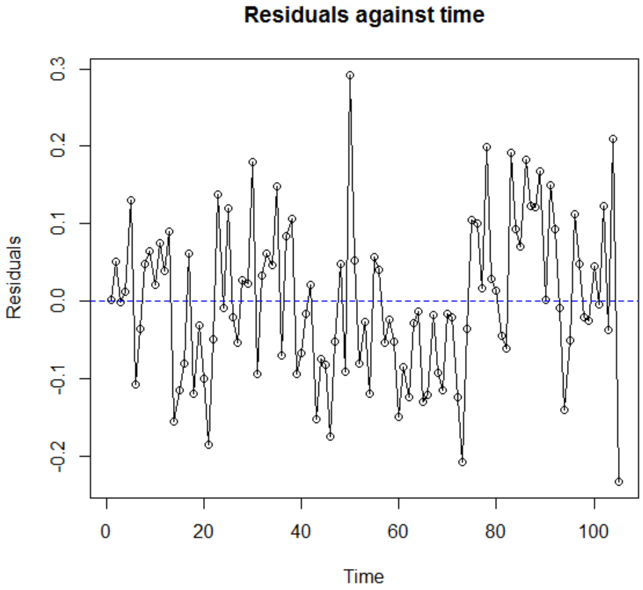

Figure 10: Plot of residuals against time for the log-linear functional form with adstock model considered for SOCIALMEDIA variable (equation 6.9)

The residuals in Figure 10 show shorter and longer runs on either side of the mean value. The D-W Statistic is 1.582541, and the p-value associated with this statistic is 0.018, indicating that the residuals are positively autocorrelated. The estimation of the autocorrelation parameter is 0.1833852, meaning that Durbin Watson test assumes that the errors are driven by the following first order autocorrelation process: $u_t = 0.1833852 * u_{t-1} + \varepsilon_t$.

An approach that specifically considers the autocorrelation structure is the Cochrane-Orcutt method described in [17]. The procedure behind this method is based on the estimation of the autocorrelation coefficient, and then the transformation of variables. With the autocorrelation parameter estimated as 0.1833852, the variables are transformed in the following way:

$$y'_t = y_t - 0.1833852 * y_{t-1}, \quad t = 2, \ldots, T$$

$$x'_{kt} = x_{kt} - 0.1833852 * x_{kt-1}, \quad t = 2, \ldots, T, \quad k = 1, \ldots, K$$

Table 14 summarizes the results of fitting the transformed data with linear regression. For the transfomed regression model the D-W Statistic is 1.889608 and the p-value is 0.576, indicating that there is no problem of autocorrelation in the transformed model.

Table 14: Coefficient estimates for the Cochrane-Orcutt Method, log-linear functional form with adstock model considered for SOCIALMEDIA variable (equation 6.9)

```
##
## Call:
## lm(formula = "SALES_TOT ~ TV+DR+DR.POSTEN+OUTDOOR+RADIO+PRINT+SOCIALMEDIA+Rain..
    mm.+sal+HOLIDAY",
##     data = regdata)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -23290744  -7598418  -1641095   7793738  82440104
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.226e+07  3.988e+06  10.598  < 2e-16 ***
## TV           2.239e+01  2.567e+00   8.721 9.59e-14 ***
## DR           1.453e+01  8.034e+00   1.809  0.07369 .
## DR.POSTEN    1.742e+01  5.417e+00   3.216  0.00178 **
## OUTDOOR      3.756e+01  1.035e+01   3.631  0.00046 ***
## RADIO       -6.633e+01  2.957e+01  -2.243  0.02724 *
## PRINT        6.304e+00  5.876e+00   1.073  0.28610
## SOCIALMEDIA  1.832e+02  2.400e+01   7.636 1.84e-11 ***
## Rain..mm.    2.796e+05  1.461e+05   1.913  0.05877 .
## sal          6.885e+06  3.734e+06   1.844  0.06834 .
## HOLIDAY      2.906e+06  6.383e+06   0.455  0.64999
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14380000 on 94 degrees of freedom
## Multiple R-squared:  0.8203, Adjusted R-squared:  0.8012
## F-statistic: 42.92 on 10 and 94 DF,  p-value: < 2.2e-16
```

An alternative estimation method that deals with the problem of autocorrelation is the maximum likelihood method. As mentioned in [17] this method is attractive, because it can be used when the structure of the errors is more complicated than the autoregressive process of order one. Table 15 shows the output from the maximum likelihood estimation assuming first order autoregressive process of the residuals, using `gls` function in `R` .The autocorrelation parameter is estimated to be 0.234079, which is close to the value retrieved by the D-W test.

Table 15: Coefficient estimates shown for the maximum likelihood estimation, log-linear functional form with adstock model considered for SOCIALMEDIA variable (equation 6.9)

```
## Generalized least squares fit by maximum likelihood
##   Model: SALES_TOT ~ .
##   Data: transformedregdata
##   Log-likelihood: 94.48966
##
## Coefficients:
##   (Intercept)            TV            DR     DR.POSTEN       OUTDOOR
##  1.787338e+01  1.454923e-07  1.241126e-07  1.450152e-07  2.815697e-07
##         RADIO         PRINT   SOCIALMEDIA     Rain..mm.           sal
## -5.038014e-07  9.843381e-08  1.850852e-06  1.631049e-03  5.712213e-02
##       HOLIDAY
## -4.609447e-02
##
## Correlation Structure: AR(1)
##  Formula: ~1
##  Parameter estimate(s):
##      Phi
## 0.234079
## Degrees of freedom: 105 total; 94 residual
## Residual standard error: 0.101172
```

6.4 Variable selection

Next step in building the model is variable selection. Since in this case a small number of variables is used, it is possible to perform best subset selection using the `regsubsets()` function in `R`. The best model that contains a given number of predictors (using RSS) is shown in Table 16.

Table 16: Best Subset Selection. The *best* model that contains a given number of predictors is chosen according to RSS

```
##           TV  DR  DR.POSTEN OUTDOOR RADIO PRINT SOCIALMEDIA Rain..mm. sal HOLIDAY
## 1  ( 1 )  "*" " " " "       " "     " "   " "   " "         " "       " " " "
## 2  ( 1 )  "*" " " " "       " "     " "   " "   "*"         " "       " " " "
## 3  ( 1 )  "*" " " "*"       " "     " "   " "   "*"         " "       " " " "
## 4  ( 1 )  "*" " " "*"       "*"     " "   " "   "*"         " "       " " " "
## 5  ( 1 )  "*" " " "*"       "*"     " "   " "   "*"         " "       "*" " "
## 6  ( 1 )  "*" " " "*"       "*"     "*"   " "   "*"         " "       "*" " "
## 7  ( 1 )  "*" " " "*"       "*"     "*"   " "   "*"         "*"       "*" " "
## 8  ( 1 )  "*" "*" "*"       "*"     "*"   " "   "*"         "*"       "*" " "
## 9  ( 1 )  "*" "*" "*"       "*"     "*"   "*"   "*"         "*"       "*" " "
## 10  ( 1 ) "*" "*" "*"       "*"     "*"   "*"   "*"         "*"       "*" "*"
```

Figure 11 displays the plots of RSS, adjusted R^2, C_p, and BIC for all of the models at once. It can be seen that both adjusted R^2 and C_p choose the model with 9 variables, while BIC chooses the model with 6 variables. As mentioned in [6], the BIC statistic generally places a heavier penalty on models with many variables, and hence results in the selection of smaller models than C_p.

