



# Master's thesis

Department of Statistics

*Masteruppsats, Statistiska institutionen*

Nr 2019:02

## Online pay-per-click marketing – response function and marketing mix

*Karl Sigfrid*

Master's Thesis 30 ECTS, HT 2018

Supervisor: Jolanta Pielaszekiewicz

---



## Abstract

An advertiser will often have several optional online marketing channels to either choose between or to combine. An obvious question from the advertiser is how to allocate a fixed budget to maximize a desired output, such as the number of paid clicks leading the user to a landing page. We have here focused on two marketing channels for dentists with small practices, for each of which we developed a response function. These functions were subsequently combined into one joint function which can be used to get point estimates, optimal marketing weights and prediction intervals. We calculate the optimal mix of marketing channels, based on our models, to either maximize total response or minimize the variance.

In the thesis we get a considerably better model fit by basing the marginal response functions on the underlying logic of the ad auctions compared to a model that assumes linearity. While response functions developed for traditional marketing are less than ideal for pay-per-click advertisement, models suitable for production functions – in this case the generalized logistic function – is well suited to model Google ads auctions.

The number of paid clicks for Google ads is modeled using a piecewise function and a generalized logistic function that both manifest the diminishing returns on advertising spending. They also give similar accuracy on 10-fold cross validation on the training data.

We found that the Google ads marginal function increases with the number of healthcare professionals, and the effect is greater at large levels of spending. Moreover, increasing population density decreases the number of paid clicks decreases given the other variables of the model.

For the Facebook marketing channel, a linear model is appropriate within the variable space of the training data. We assume that the diminishing returns are unobserved i.e. that the Facebook model would diminish in a similar manner as the Google ads function with larger levels of spending.

The point estimate that maximizes the number of paid visits to a landing page for a potential client are found using a customized gradient descent algorithm. All implementation and analysis were done in R.

# Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
1.1	Research problem . . . . .	5
1.2	Outline . . . . .	5
<b>2</b>	<b>Data</b>	<b>6</b>
2.1	Google ads variables . . . . .	6
2.2	Facebook variables . . . . .	8
<b>3</b>	<b>Method and assumptions</b>	<b>10</b>
3.1	Assumptions on marketing response functions . . . . .	10
3.2	No-intercept models . . . . .	12
3.2.1	Measures of model accuracy . . . . .	13
3.2.2	Measuring multicollinearity . . . . .	13
3.2.3	Transformation for homoskedasticity . . . . .	13
3.2.4	A note on how the back transformation influences $\hat{\beta}$ and $\hat{\sigma}_{\beta}^2$ . . . . .	14
3.3	Fitting parameters and maximization of output with gradient descent . . . . .	15
3.4	Estimation of the variance of the error term . . . . .	17
3.5	Tests of the normality assumption . . . . .	18
<b>4</b>	<b>The Google response function</b>	<b>18</b>
4.1	Google ads auctions . . . . .	19
4.2	The Google response function . . . . .	20
4.2.1	Google response function – candidate 1 . . . . .	20
4.2.2	Google response function – candidate 2 . . . . .	22
4.3	Comparison of models 1 and 2 . . . . .	27
4.4	Transforming the Google ads variables for homoscedasticity and normality . . . . .	28
4.5	Results and comments . . . . .	31
<b>5</b>	<b>The Facebook response function</b>	<b>32</b>
5.1	Transforming the Facebook ads variables for homoscedasticity and normality . . . . .	33
5.2	Results . . . . .	35
<b>6</b>	<b>An aggregated model, optimal mix and predictions</b>	<b>36</b>
6.1	Additional assumptions for the joint model . . . . .	37
6.2	Point estimate of the total number of paid clicks . . . . .	37
6.2.1	An analytical solution for finding the weight mix with a fixed budget . . . . .	38
6.3	Prediction intervals for the marginal functions . . . . .	40
6.3.1	Back transformation . . . . .	40
6.4	Prediction intervals for the aggregated number of paid clicks . . . . .	40
<b>7</b>	<b>Applying the model - an example</b>	<b>41</b>
7.1	Point estimate of the maximal number of clicks . . . . .	41
7.2	Prediction interval for the maximal total number of paid clicks . . . . .	43
7.3	Minimizing the variance for marketing mix . . . . .	44
<b>8</b>	<b>Conclusions and discussion</b>	<b>44</b>



# 1 Introduction

Universal Avenue (UA) is an upstart company engaged in digitization of small businesses. UA offers, among other services, online marketing. Medical doctors and dentists with small practices are a major client category.

UA wants to explore whether they can use available data, generated within the company or available elsewhere, to make recommendations on an optimal mix of online marketing channels. The primary marketing channels offered are Google ads and Facebook ads.

We here address the question about allocation of money between marketing channels for a potential client who comes to Universal Avenue with the intention to spend a certain amount each month on attracting clients through online advertising.

The model is not assumed to be applicable to every type of business, but only to one specific business type. It is possible that a similar model with adjusted parameter values could capture the response functions for other markets as well, but that question falls outside of the scope of this work.

## 1.1 Research problem

The task will be to optimize a function with a multidimensional variable space under conditions of uncertainty, which will require data transformation and construction of marginal response functions.

Formally, the problem can be presented as follows:

Let  $\underline{w}$  be a vector of or weights of length  $k$ , where  $k$  is the number of marketing channels used. Then  $w_i$  represents the proportion of the marketing budget that is allocated to marketing channel  $i$ . What combination of values in  $\underline{w}$  will maximize the expected output?

The output to be maximized can be measured in various ways. It could for instance be the number of visits to a landing page (the destination when clicking an ad) generated by the campaign, or it could be the value in SEK generated by the advertisement. If the response variable is measured in SEK, the value may be determined by several factors. For instance, how likely is a website visitor to book an appointment, and what is the expected lifetime value of a new customer?

In addition to website visits, impressions (views) can also be attributed value in themselves as it increases brand recognition.

*In this thesis, we will measure the number of landing page visits, which will also be referred to as the number of clicks.*

Our goal is to find a *suitable model* that allows us to obtain a *weight vector* for marketing channels with regards to the observed variation. The result is not limited to a point estimate but will express the uncertainty that goes into the model as a consequence of randomly distributed parameters.

## 1.2 Outline

The thesis has the following outline.

Firstly, the discussion on available data and choice of assumptions and methods is given in Chapters 2 and 3. Here we rely on marketing research articles such as the work by J. Saunders (1987) that explores response functions and the conditions under which they may be reasonable. [14]

Secondly, the response functions for the respective marketing channels are defined in the Chapters 4 and 5. The response function is the function that maps an input, such as spending on a particular marketing channel, to an output, for instance the resulting number of paid clicks.

At this stage we will determine how well the dependent variable can be explained with available data. Here several different types of functions will be used as response functions – all of which are consistent with the basic assumptions.

In Chapter 6 the response functions for the different marketing channels will be merged into one single model. This merged model will be used to obtain the weight for each respective marketing channel.

Chapter 7 presents the results regarding an optimal marketing mix for a small dental practice (4 healthcare professionals) in Solna.

Finally, in Chapter 8, we conclude and discuss the results.

## 2 Data

Through Universal Avenue (UA) we have access to data from current clients’ Google ads accounts. Facebook advertisement data are also available. The analysis concerns the population of UA clients in the dentist sector and is not necessarily generalizable to businesses in other fields or other countries.

The limitations set by available data will be a central part of the thesis. An analysis of what additional data that could improve the model will result in a recommendation for Universal Avenue’s future data collection activities.

The identities of clients who advertise cannot be disclosed, a restriction which should not matter for the purposes of this project. The data was generated in 2017 and 2018. It has been gathered and organized for this thesis in January 2019.

### 2.1 Google ads variables

The Google Ads variables were collected from the Google Ads administrative system. The dataset has 491 observations generated during the period November 2017 - December 2018.

	Number.of.clicks	Monthly.spending	Population.density	Number.of.qualified.employees
Mean	138.27	1297.57	676.86	6.44
Median	131.00	1375.52	111.00	5.00
Standard deviation	52.63	555.98	1249.52	4.28
Minimum	5.00	84.18	8.80	3.00
Maximum	321.00	4329.79	5074.70	25.00

Table 1: Key statistics of Google ads variables

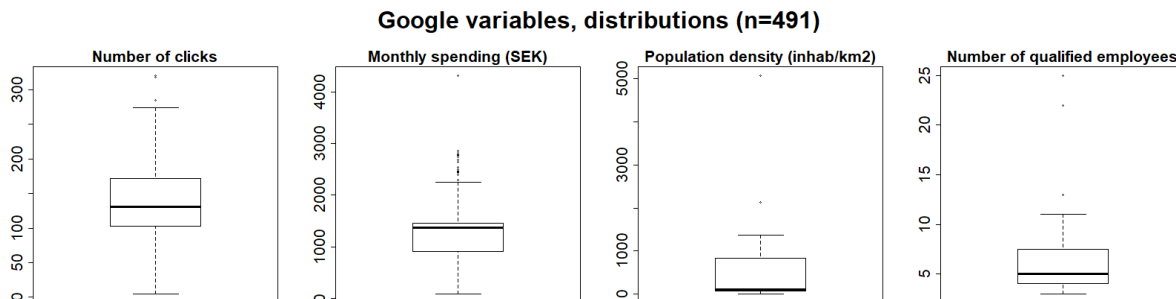


Figure 1: Google ads variable distributions.

## Google Ads variables, pairwise plots

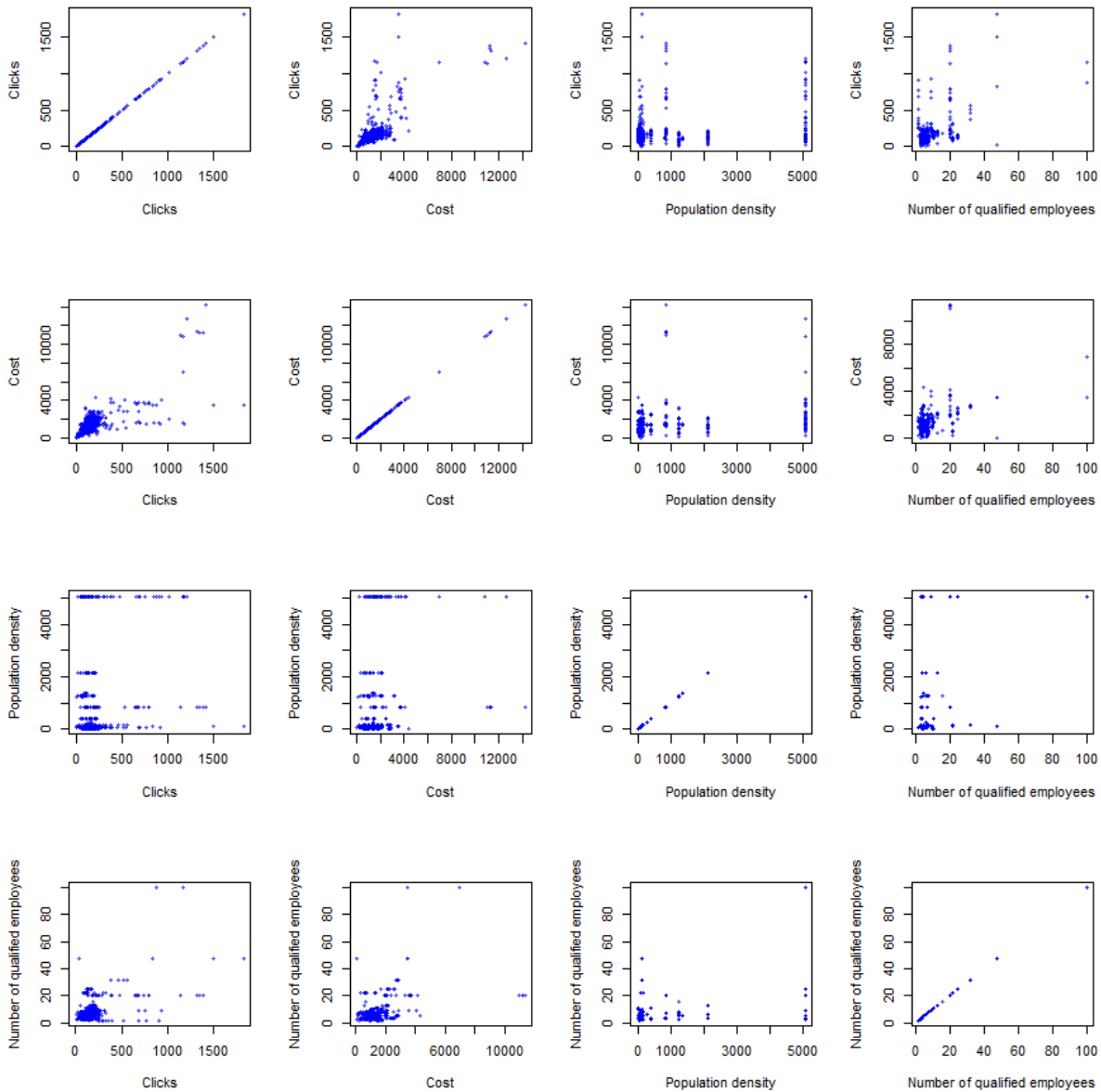


Figure 2: Relationships between pairs of variables for Google ads.

Of main interest is how the number of clicks can be modelled as a function of cost. In Figure 2 we see that the number of clicks increase with cost. However, as cost increases there seems to be a diminishing marginal return. A line that captures the relationship would therefore not be a straight line but rather a line with a diminishing slope.

We see that cost increases with the number of healthcare professionals, which is not surprising. Larger dentist's offices tend to spend more on Google ads than smaller ones and get more clicks. Whether the positive correlation between the number of clicks with the size of the doctor's office is only due to more spending or also due to other factors cannot be easily seen visually from these plots, but we will explore further the significance of this variable.



The Google ads variables that we will use in our model are:

- Clicks (response variable);
- Cost;
- Population density in the municipality;
- Number of healthcare professionals at the clinic.

## 2.2 Facebook variables

The Facebook data comes with different variables than the Google data. In the following plot we will look at Impressions, Clicks, Number of conversions and Cost.

The dataset has 473 observations generated during the year 2017. They were accessed through Facebook ads administrative system, where key figures from current and previous campaigns are given. We can see in Figure 3 that we have a large number of outliers in the upper end of the scale for all variables.

We can note in Table 2 the large number of impressions generally necessary to generate a modest number of clicks and an even smaller number of conversions.

	Number.of.impressions	Number.of.clicks	Number.of.conversions	Monthly.spending
Mean	207692.05	30.21	3.16	511.68
Median	98606.00	12.00	1.00	196.60
Standard deviation	327948.91	46.17	5.15	808.08
Minimum	1884.00	1.00	0.00	1.80
Maximum	3052003.00	353.00	60.00	6399.50

Table 2: Key statistics of Facebook ads variables

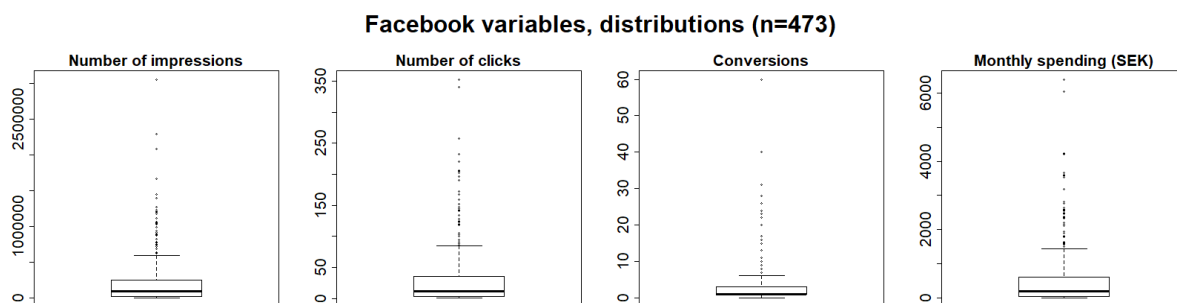


Figure 3: Facebook ads variable distributions.

### Facebook variables, pairwise plots

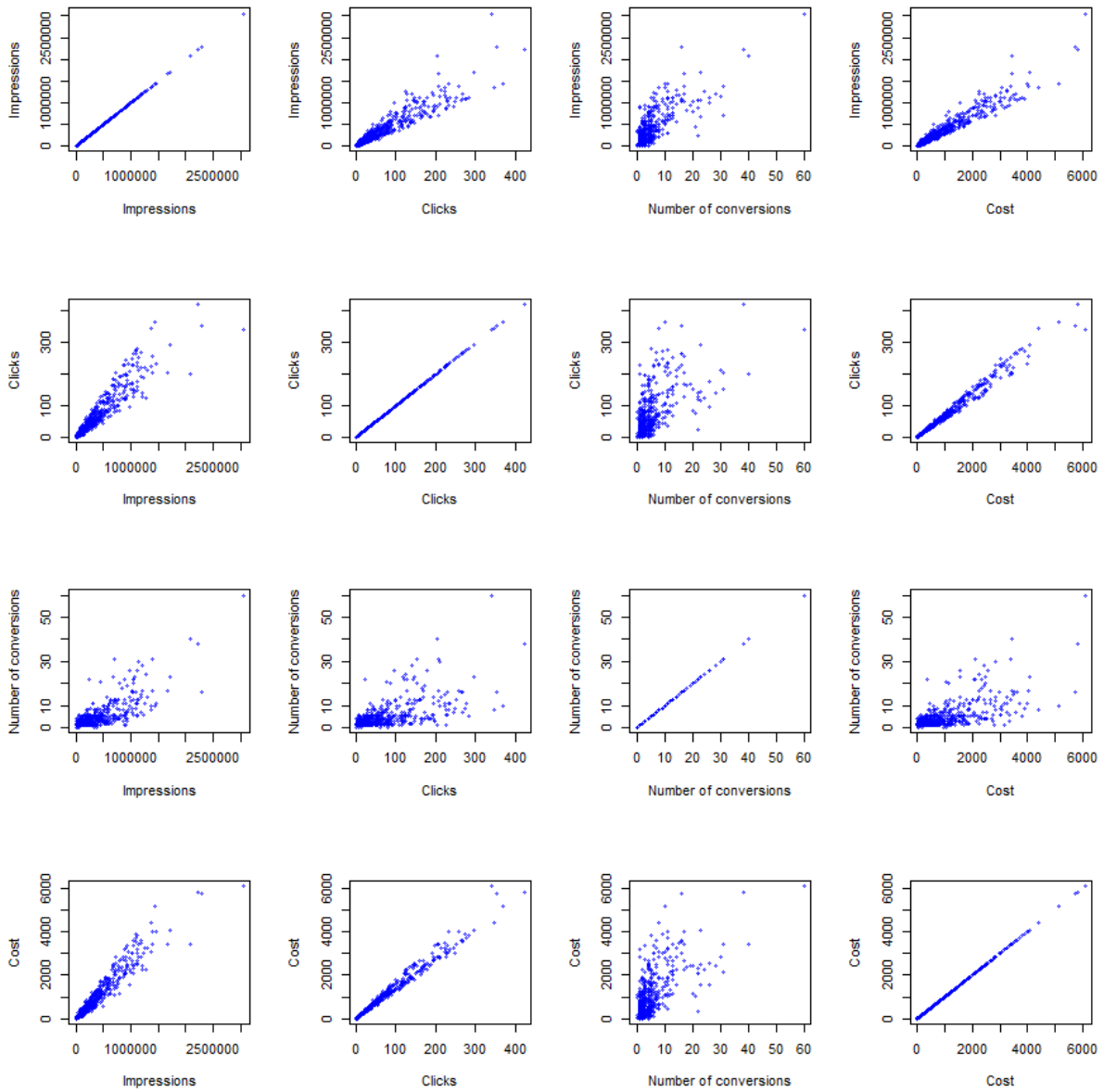


Figure 4: Pairs for facebook variables plotted against each other.

