Grote Bickersstraat 74
1013 KS Amsterdam

PO Box 247
1000 AE Amsterdam
The Netherlands

t   +31 20 522 54 44
f   +31 20 522 53 33
e   info@tns-nipo.com
www.tns-nipo.com

Report

Marketing Data Analysis Techniques (M.D.A.T.)

Maikel Groenewoud

June 2005

Preface

This paper is the result of the research conducted during my internship at TNS NIPO Healthcare in Amsterdam. The internship was the final project I had to complete before I could receive my Master's degree in Business Mathematics and Informatics from the Free University in Amsterdam.

I would like to thank my supervisors at TNS NIPO Healthcare, Carolien Hendrix and Linda Abrams, for giving me the opportunity to perform this research within this organization. I would also like to thank my supervisors at the university, Geurt Jongbloed and René Swarttouw, for their advice concerning the mathematical components of my research.

Maikel Groenewoud

Table of contents

1 Introduction

My research was conducted at the Dutch Institute for the Public Opinion and Marketing Research, better known as TNS NIPO. TNS NIPO was established in 1945 and has the highest overall market share in the Netherlands. Since 1999 this organization formerly known as NIPO is part of the international marketing research agency Taylor Nelson Sofres (TNS). TNS has approximately 200 offices in about 70 countries with more than 14.000 employees worldwide. They also have a very large pool of interviewers to perform the fieldwork (data collection).

The business unit TNS NIPO Healthcare, established in 1994, focuses on marketing research for the medical and pharmaceutical industry. This means research in the field of healthcare among private practitioners, medical specialists and paramedics but also among (potential) patients. At the business unit level the strategic decisions are about developing and sustaining a competitive advantage for the provided goods and delivered services. The goal of marketing research is to provide managers with the information they need to support them in the process of making important marketing decisions such as the introduction of new products and services. My research specifically focused on the business unit TNS NIPO Healthcare and their market.

The analysis of the situation within the own company is referred to as internal analysis and outside of the own company as external analysis. They are of strategic importance to a firm because they allow a firm to better match their resources and capabilities to (developments in) the competitive environment in which it operates. They provide information that is of vital importance for the formulation and selection of a business and marketing strategy. The firm's internal attributes are classified as strengths and weaknesses and the external environment presents opportunities and threats for the firm. The resources and capabilities are a firm's strengths and can be used as a basis for developing a competitive advantage. The absence of certain resources and/ or capabilities can be viewed as weaknesses. In some cases a weakness can also be the flip side of a strength. The external analysis can reveal new opportunities for profit and growth. However, changes in the external environment may also present threats to the firm. When making a strategy a firm should pursue opportunities that are a good fit to its strengths and try to overcome weaknesses that might prevent them from doing so. Another thing that a firm should do, is use its strengths to reduce its vulnerability to external threats and find a way to prevent its weaknesses from making it vulnerable to external threats.

The main goal of my research is to illustrate how data analysis techniques can be used to analyze large amounts of business data. The project managers of TNS NIPO Healthcare are particularly interested in predicting characteristics of competitors, clients and respondents and in segmenting them. A lot of data are available at TNS NIPO but there more can be done with them. A great deal of strategically relevant information can still be uncovered when using the right techniques and tools for analysis. The project managers of TNS NIPO Healthcare want to know if data analysis techniques can be used to extract the following types of information concerning their competitors:

- Position of their own business unit relative to them
- Different types of competitors
- Strongest competitors
- Characteristics such as revenue, number of employees etc.
- Best predictors for revenue, number of employees etc.

There are several other agencies in the Netherlands performing similar activities as TNS NIPO Healthcare. To determine the position of one's own firm, it is important to know what the strengths and weaknesses of these competitors are. TNS NIPO Healthcare has many clients ranging from pharmaceutical companies to hospitals. The same analysis applied to the competitors could of course also be applied to other companies such as for instance these clients of TNS NIPO Healthcare.

Concerning the respondents, which form the market of their clients, the project managers want to know if and how the following information can be extracted:

- Segments based on 'naturally' occurring patterns and relationships in the data
- Best predictors for chronic illnesses etc.

Market segmentation means that a heterogeneous market is divided into smaller homogenous groups (segments). This allows for a more specific targeting of each group. The big advantage of determining segments by searching for 'naturally' occurring patterns and relationships is that no prior knowledge of the data is needed. Predefined segments often contain a lot of vague unscientific assumptions about the data which really do not necessarily have to be true. Those segments are in fact nothing more than guesses that are not based on quantifiable evidence.

The modeling of missing values and outliers is also a big part of my research because I sometimes had to deal with incomplete and/or irregular data. After missing values had been replaced by 'plausible' predictions, techniques for the analysis of complete data could be used. The competitor data that have been analyzed was gathered through an external analysis of TNS NIPO Healthcare. The external analysis consists of three components:

- Competitors
- Clients (and their market)
- Developments & Trends in the market

Statistical data analysis was used in the first two phases. The analysis of developments & trends is also of great strategic importance because the rules healthcare and marketing research are constantly changing. This means that long-term strategic planning is only possible if these developments and trends are regularly monitored. Because of the constant changes, a firm's strategy has to be updated regularly too.

The next three chapters are about data analysis. The first, Analysis Techniques, looks at this analysis from a theoretical point of view. The second, Analysis Problems, discusses the specific problems that were encountered while analyzing the data. In the third, Analysis Results, the results from the analysis of some competitor and respondent data are discussed. In the final chapter my overall findings are presented. The first appendix contains fictitious competitor data that were analyzed. The second appendix contains so-called star plots of these competitor data.

2 Analysis Techniques

The data gathered through the external analysis of TNS NIPO Healthcare can be further analyzed using data analysis techniques. In the introduction it was already stated that an important aspect of my research is to illustrate how data analysis techniques can be used to predict characteristics of competitors, respondents etc. and to segment them. Some of these techniques require complete data which means that for these data to be analyzed, it cannot contain any missing values. In practice the data however are not always complete but fortunately there are also techniques available for handling missing values. Some of these techniques generate ‘plausible’ predictions to replace the missing values and will be discussed in the next chapter.

In this chapter some data analysis techniques will be discussed from a more or less ‘theoretical’ point of view. Here is an overview of the topics that will be discussed:

- 2.1 Multiple regression
- 2.2 Hierarchical clustering
- 2.3 Multidimensional scaling
- 2.4 Recursive partitioning

Multiple regression was used to generate predictions. Hierarchical clustering and multidimensional scaling were used to visualize and segment the individual cases in the data. Recursive partitioning also segments the data.

2.1 Multiple regression

The term multiple regression has been around for a long time and was first used at the beginning of the last century. The general purpose is to gain further insight into the relationship between several independent explanatory/ predictor variables and a dependent target/ response variable. The project managers of TNS NIPO Healthcare were for instance quite interested in predicting the revenue of competitors based on other variables.

There are several types of regression but this section solely focuses on the linear variant. The linear model to handle multiple explanatory variables is expressed as follows:

$$y_i = \alpha + \theta_1 x_{i1} + \ldots\ldots + \theta_n x_{in} + \varepsilon_i$$

with $i = 1,\ldots,N$,
n = number of explanatory variables,
N = number of cases,
y_i = response variable i,
α = intercept or constant term,
$\theta_1,\ldots\ldots\ldots, \theta_n$ = regression coefficients,
$x_{i1},\ldots\ldots\ldots,x_{in}$ = explanatory variables,
and ε_i = error corresponding to case i

This model can also be expressed as:

$$Y = X\theta + \varepsilon,$$

Where X is a Nx$(n+1)$ matrix containing a column with ones and the explanatory variables and θ contains the intercept α and the regression coefficients. The Y in the formula is a vector containing the N response variables and ε denotes the vector of errors. (If there would be multiple dependent variables, Y would be a matrix.)

To estimate the regression coefficients classical least squares (LS) is used. The goal is to minimize the squared deviations of the cases in the data to the regression model. The deviations are also referred to as residual and LS minimizes the sum of the squared residuals:

$$\theta \approx \hat{\theta}_{LS} = \arg\min_{\theta} \sum_{i=1}^{N} r_i^2$$

The amount of variance of the observed response y_i explained by the variance of the residuals r_i, is referred to as the coefficient of determination R^2:

$$0 \le R^2 = \frac{\mathrm{Var(y_i)} - \mathrm{Var}(r_i)}{\mathrm{Var(y_i)}} = 1 - \frac{\mathrm{Var(r_i)}}{\mathrm{Var(y_i)}} \le 1$$

with $r_1, \ldots\ldots, r_n$ = the residuals,
and $y_1, \ldots\ldots, y_n$ = the response variables

So this statistic tells the user how well the model fits the data. To determine the least squares estimators of the regression coefficients, a solution must be found for the following equations:

$$X'X\theta = X'Y$$

These are called the normal equations and if X is of rank $n+1$, a solution for these equations is given by:

$$\theta = (X'X)^{-1}X'Y$$

An assumption that is often made is that the errors ε are independent with a normal distribution. This should be checked after having constructed the model for instance by creating QQplots of the residuals. These plots give a good indication of the validity of the normality assumption. In linear regression all the variables are only allowed to assume numerical values.

If models could be constructed with which for instance the revenue of a company could be predicted based on some other variables, most likely values could be found for this revenue for various companies by using these models. One would prefer a model with just a few explanatory variables but still a large predictive value for the response. There are various methods for finding such model, one of which is stepwise model selection. In each step a different set of explanatory variables is used to construct the model until the 'best' one is found. If the omission or inclusion of a single variable has no or very little influence on the residual sum of squares, then that variable is left out of the model.

The major conceptual limitation of all regression techniques is that one can only ascertain relationships/ correlations, but can never be completely sure about the underlying causal mechanisms. The reason for this is that often in fact only part of the variables which would potentially be of interest is available in the data used to construct the various models.

Multiple regression has been used to generate various predictions but during that process several problems such as missing values and outliers were encountered. These problems need to be handled with care and my approach to them will be discussed in the chapter Analysis Problems. The following three sections are about techniques that require complete data for an analysis to take place.

