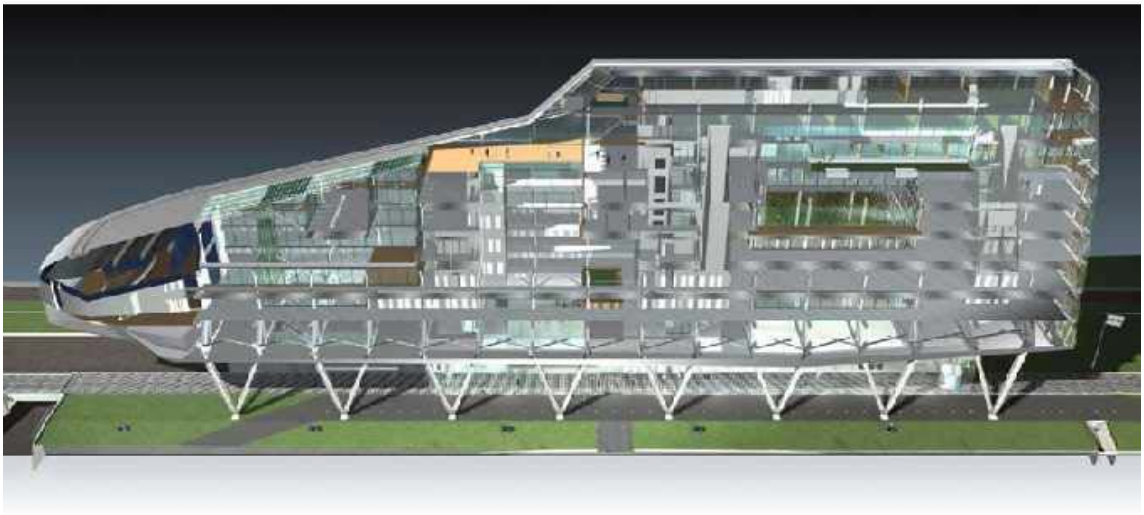
Risk assessment of IT-enabled Business Investments

Public Document



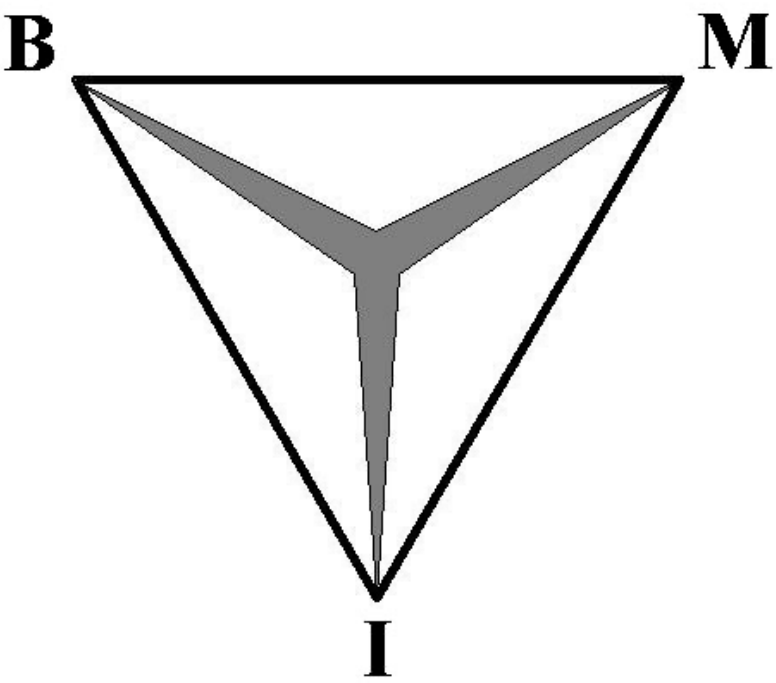
Master Thesis of Joeri van Hoeve

June 2004

ING

Risk assessment of IT-enabled Business Investments

"In search of quantitative risk prediction models"



Joeri van Hoeve

Master Thesis

Vrije Universiteit
Faculteit Exacte Wetenschappen
Business Mathematics and Informatics (BMI)
De Boelelaan 1081
1081 HV Amsterdam

Company:

ING Group / Corporate IT / IT Performance and Investment Management
Atrium Tower – Strawinskylaan 2631
Amsterdam

June 2004

Distribution List

ING:

Wilmar Hassoldt
Steven Raekelboom
John Spangenberg
Michael Vincent

Vrije Universiteit:

Bert Kersten
Rob Peters
Chris Verhoef

ING

Preface

The final part of the study program of Business Mathematics and Informatics (BMI) at the VU in Amsterdam consists of a working period. The purpose of such a working period is to gain experience in a business area and to solve a real business problem by using the knowledge acquired over the years. The problem needs to contain economical, mathematical and IT aspects.

This master thesis describes my master project within Corporate IT (CIT), a staff department of the de ING Group in Amsterdam. I have been working within the sub-group that is responsible for IT performance and investment management (ITPM). Their main activities are the collection and analysis of IT data and subsequently deliver management report to the higher management. These main activities take place from November till April; in the remaining months, the data collection process is assessed and improved. Furthermore research for new metrics is conducted in this period.

My project started in May and aimed at developing quantitative models to assess risk of IT development projects. It includes IT aspects, statistical data analysis and economical analysis of the models. The research was a challenging, but also *long and winding* road due to the unavailability of data, the closure of the department during August and September and a little writing block of the author in the closing stage.

I would like to thank John Spangenberg, my manager in corporate IT, who gave me the opportunity to gain experience in the field and also showed a lot of confidence in me during my internship. I should also offer my gratitude to my supervisor at the VU, Chris Verhoef. His article on IT portfolio management triggered my attention to this research field and his expertise and guidance helped me a lot in writing this thesis.
I would also show my special appreciation to Rob Peters, who showed me the way after a difficult start of this project and provided me with a lot of useful feedback during our several meetings. Last but not least I would like to thank Bert Kersten for reviewing my thesis and his useful comments during my end presentation.

I have enjoyed working at ITPM and would also like to thank my colleagues for a valuable working experience. I would especially like to thank Steven Raekelboom for sharing his experience on predictive modelling and his help on the final presentation. Another word of appreciation is to Wilmar Hassoldt for his help on the IT audit issues in this thesis.

Amsterdam 28 June 2004

Joeri van Hoeve

Executive Summary

This master thesis describes the research towards an Information Technology (IT) project risk assessment within ING Corporate IT (CIT). The goal of this research is the development of a formal and objective methodology that can be used as a quantitative and objective risk prediction tool of starting IT projects. CIT collects the top five of completed IT-enabled business investments from business units (BU) throughout ING. These investments are considered as business projects for which at least 25% of the budget is spent on IT. The actual performance of a project is measured by three high-level success criteria. Is the project delivered on time, on budget and with the desired business functionality? The research focuses on the risks on budget overrun, time overrun and less delivered functionality separately.

Our main conclusion is that logistic regression is the most appropriate modelling technique with respect to the collected data. Logistic regression is a frequently used model in medical studies. The main merit of logistic regression in these medical studies was the prediction of a certain risk by using a simple and straightforward formula of risk drivers, which enabled a clinical interpretation of these risk drivers. Risk drivers are risk factors that can be influenced before or in the early stages of a project. We developed logistic regression models for all three risks and assessed the quality of these models. The relatively large response errors (difference between predicted and observed risks) point out that our models do not provide exact risk probabilities for individual projects. We therefore use our models as a classification technique; the projects are classified into risky projects and not risky projects at the hand of the predicted risk probabilities. Another important model quality issue is the statistical significance of the risk drivers in the logistic regression equation. The best regression equation consists of uncorrelated risk drivers with a significant positive or negative impact.
The budget and functionality models show much better classification performances than the duration model. The budget model has the most stable regression equation with also the least correlations between the risk drivers. This budget model is thus considered as the best logistic model and we assume that the main focus in project management within ING is on meeting the original budget of the project.
The most important risk driver is the development department size (DDS), which increases the budget risk. Our model also indicates that good project management decreases the risk on budget overrun. Another finding was the relation between the budget risk of projects and the various CMM levels.