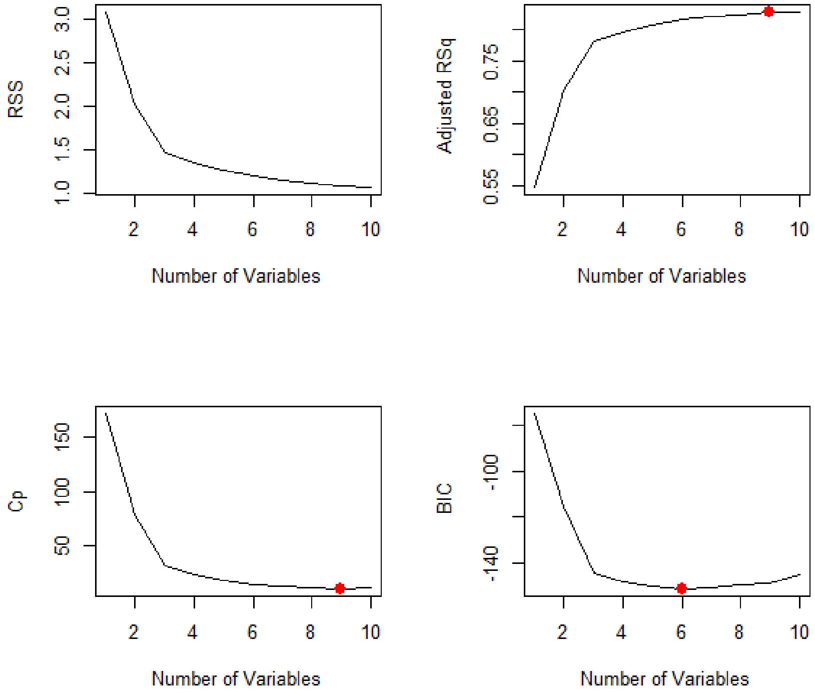

Figure 11: RSS, adjusted R^2, C_p, and BIC shown for the best models of each size

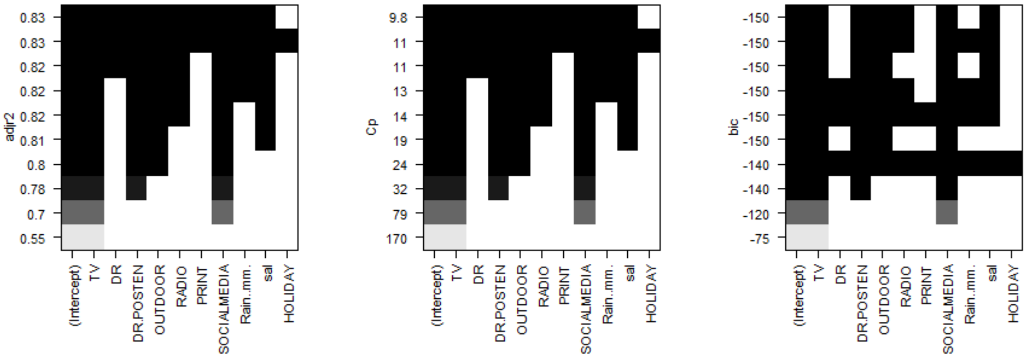

Figure 12: Adjusted R^2, C_p and BIC for log-linear functional form with adstock model considered for SOCIALMEDIA variable (equation 6.9)

Figure 12 displays the selected variables for the best model with a given number of predictors, ranked according to adjusted R^2, C_p and BIC.

One can also choose among a set of models of different sizes using the validation set and cross-validation approaches. For these approaches to yield accurate estimates of the test error, only the training observations must be used. The observations must be split into a training set and a test set. Next, best subset selection only on the training observations should be performed. The validation set error is computed for the best model of each model size, getting the results from Table 17.

Table 17: Validation set errors for the best model of each model size

Model size	**Validation set error**
1	0.03964116
2	0.03523362
3	0.04432925
4	0.02418651
5	0.02353050
6	0.01786729
7	0.01690079
8	0.01579381
9	0.01565193
10	0.01540573

The best model is found to be the one that contains ten variables. Next we would have to perform best subset selection on the full dataset and select the best ten-variable model, but since it is the full model, we just re-estimate the coefficients on the full dataset. Since the full model was selected, the estimates will be the ones from Table 10.

To choose among the models of different sizes using cross-validation, best subset selection is performed within each of the k training sets. First, each observation is allocated to one of $k = 10$ folds. Next each of the folds is used as a test set for the best subset selection procedure, and the rest of the data is used as the training set. The test errors are stored in a matrix, and then the average is calculated over the columns of this matrix in order to obtain a vector for which the jth element is the cross-validation error for the j-variable model, $j = 1, \dots, k$, ($k = 10$, number of folds). Figure 13 shows that cross-validation selected the three-variable model. Performing cross-validation multiple times for the dataset provided in the current case study, the cross-validation error always decreased for the three-variable model, followed by a growth, finally to decrease again by ten-variable model, as shown in Figure 13. If Nepa preferred a retracted model, a three-variable model would have been selected, otherwise it is also possible to chose the full model.

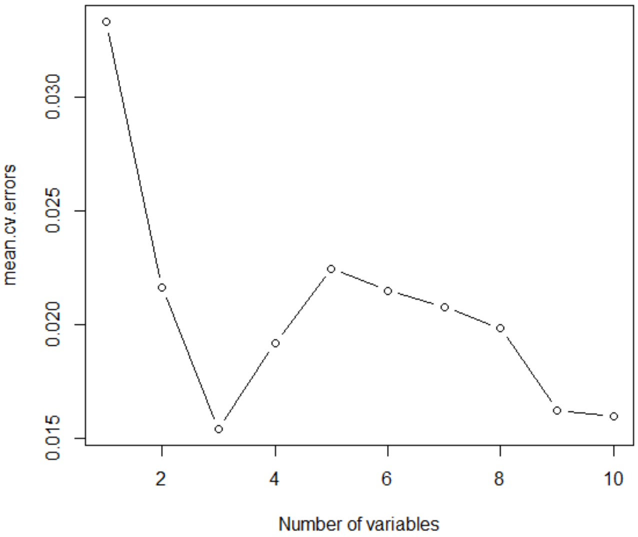

Figure 13: Cross-validation errors for the log-linear functional form with adstock model considered for SOCIALMEDIA variable (equation 6.9)

It is important to perform best subset selection on the full dataset to obtain reliable estimates for the three-variable model. Results are shown in Table 18.