2.2 Hierarchical clustering

The term cluster or clustering analysis has been around quite a while and was first used in the 1930's. Hierarchical clustering is one of the forms of cluster analysis and it allows the researcher to visualize and segment cases in the data X (N x n data matrix) of interest. The number of cases is denoted by N and the number of variables by n. Unlike the regression techniques described in the previous section, there is not any distinction between explanatory and response variables now. Hierarchical clustering is of interest for TNS NIPO healthcare because the project managers would like to determine the nature of their competition. By definition the following assumptions are made:

1. The distance of case i to case j is non-negative

   $$\delta_{ij} \geq 0$$

2. The distance of a case to itself is equal to zero

   $$\delta_{ii} = 0$$

3. The distance of case i to case j is equal to the distance of case j to case i

   $$\delta_{ij} = \delta_{ji}$$

4. The triangle inequality holds

   $$\delta_{ij} \leq \delta_{it} + \delta_{tj}$$

The Euclidean distance is commonly used to measure the distances between cases:

$$\delta_{ij} = \{ \sum_{h=1}^{n} (x_{ih} - x_{jh})^2 \}^{1/2}$$

with $x_{i1}, \ldots\ldots\ldots\ldots, x_{in}$ = the variables of case i,
and $x_{j1}, \ldots\ldots\ldots\ldots, x_{jn}$ = the variables of case j

The distances between cases are also referred to as dissimilarities. Often the variables are standardized before the dissimilarities are calculated. This is done to compensate for differences in unit or scale of measurement between variables. Company revenue is for instance measured in a completely different unit of measurement than the number of employees. The distances can be greatly affected by differences in unit or scale of measurement among the variables from which the distances are computed. That is why in practice it generally is a good habit to standardize the variables. If the data are completely categorical in nature, the percentage of disagreement is a very useful alternative distance measure:

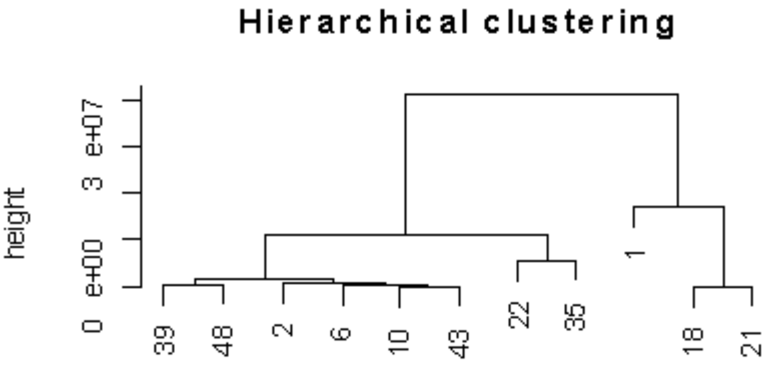
$$\delta_{ij} = \frac{\sum_{h=1}^{n} \Xi(x_{ih} \neq x_{jh})}{n}$$

with n = number of variables,

$$\text{and } \Xi = \text{an indicator function} \begin{cases} 1 \text{ if condition is satisfied,} \\ 0 \text{ if condition is not satisfied} \end{cases}$$

This measure gives the fraction of variables for which two cases have dissimilar values.

The results of a hierarchical clustering are often presented in tree-like structures known as dendrograms. Here is an example:



In these plots, the height denotes the linkage distance. For each node in the graph we can determine at which value of this distance measure the respective elements were linked together into a new single cluster. When the data contain clear structures in terms of clusters of cases that are similar to each other, then this structure will be reflected in the hierarchical tree as distinct branches. Groups consisting of a single observation such as observation '1' could be seen as outliers, however that does not always have to be the case. More will be said about outliers in the next chapter.

In Hierarchical clustering each case is initially defined as a cluster in itself and then iteratively at each stage the two most similar clusters are joined until there is just a single cluster. So one links more and more cases together and aggregates (*amalgamates*) larger and larger clusters of increasingly dissimilar elements. Then finally in the last step all cases are joined together. At the first step each case represents its own cluster and the distances between cases are (usually) defined by the Euclidean distance measure. When several cases have been linked together the distances between the new clusters have to be determined. A linkage or amalgamation rule is needed to determine when two clusters are sufficiently similar to be linked together. There are various rules to achieve this but here are the three most commonly used (Jain, Murty and Flynn, 1999):

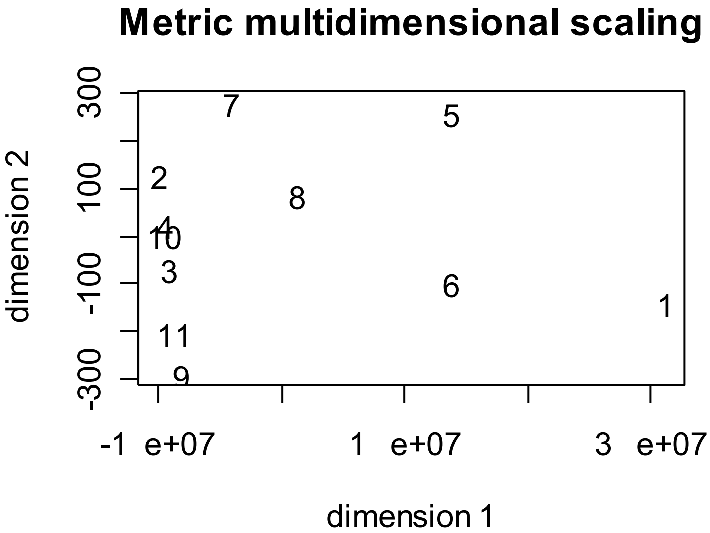
- Single linkage (nearest neighbor)
  Two clusters are linked together when any two cases in the two clusters are closer together than the respective linkage distance. The 'nearest neighbors' across clusters are used to determine the distances between clusters.
- Complete linkage (furthest neighbor)
  One could also evaluate the neighbors across clusters that are furthest away from each other. The distances between clusters are determined by the greatest distance between any two cases in the different clusters.
- Pair-group average
  In this method the distance between two clusters is calculated as the average distance between all pairs of objects in the two different clusters.

2.3 Multidimensional scaling

Multidimensional scaling (MDS) is a technique used to visualize and segment the data. There are in fact two types of MDS: metric and non-metric. Metric MDS will be discussed in the first part of this section and nonmetric MDS in the second part. This section is partly based on Cox and Cox (1994).

2.3.1 Metric multidimensional scaling

Metric MDS was the first type of MDS to be developed and that is why it is also referred to as classic MDS. Metric MDS takes a set of dissimilarities (distances) between cases and returns a set of points preferably in a low dimensional space such that the dissimilarities between these points are approximately equal to the given or 'observed' dissimilarities. The goal is to detect meaningful underlying dimensions to visualize observed dissimilarities. MDS attempts to visually arrange cases in a space with a certain dimension so as to reproduce the observed dissimilarities as well as possible. As a result it is possible to explain those dissimilarities in terms of the underlying dimensions. Cases nearer to each other are more similar. MDS could for instance be used to visualize the competitors of TNS NIPO Healthcare so one could determine how similar they are. Business data often consist out of a lot of variables which makes it increasingly difficult to quickly find valuable information. That is why one would hope for the number of dimensions for a configuration to be (considerably) less than the number of variables. Here is an example of an MDS plot:



Case '1' is clearly very different from the majority and clusters can also be identified. These types of plots are very interesting for the project managers of TNS NIPO Healthcare because they can help them to determine the nature of their competition. The same can also be said about the hierarchical clustering plots described in the previous section.

MDS rearranges cases in a manner so as to arrive at a configuration that best approximates the observed distances. It 'moves' cases around in the space defined by the specified number of dimensions, and checks how well the observed distances between them can be reproduced by the new configuration. A function minimization algorithm is used that evaluates different configurations with the goal of minimizing the lack-of-fit (equivalent to maximizing goodness-of-fit). This is an iterative process and to detect convergence a record of previous configurations is kept.

For a given configuration the approximation error in representing the dissimilarity between two cases *i and j is given by:*

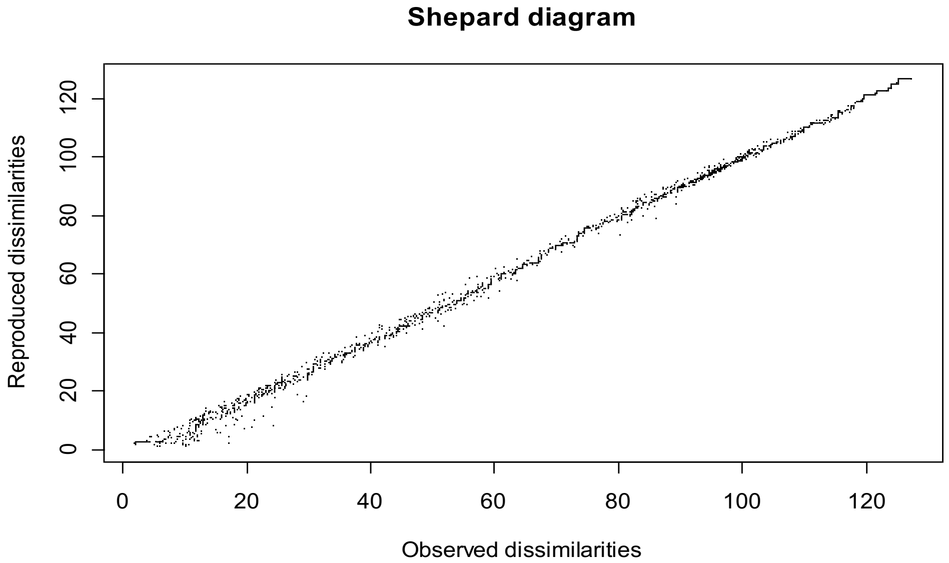
$$e_{ij} = | d_{ij} - \delta_{ij} |$$

with d_{ij} *= reproduced dissimilarities given a certain dimension k,*
and δ_{ij} *= observed dissimilarities*

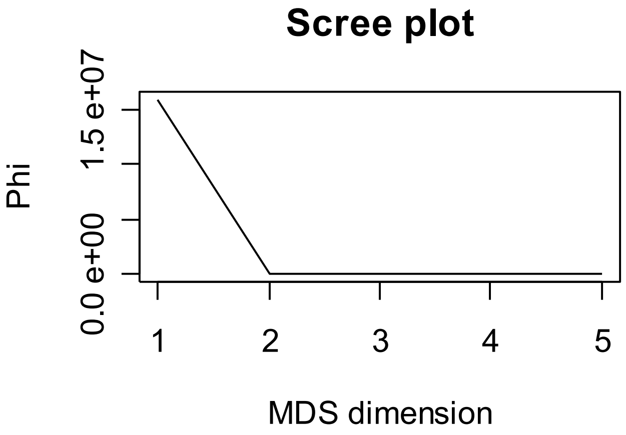
A commonly used measure to evaluate the accuracy of a particular configuration is the stress value Φ:

$$\Phi = \sum_{i \neq j} [d_{ij} - \delta_{ij}]^2$$

This is also referred to as the loss function and it accumulates the squared representation errors. The smaller the stress value, the better the representation of the observed dissimilarities by the reproduced dissimilarities. When minimizing stress one minimizes absolute error which means that errors in representing large and small dissimilarities are penalized equally. In certain cases this could mean that local details between objects are not preserved that well. A shepard diagram can be used to determine the quality of a representation. In this type of diagram, the reproduced dissimilarities for a particular number of dimensions are plotted against the observed dissimilarities. Here is an example:



This figure shows that for this particular configuration it is possible to draw a more or less straight line through the points. This means that the ‘observed’ dissimilarities are reproduced quite well. To determine the optimal number of dimensions to represent the data, one can also plot the stress value against different numbers of dimensions. This is called a scree plot and here is an example:



The stress value in this plot clearly levels out to zero at two dimensions. For the particular configurations this means that it is possible to give a good representation of the observed dissimilarities in a 2-dimensional space.

In MDS the same assumptions hold for distances as in the section about hierarchical clustering. An important note to make here is that since MDS only requires a matrix containing dissimilarities, it is not needed to have the original data matrix if one already has these dissimilarities. However if this is not the case, these dissimilarities between the objects in the data need to be calculated first.

2.3.2 Nonmetric multidimensional scaling

Nonmetric scaling only tries to fit the rank ordering of the 'observed' dissimilarities to the reproduced ones, whereas classical metric scaling attempts to fit the absolute values of the 'observed' dissimilarities to the reproduced ones. As a results the ratios between the 'observed' dissimilarities are reproduced too which doesn't have to be the case in nonmetric scaling. In nonmetric scaling the observed dissimilarities δ_{ij} are not used directly, they first undergo a monotone transformation $f(\delta_{ij})$:

$$\hat{\delta}_{ij} = f(\delta_{ij})$$

This function reproduces the rank ordering of dissimilarities between cases and the transformed dissimilarities are called disparities. There are two types of monotonicity:

Strong: $\delta_A < \delta_B \rightarrow \hat{\delta}_A < \hat{\delta}_B$
Weak: $\delta_A < \delta_B \rightarrow \hat{\delta}_A \leq \hat{\delta}_B$

If $\delta_A = \delta_B$ *there are two commonly used options:*

1. *no restriction on the relationship between* $\hat{\delta}_A$ and $\hat{\delta}_B$
2. $\hat{\delta}_A = \hat{\delta}_B$

However it must be said that out of these two options, the second though more restrictive one clearly makes more sense than the first.

