DEGREE PROJECT IN THE FIELD OF TECHNOLOGY
ENGINEERING PHYSICS
AND THE MAIN FIELD OF STUDY
MATHEMATICS,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, SWEDEN 2018

Regression Modeling from the Statistical Learning Perspective

with an Application to Advertisement Data

MAX ÖWALL

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ENGINEERING SCIENCES

Abstract

Advertising on social media, and on Facebook in specific, is a global industry from which the social media platforms get their biggest revenues. The performance of these advertisements in relation to the money invested in the advertisement can be measured in the metric cost per thousand impressions (CPM). Various regression modelling strategies combined with statistical learning approaches for model assessment are explored in this thesis with the objective of finding the model that best predicts CPM. Using advertisement data for 540 companies in Sweden during 2017, it is found that the data set comprising of 12 covariates suffers from a high degree of multicollinearity. To tackle this problem efficiently we apply different shrinkage regression methods. Starting from the Ridge and Lasso regression methods, combining the two by an elastic net and then finally expanding Lasso to adaptive Lasso, using cross-validation we find that the elastic net with approximately equal weights on Ridge and Lasso component is the best performing model. In conclusion, when regressing a metric such as CPM, on a set of variables which suffers from severe problems of multicollinearity, the shrinkage regression techniques are needed.

Sammanfattning

Annonsering på sociala medier, och speciellt på Facebook, är en global industri som de sociala medieplatformarna har som största intäktskälla. Hur lyckosamma dessa annonser är i förhållande till hur mycket pengar som investeras i dem kan mätas med nyckeltalet kostnad per tusen intryck (eng: Cost per thousand impressions, CPM). I den här uppsatsen är olika regressionmodeller av statistisk inlärning byggda för prediktering av CPM med syftet att hitta den modell som bäst kan prediktera CPM. Genom att använda 540 företags annonsdata i Sverige under 2017 upptäcks det att de 12 förklaringsvariablerna kraftigt samvarierar varav olika shrinkage regressionsmodeller byggs. Genom att först använda Ridge och Lasso, vilka sen kombineras i ett elastiskt nät och slutligen genom att utvidga Lasso till elastisk Lasso, upptäcks det att den modell som presterar bäst utifrån cross-validation är det elastiska nätet där ungefärligen lika stora vikter läggs på Ridge och Lasso. Slutsatsen är att för att regressera ett nyckeltal som CPM, där det är sannolikt att förklaringsvariablerna samvarierar, är shrinkage regressionsmodeller att föredra.

Acknowledgements

First of all, I would like to thank my supervisor at KTH Royal Institute of Technology, Associate professor of mathematical statistics Tatjana Pavlenko, for the support you gave me in the creation of this thesis. I also want to show my gratitude to Whispr Group, and Axel Martinsson and Dag Strandberg in specific, for always being helpful and staying positive throughout this project.

This thesis marks the end for my studies at KTH Royal Institute of Technology. It has been five fun and challenging years that could not have been completed without the help from family and friends. A big thanks to you!

Stockholm, May 2018
Max Öwall

Abbreviations

CPC	Cost per click
CPI	Cost per impression
CPM	Cost per thousand impressions
CTR	Click through rate
OLS	Ordinary least squares
MSE	Mean squared error
AIC	Akaike information criterion
BIC	Bayesian information criterion
VIF	Variance inflation factor
PCR	Principal component regression
CV	Cross-validation
ANOVA	Analysis Of variance
SS_T	Sum of squared total
SS_R	Sum of squared regression
SS_{Res}	Sum of squared residual
MS_T	Mean squared total
MS_R	Mean squared regression

Notation

$\mathbf{A}$	The matrix $\mathbf{A}$.
$\mathbf{A}^\intercal$	The matrix $\mathbf{A}$ transposed.
$\mathbf{A}^{-1}$	The matrix $\mathbf{A}$ inverted, given that the inverse exists.
$\mathbf{A} \succ 0$	The matrix $\mathbf{A}$ is positive definite.
$\mathbf{a}_j$	The jth column vector of the matrix $\mathbf{A}$.
a_{ij}	The element on the ith row and jth column of $\mathbf{A} \Leftrightarrow$ The element on the ith row of $\mathbf{a}_j$.
$\mathbf{a}$	The vector $\mathbf{a}$.
$\mathbf{a}_i$	The element on the ith row of $\mathbf{a}$.
a	The scalar a.
$\ln(a)$	The natural logarithm of the scalar a.
$\hat{a}$	The estimated value of a from a model.
$\mathcal{A}$	The set $\mathcal{A}$.
$\lvert\mathcal{A}\rvert$	The number of elements in the set $\mathcal{A}$.
$\arg\min_a f(a)$	The value a that minimizes $f(a)$.
$\mathbb{E}[Z]$	The expected value of a random variable Z.
$\mathrm{Var}[Z]$	The variance of a random variable Z.
$\mathrm{SD}[Z]$	The standard deviation of a random variable Z.
m_x	The arithmetic mean of x.
S_x	The standard errror of x.
$Z \sim \mathcal{N}(a, b^2)$	The random variable Z is normally distributed with expected value a and variance b^2.
$Z \xrightarrow{d} \mathcal{N}(a, b^2)$	The random variable Z converges in distribution to the normal distribution.
$Z \sim \chi^2_a$	The random variable Z is χ^2-distributed with a degrees of freedom.
$Z \sim F_{a,b}$	The random variable Z is F-distributed with parameters a and b.
$Z \sim t_a$	The random variable Z is Student t's-distributed with a degrees of freedom.

Contents

List of Figures

List of Tables

1 Introduction

On February 4th 2004, the 20-year-old Harvard student Mark Zuckerberg created one of the world's biggest social media platforms, Facebook, with the objective to "Give people the power to build community and bring the world closer together".[1] What he at that moment did not know, was that he had created a social media platform that would continuously change the way people live all over the world for many years ahead. Facebook contributes to the saying that since social media platforms were launched, people check their smartphone the first thing in the morning and the last thing in the evening. As of December 31st 2017 Facebook has 2.13 billion monthly active users, which is equivalent to 30% of the world's total population.[2] With such many users a big opportunity comes of selling online advertising space.

Advertising on Facebook and in a broader sense advertising on online social media platforms, is a global industry that in many countries is bigger than advertising on television according to Chen, Yu, Guo and Jia (2016) [5]. For Facebook in specific, their primary source of revenue comes from other companies that advertise on their platform. Because of that, Facebook has their own ads manager platform that is used in the process of creating advertisements. The advertising companies are asked when creating advertisements, among many different settings, to provide who their core customers are and how much they are willing to spend on this advertising campaign. From that, Facebook has algorithms that optimize the advertisement in order for it to show up for the right target group. Since the size of the target group can change drastically between different advertisements, key metrics, for instance cost per click (CPC), cost per impression (CPI), cost per thousand impressions (CPM) and click through rate (CTR), will vary.

This thesis aims to examine, from a mathematical statistics point of view, the properties of CPM and what factors contribute to explain that metric. The setup is to explain CPM with different regression models based on statistical learning and from that assess which regression model that best explains CPM. The baseline regression model will be the standard multiple linear regression and other statistical learning models such as Ridge, Lasso, elastic net, adaptive Lasso and principal component regression (PCR) will be extensions of the multiple linear regression.

1.1 In Cooperation with Whispr Group

This thesis is conducted in cooperation with digital insights and strategy partner Whispr Group. Whispr Group is a company of data scientists, business analysts and marketing analytics experts who provide critical digital insights to brand professionals. As of 2018, Whispr Group has offices in

[1] Facebook Inc, "Company Info", 2004
[2] Facebook Inc, "Company Info", 2004

Stockholm, New York and Oslo and provide their services to multinational clients. The business idea is to create solid business strategies based on large-size consumer data from digital and traditional social media.[3]

1.2 Problem Formulation

This thesis aims to construct statistical models to explain CPM on Facebook. CPM can be considered as a metric that describe how much online traffic advertisements generate. The results can hopefully help companies interested in creating awareness for their brand. In the models the dependent variable will be CPM and the covariates include characteristics regarding the advertisement, for instance CPC, CTR and impressions. In terms of regression modelling, different statistical learning regression methods will be tested in order to investigate which model is the most appropriate. See the method section in which this is more explained. The research question is formulated as:

Research question: Which regression model is the best for estimating CPM?

1.3 Purpose

The purpose of this thesis is to create regression models for explaining CPM on Facebook. The data for advertisements on Facebook is granular and exists for many different companies. Hence, there is no lack of data. A similar model cannot be found in an academic paper, as far as I am aware. Furthermore, if the model will be successful, Whispr Group can use it in their daily operations which will be a service with a potential to outperform their competitors.

1.4 Limitations

This analysis builds upon social media data from the calender year 2017 for a subset of Whispr Group's clients. The clients in the data set are those for which Whispr Group have complete data sets of. It cannot be revealed who these clients are because of confidentiality reasons.

1.5 Outline

The thesis starts with providing the mathematical theory used in the models. For a reader with less mathematical knowledge a more basic theoretical framework can be found in the appendix section. The mathematical theory is followed by a literature review in which previous research regarding the subject is outlined. After that, preprocessing of the data follows and an outline of the model building is found. Then the results follow, in which the model building is done, and the performance of the

[3]Whispr Group, "We deliver actionable insights to optimize marketing PR, product development and investment activities", 2015

different models is tested. The thesis ends with discussion, in which the results are discussed, followed by concluding remarks.

2 Mathematical Theory

This section explains the mathematical theory used in this thesis. The theory for the multiple linear regression model is stated in simpler terms in the appendix for a reader with less mathematical knowledge.

2.1 Multiple Linear Regression Models

The multiple linear regression model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.1}$$

with

$$\begin{aligned}
\mathbf{y} &= (y_1, \dots, y_n)^\intercal \in \mathbb{R}^n, \\
\mathbf{X} &= \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \in \mathbb{R}^{n \times p}, \\
\boldsymbol{\beta} &= (\beta_0, \beta_1, \beta_2, \dots, \beta_k)^\intercal \in \mathbb{R}^p, \\
\boldsymbol{\epsilon} &= (\epsilon_1, \dots, \epsilon_n)^\intercal \in \mathbb{R}^n,
\end{aligned} \tag{2.2}$$

where $p = k + 1$. In the ordinary least squares (OLS) approach for estimating $\boldsymbol{\beta}$, using the approach outlined by Montgomery, Peck and Vining (2012) [17], the least-squares function is introduced as

$$S(\boldsymbol{\beta}) = \boldsymbol{\epsilon}^\intercal\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}^\intercal\mathbf{y} - 2\boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{y} + \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta}. \tag{2.3}$$

The approach then builds upon finding the $\boldsymbol{\beta}$ that minimizes $S(\boldsymbol{\beta})$. That is, by solving

$$\arg\min_{\boldsymbol{\beta}} \{S(\boldsymbol{\beta})\} = \arg\min_{\boldsymbol{\beta}} \{\mathbf{y}^\intercal\mathbf{y} - 2\boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{y} + \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta}\}. \tag{2.4}$$

Using a differentiation approach, we get the so called normal equations

$$\begin{aligned}
\left.\frac{\partial S}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\text{OLS}}} &= -2\mathbf{X}^\intercal\mathbf{y} + 2\mathbf{X}^\intercal\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{OLS}} = \mathbf{0}, \\
&\implies \mathbf{X}^\intercal\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{OLS}} = \mathbf{X}^\intercal\mathbf{y}.
\end{aligned} \tag{2.5}$$

We need to have that the matrix $\mathbf{X}^\intercal\mathbf{X}$ is invertible in order to solve equation (2.5) for $\hat{\boldsymbol{\beta}}_{\text{OLS}}$, which is equivalent to that the columns of $\mathbf{X}$ are linearly independent. That means in practise that no data point can be a perfect linear combination of the other data points. Assuming that this holds, we get the OLS-estimate of $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}. \tag{2.6}$$

Hence, combining equations (2.1) and (2.6) we get

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y} = \mathbf{H}\mathbf{y}, \tag{2.7}$$

where the so called hat matrix is defined as

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal. \tag{2.8}$$

Given the Gauss-Markov assumptions, see Hastie, Tibshirani and Friedman (2008) [10], it can be shown that

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\beta}}_{\text{OLS}}] = \mathbb{E}[(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}] = \mathbb{E}[(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon})] = \\ = \underbrace{(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{X}}_{=\mathbf{I}}\boldsymbol{\beta} + (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\underbrace{\mathbb{E}[\boldsymbol{\epsilon}]}_{=\mathbf{0}} = \boldsymbol{\beta},\end{aligned} \tag{2.9}$$

$$\begin{aligned}\text{Var}[\hat{\boldsymbol{\beta}}_{\text{OLS}}] = \text{Var}[(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon})] = \text{Var}[\boldsymbol{\beta} + (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\epsilon})] = \\ = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\text{Var}[\boldsymbol{\epsilon}][(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal]^\intercal = \underbrace{(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\underbrace{\sigma^2\mathbf{I}}_{=\mathbf{X}\sigma^2}\mathbf{X}}_{=\sigma^2\mathbf{I}}(\mathbf{X}^\intercal\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}.\end{aligned} \tag{2.10}$$

Thus $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ is unbiased, if the model is correct, and if the covariates are orthogonal, which gives that $(\mathbf{X}^\intercal\mathbf{X})^{-1}$ is a diagonal matrix, then the estimates of β_i and β_j are uncorrelated. If the covariates are not orthogonal, which in practise almost always is the case, then the variance of the estimates increases as the degree of multicollinearity increases. It was shown in Hastie et al. (2008) [10] that $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ according to equation (2.6) is the unbiased estimate with the smallest variance. Nonetheless, there are estimators of $\boldsymbol{\beta}$ with smaller variance that have some bias according to Montgomery et al. (2012) [17]. This trade-off can be visualized in the mean squared error (MSE) by utilizing the definition of variance

$$\text{MSE}[\hat{\boldsymbol{\beta}}] = \mathbb{E}[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^2] = \text{Var}[\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}] + \Big(\mathbb{E}[\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}]\Big)^2 = \text{Var}[\hat{\boldsymbol{\beta}}] + \underbrace{\Big(\mathbb{E}[\hat{\boldsymbol{\beta}}]-\boldsymbol{\beta}\Big)^2}_{=(\text{bias in }\hat{\boldsymbol{\beta}})^2}. \tag{2.11}$$

2.2 Model Improvements

In this section improvements to the OLS regression are described. Improvements are limited to transformations, analysis of residuals and problems of multicollinearity.

2.2.1 Transformations

A transformation of a covariate might be needed if it turns out that there is not a linear trend of $\mathbf{y}$ with that covariate. The easiest way to spot this is simply by plotting $\mathbf{y}$ against $\mathbf{x}_j$ and from that

try to spot the behaviour of a fitted function. For instance, if there is a clear quadratic behaviour of $\mathbf{y}$ with a specific $\mathbf{x}_j$ and a linear behaviour of $\mathbf{y}$ with every other $\mathbf{x}_j$, then $\mathbf{X}$ in (2.2) is updated to

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1(j-1)} & x_{1j}^2 & x_{1(j+1)} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2(j-1)} & x_{2j}^2 & x_{2(j+1)} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n(j-1)} & x_{nj}^2 & x_{n(j+1)} & \cdots & x_{nk} \end{pmatrix}. \tag{2.12}$$

The conclusion from this is that the only requirement in terms of linearity, is that the model is linear with the transformed data. If one believes that the true model is not linear but in fact a product of the covariates, which is in practise quite common, the following transformation is adequate

$$\begin{aligned} y_i &= \beta_0 \Big(\prod_{j=1}^{k} e^{\beta_j x_{ij}} \Big) \epsilon_i, \\ \implies \ln y_i &= \ln \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \ln \epsilon_i, \end{aligned} \tag{2.13}$$

and the transformed model is linear. It does of course exist many other linearizations that can be used by basic calculus as outline by Montgomery et al. (2012) [17].

One of the assumptions for linear regression is that the dependent variable is normally distributed. There are of course cases where this assumption does not hold and a solution to the problem is needed. A common solution is the power transformation, in which the dependent variable is taken to a power λ. Then, the most optimal value of λ can be found from maximizing the likelihood as a function of λ. However, problems arise when $\lambda = 0$ since then all transformed data will be identically $= 1$. In Box and Cox (1964) [3], the Box-Cox transformation, named after its founders, was suggested as

$$y' = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \ln y, & \text{if } \lambda = 0, \end{cases} \tag{2.14}$$

where λ is a parameter to be determined such that the transformed data obtains the wanted characteristics. Statistical software, as R, has built-in functions to determine λ by maximizing the likelihood as a function of λ. To assess normality of the data both the original and transformed data can be plotted in histograms to see which data that fits the normality best.

2.2.2 Influential and Leverage Points

To begin with, we must distinguish between the difference of a leverage point and an influential point. A leverage point is defined in Montgomery et al. (2012) [17] as a data point that has an unusual combination of the covariates, or in the case with only one covariate, a very high or low x-value. An influential point on the other hand is defined as a data point that heavily influences the estimated

coefficients, and it should consequently be questioned whether that data point is correctly measured. The difference between leverage and influential point is visualized as:

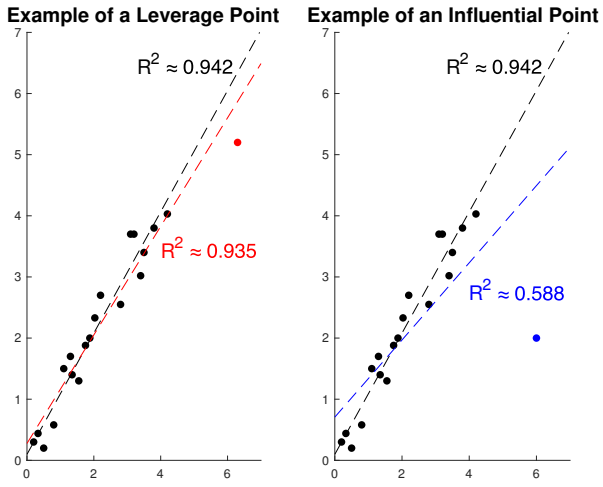


Figure 2.1: An example of a leverage point and an influential point. The black dashed line is the fitted curve without the leverage and influential point. The red and blue dashed lines are with the leverage and influential point respectively.