Table 18: Parameter Estimates for the three-variable model

(Intercept)	1.793679e+01
TV	1.739107e-07
DR.POSTEN	2.131372e-07
SOCIALMEDIA	1.948903e-06

6.5 Ridge regression

As it can be seen in section 6.3, although some of the variables show correlation, there is no strong evidence of severe multicollinearity. But even though the best linear unbiased estimator of the coefficients is given by the ordinary least squares (OLS) estimator (the Gauss-Markov Theorem), the least squares estimates might have high variance, making the estimates inefficient for out of the sample data. As the purpose of the thesis is to develop a general model building strategy which Nepa will use for future projects, the next step is to compare the performance of OLS estimation with different shrinkage methods. To select the method which is most suitable for the current data all regularization methods presented in section 5 will be applied to equation 6.9 and their prediction accuracy will be compared.

Ridge regression was performed using the function `glmnet` in `R`, over a grid of values ranging from $\lambda = 10^{10}$ to $\lambda = 10^{-2}$. This grid essentially covers all scenarios from the model containing only the intercept, to the least squares fit. It is recommended to standardize the variables before performing ridge regression, so that all the variables would be on the same scale. The function `glmnet` does it automatically, returning the coefficient estimates of the variables in the original scale.

In order to estimate the optimal parameter λ, ten-fold cross-validation was performed, using `cv.glmnet` function in `R`. First ridge regression model is fitted on the training set. Next cross-validation is used to choose the tuning parameter λ that gives the smallest cross-validation error. For each value of λ the test MSE is calculated. Finally, ridge regression model is refitted on the full dataset, using the value of λ chosen by cross-validation.

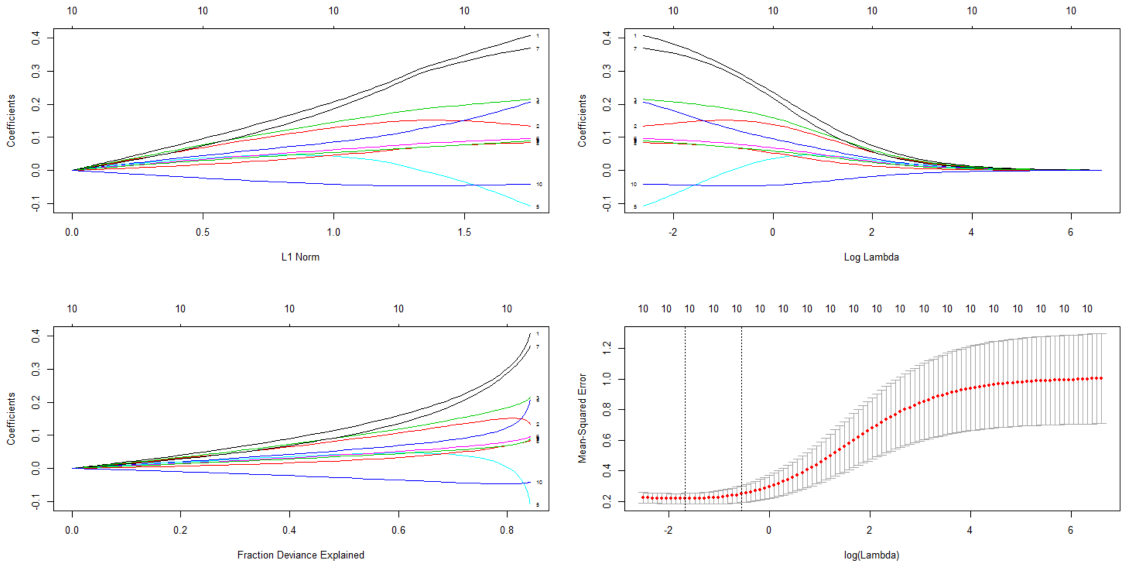

Figure 14: Ridge regression on the full dataset. Upper left: L1 norm against scaled coefficients. Upper right: Log lambda against scaled coefficients. Lower left: Fraction of deviance explained against scaled coefficients. Lower right: Log(lambda) against MSE.

Figure 14 shows the plots of l_1 norm, $\ln(\lambda)$ and fraction of deviance explained against coefficient estimates, as well as the plot of $\ln(\lambda)$ against mean squared error (lower right). At the top of each graph the number of nonzero coefficients is indicated.

Besides choosing the λ value that gives the smallest cross-validation error, one can also choose the value of λ which gives the most regularized model such that error is within one standard error of the minimum. In Table 19 are presented the estimated coefficients for lambda.min and lambda.1se chosen by cross-validation. As expected, none of the coefficients is exactly zero, as ridge regression does not perform variable selection. Figure 15 shows the plots of predicted values against actual values of sales, for ridge coefficients above from Table 19. The coefficients corresponding to lambda.min predict sales more accurately, since they are chosen in such a way that the cross-validation error is minimal.

Table 19: Ridge regression coefficient estimates for lambda.min and lambda.1se chosen by cross-validation

```
## lambdaminridge= 0.0160694261981513

## lambda1seridge= 0.124419731067532

## msetest.ridge.lambdamin= 0.0188158022878513

## msetest.ridge.lambda1se= 0.0131380574237442

## 11 x 2 sparse Matrix of class "dgCMatrix"
##             ridge.coef.lambdamin ridge.coef.lambda1se
## (Intercept)         1.788044e+01         1.793414e+01
## TV                  1.390966e-07         1.014373e-07
## DR                  1.338492e-07         1.514207e-07
## DR.POSTEN           1.372515e-07         1.164731e-07
## OUTDOOR             2.476422e-07         1.450592e-07
## RADIO              -3.331841e-07         2.032320e-08
## PRINT               8.226841e-08         6.783974e-08
## SOCIALMEDIA         1.821864e-06         1.402040e-06
## Rain..mm.           2.124742e-03         1.716374e-03
## sal                 5.090904e-02         3.908197e-02
## HOLIDAY            -4.598482e-02        -5.182046e-02
```