The disparities can be modeled by a priori taking the rank orders instead of the absolute values as dissimilarities and then applying a linear transformation to these dissimilarities:

$$\hat{\delta}_{ij} = f(\delta_{ij}) = a\delta_{ij}$$

with a>0

The optimal *a* can be found analytically through differentiation. Alternated with an iterative improvement to the coordinates, this will lead to an optimal configuration with minimal stress. In nonmetric scaling one starts with a metric scaling configuration and to find an optimal configuration, the following iterative algorithm is used:

1. Determine the dimension of an initial coordinate matrix.
2. Calculate distances of configuration.
3. Search for the 'optimal' monotonic transformation of the dissimilarities/ Calculate disparities.
4. Search for the optimal coordinates.
5. Determine the goodness of fit by calculating the stress value of the current configuration and comparing it to the stress value of the previous configuration. If the difference is larger than a predefined threshold ε, then update the coordinates and go back to step 2. If the difference is smaller than ε, then stop here because the optimal configuration has been found.

The following (stress) measure is often used in nonmetric scaling to determine the accuracy of a particular configuration:

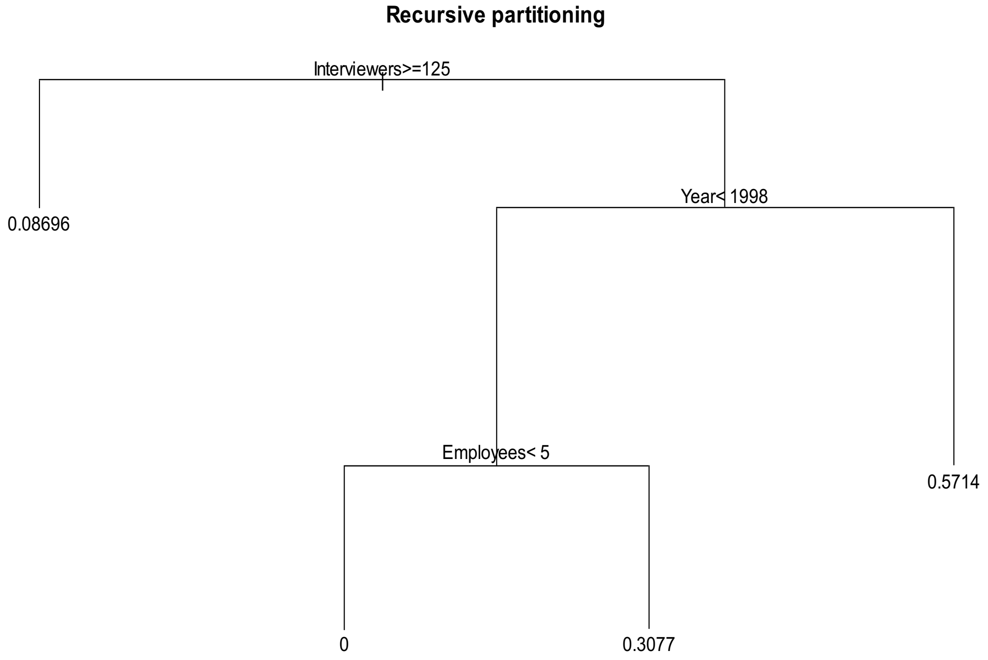
$$\Phi = \frac{\sum_{i \neq j} [d_{ij} - \hat{\delta}_{ij}]^2}{\sum_{i \neq j} d_{ij}^2}$$

with d_{ij} = reproduced dissimilarities given dimension k,
and $\hat{\delta}_{ij} = f(\delta_{ij})$ = disparities

2.4 Recursive partitioning

Recursive partitioning (Breiman, Friedman, Olshen, & Stone, 1984) is very interesting for the project managers of TNS NIPO Healthcare because it can be used to determine 'naturally' occurring patterns and segments in for instance competitor or respondent data. These patterns can be used to get an indication of the relationships between variables of interest and also to get an indication of what their possible values might be. The method can be used to deal with continuous and categorical variables. One attempts to determine the values of a continuous or categorical dependent variable from one or more continuous and/or categorical predictor variables.

In recursive partitioning tree structures are produced and it does not provide an explicit global linear model for prediction or interpretation. It seeks to split or bifurcate the data recursively at critical points of the explanatory variables. This is achieved by determining a set of if-then logical splitting conditions that permit accurate predictions or classifications. It is very useful in a Marketing context for segmentation purposes. Because of the tree-formed representation in most cases it is very easy to interpret the results. When analyzing business problems it is much easier to present a few simple if-then statements to management than for instance some elaborate regression equations. It is much easier now to comprehend why observations are classified or predicted in a particular manner. Here is an example:



In this figure it is easily determined what the splitting conditions are. At the bottom of each branch or terminal node, the likely or mean value of the response variable is printed.

The process of computing recursive partitioning trees can be characterized as involving three basic steps:

1. Selecting splits
2. Determining when to stop splitting
3. Testing on ‘new’ data

In recursive partitioning the aim is to split at each node to achieve the greatest level of predictive accuracy, measured by the node impurity. Node impurity measures the quality of the prediction or classification at a terminal node. With terminal nodes the final branches are meant. If all cases in each terminal node have an identical value, then the node impurity is minimal. When the response variable is categorical, the gini measure is commonly used to determine the impurity of a node. The gini measure is computed as the sum of products of all pairs of class proportions for classes present at a node:

$$G(t) = \sum_{j \neq i} C(i|j)\, p(j|t)\, p(i|t)$$

with C(i|j) = costs of misclassifying a category j case as category i,
C(i|j) = 1, if costs of misclassification are not specified,
p(j|t) = proportion of category j at node t,
and p(i|t) = proportion of category i at node t

Minimizing costs instead of just the proportion of misclassified cases is useful when some predictions that fail are more catastrophic than others or when some predictions that fail occur more frequently than others. In my research however, there were no costs of misclassification specified and these costs defaulted to ‘*1*’. The probabilities used in the formula are based on the class proportions at each node. The gini measure reaches a value of zero when all cases in a node belong to the same class. Now a short example will be given of the gini measure. Say there are two possible categories at a node *t*, ‘0’ and ‘1’. The number of occurrences of the first category is *1* and of the second *3*, so in total there are *4* cases at this node. Because there are more occurrences of the second category, the node gets assigned the value ‘1’. The impurity of this node *t* is:

$$C(0|1) = C(1|0) = 1$$
$$p(0|t) = 0.25$$
$$p(1|t) = 0.75$$

$$\begin{aligned} G(t) &= C(0|1)\, p(1|t)\, p(0|t) + C(1|0)\, p(0|t)\, p(1|t) \\ &= 1 \text{ x } 0.75 \text{ x } 0.25 + 1 \text{ x } 0.25 \text{ x } 0.75 = 0.375 \end{aligned}$$

The maximum value the node impurity can take in the case of a binary response variable is 0.5. This value is achieved when class sizes at a terminal node are equal. If the response variable is continuous the gini measure cannot be used. When dealing with this type of variables, a variant of the least squares method is often used to determine the impurity of a node:

$$R(t) = \frac{1}{N_w(t)} \sum_{i \in t} w_i \; (y_i - \; \overline{y}(t) \;)^2$$

with $N_w(t)$ = weighted number of cases in node ‘t’,
w_i = value of the weighting variable for case ‘i’,
w_i = 1, if no weighting is applied,
y_i = value of the response variable i,
and $\overline{y}(t)$ = weighted mean for node t

In my research no weighting was applied so the values for the weighting variables defaulted to *‘1’* and *$N_w(t)$* to *N(t),* which denotes the number of cases at node *‘t’*.

The second important issue in recursive partitioning is when to stop splitting. A data set containing *N* cases partitioned by *N - 1* splits can perfectly fit every single case in the data. If one splits a sufficient number of times one eventually will be able to perfectly predict or classify the original data. New data however will most likely be predicted or classified very poorly. Too many splits will incorporate information that cannot be predicted for the general population (random or noise variation). One way to control splitting is to allow splitting to continue until node impurity is minimal at all terminal nodes or all terminal nodes at least contain a specified minimum number or fraction of cases. The issue can also be dealt with by stopping with the generation of new split nodes when subsequent splits only result in very little overall improvement of the prediction. If for instance the addition of a certain split only raises the amount of correctly predicted or classified cases by say 0.5 %, then it makes little sense to add this split to the tree.

It is always recommendable to evaluate the quality of the prediction of a current tree in samples of observations that weren’t included in the construction of this tree. One can then ‘prune back’ the tree, to obtain a simpler tree that is equally accurate for predicting or classifying both ‘old’ and ‘new’ cases. Crossvalidation is very useful in this context. In this method one applies the tree computed from one set of cases (learning sample) to another completely independent set of observations (test sample). If most (or all) of the splits determined by the analysis of the learning sample are essentially based on ‘random noise’ then the prediction for the test sample will be very poor so the found tree would not be very good then. The test and learning samples can be formed by collecting two independent data sets but usually only one dataset is available. However if it is sufficiently large, one could reserve a randomly selected portion of the cases as test sample. The original dataset $Z=(x,y)$ containing N cases is as equally partitioned as possible into subsamples Z_1 and Z_2 of sizes N_1 and N_2 respectively. The cases in Z_1 are used as learning sample and those in Z_2 as test sample.

The analysis could also be repeated many times over using multiple randomly drawn samples from the same data. This is known as V-fold crossvalidation. This type of cross-validation is particularly useful when no test sample is available and the learning sample is too small to have the test sample taken from it. The subsamples should be as equal in size as possible. A tree of the specified size is computed V times, each time leaving out one of the subsamples from the computations Each subsample is used $(V - 1)$ times in the learning sample and V times as test sample. The original data $Z=(x,y)$ of size N is partitioned into V sub samples $Z_1, Z_2, ..., Z_V$ of sizes $N_1, N_2, ..., N_V$ respectively. There are V partitioning trees constructed from the learning sample $Z\text{-}Z_U$, *with* $U = 1,\ldots\ldots,V$. Each partitioning tree is used to predict or classify the V sub samples. The tree that on average produces the best predictions or highest classification rate is chosen.

In the next chapter analysis problems that were encountered and how was dealt with them will be discussed.

3 Analysis Problems

Missing values and outliers were the problems encountered concerning the quality and validity of the data about the respondents and competitors of TNS NIPO Healthcare. Here is an overview of this chapter:

- 3.1 Missing values
- 3.2 Outliers

3.1 Missing values

When collecting data by questionnaire respondents may sometimes be unwilling or unable to respond to certain questions. And the revenue for instance was only available for part of the competitors. This means that the collected data will sometimes be incomplete. Missing values pose a serious problem when constructing data analysis models because these models sometimes require complete data. In my research three measures to assess the amount of missing values were used:

$$\textit{fraction complete cases} = \frac{\mathrm{N_c}}{N}$$

$$\textit{fraction available values} = \frac{\sum_{\mathrm{j}=1}^{\mathrm{N}}\sum_{i=1}^{n}\Xi(x_{ij} \neq NA) + \sum_{i=1}^{N}\Xi(y_i \neq NA)}{N \times (n+1)}$$

$$\textit{fraction available response values} = \frac{\sum_{i=1}^{N}\Xi(\mathrm{y_i} \neq NA)}{\mathrm{N}}$$

with N = number of cases,
N_c = number of complete cases,
n = number of explanatory variables,
'NA' denoting a missing/not available value,

$$\textit{and } \Xi = \textit{an indicator function} \begin{cases} 1 \text{ if condition is satisfied,} \\ 0 \text{ if condition is not satisfied} \end{cases}$$

There are basically two ways of dealing with missing values, they are either left out/ ignored or they are substituted by 'plausible' values. In this section some techniques will be discussed for handling missing values by substituting them by 'plausible' values. This substitution is also referred to as imputation.