Consequently, influential points need to be detected in order to estimate the correct model. The metric Cook's Distance D measures the influence of each data point i [17]

$$\text{Cook's Distance:} \qquad D_i = \frac{\left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)^{\intercal} \mathbf{X}^{\intercal}\mathbf{X}\left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)}{p \cdot MS_{Res}}. \tag{2.15}$$

An alternative metric that lead to the same conclusion is

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{S^2_{(i)} H_{ii}}}, \tag{2.16}$$

where S^2 is an estimate of standard error squared and the notation $a_{(i)}$ means that data point i is excluded in the regression. $D_i > 1$ is usually considered an influential point, whereas according to Belsley, Kuh and Welsch (1980) [2], if $|DFFITS_i| > 2\sqrt{p/n}$ then the ith observation needs to be extra investigated. Observe that Cook's Distance and $DFFITS$ exist for each of the n data points.

2.2.3 Multicollinearity

It is possible that the covariates included in equation (2.1) suffer from high degree of multicollinearity, which will lead to an unstable model in the sense that the standard errors are large and a change in one data point will lead to a large change in the estimate. There exists many ways of testing for multicollinearity, and variance inflation factor (VIF) is one of them. VIF is calculated by first

regressing $\mathbf{x}_j$ against the other covariates. That is, run the regression

$$x_{ij} = \beta_0 + \sum_{j'=1, j' \neq j}^{k} \beta_{j'} x_{ij'} + \epsilon_i, \tag{2.17}$$

and then obtain the R_j^2, as equation (8.12) instructs, from that regression. The VIF for covariate j is then defined as

$$\text{VIF}_j = \frac{1}{1 - R_j^2}. \tag{2.18}$$

If R_j^2 is high, implying that there is a strong linear relationship of $\mathbf{x}_j$ with one or many of the other covariates, VIF_j will conversely be large which is an evidence of multicollinearity. According to Montgomery et al. (2012) [17] the model is said to suffer from severe problems of multicollinearity if one or more VIFs are larger than 10. A cutoff value of 5 can also be used for detection, which will be the case in this thesis. The model should be adjusted if that is the case, with one option being to exclude the covariate with the highest VIF and test if the problem of multicollinearity is improved by that adjustment.

Another sign of the degree of multicollinearity can be found by an analysis of eigenvalues. First, scale each covariate such that the matrix $\mathbf{X}^\intercal\mathbf{X}$ is in correlation form and then find the eigenvalues of $\mathbf{X}^\intercal\mathbf{X}$, i.e. the roots λ to the equation $\det(\mathbf{X}^\intercal\mathbf{X} - \lambda\boldsymbol{I}) = 0$. In the case of high multicollinearity, one or more of the roots will be small in relation to the largest eigenvalue. Therefore, the condition number κ is defined as

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}}. \tag{2.19}$$

The model suffers from a high degree of multicollinearity if $\kappa > 100$ according to Montgomery et al. (2012) [17]. It exists even more metrics and methods to spot high degrees of multicollinearity. However, this thesis will only incorporate VIF and conditional number to assess whether there is a problem.

2.3 Method Validation

This section will explain statistical learning methods for validating and choosing the best regression model.

2.3.1 k-Fold Cross-validated Training and Test Error

When estimating a model and one wants to assess that model in comparison to another similar model, the performances need to be tested in some way. As described in Hastie et al. (2008) [10], a common method for that matter is to first shuffle and randomly split the original data in a training set, with for instance 2/3 of the original data, and a test set, with the remaining 1/3 of the data. The model is

then estimated only with the use of the training data and its prediction performance is tested on the test data. The performance metrics can then be compared between the different models to conclude which model that performs the best. The error metric that will be used in this thesis is mean square error.

Even though it seems straightforward to randomly split the data in a test and training set, there is a risk that some models perform better on a certain choices of data set. k-Fold cross-validation (CV) is a method to avoid such problems. At first, the original data set is randomly divided in k subsets of data of equal size. Then, k models are estimated using $k-1$ subsets as training data and testing the prediction performance on the last kth subset that function as test data. Consequently k MSEs are collected, which then are averaged to get the cross-validated MSE of the model.

Of course the parameter k must be decided before. Cross-validation is a computer-intensive statistical learning method, consequently it is unpractical in terms of implementation if choosing k too large. In this thesis k will be chosen to 10 throughout.

2.3.2 Variable Selection

Even though having access to a data set of k covariates, the most suitable model in terms of prediction may not incorporate all available covariates. Also, it may be tedious to interpret a model that includes all covariates if k is large. Because of these reasons, the best model may not incorporate all available covariates according to Hastie et al. (2008) [10]. There are in fact 2^k-1 combinations of covariates that could create a model in the case of k available covariates. The term -1 comes from the requirement of including at least one covariate in the model. If k is large, one relies heavily on the computer's power to find the model estimates in order to compare them. Common metrics to compare models in this analysis are $R^2_{Adj.}$, Akaike information criterion (AIC) and MSE.

2.4 Shrinkage Regression Methods

Shrinkage regression methods are modified versions of the OLS-regression that shrinks the coefficients. This section starts off from the common shrinkage methods Ridge and Lasso and is then followed by combinations and extensions of those two.

2.4.1 Ridge

In this case of modelling CPM, with high-dimensional data from many available metrics, a risk of having high degree of multicollinearity is prevalent. Without loss of generality, assume a linear model without intercept and two covariates where all covariates are scaled to correlation form. In this case,

the setup and solution of equation (2.5) is

$$\begin{aligned} \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} &= \begin{pmatrix} r_{1y} \\ r_{1y} \end{pmatrix}, \\ \implies \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} &= \frac{1}{1-r_{12}^2} \begin{pmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{pmatrix} \begin{pmatrix} r_{1y} \\ r_{1y} \end{pmatrix}, \end{aligned} \tag{2.20}$$

with r_{12} being the sample correlation between covariate 1 and 2. This shows that if there is a high degree of multicollinearity, or in other words that r_{12} is large, then using equation (2.9) shows that the variance increases, giving unstable estimates. Note that this is only an example in the 2-dimensional case, however the same problem of multicollinearity also prevails in a multidimensional case. Hence, OLS-regression with presence of high degree of multicollinearity may not be appropriate according to Montgomery et al. (2012) [17].

Having the bias-variance trade-off from equation (2.11) in mind, an approach that is less penalizing for the variance is the Ridge regression estimates of $\boldsymbol{\beta}$. The estimate arises from solving the modified normal equation with a chosen $\lambda \geq 0$ according to Hastie et al. (2008) [10], that is

$$\begin{aligned} (\mathbf{X}^\intercal \mathbf{X} + \lambda \mathbf{I}) \hat{\boldsymbol{\beta}}_{\text{Ridge}} &= \mathbf{X}^\intercal \mathbf{y}, \\ \implies \hat{\boldsymbol{\beta}}_{\text{Ridge}} &= (\mathbf{X}^\intercal \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\intercal \mathbf{y}. \end{aligned} \tag{2.21}$$

Equation (2.21) comes from solving a modified version of equation (2.4) with a bounded feasible region

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{Ridge}} &= \arg\min_{\boldsymbol{\beta}} \left\{ \mathbf{y}^\intercal \mathbf{y} - 2\boldsymbol{\beta}^\intercal \mathbf{X}^\intercal \mathbf{y} + \boldsymbol{\beta}^\intercal \mathbf{X}^\intercal \mathbf{X} \boldsymbol{\beta} \right\}, \\ &\text{subject to } \sum_{j=1}^{k} \beta_j^2 \leq d^2. \end{aligned} \tag{2.22}$$

Equation (2.22) is a quadratic optimization program where the feasible region, according to Griva, Nash and Sofer (2009) [9], can be relaxed using Lagrangian relaxation and the parameter $\lambda \geq 0$. An equivalent formulation of equation (2.22) is then

$$\hat{\boldsymbol{\beta}}_{\text{Ridge}} = \arg\min_{\boldsymbol{\beta}} \left\{ \mathbf{y}^\intercal \mathbf{y} - 2\boldsymbol{\beta}^\intercal \mathbf{X}^\intercal \mathbf{y} + \boldsymbol{\beta}^\intercal \mathbf{X}^\intercal \mathbf{X} \boldsymbol{\beta} + \lambda \boldsymbol{\beta}^\intercal \boldsymbol{\beta} \right\} = \arg\min_{\boldsymbol{\beta}} \left\{ \mathbf{y}^\intercal \mathbf{y} - 2\boldsymbol{\beta}^\intercal \mathbf{X}^\intercal \mathbf{y} + \boldsymbol{\beta}^\intercal (\mathbf{X}^\intercal \mathbf{X} + \lambda \mathbf{I}) \boldsymbol{\beta} \right\}. \tag{2.23}$$

To obtain the solution in equation (2.21), a differentiating approach as for solving for the OLS-estimates is used. The parameter λ is called by many different names, for instance Ridge parameter and Lagrangian multiplier. There is a relationship between λ and d, which according to Hastie et al. (2008) [10] is a one-to-one relationship. This implies that there is a unique solution for the optimization program for each introduction of feasible region and therefore, the Ridge estimate is uniquely determined by λ. Furthermore, it is easy to see that OLS regression is the special case of

Ridge regression by letting $\lambda \to 0$. The key characteristics with Ridge regression are

$$\begin{aligned} \mathbb{E}[\hat{\boldsymbol{\beta}}_{\text{Ridge}}] &= (\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^\intercal\mathbf{X})\boldsymbol{\beta}, \\ \text{Var}[\hat{\boldsymbol{\beta}}_{\text{Ridge}}] &= \sigma^2(\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\intercal\mathbf{X}(\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbf{I})^{-1}. \end{aligned} \tag{2.24}$$

Hence, the Ridge estimate is biased and the variance is smaller than for OLS since $\lambda \geq 0$, because there is a possibility to obtain a lower MSE. Also, the bias increases and the variance decreases for larger values of λ. It can be seen from equation (2.21) that

$$\hat{\boldsymbol{\beta}}_{\text{Ridge}} = (\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\intercal\mathbf{y} = (\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\intercal\mathbf{X}\underbrace{(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}}_{\hat{\boldsymbol{\beta}}_{\text{OLS}}} = \left(\mathbf{I} + \lambda(\mathbf{X}^\intercal\mathbf{X})^{-1}\right)^{-1}\hat{\boldsymbol{\beta}}_{\text{OLS}}. \tag{2.25}$$

Equation (2.25) gives two interesting results. The first is that the Ridge estimate is a linear combination of the OLS-estimate. More importantly, the estimates will shrink in absolute value towards the origin as λ increases. This is why Ridge regression, and other similar methods, are called shrinkage regression methods. Furthermore, as λ handles how much shrinkage to include in the model, another commonly used name of λ is shrinkage parameter. Now, a key feature in Ridge regression is how one chooses λ. The common way of doing this is to find the k-folded cross-validated mean squared error as a function of λ. The most suitable λ is the λ that minimizes the mean squared error.

2.4.2 Lasso

Another shrinkage method is Lasso regression, in which Lasso stands for Least Absolute Shrinkage and Selection Operator. Lasso has shrinkage and variable selection features, as indicated by its name. The Lasso estimate is found by solving another modification of equation (2.4) in which the feasible region is of type L_1 instead of L_2

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{Lasso}} = \arg\min_{\boldsymbol{\beta}} &\left\{\mathbf{y}^\intercal\mathbf{y} - 2\boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{y} + \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta}\right\}, \\ &\text{subject to } \sum_{j=1}^{k}|\beta_j| \leq t. \end{aligned} \tag{2.26}$$

Equation (2.26) is a nonlinear optimization program with well-defined feasible region and, according to Griva et al. (2009) [9], it can be equivalently written with a Lagrangian multiplier $\lambda \geq 0$ as

$$\hat{\boldsymbol{\beta}}_{\text{Lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{\mathbf{y}^\intercal\mathbf{y} - 2\boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{y} + \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} + \lambda\sum_{j=1}^{k}|\beta_j|\right\}. \tag{2.27}$$

Solving equation (2.27) does not provide a closed form expression as for solving equation (2.23). Instead, one could rely on optimization algorithms for quadratic programming to solve the program in equation (2.27). However, another algorithm introduced is introduced in Hastie et al. (2008) [10], called Least Angle Regression with Lasso Modification, in which a complex quadratic program does not have to be solved:

1. Set all regression coefficients equal to zero: $\hat{\boldsymbol{\beta}} = \mathbf{0}$. Start with the residual $\mathbf{e} = \mathbf{y} - \bar{\mathbf{y}}$.

2. Find the covariate $\mathbf{x}_j$ that is most correlated with the residual.

3. Increase the coefficient β_j in the direction of the sign of the correlation between the residual and $\mathbf{x}_j$ and collect new residuals along the way. Continue to increase β_j until another covariate $\mathbf{x}_k$ correlates equally much with the residual as $\mathbf{x}_j$ does.

4. Increase the coefficients β_j and β_k in the direction of their joint least squares coefficient of the current residual, until another covariate $\mathbf{x}_m$ correlates equally much with the residual as $\mathbf{x}_j$ and $\mathbf{x}_k$.

    - If a non-zero coefficient passes zero, drop the covariate corresponding to that coefficient from the active set of covariates and continue the algorithm.

5. Continue until all covariates have entered to the model.

In terms of implementation, the proper way of finding the optimal shrinkage parameter λ is to cross-validate different values of λ and then finding the λ that minimizes the mean-squared error. λ in Lasso regression has the same shrinkage interpretation as for Ridge, however the scaling of the parameter might not correspond. Graphically, estimating Ridge and Lasso from the bias-variance trade off can be seen as

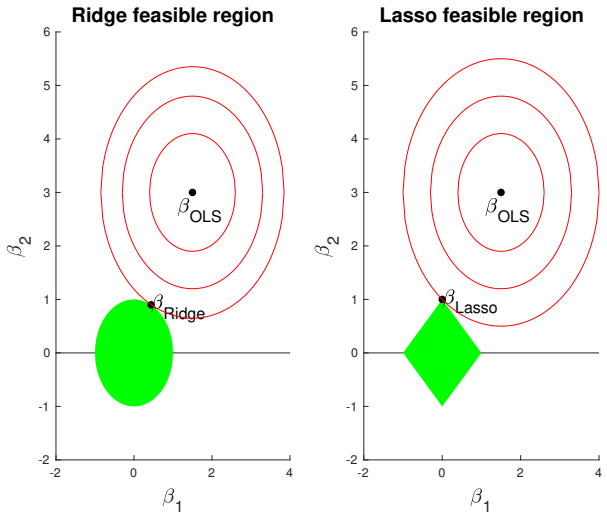


Figure 2.2: The green region represents the feasible region for the Ridge and Lasso optimization program respectively for a regression model with two explanatory variables. The red contours represents different level curves of the objective function $S(\boldsymbol{\beta})$. The estimate of $\boldsymbol{\beta}$ will be found on the edge of the feasible region, as expected by optimization theory.

Lasso regression possesses the same penalizing effects of bias, rewarding effects of variance, the one-to-one relationship between feasible region and shrinkage parameter and similar shrinkage capabilities

as Ridge regression. However, more and more coefficients will be identically zero in Lasso regression for increasing values of λ according to Hastie et al. (2009) [10]. According to Griva et al. (2009) [9], this is because the solution to the optimization program will occur at the extreme points of the feasible region, which are at the corners of the feasible region in figure 2.2, since the feasible region is not differentiable at the corners. Thus, Lasso regression also performs inevitable variable selection. Because of these reasons Lasso regression can be suitable in a case with high degree of multicollinearity and/or high-dimensional data.

Define the true support as

$$\mathcal{A} = \{j : j = [1, \dots, k], \quad \beta_j \neq 0\}, \tag{2.28}$$

where each β_j is from the true model. Now, assume $|\mathcal{A}| = k_0 < k$, or in other words that one or more of the true predictor coefficients are identically zero. Introduce the support of $\hat{\boldsymbol{\beta}}$ as $\hat{\mathcal{A}} = \{j : j = [1, \dots, k], \quad \hat{\beta}_j \neq 0\}$. Since Lasso performs variable selection, we want the Lasso estimator to perform the correct estimation. That is, we want

$$\hat{\mathcal{A}} = \mathcal{A}, \tag{2.29}$$

with a high probability. This is in practise too ambitious, as Bühlman (2017) [4] showed that very strong necessary conditions need to be satisfied. The work of Bühlmann (2017) [4] also showed that when relaxing these strong necessary conditions for obtaining equation (2.29), one instead gets an effective but not as ambitious property for Lasso stating that

$$\hat{\mathcal{A}} \supseteq \mathcal{A}. \tag{2.30}$$

Equation (2.30) tells that Lasso estimation does not set some coefficients to zero when they in fact should be non-zero. On the other hand, Lasso has the potential drawback of not setting sufficiently many coefficients to identically zero.

2.4.3 Elastic Net

According to Hastie et al. (2008) [10], both Ridge and Lasso regression can be said to be special cases of the more general L_q optimization program

$$\hat{\boldsymbol{\beta}}_q = \arg\min_{\boldsymbol{\beta}} \left\{ \mathbf{y}^\intercal \mathbf{y} - 2\boldsymbol{\beta}^\intercal \mathbf{X}^\intercal \mathbf{y} + \boldsymbol{\beta}^\intercal \mathbf{X}^\intercal \mathbf{X} \boldsymbol{\beta} + \lambda \sum_{j=1}^{k} |\beta_j|^q \right\}, \tag{2.31}$$

where $q = 2$ and $q = 1$ are equivalent to Ridge and Lasso respectively and the parameter $q \geq 0$. In terms of a feasible region, the estimate can be found by solving equation (2.4) with feasible region of

type L_q

$$\begin{aligned} &\arg\min_{\boldsymbol{\beta}} \left\{ \mathbf{y}^\intercal\mathbf{y} - 2\boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{y} + \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} \right\} \\ &\text{subject to } \sum_{j=1}^{k} |\beta_j|^q \leq t. \end{aligned} \tag{2.32}$$

Also, if $q = 0$ then

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{q=0} =& \arg\min_{\boldsymbol{\beta}} \left\{ \mathbf{y}^\intercal\mathbf{y} - 2\boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{y} + \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} + \lambda \underbrace{\sum_{j=1}^{k} |\beta_j|^0}_{=k, \text{ independent of } \boldsymbol{\beta}} \right\} = \\ =& \arg\min_{\boldsymbol{\beta}} \left\{ \mathbf{y}^\intercal\mathbf{y} - 2\boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{y} + \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} \right\} = \hat{\boldsymbol{\beta}}_{\text{OLS}}. \end{aligned} \tag{2.33}$$

The direct question here is how one should choose q, or in other words how one should define the feasible region. The following figure shows the feasible region in two dimensions for four different values of q (which could be generalized to higher dimensions):

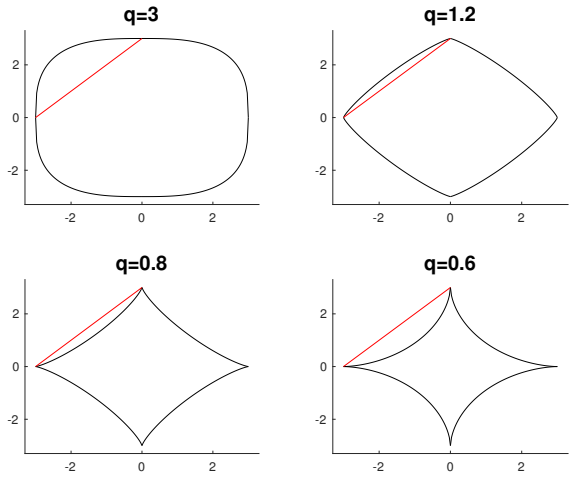


Figure 2.3: The feasible region for the general optimization program of the L_q estimate. The straight red line is plotted to indicate what values of q that yield convex feasible regions and therefore convex optimization programs.