An "impression" is an occasion at which the ad is displayed. "Conversion" refers to some action that the user can take at the landing page and that is measured. Here, conversion means that the user has used a tool on the landing page to contact the business for further information about the service, which brings them a step closer to ordering the service.

The relationship between clicks and cost, which are the key variables in this study, is to a high extent linear, and the variance appears quite small. We also see that the variance increases both with the cost and with the number of clicks.

The relationship between clicks and impressions is also linear, which means that we have a roughly constant click-through rate (clicks/impressions).

From a visual inspection, the number of conversions per click appear to be roughly constant or perhaps slightly exponential.

While the conversion rate is an important tool for transforming the number of clicks into an expected value that can be compared to the investment, we will not use it here as our response variable since the variable is not available for our Google ads data. Using different response variables for the Google ads model and the Facebook ads model would open up an entirely new question about the relative value of a conversion and a paid click.

Our Facebook data is received in an anonymized format, which means that we cannot connect data points to specific businesses. This means that we are not able to add variables such as Population density and Number of healthcare professionals. However, as we can see from the plots above, we are still in a good position since the number of clicks can be well modelled as a function of cost.

The Facebook variables that we will use in our model are:

- Clicks (response variable)
- Cost

### **3 Method and assumptions**

Throughout the thesis we will develop three functions:

- A Google ads response function
- A Facebook ads response function
- A joint Google ads-Facebook ads response function

Each of these functions will be developed in their own sections. In this section, we will look at the methodology that will be used to determine functional forms and to fit parameters.

Moreover, we look at the assumptions used as a basis for our functions.

#### **3.1 Assumptions on marketing response functions**

Many of the response functions described in available marketing literature are constructed to fit the logic of traditional marketing campaigns on tv and in newspapers. In the following, we will discuss to what extent they are applicable to pay-per-click online marketing.

J Saunders (1987) has listed some of the most common assumptions on which to base marketing response functions. [14] The assumptions will be categorized according to their usefulness in a pay-per-click context.

We will make the same assumptions for the Google ads response function as for the Facebook ads response function.

## Model assumptions

- A1. **Effect is zero when effort is zero.** When the effect of a marketing campaign is measured in total sales, as it often is in traditional models, zero marketing normally does not correspond to zero sales.

In our case, however, the response variable measures not total sales but the number of visits to a landing page through clicks on a paid link. While a dentist will certainly have clients even without online marketing, the response measured as the number of paid clicks will always be zero when the online marketing spending is zero.

- A2. **There are decreasing returns to scale with effort.** For advertising that aims to increase awareness, the spending will at first result in an increased number of unique recipients of the ads, but as the budget increases an increasing proportion of the ad recipients will be people who have already seen the ad. After a number of exposures, the ad will likely have little effect on the recipient.

With pay-per-click search advertising, there is a limit to the ad market size, and the cost per click can increase as a result of one advertiser driving it up in the Google ads auction or Facebook ads auction.

Note that even if we assume a diminishing return on investment, the data points may fall in a variable space in which we do not see this diminishing shape. In such cases, we may have to model a linear relationship within the known variable space while we are aware that the function will behave differently at variable values outside of the known variable space.

- A3. **Saturation.** Saturation occurs when an advertiser has taken a 100 percent market share and there is no potential for the market to grow. At that point, additional advertising will make no difference.

We do not expect a dentist who advertises online to capture the entire local market and reach saturation in that respect. However, we can reach saturation in terms of ranking on top at every ad auction on Google or Facebook, in which case additional spending will not be possible.

In the case of Google ads that show when a search is performed, these limited markets can indeed be saturated. The market for Facebook ads targeted towards a group of users, for instance based on residence, will typically be larger and saturation therefore less likely.

## Common response function assumptions not applicable to the model

- **There is a linear relationship between cause and effect.** A linear relationship implies that the marginal benefit of increased spending stay constant for any level of spending. This means that a successful marketing campaign can be scaled infinitely and keep giving the same return on investment, which is unreasonable.

Locally, however, a linear function may be a good approximation even if the true relationship is more complicated. Therefore, the rejection of this assumption does not rule out using a simple linear model while acknowledging that the model is only valid within a certain variable space.

- **There are increasing returns to scale with effort, or there are first increasing and then decreasing returns to scale.**

Increasing marginal returns, primarily at small levels of spending, has been motivated by large-scale marketing channels being more economical. It is difficult, though, to find empirical evidence of exponentially increasing response functions.

- **Threshold effects**

It has been claimed that marketing efforts must pass a threshold before any real effect occurs. This is an assumption, implicit or explicit, when fitting an S-shaped response function where the initial spending gives almost no result.

Reviews of the empirical evidence reject the notion of these thresholds in general. In the specific case of online marketing that follows a pay-per-click logic there can be no cost without a corresponding benefit in terms of landing page visits. Furthermore, there is no argument to be made for an initially high cost per click.

- **Supersaturation.**

Supersaturation occurs when the marketing effort is so massive that the marginal effect of additional spending is negative. A reason for such effect could be that people react negatively to marketing that is considered too aggressive.

Regardless of whether the effect can be seen in traditional marketing, the concept cannot be applied in our context where additional spending is always accompanied by additional clicks.

### 3.2 No-intercept models

There are several test-based approaches to choosing whether to use a model with or without intercept. One is to fit a regression model with an intercept and then decide whether the intercept parameter is significant (Eisenhauer, 2005). [5] Another, is to include a false datapoint – a leverage point – that draws the intercept through the origin (Casella, 1983). [3] If the false datapoint needs to be placed so that it becomes a clear outlier, it may be an argument against skipping the intercept term. If, on the other hand the false data point appears plausible, a regression through the origin is acceptable.

The test-based approaches could be described as context neutral. They consider data points while they do not need to consider the underlying reality.

When building a model that emphasizes interpretability, and that aims to reflect underlying mechanisms, there can be strong apriori reasons for omitting the constant. We sometimes know for certain that when  $X = 0$ , then  $Y = 0$ . (Eisenhauer, 2005) [5]. An example is the Cobb Douglas production function in which an output depends on a mix of resources, typically capital and labor, which we have to our disposal. We don't need to study empirical data to know that when the resources are zero, the output is also zero. Other similar observations are that a land of size zero will produce no crop.

As explained in assumption A1 in Section 3.1, we know for certain that zero advertisement spending results in zero advertisement value. Therefore, a model that claims to reflect the underlying realities needs to go through the origin. *We will here use a non-intercept model.*

No-intercept models are widely discussed in literature, and it is possible to criticize such models if the inclusion of an intercept improves the fit. Ultimately it is a choice to be made.

As we will see, the decision to drop the intercept will have consequences when we evaluate the model, since the intercept is central to some evaluation metrics such as R-squared and the variance inflation factor (VIF). It is well established that with a non-intercept model, measures of R-squared and VIF can be highly misleading.

### 3.2.1 Measures of model accuracy

A common measure of the goodness of a regression model is R-squared, which tells us how much of the total variance in the data that can be explained by the model. It has, however been pointed out that in some situations R-squared is a measure from which it is easy to draw false conclusions.

Throughout the literature, there are many different mathematical expressions for R-squared. With a linear model with an intercept, these expressions are all equivalent. However, as Kvålseth (1985) has pointed out:

'If the models are anything but linear with an intercept term, it is not unlikely that the analyst will use an inappropriate R-squared statistic and end up with possibly misleading results', (p. 279, [10]).

For non-intercept models the the different expressions for R-squared lead to different outcomes. As different software packages implement the calculation of R-square differently, it is often not obvious which expression a no-intercept R-squared has resulted from.

For the reasons mentioned above, R-squared is not an appropriate measure for our no-intercept model. Instead we fit our parameters and evaluate the model using the Mean squared error (MSE) as our loss function.

We could also have opted for the Root mean squared error (RMSE) or the Mean absolute error (MAE). RMSE and MSE are equivalent in this context, since they are both minimized at the same value in the variable space.

According to Chai and Draxler (T. Chai & R. R. Draxler, 2014), MSE/RMSE is a more appropriate measure than MAE for models with normally distributed error terms: 'The RMSE is more appropriate to represent model performance than the MAE when the error distribution is expected to be Gaussian.' [4]

### 3.2.2 Measuring multicolliniarity

Multicollinearity in a regression model means that an explanatory variable can be linearly predicted by at least one other another explanatory variable. Multicolliniarity can be measured as the Variance of inflation factor(VIF), which is calculated for each explanatory variable.

The VIF for one explanatory variable gives a measure of how well that specific explanatory variable can be explained by all the other explanatory variables in the model. Particularly, for the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

VIF can be calculated for  $X_1$  by fitting the regression

$$X_1 = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

and then calculating

$$\frac{1}{1 - R_{X_1}^2}.$$

### 3.2.3 Transformation for homoskedasticity

Let the standard deviation of the model's residuals be proportional to a power of the estimated response variable  $\hat{Y}$

$$\sigma_Y = k\hat{Y}^a$$

If we then transform the response variable so that

$$Y^* = \frac{Y}{\hat{Y}^a},$$

Montgomery (2009) points out that we will achieve constant variance: [12]

$$\sigma_{Y^*} = k \frac{\hat{Y}^a}{\hat{Y}^a} = k$$

The special case of setting  $a$  to 0.5 is known as a square root transformation.

We can note that the principle described by Montgomery also applies when the variance of the error term is correlated with a power of an explanatory variable  $X_j$ :

$$\sigma_Y = kX_j^a, Y^* = \frac{Y}{X_j^a} \rightarrow \sigma_{Y^*} = k \frac{X_j^a}{X_j^a} = k$$

How then to set  $a$  to an appropriate value? According to Montgomery (2009), 'many experimenters select the form of the transformation by simply trying several alternatives and observing the effect of each transformation on the plot of residuals versus the predicted response. The transformation that produced the most satisfactory plot is then selected.' [12]

It is however possible to undertake a more systematic search for an  $a$  that results in homoskedasticity. We can, for instance, let  $a$  vary within an interval and note how the p-value for a chosen homoskedasticity test varies with  $a$ . This is how we will find a value for  $a$  that lets us satisfy both the homoskedasticity assumption and the normality assumption.

### 3.2.4 A note on how the back transformation influences $\hat{\beta}$ and $\hat{\sigma}_\beta^2$

In this section we discuss the back transformation of the transformation for homoscedasticity presented in the previous section. Notice that although the point estimate and the variance of the  $\hat{\beta}$  change when we backtransform our results, they are still unbiased estimators of the same underlying regression parameters.

When we say that  $\beta$  parameters do not change with the back transformation, this does not mean that the  $\beta$  estimates  $\hat{\beta}_i$  received with the transformed variables are the same as when the regression is performed with the original variables. The same is true for the variance. As the left plot in Figure 5 shows, the estimated regression line does differ depending on whether it has been fit using original variables or transformed variables.

This difference is a result of the data points changing their relative weights. In the following transformation, we divide the regression model by the  $j$ :th explanatory variable to the power of  $a$ .

$$Y = \sum_{i=0}^p \beta_i X_i + \epsilon \rightarrow \frac{Y}{X_j^a} = \sum_{i=0}^p \beta_i \frac{X_i}{X_j^a} + \frac{\epsilon}{X_j^a} \quad (1)$$

While the transformation changes the  $\beta$  estimates, they do so in a random manner as long as the the linear assumption and the normality assumption hold. The homoskedasticity assumption, however, does not need to hold.

An crucial property of the transformation is that the "true" values of the  $\beta$  parameters do not change. We can see this from Equation 1, and it is also illustrated by the right-hand plot in Figure 5. The plot shows the convergence of two  $\beta$  parameters from a linear regression model. We can see that asymptotically, as the number of data points increases, the  $\beta$  estimates when using the transformed variables converge with those that result from using the original variables.

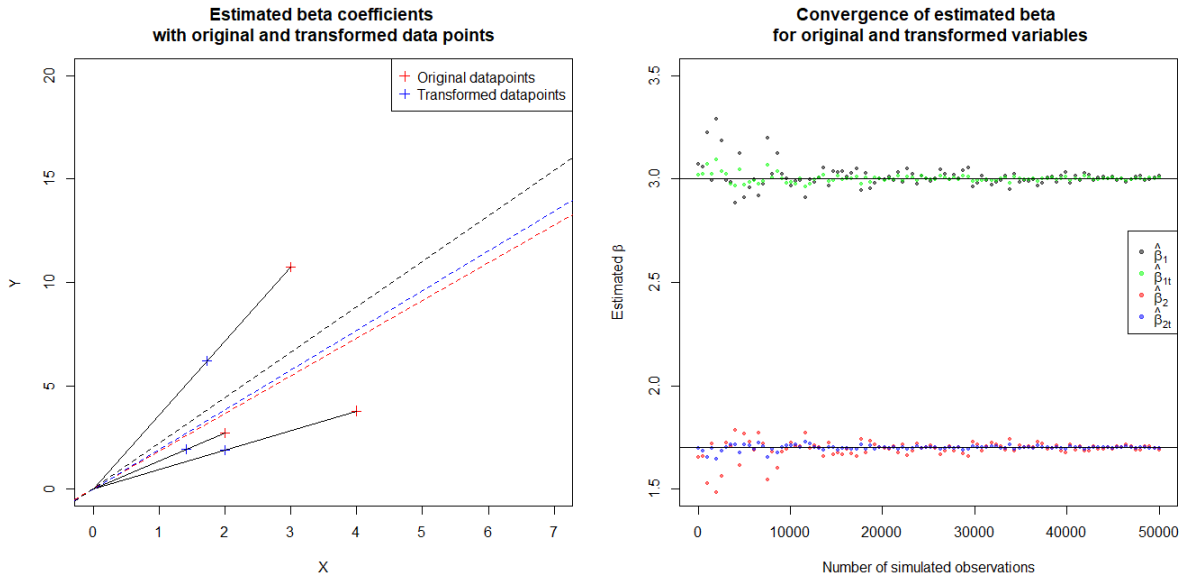


Figure 5: Left: The transformation shifts the positions of each data point along a straight line between the original data point and the origin. While the ratio  $\frac{Y}{X}$  remains the same, the relative weights of the points change. Right: The  $\hat{\beta}$  estimates for the original variables and their transformed variables converge toward the true parameter values, which are the same before and after transformations.

This is a very useful property of the chosen variable transformation which many other variance stabilizing transformations do not share.

### 3.3 Fitting parameters and maximization of output with gradient descent

In this thesis, the gradient descent algorithm plays an important role, and it will be used for two separate optimization purposes:

- Fitting parameters in variable transformations.** After choosing our models we need to fit the parameters. While the  $\beta$  parameters for the Google ads model and the Facebook ads model respectively are fit with a regular linear regression, we must also fit the parameters used for variable transformation. A straightforward way of doing this, which does not depend on the mathematical properties of the function, is by applying a gradient descent algorithm. We fit the parameters to the data by minimizing MSE.
- Optimizing the marketing mix.** When the marginal models are defined, we want to set the values of the variables for Google ads spending and Facebook ads spending respectively to maximize the number of paid clicks.

The gradient descent algorithm uses the first derivative of each dimension, the gradient, of the variable space. After calculating the gradient we take a step in the opposite direction of the gradient as this is the direction in which the descent is at its steepest.



$$\underline{x}^{k+1} = \underline{x}^k - \underline{\eta} \nabla f(\underline{x}),$$

where  $\underline{x}$  is a vector giving a position in the variable space or parameter space,  $k$  is the index and  $\underline{\eta}$  is a step size which should be calibrated so that the algorithm ideally does not overshoot the target while also being reasonably fast.

An alternative to differentiating the function is to approximate the gradient. Recall that the derivative of  $g(a)$  is  $\lim_{\epsilon \rightarrow 0} \frac{g(a + \epsilon) - g(a)}{\epsilon}$

We can approximate the derivative by setting  $\epsilon$  to a very small number, being aware that this can be a computationally inefficient method. An advantage, however, is that we get a gradient descent algorithm that is function agnostic.

An addition to the gradient descent algorithm allows it to find not only the global minimum where the gradient is zero but also the global minimum when it is located along a constraint line or at a corner point.

Whenever the standard algorithm takes a step from  $k$  to  $k + 1$  and thereby places us outside of the permitted variable space, we follow a special procedure which here, for simplicity, assumes a two-dimensional variable space:

1. Calculate  $f(x_1^{k+1}, x_2^*)$ , where  $x_2^*$  is a value of  $x_2$  as close as possible to  $x_2^{k+1}$  given the constraints and the value of  $x_1^{k+1}$ .
2. Calculate  $f(x_1^*, x_2^{k+1})$ , where  $x_1^*$  is a value of  $x_1$  as close as possible to  $x_1^{k+1}$  given the constraints and the value of  $x_2^{k+1}$ .
3. Of the vectors above, choose that which gives the lowest value of the loss function. If the loss function value at the new coordinates is higher than the value at the previous coordinates, divide the stepsize by 2.

The algorithm stops, as usual, when the last step taken is smaller than a chosen threshold value.