The regression models discussed in the previous chapter can be used to handle missing values by generating predictions for them. The 'best' model selected through a stepwise approach is very useful when dealing with missing values because it only contains the 'most influential' explanatory variables. That enables the researcher to discard the other explanatory variables which decreases the number of missing values one has to deal with. If all variables would have been taken into account, regression can usually only be used as a predictive tool for a very small number of cases because often only a few of these cases do not contain any missing values at all.

Because the missing values were scattered among the various variables in the data about the competitors of TNS NIPO Healthcare, several robust regression models based on different (sub)sets of explanatory variables from the 'best' model had to be constructed. For instance, for some companies the number of employees was known but the number of interviewers was not or vice versa. But for other companies the values of both of these variables were known or both of them were missing. So various regression models were constructed, with and without these variables. The models based on subsets of the 'best' model M_b, will be referred to as partial 'best' models M_{p_1}, M_{p_2}, etc. These models were used in succession to predict the (missing) values, starting with the 'best' model found through stepwise model selection. Then a next model is used to try to predict the values that the previous model was not able to. So the predictions from various regression models are combined to get one final prediction. That is why I call this method 'combined sequential prediction'. One could also first impute all the missing values in the explanatory variables and then generate predictions for the response based on this 'complete' data. Only one model is needed to generate predictions for the response then. However the drawback of this method is that these predictions are based on other predictions, so one in fact indirectly predicts the response from the observed data. An advantage of combined sequential prediction is that all the predictions are directly based on the observed data. A drawback of using combined sequential prediction is that if variables are strongly correlated, only one of them is included in the 'best' model. When evaluating partial best models only the variables from this 'best' model are taken into account. It sometimes might lead to better predictions if the partial best models also included variables that were left out of the best model. If these variables have a strong correlation to the variables in the best model, they also have predictive value for the response.

Multiple imputation (MI: Rubin, 1987) is another way of dealing with missing values. In MI the missing values are replaced by not one but multiple predictions. Each missing value is replaced by $m > 1$ 'plausible' values so *m* different 'complete' datasets are produced. The 'plausible' values are drawn from the estimated distribution of the missing values. The reason that multiple imputations are calculated is to determine the amount of variance in the predicted values so one gets a good notion of their accuracy. The uncertainty with which the missing values can be predicted from the observed values is reflected in the variance among the 'm' imputations. Regardless of the imputation method being used, one should always keep in mind that the imputed values are only estimates of the missing values and therefore always contain a certain level of uncertainty. If one would use the imputed datasets as explanatory variables to predict a certain response variable, one might also consider the variance between the *m* 'plausible' values for that response. Big or small variance in the *m* imputations for the explanatory variables does not necessarily have to be reflected in the response by a big or small variance.

Each imputed dataset must undergo the exact same statistical analyses and afterwards all the results are combined to produce overall estimates and standard errors. These overall statistics reflect the uncertainty in the missing data. The 'complete' datasets can be analyzed by using methods commonly used for complete data analysis such as those discussed in the previous section.

Rubin uses the terms Missing Completely at Random (MCAR), Missing at Random (MAR) and Not Missing at Random (NMAR) to describe the degree in which the pattern of missing data is random. In the remainder of this chapter these terms will be used when modeling the missing values. Now a predictive model will be described when dealing with continuous missing explanatory values. One has the following data: explanatory variables X (Nxn data matrix) and response Y (Nx1 data vector). The goal is then to predict response Y from explanatory variables X:

$X = (X_A, X_{NA})$

with X_A *= portion of X observed for all items,*
and X_{NA} *= portion of X containing missing values*

The notation $X_{A,j}$ and $X_{NA,j}$ is used to respectively represent the observed and missing portion of variable *j*. The notation $X_{i(A)}$ and $X_{i(NA)}$ is used to respectively represent the observed and missing portion of case *i*.

There are two main assumptions made:

1. The first main assumption that is made is that the missing values are Missing at Random (MAR). This however does not mean that the probability that the values are missing in a certain pattern *M*, is completely independent of what those values really were and of what the observed values are. If that would be the case, the missing values would be Missing Completely at Random (MCAR). MAR means that which values are missing does not depend on what their actual values were, given the observed data. The event wether or not an observation is missing, is conditionally independent of the unobserved values given the observed values. MAR is in fact a relaxation of MCAR in the sense that it is less restrictive.

   An *Nxn* matrix *M* is used to indicate which values are missing/ describe the pattern of missing data:

$$M_{ij} \begin{cases} 1 \text{ if } x_{ij} \text{ is available,} \\ 0 \text{ if } x_{ij} \text{ is not available} \end{cases}$$

   The columns of *M* are referred to as indicator variables. The joint probability distribution of the data generation process and the missing data mechanism is given by:

$$P_{\theta,\phi}(X, M) = P_{\theta}(X)\, P_{\phi}(M \mid X) = P_{\theta}(X_A, X_{NA})\, P_{\phi}(M \mid X_A, X_{NA})$$

   with θ *denoting the set of parameters for the distribution of the explanatory variables X (data generating process),*
   and ϕ *denoting the set of parameters for the distribution of the indicator variables (missing data mechanism)*

   Under the assumption of MCAR the probability that the values are missing in the pattern *M* is:

$$P_{\phi}(M \mid X) = P_{\phi}(M \mid X_A, X_{NA}) = P_{\phi}(M)$$

   This means that the probability that the values are missing in the pattern *M*, is independent of what those values really were but also of what the observed values are. But again, MCAR is a much stronger assumption than MAR and that is why only MAR is assumed for the missing values. Under the assumption of MAR the probability that the values are missing in a certain pattern *M* is independent of what those values really were:

$$P_{\phi}(M \mid X) = P_{\phi}(M \mid X_A, X_{NA}) = P_{\phi}(M \mid X_A)$$

This pattern may however be dependent of the observed values X_A. For instance, it could be that x_{ij} is always missing for certain values of x_{ik} provided that x_{ik} is always observed with $j \neq k$. A pattern of missing data specifically relevant for marketing research is that of monotone missing data. In the case of monotone missing data it is so that if x_{ij} is missing, so is x_{ik} with *(k>j)*. This is a common phenomenon when dealing with questionnaires. When respondents stop before the questionnaire is finished, the answers to the remaining questions will all be missing. This missing data pattern also occurs in longitudinal research. It often happens that participants drop out before the research period is completed so the remaining measurements will be missing then.

If the data isn't Missing (Completely) at Random, it is referred to as Not Missing at Random (NMAR).

2. The second main assumption that is made is that the continuous explanatory variables have a normal marginal distribution (which in practice however is not always the case):

$$\mathrm{N}(\mu_{\mathrm{i}}, \sigma_i^2)$$

with μ_j *= expectation of variable j,*
and σ_j^2 *= variance of variable j*

The simultaneous distribution is multinormal which also means that by definition the marginal distributions are normal:

$$\mathrm{N}(\mu, \Sigma)$$

with $\mu = (\mu_1, \mu_2, \ldots\ldots\ldots, \mu_n)$,

and Σ *= symmetric covariance matrix =* $\begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{pmatrix}$;

$\sigma_{ii} = \sigma_i^2$
$\sigma_{ij} = \sigma_{ji}$

When data are incomplete, the full probability model to describe the data is the joint probability model $P_{\theta,\phi}\ (X_{i(A)}, X_{i(NA)}, M_i)$. Since $X_{i(N(A)}$ is unknown $P_{\theta,\phi}(X_{i(A)}, M_i)$ is evaluated instead. By definition, the observed-data likelihood function is proportional to the marginal distribution of the joint distribution integrated over $X_{i(NA):}$

$$L(\theta,\phi \mid X_{i(A)}, M) \propto P_{\theta,\phi}(X_{i(A)}, M_i)$$

$$P_{\theta,\phi}(X_{i(A)}, M_i) = \int P_\theta(X_{i(A)}, X_{i(NA)}) P_\phi(M_i \mid X_{i(A)}, X_{i(NA)}) dX_{i(NA)}$$

If the data are MAR then:

$$P_\phi(M_i \mid X_{i(A)}, X_{i(NA)}) = P_\phi(M_i \mid X_{i(A)})$$
$$P_{\theta,\phi}(X_{i(A)}, M_i) = P_\phi(M_i \mid X_{i(A)}) \int P_\theta(X_{i(A)}, X_{i(NA)}) dX_{i(NA)} = P_\phi(M_i \mid X_{i(A)}) P_\theta(X_{i(A)})$$

So if the data are MAR then the likelihood can be factored. The data distribution $P_{\theta,\phi}(X_{i(A)}, M_i)$ can be replaced by the marginal data distribution $P_\theta(X_{i(A)})$ for inferences on θ without concerning $P_\phi(M_i \mid X_{i(A)})$ since θ and ϕ are distinct parameters. The observed-data likelihood function when ignoring the missing data mechanism is proportional to $P_\theta(X_{i(A)})$:

$$L(\theta \mid X_{i(A)}) \propto P_\theta(X_{i(A)})$$

Therefore the parameters ϕ of the missing data mechanism can be ignored for the purpose of estimating the parameters θ of the data generation process. This means that maximizing (over θ) of $L(\theta \mid X_{i(A)})$ is equivalent to maximizing $L(\theta, \phi \mid X_{i(A)}, M_i)$.

$$\theta \approx \hat{\theta} = \arg\max_{\theta} L(\theta, \phi \mid X_{i(A)}, M_i) = \arg\max_{\theta} L(\theta \mid X_{i(A)})$$

Often the loglikelihood function denoted by *l* is maximized because it is computationally easier.

The expectation maximization algorithm (EM: Dempster, Laird and Rubin, 1977) can be used for finding maximum-likelihood parameter estimates. EM finds the maximum-likelihood estimate of the parameters of an underlying distribution from a given data set containing missing values. This algorithm consists of two steps:

- E-step: Evaluation of the expectation
  In this step the expected value of the complete-data log-likelihood with respect to the missing data given the observed data and the distribution corresponding to the current parameter estimates θ_k is determined:

  $$Q(\theta \mid \theta_k) = \mathrm{E}[l(\theta \mid X_i, M_i) \mid X_i, \theta_k] = \mathrm{E}[l(\theta \mid X_{i(A)}, X_{i(NA)}, M_i) \mid X_{i(A)}, \theta_k]$$

- M-step: Maximization of the expectation computed in the E-step by finding the parameters that maximize *Q*:

  $$\theta_{k+1} = \arg\max_{\theta} Q(\theta \mid \theta_k)$$

Through iteration these two steps are repeated as many times as necessary. Each iteration increases the log-likelihood and the algorithm converges to a local maximum of the likelihood function. For a normal distribution it converges to the sample mean and variance. The found parameter estimates are used to draw the *m* imputations. If there were no missing values at all, this algorithm would converge immediately. In general it can be said that the more missing values, the more iterations are needed before convergence sets in.