A region Ω, as defined by Griva et al. (2009) [9], is convex if $x, y \in \Omega$ and $\delta x + (1-\delta)y \in \Omega, \forall \delta \in [0, 1]$. It can clearly be seen that the red line leaves the feasible region, if choosing $\delta = 1/2$ for $q = 0.8$ and $q = 0.6$. Consequently, as stated in Hastie et al. (2008) [10], concave optimization program arises for values of $q < 1$, which are far more complicated to solve than their convex counterparts. q is for these reasons restricted to $q \geq 1$. Finding the appropriate value of q can often be done from the specific data set, however according to Hastie et al. (2008) [10] "it is not worth the effort for the extra variance incurred". They instead suggest to choose $q \in [1, 2]$, i.e. to choose a compromise between

Ridge and Lasso. In this approach, one should pay attention to how the variable selection features of Lasso evolves in the case of combining Ridge and Lasso. In fact, the variable selection feature does completely disappear if $q > 1$ since the feasible region is then differentiable at all points as stated by Hastie et al. (2008) [10]. To still have the variable selection feature of Lasso and still combine Lasso with Ridge the elastic net feasible region was defined by Zou et al. (2005) [22] as:

$$\sum_{j=1}^{k} ((1-\alpha)\beta_j^2 + \alpha|\beta_j|) \leq d^2, \quad \alpha \in [0,1], \tag{2.34}$$

which can be interpreted as a weighted split between Ridge ($\alpha = 0$) and Lasso ($\alpha = 1$). The elastic net is then used as a penalizing term to obtain the elastic net estimate

$$\hat{\boldsymbol{\beta}}_{\text{Elastic net}} = \arg\min_{\boldsymbol{\beta}} \left\{ \mathbf{y}^\intercal\mathbf{y} - 2\boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{y} + \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} + \lambda \sum_{j=1}^{k} ((1-\alpha)\beta_j^2 + \alpha|\beta_j|) \right\} \tag{2.35}$$

In this case, one do not need an advanced method to choose a parameter q. What is needed is a choice of α, i.e. a decision upon how much weight the estimate should put on Ridge in comparison to Lasso. The following figure shows that for a given choice of $1 < q < 2$ in the L_q estimate, the feasible region can to a large extent be replicated by the elastic net:

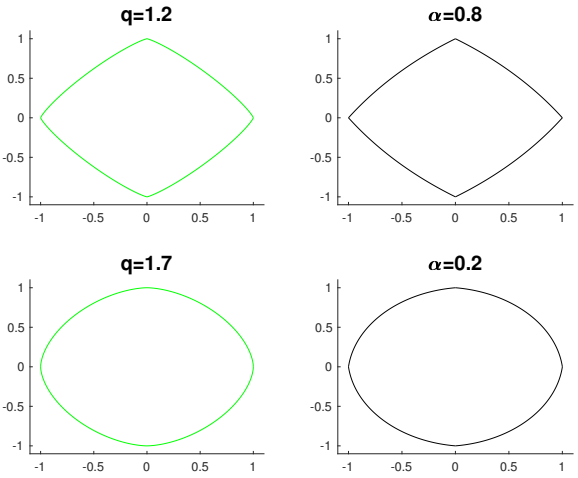


Figure 2.4: The green plots are for the L_q feasible region and black are for the elastic net. The feasible regions for the upper and lower row respectively appear to be visually similar. However, note the sharp corners in elastic net due to the variable selection features of Lasso.

Figure 2.4 shows that the L_q feasible region can be replicated by the elastic net and by still keeping the variable selection features of Lasso. Since the elastic net is not differentiable at the corners of its feasible region, by recalling for instance that $f(x) = |x|$ is not differentiable at $x = 0$ as shown in Persson and Böiers (2010) [18], the variable selection features of Lasso will be kept by all choices of α.

2.4.4 Adaptive Lasso

A desirable property for an estimator of $\hat{\boldsymbol{\beta}}(\delta)$ is the support recovery and screening property. In order to define that property as done in Fan and Li (2001) [8], let the data be mean-centered such that the intercept for the true model is neglected. Let the Lasso estimate in equation (2.27), as a function of the number of data points n, be defined as

$$\hat{\boldsymbol{\beta}}^{(n)}_{\text{Lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \mathbf{y}^\intercal \mathbf{y} - 2\boldsymbol{\beta}^\intercal \mathbf{X}^\intercal \mathbf{y} + \boldsymbol{\beta}^\intercal \mathbf{X}^\intercal \mathbf{X} \boldsymbol{\beta} + \lambda_n \sum_{j=1}^{k} |\beta_j| \right\}. \tag{2.36}$$

That implies, of course, that the Lagrangian multiplier λ_n is a function of n. Also let $\mathcal{A}_n = \{j : j = [1, \dots, k], \quad \hat{\beta}^{(n)}_j \neq 0\}$. The support recovery and screening property, as defined by Fan et al. (2001) [8], is satisfied by a procedure δ if $\hat{\boldsymbol{\beta}}(\delta)$ satisfies

- $\lim_{n\to\infty} \mathbb{P}(\mathcal{A}_n = \mathcal{A}) = 1$, i.e. the model identifies in limit the right subset of coefficients almost surely,
- $\sqrt{n}(\hat{\boldsymbol{\beta}}(\delta) - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the covariance matrix for the true model, i.e. the model has the right convergence rate.

The first condition in this oracle procedure is denoted as that the variable selection is consistent. Now, without loss of generality assume a linear regression model according to equation (2.1) and that

$$\begin{aligned} \mathcal{A} &= \{1, 2, \dots, k_0\} \\ , \lim_{n\to\infty} \frac{1}{n} \mathbf{X}^\intercal \mathbf{X} = \mathbf{C} &= \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix} \succ 0, \qquad \mathbf{C}_{11} \in \mathbb{R}^{k_0 \times k_0}. \end{aligned} \tag{2.37}$$

As shown in Zou and Hastie (2006) [21], a necessary condition for consistency is that there exists a vector $\mathbf{s} = (\pm 1, \dots, \pm 1)^\intercal \in \mathbb{R}^{k_0}$ such that

$$\left| \mathbf{C}_{21} \mathbf{C}_{11}^{-1} \mathbf{s} \right| \leq 1 \quad \text{(Interpreted componentwise)}. \tag{2.38}$$

Thus, an estimation procedure is inconsistent if condition (2.38) fails. It was shown in Zou (2006) [21] that Lasso estimation is an inconsistent variable selection procedure and consequently not an support recovery and screening procedure. Therefore, Zou (2006) [21] suggested a modified Lasso regression which is shown to satisfy the support recovery and screening procedure, called adaptive Lasso that is defined as

$$\hat{\boldsymbol{\beta}}^{(n)}_{\text{Adaptive Lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \mathbf{y}^\intercal \mathbf{y} - 2\boldsymbol{\beta}^\intercal \mathbf{X}^\intercal \mathbf{y} + \boldsymbol{\beta}^\intercal \mathbf{X}^\intercal \mathbf{X} \boldsymbol{\beta} + \lambda_n \sum_{j=1}^{k} \hat{w}_j |\beta_j| \right\}, \tag{2.39}$$

where the vector $\hat{\mathbf{w}}$ is defined as $\hat{\mathbf{w}} = 1/|\hat{\boldsymbol{\beta}}_{\text{OLS}}|^\gamma$ for a chosen $\gamma > 0$. For computational matters, in Efron, Hastie, Johnstone and Tibshirani (2004) [7] the LARS algorithm was suggested for finding the adaptive Lasso estimate:

1. Let $\mathbf{x}'_j = \mathbf{x}_j/\hat{w}_j, \quad j = 1, \ldots, k.$

2. Find the Lasso regression for

$$\hat{\boldsymbol{\beta}}'_{\text{Lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \mathbf{y}^\intercal\mathbf{y} - 2\boldsymbol{\beta}^\intercal\mathbf{X}'^\intercal\mathbf{y} + \boldsymbol{\beta}^\intercal\mathbf{X}'^\intercal\mathbf{X}'\boldsymbol{\beta} + \lambda\sum_{j=1}^{k}|\beta_j| \right\}. \tag{2.40}$$

3. Compute $\hat{\beta}_{(\text{Adaptive Lasso, j})} = \hat{\beta}'_{(\text{Lasso, j})}/\hat{w}_j$.

The LARS algorithm contains two parameters that need to be estimated, namely λ and γ. A two-dimensional cross-validation approach to find the most appropriate choices of coefficients is suggested here by Zou et al. (2006) [21]. Furthermore, the adaptive Lasso can be slightly modified by choosing other consistent estimators than $\hat{\boldsymbol{\beta}}_{\text{OLS}}$. The cross-validation would in that case be three-dimensional and the computational complexity would increase. In this thesis the OLS will be the only choice of consistent estimator for finding adaptive Lasso estimates.

2.5 Derived Inputs Regression: PCR

PCR can help to solve the multicollinearity problem, if such a problem would occur, by reducing the dimension of the covariates as stated by Montgomery et al. (2012) [17]. The idea with PCR is to derive new covariates as linear combinations of the existing covariates such that the derived data will be orthogonal which completely remove the problem of multicollinearity.

Montgomery et al. (2012) [17] stated that PCR will yield biased estimates with less covariates included in the final model, and thus a model easier to interpret. To begin with, let the data in $\mathbf{X}$ be mean-centered. Since $\mathbf{X}^\intercal\mathbf{X}$ is a scaled version of Var[$\mathbf{X}$], the matrix $\mathbf{X}^\intercal\mathbf{X}$ is symmetric and then according to the spectral theorem of linear algebra $\mathbf{X}^\intercal\mathbf{X}$ is diagonalizable. Let $\mathbf{T}$ be the matrix of eigenvectors and $\boldsymbol{\Lambda}$ be the diagonal matrix of corresponding eigenvalues to $\mathbf{X}^\intercal\mathbf{X}$ such that

$$\mathbf{T}^\intercal(\mathbf{X}^\intercal\mathbf{X})\mathbf{T} = \boldsymbol{\Lambda}. \tag{2.41}$$

Without loss of generality, arrange the eigenvalues such that $\boldsymbol{\Lambda}$ =diag$(\lambda_1, \ldots, \lambda_p)$ with $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_p|$ and define $\mathbf{T}$ from that. Now, introduce

$$\mathbf{Z} = \mathbf{X}\mathbf{T}, \quad \boldsymbol{\alpha} = \mathbf{T}^\intercal\boldsymbol{\beta}. \tag{2.42}$$

From this, introduce the model

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\alpha} + \epsilon, \tag{2.43}$$

and regress, according to equation (2.6), the OLS-estimate as

$$\hat{\boldsymbol{\alpha}} = (\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal\mathbf{y}. \tag{2.44}$$

Now comes the procedure of choosing what principal components to include in the analysis. It is suggested in the work of Hult, Lindskog, Hammarlid and Rehn (2012) [11] to regard the quantity

$$\frac{\sum_{j=1}^{s} \lambda_j}{\sum_{j=1}^{p} \lambda_j}, \tag{2.45}$$

as a function of s. As the quotient is approaching 1, the last eigenvalues and those corresponding components will be left out of the model. Empirically, it is often the case that the quotient is close to 1 for a relatively small value of s, according to Hult et al. (2012) [11]. A similar quotient is suggested in Izenman et al. (2008) [12]

$$1 - \frac{\sum_{j=1}^{s} \lambda_j}{\sum_{j=1}^{p} \lambda_j}, \tag{2.46}$$

which yields the same results. s is then chosen accordingly and from that $\hat{\boldsymbol{\alpha}}$ is modified to

$$\hat{\boldsymbol{\alpha}}_{\mathrm{PCR}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_s, 0, \ldots, 0)^{\intercal} \in \mathbb{R}^p. \tag{2.47}$$

It can here be seen that the nonzero elements of $\hat{\boldsymbol{\alpha}}_{\mathrm{PCR}}$ will be less than the dimension of $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$, which implies that PCR reduces the dimension. By using equation (2.47) the estimated regression coefficients are then calculated as

$$\hat{\boldsymbol{\beta}}_{\mathrm{PCR}} = \mathbf{T}\hat{\boldsymbol{\alpha}}_{\mathrm{PCR}}. \tag{2.48}$$

Furthermore, another approach of deciding how many components to include in the final model is to cross-validate the mean squared error for different number of components in the PCR, and from that choose an optimal number of components.

3 Literature Review

The current state of knowledge regarding advertising on social media platforms is quite wide, in the sense that it exists research of a lot of different topics. The research ranges across machine learning and neural networks to factors for determining a winning advertisement from a business perspective. Nonetheless, specific models within mathematical statistics for predicting CPM on social media platforms do not yet exist to my knowledge as academic papers..

As explained in the introduction, advertising on social media platforms is done by describing who the core customer is in order for the advertisement to be shown. As described in Liu, Kliman-Silver, Bell, Krishnamurthy and Mislove (2014) [15], factors of interest are for example location, gender, age and relationship status. In that paper, they find critique for the algorithm used by Facebook for finding the right target group, by stating that the algorithm only uses a small sample of the data. They also lift the problem that the owners of the data are the social media platforms and not the advertising companies which makes it hard to verify what is actually successful in campaigns. Regarding metrics for advertisement, it is explained in Liu et al. (2014) [15] that advertisers can choose the minimum, median and maximum CPM for the campaign when posting it to Facebook.

Even though advertisers can choose levels of CPM for their campaign it is showed in Kolesnikov, Logachev and Topinskiy (2012) [13] that it is hard to predict what the outcome of CPM will be. The factors determining how successful the prediction will be is how many clicks similar ads have gotten. They found that if similar ads have not got many clicks, the prediction will be estimated with large confidence intervals. Instead they provide an algorithm for predicting CPM for advertisements with sparse click history by extending the criteria for what can be considered a similar ad and giving these ads a smaller weight in the prediction. The model built in Kolesnikov et al. (2012) [13] outperforms baseline estimate and previous research for estimation.

Predicting CPC is done in Wang and Chen (2012) [20]. They propose various models for predicting CPC not on social media platforms but for ad words, however not from a statistical point of view. They found different semantic segments that could be of use when predicting CPC. Nonetheless, they finish their paper by stating that the segments are hard to generalize to other advertisement forms than ad words.

Other features of interest for predicting CPM is researched in Cheng, van Zwol, Azimi and Manavoglu (2012) [6]. They propose that future modelling of CPM should include multimedia features of the ad, for instance brightness, pixel rate, background colors, number of characters and more. They

found significant results, for example that "flash ads with audio generate more clicks than flash ads without audio" and "Large background image ads receive less clicks than small background image ads". Nonetheless, they point out the problem of finding correct and reliable data for replicating their results and extend the study by including other multimedia features.

Investigating advertising on search engines, for example Google, was done in Tang, Yang and Pei (2013) [19] through regression modelling. The dependent variable was the amount of price information in an advertisement and the covariates in the model was divided in the three categories query, ad content and who the advertiser was (by using dummy variables), in order to present the results to companies and people with conflicting interests. Interesting for this thesis is that they included both CPC and CPC^2 as covariates, since they empirically find a quadratic behaviour of price information versus CPC. Hence, if regressing CPC or another cost-based metric as CPM, against other covariates in another setting, it is likely that some transformations of the covariates is needed. On the contrary, one should note that the modelling in the work of Tang et al. (2013) [19] is done for search engines and not for social media platforms, where the results can end up differently.

Online advertising from a mathematical point of view is also modelled in Krushevskaja, Simpson and Muthukrishnan (2016) [14]. They highlight that "advertisers pay per click (...), but require the average cost of a conversion to be below some threshold", and there is therefore a discrepancy between the advertiser and the social media platforms. They solve this by proposing a dynamic programming approach to optimize the best response strategy.

A different topic of interest when studying CPM on social media platforms is the problem of click fraud, having in mind that the cost of advertising is in some sense based on number of clicks. Click fraud is when the clicks are not done by social media users or that the number of clicks are not measured correctly. In fact, this is a major problem that according to Google Inc.'s former chief financial officer George Reyes "is the biggest threat to the Internet economy".[4] The problem of click fraud is highlighted in Midha (2009) [16] by finding evidence that "70 percent of advertisers are worried about click fraud" and some companies experience a click fraud rate as high as 35 percent. One could handle the problem of click fraud similar to how missing data points is handled in mathematical statistics. However, since the click fraud rate is so high as indicated in Midha (2008) [16], the models suggested will inevitably be affected by click fraud.

Having in mind the problems caused by click fraud, different models other than the CPM-model for determining the price of an advertisement on social media platforms was suggested in Amirbekian,

[4]Crawford, K., CNN, 2004

Chen, Li, Yan and Yin (2012) [1]. The models suggested incorporate the quality of the clicks and by that try to neglect the effect caused by click fraud. They are created from logistic regression, random forest modelling and a two-stage regression and all of the three models are in need of less historical data than other similar models. To sum up, the problems of click fraud that was highlighted by Midha (2008) [16] in 2008 was solved in 2012 by Amirbekian et al. (2012) [1].