This master thesis introduces a formal methodology that enables us to develop predictive models for our risky projects. Although we only had a small amount of available projects and risk drivers, we have developed a logistic model that predicts the budget risk and that is useful as project selection tool in the IT audit department. We notice that this audit tool is only valid for a general group of projects. This model is for example not useful for a set of only projects from EC Americas. We conclude this summary with our recommendations to improve the current logistic models. The current data collection process should be expanded. We can first of all improve our logistic models by collecting data on more projects and on a more frequent base. Secondly we should expand our set of risk drivers with specific IT–enabled project characteristics, such as the size of the IT component of a project or the amount of staff used for a project. We expect that more collected data leads to models that are up-to-date and valid for specific groups of projects as EC Europe projects. We will then also be capable to develop good duration and functionality models.

Table of Contents

1 Introduction

Within the ING Group IT projects are no longer viewed as a cost centre, but as an investment centre that drives value creation. ING has paid special attention to develop metrics to follow-up IT projects and has come to view IT projects as IT-enabled business investments. IT-enabled business investments are business projects for which at least 25% of the project budget is spent on IT.

Financial transparency and risk/return metrics of these projects are essential in order to make sound decisions about these IT projects. The decision process on IT projects and proposals is currently supported by qualitative risk assessments. The wish of ING is to obtain more objective quantitative risk assessment methods.

The goals of this master project are to investigate the risk impact of IT project features and to develop a *predictive* early warning system for high-risk projects. Project failures are defined by three main project *success* criteria, e.g. project is within time, within budget and delivered 95% functionality.

The benefits for ING will be two-fold. First, critical insights are gained into their current risky IT projects. Secondly, the predictive model is a more objective manner of risk assessment than the deployed qualitative models and provides extra information to the decision process on IT project proposals.

Sections 1.1 and 1.2 describe the ING Group and its IT governance in general to place the research in a business context. The research is conducted within a sub-group *IT Performance and Investment Management* (ITPM) of the staff department Corporate IT. In Section 1.3 the work of this group will be summarised. In the last section we will present the outlook of this master thesis.

1.1 ING Group

ING Group is a global financial services institution of Dutch origin offering banking, insurance and asset management to 60 million private, corporate and institutional clients worldwide. It is a multi-product, multi-distribution company, approaching the customer through his or her channel of choice. ING group is very much a multi-brand company as well. So much so that ING companies realized only a minority of our revenues. That is changing rapidly, however. A lot has been going on to build the global awareness of the ING brand. Well-known local brands as Mercantile Mutual, Reliastar, Seguros Comercial Americas, Bank Slaski, BHF and BBL have been or are being replaced by the ING lion brand.

ING employs over 112,000 people and 70% of its stock is held outside the Netherlands. In today's depressed financial markets it has a current market capitalization of 36 billion euros. Total assets amount to over 700 billion Euros. The asset management business has 450 billion euros of assets under management.

By all measures ING is a large global and diverse business, which has grown very significantly in recent years through a combination of autonomous growth and targeted acquisitions. Like all global financial services organizations ING is totally dependent on IT, not just to support and enhance the business, but also increasingly to enable it. Without IT the ING Group has no business.

1.2 IT governance

The IT governance structure meshes with the overall corporate governance structure of ING. This IT structure aims at ensuring the strategic alignment of IT with the business. This structure is meant not only to improve the quality of the IT functions but also to speed up decision-making. The ING IT governance structure is depicted in Figure 1.1 and is necessary for the executive company board of ING in their quest for answers to important IT-related questions; e.g. How often do IT projects fail to deliver what they promised? How does IT add value to the business?

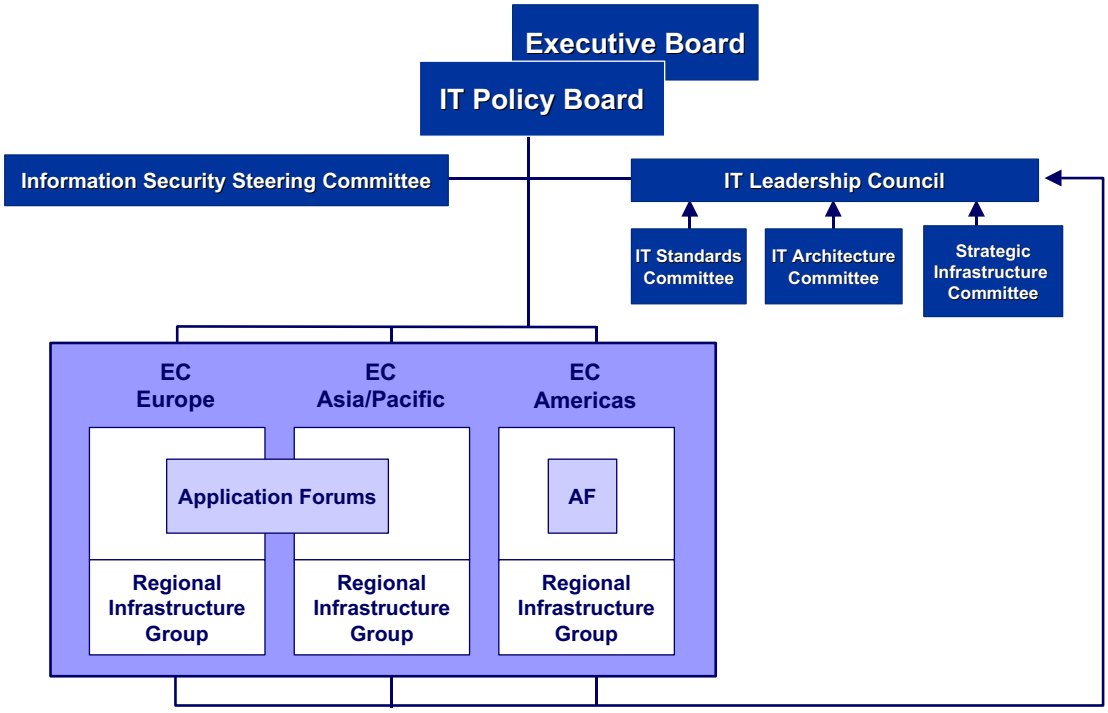


Figure 1.1: global ING IT governance model

The *IT policy board* is responsible for the global IT procedures. This board consists of three executive board members and the Operations / IT (OPS/IT) portfolio keepers of the three *Executive Centers* (Europe, Asia Pacific and Americas). ING defines Operations as the *management of the process itself*. Operations involve all activities after the sales to the client until final settlement. This can encompass, for example, the processing of domestic and international payments, securities, mortgages, insurance products, claims, and data processing. These activities are mainly supported by IT, hence OPS/IT. The director of the staff department Corporate IT also sits in the IT policy board. This Group staff department is responsible for IT policy preparation, the provision of IT advice to the businesses and monitoring the IT activity within the entire organization. The *IT Leadership Council* includes mainly business CIOs and provides advice to the Policy Board. It has three *sub-groups*, dealing with respectively IT standards, IT architecture and IT infrastructure. In line with the importance of information security the separate *Information Security Steering Committee* reports directly on this subject to the IT policy board. Within the ECs the Application Forums and Infrastructure Groups are responsible for standardization of application areas and for integration of systems and infrastructure.

1.3 IT Performance and Investment Management

The IT Performance and Investment Management Team (ITPM) main task is to monitor the IT activity within the entire organization. Their main operational activity is the deliverance of management reports on IT performance to the Executive and IT Policy Board and to the Executive and Management Centres. These reports deliver valuable information to support decision-making with respect to the IT policy.

The underlying data for these reports is collected by means of the IT section-reporting template as part of the yearly MTP (Middle Term Planning) reporting cycle. This template has been providing general IT cost and staff data and other IT-related features from all Business Units (BU) for the past few years. IT performance indicators were developed that provided valuable insights into the effectiveness and efficiency of the use of IT within the organisation. Benchmarking studies offered similar indicators for competitors in the same lines of businesses as the BUs of ING. The comparisons provided a good overview of the IT position of ING with respect to its main competitors.