3.2 Outliers

Outliers are points that deviate so much from the majority of the observations that they probably were generated by some other mechanism than the ‘regular’ data generating mechanism. They can for instance seriously bias the results of a regression analysis by “pulling” or “pushing” the regression line in a certain direction. This means that the obtained regression coefficients will be biased too. Certain competitors of TNS NIPO Healthcare for instance have an extremely high revenue in comparison to the majority of the competition. Often it is so that the exclusion of just a single such extreme outlier will give completely different results. This is so because the selection of the variables to be included in a standard linear regression is determined by looking at the residual sum of squares and not on the sum of the absolute residuals. Because the square values of the residuals are taken, this can easily lead to a very different selection of variables and estimated coefficients. So even if the sample size would be very large, outliers could still negatively influence the results a great deal.

Scatterplots of the variables often give a very good indication of the outliers in the dataset. But when there are more than 3 dimensions, the data cannot be visually represented like this anymore. Visualization techniques such as for instance the ones discussed in the previous chapter can be used to segment data but also to detect outliers. Hierarchical clustering can be used for this purpose regardless the number of variables. If the majority of the segments contain a lot of points and some segments only consist out of a single point, these points are quite possibly outliers.

There are three types of outliers (Rousseeuw, 1997):

- vertical outliers: points (x_i,y_i) whose x_i is not outlying but y_i is
- good leverage points: points (x_i,y_i) whose x_i is outlying but y_i is not
- bad leverage points: points (x_i,y_i) whose x_i and y_i are both outlying

One could also refer to the second type as horizontal outliers and the last type as complete outliers.

To determine the effects of outliers in regression, one could build models with and without these points and compare the results. Robust or resistant regression can be used to fit the model to the ‘regular’ points in the dataset. This means that the regression estimator is resistant to the influence of extreme outliers in the data. As stated in the previous chapter, classical least squares (LS) is used to estimate the model parameters in regression. Rousseeuw presents several robust methods for estimating the model parameters θ, two of these will be discussed now. One could for instance minimize a trimmed sum, such as in the least trimmed sum of squares (LTS) method:

$$\theta \approx \hat{\theta}_{LTS} = \arg\min_{\theta} \sum_{i=1}^{h} (r^2)_{i:N}$$

with $n \leq h \leq N$,
and $(r^2)_{1:N} \leq (r^2)_{2:N} \leq \ldots\ldots\ldots\ldots \leq (r^2)_{N-1:N} (r^2)_{N:N}$ *are the ordered squared residuals*

LTS is based on the subset of *h* cases out of *N* whose least squares fit has the smallest sum of squared residuals.

Another option is to minimize a certain quantile which is done in the least quantile of squares (LQS) method:

$$\theta \approx \hat{\theta}_{LQS} = \arg\min_{\theta} (r^2)_{h:N} = \arg\min_{\theta} |r|_{h:N}$$

$$\textit{with } n \leq h \leq N$$

This is a generalization of the least median of squares (LMS) method in which the median of the squared residuals is minimized. If N is an odd number and $h = [\frac{N}{2}] + 1$, LQS is equal to LMS.

In the next chapter the results from the analysis of some competitor and respondent data will be discussed.

4 Analysis Results

In the previous two chapters certain techniques for analyzing quantitative data have been discussed. This chapter illustrates the practical application of some of these techniques. To further analyze the collected competitor and respondent data, software specifically designed for data analysis was used. The software package used was R, a language and environment for statistical computing. Here is an overview:

- 4.1 Competitors
- 4.2 Respondents

In practice not only continuous response variables but also discrete ones were encountered. Continuous predictions were generated for these discrete variables. After the predictions for these discrete variables were generated, they were rounded to correspond with the observed discrete values. So they were in fact treated as discretized continuous variables. This method can be applied to variables that follow some sort of meaningful hierarchical ordering. With meaningful is meant that a higher/lower value means that there is more/less of something. This method can for instance be applied to binary and hierarchically ordered categorical variables such as age categories. But certain variables without this kindof ordering, such as city names, cannot be handled in this manner.

Now first the analysis of the competitor data will be discussed, followed by the analysis of the respondent data.

4.1 Competitors

In this chapter some results will be discussed from the analysis of a dataset about the competitors of TNS NIPO Healthcare. The following five competitor characteristics were included in the analysis:

- Specialized → (1: Specialized, 0: Not specialized)
- Year (Founded)
- (Number of) Employees
- (Number of) Interviewers
- (Total) Revenue

The competitors have the following general characteristics:

- The number of employees is always less than the number of interviewers.
- Agencies with a larger number of employees usually also have a higher revenue.
- Certain agencies are completely specialized in healthcare related marketing research while others are also active in other branches.
- The specialized agencies are usually smaller than the general ones in terms of number of employees, revenue etc.
- There are more general than specialized companies.

Because of the sensitive nature of a competitor analysis, individual data about existing agencies such as their names will not be given. Fictitious data was used that contained *N=46* agencies (also see appendix 1).

Not every company wants all of its information to be public so I had to look for ways to deal with these missing values. Before imputation, there was a great deal of missing values which is reflected in the following statistics:

fraction complete cases = 0.20

fraction available values = 0.80

fraction available response values = 0.63

Revenue was the response variable in the analysis. Regression imputation was used to deal with missing values.

Through stepwise model selection *Employees* and *Interviewers* were identified as 'most influential' explanatory variables. These two variables contain the number of employees and interviewers of each agency. Because both of these variables were not available for all agencies, combined sequential prediction was used. The 'best' model contained both these variables, the second 'best' only *Employees* and the third 'best' only *Interviewers*.

The predicted revenue is sometimes much lower than the 'observed' value such as is the case with agency '2'. This might imply that the 'observed' value is in fact incorrect and in reality the revenue is much lower. Such a high revenue could however also be caused by exceptional business processes. The predicted revenue could also be higher than the 'observed' one. This might imply that the 'observed' revenue is incorrect or that one can expect the revenue of this agency to grow considerably given the resources it has.

It could occur that for certain companies there is so little information available that it is not possible to generate a reliable prediction for them. That is the case when none of the explanatory variables found through stepwise model selection have a value for those cases. One could then of course take the average revenue as the best possible prediction given the observed data. This can be very valuable in certain situations which will be illustrated with an example. If one would like to get an estimate of the total revenue of an industry, one would usually take the sum of the revenues of the individual companies. But if certain values are missing, even after imputation, this can be a problem. By substituting the mean revenue (8 x 10^5) for these missing values, one can still get a reasonable estimate of the total revenue of a particular industry. For the revenue in the analyzed data however mean or average substitution was not needed because after the combined sequential prediction, all agencies had been assigned a value.

While collecting data about the firm's competitors a serious problem encountered was that of outliers. Some data were for instance outdated which increases the likeliness of outliers being present in the data. The robust models that have been used to deal with these problems were based on the least trimmed sum of squares (LTS) and least quantile of squares (LQS). For the analyzed data, the LTS and LQS predictions were nearly identical.

From a practical point of view, robust regression enables the user to detect irregularities in the data. The model determines the most likely values for the variables and by comparing them to the observed data, one can find values or cases that deviate from the other observations. If for instance the real revenue of a company is much higher than the predicted one (high residual value), this might imply that the high revenue was caused by incidental factors such as the sale of some offices. In such a case it is very likely that the expected revenue gives a far better approximation of the position of that company. It could also be that the real revenue is much lower than would be expected from the model. This could imply that the company is still growing and will reach the predicted revenue in the future. These examples illustrate the real practical value of robust analysis, not only are they resistant to the influence of outliers but they also enable the researcher to predict more likely values for these outliers.

To assess the overall quality of the prediction the residual sum of squares and the coefficient of determination were used. The following measure was also used:

$$\textit{average absolute residual} = \frac{\sum_{i=1}^{N} | r_i |}{N}$$

with r_i = residual corresponding to response i,
and N = number of cases

This gives the average absolute deviance of the observed response to the predicted values. The average residual was used because unlike the other two statistics, it is measured in the same unit as the response. The quality of the prediction is given by:

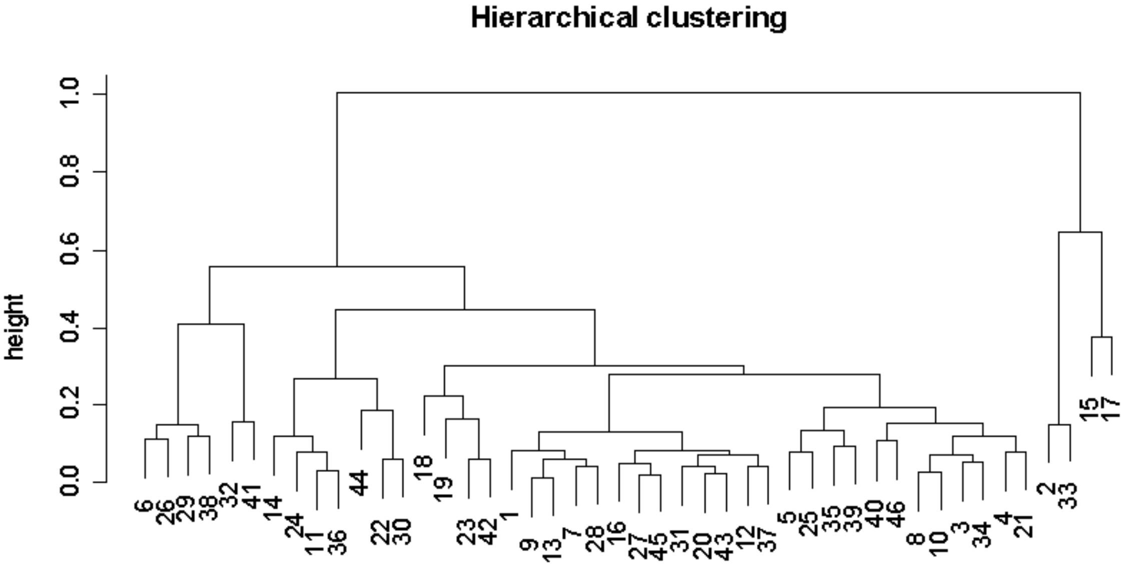
residual sum of squares = 10×10^{12}
average absolute residual = 3×10^{5}
coefficient of determination = 0.81

If the three biggest vertical outliers (‘2’, ‘17’ and ‘26’) are left out, the quality of the prediction is given by:

residual sum of squares = 1×10^{12}
average absolute residual = 1×10^{5}
coefficient of determination = 0.96

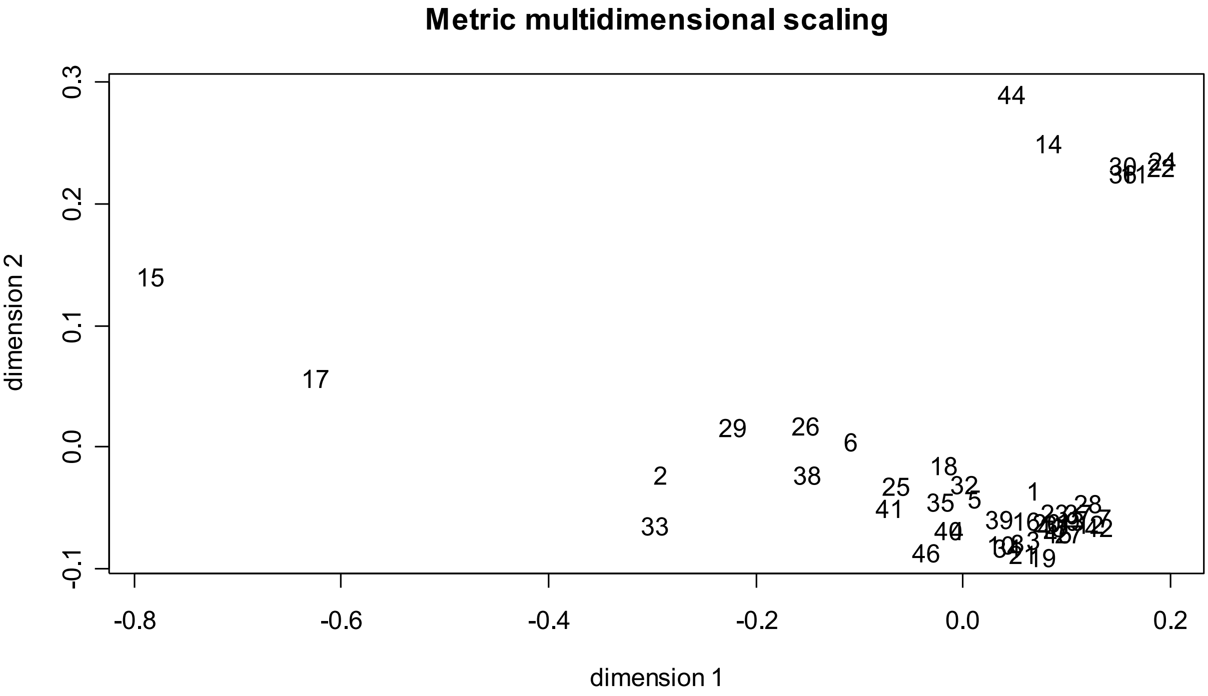
So just leaving out a few cases greatly reduces the error in prediction.