4 Data and Model Building

This section explains the data used in the thesis and how the models are built.

4.1 Explorative Data Analysis

The data used in this thesis are advertisement data from Whispr Group's clients. It consists of performance metrics of advertisements displayed in Sweden on Facebook from January 1st to December 31st 2017. The data is collected on aggregated level for Whispr Group's clients and split by the campaign. Each client can decide through its specific account on Facebook's ads manager to create advertisements on certain campaigns. Hence, each data point is the aggregated advertisement performance on Facebook for a specific campaign for one of Whispr Group's client. The names of the clients and the campaigns cannot be revealed because of reasons of confidentiality between Whispr Group and its clients. The data in the data set is (with description defined on Facebook's Ads manager):

Variable name	Description
`CPM`	The average cost for 1000 impressions.
`Reach`	The number of people who saw the advertisement at least once.
`Frequency`	The average number of times each person saw the advertisement.
`Impressions`	The number of times the advertisement was on screen.
`Social_reach`	The number of people who saw the advertisement when it was displayed with social information of other Facebook friends that have engaged with the advertisement.
`Social_impressions`	The number of times the advertisement was displayed with social information of other Facebook friends that have engaged with the advertisement.
`Actions`	The sum of likes, comments, shares and link clicks on the advertisement.
`Amount_spent`	The total amount of money spent on advertisement and the Facebook page. Measured in SEK, not taken into account monetary inflation.
`Cost_1000_reached`	The average cost to reach 1000 people.
`Page_engagements`	The sum of likes, comments, shares and link clicks on the Facebook page that are attributed to the advertisement.
`Link_clicks`	The number of clicks on links shown in the advertisement.
`CPC`	The average cost for each link click.
`CTR`	The fraction of the times people saw your ad and performed a click.

Table I: Description of the dependent variable and the covariates in the data set. The dashed line mark the distinction between dependent variable and covariates.

The difference between `Reach` and `Impressions`, which in some sense is confusing, can preferably be explained by the following example: If the same advertisement appears on the same person's news feed twice, it has *reached* that person once but it *was on screen* twice. Therefore, in this example is `Reach` $= 1$ and `Impressions` $= 2$. The difference between `Reach` and `Social_reach` is that `Social_reach` is only effected if someone of your Facebook friends have liked that advertisement before it appears on your news feed. The distinction between `Impressions` and `Social_impressions` is defined in the same manner.

All dependent variables are attributable to the advertisements solely, except for `Amount_spent` and `Page_engagements`. `Amount_spent` is attributable to a combination of the Facebook page of the advertising company and the advertisements whereas `Page_engagements` is attributable to only the Facebook page of the advertising company.

4.2 Data Preprocessing

The data was cleaned before any statistical modelling could be started. At first, the data set contained 858 data points. The following cleaning was made, in this specific order, which reduced the data set with (x) data points:

1. Delete data points with `Amount_spent` $= 0$. (280)
2. Delete data points with `Reach` $= 0$. (6)
3. Delete data points with `Reach` $=$ "Cannot be found". (9)
4. Delete data points with `Impressions` $> 3,000,000$. (23)

All removals are straightforward, except possibly for number 4 which can be motivated by considering those data points as outliers. For instance, the maximal of `Impressions` before the cleaning of the data was 11.7M which can be related to Sweden's population of 10M at the time.[5] The data contained no missing values after this cleaning. The scale and spread of the variables after the data cleaning differs quite a lot, as this summary statistics shows:

[5] Statistics Sweden, "Sveriges Folkmängd från 1749 och fram till idag", 2018

Statistic	N	Mean	St. Dev.	Min	Max
CPM	540	17.695	29.479	0.260	199.820
Reach	540	190,490.800	271,799.900	8	2,227,066
Frequency	540	2.608	4.348	1.000	58.950
Impressions	540	456,607.200	624,144.600	9	2,997,043
Social_reach	540	116,115.700	183,570.700	0	1,980,593
Social_impressions	540	261,951.600	394,527.100	0	2,451,203
Actions	540	158,013.200	287,743.400	0	3,287,375
Amount_spent	540	4,091.011	9,664.541	0.010	100,566.400
Cost_1000_reached	540	38.598	68.867	0.300	500.600
Page_engagements	540	22,051.520	68,466.790	0	742,596
Link_clicks	540	3,549.957	7,991.678	0	112,439
CPC	540	2.036	5.401	0.000	73.710
CTR	540	1.794	1.867	0.000	14.760

Table II: Summary statistics for the dependent variable and the covariates in the data set. The dashed line mark the distinction between dependent variable and covariates.

Table II raises doubt whether the variables should be rescaled to obtain a data set more coherent in terms of scale. Nonetheless, Facebook Business Manager that provides the data set has no built-in feature to define new variables. And in order not to require too much analysis of the data before applying it to the model, it is more convenient to define the model in terms of non-scaled variables. One could argue here that transformations, if applied to the data, also are methods of rescaling the data. However, transformations are necessary to obtain a better performing model whereas rescaling only is a change of interpretation. Hence, transformations will be done if necessary and rescaling will not be done.

4.3 Model Building Approaches

This section present various model building strategies to be considered and studied. The strategy follows to some extent the strategy described in Montgomery et al. (2012) [17].

1. A model is fitted without any transformations.

2. All covariates are investigated to see whether any transformations are needed. The dependent variable is assessed through Box-Cox transformation. A model with the transformed data is fitted and hereafter is the transformed data used.

3. The residuals are assessed to find potential influential and leverage points. Influential points are deleted and a new model is fitted. The residuals in the obtained model are also assessed.

4. The data set is investigated for multicollinearity, by the metrics VIF and condition number.

5. All possible regression is performed in order to find if any covariates can be excluded out of the model, and in that case which covariates to exclude.

6. The first shrinkage method is tested, Ridge regression. Cross-validation is performed to find the optimal λ.

7. Lasso regression is tested. Cross-validation is also here performed to the optimal λ.

8. The combination of Ridge and Lasso, elastic net is tested. Two-dimensional cross-validation is performed to find the optimal weights for Ridge and Lasso and to find the optimal λ.

9. Adaptive Lasso is tested and compared to regular Lasso. Two-dimensional cross-validation is also here performed, but in this case to find the optimal γ and λ. Cross-validation using other estimators than OLS, as suggested in implementation for adaptive Lasso, is not performed.

10. Using derived inputs is PCR performed. The performance for including different number of principal components are compared.

5 Results and Analysis

5.1 Standard Regression Modelling

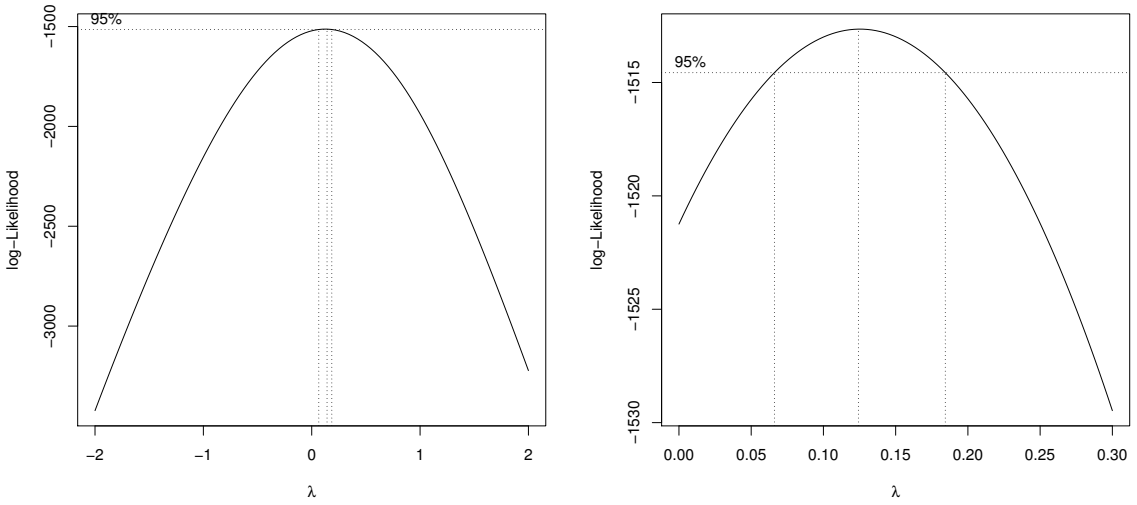
The standard regression model, as defined in the mathematical theory, will be created in this section.

5.1.1 Transformations

Without any transformations, the standard model to be estimated is

$$\begin{aligned}\texttt{CPM}_i = \beta_0 &+ \beta_1\texttt{Reach}_i + \beta_2\texttt{Frequency}_i \\ &+ \beta_3\texttt{Impressions}_i + \beta_4\texttt{Social\_reach}_i \\ &+ \beta_5\texttt{Social\_impressions}_i + \beta_6\texttt{Actions}_i \\ &+ \beta_7\texttt{Amount\_spent}_i + \beta_8\texttt{Cost\_1000\_reached}_i \\ &+ \beta_9\texttt{Page\_engagements}_i + \beta_{10}\texttt{Link\_clicks}_i \\ &+ \beta_{11}\texttt{CTR}_i + \beta_{12}\texttt{CPC}_i + \epsilon_i.\end{aligned} \tag{5.1}$$

Some type of transformations might be required to the data as outlined in the mathematical theory. Box-Cox transformation of the dependent variable give the following figure:

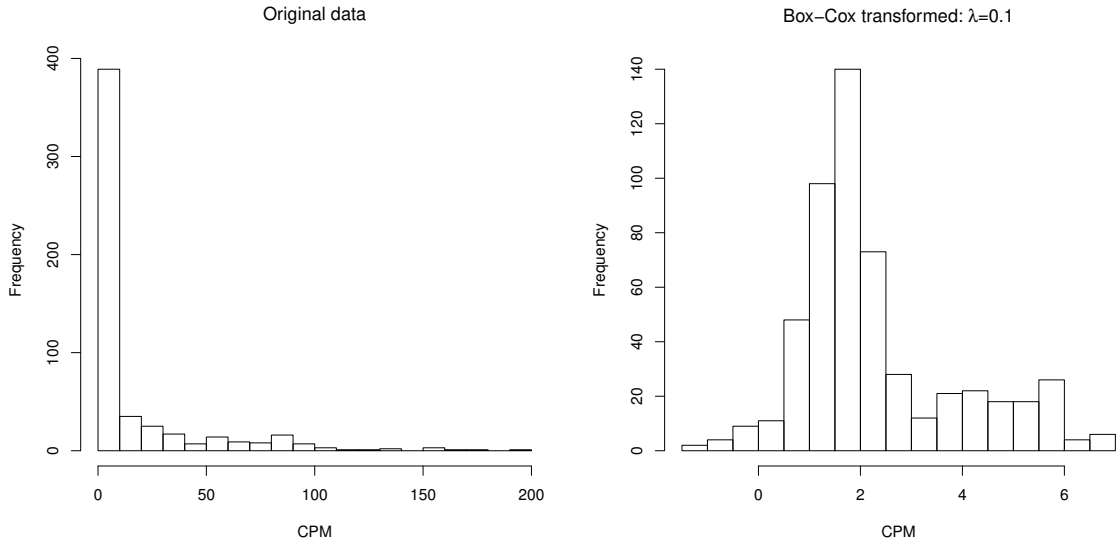


(a) $-2 \leq \lambda \leq 2$ *(b)* $0.0 \leq \lambda \leq 0.3$

Figure 5.1: Box-Cox transformation of the dependent variable `CPM` *as a function of the Box-Cox parameter* λ*. The three horizontal lines denote the optimal value and the 95% confidence interval.*

As suggested in the work of Box et al. (1964) [3] is the Box-Cox parameter λ not chosen exactly to the λ that maximizes the likelihood, but instead for a λ close to the maximizer such that the

transformations become easy to interpret. Hence, in this case chosen to $\lambda = 0.1$. The non-transformed and Box-Cox transformed data can then be seen in the histograms:



(a) Original data

(b) Box-Cox transformed data

Figure 5.2: Histogram for the dependent variable `CPM`, first for the non-transformed data and then for the Box-Cox transformed data.

As indicated by figure 5.2 the normality assumption of the dependent variable is better fulfilled for the Box-Cox transformed version of `CPM` since the histogram in figure 5.2(b) is more similar to the normal density function. Therefore, hereafter will the transformation

$$\texttt{CPM}_i \rightarrow \frac{\texttt{CPM}_i^{0.1} - 1}{0.1}, \tag{5.2}$$

be used. Each covariate is separately plotted against the (Box-Cox transformed) dependent variable `CPM` in order to see if any transformations are needed for the covariates.

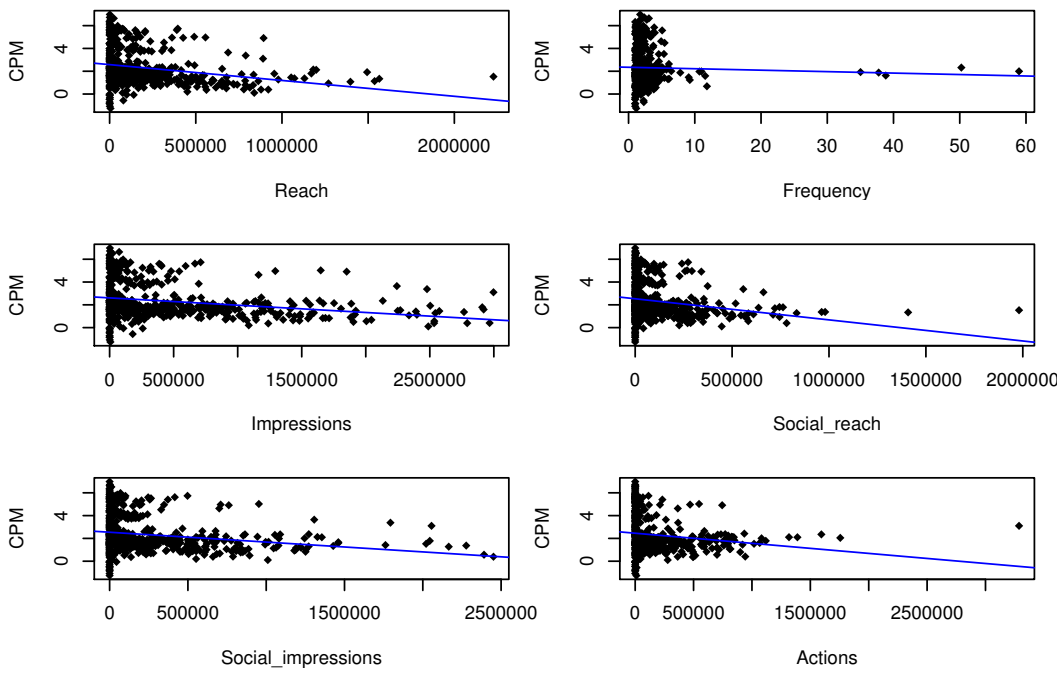


Figure 5.3: Scatter plot for the covariates `Reach`, `Frequency`, `Impressions`, `Social_reach`, `Social_impressions` *and* `Actions`. *The blue line in each plot denote the linear trend, estimated by OLS.*

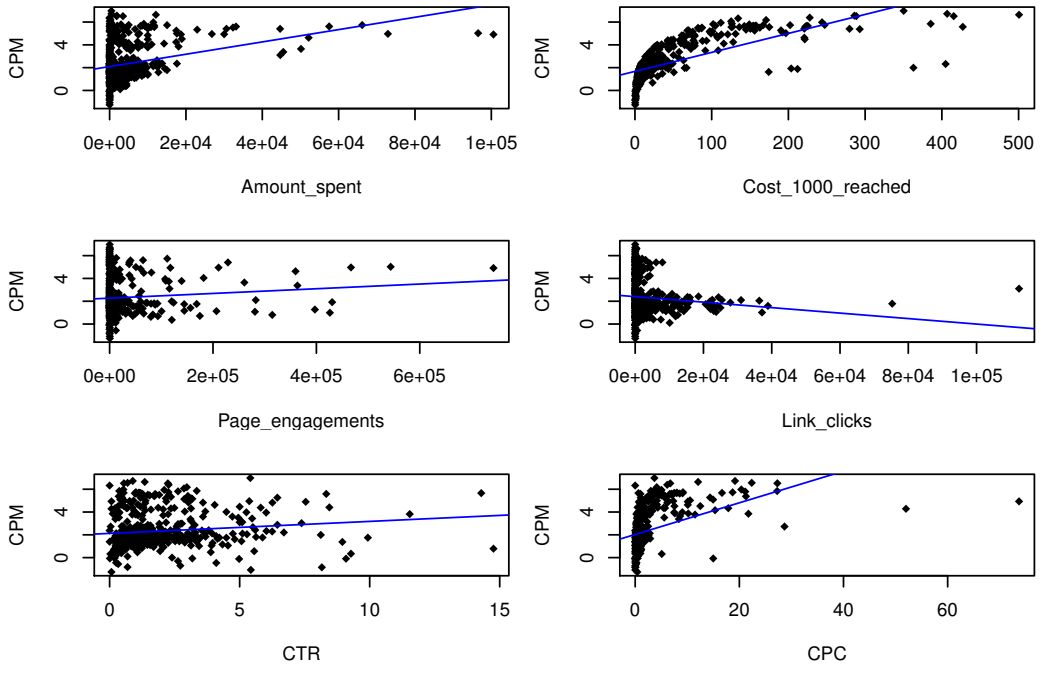


Figure 5.4: Scatter plot for the covariates `Amount_spent`, `Cost_1000_reached`, `Page_engagements`, `Link_clicks`, `CTR` *and* `CPC`. *The blue line in each plot denote the linear trend, estimated by OLS.*

From figures 5.3 and 5.4 it can be seen that some transformations are most likely needed. For instance, it appears that the square roots of `Social_reach` and of `Actions` should be used in the model. One way to find the most appropriate transformations is to first decide upon certain transformations, fit a linear trend of the transformed data and the dependent variable and from that model collect a performance metric, for instance R^2. Then, for each covariate, the transformation that maximizes R^2 is chosen. It appears that different power of reciprocal transforms are needed, according to figures 5.3 and 5.4. In some sense can $\ln(\cdot)$ be considered as the power transformation of order zero, and that transformation will consequently also be tested. To summarize, the transformations that will be tested are: $\sqrt{x}$, $\ln x$, x^{-1}, $x^{-\frac{3}{2}}$ and x^{-2}. Note that if at least one value is zero in the non-transformed data will some of these transformations not be defined.