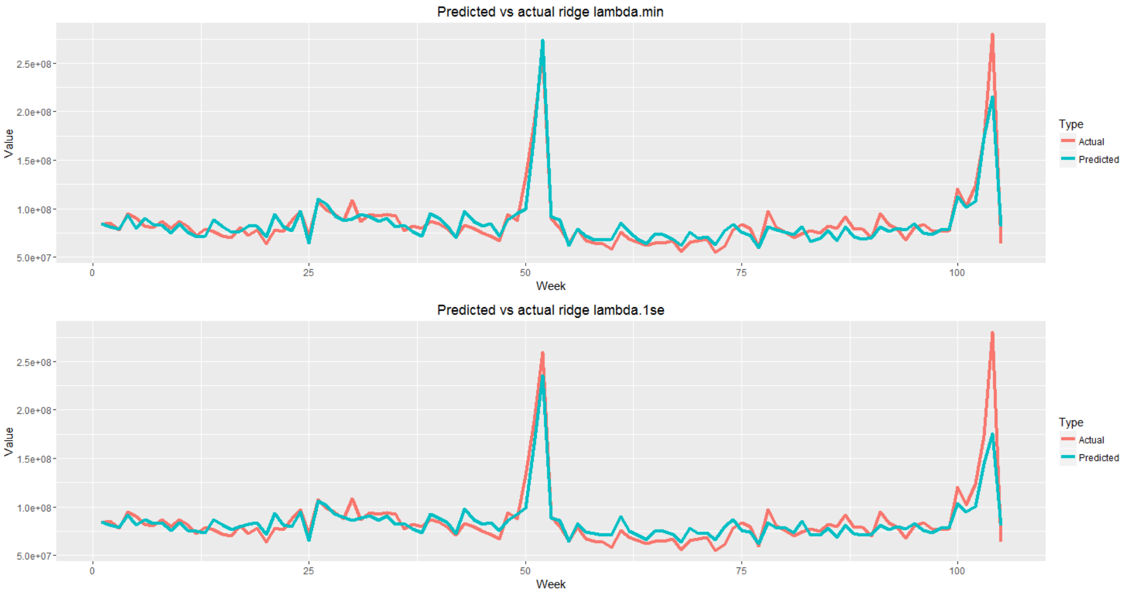

Figure 15: Ridge regression fit for lambda.min (*top*) and lambda.1se (*bottom*) chosen by cross-validation

6.6 The Lasso

To perform the Lasso, the `glmnet` function in `R` is used again for the same range of λ. Following the same cross-validation procedure as for ridge regression, the coefficient estimates and test MSE are obtained for lambda.min and lambda.1se chosen by cross-validation. The lasso test MSE is close to the ridge test MSE. However, the lasso has a substantial advantage over the ridge regression that it performs also variable selection. The results are shown in Table 20. For the largest λ at which the MSE is within one standard error of the minimal MSE two coefficient estimates are zero: RADIO and HOLIDAY.

Table 20: Lasso coefficient estimates for lambda.min and lambda.1se chosen by cross-validation

```
## lambdaminlasso= 0.00294189136833369

## lambda1selasso= 0.0207544620232037

## msetest.lasso.lambdamin= 0.0192751123266679

## msetest.lasso.lambda1se= 0.0190073954177291

## 11 x 2 sparse Matrix of class "dgCMatrix"
##             lasso.coef.lambdamin lasso.coef.lambda1se
## (Intercept)         1.788390e+01         1.796400e+01
## TV                  1.509192e-07         1.479784e-07
## DR                  1.135039e-07         7.858522e-08
## DR.POSTEN           1.390981e-07         1.317557e-07
## OUTDOOR             2.518648e-07         1.071373e-07
## RADIO              -3.414987e-07         .
## PRINT               6.911367e-08         9.852782e-09
## SOCIALMEDIA         1.879224e-06         1.606965e-06
## Rain..mm.           1.902475e-03         3.636966e-05
## sal                 5.072095e-02         1.575444e-02
## HOLIDAY            -3.266459e-02         .
```

Figure 16 shows each curve's path of its coefficient against the l_1-norm , $\ln(\lambda)$ and fraction of deviance explained, as well as the plot of $\ln(\lambda)$ against mean squared error (lower right). At the top of each graph the number of nonzero coefficients is indicated. Figure 17 shows the plots of predicted against actual sales, using lasso coefficients from Table 20.

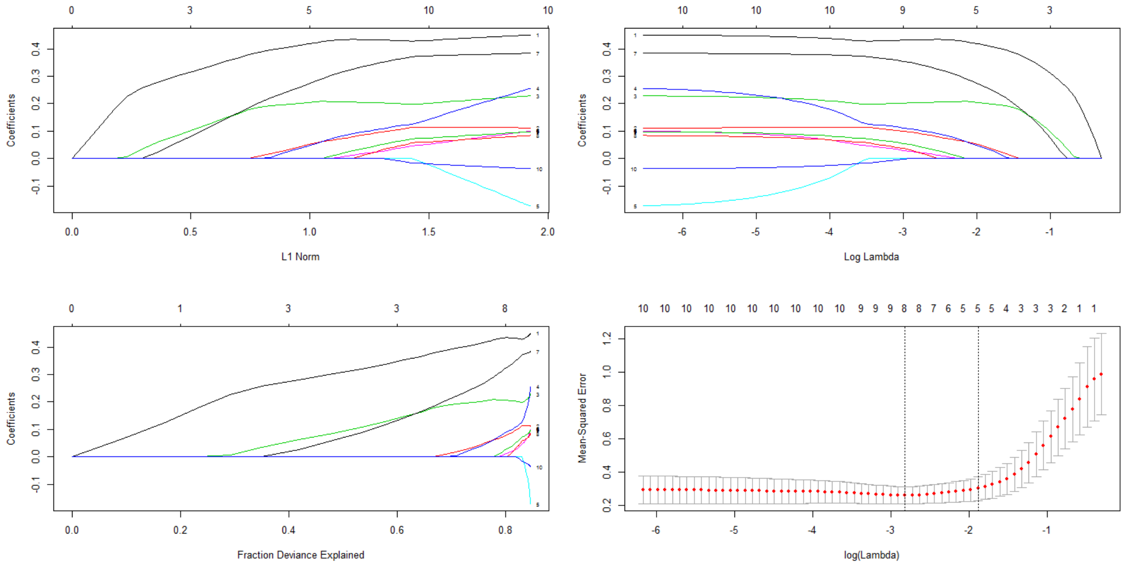

Figure 16: The lasso on the full dataset. Upper left: L1 norm against scaled coefficients. Upper right: Log lambda against scaled coefficients. Lower left: Fraction of deviance explained against scaled coefficients. Lower right: Log(lambda) against MSE.