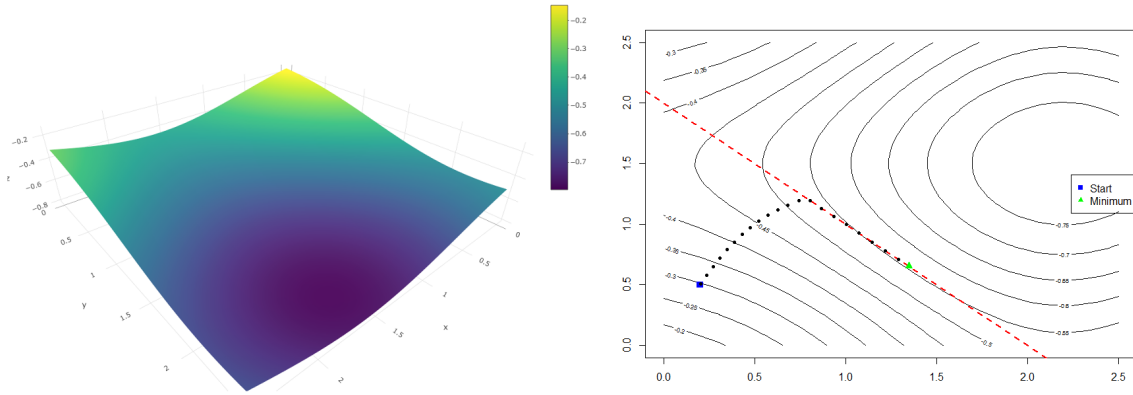


Figure 6: Left: The surface of function to minimize. Right: The adjusted gradient descent function can follow constraint lines to find the minimum within the constrained variable space. This is an illustrated example of a function which has its minimum outside the allowed variable space. The optimal point is found on one of the constraint lines.

In our case, we will want to optimize the allocation of funds between two different marketing channels. In that context, the constraint line represents the total amount that can be spent. If it is optimal to spend the entire sum on online marketing, we will end up somewhere on the constraint line, and our optimization method must be able to handle that.

*The R code for the modified gradient descent algorithm described in this section can be found in Appendix 1.*

### 3.4 Estimation of the variance of the error term

The variance  $\sigma^2$  of the error term  $\epsilon$  can be estimated as

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=0}^n (Y_i - \hat{Y}_i)^2,$$

where  $p$  is the number of parameters in the regression model and  $n$  is the sample size.

The standard deviation of the error term is estimated by the standard error (B.S. Everitt & A. Skrondal, 2010). [6]

$$s.e. = \hat{\sigma} = \sqrt{\hat{\sigma}^2}.$$

One way to calculate the standard error is to measure the squared differences between the data points in the training set and their corresponding estimated values according to the model. However, if the model is overfitted the measured error on the training set may underestimate the error on a new dataset. For instance, if we introduce new variables with no or little explanatory value, these variables will enable the model to adapt closer to the random noise, thus decreasing the error on the training set while making the model less generalizable (J. Brownlee, 2014). [2]

Since we are interested in estimating the error on new data, we want an estimate of the error term variance based on a test set that is separate from the training set.

One option is to divide the data into two sets, one for training and one for evaluation. Another option, which makes better use of the available data, is K-fold cross validation. The K-fold cross validation algorithm is as follows:

1. Divide the dataset into  $K$  (almost) equally sized subsets.  $K = 10$  is often considered a default.
2. For each subset  $i$ , where  $i$  goes from 1 to  $K$ :
  - (a) Train the model using all data except the data in subset  $i$ .
  - (b) Calculate the accuracy metric of interest, in our case  $s.e._i$ , using subset  $i$  as testset.
  - (c) Store  $s.e._i$ .
3. Calculate the mean of the chosen accuracy measure  $\frac{1}{K} \sum_{i=1}^K s.e._i$

*In the R programming language, K-fold cross validation for a variety of model types is included in the "caret" package.*

### 3.5 Tests of the normality assumption

The Kolmogorov Smirnov (K-S) test for normality compares a sample to a normal reference distribution with the same mean and standard deviation. The K-S statistic is the absolute of the maximum distance between the empirical CDF and the reference CDF. For small sample sizes, the Kolmogorov Smirnov test has little power and seldom rejects the normality null hypothesis. However, at a sample size of 50 or larger, the power of the test will be greater than 0.95. (A. Vexler, A. D. Hutson & X. Chen, 2017) [15]

The Shapiro-Wilk (S-W) test is more complicated than the K-S test and cannot be generalized to non-normal distributions. However, it has the greatest power of the commonly used tests of normality. This makes it a good test if we want to be conservative in our assumptions (Lohninger, 2012). [11]

We will in some instances present both the K-S and the S-W statistic. However, when we prepare the dataset for a linear-regression analysis, we will use the more demanding S-W test to ensure that the normality assumption is met.

## 4 The Google response function

In this section we will:

1. Describe the Google ads auctions.
2. Develop two alternative models which both aim to capture the Google ads logic. The difference between the two models consists of the transformations of the explanatory variables.
3. Evaluate and choose one of the two candidate models.
4. Perform further variable transformations to ensure compliance with the assumptions of the linear regression model.

## 4.1 Google ads auctions

When a user searches with a specific query, advertisers that bid on keywords relevant to that query get their ads ranked. The ranking determines the order in which the ads are displayed.

Google ads advertisers pay per click, not per impression. When your ad is displayed but not clicked, you do not pay anything. If an ad turns out to be unattractive and rarely clicked, it will be displayed less often.

The ad position is based not only on the bid but also on quality factors, the purpose of which are to prioritize ads that contribute to a better user experience.

The quality factor involves the expected clickthrough rate, the landing page experience and the relevance of the ad to the search. The formatting of the ad is also scored to incentivize formatting that complies to Google standards.

The Ad Rank score, which determined the order in which ads are displayed, is a function of the bid, the quality score and the formatting score. Ads with insufficient Ad Rank scores may not be displayed at all.

Google

tandläkare bromma

Allt Kartor Bilder Nyheter Shopping Fler Inställningar Verktyg

Ungefär 701 000 resultat (0,40 sekunder)

**Tandläkare Bromma | Hög kvalitet till låga priser | tandläkarebromma.nu**  
[\[Annon\] www.tandläkarebromma.nu/](#) ▼  
Solna Dental - Välkommen till vår toppmoderna klinik, tio minuter från Bromma. Boka tid hos din...  
Kontakta oss · Tandhygienist · Om oss · Behandlingar

**Tandläkare i Bromma | Undersökning 199kr**  
[\[Annon\] www.novodental.se/tandläkare](#) ▼ 08-505 054 17  
Just nu basundersökning 199 kr för nya patienter, boka din tid redan idag. Välkommen in! Få professionell tandvård och individuellt anpassade behandlingar av högsta kvalitet. Senaste tekniken.  
Kontakta oss · Om oss · Våra priser · Avbetala din behandling · Våra aktuella erbjudanden  
📍 Landsvägen 79, Sundbyberg - Stängt nu · Tider ▼

**Privat tandläkare Stockholm | Tandvård utan långa väntetider**  
[\[Annon\] www.aquadental.se/tandläkare/stockholm](#) ▼ 010-555 97 15  
Tandvård med kvalitet till rätt pris. Öppet 365 dagar om året, kontakta oss! Generösa öppettider. Prisivård tandvård. Betala med Klarna. Tjänster: Akut tandvård, Estetisk tandvård, Käkkirurgi, Rotfyllning, Tandreglering, Tandställning, Tandimplantat, Skalfasader.  
Våra priser · Boka tid · Kliniker i Stockholm · Behandlingar · Akut tandvård Stockholm

Figure 7: Ranked ads on Google search. The ranking order is determined by assessed ad quality and bid size.

A consequence of the quality and formatting scores is that the number of clicks that two ads receive can be very different even if they spend the same amount of money. The ad with higher quality score will be treated more favorably. Since ad quality can only be predicted to a certain extent, we will here assume that all ads in our model are of the same quality and that the quality is fixed so that only the bid size determines their position. This is a reasonable assumption since all ads in the model are produced by the same ad agency in a standardized manner.

Note that the assumption of equal quality only concerns the comparisons between ads administered by UA. We do not need to assume equal quality for all competing ads on the Google network.

The advertiser pays the minimum amount necessary to retain their position in the ranking.

The amount paid by the advertiser is given by the answer to the following question:

Given the quality and formatting scores of the ranked ads and given the bids of the other ads, what is the lowest bid that the clicked advertiser could have given and still retained the current position in the ranking? This lowest possible bid is the cost of the click.

## 4.2 The Google response function

By testing the significance of available variables, we have decided on an overall structure for the Google ads response function:

$$Y_1 = \beta_{11}X_{11} + \beta_{12}X_{12} + \epsilon = (\beta_{11} + \beta_{12}D)X_{11} + \epsilon, \quad (2)$$

where  $Y_1$  is the number of paid clicks,  $X_{11}$  is a transformed cost variable and  $X_{12}$  is an interaction between  $X_{11}$  and the population-density variable, i.e.,

$$X_{12} = D \cdot X_{11},$$

where  $D$  is the population density in the municipality. As we shall see in the subsequent parameter estimates, the population density variable  $X_{12}$  has a negative impact on the number of clicks, given  $X_{11}$ . This could possibly be an effect of competition that increases with population density. The fact that the interaction variable  $X_{12} = D \cdot X_{11}$  gives a better fit than  $X_{12} = D$  would imply that the negative impact of high population density increases with the amount spent on Google ads. An additional benefit of using the  $X_{12}$  interaction rather than the original  $D$  variable is that we force the expected value of the total model through the origin, which we want to do in accordance with assumption A1 in Section 3.1.

In the following sections we will develop two alternative models to serve as the Google ads response function. They will both fit into the overall structure of Equation (2). Both candidate models will also aim to model the logic of the Google ads auctions described above. Notice that **the difference between the candidate models consists only in different transformations of the cost variable  $Q$  to  $X_{11}$** . We will refer to Equation (2) as our **model framework**, which both Google ads candidate models have in common.

The transformation of  $Q$  to  $X_{11}$  will be the topic of Sections 4.2.1 and 4.2.2.

### 4.2.1 Google response function – candidate 1

An assumption for the first model candidate – which is consistent with previous assumptions made in Section 3.1 – is that the market-price impact of smaller purchases of pay-per-click advertisement is non-existent or negligible. The cost-per-click is therefore approximately constant for small purchases. However, as the spending increases beyond a certain threshold point, we get diminishing returns (see assumption A3). Where this threshold point is located would depend on the market size. It is worth to notice that the markets for ads associated to certain search queries can be quite small. Not that many searches are made every month for dentists in a small town. Therefore, a single dentist purchasing ads may drive up the market price of a click for that query.

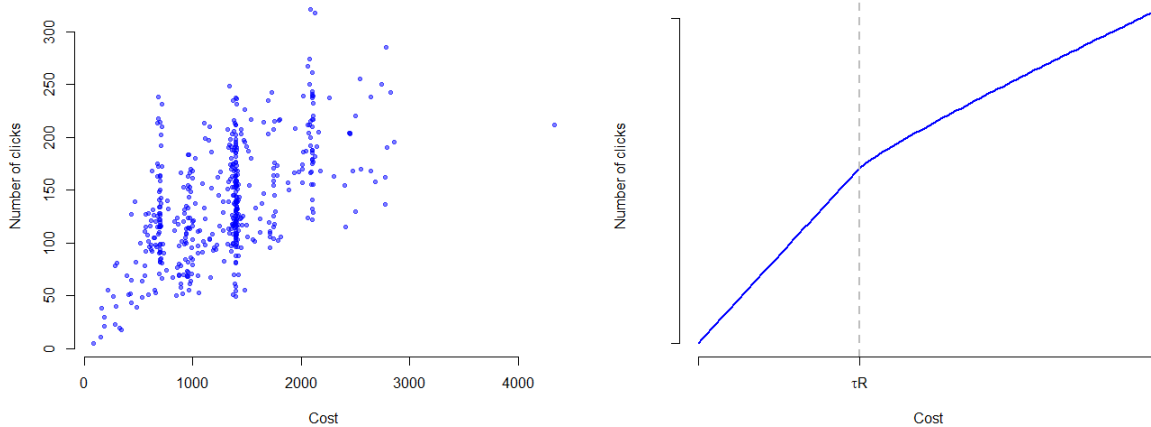


Figure 8: To the left, the data points. To the right, the proposed transformation of the cost variable in the Google ads response function. The threshold point is represented by  $\tau R$  in Equation (3).

In this model, we assume the number of clicks initially increases linearly, but beyond the threshold point the market begins to saturate and the marginal return diminishes. The position of the threshold is in this model candidate a function of the market size which we approximate with the number of healthcare professionals employed at the clinic.

A larger market size – in terms of a broader scope of services, more available time slots for visits, etc – will likely correspond to a larger ad market on Google ads. The larger market size, the more we expect the advertiser to be able to spend before the marginal returns begin to diminish.

The variable  $X_{11}$  will in candidate model 1 be the result of the following transformation, according for the reason stated above:

$$X_{11} = \begin{cases} Q, & \text{if } 0 \leq Q \leq \tau R \\ \tau R + (Q - \tau R + 1)^\gamma - 1 & \text{if } Q > \tau R \end{cases}, \quad (3)$$

where  $Q$  is the original cost variable,

$R$  is the number of healthcare professionals at the worksite,

$\tau$  is a parameter that determines the (linear) relationship between the number of healthcare professionals  $R$  and the threshold point  $\tau R$ ,

$\gamma$  is a parameter controlling at which rate the marginal number of clicks diminishes when  $Q > \tau R$ .

A model which assumes diminishing returns from the start is the special case of our model where  $\tau = 0$ . In the case where  $\gamma = 1$  we get that  $X_{11}$  is identical to the untransformed cost variable  $Q$ .

A response function in which the cost per click increases with spending would in theory approach infinite marginal cost to obtain additional clicks. However, in practice investors will limit their spending when marginal return is considered insufficient. Even an advertiser willing to pay very large sums of money for additional clicks is generally not able to do so. Because of the nature of the auctions – recall that the advertiser pays only the minimum amount to retain a certain position in the ranking – an extreme bid will not result in an extreme cost per click assuming that other bidders are rational.

Instead, the ability to increase the number of clicks by increasing spending comes to a stop as the market for relevant search queries saturates completely (assumption A3). Complete saturation means that there are no more opportunities to have the ad displayed regardless of how much the advertiser is willing to pay. In other words, the variable space for the number of clicks,  $Y_1$ , as well as for  $Q$  is limited.

We will here assume that the advertiser does not reach this limit. Without such assumption, we would need to build an estimate of this limit into our model.

Fully expanded model given by Eq. (2) with candidate model 1 is given as follows:

$$Y_1 = \begin{cases} (\beta_{11} + \beta_{12}D)Q + \epsilon, & \text{if } 0 \leq Q \leq \tau R \\ (\beta_{11} + \beta_{12}D)(\tau R + (Q - \tau R + 1)^\gamma - 1) + \epsilon, & \text{if } Q > \tau R \end{cases}. \quad (4)$$

#### 4.2.2 Google response function – candidate 2

While the piecewise function outlined above is easily interpretable, it is less elegant than a smooth function.

One strong candidate to fill the role of a smooth replacement function is the Sigmoid function

$$f(t) = \frac{e^t}{1 + e^t} = \frac{1}{1 + e^{-t}}.$$

The full S-shape of the Sigmoid function would not be appropriate – we have already made clear that there is no basis for assuming an initial increase of the marginal benefit. Here we will only be concerned with the variable space where  $t$  is zero or greater.

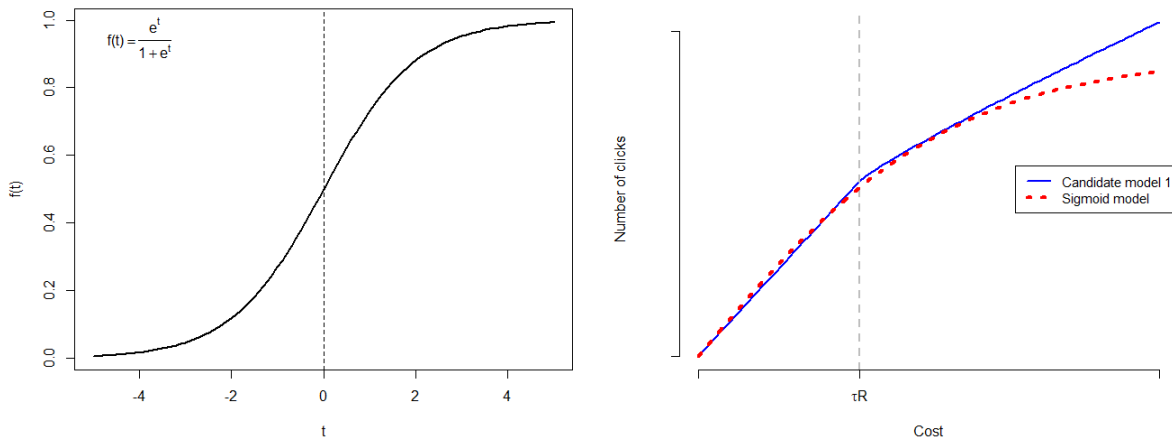


Figure 9: Left: Standard Sigmoid function, Right: The positive part of a Sigmoid function approximates the initial model.

While a Sigmoid function  $f(t)$  has a constantly decreasing slope when  $t$  is positive, it can approximate a linear function within a limited variable space. Its shape makes it suitable to model a function with a slope that is first constant and then decreasing, as is the case with candidate function 1.

#### The generalized logistic function

We generalize the Sigmoid function to allow for additional flexibility. The generalized logistic function is formulated:

$$f(t) = \rho + \frac{\delta - \rho}{(\kappa + \iota e^{-\lambda t})^{1/\theta}},$$

where

$t$  is an argument of the function  $f(\cdot)$ ,

$\rho$  is the lower asymptote, the low value which the function will approach but never reach,

$\delta$  is the upper asymptote,

$\lambda$  is the growth rate,

$\theta$  affects where maximum growth occurs,

$\iota$  is related to the value  $f(0)$

and  $\kappa$  will in most cases take the value 1.

One can note that the Sigmoid function is the special case of the generalized logistic function where  $\rho = 0$ ,  $\delta = \lambda = \theta = \iota = \kappa = 1$  namely,

$$0 + \frac{1 - 0}{(1 + 1e^{-1t})^{1/1}} = \frac{1}{1 + e^{-t}}.$$

Let us call these parameter values the *default* parameter values. For each parameter, we can now choose whether to leave it at its default value or whether to allow it to take other values:

- $\rho$  need to be adjusted to ensure that  $X_{11}$  obtains value zero when the cost is zero.
- $\delta$  need to be adjustable. With a greater market size, estimated as a function the number of healthcare employees at the clinic, we want the function to reach higher values before the slope diminishes. (This will be explained further).
- $\lambda$  needs to be adjustable so that we can fit the slope of the function to the data.

The remaining parameters in the generalized logistic function, i.e.  $\theta$ ,  $\iota$  and  $\kappa$  will be kept at their default levels.

We have now obtained a rather flexible function that can fulfill assumptions A1-A3,

$$f(t) = \rho + \frac{\delta - \rho}{(1 + e^{-\lambda t})}.$$

The parameter  $\lambda$  controls the shape of the curve. A high value of  $\lambda$  gives an initial high slope, which will then more abruptly flatten as  $f(t)$  approaches its upper asymptote. A lower value of  $\lambda$  gives a slope which is initially lower but also more enduring. See Figure 10.