The IT investment approach was the follow-up and broadened the focus from general IT performance to IT project performance and the IT section template was expanded in 2001 with a section with data on the top five of completed and running IT projects (IT-enabled Business investments). In 2001 the transparency of this project data was not high, which means a lot of missing or unreliable data was reported. The BUs were clearly not ready to deliver this kind of data. The IT section template of the MTP report in November 2002 already provided a lot more project data. The management reports were enriched with project performance information, such as the percentage of total reported projects that were delivered on budget, on time and with 95% functionality. These three criteria can be seen as the main management risks of a project in the development phase.

The research towards new IT performance metrics and the continuous improvement of the whole data collection process are thus also important activities within ITPM.

1.4 Outline of the master thesis

After having introduced the general purpose of this research briefly and having given a description of the research environment, the goal of this master project is described based on the research questions in the next chapter.
The rest of the thesis consists of three main parts. In Chapter 3 we will discuss the data collection process and quality topics concerned with the collected project risk drivers and project risks. Chapter 4 represents the mathematical part of the master thesis and describes the steps from the rough data from Chapter 3 to a risk prediction model. The reader who has no interest in the math can proceed with Chapter 5. In Chapter 5 the results of the research as well as the practical purposes of the models for ING are presented. The final conclusions of our report are summarised in Chapter 6.

2 Research Questions

The research toward project risk assessment metrics is an important research topic within ITPM. Qualitative risk assessments methods are used that are based on questionnaires completed by a project manager. This master thesis addresses the need for a more objective and quantitatively based risk assessment method. The three main risks of budget overrun, time overrun and less functionality can be quantified using the main project success criteria:

- Completing the project within the agreed budget.
- Completing the project within the agreed duration.
- Delivering more than 95% of the agreed functionality.

The research aims at answering the following general research questions:

- Can we develop a formal mathematical methodology that generates early warning signals for not meeting the three main project success criteria?

- Which project and BU characteristics can be considered as important risk drivers?

- What is the alignment of the outcome of the mathematical models with the real ING business situation?

These questions are reflected throughout the research by the research objectives. These objectives represent a classical modelling approach. In the first stage the input data of the model is checked. These data should represent the real situation well, because otherwise no model has meaning. In the following phase, a model is build that reflects the real business problem. Finally the outcomes of the mathematical model should be transformed and into practical results.

The research objectives therefore are:

- Determine and assess project and BU characteristics, which are measurable in an early project phase, as possible risk drivers.

- Develop a mathematical model that is intuitively easy to interpret and predicts a project risk based on related risk drivers.

- Translate the outcome of the models into the business context (ING).

Even if the developed predictive models turn out to be rather indicative, the outcome of the research will provide valuable insights into the current data collection process and the possibilities of predictive risk assessment. This master thesis is not only written for members of Corporate IT, but also for project managers throughout ING. They are responsible for the deliverance of reliable project data and this master thesis will show them the importance a sound reporting on this project data.

3 Data collection

Within Corporate IT project features and general BU features are collected and these features represent the possible input data for the predictive modelling. This chapter deals with the search and assessment of possible risk drivers that can be assessed in an early project phase.

In Section 3.1 an overview of important risk drivers according to the literature is presented. Section 3.2 discusses the data collection process within ITPM. The collected ING data is summarised and compared with ideal risk drivers. The quality issues concerning the research data are presented in Section 3.3.

3.1 Perfect situation

In a perfect situation a data analyst is involved in all phases of the research. In this research the project data have already been collected and the scope of the projects is confined to the development phase. We have to focus on the risks of budget and time overrun and less delivered functionality. What kind of possible risk drivers would we ideally have collected with respect to these risks in the development phase?

The search for risk drivers in the literature for our specific projects was not straightforward. The reported risk factors in most studies are not necessarily risk drivers because several reported risk factors cannot be influenced before or in the early stages of a project. We conclude this section by summarising the risk factors that can be considered as pure risk drivers.

Another obstacle is that most risk factors that are mentioned in the literature refer to software development projects. Our projects under study are no pure IT projects, but IT-enabled business projects. The IT component plays an important role, however, and the found software risk factors could therefore be well used.

The risk factors were derived from the findings of empirical studies in the field.
We note the important risk factors as well as the research method of the study.

In their CHAOS reports the Standish group [6,7,8] does report on important failure and success factors of certain types of projects.

- **Successful project**
  The project is completed on time and on budget, with all features and functions as initially specified.

- **Challenged project**
  The project is completed and operational, but over budget, over the time estimate, and offers fewer features and functions than originally specified.

- **Cancelled project**
  The project is cancelled at a certain point during the development cycle.

In their CHAOS Report from 1994 [8], a top ten of success factors was presented for the successful projects and consequently two top tens of risk factors were given for the challenged and cancelled projects.

The list of success factors and risk factors were very similar and in the CHAOS report in 1999 and 2001 [6,7] only the top ten of success criteria was reported. We mention the top ten of success factors in the latest Chaos report [6].

1. Executive Support (18 Points)
2. User Involvement (16 Points)
3. Experienced Project Manager (15 Points)
4. Clear Business Objectives (12 Points)
5. Minimised Scope (10 Points)
6. Standard Software Infrastructure (8 Points)
7. Firm Basic Requirements (6 Points)
8. Formal Methodology (6 Points)
9. Reliable Estimates (5)
10. Other (5)

The number between the brackets shows the weight, which has to be attached to each available success factor. The total score of a project can sum up to 100 and gives an indication of its successfulness. A low score thus means that the project is potentially risky and the success factors are easily transformed into risk factors.

The Standish Group research is done through focus groups, in-depth surveys and extensive interviews with Fortune 500 companies. The research focuses on mission-critical software applications, management techniques and technologies.

The Standish Group CHAOS research is the largest body of primary research in the IT community according to their own website. An important source of information for the CHAOS is a yearly web questionnaire filled in by the so-called SURF members. Standish User Research Forum (SURF) is a collection of IT executives from various user organizations throughout the world. These executives represent a cross section of the IT community from different industries, organization sizes, and geographic locations.

In [17] three Delphi surveys (Figure 3.1 on page 9) were deployed to identify a ranked list of project risk factors. These surveys were conducted in three different countries: Hong Kong (HK), Finland (FIN) and the United States (US). The three panels (one for each country) consisted of experienced project managers.
The Delphi Survey process in this study consisted of three phases. In the first phase of *brainstorming* each panelist made a list of possible risk factors (at least six). The collected risk factors were compared with each other and combined into an extensive list of risk factors (without exact doubles and *similar* risk factors). In the next step each panel, independently of the others, narrowed this list down into a more manageable list. Each panelist chose the most important factors (at least ten) from a random set of risk factors based the previous extensive list. Each factor that was picked by at least 50% of the participant was taken into account in the final phase. The initial list of more than 150 items was now reduced into three smaller lists (HKG – 15, FIN – 23, USA – 17).In the last stage each panellist ranks the remaining risk factors for their panel in order of importance, which results in an overall project risk-ranking list per separate panel.
In order to get an international ranking, a composite ranking was made. All eleven risk factors that were present in all three ranked lists were ordered by their average relative ranks. An important risk factor that is left out of this composite ranking was *the lack of effective project management skills*. This factor was high-ranked in FIN (1) and USA (5), but was not selected at all.