	`Reach`	`Frequency`	`Impressions`	`Social_reach`
x	0.058	0.001	0.065	0.047
$\sqrt{x}$	**0.077**	**0.002**	**0.077**	**0.062**
$\ln x$	0.053	0.001	0.048	N/A
x^{-1}	0.020	0.000	0.018	N/A
$x^{-\frac{3}{2}}$	0.000	0.001	0.013	N/A
x^{-2}	0.014	0.000	0.011	N/A

	`Social_impressions`	`Actions`	`Amount_spent`	`Cost_1000_reached`
x	0.047	0.026	0.111	0.536
$\sqrt{x}$	**0.059**	**0.072**	**0.119**	0.735
$\ln x$	N/A	N/A	0.049	**0.803**
x^{-1}	N/A	N/A	0.013	0.235
$x^{-\frac{3}{2}}$	N/A	N/A	0.007	0.107
x^{-2}	N/A	N/A	0.006	0.065

	`Page_engagements`	`Link_clicks`	`CTR`	`CPC`
x	**0.008**	0.015	0.016	0.229
$\sqrt{x}$	0.002	**0.052**	**0.025**	**0.442**
$\ln x$	N/A	N/A	N/A	N/A
x^{-1}	N/A	N/A	N/A	N/A
$x^{-\frac{3}{2}}$	N/A	N/A	N/A	N/A
x^{-2}	N/A	N/A	N/A	N/A

*Table III: Each column is the R^2 of a fitted model of the dependent variable `CPM` and the covariate of the column. The **bold** numbers are the maximal R^2 for each covariate. N/A = Not Applicable.*

Table III gives indication about how the transformations could be done. These transformations will be done after maximizing R^2 and then subtracting mean and dividing by standard error to set the covariates on correlation form

$$
\begin{aligned}
\texttt{Reach}_i \to \sqrt{\texttt{Reach}_i} &\to \frac{\sqrt{\texttt{Reach}_i} - m_{\sqrt{\texttt{Reach}}}}{S_{\sqrt{\texttt{Reach}}}},\\
\texttt{Frequency}_i \to \sqrt{\texttt{Frequency}_i} &\to \frac{\sqrt{\texttt{Frequency}_i} - m_{\sqrt{\texttt{Frequency}}}}{S_{\sqrt{\texttt{Frequency}}}},\\
\texttt{Impressions}_i \to \sqrt{\texttt{Impressions}_i} &\to \frac{\sqrt{\texttt{Impressions}_i} - m_{\sqrt{\texttt{Impressions}}}}{S_{\sqrt{\texttt{Impressions}}}},\\
\texttt{Social\_reach}_i \to \sqrt{\texttt{Social\_reach}_i} &\to \frac{\sqrt{\texttt{Social\_reach}_i} - m_{\sqrt{\texttt{Social\_reach}}}}{S_{\sqrt{\texttt{Social\_reach}}}},\\
\texttt{Social\_impressions}_i \to \sqrt{\texttt{Social\_impressions}_i} &\to \frac{\sqrt{\texttt{Social\_impressions}_i} - m_{\sqrt{\texttt{Social\_impressions}}}}{S_{\sqrt{\texttt{Social\_impressions}}}},\\
\texttt{Actions}_i \to \sqrt{\texttt{Actions}_i} &\to \frac{\sqrt{\texttt{Actions}_i} - m_{\sqrt{\texttt{Actions}}}}{S_{\sqrt{\texttt{Actions}}}},\\
\texttt{Amount\_spent}_i \to \sqrt{\texttt{Amount\_spent}_i} &\to \frac{\sqrt{\texttt{Amount\_spent}_i} - m_{\sqrt{\texttt{Amount\_spent}}}}{S_{\sqrt{\texttt{Amount\_spent}}}},\\
\texttt{Cost\_1000\_reached}_i \to \ln \texttt{Cost\_1000\_reached}_i &\to \frac{\ln \texttt{Cost\_1000\_reached}_i - m_{\ln \texttt{Cost\_1000\_reached}}}{S_{\ln \texttt{Cost\_1000\_reached}}},\\
\texttt{Page\_engagements}_i &\to \frac{\texttt{Page\_engagements}_i - m_{\texttt{Page\_engagements}}}{S_{\texttt{Page\_engagements}}},\\
\texttt{Link\_clicks}_i \to \sqrt{\texttt{Link\_clicks}_i} &\to \frac{\sqrt{\texttt{Link\_clicks}_i} - m_{\sqrt{\texttt{Link\_clicks}}}}{S_{\sqrt{\texttt{Link\_clicks}}}},\\
\texttt{CTR}_i \to \sqrt{\texttt{CTR}_i} &\to \frac{\sqrt{\texttt{CTR}_i} - m_{\sqrt{\texttt{CTR}}}}{S_{\sqrt{\texttt{CTR}}}},\\
\texttt{CPC}_i \to \sqrt{\texttt{CPC}_i} &\to \frac{\sqrt{\texttt{CPC}_i} - m_{\sqrt{\texttt{CPC}}}}{S_{\sqrt{\texttt{CPC}}}},
\end{aligned}
\tag{5.3}
$$

where $m_x = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the arithmetic mean and $S_x = \sqrt{\sum_{i=1}^{n}(x_i - m_x)^2}$ is the standard error. There is a risk that these transformations are not the most appropriate since it has not been taken into account how transformations of multiple covariates affect the dependent variable. That means that the linearity of the dependent variable as a function of multiple covariates could be worsen by the transformation. Hence it could be the case that transformations of more advanced formulas could be even better. Nonetheless these transformations are easily interpreted and the model itself will consequently be easier to interpret. The transformations in equation (5.3) will be used for the data

hereafter. The updated version of equation (5.1) is

$$
\begin{aligned}
\frac{\texttt{CPM}_i^{0.1}-1}{0.1} =& \beta_1 \frac{\sqrt{\texttt{Reach}_i}-m_{\sqrt{\texttt{Reach}}}}{S_{\sqrt{\texttt{Reach}}}} \\
&+ \beta_2 \frac{\sqrt{\texttt{Frequency}_i}-m_{\sqrt{\texttt{Frequency}}}}{S_{\sqrt{\texttt{Frequency}}}} \\
&+ \beta_3 \frac{\sqrt{\texttt{Impressions}_i}-m_{\sqrt{\texttt{Impressions}}}}{S_{\sqrt{\texttt{Impressions}}}} \\
&+ \beta_4 \frac{\sqrt{\texttt{Social\_reach}_i}-m_{\sqrt{\texttt{Social\_reach}}}}{S_{\sqrt{\texttt{Social\_reach}}}} \\
&+ \beta_5 \frac{\sqrt{\texttt{Social\_impressions}_i}-m_{\sqrt{\texttt{Social\_impressions}}}}{S_{\sqrt{\texttt{Social\_impressions}}}} \\
&+ \beta_6 \frac{\sqrt{\texttt{Actions}_i}-m_{\sqrt{\texttt{Actions}}}}{S_{\sqrt{\texttt{Actions}}}} \\
&+ \beta_7 \frac{\sqrt{\texttt{Amount\_spent}_i}-m_{\sqrt{\texttt{Amount\_spent}}}}{S_{sqrt\texttt{Amount\_spent}}} \\
&+ \beta_8 \frac{\ln \texttt{Cost\_1000\_reached}_i - m_{\ln \texttt{Cost\_1000\_reached}}}{S_{\ln \texttt{Cost\_1000\_reached}}} \\
&+ \beta_9 \frac{\texttt{Page\_engagements}_i - m_{\texttt{Page\_engagements}}}{S_{\texttt{Page\_engagements}}} \\
&+ \beta_{10} \frac{\sqrt{\texttt{Link\_clicks}_i}-m_{\sqrt{\texttt{Link\_clicks}}}}{S_{\sqrt{\texttt{Link\_clicks}}}} \\
&+ \beta_{11} \frac{\sqrt{\texttt{CTR}_i}-m_{\sqrt{\texttt{CTR}}}}{S_{\sqrt{\texttt{CTR}}}} \\
&+ \beta_{12} \frac{\sqrt{\texttt{CPC}_i}-m_{\sqrt{\texttt{CPC}}}}{S_{\sqrt{\texttt{CPC}}}} + \epsilon_i,
\end{aligned}
\tag{5.4}
$$

where the intercept is taken out of the model since the data is mean-centered. The following regression results is obtained if estimating the models according to equations (5.1) and (5.4):

	Dependent variable:	
	CPM	
	(5.1)	(5.4)
Reach	$-3.990{\cdot}10^{-6}$ $(8.405{\cdot}10^{-6})$	0.012 ** (0.048)
Frequency	2.757 *** (0.141)	−0.308 *** (0.010)
Impressions	$2.994{\cdot}10^{-6}$ $(3.353{\cdot}10^{-6})$	−0.170 *** (0.047)
Social_reach	$3.976{\cdot}10^{-6}$ $(1.206{\cdot}10^{-5})$	0.112 ** (0.047)
Social_impressions	$-3.465{\cdot}10^{-6}$ $(5.192{\cdot}10^{-6})$	−0.132 *** (0.045)
Actions	$-3.458{\cdot}10^{-6}$ $(3.466{\cdot}10^{-6})$	−0.035 ** (0.014)
Amount_spent	$-1.662{\cdot}10^{-5}$ $(8.294{\cdot}10^{-5})$	0.022 (0.013)
Cost_1000_reached	0.404*** $(9.987{\cdot}10^{-3})$	0.980 *** (0.011)
Page_engagements	$-1.192{\cdot}10^{-5}$ $(1.125{\cdot}10^{-5})$	0.010 (0.009)
Link_clicks	$-2.233{\cdot}10^{-5}$ $(1.154{\cdot}10^{-4})$	−0.025 * (0.013)
CPC	0.373 *** (0.114)	0.061 *** (0.011)
CTR	1.119 *** (0.298)	0.049 *** (0.008)
Constant	7.315 *** (0.994)	
Observations	540	540
R^2	0.846	0.980
$R^2_{Adj.}$	0.842	0.980
F_0 (df = 12; 527)	240.869***	2,195.213***
AIC	4,204.338	−563.584
BIC	4,264.420	−503.502
Note:		*p<0.1; **p<0.05; ***p<0.01

Table IV: OLS regression without and with transformed covariates, with the standard errors of the estimates in parenthesis. Note that the interpretation of the coefficients are different between the two models due to transformations of the data, and thus it is not possible to compare the estimates of the coefficients. Also note that model (5.4) is defined without intercept.

Table IV shows, as expected, that model (5.4) with transformed data outperforms model (5.1) without transformed data. This is seen from, for example, that both R^2 and $R^2_{Adj.}$ are higher and that AIC and Bayesian information criterion (BIC) are lower for model (5.4). Also, the model is improved in the sense that more covariates are statistically significant on high confidence levels, which indicate in some way that the transformations have contributed to stabilize variances. Consequently, the model is

improved by using transformed data and henceforth will the transformed data according to equations (5.2) and (5.3) be used.

5.1.2 Leverage and Influential Points

In this section it will be assessed whether the model suffers from influential points. Any influential points will be detected and thereafter deleted to see whether the model is improved from the deletion. The residuals for model (5.4) are:

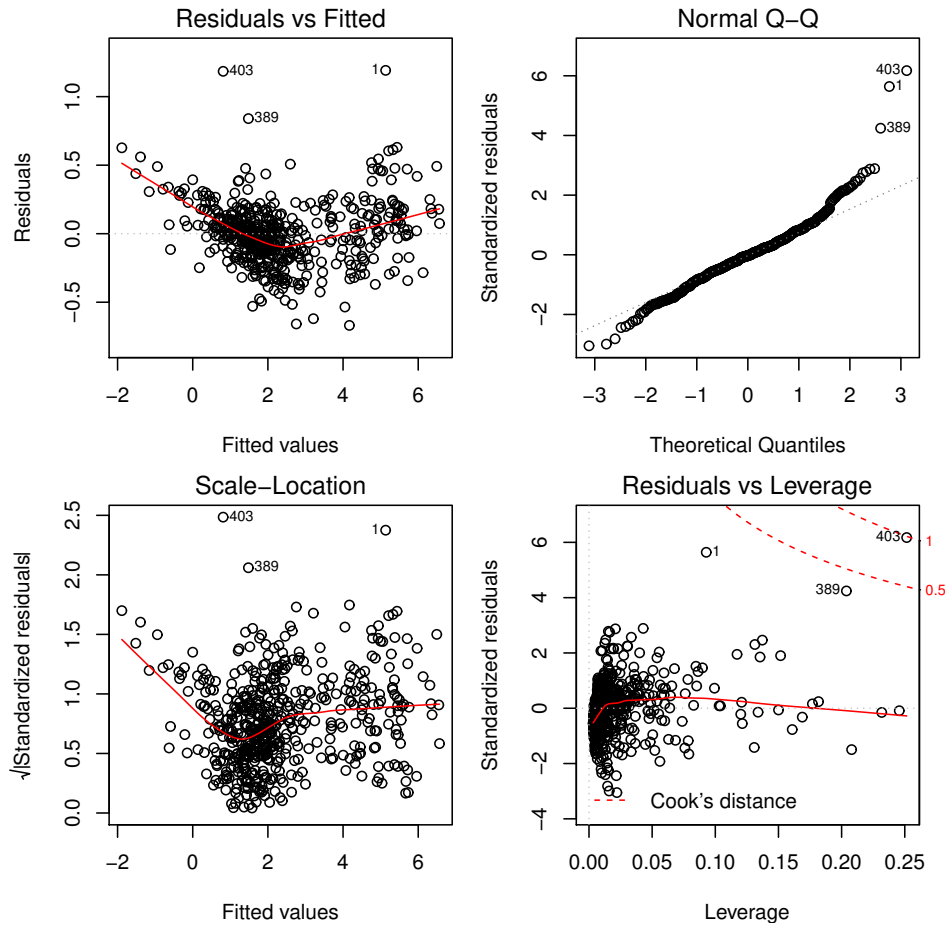


Figure 5.5: Residuals plot for model (5.4).

Figure 5.5 shows that some of the residuals are too large in order for the model to function properly. This can, for instance, be seen in the tails for the normal Q-Q plot where standardized residuals are too large in absolute vale at the tails. The points in the tails could possibly be influential, and therefore the DFFITS and Cook's distance are plotted:

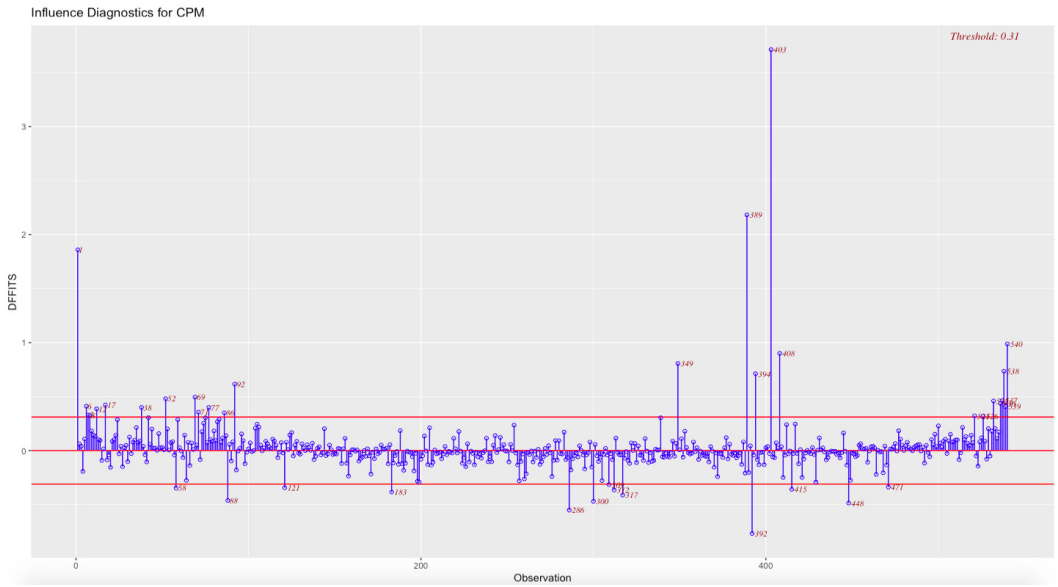


Figure 5.6: DFFITS for model (5.4). The red threshold level is calculated as $2\sqrt{(k+1)/n}$. On top of each bar is the ordering number.

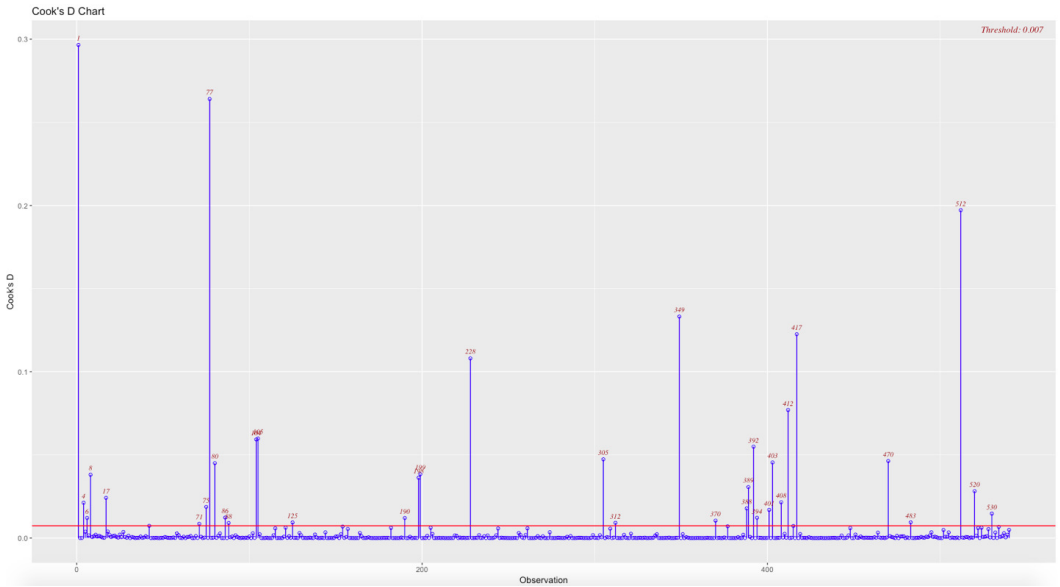


Figure 5.7: Cook's distance for model (5.4). The red threshold level is calculated from R as the level at which the same data points will be considered as influential points. On top of each bar is the ordering number.