## Variations in $\lambda$

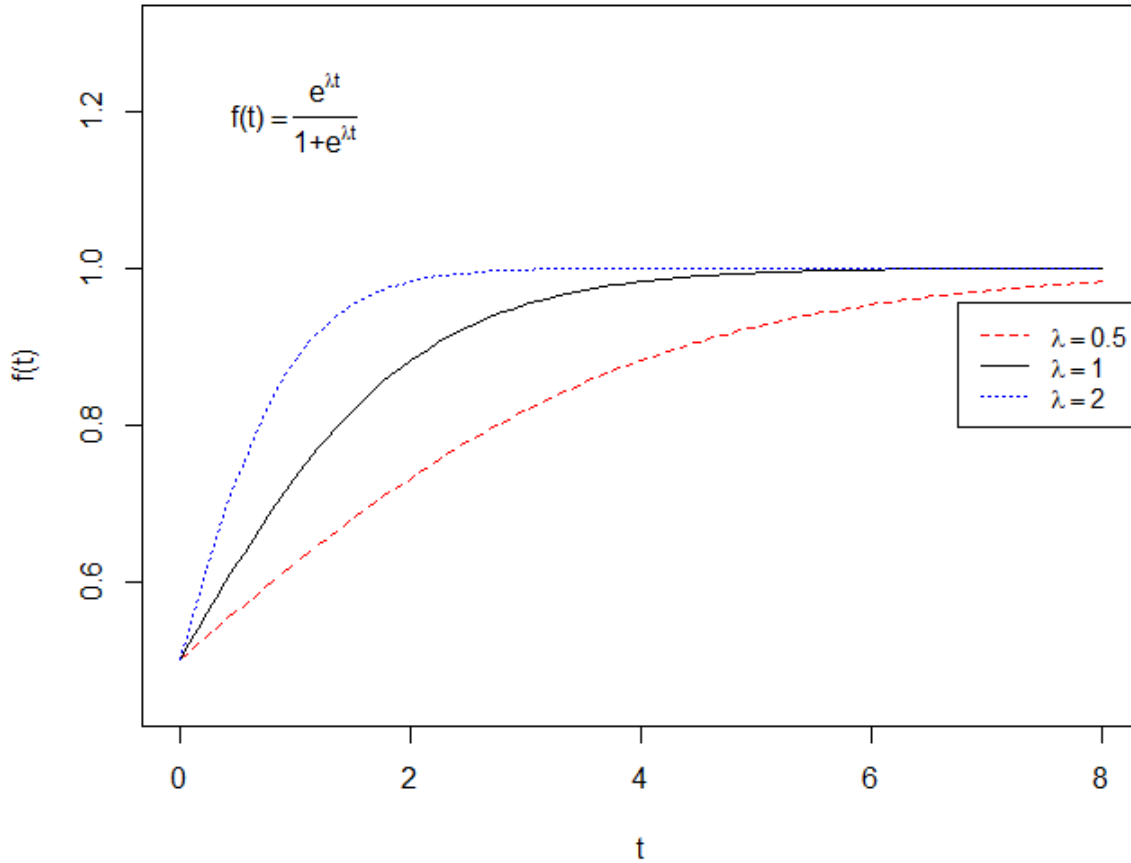


Figure 10: When we change  $\lambda$ , the upper asymptote remains constant. A more enduring slope is associated with a lower initial slope.

To enforce Assumption A1 that  $f(0) = 0$ , we set

$$f(0) = \rho + \frac{\delta - \rho}{1 + e^{-\lambda \cdot 0}} = \rho + \frac{\delta - \rho}{1 + 1} = \frac{\delta}{2} + \frac{\rho}{2} = 0 \Rightarrow \rho = -\delta.$$

Hence, setting  $\rho$  to  $-\delta$  gives us

$$f(t) = \frac{2\delta}{1 + e^{-\lambda t}} - \delta.$$

Recall candidate function 1, which is linear when  $0 \leq Q \leq \tau R$  while the slope is diminishing when  $Q > \tau R$ , where the variable  $R$  is the number of healthcare professionals, which serves as a measure of the market size. With a greater market size,  $\tau R$  will be larger and the spending can be larger before the marginal return start to diminish.

We want to achieve a similar effect with our general logistic function, see assumption A3. The slope should diminish more slowly and  $f(t)$  should be allowed to reach higher values when the

number of healthcare professionals is larger. We want the initial slope  $f'(0)$  to stay constant since the market size is assumed not to affect the initial cost-per-click.

With a larger market size, we want the cost per click should increase more slowly, and it should be possible to obtain a higher number of clicks before we reach complete saturation.

Simply decreasing the slope variable  $\lambda$  would make the slope diminish more slowly, but the asymptote will remain the same and the initial slope will decrease. Increasing  $\delta$  increases the asymptote as well as the increases the initial slope. See Figure 11.

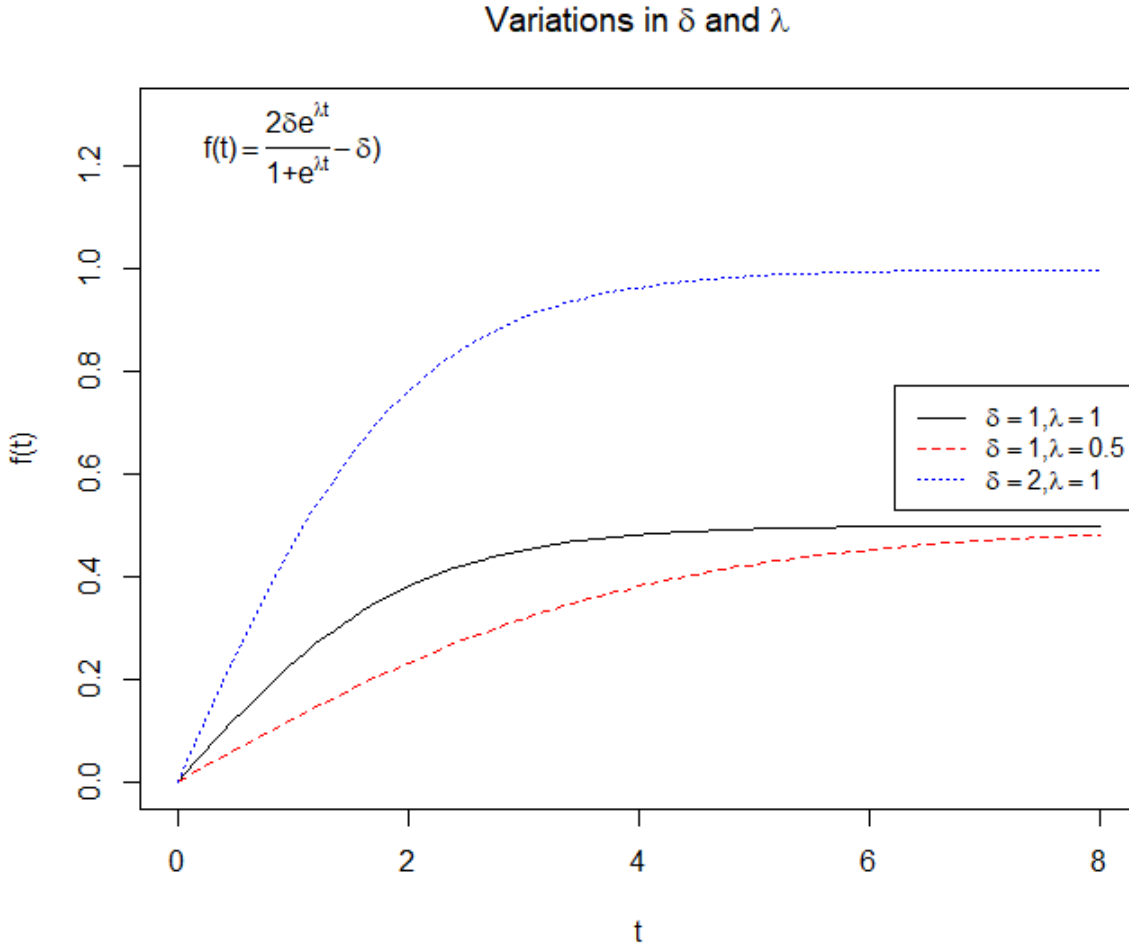


Figure 11: Changing either parameter  $\delta$  or  $\lambda$  in the unaltered generalized logistic function changes the initial slope.

In fact, there are no parameter in the generalized logistic function that by itself will make the slope diminish slower while leaving the initial slope  $f'(0)$  unchanged. However, this can be achieved with a minor extension of the function  $f(t)$  discussed above.

Let's start by finding the function that defines the initial slope,  $f'(0)$ .

$$f(t) = \frac{2\delta e^{\lambda t}}{1 + e^{\lambda t}} - \delta \implies f'(t) = \frac{2\delta \lambda e^{\lambda t}}{(1 + e^{\lambda t})^2} \implies f'(0) = \frac{\delta \lambda}{2}.$$

Thus, we confirm that the initial slope is dependent on both  $\lambda$  and  $\delta$ , and we can also see a simple

remedy. If we introduce a new parameter  $\psi = \delta\lambda$ , it gives us  $\lambda = \frac{\psi}{\delta}$ :

$$f'(0) = \frac{\delta \frac{\psi}{\delta}}{2} = \frac{\psi}{2}.$$

As we substitute  $\frac{\psi}{\delta}$  for  $\lambda$ , the initial slope is no longer directly dependent on  $\delta$ . Now we can increase or decrease  $\delta$  to change the endurance of the slope while  $\psi$  alone controls the initial slope that represents the cost per click when the ad buyer of interest has not entered the market.

Putting it all together gives us the slightly modified generalized logistic function:

$$f(t) = \frac{2\delta}{1 + e^{-\psi t/\delta}} - \delta.$$

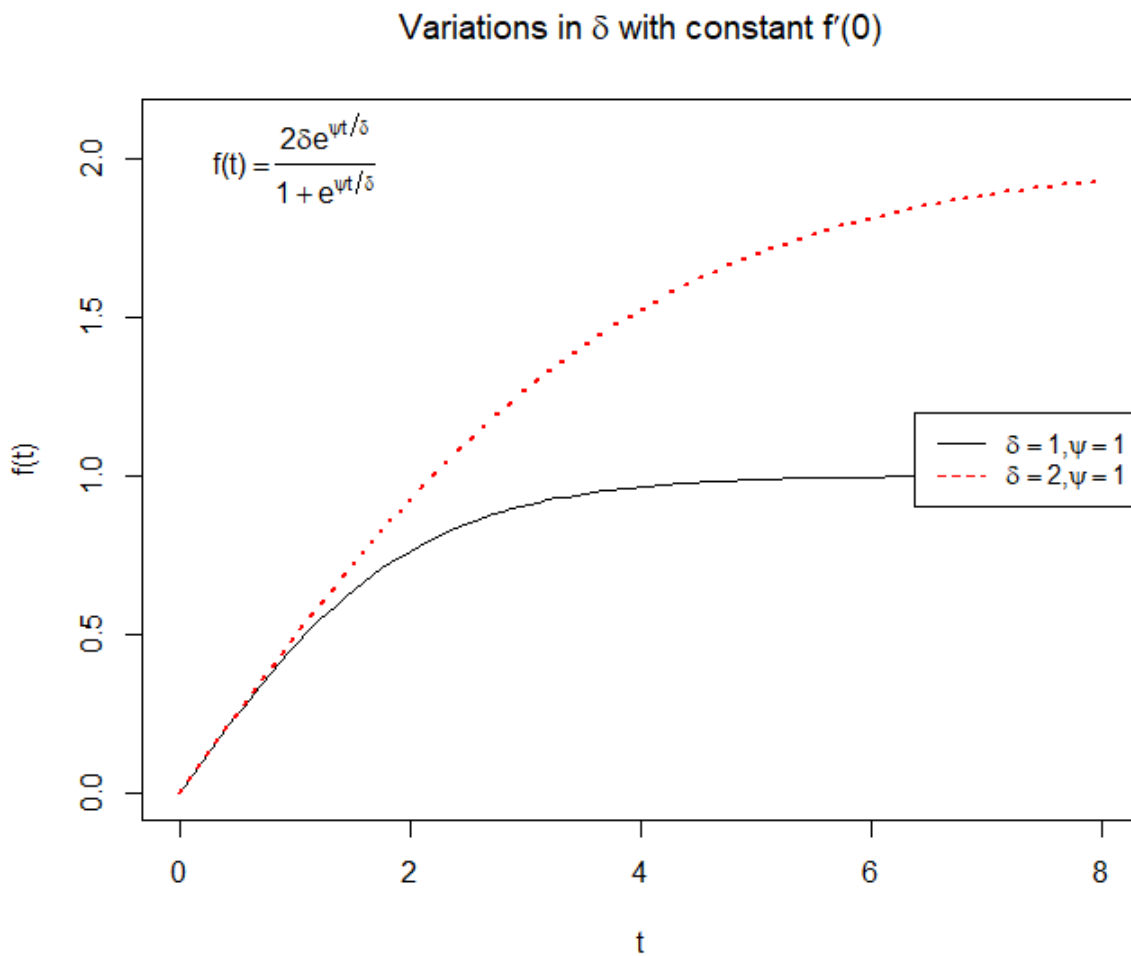


Figure 12: The extended generalized logistic function allows increased endurance of the slope through the parameter  $\delta$  without affecting the initial slope.

As mentioned above,  $\delta$  should be the function of the number of healthcare employees at the clinic. By setting

$$\delta = R^\omega,$$

we do not have to assume a linear relationship between the number of healthcare professionals and the factor by which we want to increase the endurance of the function. As in the previous model, the variable  $R$  stands for the number of healthcare professionals.

Our updated transformation of the cost variable is hence:

$$f(t) = \frac{2R^\omega}{1 + e^{-\psi t/R^\omega}} - R^\omega.$$

Hence, our fully model given by Eq. (2) expanded with candidate model 2 obtain following form:

$$Y_1 = (\beta_{11} + \beta_{12}D) \left( \frac{2R^\omega}{1 + e^{-\psi Q/R^\omega}} - R^\omega \right) + \epsilon. \quad (5)$$

### 4.3 Comparison of models 1 and 2

In this section, we compare candidate model 1 (given by Eq. (4)) and candidate model 2 (given by Eq. (5)).

In the comparison, we also include a simpler model that assumes linearity between each explanatory variables and the response variable. It includes the same original variables as the two candidate models:  $Q$ =cost,  $R$ = the number of healthcare professionals and  $D$ =population density.

$$Y_1 = \beta_1 Q + \beta_2 R + \beta_3 D + \epsilon. \quad (6)$$

This third model serves as a benchmark to confirm that our variable transformations indeed does improve the accuracy, measured as MSE.

	Parameter	Estimate
1	$\beta_{11}$	2.417
2	$\beta_{12}$	-0.000211
3	$\tau$	1.043
4	$\gamma$	0.563

Table 3: Parameters for candidate model 1

	Parameter	Estimate
1	$\beta_{11}$	108.670
2	$\beta_{12}$	-0.0105
3	$\omega$	0.363
4	$\psi$	0.00290

Table 4: Parameters for candidate model 2

	Model 1	Model 2	Linear benchmark
MSE	1477	1482	1996
Shapiro-Wilk W-stat	0.985	0.989	0.986
Shapiro-Wilk p-value	0.000057	0.0012	0.00009
Kolmogorov Smirnov D-stat	0.0395	0.0358	0.0555
Kolmogorov Smirnov p-value	0.428	0.555	0.0974

Table 5: A comparison between the two candidate models and a benchmark model. MSE is measured using 10-fold cross validation.

Table 5 shows that candidate model 1 has a slightly lower MSE than candidate model 2. Neither candidate model has residuals that are normally distributed according to the Shapiro-Wilk test at a 0.05 significance level. According to the Kolmogorov Smirnov test, both models are normally distributed at the same significance level.

As far as accuracy is concerned the models, which differ only in their variable transformations, are almost equal. The slight advantage of model 1 has little impact on the decision. Neither is the difference in proximity to a normal distribution a determining factor in itself. Whichever model we choose to go ahead with, further variable transformations will be necessary to ensure that the necessary assumptions for linear regression are met. In addition to linearity, the residuals should be independent, normally distributed and homoscedastic (Nau, 2018). [13]

In the MSE comparison in Table 5, we can see that both model 1 and model 2 outperform the linear benchmark model.

**We will move forward with candidate model 2.** While the fit to the data is approximately equal for the two candidate models and both models are founded on the logic of Google ads auctions, candidate model 2 has the advantage of not being a piecewise function. It is therefore easier to work with. It is a useful and intuitive principle to select the simpler function when two or more optional functions capture the data equally well.

#### 4.4 Transforming the Google ads variables for homoscedasticity and normality

After fitting parameters  $\psi$  and  $\omega$  we plot the residuals of the linear model in which we transformed the cost variable with an extended generalized logistic function.

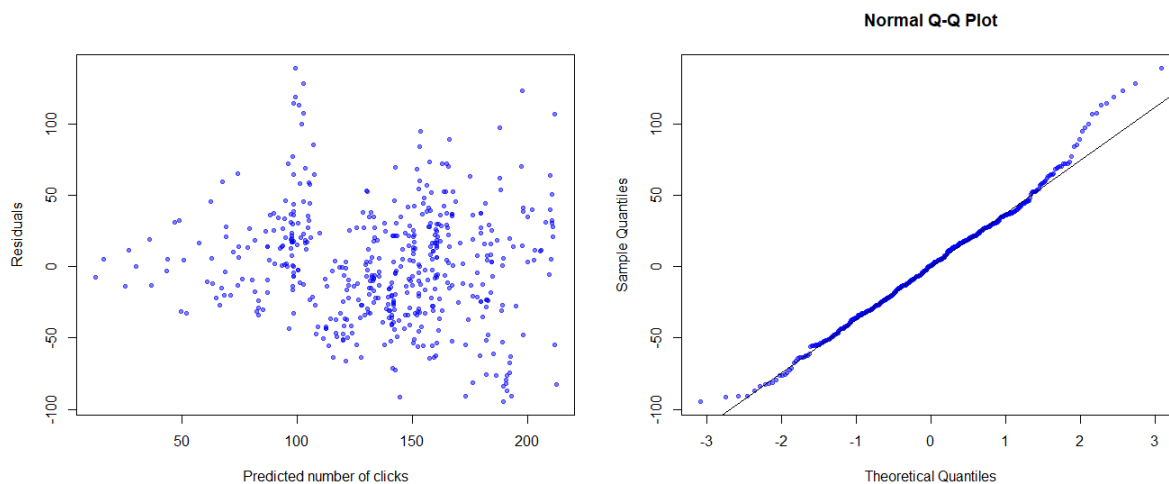


Figure 13: Left: Residuals vs predicted number of clicks. Right: Q-Q-plot for the residuals.