Now the techniques will be discussed that have been used to visualize and segment the data. Hierarchical clustering was used to determine which competitors are most similar to each other:



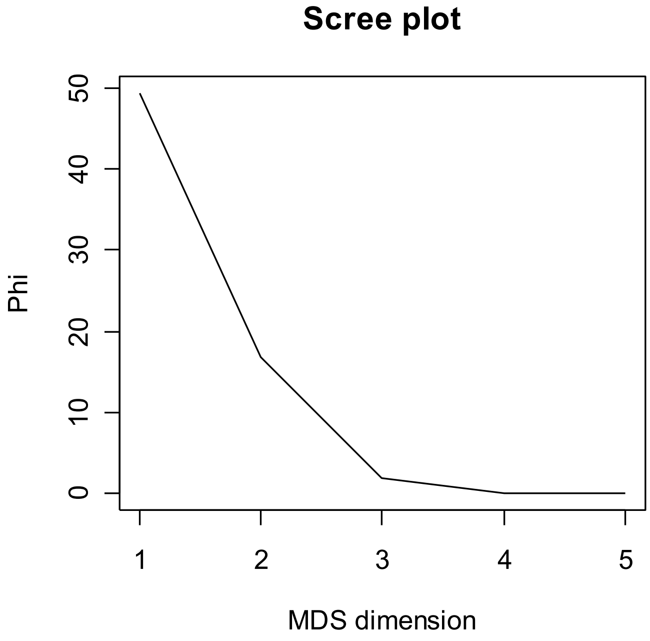
The agencies are represented by their respective index numbers in the data. This type of output facilitates the process of determining which agencies are most similar to each other and which companies are very different from the majority. To determine how similar two agencies are, their values for all the measured characteristics are compared. The differences in these characteristics are then taken together to allow a comparison between agencies. The level at which a split occurs in the above figure indicates the degree in which agencies are different from each other.

Classical (metric) multidimensional scaling was used to try to map the data in a lower dimension than the number of variables (*n=5*):



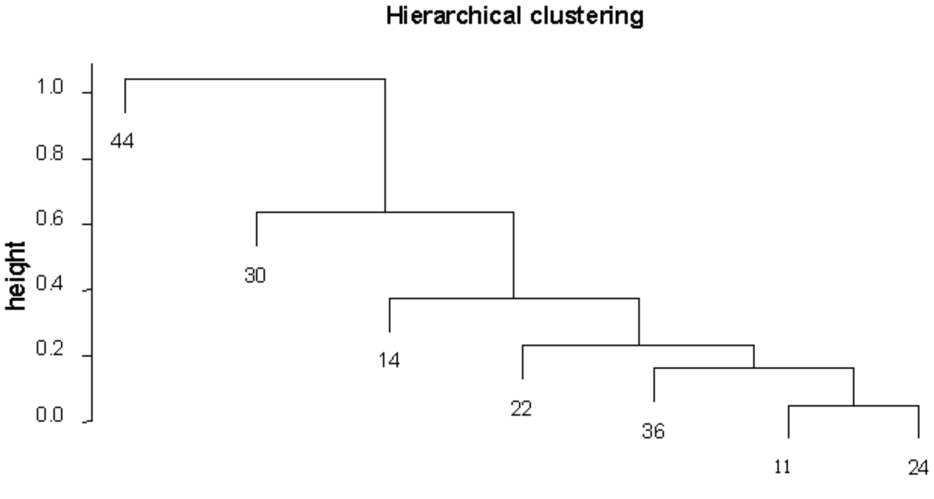
The agencies are again represented by their respective index numbers in the data. This type of output also facilitates the process of determining which agencies are most similar to each other and which companies are very different from the majority. If the own firm would also be part of the analysis, one could roughly establish one's competitive position to the other agencies. In the plot there is a clear separation between two segments of agencies. The agencies in the top right corner are specialized in healthcare and those in the bottom right corner are not. Most competitors are also active in other branches than healthcare which is reflected in this plot. It is not always clear what the meaning of the dimensions is. When trying to interpret the dimensions one could make scatterplots of the various possible combinations of the dimensions and original variables. So-called star plots (Chambers, Cleveland, Kleiner and Tukey, 1983) are also very useful for interpreting the dimensions (see appendix 2). For this particular data, dimension 2 gives an indication of whether an agency is or is not specialized in healthcare related marketing research. Dimension 1 gives an indication of the size of an agency in terms of its revenue. Larger companies are more to the left than smaller ones. In general a large or small revenue for a company also means that it has a large or small number of employees and interviewers. The same agencies that are clearly very different from the majority in the hierarchical clustering ('15' and '17') are also clearly very different in the multidimensional scaling. In both of these techniques the same agencies are clustered together which says something about how similar these agencies are. An important note to make is that even though agency '15' is quite different from the majority, based on the values for the explanatory variables however it is not a vertical outlier. By further analyzing the competitors by looking at their profiles, one could try to find reasons to explain the results of these analyses. These profiles also contain a lot of qualitative data which is much harder to model, but one now has a better idea of which companies to pay special attention too.

A scree plot was used to determine the optimal MDS dimension to represent the data:

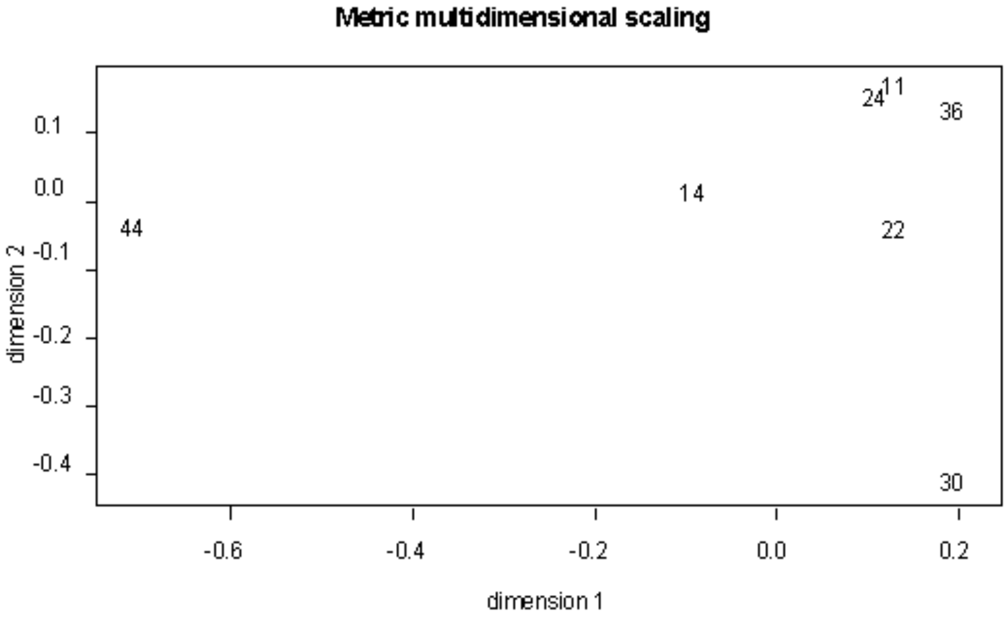


Because the stress value *Φ (Phi)* is already close to zero when two dimensions are used, it is quite likely that a two-dimensional plot gives a good representation. When the data were not standardized, the scree plot completely leveled out at two dimensions. This was probably caused by the fact that the variable revenue dominated the other variables as a result of the large unit of measurement.

Because the agencies that are completely specialized in healthcare related marketing research are of particular interest, a closer look was taken at them. First here is the hierarchical clustering of these agencies:

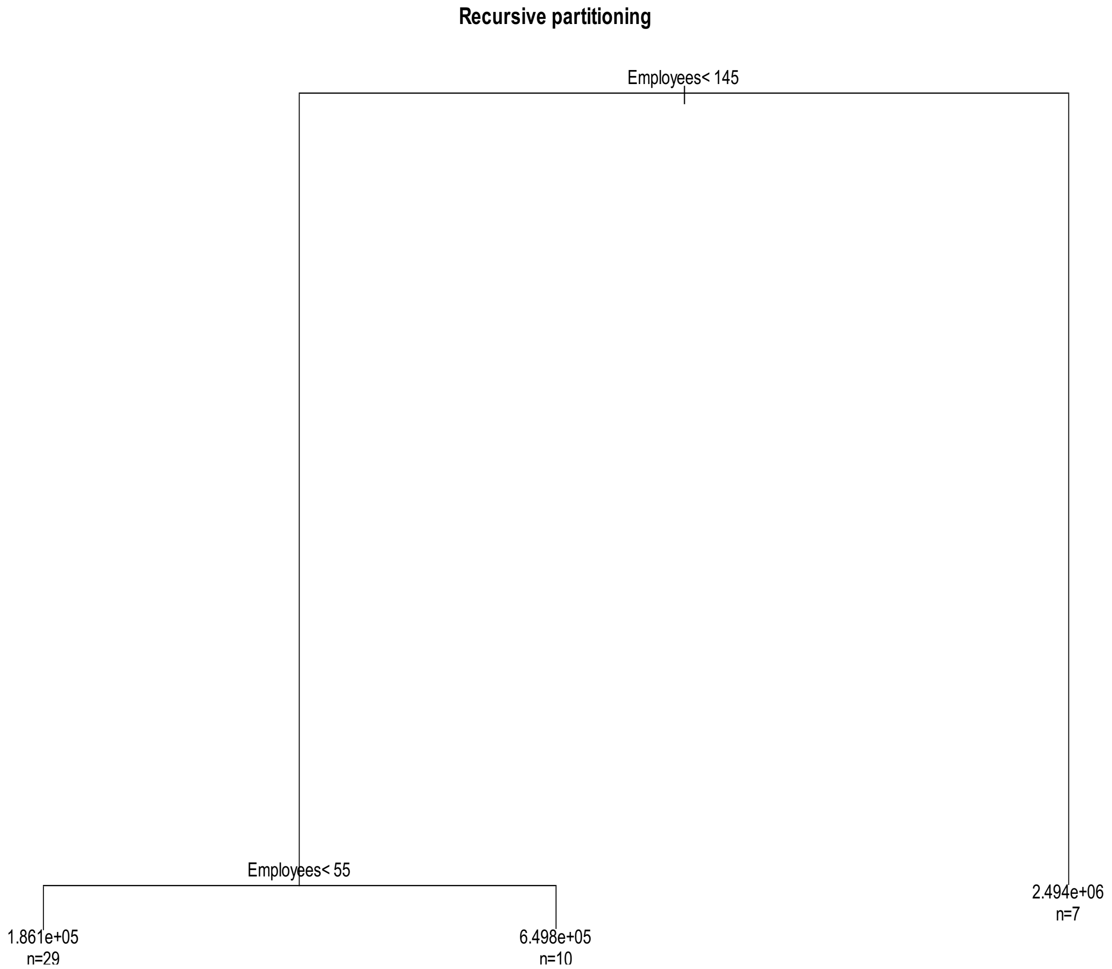


The following figure illustrates the multidimensional scaling of these agencies:



Agency '44' is by far 'the biggest outlier' in both plots and a look at the data shows that it is the largest specialized competitor in terms of revenue and number of employees. Just as in the previous MDS plot, Dimension 1 also gives an indication of the size of an agency now. Dimension 2 however now says something about the year in which the agency was founded. Agency '30' is the oldest specialized competitor and '24' the newest. The scree plot for this scaling completely leveled out at 2 dimensions.