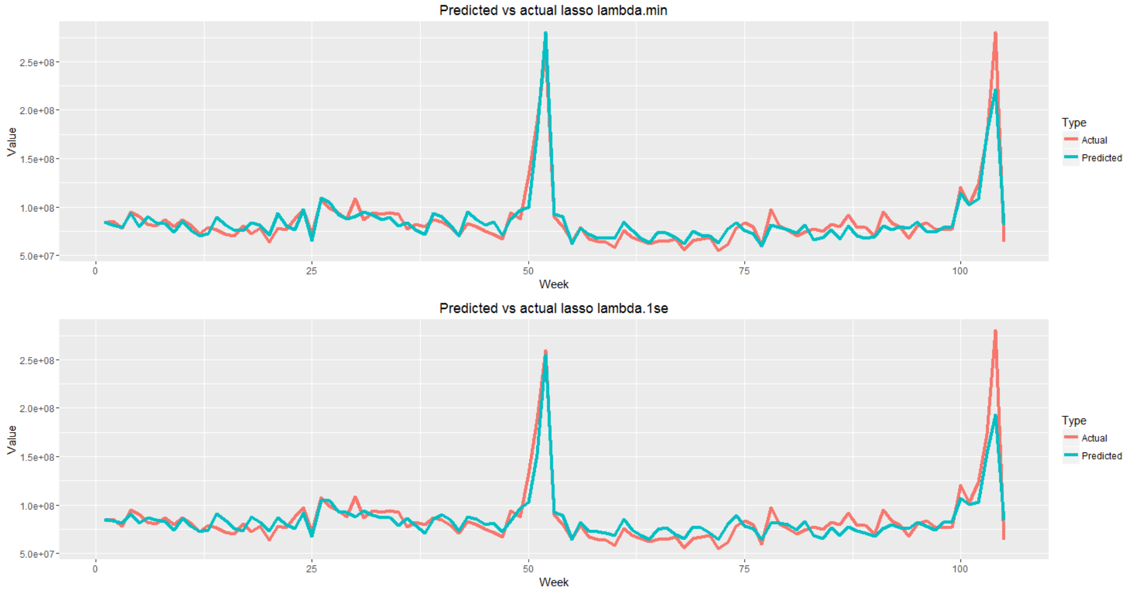

Figure 17: Lasso fit for lambda.min (*top*) and lambda.1se (*bottom*) chosen by cross-validation

6.7 Naive elastic net

As described in section 5.2.6, the naive elastic net penalty is a convex combination of the lasso and ridge penalty.

$$(1-\alpha)\sum_{j=1}^{p}|\beta_j|+\alpha\sum_{j=1}^{p}\beta_j^2 \leq s$$

To choose the optimal parameter α, the function `cv.glmnet` was called with a pre-computed vector `foldid`, and then this same fold vector was used in separate calls to `cv.glmnet` with different values of α. Note that in the `glmnet` package in `R` the penalty is defined as

$$(1-\alpha)/2\sum_{j=1}^{p}\beta_j^2+\alpha\sum_{j=1}^{p}|\beta_j|\leq s$$

It can be seen in the Figure 18 that ridge does about the best for the given dataset, so it seems reasonable to choose a value of α closer to ridge. Calling the `cv.glmnet` function with parameter `alpha` 0.1 yields the following results shown in Table 21 and Figures 19 and 20.

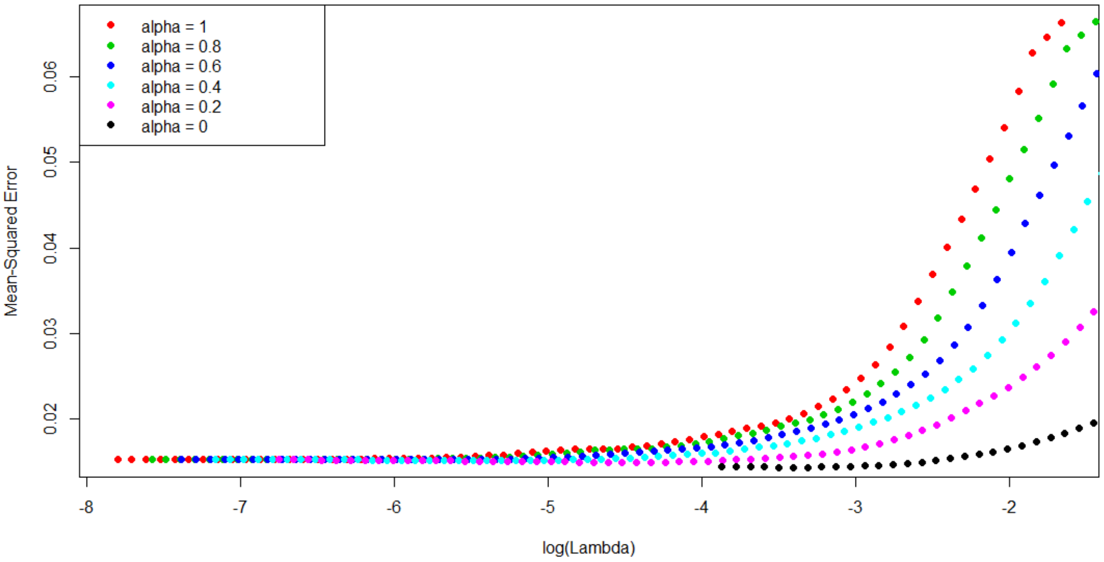

Figure 18: The standardized coefficients as a function of λ, displayed for several values of α

Table 21: Naive elastic net coefficient estimates for lambda.min and lambda.1se chosen by cross-validation

```
## lambdaminelnet= 0.0202773160550586

## lambda1seelnet= 0.108213937001959

## msetest.elnet.lambdamin= 0.0189826745876587

## msetest.elnet.lambda1se= 0.0120703463671487

## 11 x 2 sparse Matrix of class "dgCMatrix"
##              elnet.coef.lambdamin elnet.coef.lambda1se
## (Intercept)          1.789012e+01         1.795649e+01
## TV                   1.380539e-07         1.065173e-07
## DR                   1.321510e-07         1.361582e-07
## DR.POSTEN            1.335377e-07         1.171296e-07
## OUTDOOR              2.179913e-07         1.278893e-07
## RADIO               -2.296430e-07         .
## PRINT                7.185855e-08         4.996454e-08
## SOCIALMEDIA          1.805622e-06         1.357760e-06
## Rain..mm.            1.947446e-03         1.036418e-03
## sal                  4.743022e-02         2.685974e-02
## HOLIDAY             -3.979584e-02        -2.557000e-02
```