We see that the residuals are clearly heteroscedastic and that they increase with the predicted number of clicks. A similar pattern is true when we plot the residuals against the transformed cost variable.

The Q-Q-plot shows that we do not have a normal distribution, which is confirmed by a Shapiro-Wilk p-value of approximately 0.001.

We will start by addressing the heteroscedasticity as described in Section 3.2.3 due to the proportionality between standard deviation and cost. We utilize fact that the standard deviation of the

residuals  $\epsilon_i$  are proportional to  $\hat{x}_{11i}^a$ , where  $a$  is an unknown parameter. If this is indeed the case, we can get a constant variance by dividing our regression model by  $x_{11i}^a$ :

$$\frac{y_{1i}}{x_{11i}^a} = \beta_1 \frac{x_{11i}}{x_{11i}^a} + \beta_2 \frac{x_{12i}}{x_{11i}^a} + \frac{\epsilon_i}{x_{11i}^a} \quad (7)$$

We approach this by testing how the homoscedasticity, measured as the p-value of the F-test for homoscedasticity, changes with the value of  $a$ . We also test how the normality, measured as the p-value of the Shapiro-Wilk test, changes with the value  $a$ .

Not to reject that the homoscedasticity and normality assumptions are met, we will require a p-value of at least 0.05 for each of the tests.

In each of the following residual plots,  $a$  takes a different value.

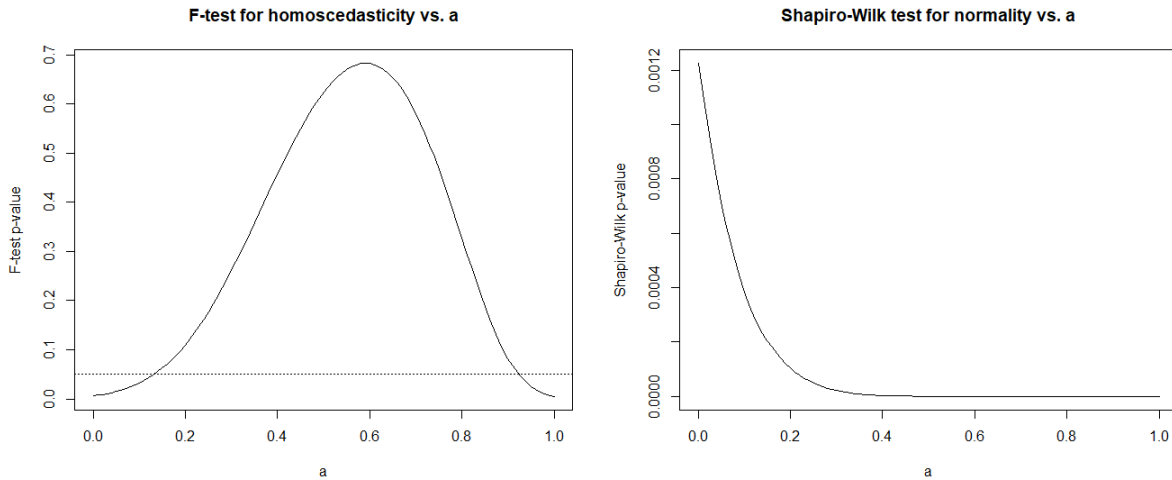


Figure 14: Left: The p-value of the F-test of homoscedasticity changes with  $a$ , when the model is divided by  $x_{11i}^a$ . Right: The p-value of the Shapiro-Wilk normality test changes with  $a$ . We see that we do not reach a p-value of 0.05 for any value of  $a$ .

Figure 14 shows that while the homoskedasticity requirement is fulfilled for a broad range of  $a$  values, there is no value of  $a$  that does not lead to rejection of the normality assumption at a 0.05 significance level. We can also note that the distribution of the residuals becomes less normal as the value of  $a$  increases.

We proceed by identifying data points that are detached from the main cluster.

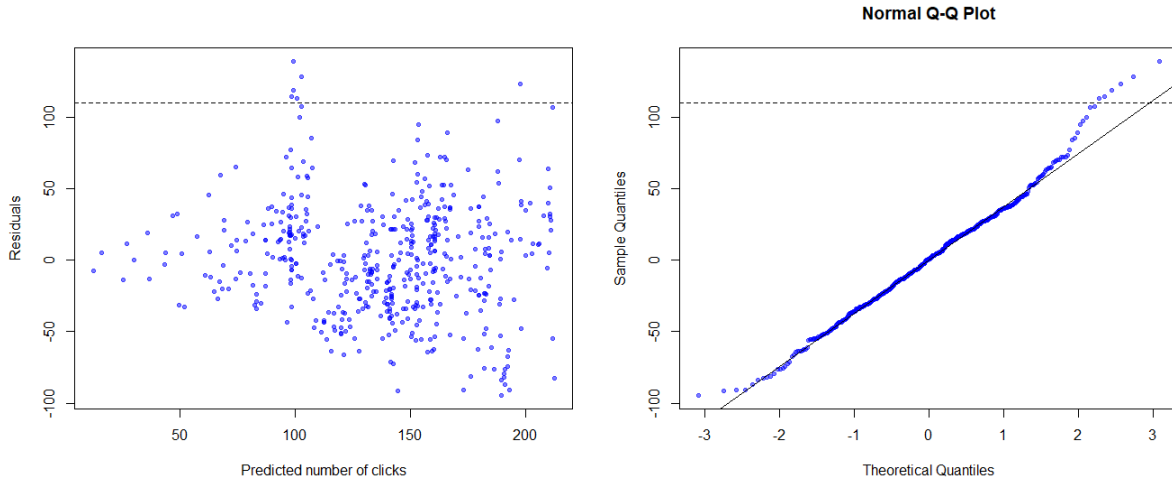


Figure 15: The dotted line separates outliers from the main cluster.

In Figure 15 we treat values with residuals that deviate more than 2.8 standard deviations, as outliers. When we study these points in detail, we can see that all but one of them is associated with the same dentist (Recall that each observation corresponds to the marketing of one clinic during one month). Apparently, this clinic has some special property that is not captured by our explanatory variables.

We remove the outliers, and again we test our variable transformation for homoskedasticity.

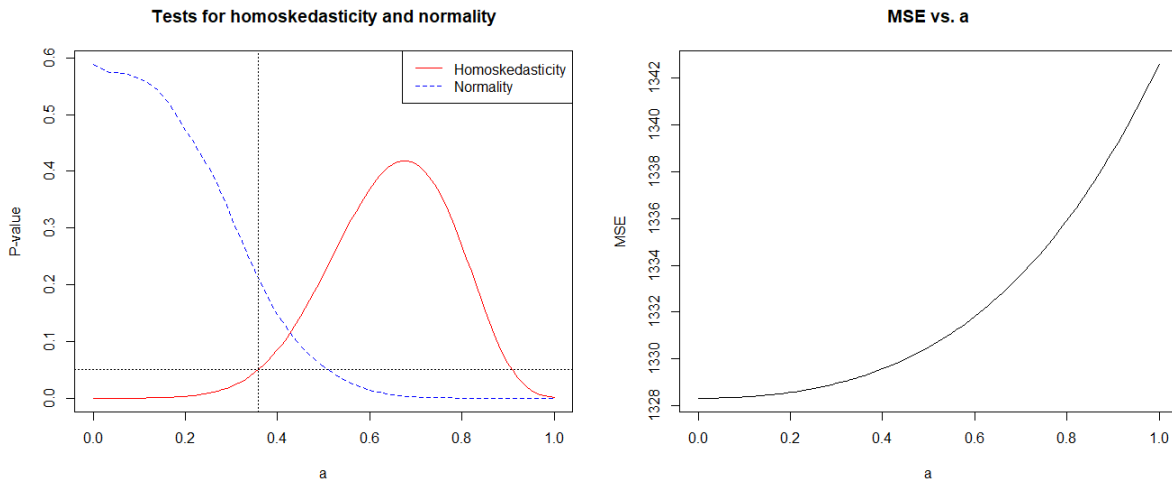


Figure 16: Left: The p-value of the F-test of homoskedasticity and the Shapiro-Wilk test of normality change with  $a$  when the model is divided by  $x_{11i}^a$  Right: The model MSE increases with  $a$ .

To the left in Figure 16, we see how the F-test for homoskedasticity p-value changes as  $a$  in Equation (7) increases. We also see how the Shapiro-Wilk p-value for normality decreases with  $a$ .

The plot shows that there is an interval around  $a = 0.4$  where the p-values for both normality and homoskedasticity are above 0.05, which is the threshold that we have decided to use. Any value

of  $a$  within this interval is acceptable. However, as we see in the right-hand plot of Figure 16, the model's MSE increases with  $a$ .

We will choose  $a$  so that we minimize model MSE given that the assumptions of normality and homoskedasticity are met. This means that we select the smallest value of  $a$  where both p-values are at least 0.05. This is the value of  $a$  where the homoscedasticity p-value is 0.05 and rising. This value of  $a$  is marked with a dotted line in the left-hand plot. To find it, we formulate a loss function  $g(a) = (f(a) - 0.05)^2$ , where  $f(a)$  is the homoscedasticity p-value for  $a$ . We minimize this loss function and find that the minimum occurs when  $a = 0.36$ .

This gives us the regression model:

$$Y_1^* = (\beta_1 + \beta_2 D) X_{11}^{1-0.36} + \tilde{\epsilon}, \quad (8)$$

where  $X_{11}$  is transformed according to the candidate 2 model.

Note that the response variable has been transformed, so back transformation will be required to interpret the the result in a meaningful way.

#### 4.5 Results and comments

The final marginal model for Google ads marketing is built on theoretical assumptions and also satisfies the normality and homoskedasticity assumptions about the residuals.

Fitting the model gives us point estimates of the model parameters.

Parameter	Estimate
$\beta_{11}$	110.1
$\beta_{12}$	-0.0106
$\omega$	0.362
$\psi$	0.003

Table 6: Parameters for the Google ads Marginal model

Estimated residual standard error (483 d.f.)	33.06
Shapiro-Wilk W-stat	0.996
Shapiro-Wilk p-value	0.21
Kolmogorov Smirnov D-stat	0.021
Kolmogorov Smirnov p-value	0.981
Model significance, F statistic (2, 483 d.f.)	3437
Model significance, p-value	0.000
Variance inflation factor(VIF), $X_{11}$ , $X_{12}$	1.33

Table 7: Selected results of fitting the Google ads model

Since the  $R^2$  value has an uncertain interpretation for no-intercept models (it is calculated differently by different software packages) we do not include that statistic.

The threshold for a problematic VIF value range between 2.5 and 10 depending on the author. (Glen, 2015) [8] Our VIF value of 1.33 should not be a reason for concern.

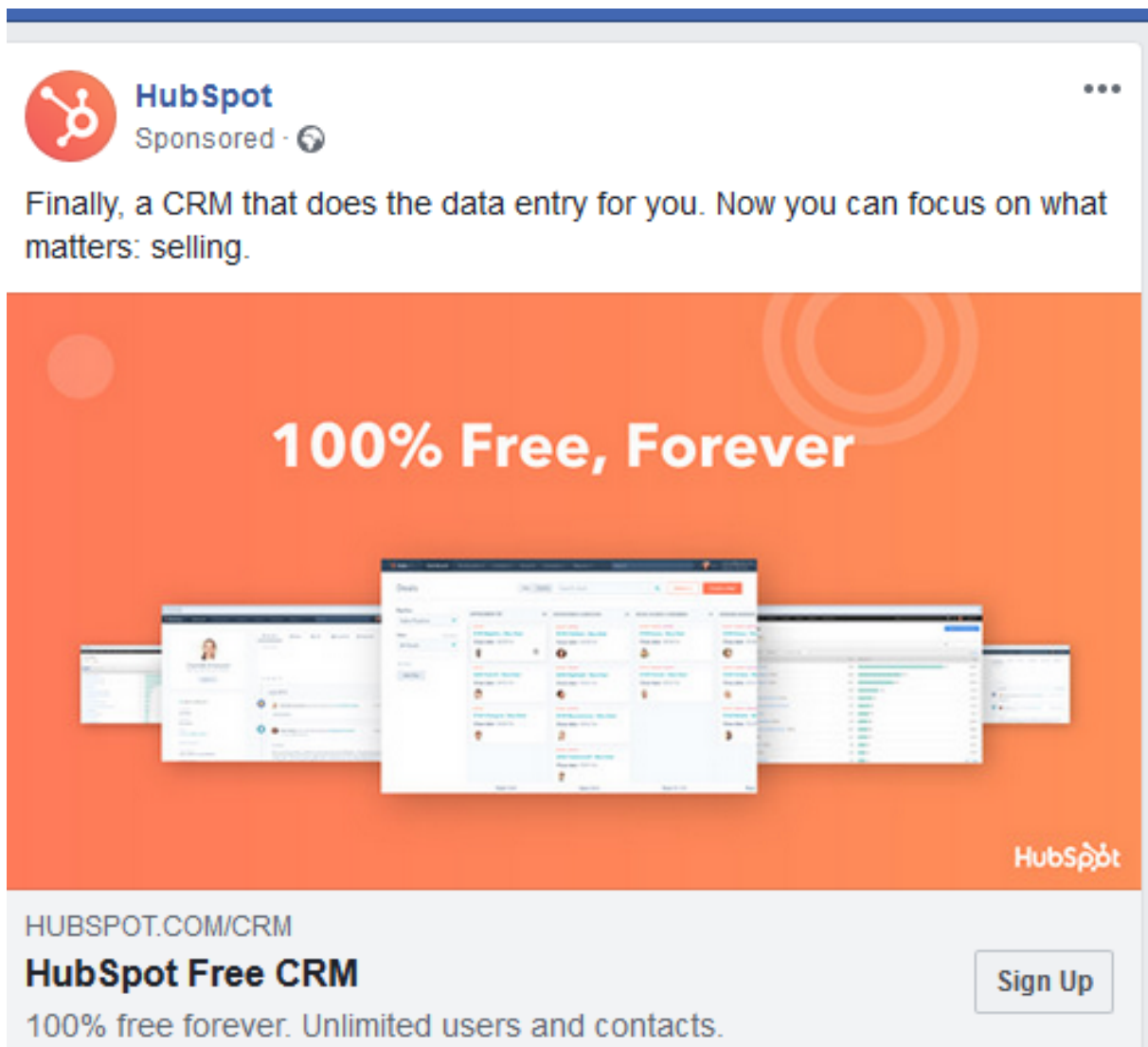


The R code for each step in developing the model (excluding the code used to generate plots) can be found in Appendix 1.

## 5 The Facebook response function

On Facebook, the advertiser decides whether to set a budget per day or a lifetime budget. A lifetime budget makes sense if the ad campaign concerns a project limited in time.

You can also select a bid type according to what you want to accomplish with the ad. The options are “Bid for clicks” or “Bid for impressions”. “Bid for clicks” means that you pay per click while “Bid for impressions” means that you pay for the number of times that the ad shows regardless of the number of clicks.



HubSpot  
Sponsored · 🌐

Finally, a CRM that does the data entry for you. Now you can focus on what matters: selling.

**100% Free, Forever**

HUBSPOT.COM/CRM  
**HubSpot Free CRM**  
100% free forever. Unlimited users and contacts.

Sign Up

Figure 17: Facebook ads show up among the updates in the feed. The Ads allow for graphical elements.

Facebook recommends the “Bid for clicks” option if the goal is to direct as many users as

possible to a landing page (Facebook, 2019). [7]

If the advertiser pays per click, the cost for each click is limited by the bid value. As with Google ads, there is a quality score assigned to each ad which will manifest itself as part of the model's error term if the ads in the sample have varying quality scores.

Our Facebook data shows an almost perfectly linear relationship between the cost and the number of clicks. For this reason, we do not need any variable transformation to make the relationship linear as in the Google case.

Furthermore, the cost variable is the only explanatory variable that we have access to. (Impressions, conversions and number of clicks are unknown variables until after the advertisement campaign.)

Thus, our Facebook model is

$$Y_2 = \beta_{21}X_{21} + \epsilon. \tag{9}$$

### 5.1 Transforming the Facebook ads variables for homoscedasticity and normality

We first fit a linear no-intercept regression model with the unmodified variables  $X_{21}$ =Cost and  $Y_2$ =Number of clicks. We plot the residuals against  $\hat{Y}_2$  and generate a Q-Q-plot.

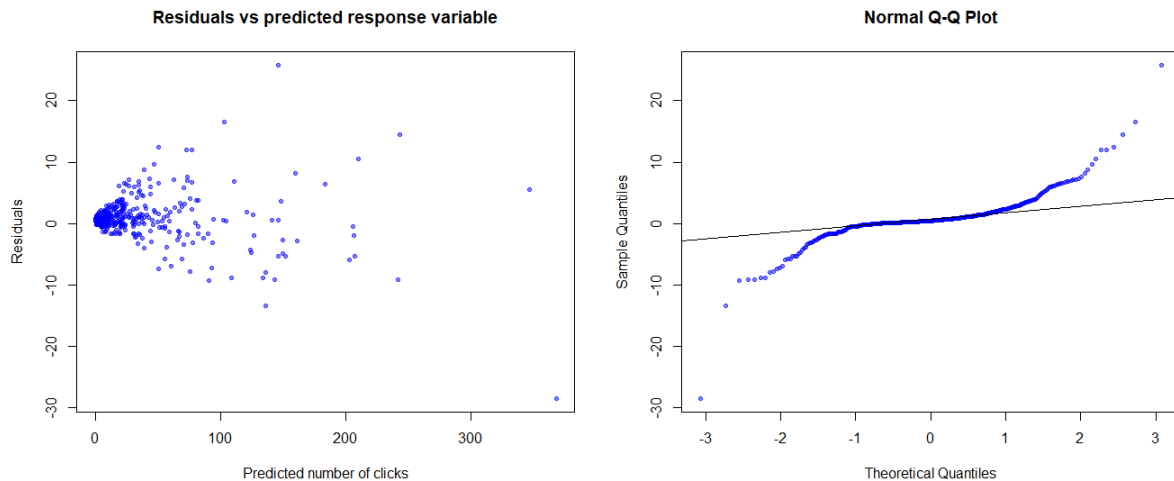


Figure 18: Linear regression using the unmodified data. Left: Residuals vs predicted number of clicks. Right: Q-Q-plot of the residuals.

In Figure 18, we see a clear case of heteroscedasticity. The spread of the residuals increases with the estimated response variable. The Q-Q-plot shows that the data does not follow a normal distribution, which is confirmed by a Shapiro-Wilk test that gives a p value smaller than  $2.2 \cdot 10^{-16}$ .