Now the recursive partitioning of the competitors will be discussed. Recursive partitioning was applied to the data, to find splitting conditions to predict the value of a response variable of interest. These splitting conditions also make it possible to place each competitor in a particular segment. The response variable in the recursive partitioning was again 'Revenue':



This technique ‘searches’ for the general pattern in the data in the sense that it establishes ‘naturally’ occurring segments. The type of output makes it very easy to determine the splitting conditions, which allows one to quantify the differences between segments. At the end of each segment, the average value for the response variable is given along with the number of agencies. This figure can be used to get a quick indication of the value of the variable ‘Revenue’ for a particular competitor. In the following table the results of the recursive partitioning are summarized along with the impurity of each terminal node:

Branch (Left to Right)	Response value	Impurity	Number of agencies
1	186.091	3.061558e+11	29
2	649.785	7.945841e+11	10
3	2.493.878	7.197496e+12	7

4.2 Respondents

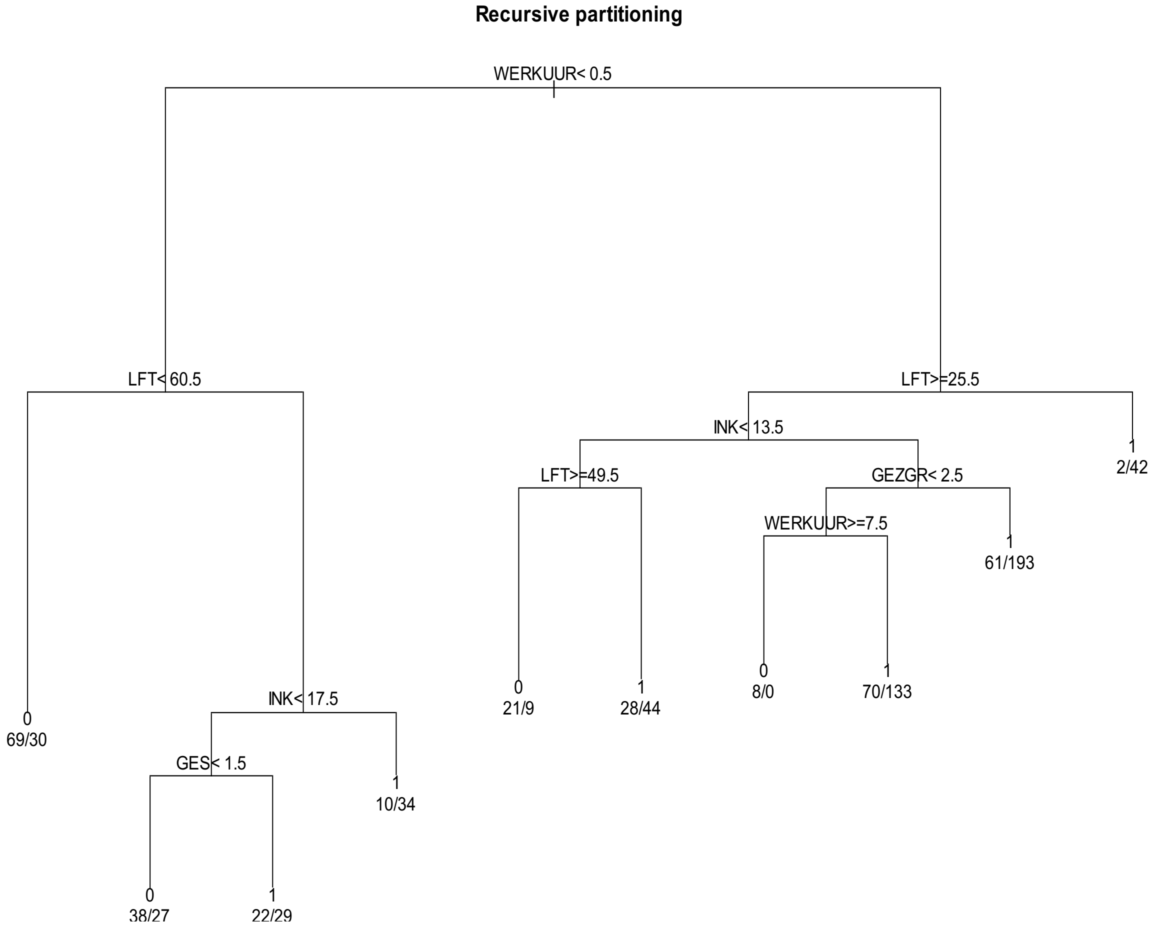
As part of my research data about the respondents, the people that form the market of the clients of TNS NIPO Healthcare, were analyzed too. These data were gathered through the numerous questionnaires over the years. The pool from which the respondents are chosen is known as the sampling frame. The sample plan refers to the technical process used to select units from this frame to be included in the sample. This in part determines how representative the sample is for the population. Often the telephone book is used as sampling frame but there are some disadvantages to doing this. People who do not have a telephone or have unlisted numbers are always excluded. It is also so that certain numbers listed in a telephone book are out of service. Such sampling biases can in part be overcome by using random digit dialing. This method allows for all possible telephone numbers to be selected by randomly choosing the digits. The sample size refers to how many elements of the population should be included in the sample. Usually it is so that the larger the sample the better, because it will be more representative of the population then (if there is also a good sample plan).

Sampling errors are made when choosing a sample due to the fact that the sample size is less than the size of the population being studied. A larger sample size leads to a smaller sampling error but also to higher costs. There are also so-called non-sampling errors such as selecting the wrong group of people to interview. For example, the objective is to study the behavior of internet users but the people selected never use the internet. An interviewer can also intentionally cause errors by introducing some bias which will lead the respondent to provide certain answers. Errors can also be caused by a lack of understanding of the questions by the respondent and/or interviewer. Other non-sampling errors are faulty coding, untruthful responses, respondent and/or interviewer fatigue etc.

Below the *n=8* characteristics that were considered are given followed first by their description in Dutch and then in English:

- GES: Geslacht (Gender) → Binary (1: Man, 2: Woman)
- LFT: Leeftijd (Age) → Ratio (18, …, 98)
- WERKUUR: Aantal uur per week werkzaam (Total number of working hours per week) → Ordinal (0, …, 9)
- VOLTOOID: Opleiding met een diploma voltooid (Highest completed education) → Ordinal (1, …, 4)
- GEMGR: Gemeentegrootte (Community size) → Ordinal (1, …, 5)
- GEZGR: Gezinsgrootte (Family size) → Ratio (1, …, 24)
- INK: Bruto jaarinkomen van het huishouden (Total annual household income) → Ordinal (0, …, 29)
- ZIEK: Chronische ziekte (Has chronic illness?) → Binary (1: Sick, 0: Not Sick)

The response variable was ‘ZIEK’. This binary variable indicates if a person has a chronic illness or not which is very interesting for an agency specialized in healthcare related marketing research. For all the variables the range of values they can assume has a meaningful hierarchical ordering. The following figure illustrates the results of the recursive partitioning of the data which contained *N=1000* respondents:



The figure shows how each characteristic is different within the two segments listed below that characteristic. So this figure in fact illustrates which patterns ‘naturally’ occur in the data. By looking at the splitting conditions one can determine segments with their specific characteristics. This is preferable to choosing segments beforehand based on certain vague assumptions about a population because prior knowledge of the data is not needed. Predefined segments often contain a lot of unscientific assumptions about the data which really do not necessarily have to be true. One can also determine how many people of each class there are within each segment which allows one to assess the impurity at each branch. From a financial point of view, it often is not profitable to target very small segments.

In the previous section it was discussed how the revenue of competitors was predicted from other variables. One could of course also apply this technique to the respondents to for instance predict their income based on certain other variables such as age, number of cars, level of education etc. If for a number of respondents all the variables one is interested in are known, one could then construct a predictive model based on these variables. This model could now be used to predict unknown characteristics such as for instance the income of other respondents. Instead of performing new marketing research every time, one could use the data already available to get some insight.

In the next and final chapter my overall findings will be presented.

5 Conclusion

The external analysis of TNS NIPO Healthcare consisted of three components:

- Competitors
- Clients
- Developments & Trends in the market

The main goal of the research described in this paper was to illustrate how to analyze large amounts of business data to get strategically relevant information. The data analyzed contain missing values and outliers which had to be handled with care. Outliers can have a very negative influence on regression models which would reduce the reliability of their predictions. That is why several robust methods of regression were used. These regression methods were also used to generate 'plausible' estimates to replace the missing values. Now analysis methods could be used that require complete data: hierarchical clustering, multidimensional scaling and recursive partitioning. The main strength that these three techniques have in common in a marketing context is that they segment the data. Hierarchical clustering is very useful for a qualitative segmentation of the data. Recursive partitioning can be applied for a quantitative segmentation of the data in terms of specific variable values. Multidimensional scaling does a little of both because one can qualify the proximity between individual cases but the dimensions sometimes also clearly quantify this proximity. Sometimes it is possible to interpret the dimensions in terms of specific variables.

In recursive partitioning the segmentation is based on a response variable which is not so in hierarchical clustering and MDS. In both hierarchical clustering and MDS, the Euclidean distance is often used to measure the dissimilarities between cases. That is why the output of these techniques often also leads to similar conclusions. Cases that are grouped together in hierarchical clustering are often also close to each other in MDS and outliers in hierarchical clustering are often also outliers in MDS. This was also so for the competitor data that were analyzed. However it must be noted that with MDS it is not always possible to represent the data in a small number of dimensions. The larger the number of variables the more likely a good configuration cannot be found to represent the data in a small number of dimensions.

Data about companies such as competitors and clients and data about respondents can be analyzed by using data analysis tools. Valuable strategic information can be extracted from these data which can support project managers in their strategic decision making process. By further examining the competitors by looking at their profiles which also contain qualitative data, the project managers of TNS NIPO Healthcare could try to find reasons to explain the results of these analyses. This information allows the project managers to better service their clients and to determine and strengthen their market position relative to their competitors.