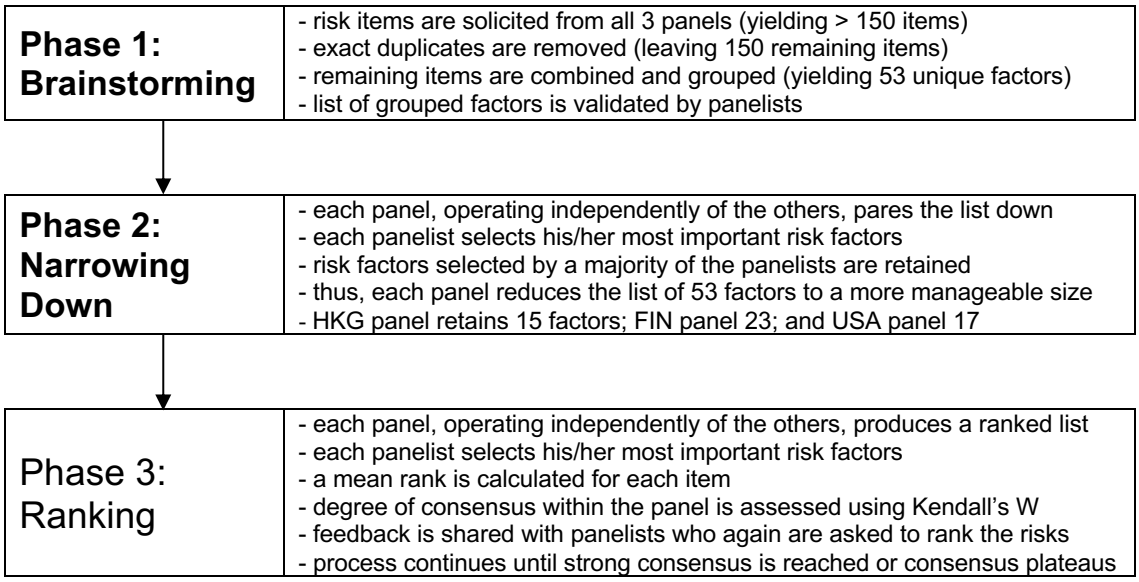


Figure 3.1: Description of Delphi Survey Process Used in [17]

The composite top eleven of most important risk factors from [17].

1. Lack of top management commitment to the project
2. Failure to gain user commitment
3. Misunderstanding the requirements
4. Lack of adequate user involvement
5. Lack of required knowledge/skills in the project personnel
6. Lack of frozen requirements
7. Changing scope/objectives
8. Introduction of new technology
9. Failure to manage end user expectations
10. Insufficient / inappropriate staffing
11. Conflict between user departments

All mentioned risk factors were so far assessed on a high management level and thus represent software project management issues. We also looked at operational software risk factors.

In his book Jones [10, p 115] stated that software projects are influenced by as many as 250 different factors that can affect schedule, cost, quality and user satisfaction, but only ten to twenty major issues usually affect individual projects.

Jones furthermore recognised 36 key data factors that should be included in all assessment and benchmark studies and illustrate the kinds if things that can influence whether software projects will be successful or unsuccessful. These key factors are divided into six main categories of software factors:

- **Classification factors:** represent the specific nature of the software project.
- **Project Specific factors:** depict characteristics of individual projects.
- **Technology factors:** record specific tool and methodologies utilized.
- **Sociological factors:** indicate sociological and experience factors.
- **Ergonomic factors:** indicate the working circumstances.
- **International factors:** are of importance to large global software projects.

We summarise a set of key factors from [10].

- **Classification:** a project can be a high-risk strategic project with high impact on the business or a low risk informational project.
- **Project Specific:** Size, complexity, constraints, class and type and scope of the software applications. An example of a constraint of a software application is a staffing or team size limitation.
- **Sociological:** experience level of the development team and the project managers or the capability maturity level of the development organisation.
- **Technology:** Any formal development methodology used, Project management and development tool suites used on the project.
- **Ergonomic:** Size of the office space, availability of meeting spaces.
- **International:** Local laws that affect international projects, variations in compensation levels for the different countries involved.

Our research focus is on quantifiable risk drivers in an early stage of an individual project. Risk drivers can be regarded as risk factors that can be influenced before or in the early stage of a project. The classification of a project into a software type cannot be influenced and is not viewed a possible risk driver. The commitment of top management can be influenced, but is very hard to quantify in the initial stage of a project and is thus not suited as risk driver for our research.

The following factors are considered as quantifiable risk drivers in an early project phase of our IT-enabled Business investments.

- Size of the pure IT part of the project, e.g. function points, lines of code or number of applications.

- Staffing essentials, e.g. number of FTEs working on the project, ideally broken up into activities.

- Project constraints, e.g. Does the project have a fixed date or a fixed price contract?

- Is a formal development or project management methodology used?

- The capability maturity level of the software development organization, which is more generally know as the CMM level.

The following risk drivers are mentioned in the previous section, but the added quantifiable measures are recommendations from the author.

- Scope of the project, e.g. the number of milestones in the project or the number of changes in the user requirements.

- The experience level of the project team, which can be measured by total years of experience or by percentage of successful projects.

- The experience level of the project manager, which can be measured by the percentage of successful projects or total years of experience.

- The experience level of the user, percentage of successful projects done for a certain user.

3.2 Available data within ING

This section is not published due to confidential information.

3.3 Data completeness

This section is not published due to confidential information.

4 Data analysis

Statistical data analysis consists of collecting, analysing and interpreting data. The collection process and the available research data have already been described in the previous chapter. This chapter thus describes the mathematical part of this master thesis and focuses on the statistical data analysis of the research data and the development of predictive modelling. The first section contains a brief and general introduction to frequently used statistical terms and methods. We explain our choice for logistic modelling and also discuss our specific approach of the data analysis in Section 4.2. Section 4.3 demonstrates the practical data analysis model building issues based on the used statistical software package Splus.

4.1 Introduction to statistics

We introduce a few basic statistical terms. Instead of data we use the term v*ariable,* which represents a certain characteristic of a project, e.g. the planned budget.

The variables can be classified into several pairs:

- *Response variable:* is the actual object of the research and is often called a dependent variable.
- *Explanatory variable:* is also called an independent variable and the goal of the analysis is to determine its influence on the dependent variable.

- *Quantitative variable:* measured on a numeric or quantitative scale.
- *Qualitative variable:* no natural sense of ordering (also: categorical variable).

- *Continuous variable:* with a quantitative scale, having a continuous and infinite range of values.
- *Discrete variable:* having a finite number of values. A variable with enough discrete values can be considered as continuous for practical purposes.

We should note that all qualitative variables are discrete, but that quantitative variables can be discrete. An example is the CMM level that is rated as 1,2,3,4,5.

We can further categorise the variables according to their measure level (Table 4.1). These measure levels are classified from high to low (one to four). The higher the level, the more mathematical operations on the variables are allowed.

Class	Measure level	Description	Example in our data set
1	Ratio	Variable on a quantitative scale with an absolute zero point.	Planned Budget
2	Interval	Variable on a quantitative scale with an arbitrary zero point.	Budget Deviation (+/-)
3	Ordinal	Variable on an ordered, but not truly quantitative scale.	CMM Level
4	Nominal	Variable on a scale that consists of non-ordered labels.	Project Category

Table 4.1: Different Measure levels

The purpose of a proper statistical analysis is to examine and summarise certain aspects of each variable. We need to know how a certain variable is distributed and need to check if certain observations are abnormal. This step is very important since the distribution of variable determines which statistical methods are more or less suited to use. Another important step in the exploratory analysis is to determine the relation between pairs of variables, which is necessary to determine a set of independent variables that can be used in the predictive modelling phase.

This exploratory phase of checking distributions and simple relations can be done in several ways by looking at g*raphical plots and numerical summary statistics.*

The plots and summaries provide indications of our assumptions concerning the variables. In order to take a more formalised decision we verify our assumptions with statistical tests.

These tests use a statistical model to either reject or accept a certain assumption, which is called the *null* hypothesis. The main objective of each test is to reject this null hypothesis and accept the *alternative* hypothesis, which is considered as a strong conclusion. The other situation of not rejecting, but also not accepting the null hypothesis cannot be regarded as an actual conclusion. The error of incorrectly rejecting the null hypothesis should be as small as possible.

The choice of the right test is very important since the test is based on certain assumptions. The use of an improper test will lead to wrong conclusions about the tested hypothesis.

The next phase is the development of statistical model in which we look at more complex relations between variables. In this modelling phase a response variable is related to a certain function of independent explanatory variables. The objective of the modelling phase is to determine the best function of variables that predicts the response with the lowest error.

All assumptions concerning distributions and relations, statistical tests and graphical plots are meant to support us in the modelling phase.

With this background information on statistics the following sections should be easier to read.