The heteroscedasticity could be seen in the preliminary investigation of the data. It is apparent from the Clicks vs. Cost plot in Figure 4 that the variance increases with increasing cost.

To achieve homoskedasticity, we follow the same transformation procedure as with the Google ads model i.e. we divide the model by  $x_{21i}^a$  after identifying an appropriate value for  $a$ .

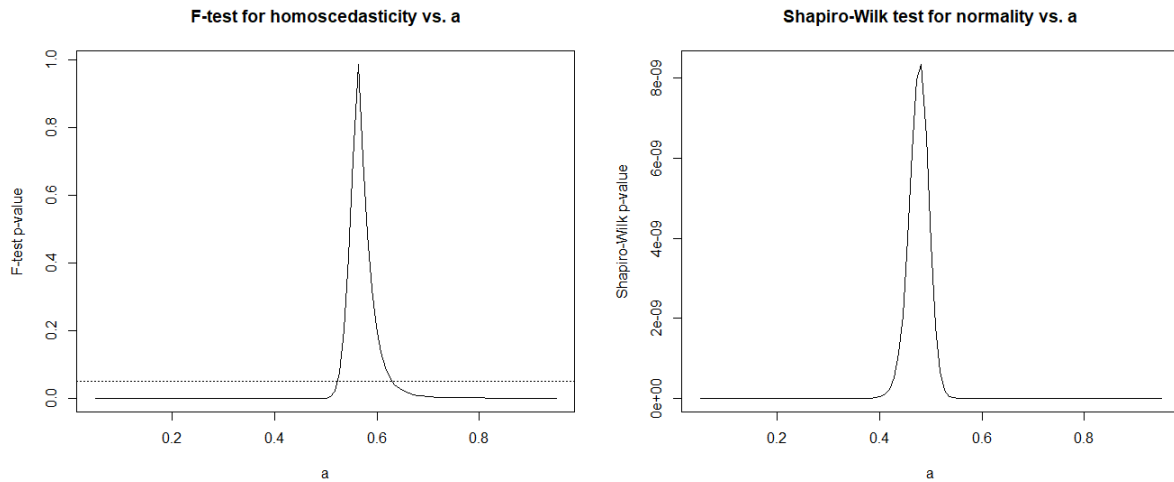


Figure 19: Left: The p-value of the F-test of homoscedasticity changes with  $a$ , when the model is divided by  $x_{11i}^a$ . Right: The p-value of the Shapiro-Wilk normality test changes with  $a$ . We see that we do not reach a p-value of 0.05 for any value of  $a$ . Note the different scales.

Figure 19 shows that while the homoskedasticity requirement is fulfilled for  $a$  values around 0.6, there is no value of  $a$  that does not lead to rejection of normality at a 0.05 significance level.

We proceed by identifying data points that are detached from the main cluster.

We want to identify observations which produce abnormal residuals after the variable transformation. Looking only at the residual values plotted in Figure 18 will not help much. We need to know which residuals stand out in relation to the predicted number of clicks. We illustrate this in Figure 20.

## Residuals in relation to predicted number of clicks

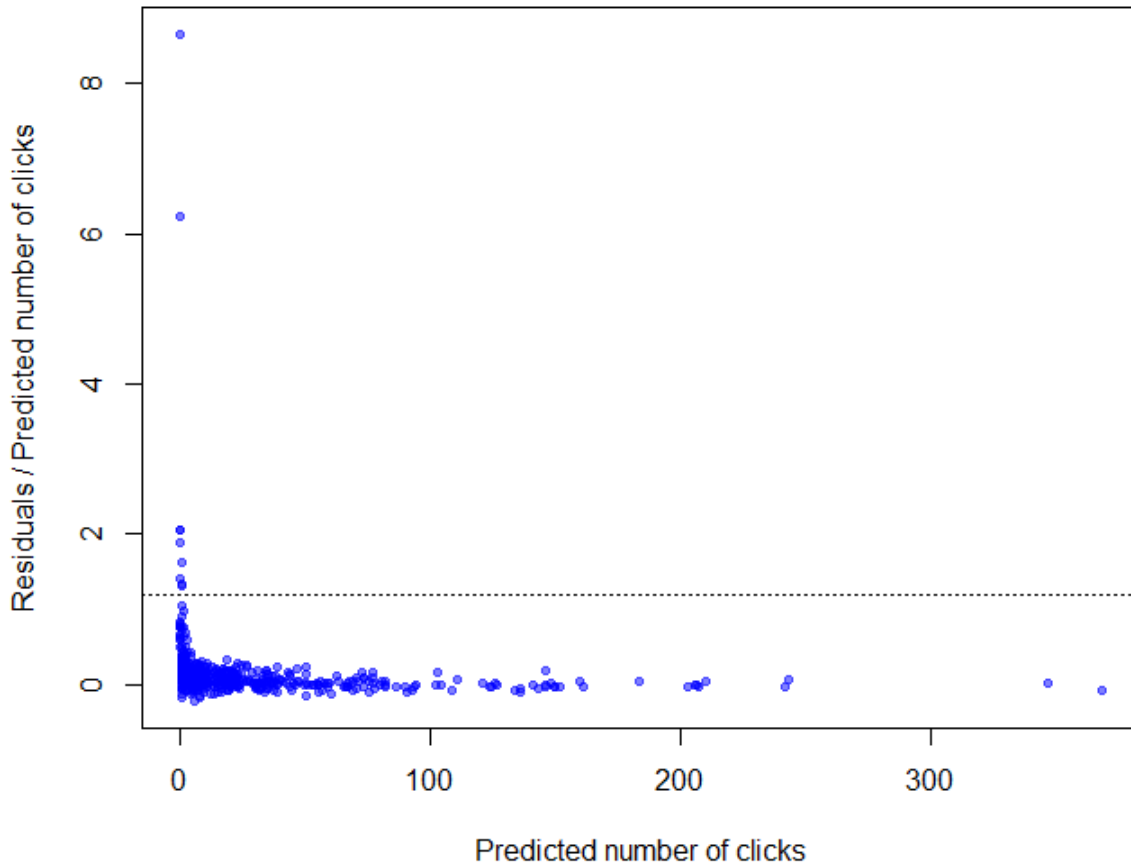


Figure 20: The dotted line separates outliers from the main cluster.

We remove outliers. The points affected are data points representing one or two clicks. These data points can be the result of an advertisement that ran only for a day or a few hours. As such they are not representative of the scenarios that we want to predict.

After removing the outliers, we repeat the procedure of transforming the variables for homoskedasticity.

Figure 21 shows an interval where  $a$  is slightly larger than 0.6 where the p-values for both normality and homoskedasticity are above our significance level 0.05.

As with the Google ads model, We will choose  $a$  so that we minimize model MSE given that the assumptions of normality and homoskedasticity are met. This is the case where  $a = 0.65$ . This gives us the regression model:

$$Y_2^* = \beta_{21} X_{21}^{1-0.65} + \tilde{\epsilon}. \quad (10)$$

## 5.2 Results

The marginal model for the number of clicks due to Facebook ads marketing is given by Equation 10, where  $\tilde{\epsilon}$  can be assumed to be normally distributed and homoskedastic. See Table 9. Estimate

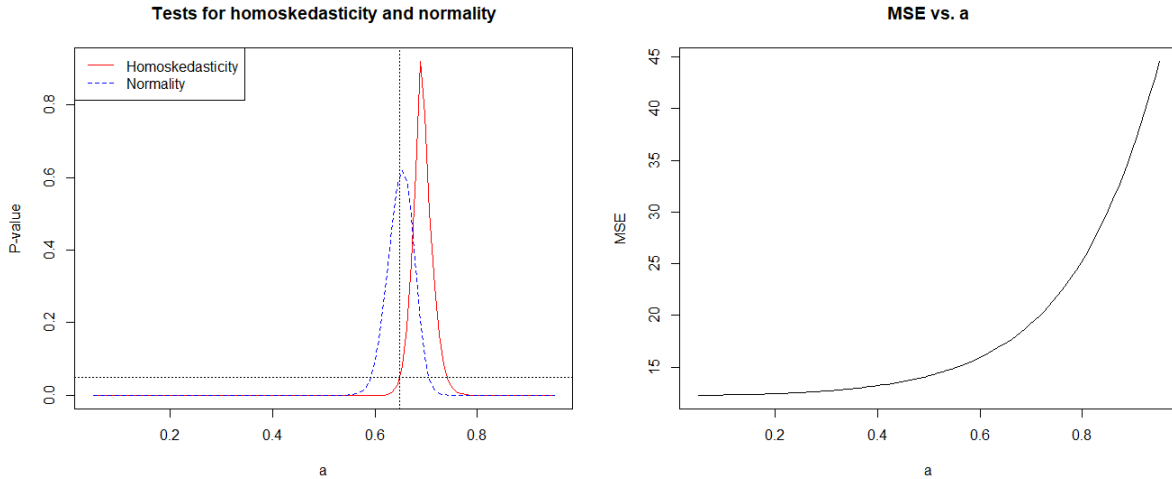


Figure 21: Left: The p-value of the F-test of homoscedasticity and the Shapiro-Wilk test of normality change with  $a$  when the model is divided by  $x_{21i}^a$ . Right: The model MSE increases with  $a$ .

of the model parameter is given in Table 8.

Parameter	Estimate
$\beta_1$	0.0599

Table 8: Parameters for the Facebook ads model

Estimated residual standard error (463 d.f.)	0.0499
Shapiro-Wilk W-stat	0.997
Shapiro-Wilk p-value	0.615
Kolmogorov Smirnov D-stat	0.0341
Kolmogorov Smirnov p-value	0.653
Model significance, F statistic (2, 481 d.f.)	44370
Model significance, p-value	0.000

Table 9: Selected results of fitting the Facebook ads model

The R code for each step in developing the model (excluding the code used to plot) can be found in Appendix 1.

## 6 An aggregated model, optimal mix and predictions

We have developed two marketing response models, one for Google ads and one for Facebook ads. In this section we will combine them into a joint function.

The joint function will have a two dimensional variable space or, as we will see, a one-dimensional variable space under the condition that we will spend a fixed marketing budget.

However, it is an advantage if the model can easily handle response functions for an arbitrary number of marketing channels. For that reason we will employ numerical methods to optimize the variables even when there is an analytical solution, and we will do it without reducing the number of dimensions.

In section 6.2.1 we will briefly show that an analytical solution is possible in the one dimensional case.

Building an aggregated model will involve the following steps:

1. State assumptions for the joint model.
2. Formulate the joint function for the point estimate of the number of clicks, and choose a method to maximize the number of clicks.
3. Compare the numerical solution to an analytical solution.
4. Formulate the function for the confidence interval and prediction interval.
5. Apply the full model to a test case.

## 6.1 Additional assumptions for the joint model

In addition to assumptions A1-A3, which we have already established for the marginal functions:

- We assume independence between the error terms of the response functions for the different marketing channels. Since our Google ads data and our Facebook ads data do not pertain to the same companies, we cannot estimate a covariance. Furthermore, the variance in the Facebook model is quite small, which means that the covariance has less impact.
- We assume that the number of Google ads clicks does not depend on any of the variables in the Facebook ads response function, and vice versa. The current data does not indicate whether, for instance, a Google ads click affects user's likeliness of clicking a subsequent Facebook Ad from the same business. On the one hand, the user has already shown interest in the business. On the other hand, the user may have already explored the website and may have less reason to visit the website a second time. In a future where data make it possible to estimate these conditional probabilities, the assumption of independence can be reconsidered.

## 6.2 Point estimate of the total number of paid clicks

We begin by formulating the point estimate of the joint function and leave the error term for later.

Our linear regression for the Google ads gives the point estimate:

$$\hat{Y}_1^* = \hat{\beta}_{11}X_{11}^* + \hat{\beta}_{12}X_{12}^*,$$

where  $\hat{Y}_1^*$  is the estimated response variable for the Google ads regression.

If we formulate the regression with our original variables, we get:

$$\frac{\hat{Y}_1}{\left(\frac{2R^\omega e^{\psi Q/R^\omega}}{1+e^{\psi Q/R^\omega}} - R^\omega\right)^{0.36}} = \hat{\beta}_{11} \frac{\left(\frac{2R^\omega e^{\psi Q/R^\omega}}{1+e^{\psi Q/R^\omega}} - R^\omega\right)}{\left(\frac{2R^\omega e^{\psi Q/R^\omega}}{1+e^{\psi Q/R^\omega}} - R^\omega\right)^{0.36}} + \hat{\beta}_{12} \frac{\left(D \left(\frac{2R^\omega e^{\psi Q/R^\omega}}{1+e^{\psi Q/R^\omega}} - R^\omega\right)\right)}{\left(\frac{2R^\omega e^{\psi Q/R^\omega}}{1+e^{\psi Q/R^\omega}} - R^\omega\right)^{0.36}},$$

where  $\hat{Y}_1$  is the expected number of clicks,  $Q$  is the Google ads spending,  $R$  is the number of healthcare professionals at the clinic,  $D$  is the population density in the municipality.  $\omega$  and  $\psi$  are shape parameters fit to the data.

We perform a back transformation so that we estimate  $\hat{Y}_1$  by multiplying both sides of the equation with the common denominator:

$$\begin{aligned}\hat{Y}_1 &= \hat{\beta}_{11} \left( \frac{2R^\omega e^{\psi Q/R^\omega}}{1 + e^{\psi Q/R^\omega}} - R^\omega \right) + \hat{\beta}_{12} D \left( \frac{2R^\omega e^{\psi Q/R^\omega}}{1 + e^{\psi Q/R^\omega}} - R^\omega \right) \\ &= (\hat{\beta}_{11} + \hat{\beta}_{12} D) \left( \frac{2R^\omega e^{\psi Q/R^\omega}}{1 + e^{\psi Q/R^\omega}} - R^\omega \right).\end{aligned}$$

Our linear regression for the Facebook ads gives the point estimate:

$$\hat{Y}_2^* = \hat{\beta}_{21} X_{21}^*$$

which we can formulate with the original variables as:

$$\frac{\hat{Y}_2}{X_{21}^{0.65}} = \hat{\beta}_{21} \frac{X_{21}}{X_{21}^{0.65}} \quad \Rightarrow \quad \hat{Y}_2 = \hat{\beta}_{21} X_{21},$$

where  $\hat{Y}_2^*$  is the estimated response variable for the Facebook ads regression,  $\hat{Y}_2$  is the expected number of Facebook ad clicks and  $X_{21}$  is the Facebook ads spending.

Finally we obtain the point estimate of the total number of clicks  $\hat{Y}_{tot}$  as

$$\hat{Y}_{tot} = \hat{Y}_1 + \hat{Y}_2 = (\hat{\beta}_{11} + \hat{\beta}_{12} D) \left( \frac{2R^\omega e^{\psi Q/R^\omega}}{1 + e^{\psi Q/R^\omega}} - R^\omega \right) + \hat{\beta}_{21} X_{21}.$$

Note that it is the values of  $Q$  and  $X_{21}$  that we want to set so that we maximize  $\hat{Y}_{tot}$ . All other parameters and variables will be given when we apply the model to the data of a new clinic.

### 6.2.1 An analytical solution for finding the weight mix with a fixed budget

In the case where we want to spend a fixed amount of money on marketing and we only have two marketing channels, the optimization problem is one dimensional.

$$\begin{aligned}\hat{Y}_1 &= \hat{\beta}_{11} \left( \frac{2R^\omega e^{\psi Q/R^\omega}}{1 + e^{\psi Q/R^\omega}} - R^\omega \right) + \hat{\beta}_{12} D \left( \frac{2R^\omega e^{\psi Q/R^\omega}}{1 + e^{\psi Q/R^\omega}} - R^\omega \right) \\ \hat{Y}_2 &= \hat{\beta}_{21} X_{21}\end{aligned}$$

Moreover we set  $W$  as the total amount that will be spent on online marketing:

$$W = Q + X_{21} \quad \Rightarrow \quad X_{21} = W - Q.$$

Thus, we can replace the Facebook ads spending variable  $X_{21}$  with a function of the Google ads spending variable  $Q$ , so we have a (negative) loss function with  $Q$  as the only variable. All other values in the equation are given when we predict the outcome in a specific case.

$$\hat{Y}_{tot} = \hat{Y}_1 + \hat{Y}_2 = \hat{\beta}_{11} \left( \frac{2R^\omega e^{\psi Q/R^\omega}}{1 + e^{\psi Q/R^\omega}} - R^\omega \right) + \hat{\beta}_{12} D \left( \frac{2R^\omega e^{\psi Q/R^\omega}}{1 + e^{\psi Q/R^\omega}} - R^\omega \right) + \hat{\beta}_{21} (W - Q)$$

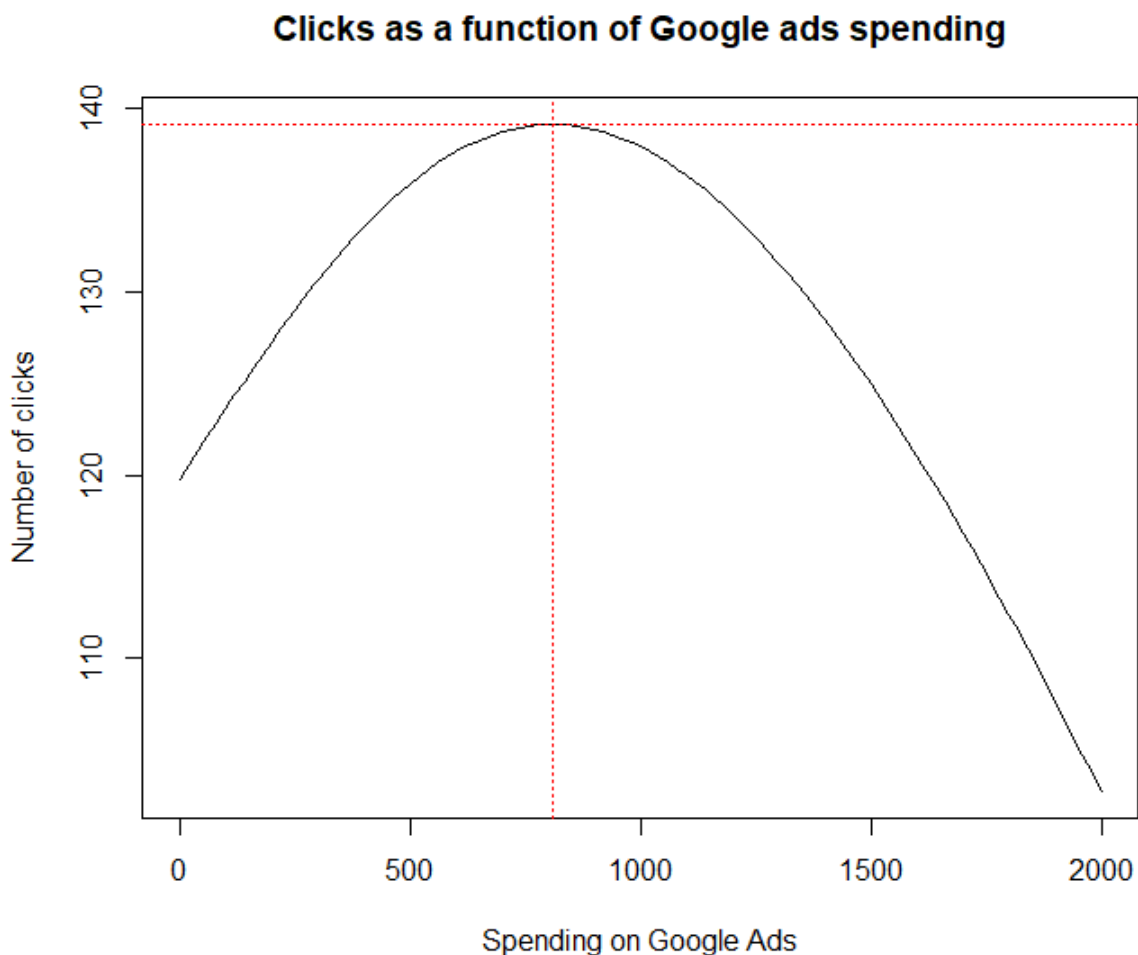


Figure 22: Predicted number of clicks as a function of Google ads spending, in a case where  $R = 4$ ,  $D = 4198.7$ . At 808 SEK spent on Google ads, we get the maximum 139 clicks, as shown in section 7.1.