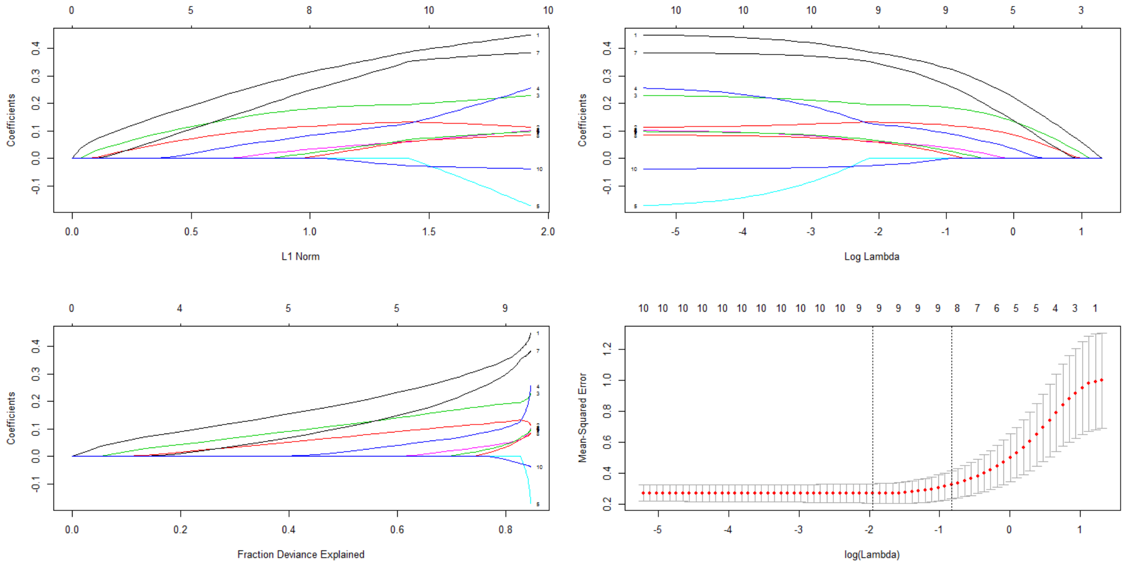

Figure 19: Naive elastic net on the full dataset. Upper left: L1 norm against scaled coefficients. Upper right: Log lambda against scaled coefficients. Lower left: Fraction of deviance explained against scaled coefficients. Lower right: Log(lambda) against MSE.

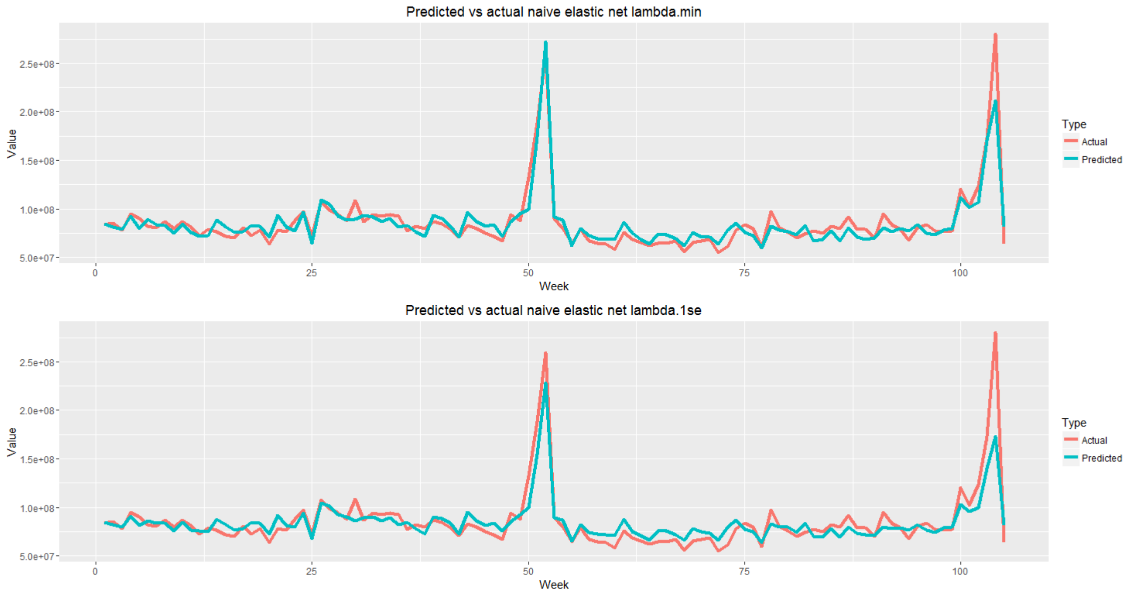

Figure 20: Naive elastic net fit for lambda.min (*top*) and lambda.1se (*bottom*) chosen by cross-validation

It is clear that even naive elastic net outperforms ridge and lasso.

6.8 Elastic net

The elastic net has the advantages of both variables selection and continuous shrinkage, similar to the lasso. The elastic net method was performed using the function `elasticnet` package in `R`. The optimal parameters were chosen by ten-fold cross-validation, as described in section 5.2.7. Figure 21 shows the elastic net estimates and the solution path for $\lambda_2 = 1$ as a function of s, where s refers to the ratio of the l_1 norm of the coefficient vector, relative to the norm of the full LS solution. The minimum cross-validation error is obtained around value $s = 0.5$. Table 22 shows the elastic net coefficients for $\lambda_2 = 1$ and $s = 0.47$. The predicted sales using elastic net method together with the actual sales are plotted in Figure 22. The mean-squared test error for $\lambda_2 = 1$ and $s = 0.47$ is 0.01188888.

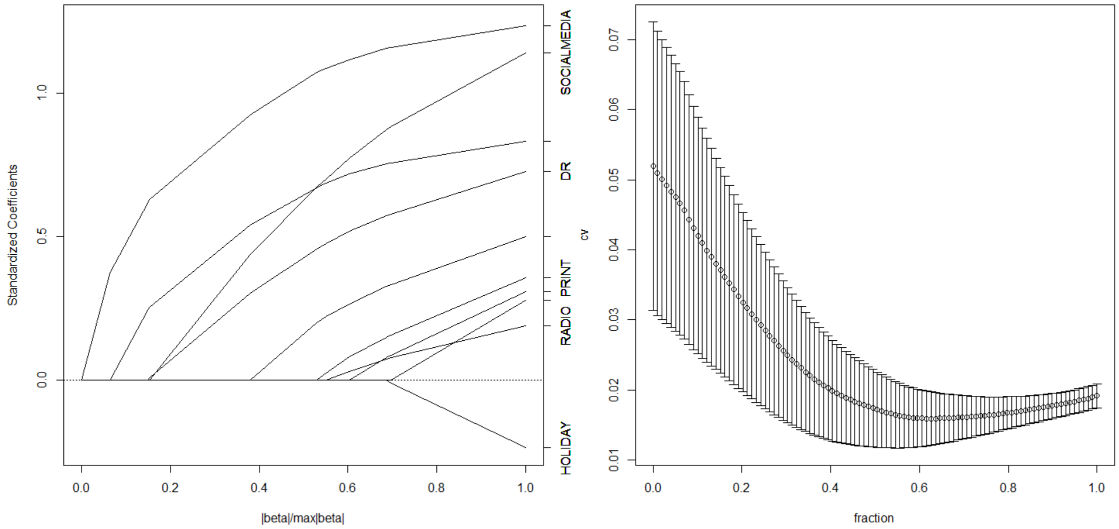

Figure 21: *Left*: Elastic net estimates ($\lambda_2 = 1$) as a function of s. *Right*: solution path ($\lambda_2 = 1$) as a function of s