4.2 Preliminary analysis

In this section we first of all explain the choice for an appropriate modelling technique and subsequently explain our limitations with respect to the exploratory data analysis. This section concludes with an assessment whether the overall research sample is representative or not.

4.2.1 Modelling techniques

We recall the objective of this section from Chapter 2.

- Develop a *mathematical* model that is intuitively easy to interpret and predicts a project risk based on related risk drivers.

The choice of the modelling technique depends on the outlook and type of variables. In Chapter 3 we have already discussed the data collection and formulated a set of possible research risk drivers. These risk drivers are the explanatory variables in our model. The only uncertainty in the objective is the choice of how to represent the project risk in our model. We will therefore first focus on the determination of the risk response variables.

As stated in Chapter 2 the goal of our research is to predict the risk that a certain project does not meet one of the project success criteria. ING identifies three main project success criteria:

- Completing the project within the agreed budget
- Completing the project within the agreed duration
- Delivering more than 95% of the agreed functionality

These criteria suggest the use of binary categorical variables as response variables; a project is classified into a budget failure (BF), a duration failure (DF), a functionality failure (FF) and an overall failure (OF).

$$BF = \begin{cases} 1 & \textit{The project did not meet the budget success criterium} \\ 0 & \textit{The project did meet the budget success criterium} \end{cases}$$

$$DF = \begin{cases} 1 & \textit{The project did not meet the duration success criterium} \\ 0 & \textit{The project did meet the duration success criterium} \end{cases}$$

$$FF = \begin{cases} 1 & \textit{The project did not meet the functionality success criterium} \\ 0 & \textit{The project did meet the functionality success criterium} \end{cases}$$

$$OF = \begin{cases} 1 & \textit{The project did not meet all 3 success criteria} \\ 0 & \textit{The project did meet all 3 success criteria} \end{cases}$$

Another possibility is to look at the actual overruns with respect to budget, duration and functionality (BO,DO,FO). These are continuous response variables, which can be turned in to risk variables by using a certain calculation, e.g. the risk is the percentage relative expected budget overrun with respect to the largest overrun.

We considered the following three modelling techniques, which could be executed by our statistical package Splus.

Multiple Linear Regression

The simplest model that enables the prediction of the amount of risk using a set of continuous or categorical variables is a multiple linear regression model. This technique assumes in general a continuous and normally distributed response variable that is expected to be linearly dependent on a set of explanatory variables.

Generalised Linear Models (GLM)

In 1972 Nelder and Wedderburn [13] introduced a generalised class of linear models that enables us to fit a model to a response variable that depends on a function of a linear model. The response does not have to be normally distributed variable or a continuous variable.

Logistic regression

A popular example of a generalised model is the logistic regression model, that is a frequently used model in medical studies. In [2] logistic regression is used in a clinical research to determine the influence of certain characteristics as gestational age on the mortality risk of babies.
This technique thus expands the possibilities of the normal linear regression. The use of logistic regression in medical studies started in 1962 according to [3], although a framework for the generalised linear models was not provided until 1972 (Nelder and Wedderburn).

Logistic regression provides the possibility to predict the outcome of a categorical (binary) response variable. Logistic regression does this by taking a certain transformation of the binary response variable (**Logit** transformation: see Appendix B for the exact mathematical notation) and then by fitting a linear relation of explanatory variables to this transformed response variable. In our study the output of the model is a formula from which the probability of the failure of a project can be predicted by using the values of a set of risk drivers.

Tree modelling

We explain tree-based modelling using [11]. We make a distinction between a *classification* tree to predict categorical response variables and a *regression* tree to predict continuous response variables. The response variable is predicted by a set of rules. We thus distinguish two types of rules:

1 *Classification* rule with a categorical response variable (failure, no failure).
2 *Regression* rule with a continuous response variable.

We illustrate these two types of rules with imaginary examples:

1. If BP (budget – planned) < 2 and EC is Europe, then project fails (BF = 1).
2. If BP < 2 and EC is Europe, then BO = 25%.

The rules are categorised in a tree like fashion, in which each node contains a rule, which leads to maximal two new rules (branches) or solutions.

4.2.2 Choice of the modelling technique

The most appropriate technique for our specific data set and research questions turned out to be logistic regression. We explain our choice based on our research objective as stated earlier in this chapter, which delivers us two guidelines:

1. The response variable should represent the risk on a certain project failure
2. The set of explanatory or predictor variables should be insightful.

This first guideline enables us to make a choice between linear and logistic regression. We prefer logistic regression based on the following arguments:

- A multi-linear regression model has a continuous response variable, which for example can predict the actual overrun in budget. The risk of each project is then derived from the difference between planned and actual. This model nevertheless is not preferred, since our continuous response (overrun) variables are not normally distributed. This assumption of normality is a basic requirement for a multi-linear regression model.

- Logistic regression requires a categorical response variable and turns out to be more suitable for the purposes of the early warning system, while this response variable represent the probability of failure of a project directly.

- Moreover the approach with categorical risk responses enables us to look at the overall risk of a project, which is not possible with continuous response variables.

Classification tree modelling and logistic regression are the two remaining techniques with a categorical response. The desire for a more insightful formula again leads to a preference of logistic regression.

- Tree modelling leads to more complex models and to a less transparent overview of the risk drivers. The models can be complex combinations of binary trees, which lead to ambiguous combination of decision rules.

- The logistic regression models are far more intuitive. The linear formula of explanatory risk drivers with corresponding coefficients directly indicates the influence of the risk drivers according to [2]. The choice of logistic regression to determine the mortality risk was mainly because *the parameters in a logistic equation had a simple and direct clinical interpretation* [2].

- Trees tend to over-fit the model more than logistic regression models do.

In the light of the both guidelines for the risk response and the outlook of the formula of explanatory variable, logistic regression turns out to be the best choice for our modelling technique.

4.2.3 Analysis approach

The choice of a proper modelling technique is only an initial step. The model has to be meaningful and representative for ING because the outcomes of the model have to be placed into the ING business context.

The main condition for developing a meaningful model is a representative data set. The famous quotation *garbage in - garbage out* describes the truism danger of a lack of representative data. A model based on unreliable data will provide unreliable predictions.

In the ideal situation we should have checked the following data issues before the actual logistic model building starts.

1. Is the data sample representative for ING?

   The collected project data should be a decent reflection of all project data within ING in order to avoid developing models without any practical meaning.

2. Is the collected data homogeneous?

   The data should be homogeneous, which means the quality of being similar or comparable in kind or nature. An inhomogeneous dataset threatens the validity of our model.

3. Are the explanatory variables independent?

   The set of explanatory variables included in the modelling phase should not be mutually dependent. These dependencies cause *multicollinearity* between the explanatory variables, which among others lead to less reliable estimates of the coefficients in the logistic equation.

The rest of this section contains confidential information.

4.3 Logistic modelling

This section describes the mathematical modelling in this master thesis. We start with a general explanation of unsettled data topics in our specific situation and discuss the consequences for the modelling approach. The development of the logistic models is discussed by using a few examples. This demonstration is backed up by the used commands in Splus, which is the statistical package used for our data analysis. We conclude with a discussion on the validity of the developed models.

4.3.1 Overall Model building approach

This logistic modelling was not a straightforward process, but was an iterative process of trial and error to find the most appropriate model. The undone issues in the previous section were the checks for homogeneity and independence in the set of risk drivers. Clustering analysis is a proper research method to determine sets of homogeneous data, but this technique is inappropriate when we consider our small data set and presence of a lot of categorical variables. Simple outlier detection for all continuous research variables is more suited in our situation.

We note that the overall data set consists of so-called *pooled* data. The original data is collected from various different samples (different Bus or ECs). We cannot develop individual models on these small samples and therefore take all data together; all data is *pooled* into an overall set. We are aware of the danger of looking at all data together and therefore also look at smaller subsets. Possible subsets can be derived from the classification factors in Chapter 3. We depict the subsets with their subtotals.