By differentiating our loss function and setting the derivative to 0, we can identify the position of the maximum in Figure 22.

The derivative of  $\hat{Y}_{tot}$  with respect to  $Q$  is

$$\hat{Y}'_{tot} = \frac{2\psi e^{\psi Q/R^\omega}}{(1 + e^{\psi Q/R^\omega})^2}(\hat{\beta}_{11} + \hat{\beta}_{12}D) - \hat{\beta}_{21}.$$

Setting  $\hat{Y}'_{tot} = 0$  and solving for  $Q$  gives

$$Q = \frac{R^\omega}{\psi} \log \left[ \frac{\psi(\hat{\beta}_{11} + \hat{\beta}_{12}D) - \hat{\beta}_{21}}{\hat{\beta}_{21}} \pm \sqrt{\left( \frac{\hat{\beta}_{21} - \psi(\hat{\beta}_{11} + \hat{\beta}_{12}D)}{\hat{\beta}_{21}} \right)^2 - 1} \right]$$

We will only consider values in the range  $0 \leq Q \leq W$ .



### 6.3 Prediction intervals for the marginal functions

A linear regression prediction interval for one marginal function  $i$  is calculated as

$$y_0^* = \hat{y}_0^* \pm t_{\alpha/2, v} s.e._i^* \sqrt{1 + X_0^T (X^T X)^{-1} X_0}, \quad (11)$$

where  $t$  is a Student's t distribution,  $s.e._i^*$  is the standard error of the  $Y_i$  variable,  $X$  is the design matrix,  $X_0$  is a vector with the explanatory variables for the observation of interest. (C. Zaiontz, 2019) [16]

An alternative expression for the prediction interval, which clarifies the source of the uncertainties in the prediction, is

$$y_0^* = \hat{y}_0^* \pm t_{\alpha/2, v} \sqrt{\hat{\sigma}_\beta^2 + \hat{\sigma}^2},$$

where  $\hat{\sigma}_\beta^2$  captures the uncertainty of the model's  $\beta$  parameters and  $\hat{\sigma}^2$  captures the variability represented by the error term  $\epsilon$ . (Note that  $\hat{\sigma}_\beta^2$  is not the variance of the  $\beta$  parameters themselves. It is the contribution to the variance of the response variable that result from the uncertainty of the  $\beta$  parameters.)

The estimated standard deviation of  $\tilde{\epsilon}$  for the marginal models is here estimated by  $s.e._i^*$  in Equation (11).

#### 6.3.1 Back transformation

Due to the transformations, the response variable  $Y^*$  does not give us the value of interest. Instead, we are interested in the back transformed variable  $Y = Y^* \cdot T_i$  that represents the number of paid clicks.  $T_i$  is the transformation factor, which for our Google ads model is  $T_1 = X_{11}^{0.36}$  and for Facebook ads model  $T_2 = X_{21}^{0.65}$ .

Likewise, the standard error of our regression model is not the standard error regarding the number of clicks, but rather the standard error of the modified response variable.

Back transformation of the model:

$$Y_i^* = \sum_{k=1}^p \left( \frac{\beta X_k}{T_i} \right) + \tilde{\epsilon} \quad \Rightarrow \quad Y_i = \sum_{k=1}^p (\beta X_k) + \tilde{\epsilon} T_i.$$

Our assumption is that the error term is normally distributed:  $\tilde{\epsilon} \sim N(0, \hat{\sigma}_{Y^*})$ .

Using notation introduced earlier,  $\hat{\sigma}_{Y_i^*} = s.e._i^*$  and thus  $\hat{\sigma}_{Y_i} = s.e._i^* T_i$ .  $\sigma_\beta$  is estimated by  $s.e._i^* \cdot T_i \sqrt{X_0^T (X^T X)^{-1} X_0}$ .

The total prediction interval for each marginal distribution can be formulated:

$$\begin{aligned} Y &= \hat{Y}_0 \pm t_{\alpha/2, v} \sqrt{\hat{\sigma}^2 + \hat{\sigma}_\beta^2} \\ &= \hat{Y}_0 \pm t_{\alpha/2, v} \sqrt{(s.e._i^* T_i)^2 + (s.e._i^* \cdot T_i)^2 X_0^T (X^T X)^{-1} X_0} \\ &= \hat{Y}_0 \pm t_{\alpha/2, v} s.e._i^* \cdot T_i \sqrt{1 + X_0^T (X^T X)^{-1} X_0}. \end{aligned} \quad (12)$$

### 6.4 Prediction intervals for the aggregated number of paid clicks

Adding two t-distributed variables is often done by approximation. An approximation that work well with sufficient degrees of freedom is the normal approximation (D. S. Bhoj & D. Kushary,

2012). [1] Normal approximation is considered acceptable with more than 30 degrees of freedom. This method treats each t-distributed variable  $t_i$  as a normally distributed variable. Two independent normal distributed variables that are added results in a normal distribution in which the means and the variances are added:

$$A + B \sim N(\mu_A + \mu_B, \sigma_A^2 + \sigma_B^2)$$

Hence, to capture the sum of the number of clicks from the marketing channels, we add the normal approximation of the Google ads error term  $e_1$  to the normal approximation of the Facebook ads error term  $e_2$ . Our combined error term will be  $e_{tot} \sim N(0, \sqrt{\sigma_1^2 + \sigma_2^2})$ , where  $\sigma_1^2 = \sigma_{1\beta}^2 + \sigma_{1,0}^2$  and  $\sigma_2^2 = \sigma_{2\beta}^2 + \sigma_{2,0}^2$ .

The total number of clicks  $Y_{tot}$  can then be modeled by

$$Y_{tot} = \hat{Y}_{tot} + e_{tot}, \quad \text{where} \quad e_{tot} \sim N(0, \hat{\sigma}_{tot}),$$

$$\begin{aligned} \hat{\sigma}_{tot} &= \sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_{1,\beta}^2 + \hat{\sigma}_2^2 + \hat{\sigma}_{2,\beta}^2} \\ &= \sqrt{(s.e._1)^2 \left( T_1^2 + T_1^2 \cdot \mathbf{X}_{1,0}^T (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_{1,0} \right) + (s.e._2)^2 \left( T_2^2 + T_2^2 \cdot \mathbf{X}_{2,0}^T (\mathbf{X}_2^T \mathbf{X}_2)^{-1} \mathbf{X}_{2,0} \right)}. \end{aligned}$$

With sufficient degrees of freedom, we can use this normal approximation to construct a prediction interval.

An alternative approach to approximating the distribution of the sum of two t-distribution is to simulate the combined distributions. To construct a prediction interval, we simply extract the quantiles for the probabilities which we want to use as the limits of the interval (i.e.  $p=0.15$  and  $p=0.85$  for a 70 percent prediction interval).

*We will use the normal approximation since this allows us to describe the entire function, in mathematical notation.*

We have well above 400 degrees of freedom for each of the random variables  $Y_1$  and  $Y_2$ . Thus, we can expect the normal approximation of  $Y_{tot} = Y_1 + Y_2$  to be an accurate approximation.

## 7 Applying the model - an example

We will apply the model to a fictive potential client.

Firstly we find the point estimate for the maximum number of clicks, and then we add a 70 percent prediction interval for this variable.

We give the weights for each marketing channel that lead to the maximum estimated number of clicks. For comparison we also calculate the weights that minimize the variance of our marketing mix portfolio.

The choice to use a 70 percent prediction interval rather than the more common 95 percent interval is pragmatic. Since the variance in the Google ads model is quite large, a 95 percent confidence interval will be wide. In the choice between on the one hand a wide high-confidence interval and on the other hand a narrower lower-confidence interval which can still be described as "very likely", the latter is considered more informative.

Our fictional dentist will be located in Solna and employ 4 healthcare professional, which makes it a quite typical private clinic. The population density in Solna is, according to SCB, 4198.7 inhabitants per square kilometer.

The clinic is willing to spend at most  $W = 2000$  SEK per month on online marketing.

### 7.1 Point estimate of the maximal number of clicks

In section 6.2, we found the following estimate of the number of clicks.

$$\hat{Y}_{tot} = \hat{Y}_1 + \hat{Y}_2 = \hat{\beta}_{11} \left( \frac{2R^\omega e^{\psi Q/R^\omega}}{1 + e^{\psi Q/R^\omega}} - R^\omega \right) + \hat{\beta}_{12} D \left( \frac{2K^M e^{\psi Q/R^\omega}}{1 + e^{\psi Q/R^\omega}} - R^\omega \right) + \hat{\beta}_{21} X_{21},$$

where for our client:  $R = 4$  (number of healthcare professionals),  
 $D = 4198.7$  (population density),  
 $W = 2000$  (maximum monthly budget).

The parameter values are estimated to:

$$\begin{aligned} \hat{\beta}_{11} &= 110.1, \\ \hat{\beta}_{12} &= -0.0106, \\ \hat{\beta}_{21} &= 0.0599, \\ \hat{\omega} &= 0.362, \\ \hat{\psi} &= 0.003. \end{aligned}$$

Estimate  $\hat{Y}_{tot}$  is now a function of Google ads spending,  $Q$ , and Facebook ads spending,  $X_{21}$ . All other values are fixed.

We want to solve following optimization problem:

$$\operatorname{argmax}_{Q, X_{21}} (\hat{Y}_{tot}) \quad (13)$$

$$\text{subject to } \begin{cases} 0 \leq Q \leq 2000 \\ 0 \leq X_{21} \leq 2000 \\ Q + X_{21} \leq 2000 \end{cases}, \quad (14)$$

and to do so, we formulate a loss function in which we minimize  $-\hat{Y}_{tot}$  within the given constraints. We do this with the modified gradient descent function described in section 3.3.

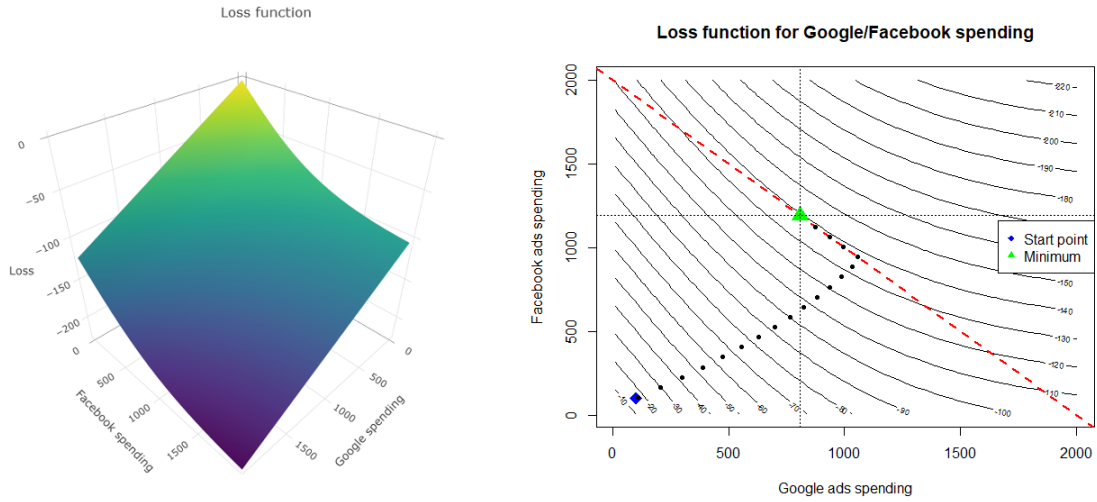


Figure 23: Left: The gradient descent loss function with  $-\hat{Y}_{tot}$  as a function of  $Q$  and  $X_{21}$ . Right: The loss function as a contour plot showing the path of the gradient descent algorithm.

The minimum value within the constraints were found at  $Q = 808$  and  $X_{21} = 1192$ . The value at those coordinates was  $-139$ . (Values have been rounded to integers).

This means that we get the highest expected number of clicks by spending 808 SEK on Google ads and 1192 SEK on Facebook ads. The expected number of clicks per month are then  $\hat{Y}_{tot} = 139$ .

This can be compared to 103 expected clicks if we spend all money on Google ads or 120 expected clicks if we spend all money on Facebook ads.

Since we spend the full 2000 in this example, the problem can be reduced to one dimension as described in section 6.2.1. Presented solution

$$Q = \frac{R^\omega}{\hat{\psi}} \log \left[ \frac{\hat{\psi}(\hat{\beta}_{11} + \hat{\beta}_{12}D) - \hat{\beta}_{21}}{\hat{\beta}_{21}} \pm \sqrt{\left( \frac{\hat{\beta}_{21} - \hat{\psi}(\hat{\beta}_{11} + \hat{\beta}_{12}D)}{\hat{\beta}_{21}} \right)^2 - 1} \right]$$

for given client leads to  $Q = \pm 808$ . Thus, the numerical gradient descent optimization and the analytical optimization confirm each other.

## 7.2 Prediction interval for the maximal total number of paid clicks

To be able to give a prediction interval, we calculate variance of our point estimate of the total number of paid clicks.

$$\hat{\sigma}_{tot} = \sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_{1,\beta}^2 + \hat{\sigma}_2^2 + \hat{\sigma}_{2,\beta}^2} \quad ,$$

where

$$\begin{aligned} \hat{\sigma}_1^2 &= (s.e._1 T_1)^2 \\ &= (33.03173 \cdot 1.002831)^2 \\ &= 1097.282 \end{aligned}$$

$$\begin{aligned} \hat{\sigma}_{1,\beta}^2 &= (s.e._1 \cdot T_1 (\mathbf{X}_{1,0}^T (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_{1,0}))^2 \\ &= \left( 33.03173 \cdot 1.002831 \cdot [1.005038 \quad 4219.854816] \begin{bmatrix} 4.185 \cdot 10^{-2} & -1.842 \cdot 10^{-6} \\ -1.842 \cdot 10^{-6} & 7.736 \cdot 10^{-10} \end{bmatrix} \begin{bmatrix} 1.005038 \\ 4219.854816 \end{bmatrix} \right)^2 \\ &= 1.79308 \end{aligned}$$

$$\begin{aligned} \hat{\sigma}_2^2 &= (s.e._2 T_2)^2 \\ &= (0.04981407 \cdot 99.90324)^2 \\ &= 24.76642 \end{aligned}$$

$$\begin{aligned} \hat{\sigma}_{2,\beta}^2 &= (s.e._2 T_2 (\mathbf{X}_{2,0}^T (\mathbf{X}_2^T \mathbf{X}_2)^{-1} \mathbf{X}_{2,0}))^2 \\ &= (0.04981407 \cdot 99.90324 \cdot 99.90324 \cdot (11.93155 \cdot 1.379 \cdot 10^{-4} \cdot 11.93155))^2 \\ &= 9.545056 \cdot 10^{-3} \end{aligned}$$

$$\hat{\sigma}_{tot} = \sqrt{1097.282 + 1.79308 + 24.76642 + 9.545056 \cdot 10^{-3}} = 33.52388$$

Notice how the error term in the Google model, estimating  $\sigma_1^2$  accounts for the large part of the total error.

Under model:

$$Y_{tot} = \hat{Y}_{tot} + e_{tot}, \quad e_{tot} \sim N(0, \hat{\sigma}_{tot})$$

we obtain estimates  $\hat{\sigma}_{tot} = 33.52$  and  $\hat{Y}_{tot} = 139$ . Hence, for a two sided 70 percent prediction interval, the error margin is  $Z_{0.85} \cdot 33.52 = 34.74$ . Rounding it to natural numbers the prediction interval for maximal amount of paid clicks is given as

$$I_{Y_{tot}}^P = (139 \pm 35).$$

This gives the highest number of predicted clicks that the client can receive by choosing the optimal resource allocation calculated in section 7.1.

### 7.3 Minimizing the variance for marketing mix

Above we present the estimate for expected number of clicks generated by our Google and Facebook marketing channels. The estimate can be maximized as a function of  $Q$  and  $X_{21}$  under a budget constrain. Such maximization provides us a mix of marketing channels that maximize the response function. However, knowing the variance associated to each of the marketing channel we can also be interested in minimizing the variance of our marketing mix. In such case, , as described by Glombek (2014) [9], the weights will be a function of the covariance matrix for the marginal distributions:

$$w_{MP} = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}^T \Sigma^{-1} \mathbf{1}},$$

where  $\Sigma$  is the covariance matrix and  $\mathbf{1}$  is a 1-vector.

For our model, the variance is minimized:

$$\hat{w}_{MP} = \frac{\begin{bmatrix} \hat{\sigma}_G^2 & \hat{\sigma}_{G,FB} \\ \hat{\sigma}_{G,FB} & \hat{\sigma}_{FB}^2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix}}{\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \hat{\sigma}_G^2 & \hat{\sigma}_{G,FB} \\ \hat{\sigma}_{G,FB} & \hat{\sigma}_{FB}^2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix}} = \frac{\begin{bmatrix} 1099.2 & 0 \\ 0 & 24.8 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix}}{\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1099.2 & 0 \\ 0 & 24.8 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix}} = \begin{bmatrix} 0.022 \\ 0.978 \end{bmatrix}$$

Thus, to minimize the variance of the model, we allocate 2.2 percent of the spending on Google ads and 97.8 percent of the spending on Facebook ads.