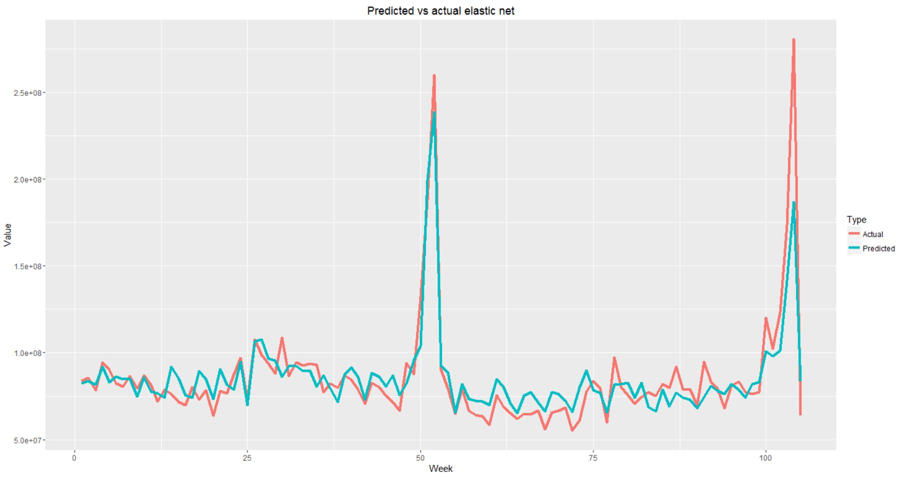

Figure 22: Elastic net fit for $\lambda_2 = 1$ and $s = 0.47$ chosen by cross-validation

To include the seasonal effect, the model was extended by adding a trend variable and 51 dummy variables for seasons, as described in section 2.3. Table 23 shows the coefficients for the extended model with $\lambda_2 = 0.1$ and $s = 0.4$. It can be seen that the method selects only some of the seasonal variables, allowing a partial seasonal adjustment.

Table 22: Elastic net coefficient estimates for $\lambda_2 = 1$ and $s = 0.47$ chosen by cross-validation

```
## $s
## [1] 0.47
##
## $fraction
##    0
## 0.47
##
## $mode
## [1] "fraction"
##
## $coefficients
##           TV           DR    DR.POSTEN      OUTDOOR        RADIO        PRINT
## 1.320210e-07 1.511765e-07 1.506639e-07 5.525586e-08 0.000000e+00 0.000000e+00
##  SOCIALMEDIA    Rain..mm.          sal      HOLIDAY
## 1.091444e-06 0.000000e+00 0.000000e+00 0.000000e+00
```

Table 23: Elastic net coefficient estimates with season and trend variables

```
## $s
## [1] 0.4
##
## $fraction
##   0
## 0.4
##
## $mode
## [1] "fraction"
##
## $coefficients
##            S1            S2            S3            S4            S5
##  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
##            S6            S7            S8            S9           S10
##  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
##           S11           S12           S13           S14           S15
##  0.000000e+00  0.000000e+00  0.000000e+00 -3.310320e-02  0.000000e+00
##           S16           S17           S18           S19           S20
##  0.000000e+00  0.000000e+00 -2.827022e-02  0.000000e+00 -2.507911e-02
##           S21           S22           S23           S24           S25
## -6.727018e-02  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
##           S26           S27           S28           S29           S30
##  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
##           S31           S32           S33           S34           S35
##  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  4.078321e-02
##           S36           S37           S38           S39           S40
##  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
##           S41           S42           S43           S44           S45
##  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
##           S46           S47           S48           S49           S50
##  0.000000e+00  0.000000e+00  2.042501e-02  0.000000e+00  1.358926e-01
##           S51         trend            TV            DR     DR.POSTEN
##  1.703252e-01  0.000000e+00  1.499851e-07  1.342277e-07  1.178116e-07
##       OUTDOOR         RADIO         PRINT   SOCIALMEDIA     Rain..mm.
##  7.377897e-08  0.000000e+00  1.595277e-08  1.445939e-06  1.432973e-04
##           sal       HOLIDAY
##  2.989432e-02  0.000000e+00
```

7 Conclusions & Recommendations

This thesis illustrates an application of modern approaches of statistical learning on a set of data provided by Nepa. The goal of the thesis is to construct a model building strategy suitable for a high level of complexity of the data, with the ambition to tackle several difficulties encountered with statistical analysis applied to marketing economics. A marketing mix model must address all elements of the problem being studied. In the specification step, one of such elements is the choice of the appropriate functional form. To find the suitable specification, which describes the relationship between the dependent and independent variables, the RESET test and the Box-Cox transformation of the response variable were used. The plot of the residuals against each predictor variable as well as the tests above suggest that the log-linear specification is appropriate. Several subset selection methods were employed on the log-linear model. The results of the validation set and cross-validation approaches justify the choice of the full model. To adapt the model to the dynamic marketing behavior, the optimal lag weight parameters can be found with the Levenberg-Marquardt algorithm, using `nlsLM` function in `R`.

Since the purpose is both explanatory and predictive analysis, in order to be able to perform statistical inferences based on obtained point estimates, the assumptions made in section 4 must hold. To sum up, the results show that we cannot assume that the error terms are uncorrelated. The solution proposed was to employ the Cochrane-Orcutt method, an approach that specifically considers the autocorrelation structure, or to use alternative estimation methods, such as the method of maximum likelihood. The testing of the assumptions also shows that the data exhibits mild degree of multicollinearity. A comparison of several estimation methods is provided, so that Nepa could use this thesis as a guideline for future marketing mix modelling projects that include data with severe multicollinearity. Regularization methods were performed using `glmnet` and `elasticnet` packages in `R`. Note that the penalty in the `glmnet` package in `R` is defined differently from the penalty in the `elasticnet` package. Table 24 shows the performance of ridge regression, the lasso, the naive elastic net and elastic net results applied to the same training set and validation set. Model fitting and tuning parameter selection by tenfold cross-validation (CV) should be carried out on the training data, and then the performance of those methods must be compared by computing their prediction mean-squared error (MSE) on the test data. Although the difference between mentioned methods is not signifficant, the lowest test MSE is achieved by elastic net, which also chose the smallest number of variables.