Figures 5.6 and 5.7 show that the model do suffer from influential points since the DFFITS and Cook's distance for 37 data points exceed the threshold level. A bold solution to this problem is to simply delete these data points from the model and estimate a new model with consequently less data. It is reasonable in this case to use such an approach since it is relatively few data points that are deleted (only $37/540 \approx 7\%$). If it would be the case that a larger share of the data points were

deleted, this approach would not be as suitable. Also, these data points could be questioned whether they originate from click fraud which Midha (2009) [16] brought up. It can also be seen in figures 5.6 and 5.7 that the influential points are far out of the threshold levels. The performance metrics for the model with deleted influential points, together with the previous model from table IV, are:

	(5.4)	(5.4) without influential points
Observations	540	503
R^2	0.980	0.994
$R^2_{Adj.}$	0.980	0.994
F_0	2,195.213 (df = 12; 527)	6,722.101 (df = 12; 490)
AIC	−563.584	−1,153.242
BIC	−503.502	−1,094.154

Table V: Performance metrics of the regression with and with without influential points.

Table V shows, as expected, that the model is improved if deleting the influential points. This can be seen from both AIC and BIC that decreases at the same time as R^2 and $R^2_{Adj.}$ increases. To further highlight the improvement, the following figure shows that the behaviour of the residuals is improved if estimating a new model without the influential points.

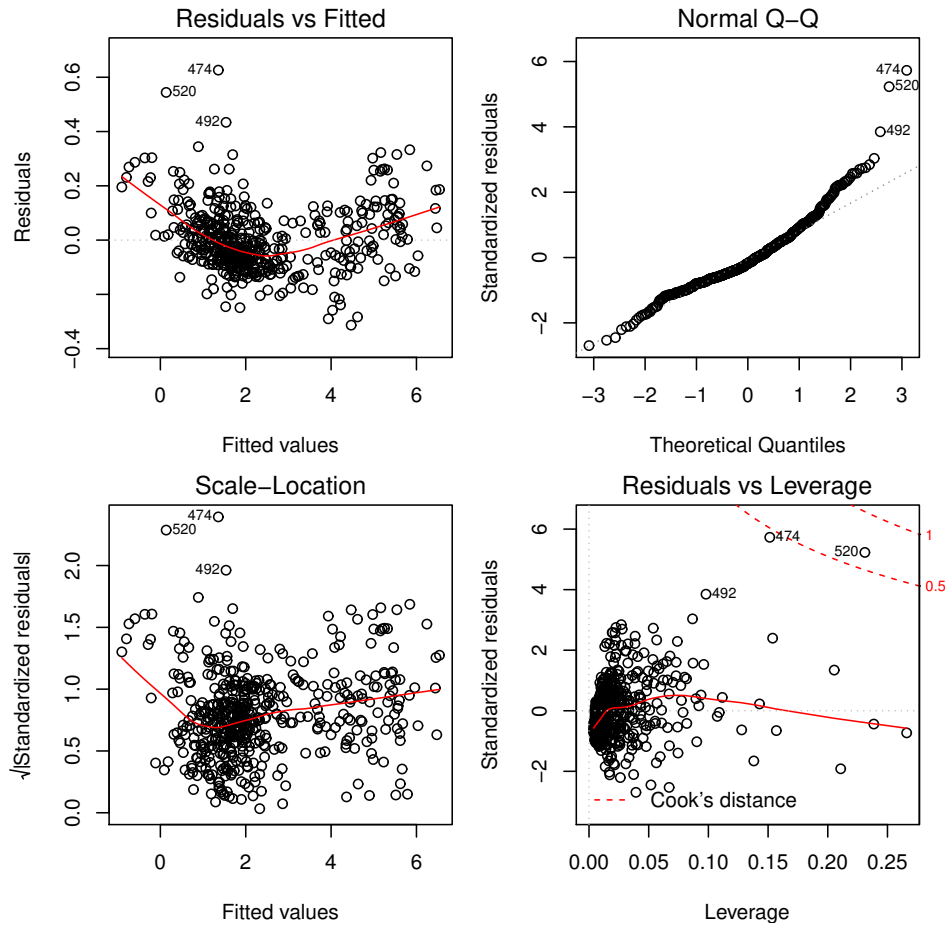


Figure 5.8: Residuals plot for model (5.4) without influential points.

Figure 5.8 shows better behaviour of the residuals as the influential points are deleted, however the residuals are not perfect since the tails for the normal Q-Q plot where the standardized residuals are too large in absolute vale at the tails.

5.1.3 Multicollinearity

Not all estimates in table IV are statistically significant, which could indicate that the model suffers from multicollinearity. This could preferably be seen from the VIFs and the conditional number κ. Here will the VIFs and κ be calculated for model (5.4) and the covariate with the highest VIF will be taken out of the model if one or more VIFs are > 5. Then the new VIFs will be calculated for the reduced model. The reduction continues until all VIFs are ≤ 5. The following VIFs are obtained:

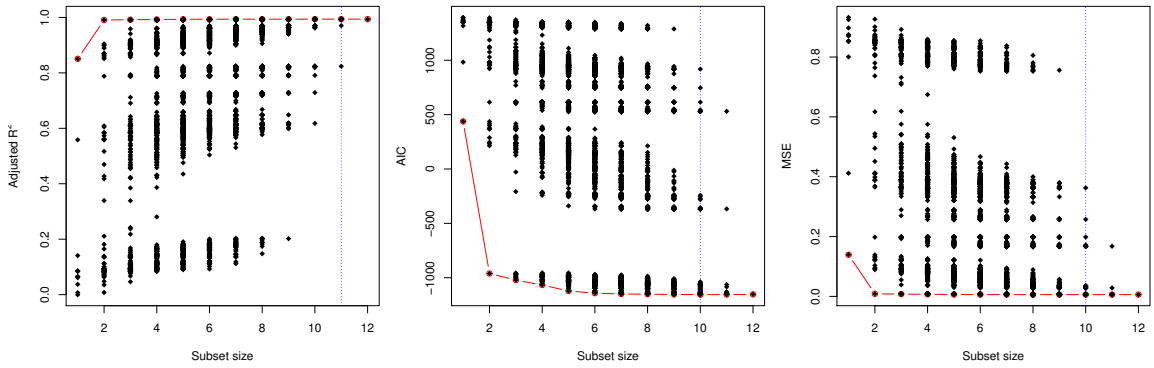
Excluded covariates:	0	1	2	3	4
`Reach`	72.633745	10.272201	**6.275053**		
`Frequency`	6.105851	3.978210	2.572427	2.273236	1.520395
`Impressions`	**74.414173**				
`Social_reach`	59.169989	**29.704821**			
`Social_impressions`	56.214911	26.618186	5.614645	4.004432	3.933260
`Actions`	5.632628	5.511324	4.945328	4.618521	4.616678
`Amount_spent`	5.034609	5.022805	4.939485	4.509109	3.987790
`Cost_1000_reached`	5.713651	5.648685	5.526354	**5.328202**	
`Page_engagements`	2.158267	2.157007	2.139489	2.132290	2.048072
`Link_clicks`	5.148061	4.946582	4.831011	4.794997	**4.743670**
`CPC`	4.845700	4.845599	4.839310	4.839304	2.327109
`CTR`	2.730614	2.723041	2.719147	2.550512	1.507258
κ	1,331.216	291.395	49.0504	41.03437	30.72596

Table VI: Starting from all covariates are the VIFs calculated and the covariate with the highest VIF is excluded. The algorithm continues until all VIF$\leq$ 5. For each round, the highest VIF-value is ***bold*** *and deleted in the next round. At the bottom of the table is the conditional number.*

Table VI shows that the model with all covariates suffers from severe problems of multicollinearity, with four VIFs > 10 and $\kappa > 100$. There is still a problem if excluding two variables according to analysis of the VIFs, however the analysis of conditional number leads to the conclusion that the problem is not as severe after excluding two covariates. Furthermore, the problem of multicollinearity is solved both according to analysis of VIF and conditional number if excluding four covariates. The covariates to exclude according to this analysis are namely `Impressions`, `Social_reach`, `Reach` and `Cost_1000_reached` in order to obtain a model without problem of multicollinearity. Nonetheless, it should here be taken into account that the problems of multicollinearity are not completely reduced by this analysis. The VIFs are after the reduction still close to the threshold level of 5.

5.1.4 Variable Selection

If performing an all possible subsets regression, the following performance metrics are obtained as a function of the number of covariates:



(a) $R^2_{Adj.}$ *(b) AIC (c) MSE*

Figure 5.9: Each black dot correspond to a regression model. The red line in each plot connects the best performing model for each subset size. The blue dotted line gives how many covariates to include in the model according to that performance metric.

Figure 5.9 shows that across the three metrics the best model does not incorporate all 12 covariates. $R^2_{\text{Adj.}}$ decides upon a model of 11 covariates which excludes the use of `Page_engagements` whereas AIC and MSE actually decide upon the same model of 10 covariates that excludes the use of `Page_engagements` and `Link_clicks`. What is interesting is that the analysis of multicollinearity in table VI leads to a completely different result, namely to exclude `Impressions`, `Social_reach`, `Reach` and `Cost_1000_reached`. The following cross-validated mean square error are obtained if comparing the OLS-models chosen by (I) VIF, (II) $R^2_{Adj.}$ and (III) AIC and MSE:

	Covariates excluded	10-fold CV
(I)	`Impressions`, `Social_reach`, `Reach` and `Cost_1000_reached`	0.177845450
(II)	`Page_engagements`	0.006792440
(III)	`Page_engagements` and `Link_clicks`	0.006799235

Table VII: Cross-validated mean square error of the three models chosen from analyzing the VIF (1) and from an all possible subsets regression method (2 and 3).

Table VII shows that the two models chosen by the all possible subsets regression clearly outperforms the model chosen from analyzing the VIFs. If comparing the two models chosen from the all possible subsets regression, it can be concluded that the performance is worsen if also excluding `Link_clicks`. Nonetheless, the difference in 10-fold cross-validated errors is small. Therefore, if one decides to exclude `Page_engagements` then `Link_clicks` should be kept in the model.

5.2 Shrinkage Regression Methods

In this section shrinkage regression models are created and tested.

5.2.1 Ridge

As outlined in the mathematical theory a solution to the multicollinearity problem, that the model clearly suffers from, is to instead use Ridge regression. By varying the Ridge parameter, the following Ridge trace is obtained:

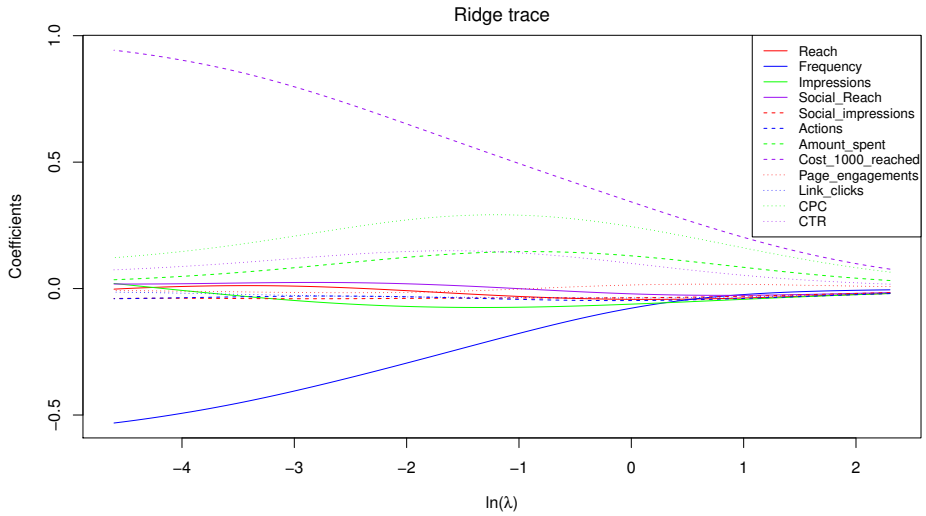
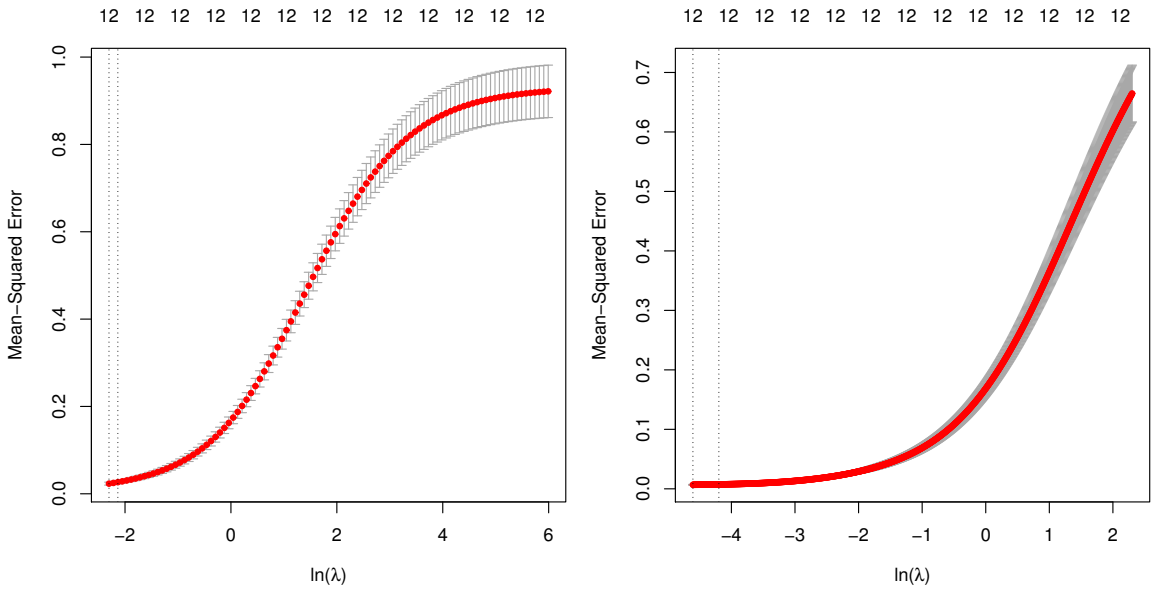


Figure 5.10: The regression coefficients plotted against $\ln\lambda$ *where* λ *is the parameter in the Ridge regression.*

As seen in figure 5.10, the Ridge regression does work as a shrinkage method since the absolute value of the coefficients decreases for higher values of λ. In order to find the most suitable value of λ a cross-validation procedure is performed, first with the default scaling of λ built in R and then with a zoomed in scaling of λ:



(a) Default λ *(b) Zoomed in λ*

Figure 5.11: 10-fold cross-validated error for varying values of λ in the Ridge regression. The red dots denote the average of the cross-validated errors and the grey bars denote the errors' range. The values on top of the diagram denote the number of non-zero coefficients that the model incorporates.

Figure 5.11 shows that relatively small values of λ is required to get a small cross-validated error. It is also seen that the cross-validated error is roughly constant when λ is sufficiently small. Even though, it exists a minimum value for the cross-validated error, which is at $\ln \lambda = \ln 0.01 = -4.60517$. Figure 5.11 is in line with the non-existing variable selection features of the Ridge regression, which is seen on top of the diagram where all variables are included in the solution for all choices of λ.