The practical added value of analysis tools is that they generate predictions and segment cases in the data. Segments do not have to be predefined based on certain 'vague' assumptions, but can be determined based on quantitative characteristics extracted from the data. The analysis has for instance shown that the number of employees and interviewers of competing agencies are good predictors for their respective revenues. Segments and extremely strong agencies could also be identified. The classification of respondents concerning the occurrence of chronic illnesses was illustrated by using recursive partitioning. This allows one to better segment and target these respondents because the 'best' predictors for chronic illnesses were identified.

As a very important concluding remark it needs to be said that even though the discussed analysis techniques can provide valuable information, this information should not be considered to be absolute facts. There is always a level of uncertainty in predictions and segmentations. It is usually never so that one can model all the variables that would potentially be of interest. As a result the models will always be simplified versions of reality. The outcomes of the analyses should be used to support the decision process or to complement other analyses. Decisions should never solely be based on these outcomes, they always have to be investigated further.

Sources

Business

- David A. Aaker, “Developing Business Strategies” ,Fifth Edition, John Wiley & Sons, (1998)
- David A. Aaker, “Strategic Market Management” ,Fifth Edition, John Wiley & Sons, (1998)
- Michael E. Porter, “Competitive Strategy: Techniques for Analyzing Industries and Competitors”
- Michael E. Porter, “Competitive Advantage: Creating and Sustaining Superior Performance”
- Kees Westerkamp, “Een marketingplan in twaalf stappen” (article)
- Philip Kotler and Gary Armstrong, “Principles of Marketing”
- Solomon, Bamossy and Askegaard, “Consumer Behaviour; A European Perspective”, Harlow: Pearson Education (2nd edition) (2002)
- Alvin C. Bush and F. Ronald, “Marketing Research”

Mathematics

- A. Slotboom, “Statistiek in woorden”, third edition, Wolters-Noordhoff Groningen, (2001)
- John Chambers, William Cleveland, Beat Kleiner and Paul Tukey, “Graphical Methods for Data Analysis”, Wadsworth, (1983)
- P. J. Rousseeuw and A. M. Leroy, “Robust Regression and Outlier Detection”, Wiley, (1987)
- P.J. Rousseeuw and M. Hubert, “Recent developments in PROGRESS, In L1-Statistical Procedures and Related Topics” (article), (1997)
- P.J. Rousseeuw, Stefan van Aelst and Katrien van Driessen, “Robust Multivariate Regression” (article), (2000)
- Mark Huisman, “Simple and Effective Methods to Treat Missing Item Responses” (article)
- Mark Huisman and Johannes van der Zouwen, “Item Nonresponse in Scale Data from Surveys: Types, Determinants and Measures” (article)
- A.P. Dempster, N.M. Laird and D.B. Rubin. “Maximum-likelihood from incomplete data via the EM algorithm”, J. Royal Statistical Society Series. (1977)
- R.J. Little and D.B. Rubin “Statistical Analysis with Missing data” ,Wiley, New York, (1987)
- D.B. Rubin. “Multiple Imputation for Nonresponse in Surveys”, Wiley, New York. (1987)
- Zoubin Ghahramani and Michael I. Jordan, “Learning from incomplete data” (article) (1994)
- Cox, T. F. and Cox, M. A. A. “Multidimensional Scaling”, London: Chapman and Hall, (1994)
- Breiman, Friedman, Olshen, and Stone. “Classification and Regression Trees”, Wadsworth, (1984)
- A.K. Jain, M.N. Murty and P.J. Flynn, ”Data Clustering: A Review” (article), ACM Computing Surveys, (1999)

Informatics

- Michael J. Crawley, Wiley, ”Statistical Computing: An Introduction to Data Analysis using S-Plus”, (2002)
- Andreas Krause and Melvin Olson, “The Basics of S and S-PLUS”, (Third edition), Springer-Verlag New York, (2002)
- W. N. Venables, D. M. Smith and the R Development Core Team, “An Introduction to R, A Programming Environment for Data Analysis and Graphics”, Version 1.8.1, (2003)
- Ian H. Witten and Eibe Frank, “Data Mining”, Morgan Kaufmann Publishers
- Paolo Giudici, “Applied Data Mining”,John Wiley & Sons, Ltd, (2003)
- Laudon and Laudon, “Management Information Systems”, Seventh Edition

Business Mathematics and Informatics

- Maikel Groenewoud, “Marketing: Predicting Success” (paper), (2004)

Appendices

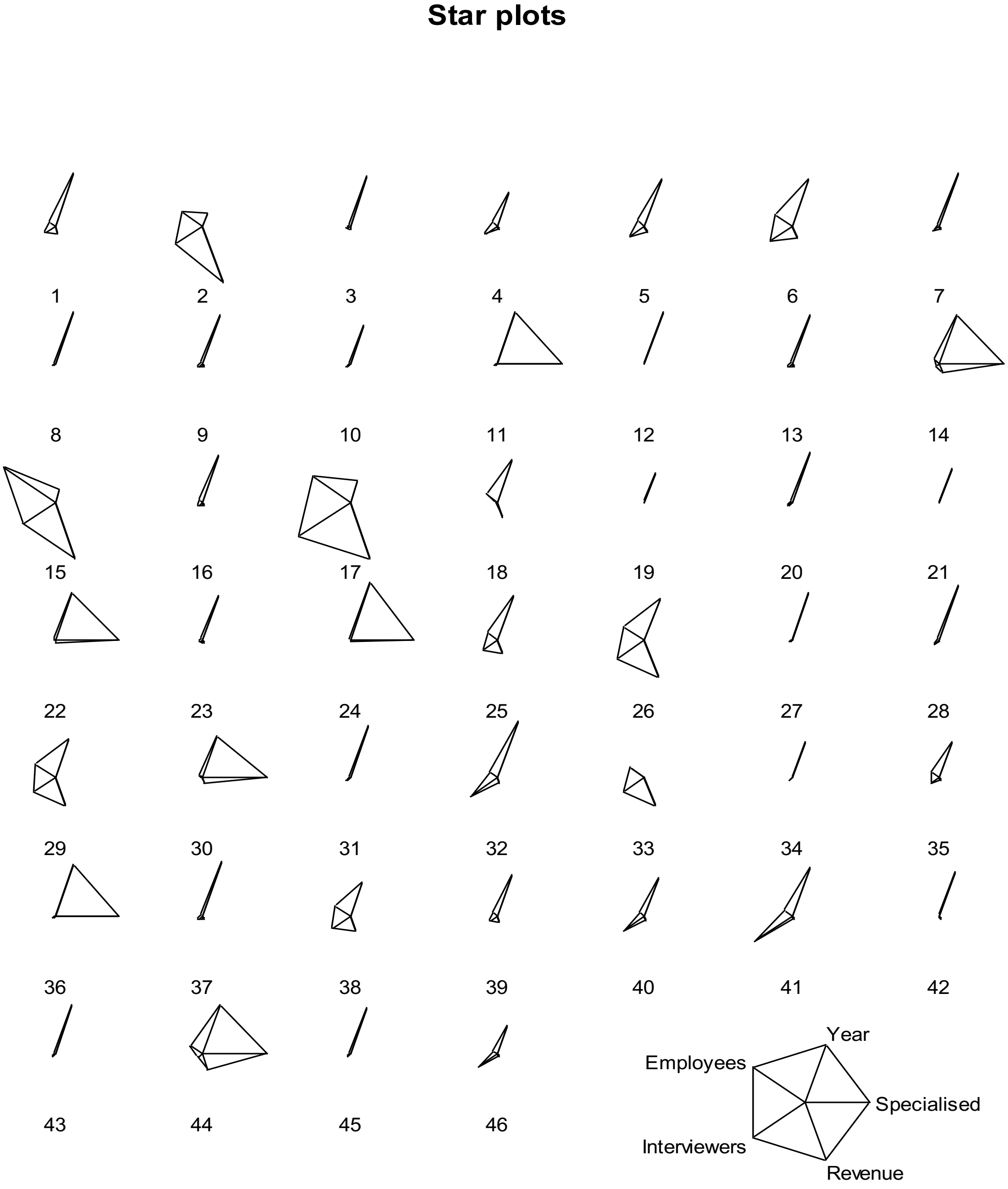
1 Competitor Data

This appendix contains the fictitious competitor data that were analyzed along with the predicted revenue.

Agency	Specialized	Year	Employees	Interviewers	Observed revenue	Predicted revenue
1	0	1996	60	NA	NA	4.9E+05
2	0	1981	200	NA	3.6E+06	1.3E+06
3	0	NA	8	NA	2.5E+05	2.0E+05
4	0	1989	50	1000	2.9E+05	3.8E+05
5	0	1994	70	1200	NA	5.0E+05
6	0	1994	160	NA	8.6E+05	1.1E+06
7	0	1996	16	600	2.5E+05	1.9E+05
8	0	NA	4	NA	7.1E+04	1.7E+05
9	0	1995	16	NA	2.5E+05	2.4E+05
10	0	1991	8	NA	NA	2.0E+05
11	1	1996	2	NA	7.1E+04	1.6E+05
12	0	1996	4	NA	7.1E+04	1.7E+05
13	0	1995	16	NA	2.5E+05	2.4E+05
14	1	1995	60	NA	NA	4.9E+05
15	0	1981	500	2400	3.6E+06	3.5E+06
16	0	1994	26	NA	2.5E+05	3.0E+05
17	0	1984	360	3800	3.6E+06	2.2E+06
18	0	1992	110	200	1.0E+06	9.7E+05
19	0	1987	6	100	NA	2.0E+05
20	0	1995	16	NA	1.1E+05	2.4E+05
21	0	1989	2	NA	7.1E+04	1.6E+05
22	1	1994	16	200	NA	2.6E+05
23	0	1993	NA	300	NA	2.7E+05
24	1	1998	6	NA	NA	1.8E+05
25	0	NA	100	NA	NA	7.3E+05
26	0	1992	200	NA	2.3E+06	1.3E+06
27	0	1994	2	NA	7.1E+04	1.6E+05
28	0	1997	28	NA	1.1E+05	3.1E+05
29	0	1991	200	1600	1.8E+06	1.4E+06
30	1	1992	30	120	NA	3.8E+05
31	0	1996	8	NA	NA	2.0E+05
32	0	1998	80	2000	4.3E+05	4.4E+05
33	0	1976	150	NA	1.8E+06	1.0E+06
34	0	1990	8	NA	7.1E+04	2.0E+05
35	0	1990	80	680	2.5E+05	6.6E+05
36	1	1996	2	NA	7.1E+04	1.6E+05
37	0	1997	16	NA	2.5E+05	2.4E+05
38	0	1989	140	1400	NA	9.9E+05
39	0	1992	42	NA	NA	3.9E+05
40	0	1991	50	1600	3.6E+05	2.8E+05
41	0	1995	80	2800	NA	3.0E+05
42	0	1993	12	120	NA	2.4E+05
43	0	1995	8	NA	7.1E+04	2.0E+05
44	1	1995	120	400	NA	1.0E+06
45	0	1994	4	NA	7.1E+04	1.7E+05
46	0	1987	40	1440	NA	2.3E+05

2 Competitor Star Plots

This appendix contains the star plots of the fictitious competitor data that were analyzed. The star plot is a method of displaying multivariate data. Each star represents a single observation. Star plots are used to examine the relative values for a single data point. This allows one to find dominant variables, locate clusters of similar points and detect outliers. The size in the star plot of a variable is proportional to the magnitude of the variable for the data point relative to the maximum magnitude of the variable across all data points. Here are the competitor star plots:


Star plots
Year
Employees
Specialised
Interviewers
Revenue