Executive Centre (EC):	Europe (x), Americas (x), Asia/Pacific (x)
Line of Business (LOB):	Banking (x), Insurance (x), Asset management (x)
Project category (PC):	Transactional projects (x), Informational projects (x), Strategic projects (x), Infrastructure projects (x)

In theory we should develop logistic models for each subset and compare them with the models of the overall set. Most subsets are simply too small for the proper development of a logistic model, so we choose the following approach:

A. We analyse the risk drivers of the overall set and depict the most obvious outliers (extreme values) and also research the dependencies between the pairs of risk drivers.

B. We build risk models for the four risk responses on the overall data set and select the models with a reasonable fit for further analysis.

C. The influence of outliers (as found in step A) is checked by looking at the predictive performance of best models on large subsets of data.

D. We choose the best model with the most general predictive performance and the simplest and most stable regression equation. We can assess the stability of this equation by examining the dependencies between the risk drivers and by looking at the change in equations of sub models (based on same subsets as in Step C).

4.3.2 Overall data summary (Step A)

This section is not published due to confidential information.

4.3.3 Splus modeling (Step B)

The statistical package Splus was used to estimate the unknown coefficients of the logistic models. We explain the automated part of the logistic model building using the used Splus commands. We use the budget risk models based on the overall data set to demonstrate this initial part of the modelling. The underlying math of this section is summarised in Appendix B3.

The Splus command to fit a logistic model is the general linear model fit (*glm*) with a binomial probability family. We show the null model with only a fixed intercept and without any variables included.

```
>  nullmodel_glm(formula = br ~ 1,family = binomial(link = logit), data = dataset)
```

We can use the *glm* function to fit models with all possible combinations of variables and then select the best model. This approach is time-consuming and the use of the Splus *step* function is much more convenient. This function searches the optimal model with two algorithms.

Forward step algorithm
Variables are sequentially added to a certain small model, based on their added value according to the so-called Akaike Information Criterion (AIC). A variable ads value if the decrease in deviance (depicting the size of the error of the model) is greater than the increase in complexity (the number of variables in the model). In Appendix B3 more detailed information can be found on this AIC.

Backward step algorithm
Variables are sequentially deleted from a certain large model based on the AIC.

Our initial thought was to start from the null model and use the forward algorithm to sequentially add all variables and interactions between variables. This forward step function is depicted by the following command.

```
> step(nullmodel, list(upper = ~.^2, lower = ~ 1), direction = "forward",trace = F)
```

The upper scope of the model (.^2) states that all variables and interactions should be included in algorithm, but Splus simply cannot handle all these possibilities.

Our other option would have been to apply a backward algorithm to a full model with all variables and mutual interaction included.

```
>  fullmodel_glm(formula = br ~ .^2,family = binomial(link = logit), data = dataset)
```

Splus is unable to fit the full model since the iterative algorithm that estimates all coefficients cannot handle the amount of variables and interactions. We thus cannot use the backward step function with a full model as starting point.

We have seen so far that we cannot simply apply the step function on either a null or full model. We have to choose sensible upper and lower bounds for the step function and also have to develop a suitable full model. We therefore fit an *approximate* full model by using a forward step function that includes interactions to the model with all continuous variables.

> allmodel_glm(formula = br ~ bp + dp + pp + io + cmm + bs + pmm + dds + ids + rqf,family = binomial(link = logit),data = dataset)
> fullmodel.approx_step(allmodel, list(upper = ~.^2 , lower = ~ 1), direction = "forward",trace = F)

This full model includes all variables in the regression equation and also four interactions. The regression equation is thus hard to interpret and we therefore choose to develop sub models. We depict these models in order of complexity.

- **model1**: only continuous variables are investigated.

- **model2**: all continuous and categorical variables were examined, but interactions between the pairs of variables are ignored.

- **model3**: all variables included in model2 and all interactions between these variables were considered.

The omission of the interactions enables us to fit full models for model1 and model2 in Splus.

>fullmodel1_glm(formula = br ~ bp + dp + pp + bs + dds + ids,family = binomial (link = logit), data = dataset)

>fullmodel2_glm(formula = br ~ bp + dp + pp + io + cmm + bs + pmm + dds + ids + rqf,family = binomial(link = logit),data = dataset)

We are now able to use the step function in a forward and backward direction.

>forwardmodel1_step(nullmodel, list(upper = ~ bp + dp + pp + bs + dds + ids, lower = ~ 1), direction = "forward")
>backwardmodel1_step(fullmodel1, list(upper = ~ bp + dp + pp + bs + dds + ids, lower = ~ 1), direction = "backward")
>forwardmodel2_step(nullmodel, list(upper = ~ bp + dp + pp + io +cmm + bs + pmm + dds + ids + rqf, lower = ~ 1), direction = "forward")
>backwardmodel2_step(fullmodel2, list(upper = ~ bp + dp + pp + io +cmm + bs + pmm + dds + ids + rqf, lower = ~ 1), direction = "backward")

The forward and backward modelling can lead to different models. Remember that each model is characterised by its deviance (error measure) and its complexity (the number of variables included). We distinguish a set of decision rules that are used to decide when the deviance and complexity of the two models are not exactly equal.

- If both models have the same deviance then the simplest model is taken as the best model
- If both models are equally complex then the model with the lowest deviance is taken as the best model.
- If one model is less complex and also has a lower deviance, then this model is obviously better.
- If one model is more complex and has a lower deviance, then a likelihood ratio test is necessary to determine the best. The outcome of this test shows whether the decrease in deviance is large enough to compensate for the increased complexity. The likelihood ratio test is described in more detail in Appendix B3.

We have written a function in Splus that applies all these rules. We will not depict this whole function, but consider the situation of a more complex backward model with lower deviance than the simpler forward model. In this case we thus have to apply the following statistical test.

```
> # apply ratio test with chi-squared test-statistic
> p.value_anova(forwardmodel,backwardmodel,test="Chisq")[2,7],4)
> # the nullhypothesis that the simple forward model is appopriate is rejected if the
> # p.value (unreliability) is less than 0.05 (the general unreliability level of a test)
> if(p.value <= 0.05) {
> # choose backward model: more complex, but with a significant lower deviance
> model_backwardmodel
> }
> else{
> # choose forward model: backward model has no significant lower deviance
> model_forwardmodel
> }
```

We now include interactions in the model by using a forward step function that expands model2 with all possible mutual interactions.

```
>model3_step(model2, list(upper = ~.^2 , lower = ~ 1), direction = "forward")
```

By building these general logistic models we have assumed a binomial distribution for the response variable. In that case the so-called scale or dispersion parameter in the variance function is supposed to be one. This assumption does not always hold, but the collection method does not allow us to check whether the dispersion in the variance of the response variable is smaller or larger than one.

Our only possibility is to fit a quasi-likelihood model that estimates this dispersion parameter. This parameter is used in the AIC of the step function and can thus probably lead to a different model. This model is then compared with the general model and the best model is chosen.

In order to check the assumption we use *quasi-likelihood* estimation in the model building by changing the family parameter within the glm() function for all three sub models.

> glm(formula = br ~….., **family = quasi(link = "logit", var = "mu(1-mu)")**, data = dataset)

The full models and the null model are changed and the best quasi-models are found analogously as the general models. Each quasi model is now compared with the general model by looking at the balance between complexity and deviance. We thus do not pay much attention to the estimated dispersion parameter, but look at the change in models.

Table 4.10 depicts which of the two estimation methods is preferred for all sub models of our four risks. Model4 represents the approximate full model, which is the most complex model according to included variables and interactions.