This allocation leads to an expected number of clicks  $\hat{Y}_{tot} = 121$ , and an error margin of 15 at 70 percent confidence.

Thus, the minimum variance outcome is:

$$I_{Y_{tot}}^P = (121 \pm 15)$$

## 8 Conclusions and discussion

Compared to a model that assumes linearity, we can get a considerably better model fit by basing the marginal response functions on the underlying logic of the ad auctions. While response functions developed for traditional marketing are less than ideal for pay-per-click advertisement, models suitable for production functions – in this case the generalized logistic function – is well suited to model Google ads auctions.

When spending is low in relation to the size of a particular ad market, as we see in the Facebook data, a linear function can be a good approximation. However, we need to be aware that such linear model may severely overestimate the response value for higher levels of spending. Therefore, the model should only be applied in situations where the spending is within or close to the variable space of the training data.

The Google ads marginal function increases with the number of healthcare professionals, and the effect is greater at large levels of spending. More helthcare professionals at the clinic delays the diminishing property of the function.

This may be a result of market forces: A larger market for the set of services offered at a clinic results in more paid clicks, and a larger market also enables clinics to grow in terms of staff numbers. In other words, both the number of healthcare professionals and the number of paid clicks may be results of market size.

When population density increases, the number of paid clicks decreases given the other variables of the model. A possible explanation to this could be that high population density means increased competition.

While the Facebook ads model is simple compared to the Google ads model, this probably does not mean that the "true" Facebook ads model is simpler. It only means that in the variable space where we have our observations, we cannot see the complexities of the true Facebook function. A reasonable guess is that with larger levels of spending, the Facebook model would diminish in a similar manner as the Google ads function.

It is important to assess the robustness of the model with regards to the assumptions. We should be aware whether small errors in the assumptions result in major changes in the end result, or whether they have little impact.

The assumptions made in our case are on quite steady ground. The logic of the ad auctions implies that we do not have an intercept, that returns are diminishing and that saturation of the market is possible.

None of the model's parameter values are assumed. Instead they are fit to the training data using a gradient descent algorithm.

What remains to scrutinize is the shapes of the response functions. In the case of the Facebook data, a different shape may – as mentioned above – better capture the response to high levels of spending. Regarding the Google response function, we noted that the chosen functional shape was superior to a linear regression without variable transformations. However, the accuracy of the chosen logistic model was similar to the more crude piecewise model to which it was compared. As with the Facebook marginal function, we know little about which functional shape that best captures levels of spending that are significantly above those in the training data.

As shown in section 6.2.1, a simple case where we allocate a fixed sum of money between two marketing channels can be reduced to a one-dimensional optimization problem. Despite this, the model is constructed as a multidimensional problem which makes it generalizable to higher dimensions. Any number of additional marketing channels can be added to the mix without changes in the overall model.

The point estimate of the total number of clicks is a simple addition of the point estimates of the marginal functions. Also the variances of the marginal models can be added assuming that we have enough observations for a normal approximation of the combined error distribution. If this assumption is not fulfilled, the combined error t-distributions can also be accurately simulated.

The error margins of the model are large, which to some degree can be a consequence of large random variations. It is possible that the model can be improved with additional relevant explanatory variables.

Within different subject areas there will be different expectations on accuracy. In marketing, the expectation of accuracy will obviously be less strict than in, for instance, some branches of natural science. When the error margins are larger, it can be reasonable to lower the confidence of the prediction interval rather than letting the error margins run too wide. A 70 percent prediction interval should be acceptable in the context of a marketing effort/response model.

To further evaluate the model, data will need to be collected from businesses who split campaigns between different marketing channels rather than, as now, running one campaign in one marketing channel.

Until then, we will have to be content evaluating the marginal models against new data.

A paid click does not have a value in itself, but it is instrumental as a means to recruit new patients to a clinic. The value of each paid click depends on the probability of the click leading to a new patient and that patient's lifetime value as a client. Also the probability that a patient enrolled through a marketing campaign would have found the clinic through other means must be taken into account.

With the number of paid clicks as response variable, the model will always recommend spending the entire proposed budget. An increased budget can never lead to a decreased number of paid clicks. However, in reality, it may be desirable to only spend a part of the proposed budget.

To know what part of the budget to spend, it should be possible to measure the output of the marketing campaign in the same unit as the input (for instance SEK, EUR or USD). We will then be able to determine at what level the expected return on additional spending is high enough to justify the extra cost.

A possible extension of the model developed in this thesis would involve determining the value of a click – the value can be different depending on the marketing channel as well as other variables – so that we can decide how much spending is ideal.

To calculate the expected value of a marketing campaign, conversion is a key metric. Collecting conversion data requires code added to the client website. This is already implemented on some client websites, and

a suggestion for the future would be to increase the efforts to implement it with as many clients as possible. In addition to this, statistics on the expected lifetime value of a new patient would contribute to an extended model with expected income as response variable rather than the number of paid clicks.

## 9 References

### References

- [1] D. S. Bhoj, D. Kushary. (2012) *Combining independent t-variables* Available from: <http://interstat.statjournals.net/YEAR/2012/articles/1210001.pdf> [Accessed 2019-05-12]
- [2] J. Brownlee (2014) *A Simple Intuition for Overfitting, or Why Testing on Training Data is a Bad Idea* Available from: <https://machinelearningmastery.com/a-simple-intuition-for-overfitting/> [Accessed 2019-05-12]
- [3] G. Casella. (1983) Leverage and Regression Through the Origin. *The American Statistician* 37 (2), 147-152.
- [4] T. Chai, R. R. Draxler. (2014) Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development* 7, 1247–1250.
- [5] J. G. Eisenhauer. (2003) Regression through the origin. *Teaching Statistics*. 25 (3), 76-80.
- [6] B.S. Everitt, A. Skrondal (2010) *The Cambridge dictionary of Statistics* , 4th ed. Cambridge, Cambridge University Press.
- [7] Facebook (n.d.) *Best Practices: Ad Bidding and Budgeting*. Available from: <https://www.facebook.com/business/a/ad-bidding> [Accessed 2019-05-12].
- [8] S. Glen (2015) *Interpreting the Variance Inflation Factor*. Available from: <https://www.statisticshowto.datasciencecentral.com/variance-inflation-factor/> [Accessed 2019-05-12].
- [9] K. Glombek (2014) Statistical Inference for High-Dimensional Global Minimum Variance Portfolios. *Scandinavian Journal of Statistics* 41 . 846.
- [10] T. O. Kvalseth. (1985) Cautionary Note about R<sup>2</sup>. *The American Statistician*. 39 (4), 279-285.
- [11] H. Lohninger (2012) *Fundamentals of statistics* Available from: [http://www.statistics4u.com/fundstat\\_eng/cc\\_normality\\_test.html](http://www.statistics4u.com/fundstat_eng/cc_normality_test.html) [Accessed 2019-05-12]
- [12] D. Montgomery. (2017) *Design and Analysis of Experiments* 85-86, Hoboken, New Jersey, John Wiley Sons.
- [13] R. Nau. (n.d.) *Statistical forecasting: notes on regression and time series analysis*. Available from <http://people.duke.edu/~rnau/411home.htm> [Accessed 2019-05-12].
- [14] J. Saunders. (1987) The Specification of Aggregate Market Models. *European Journal of Marketing*. 21 (2), 5-47.
- [15] A. Vexler , A. D. Hutson & X. Chen (2016) *Statistical Testing Strategies in the Health Sciences* p.66 Chapman & Hall/CRC Biostatistics Series. Boca Raton, Florida, Chapman and Hall/CRC
- [16] C. Zaiontz. (n.d.) *Confidence and Prediction Intervals*. Available from: <http://www.real-statistics.com/multiple-regression/confidence-and-prediction-intervals/> [Accessed 2019-05-12].

## Appendix 1 – R code

### Fitting the parameters for the Google ads model

```
##Load required libraries and the Google dataset
library(dplyr)
library(caret)
google_df <- read.csv("../ ../ ../data_ptj_google_ads/google_df.csv", encoding="UTF-8",
                      stringsAsFactors = F)

##Fit a and b for candidate model 1 using the modified gradient descent algorithm
##See the gradient descent algorithm in separate code chunk
source("../ ../ ../gradient_descent/gradient_descent_3.R")
gradient_descent(c(1, 1), f=loss_f_initial, stepsize=c(10^6,10^3))

##Calculate the std error, S-W test, K-S test for candidate model 1
##Tau and gamma are given by the gradient descent algorithm above
tau <- 1.0428121
gamma <- 0.5640559
a2 <- tau *google_df$number_qualified_employees
google_df$modified_cost <-
  ifelse(google_df$Cost > a2, a2+(google_df$Cost-a2+1)^gamma-1, google_df$Cost)
google_df$density2 <- google_df$density*google_df$modified_cost
set.seed(123)
kmod<-train(Clicks ~ modified_cost+density2-1, data=google_df, method="lm",
            trControl=trainControl(method="cv", number=10, verboseIter=F),
            tuneGrid = expand.grid(intercept = FALSE))
kmod$results$RMSE^2
shapiro.test(kmod$finalModel$residuals)
ks.test(kmod$finalModel$residuals, "pnorm", mean(kmod$finalModel$residuals),
        sd(kmod$finalModel$residuals))

## fit omega and psi for candidate model 2
source("../ ../ ../gradient_descent/gradient_descent_3.R")
gradient_descent(c(0.5, 0.01), f=loss_f_sigmoid, stepsize=c(10^3,0.1))

##Calculate the std error, S-W test, K-S test for candidate model 1
sigmoid6 <- function(x, k, psi, omega) 2*k^omega / (1 + exp(-x*psi/k^omega)) - k^omega
omega <- 0.352424292
psi <- 0.003027473
google_df$modified_cost2 <- sigmoid6(google_df$Cost,
                                    google_df$number_qualified_employees,
                                    psi, omega)
google_df$density2 <- google_df$density*google_df$modified_cost2
kmod<-train(Clicks ~ modified_cost2+density2-1, data=google_df, method="lm",
            trControl=trainControl(method="cv", number=10, verboseIter=F),
            tuneGrid = expand.grid(intercept = FALSE))
kmod$results$RMSE^2
shapiro.test(kmod$finalModel$residuals)
ks.test(kmod$finalModel$residuals, "pnorm", mean(kmod$finalModel$residuals),
        sd(kmod$finalModel$residuals))

##Fit and test the linear benchmark model
```



```

kmod<-train(Clicks ~ Cost+number_qualified_employees+'Population density'-1,
           data=google_df, method="lm",
           trControl=trainControl(method="cv", number=10, verboseIter=F),
           tuneGrid = expand.grid(intercept = FALSE))
kmod$results$RMSE^2
shapiro.test(kmod$finalModel$residuals)
ks.test(kmod$finalModel$residuals, "pnorm", mean(kmod$finalModel$residuals),
        sd(kmod$finalModel$residuals))

##Remove outliers
google_df <- google_df[kmod$finalModel$residuals<110,]

## refit omega and psi for sigmoid model without outliers
source("../.../gradient_descent/gradient_descent_3.R")
gradient_descent(c(0.5, 0.01), f=loss_f_sigmoid, stepsize=c(10^3,0.1))

omega <- 0.362847271
psi <- 0.002898401

##Set a to 0.36 based on the analysis described in section:
##"Transforming the Google ads variables for homoscedasticity and normality"
a <- 0.36

##Calculate the final std error, S-W test, K-S test for the Google model
google_df$modified_cost2 <- sigmoid6(google_df$Cost,
                                   google_df$number_qualified_employees,
                                   psi, omega)
google_df$density2 <- google_df$density*google_df$modified_cost2
google_df$clicks_sq <- google_df$Clicks/google_df$modified_cost2^a
google_df$cost_sq <- google_df$modified_cost2/google_df$modified_cost2^a
google_df$density_sq <- google_df$density2/google_df$modified_cost2^a
kmod<-train(clicks_sq ~ cost_sq+density_sq-1, data=google_df, method="lm",
           trControl=trainControl(method="cv", number=10, verboseIter=F),
           tuneGrid = expand.grid(intercept = FALSE))
kmod$results$RMSE
shapiro.test(kmod$finalModel$residuals)
ks.test(kmod$finalModel$residuals, "pnorm", mean(kmod$finalModel$residuals),
        sd(kmod$finalModel$residuals))

```

## Fitting the parameters for the Facebook ads model

```

library(dplyr)

fb <- read.csv("../.../facebook_data/fb.csv", stringsAsFactors = F)

##Facebook fit and test initial model
fbmod <- lm(fb$Clicks ~ fb$Cost-1)
shapiro.test(fbmod$residuals)

##Remove outliers
fb <- fb[fbmod$residuals/fbmod$fitted.values<1.2,]

##Set a to 0.65 based on the analysis described in section:
##"Transforming the Facebook ads variables for homoscedasticity and normality"
a <- 0.65

```

```

##Formulate and test the final, transformed model
fb$ $sq\_clicks$  <- fb$Clicks/fb$Cost^a
fb$ $sq\_cost$  <- fb$Cost/fb$Cost^a
fbmod<-train( $sq\_clicks$  ~  $sq\_cost-1$ , data=fb, method="lm",
            trControl=trainControl(method="cv", number=10, verboseIter=F),
            tuneGrid = expand.grid(intercept = FALSE))
fbmod$results$RMSE
shapiro.test(fbmod$finalModel$residuals)
ks.test(fbmod$finalModel$residuals, "pnorm", mean(fbmod$finalModel$residuals),
        sd(fbmod$finalModel$residuals))

```

## A modified gradient descent function

```

library(dplyr)

##Calculate a function agnostic numerical approximation of the gradient
get_gradient_numerical <- function(coordinates, f, epsilon=10^(-4)){
  gradients <- sapply(1:length(coordinates), function(q){
    local_coordinates1 <- local_coordinates2 <- coordinates
    local_coordinates1[q] <- local_coordinates1[q] - epsilon
    local_coordinates2[q] <- local_coordinates2[q] + epsilon
    (f(local_coordinates2) - f(local_coordinates1)) / 2*epsilon
  })
  gradients
}

##Function to constrain the variables/parameters within the allowed space
adjust_coordinates <- function(coordinates, old_coordinates, f,
                               the_min=-Inf, the_max=Inf, xplusy=F){
  if(coordinates[1] < the_min) coordinates[1] <- the_min
  if(coordinates[2] < the_min) coordinates[2] <- the_min
  if(coordinates[1] > the_max) coordinates[1] <- the_max
  if(coordinates[2] > the_max) coordinates[2] <- the_max
  if(xplusy && sum(coordinates) > the_max){
    alt_coordinates <- list()
    alt_coordinates[[1]] <- c(coordinates[1], the_max - coordinates[1])
    alt_coordinates[[2]] <- c(the_max - coordinates[2], coordinates[2])
    alt_coordinates[[3]] <- old_coordinates
    which_coordinates <- which.min(c(f(alt_coordinates[[1]]),
                                   f(alt_coordinates[[2]]),
                                   f(alt_coordinates[[3]])))
    coordinates <- alt_coordinates[[which_coordinates]]
  }
  coordinates
}

##The main gradient descent function
##Starting_coordinates should be a vector of coordinates
gradient_descent <- function(starting_coordinates,
                             get_gradient=get_gradient_numerical,
                             f=loss_f,
                             stepsize=1,
                             stop_threshold = 10^-9,
                             plot_contour=F,

```

```

        xplusy=F,
        constrains=F) {
if (plot_contour) points(starting_coordinates[1], starting_coordinates[2],
                        col="blue", pch=18, cex=2)
coordinates <- old_coordinates <- starting_coordinates
counter <- 0
repeat{
  gradient <- get_gradient(coordinates, f)
  coordinates <- coordinates - gradient*stepsize
  if (constrains) coordinates <-
    adjust_coordinates(coordinates, old_coordinates, f,
                      the_min=0, the_max=2000, xplusy=xplusy)
  print(c(coordinates, gradient, f(coordinates),
          sum((coordinates-old_coordinates)^2)))
  if (all(coordinates==old_coordinates)){
    stepsize <- stepsize / 2
    counter <- counter+1
    next
  }
  if (plot_contour && counter%%10==0){
    points(coordinates[1], coordinates[2], col="black", pch=20)
    Sys.sleep(0.1)
  }
  if (sum((coordinates-old_coordinates)^2) < stop_threshold){
    if(plot_contour) points(coordinates[1], coordinates[2],
                          col="green", pch=17, cex=2)
    return (coordinates)
  }
  old_coordinates <- coordinates
  counter <- counter+1
}
}

##Loss function for Google ads candidate model 1
loss_f_initial <- function(v){
  a <- v[1]
  b <- v[2]
  a2 <- a *google_df$number_qualified_employees
  google_df$modified_cost <-
    ifelse(google_df$Cost > a2, a2+(google_df$Cost-a2+1)^b-1, google_df$Cost)
  google_df$density2 <- google_df$density*google_df$modified_cost
  kmod <- lm(Clicks ~ modified_cost+density2-1, data=google_df)
  mean((google_df$Clicks - kmod$fitted.values)^2)
}

##Loss function for Google ads candidate model 2
sigmoid6 <- function(x, k, L, M) 2*k^M / (1 + exp(-x*L/k^M)) - k^M
loss_f_sigmoid <- function(v){
  M <- v[1]
  L <- v[2]
  google_df$modified_cost2 <- sigmoid6(google_df$Cost,
                                     google_df$number_qualified_employees, L, M)
  google_df$density2 <- google_df$density*google_df$modified_cost2
  kmod <- lm(Clicks ~ modified_cost2+density2-1, data=google_df)
}

```

```

    mean((google_df$Clicks - kmod$fitted.values)^2)
}

##Loss function for the final Google ads model
sigmoid6 <- function(x, k, L, M) 2*k^M / (1 + exp(-x*L/k^M)) - k^M
loss_f_sigmoid_trans <- function(v){
  i <- 0.36
  M <- v[1]
  L <- v[2]
  google_df$modified_cost2 <- sigmoid6(google_df$Cost,
                                       google_df$number_qualified_employees, L, M)
  google_df$density2 <- google_df$density*google_df$modified_cost2
  google_df$clicks_sq <- google_df$Clicks/google_df$modified_cost2^i
  google_df$cost_sq <- google_df$modified_cost2/google_df$modified_cost2^i
  google_df$density_sq <- google_df$density2/google_df$modified_cost2^i
  kmod <- lm(clicks_sq ~ cost_sq+density_sq-1, data=google_df)
  mean((google_df$Clicks - kmod$fitted.values*google_df$modified_cost2^i)^2)
}

```