5.2.2 Lasso

The other commonly used shrinkage method called Lasso gives the following trace:

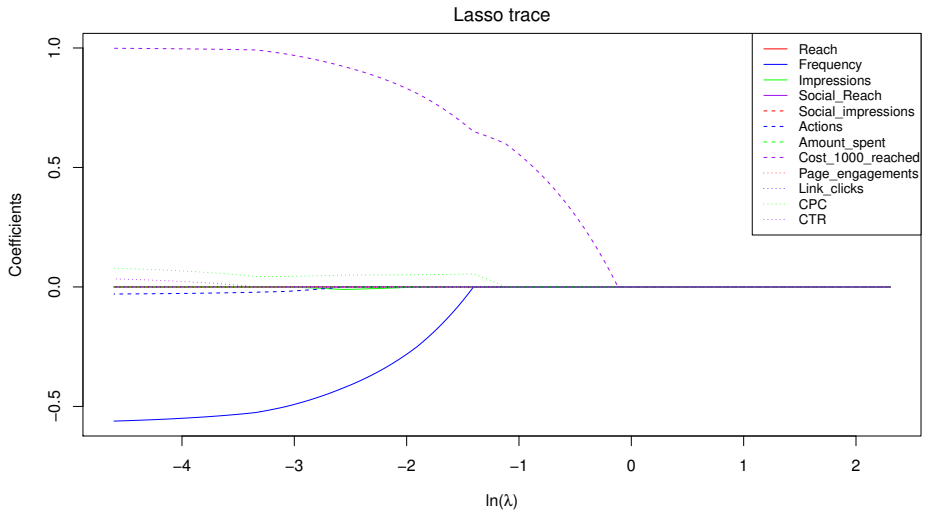
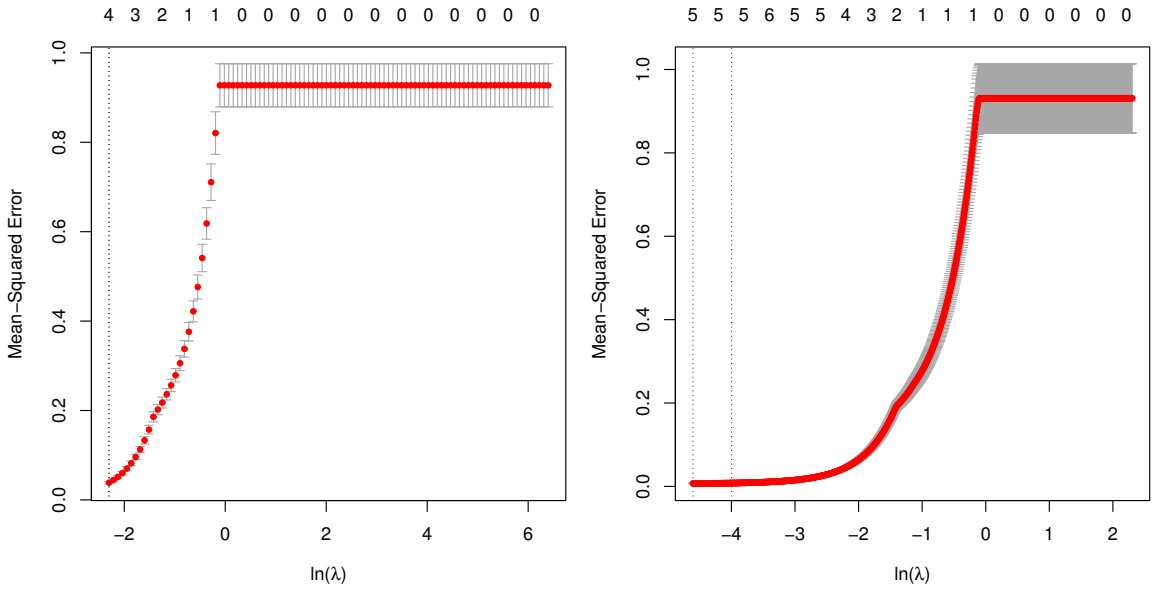


Figure 5.12: The regression coefficients are plotted against $\ln \lambda$ *where* λ *is the parameter in the Lasso regression.*

As for the Ridge trace in figure 5.10 it is clear that Lasso also works as shrinkage parameter. However, different from Ridge is that Lasso performs variable selection inevitably. It is clear from figure 5.12 that the covariate that is kept longest in the model is `Cost_1000_reached` followed by `CPC`. As for Ridge, in order to find the most suitable value of λ is a cross-validation procedure performed, first with the default scaling of λ built in R and then with a zoomed in scaling of λ:



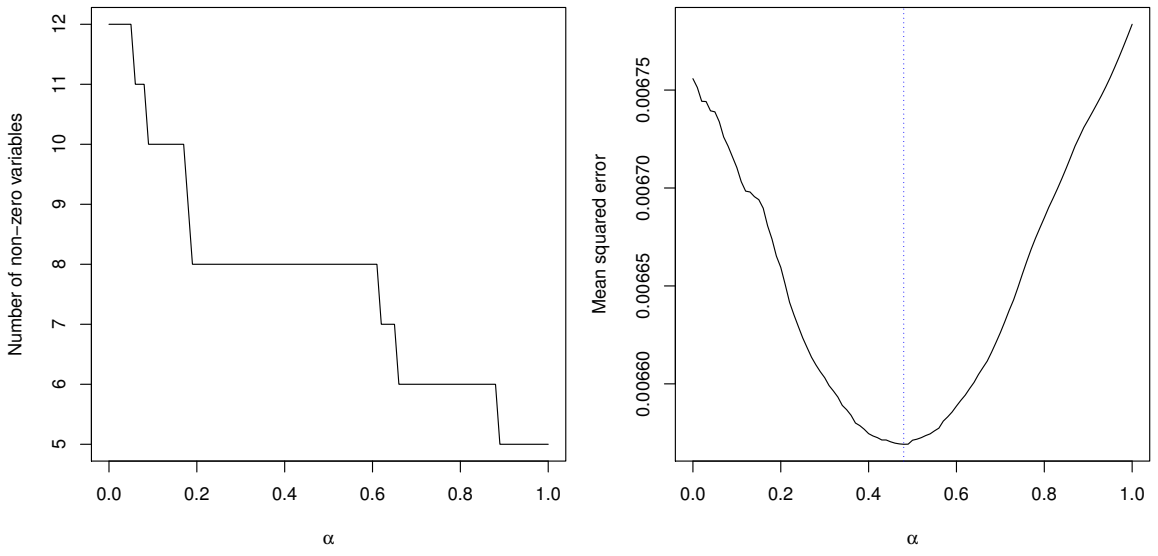
(a) Default λ *(b) Zoomed in* λ

Figure 5.13: 10-fold cross-validated error for varying values of λ *in the Lasso regression. The red dots denote the average of the cross-validated errors and the grey bars denote the errors' range. The values on top of the diagram denote the number of non-zero coefficients that the model has.*

Figure 5.13 shows that the cross-validated error decreases and that less variables are kept in the model for smaller values of λ, where the latter is as expected according to theory and also figure 5.12. The minimum value of the cross-validated error is obtained for $\ln \lambda = \ln 0.01 = -4.60517$. Pay attention to the harsh variable selection performed by Lasso since the number of non-zero covariates decreases drastically for larger numbers of λ.

5.2.3 Elastic Net

Ridge and Lasso can be combined using an elastic net as stated in the theory section. The elastic net contains the two parameters $\lambda \geq 0$ and $\alpha \in [0, 1]$, that in this case will not be decided in advance. The most optimal combination of these parameters will instead be estimated from the data. By letting $\alpha = 0, 0.01, \ldots, 1$, a model can be estimated for each value of α. The optimal value of λ for those models can be found by cross-validating the errors and choosing the λ that minimizes the mean squared error. It implies to go through the same procedure of Ridge and Lasso as earlier for each value of α. This method relies on computer power in order to function properly. If doing this algorithm, for each value value of α, plotted with its optimal combination with λ, the number of covariates in that model and the mean squared error are obtained:



(a) Number of covariates as a function of α

(b) Mean squared error as a function of α

Figure 5.14: On the left is the number of non-zero covariates kept for that choice of λ as a function of α. On the right is the cross-validated mean squared error for $\alpha \in [0, 1]$. The blue dotted line represents the α that minimizes the cross-validated mean square error.

Recall that $\alpha = 0$ is equivalent to Ridge and $\alpha = 1$ is equivalent to Lasso, which in essence gives that the optimal value of α is found more or less at equal weights of Ridge and Lasso, namely at

$\alpha = 0.48$. The mean squared error in figure 5.14 also gives that the elastic net outperforms both the pure Ridge and pure Lasso in terms of performance. Also, the variable selection features of Lasso is more prevalent when α is close to 1, i.e. close to Lasso regression, as expected. The trace for $\alpha = 0.48$ is:

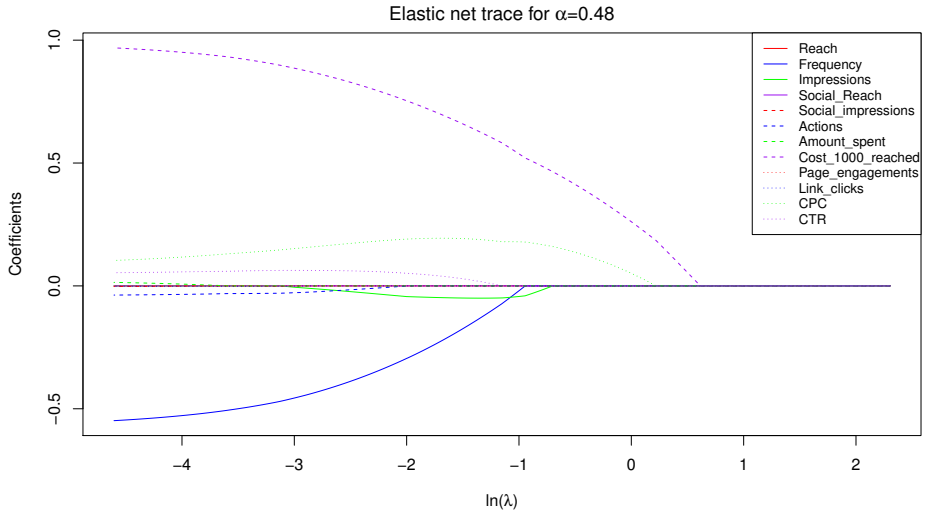


Figure 5.15: The regression coefficients are plotted against $\ln \lambda$.

The Ridge, elastic net with $\alpha = 0.48$ and Lasso regression estimates, using cross-validated optimal λ's in each case, are found in the following table:

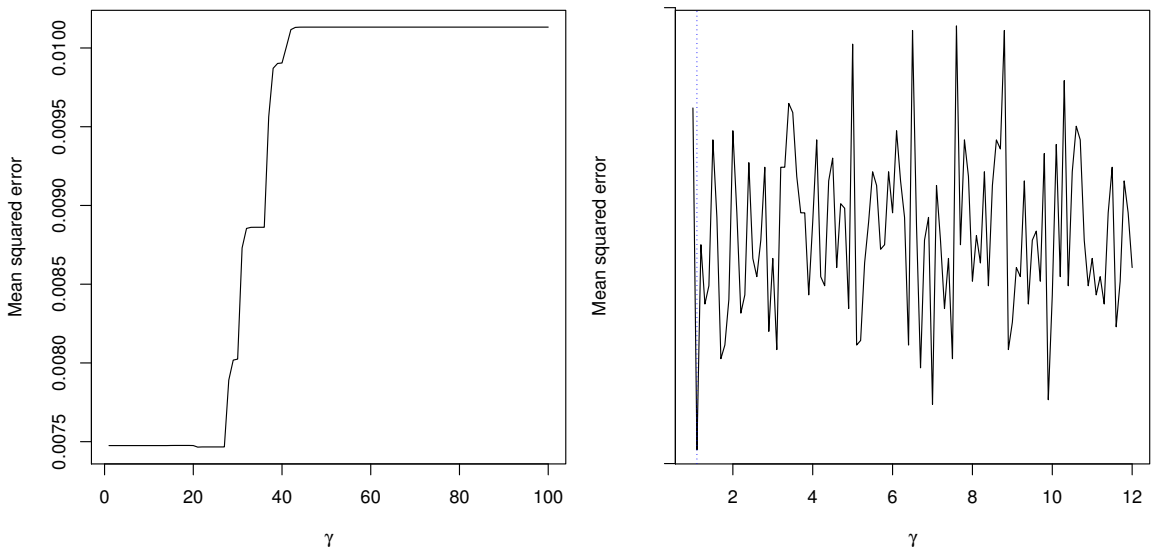
	Dependent variable:		
	CPM		
	Ridge	Elastic net ($\alpha = 0.48$)	Lasso
Reach	−0.002	0	0
Frequency	−0.532	−0.548	−0.561
Impressions	0.020	0	0
Social_reach	0.018	0	0
Social_impressions	−0.039	−0.003	0
Actions	−0.040	−0.037	−0.030
Amount_spent	0.034	0.014	0
Cost_1000_reached	0.943	0.969	0.999
Page_engagements	−0.008	0	0
Link_clicks	−0.014	−0.002	0
CPC	0.122	0.104	0.078
CTR	0.074	0.054	0.033
Constant			

Table VIII: The coefficient estimates for Ridge, elastic net and Lasso regression respectively.

Here it is obvious that Lasso have performed variable selection, since the estimates for `Reach`, `Impressions`, `Social_reach`, `Social_impressions`, `Amount_spent` `Page_engagements` and `Link_clicks` are found to be identically zero and left out of the model. Recall table VI where it was found that `Reach`, `Impressions`, `Social_reach` and `Cost_1000_reached` was the covariates that should be left out of the model first. Hence, Lasso regression and the selection of covariates through VIF leads to similar, but not identical, variable selection. The elastic net regression is similar to Lasso in terms of variable selection, since `Reach`, `Impressions`, `Social_reach` and `Page_engagements` are set to zero in both algorithms. It can also be said that the variable selection features of VIFs and elastic net lead to very similar conclusion, with the only disagreement being whether `Cost_1000_reached` and `Page_engagements` should be excluded or included.

5.2.4 Adaptive Lasso

It is problematic that Lasso is not consistent in terms of variable selection, as verified in the mathematical theory. In this section will therefore the adaptive Lasso procedure, which is consistent in variable selection, be performed. As for the elastic net, adaptive Lasso has two parameters $\gamma > 0$ and $\lambda > 0$ that need to be estimated. By discretizing γ within a predetermined interval, a model can be fitted for all discrete values of γ. The optimal value of λ for each choice of γ can then be cross-validated by minimizing the mean squared error. λ as a function of γ is found from this proedure. Then, if cross-validating the models for each choice of γ, in combination with its optimal λ, the following cross-validated mean square error is obtained:



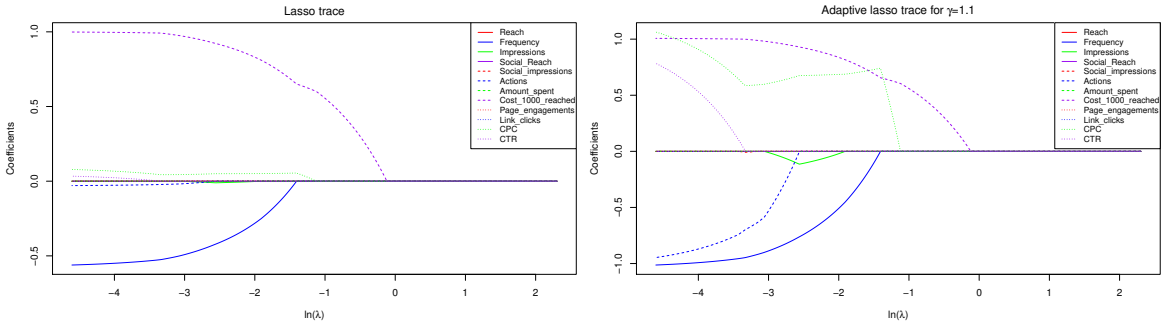
(a) MSE as a function of γ

(b) MSE as a function of γ

Figure 5.16: MSE as a function of γ for two choices of discretizations of γ. The blue dotted line denotes the γ that minimizes MSE.

Figure 5.16(a) shows that performance is worsen drastically if $\gamma > 26$. The discretization in figure 5.16(b) is consequently more useful. That figure shows that the best performing model is obtained by choosing $\gamma = 1.1$. Nonetheless, figure 5.16 shows that the performance of the model obtained if $\gamma \to 0$, which is equivalent to Lasso regression, is at similar levels as for $\gamma = 1.1$. Thus, the performance is only slightly improved by updating the Lasso so that it becomes consistent.

To highlight the difference between Lasso and adaptive Lasso, their respective traces are plotted:



(a) Lasso *(b) Adaptive Lasso*

Figure 5.17: Comparison of the traces of Lasso and adaptive Lasso for $\gamma = 1.1$ *as a function of* $\ln \lambda$.

The traces for Lasso and adaptive Lasso are in this case pretty similar, as indicated by figure 5.17.

5.3 Derived Inputs Regression: PCR

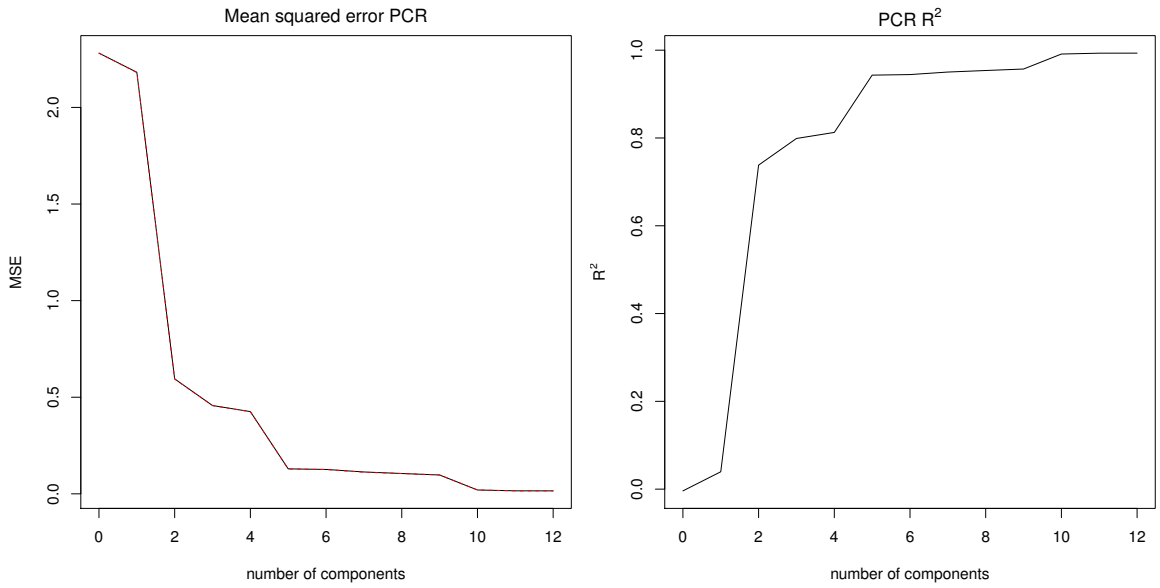
Another solution to the multicollinearity problem, that according to table VI prevails, is to orthogonalize the data and perform a PCR, see mathematical theory for further details. PCR also has the upside of mitigating the potential drawback of overfitting, as only a few principal components are chosen in the model. The following validation error and cumulative percentage error are obtained if performing PCR on model (5.4) using 10-fold cross-validation:

Cross-validated error													
	Const	1	2	3	4	5	6	7	8	9	10	11	12
CV	0.9644	0.9422	0.4914	0.4344	0.4260	0.2343	0.2308	0.2154	0.2075	0.2019	0.0911	0.0812	0.0805
adjCV	0.9644	0.9419	0.4909	0.4341	0.4251	0.2336	0.2301	0.2143	0.2070	0.2018	0.0909	0.0809	0.0802

Cumulative percentage of variance explained												
	1	2	3	4	5	6	7	8	9	10	11	12
X	48.67	67.74	78.62	87.10	92.48	95.36	96.77	97.99	98.98	99.71	99.96	100.0
`CPM`	5.04	74.25	80.16	81.93	94.53	94.70	95.35	95.63	96.03	99.17	99.37	99.40

Table IX: At the top is the 10-fold cross-validated root mean square error of the prediction (CV) and the bias-corrected cross-validated root mean square error the prediction (adjCV) as a function of the number of components. At the bottom is the cumulative percentage of the variance explained for the covariates ($\boldsymbol{X}$*) and for the dependent variable (`CPM`) as a function of the number of components.*

The performance of the PCR can be seen in the following plots:



(a) Mean squared error *(b) R^2*

Figure 5.18: Performance metrics for the PCR as function of number of components.