	Comparison between estimation methods
Budget model1	No difference between quasi- and log-likelihood estimation
Budget model2	No difference between quasi- and log-likelihood estimation
Budget model3	*Quasi-likelihood* estimation leads to less complex model
Budget model4	No difference between quasi- and log-likelihood estimation
Duration model1	No difference between quasi- and log-likelihood estimation
Duration model2	No difference between quasi- and log-likelihood estimation
Duration model3	No difference between quasi- and log-likelihood estimation
Duration model4	No difference between quasi- and log-likelihood estimation
Func. model1	No difference between quasi- and log-likelihood estimation
Func. model2	No difference between quasi- and log-likelihood estimation
Func. model3	*Log-likelihood* estimation leads to less complex model
Func. model4	No difference between quasi- and log-likelihood estimation
Overall model1	*Quasi-likelihood* estimation leads to less complex model
Overall model2	*Quasi-likelihood* estimation leads to less complex model
Overall model3	*Quasi-likelihood* estimation leads to less complex model
Overall model4	*Log-likelihood* estimation leads to a significantly better model

Table 4.10: Preferred estimation method per risk model

Table 4.10 shows that log-*likelihood* estimation often leads to the same models as *quasi-likelihood* estimation does. In only one case (overall model4) we have been able to take a strong conclusion that *log-likelihood* estimation is significantly better than *quasi-likelihood* estimation. For other models with differences between both estimation methods, we cannot take a strong conclusion, which of both methods is preferred. If we for example look at the budget model3, we notice that *log-likelihood* estimation is not significantly better than *quasi* estimation. We thus prefer the *quasi* estimation for budget model 3, since this method leads to a less complex model than the *log-likelihood* estimation.

Goodness-of-fit of the models

We now introduce a few goodness-of-fit metrics that depict the quality of the logistic model like the R^2 does for linear regression models. These metrics are used to identify the best models for further analysis. The small data set prevents a sensible division in a test and a training set. The goodness-of-fit metrics thus also depict the quality of the predictions.

The McFadden *Pseudo* R^2 depicts the amount of reduced deviance by the estimated model as a percentage of the deviance of the null model. A similar measure is the LRFC statistic, which also depicts the quality of the fit of the estimated deviance. This metric uses the exponent of the difference in log-likelihood between null model and the estimated model. The original formulas of McFadden R^2 and the LRFC statistic use a null model with an intercept included according to [23] and [4]. In Appendix B3 the reader can find the mathematical equations of both statistics. Each null model with an intercept has a different null deviance, while this intercept is different for each risk response, which makes it hard to compare between the different risk models.
We therefore prefer to use a null model without an intercept since this null model has similar null deviances for all four risks. We compare the so-called Mcfadden R^2 (0) and LRFC (0) of all four models for each risk in Table 4.11 on the next page.

Budget risk	model1	model2	model3	model4
McFadden R^2 (0)	0.192	0.259	0.343	0.373
LRFC statistic (0)	0.142	0.197	0.269	0.295
Functionality risk	**model1**	**model2**	**model3**	**model4**
McFadden R^2 (0)	0.302	0.324	0.380	0.644
LRFC statistic (0)	0.233	0.252	0.301	0.563
Duration risk	**model1**	**model2**	**model3**	**model4**
McFadden R^2 (0)	0.065	0.065	0.065	0.346
LRFC statistic (0)	0.046	0.046	0.046	0.332
Overall risk	**model1**	**model2**	**model3**	**model4**
McFadden R^2 (0)	0.117	0.117	0.155	0.359
LRFC statistic (0)	0.084	0.084	0.113	0.282

Table 4.11: Goodness-of-fit metrics of the best possible overall models

These statistics show that the simpler models for budget and functionality risks outperform those for overall and duration risks. The *approximate* full model4 is the best for the functionality risk, but this model includes all variables and several interactions between these variables, which makes the model harder to interpret. This model4 depicts the model with the lowest possible deviance as calculated by Splus with the AIC. We adapt the McFadden R^2 (0) of the simpler models (1,2,3) of Table 4.11 by dividing them by the McFadden R^2 (0) of model4.

Relative R^2 (based on R^2 model4)	Model1	model2	model3
Budget risk models	0.515	0.694	0.920
Functionality risk models	0.469	0.503	0.590
Duration risk models	0.188	0.188	0.188
Overall risk models	0.326	0.326	0.432

These relative R^2 also demonstrate that the budget and functionality models perform better than the overall and duration models. We will thus focus on these two models for further analysis.

4.3.4 Generality of the models (Step C)

We have noticed that the budget and functionality models show the best predictive potential using their Mcfadden R^2 metrics in the previous subsection. We first of all search for the most consistent model with respect to its predictive performance.

Our data analysis showed obvious outliers and also demonstrated that our data mainly consists of relatively small projects based on the project size indicators and BU size indicators. We therefore consider the following large subsets that we can use as test sets to explore the overall predictive performance of the models.

- **Outlier set**: subset without the obvious outliers in the overall data set. This data set contains data of projects with a budget smaller than x millions of euros, duration smaller than x months, project power smaller than x million a month. Also the projects from one extremely large BU are left out. This subset is still considered as an overall representation of ING and contains x projects.

- **BP set:** subset that represents all projects with a budget smaller than or equal to x millions of euros. This set contains x projects.

- **DP set:** subset that represents all projects with a duration smaller than or equal to x months. This set contains x projects.

- **PP set:** subset that represents all projects with a project power smaller than or equal to x million per month. This set contains x projects.

- **BS set**: subset that represents all projects from BUs that have a size smaller than or equal to x millions of euros. This set contains x projects.

We compare the predictive quality of the overall models (based on all research data) with the performance of these models on the different subsets. The predictive quality is depicted by the Mean Minus Log-Likelihood (MML), which shows the deviation of the predictions with the original response variable. The perfect prediction has a MML of zero. We consider a model as reasonable when its MML is smaller than –0.5. Table 4.12 obviously shows that the overall predictive budget model1 and model2 perform equally well or even better on the subsets. These models can be considered as the best performing models, because they show the most stable predictive quality.

MML	All data	Outlier Set	BP set	DP set	PP set	BS set
Budget model1	-0.56	**-0.559**	**-0.534**	**-0.526**	**-0.556**	**0.532**
Budget model2	-0.514	**-0.513**	-0.517	**-0.512**	-0.528	-0.515
Budget model3	-0.455	-0.463	-0.500	-0.48	-0.513	-0.488
Func. model1	-0.484	-0.505	-0.509	-0.516	-0.5	-0.494
Func. model2	-0.469	-0.489	-0.504	-0.505	-0.5	-0.499
Func. model3	-0.43	-0.45	-0.491	-0.468	-0.49	-0.462

Table 4.12: Performance of overall models on the different subsets

We have shown that the budget model2 is the best general model for predictive use within ING. The rest of this section contains confidential information.

4.3.5 Stability of the models (Step D)

The predictive ability of both the functionality and budget models is very reasonable and the models are valid to do general predictions within ING. The second part of the research objective was to assess the influence of the risk drivers, which can be done by examining the coefficients of the variables in the regression equation.

The rest of this section contains confidential information.

5 Results

This section describes the practical results of our research for ING. The practical meaning of the observed influences of the risk drivers on the risks will be presented in Section 5.1. The practical meaning of prediction is explained based on our most stable and budget model, which is also the most general model. Section 5.2 will discuss the predictive ability use of this model based on ING management issues.
The practical use of this budget model is demonstrated in Section 5.3.

5.1 Risk drivers of the models

This section is not published due to confidential information.

5.2 Predictive ability of the budget model

This section is not published due to confidential information.

5.3 Practical use of the budget model

This section is not published due to confidential information.

6 Conclusions

Section 6.1 contains the answers to our research questions. We depict a few factors that have limited our research in the next section. The developments for the MTP 2004-2006 reporting period are mentioned in Section 6.3 and finally the possibilities for future research are depicted in the last section.

6.1 Research conclusions

This section is not published due to confidential information.

6.2 Limitations of the research

This section is not published due to confidential information.

6.3 Current Developments

This section is not published due to confidential information.

6.4 Future Research

This section is not published due to confidential information.

7 References

[1] Boehm, B.W. (1981), Software Engineering Economics, Prentice Hall, New Jersey.

[2] Brand, R. (1990), Using logistic regression in perinatal epidemiology, Paediatric and Perinatal Epidemiology, 4, p. 234-249.

[3] Le Cessie, S. (1991), Model building techniques for logistic regression, with applications to medical data,Thesis, University of Leiden.

[4] Darlington, R. B. (1990), Regression and linear models, McGraw-Hill, New York.

[5] Everitt, B.S. (1977), The analysis of Contingency Tables,John Wiley & Sons, New York.