Method	Parameters	test MSE	Variables selected
Ridge regression	$\lambda_1 = 0$, $\lambda_2 = 0.01606943 * 2$	0.0188158	All
Ridge regression	$\lambda_1 = 0$, $\lambda_2 = 0.1244197 * 2$	0.01313806	All
Lasso	$\lambda_1 = 0.002941891$, $\lambda_2 = 0$	0.01927511	All
Lasso	$\lambda_1 = 0.02075446$, $\lambda_2 = 0$	0.0190074	(1,2,3,4,6,7,8,9)
Naive elastic net	$\lambda_1 = 0.02027732/0.1$, $\lambda_2 = 0.02027732 * 2/(1 - 0.1)$	0.01898267	All
Naive elastic net	$\lambda_1 = 0.1082139/0.1$, $\lambda_2 = 0.1082139 * 2/(1 - 0.1)$	0.01207035	(1,2,3,4,6,7,8,9,10)
Elastic net	$\lambda_2 = 1$, $s = 0.47$	0.01188888	(1,2,3,4,7)

Table 24: Comparing the mean-squared error of the regularization methods

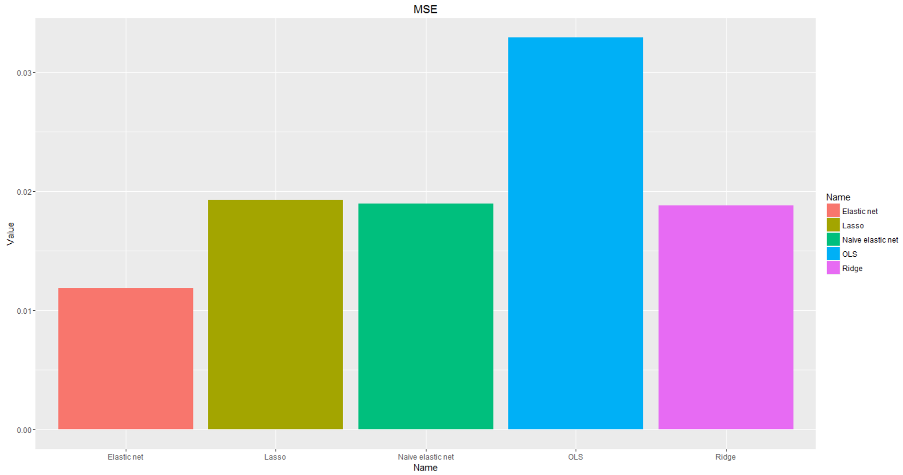

Figure 23: Mean-squared test errors illustrated for different methods. It can be seen that OLS performs worst in terms of prediction accuracy

The mean-squared test errors of the models above are also illustrated in comparison with OLS in Figure 23. The results show that while the elastic net produces a model with fewer variables, its prediction accuracy is higher compared to other estimation methods.

References

[1] David M. Blei. *Regularized Regression.* Columbia University. 2015.

[2] William D. Perreault Charlotte H. Mason and Jr. *Collinearity, Power, and Interpretation of Multiple Regression Analysis.* Journal of Marketing Research, Vol. 28, No. 3 (Aug., 1991), pp. 268-280. 1991.

[3] Peter S.H. Leeflang Csilla Horvath Marcel Kornelis. *What marketing scholars should know about Time Series Analysis: Time Series applications in marketing.* 2002.

[4] James G. MacKinnon Davidson Russell. *Estimation and Inference in Econometrics.* 1993.

[5] J. Durbin. *Testing for Serial Correlation in Least-Squares Regression When Some of the Regressors are Lagged Dependent Variables.* Econometrica , Vol. 38, No. 3 (May, 1970), pp. 410-421. 1970.

[6] Trevor Hastie Robert Tibshirani Gareth James Daniela Witten. *An Introduction to Statistical Learning: with Applications in R.* Springer Texts in Statistics. 2013.

[7] Henri P. Gavin. *The Levenberg-Marquardt method for nonlinear least squares curve-fitting problems.* 2016.

[8] Alan J. Izenman. *Modern Multivariate Statistical Techniques.* Springer Texts in Statistics. 2008.

[9] G. S. Watson J. Durbin. *Testing for Serial Correlation in Least Squares Regression: I.* Biometrika, Vol. 37, No. 3/4 (Dec., 1950), pp. 409-428. 1950.

[10] G. S. Watson J. Durbin. *Testing for Serial Correlation in Least Squares Regression: II.* Biometrika, Vol. 38, No. 1/2 (Jun., 1951), pp. 159-177. 1951.

[11] John DiNardo Jack Johnson. *Econometric methods, Fourth Edition.* 1997.

[12] Peter Kennedy. *A Guide to Econometrics, 6th Edition.* 2008.

[13] Harald Lang. *Elements of Regression Analysis.* 2016.

[14] Wittink D.R. Wedel M. Naert P.A. Leeflang P. *Building Models for Marketing Decisions.* Springer International Series in Quantitative Marketing. 2000.

[15] *Market Response Models: Econometric and Time Series Analysis.* Volume 12. 2001.

[16] *Modeling Markets: Analyzing Marketing Phenomena and Improving Marketing Decision Making.* International Series in Quantitative Marketing. 2015.

[17] Douglas C. Montgomery. *Introduction to Linear Regression Analysis, Fifth Edition.* 2013.

[18] *On the econometrics of the Koyck model.* Springer Texts in Statistics. 2004.

[19] J.B. Ramsey. *Classical model selection through specification tests.* Frontiers in Econometrics. Academic, New York. 1974.

[20] J.B. Ramsey. *Tests for specification errors in classical linear least squares regression analysis.* 1969.

[21] A. R. Pagan T. S. Breusch. *A Simple Test for Heteroscedasticity and Random Coefficient Variation.* Econometrica, Vol. 47, No. 5 (Sep., 1979), pp. 1287-1294. 1979.

[22] Halbert White. *A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.* Econometrica, Vol. 48, No. 4 (May, 1980), pp. 817-838. 1980.

[23] Wayne Winston. *Marketing Analytics: Data-Driven Techniques with Microsoft Excel, 1st Edition.* 2014.

[24] Jeffrey M. Wooldridge. *Introductory Econometrics, a modern approach,5th Edition.* 2012.

[25] Elena Yusupova. *Additive versus Multiplicative Marketing Mix Model.* 2013. URL: `http://analytics.sd-group.com.au/blog/additive-versus-multiplicative-marketing-mix-model/`.

[26] Hui Zou and Trevor Hastie. *Regularization and Variable Selection via the Elastic Net.* URL: `http://web.stanford.edu/~hastie/TALKS/enet_talk.pdf`.

[27] Hui Zou and Trevor Hastie. *Regularization and Variable Selection via the Elastic Net.* Journal of the Royal Statistical Society. Series B (Statistical Methodology), Vol. 67,No. 2 (2005), pp. 301-320. 2005.