All performance metrics are improved by including more components in the model as expected. It is wanted in the PCR to have a low cross-validated error for a low number of components, and according to figure 5.18 the mean squared error does not improve anything dramatically if including more than five components. Therefore, at least five principal components should be included in the model according to PCR. This is in line with Lasso regression, even though Lasso is not built upon derived inputs as PCR is.

5.4 Model Suggestion

One can say that 7 models have been suggested in this thesis, namely from (I) VIF, (II) all possible subset variable selection, (III) Ridge, (IV) Lasso, (V) elastic net, (VI) adaptive Lasso and (VII) PCR. The following mean squared errors are obtained if cross-validating the performance of these models:

	Model	10-fold CV
(I)	VIF	0.177845450
(II)	All subsets variable selection	0.006892440
(III)	Ridge	0.007018895
(IV)	Lasso	0.007105498
(V)	Elastic net (α=0.48)	0.006795659
(VI)	Adaptive Lasso (γ=1.1)	0.007662462
(VII)	PCR (five components)	0.053024914

Table X: 10-fold cross-validated mean square error of the models suggested.

It can be seen in table X that the performance of all subsets possible regression and the shrinkage models are very similar, whereas the VIF- and PCR-models perform worse. The performance of the shrinkage models do not differ to a large extent, but it should be noted that the elastic net is the one that performs the best. The performance of PCR is worse than the shrinkage models, with a possible explanation of only including five components in the PCR. The performance for PCR would be better if including more components.

6 Discussion and Conclusion

This section aims to evaluate the models created and assess the most appropriate type of modelling for estimating CPM. The section will end with concluding remarks for the models created followed by some suggestions for future research on this topic.

6.1 Model Building Evaluation

The model building started with quickly realizing that the Box-Cox transformed version of `CPM` was more appropriate since the normality assumption was improved by the transformation. Nonetheless, the transformed data was still considered far from being perfectly normally distributed and the problems that this cause should be noted. It was primarily the right tail in the histogram that was too thick, causing the normal Q-Q plot not being perfectly shaped. It was also seen that transformations of the covariates were needed. Further analysis demonstrated that `Page_engagements` was the only covariate that did not transform and the most common transformation for the other covariates was the square root. To transform the covariates is in line with the work of Tang et al. (2013) [19]. The Box-Cox and covariates' transformation clearly improved the performance, since R^2, $R^2_{Adj.}$, AIC and BIC all were improved after the transformation. From this it can be concluded that it is important to apply transformations of various kinds to a data set that originates from advertisement metrics on social media platforms. When the residuals were analyzed, 37 data points (or 7% of the data set) were considered as outliers and removed from the model building. This helped to improve the heavy tails of the histogram, even though the tails remained heavy. Despite this, the fit was improved from the removal since R^2, $R^2_{Adj.}$, AIC and BIC were all improved.

Since some of the covariates have similar definitions (for instance `Reach` and `Impressions`), it was not far-fetched to suspect a data set with high degree of multicollinearity. This hypothesis was verified by the VIFs and the conditional number. The problem of multicollinearity was prevalent until four covariates were excluded from the model. All subsets regression, which only maximizes performance and neglects problems of multicollinearity, also came to the conclusion of excluding covariates, even though only one or two covariates depending on according to which metric. From this, one can conclude that the data set contained a few covariates that in practise were unnecessary.

Shrinkage regression methods were implemented to solve the multicollinearity problem, first Ridge and Lasso and then by combining the two in an elastic net. In the present data set it was found that the optimal shrinkage parameter λ was small for all these three, implying that only a small penalty term was used. It was also found that the optimal Lasso and elastic net regression performed extensive variable selection as those models excluded seven and four covariates respectively. The adaptive

Lasso regression was also implemented to not neglect the problems of consistency of Lasso. In terms of outcome of adaptive Lasso, the optimal value of γ in adaptive Lasso produced similar trace as for the regular Lasso trace. This implies that adaptive Lasso, even though solving the problem of consistency, did not change extensively the model proposed by the Lasso regression. Thus, all shrinkage methods concluded that a few covariates did not improve the model, as was the case for all subsets regression and analysis of multicollinearity.

The PCR do completely solve the problem of multicollinearity by orthogonalizing the data. For this data, the PCR showed that approximately five components were needed for obtaining a model of sufficient performance. The performance was of course improved if including more components, nonetheless the increase was comparably small.

6.2 Concluding Remarks and Recommendations

It is concluded from table X that a shrinkage method is a suitable regression method for regressing CPM. The reason for this is that many covariates in the original data set were strongly dependent and should be excluded in models, which was handled by the variable selection features from Lasso. The best performing model was the elastic net model with $\alpha = 0.48$. The variables included in that model was `Frequency`, `Social_impressions`, `Actions`, `Amount_spent`, `Cost_1000_reached`, `Link_clicks`, `CPC`, `CTR`. The remaining covariates, namely `Reach`, `Impressions`, `Social_reach` and `Page_engagements` were excluded. If choosing a model with less covariates, the following should be excluded and in this order: `Link_clicks`, `Social_impressions`, `Amount_spent`.

Also, shrinkage regression methods is suitable from an implementation perspective. It would be impossible to use an all possible regression approach if the data set would contain significantly more variables than 12, since the complexity increases by $\mathcal{O}(2^k)$. To efficiently shrink a data set with significantly more variables, which could be the case of choosing other metrics from Facebook's Ads manager, the elastic net with two-dimensional cross-validation for tuning of parameters is suitable. Furthermore, as earlier stated Facebook Ads manager has many other metrics that could be incorporated in a model. Nonetheless, the analysis in this thesis shows that more metrics are not needed if the aim is to maximize performance.

6.3 Suggestion for Future Research

The models built in this thesis rely upon advertisement data from Whispr Group's Swedish clients during 2017. It could be the case that this data is not significative for advertisement data for other Swedish companies, during other time periods or at other geographical locations. In the end, adver-

tisement is linked to consumer behaviour which changes across time and geographies. Having this in mind, it would be interesting to see whether using models outlined in this thesis on another data set might lead to other results. Furthermore, it should be taken into account that some data points were excluded. These include non-complete data points, unreasonably large data points and outliers in the model. Hence, the models built can be criticized for these deletions.

A substantial part of the model building went to solving the problems of multicollinearity, which could be traced to that the data set contained too many covariates. It would therefore be interesting to build a model with less covariates to begin with and see if the results end up differently. Another similar extension would be to choose a completely different set of covariates. This would be possible since Facebook Business Manager has even more metrics of advertisement data that could be used for model building. All models in this thesis are built on a regression framework which is of course not the only method that could be suitable. For instance machine learning or deep learning models could lead to even better models.

A factor that contributes to advertisement performance, which is completely neglected in this thesis, is how the advertisements are formulated. It makes sense, and was also shown in Cheng et al. (2012) [6], that a well formulated advertisement would outperform a poorly formulated and an advertisement with an appealing picture would also meliorate performance. Consequently, future work could focus on modelling from a statistical point of view how advertisement texts should be formulated and what pictures are suitable in advertisements. Also, recall that this thesis focused on predicting CPM, which is not the only metric within the field of advertisement data that could be investigated. For instance, predicting cost for advertising by CPC could also be investigated that most likely would end up in interesting models. The possibilities within this field is broad with many different variables for which models can be built.

During the creation of this thesis, it was revealed how political consulting firm Cambridge Analytica used data from Facebook to help Donald Trump win the presidential election in the United States of America. The news, which were later rephrased to a data sharing scandal, can shortly be summarized that Facebook provided a consulting firm with personal data from 87 million Facebook users without the consent from the users. The scandal led to that Facebook's founder and CEO Mark Zuckerberg was heard by the US congress about Facebook's handling of data. The connection with the scandal and this thesis may seem somewhat unclear but it has to do with how one can get data from Facebook in the future. The Cambridge Analytica scandal has forced Facebook to be careful with how they spread data to other companies, which might make it harder for companies like Whispr Group to get advertisement data from Facebook. Therefore, the models created in this thesis could be unique

and could be completely different in the future if Facebook restricts even further how they provide companies with data.

7 References

[1] Amirbekian, R., Chen, Y., Li, A., Yan, T., Yin, L. 2012, "Traffic Quality based Pricing in Paid Search using Two-Stage Regression", *Proceedings of the 22nd International Conference on World Wide Web*, pp. 113-114.

[2] Belsley, D. A., Kuh, E. Welsch, R. E. 1980, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New York NY.

[3] Box, G. E. P., Cox, D. R. 1964, "An Analysis of Transformations", *Journal of the Royal Statistical Society*, vol. 26, no. 2, pp. 211–252.

[4] Bühlmann, P. 2017, "High-dimensional statistics, with applications to genome-wide association studies", *EMS Surveys in Mathematical Sciences*, vol. 4, pp. 45-75.

[5] Chen, Q., Yu, S., Guo, Z., Jia, Y. 2016, "Estimating Ads' Click through Rate with Recurrent Neural Network", *ITM Web of Conferences 7*.

[6] Cheng, H., van Zwol, R., Azimi, J., Manavoglu, E. 2012, "Multimedia Features for Click Prediction of New Ads in Display Advertising", *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 777-785.

[7] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. 2004, "Least Angle Regression," *The Annals of Statistics*, vol. 32, pp. 407–499.

[8] Fan, J., Li, R. 2001, "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties", *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360.

[9] Griva, I., Nash, S. G., Sofer, A. 2009, *Linear and Nonlinear Optimization*, Second edition, Society for Industrial and Applied Mathematics, Philadelphia, PA.

[10] Hastie, T., Tibshirani, R., Friedman J. 2008, *The Elements of Statistical Learning*, 2nd edition, Springer, New York, NY.

[11] Hult, H., Lindskog, F., Hammarlid, O., Rehn, C. J. 2012, *Risk and Portfolio Analysis: Principles and Methods*, 1st edition, Springer, New York, NY.

[12] Izenman, A. J. 2008, *Modern Multivariate Statistical Techniques*, 1st edition, Springer, New York, NY.

[13] Kolesnikov, A., Logachev, Y., Topinskiy, V. 2012, "Predicting CTR of New Ads via Click Prediction", *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 2557-2550.

[14] Krushevskaja, D., Simpson, W., Muthukrishnan, S. 2016, "Ad Allocation with Secondary Metrics", *2016 IEEE International Conference on Big Data*, pp. 1202-1211.

[15] Liu, Y., Kliman-Silver, C., Bell, R., Krishnamurthy, B., Mislove, A. 2014, "Measurement and Analysis of OSN Ad Auctions", *2016 IEEE International Conference on Big Data*, pp. 139-149.

[16] Midha, V. 2009, "The Glitch in On-line Advertising: A Study of Click Fraud in Pay-Per-Click Advertising Programs", *International Journal of Electronic Commerce*, vol. 13, no. 2, pp. 91–112.

[17] Montgomery, D. C., Peck, E. A., Vining, G. G. 2012, *Introduction to linear regression analysis*, 5th edition, Wiley, Hoboken, NJ.

[18] Persson, A., Böiers, L. C. 2010, *Analys i en variabel*, 3rd edition, Studentlitteratur, Lund, Sweden.

[19] Tang, G., Yang, Y., Pei, J. 2013, "Price Information Patterns in Web Search Advertising: An Empirical Case Study on Accommodation Industry", *2013 IEEE 13th International Conference on Data Mining*, pp. 737-746.

[20] Wang, C., Chen, H. 2012, "Learning to Predict the Cost-Per-Click for Your Ad Words", *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 2291-2294.

[21] Zou, H. 2006, "The Adaptive Lasso and Its Oracle Properties", *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429.

[22] Zou, H., Hastie, T. 2005, "Regularization and variable selection via the elastic net", *Journal of the Royal Statistical Society Series B*, vol. 67, no. 2, pp. 301–320.

8 Appendix

8.1 Extended Mathematical Theory

The theory in this section is referred to the work of Montgomery et al. (2012) [17]. In regression modelling, the objective is to explain a numerical quantity, denoted y and called dependent variable, with other numerical quantities, denoted $x_1, x_2, \ldots, x_k$ and called covariates or independent variables. One uses existing data from n data points to estimate the model. That is, for each $y_i \in \mathbb{R}$, $i = 1, \ldots, n$, one utilizes the values of $x_{ij} \in \mathbb{R}$, $i = 1, \ldots, n$, $j = 1, \ldots, k$. This thesis uses multiple regression, in which the dependent variable is a function of the independent variables. A regression model in its simplest form is the linear regression model

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i. \tag{8.1}$$

β_0 is called the intercept and β_j is called the regression coefficients, which will be estimated from the data and from a chosen selection criteria, i.e. we want to find the $\hat{\beta}_j$'s that best fits the data. With the regression coefficients chosen from the distribution of β_j, the fitted data $\hat{y}_i$ satisfies the equation

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij}. \tag{8.2}$$

ϵ_i in equation (8.1) is the error term which is the deviation of the observed value from the true value and is defined as

$$\epsilon_i = y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij}. \tag{8.3}$$

A common concept is residuals, e_i, which is the deviation of the estimated value and observed value

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij}. \tag{8.4}$$

In order for the OLS-approach to work properly are certain criteria, called the Gauss-Markov assumptions, imposed which are crucial for OLS-regression:

1. There is a linear relationship, or at least partly linear, between the dependent variable and the covariates.

2. $\mathbb{E}[\boldsymbol{\epsilon_i}] = \mathbf{0}, \quad \forall i = 1, \ldots, n$, or in other words that the error term is zero on average.

3. $\mathrm{Var}[\epsilon_i] = \sigma^2, \quad \forall i = 1, \ldots, n$, or in other words is the variance of the error term constant for all data points.

4. $\mathrm{corr}[\epsilon_i, \epsilon_j] = 0, \quad \forall i \neq j$, or in other words that all error terms are uncorrelated.

5. $\epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad \forall i = 1, \ldots, n$, or in other words are all error terms normally distributed.

Some performance metrics are preferably calculated in a regression model. These metrics will be defined in this section. Let $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and it can be shown that

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = SS_R + SS_{Res}. \tag{8.5}$$

The mean squared errors are defined as

$$\begin{aligned} MS_R &= \frac{SS_R}{k}, \\ MS_{Res} &= \frac{SS_{Res}}{n-k-1}. \end{aligned} \tag{8.6}$$

It can be the case that a covariate included in the model is zero, which is equivalent to that covariate not contributing to explaining the dependent variable and should hence be left out of the model. Since it can be shown that

$$\begin{aligned} \frac{SS_R}{\sigma^2} &\sim \chi^2_k, \\ \frac{SS_{Res}}{\sigma^2} &\sim \chi^2_{n-k-1}, \end{aligned} \tag{8.7}$$

it holds that the F-statistic satisfies

$$F_0 = \frac{SS_R/k}{SS_{Res}/(n-k-1)} = \frac{MS_R}{MS_{Res}}. \sim F_{k,n-k-1} \tag{8.8}$$

This gives that if the F-statistic is large then it is likely that that at least one $\beta_j \neq 0$. Large in this case is of course in comparison with values from the $F_{k,n-k-1}$-distribution. The main result of the regression is the coefficient estimates $\hat{\boldsymbol{\beta}}$. Nonetheless, since $\hat{\boldsymbol{\beta}}$ is drawn from a distribution it could be the case that the error range of the estimate is large. A measure of the error range for estimate is the standard error, defined for each β_j-coefficient as

$$\text{se}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2((\mathbf{X}^\intercal\mathbf{X})^{-1})_{jj}}. \tag{8.9}$$

Using the standard error, it can be shown that the test statistic

$$t_{0,j} = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)}, \tag{8.10}$$

can be used for testing the hypothesis

$$\begin{cases} H_0: & \beta_j = 0 \\ H_1: & \beta_j \neq 0, \end{cases} \tag{8.11}$$

in which the null-hypothesis can be rejected if $|t_{0,j}| > t_{\alpha/2;n-k-1}$, where α is the confidence level. This holds since $t_0 \sim t_{n-k-1}, \quad \forall j = 1, \ldots, k$. In this thesis will the estimate together with the standard error and an indication of the level of the p-value for the hypothesis test (8.11) be reported. Using the sum of squares as defined in equation (8.5), the metric R^2 can be defined as

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}. \tag{8.12}$$

Since equation (8.5) gives that $0 \leq SS_{Res} \leq SS_T$, the R^2-value for a regression satisfies $0 \leq R^2 \leq 1$. The interpretation of R^2 is that a model can explain 100% of the variation in the data if $R^2 = 1$ and the model cannot explain the data at all if $R^2 = 0$. Hence, a high R^2-value is desirable. Nonetheless, R^2 has the characteristic that it increases if adding another covariate to the model, regardless if the model in fact gets better from the adjustment. To adjust for this downside of R^2 exists $R^2_{Adj.}$ (read as "adjusted-R^2") which increases if the the newly added covariate in fact improves the model

$$R^2_{Adj} = 1 - \frac{SS_{Res}/(n-k-1)}{SS_T/(n-1)}. \tag{8.13}$$

Other metrics included in the regression results are the AIC and BIC, defined as

$$\begin{aligned} \text{AIC} &= -2\ln L + 2(k+1) = \{\text{If OLS}\} = n\ln\left(\frac{SS_{Res}}{n}\right) + 2(k+1), \\ \text{BIC} &= -2\ln L + p\ln(k+1) = \{\text{If OLS}\} = n\ln\left(\frac{SS_{Res}}{n}\right) + p\ln(k+1), \end{aligned} \tag{8.14}$$

where L is the estimate of the likelihood-function from the regression. Both AIC and BIC are measures of the relative quality of the model in comparison to another model. Hence, AIC or BIC does not contribute with any information of the quality of the model unless it is compared to another AIC- or BIC-value. For both AIC and BIC is the aim to choose the model with the lowest AIC or BIC. Nonetheless, the two metrics may decide upon different models as the best. The regression results will be summarized, together with the p-values for the hypothesis tests in equation (8.11) and the analysis of variance (ANOVA) information, in the following table:

	Dependent variable		
	y		
	(Model 1)	(Model 2)	(Model 3)
x1	x (y)		
⋮			
Constant	u (v)		
Observations			
R^2			
$R^2_{Adj.}$			
F_0 (df = ;)			
AIC			
BIC			
Note:		*p<0.1; **p<0.05; ***p<0.01	

Table XI: x and u are regression estimates and (y) and (v) are standard errors for that specific coefficient.