[6] The Standish Group (2001), Extreme CHAOS, Purchase via: https://secure.standishgroup.com/reports/reports.php.

[7] The Standish Group (1999), CHAOS: A recipe for Success, Available via: www.standishgroup.com/sample_research/PDFpages/chaos1998.pdf.

[8] The Standish Group (1995), The CHAOS report (1994), Available via: www.standishgroup.com/sample_research/chaos_1994_1.php.

[9] Johnston, J. (1972), Econometric Methods - 2nd Edition, McGraw-Hill, New York.

[10] Jones, C. (2000), Software Assessments, Benchmarks and Best Practices, Addison-Wesley, New Jersey.

[11] Mathsoft (1999), Splus 2000 - Guide to Statistics,Volume1.

[12] McCullagh, P. and Nelder, J.A. (1989), Generalized Linear Models (2nd ed.), Chapman and Hall, London.

[13] Nelder, J.A. and Wedderburn, R.W.M. (1972), Generalized Linear Models, Journal of the Royal Statistical Society A, 135, p. 370-384.

[14] Paulk, M.C, Weber, C.V, Curtis, B. and Chrissis, M.B. (1995), The Capability Maturity Model: Guidelines for improving the Software Process,Addison-Wesley.

[15] Rice, J.A. (1995), Mathematical Statistics and Data Analysis, Duxbury Press, Belmont.

[16] Schervish (1995), Theory of Statistics, Springer-Verlag, New York.

[17] Schmidt R, Lyytinen K, Keil M. and Cule P. (2001), Identifying software project risks: an international delphi study, Journal of Management Information Systems,17(4), p. 5-36.

[18] Tukey, J.W. (1977), Exploratory Data Analysis,Addison-Wesley, Reading.

[19] Verhoef, C. (2003), Quantitative Aspects of Outsourcing deals, working paper 2003, Available via: http://www.cs.vu.nl/~x/out/out.pdf.

[20] Verhoef, C. (2002), Quantitative IT portfolio management, Science of Computer Programming, 45, No.1 (October 2002), p. 1-96. Available via: www.cs.vu.nl/~x/ipm/ipm.pdf.

[21] Weill, P. and Broadbent, M. (1998), Leveraging the new infrastructure, Harvard Business School Press, Boston.

[22] Witten, I.H. and Frank, E. (2000), Data Mining - Practical Machine Learning Tools and Techniques with Java Implementations,Morgan Kaufmann Publishers, San Francisco.

[23] Woolridge, J.M. (2002), Econometric Analysis of Cross Section and Panel Data, The MIT Press.

Appendix A: Data definitions

A1: Description of collected data

This section is not published due to confidential information.

A2: Description of research data

This section is not published due to confidential information.

Appendix B: Mathematical methods used

B1: Summary statistics

We introduce general summary statistics from a general textbook on data analysis [18]. We first denote that $x_{(i)}$ stands for the so-called *order statistic* of the original sample with observations (x_1,……,x_n). So $x_{(1)}$ refers to the smallest observation of the n observations, $x_{(2)}$ represents the second smallest observation and so on, with $x_{(n)}$ as largest observation.

Mean: $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Median: $$med(x) = \begin{cases} x_{((n+1)/2)}, & if\ n\ is\ odd \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right), & als\ n\ is\ even \end{cases}$$

Standard deviation: $$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Correlation coefficient: $$\rho_{xy} = \frac{s_{xy}}{s_x s_y} \quad with\ s_{xy} = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})$$

1st quartile: $$kwart(x) = \begin{cases} \frac{1}{2}\left(x_{(n/4)} + x_{(n/4+1)}\right), & if\ n\ is\ a\ multiple\ of\ 4 \\ x_{((n+3)/4)}, & if\ n\ is\ a\ multiple\ of\ 4+1 \\ x_{((n+2)/4)}, & if\ n\ is\ a\ multiple\ of\ 4+2 \\ \frac{1}{2}\left(x_{((n+1)/4)} + x_{((n+5)/4)}\right), & if\ n\ is\ a\ multiple\ of\ 4+3 \end{cases}$$

3rd quartile: $$3kwart(x) = \begin{cases} \frac{1}{2}\left(x_{(3n/4)} + x_{(3n/4+1)}\right), & if\ n\ is\ a\ multiple\ of\ 4 \\ x_{((3n+1)/4)}, & if\ n\ is\ a\ multiple\ of\ 4+1 \\ x_{((3n+2)/4)}, & if\ n\ is\ a\ multiple\ of\ 4+2 \\ \frac{1}{2}\left(x_{((3n-1)/4)} + x_{((3n+3)/4)}\right), & if\ n\ is\ a\ multiple\ of\ 4+3 \end{cases}$$

B2: General theory on hypothesis testing

This section contains a general explanation of hypothesis testing from [16] as well as the theory of specific test statistics used in our research. We substantiate these tests with examples.

The testing theory is based on a statistical model. An observation X belongs to a probability distribution with parameter θ. This parameter either is part of a set Θ_0 or a disjunctive set $\Theta_1 = \Theta - \Theta_0$. We distinguish two hypotheses.

$$H_0 : \theta \in \Theta_0 \quad (null\ hypothesis)$$
$$H_1 : \theta \in \Theta_1 \quad (alternative\ hypothesis)$$

A statistical test for these hypotheses can lead to two conclusions.

- Reject H_0 (and accept H_1 as correct).
- Do not reject H_0 (but also do not accept H_1 as correct).

The first conclusion is a strong conclusion and the second is not an actual conclusion. The alternative hypothesis represents the strong conclusion and the goal of each statistical test should be to draw this strong conclusion. We notice two types of errors.

- Error Type 1: reject H_0 when it is in fact correct.
- Error Type 2: do not reject H_0 when it is in fact incorrect.

The first error is very undesirable and represents the unreliability of the test. In general the threshold of unreliability is α = 0.05 (5%). A conclusion with $\alpha \leq 5\%$ is considered as statistically significant.

Kolmogorov-smirnov distribution test

We considered this test to compare if two samples of data are similar. The test considers the following hypotheses concerning distributions F and G of both samples. In case of the one-sample test, the G distribution represents a certain known probability distribution as the normal or exponential distribution.

$$H_0 : F = G \quad (F \in F_0\ in\ one-sample\ test)$$
$$H_1 : F \neq G \quad (F \notin F_0\ in\ one-sample\ test)$$

The Kolmogorov-smirnov test is very suited, because the used *distribution-free* test statistic makes no assumptions concerning a distribution. The test uses empirical distribution functions.

$$\hat{F}_n(x) = \frac{1}{n}\sum_{j=1}^{n} 1\{X_j \leq x\} \quad with\ indicator\ 1\{X_j \leq x\} = \begin{cases} 1 & if\ X_j \leq x \\ 0 & if\ X_j > x \end{cases}$$

The test statistic for the one-sample test is:

$$D_{m,n} = \sup_{-\infty<x<\infty} \left|\hat{F}_m(x) - F(x)\right|$$

The test statistic for the two-sample test is:

$$D_{m,n} = \sup_{-\infty<x<\infty} \left| \hat{F}_m(x) - \hat{G}_n(x) \right|$$

The null hypothesis is rejected when $D_{m,n}$ has a larger value than expected by the empirical distribution. This value can be derived from a table and depends on the sample size and the unreliability of the test (in general α = 0.05). An example of this one-sample test is to test whether variable DDS in the overall data set is normally distributed (See Section 4.3.2). We then use the following command in Splus to perform the test and also show the outcome of this one–sample test.

> ks.gof(dds, distribution = "normal", mean = mean(dds), sd= stdev(dds))
ks = 0.1125, p-value = 0.0308
alternative hypothesis: True cdf is not the normal distn. with the specified parameters

We thus test whether the values of variable DDS are normally distributed with as distribution parameters the mean and standard deviation (sd) of this variable. The outcome of this test is a p-value of 0.03. This p-value means that the test statistic $D_{m,n}$ (ks = 0.11) is greater than the expected values for each α > 0.03 We have assumed that a test with α = 0.05 is still significant. Our p-value is thus smaller than this 0.05 and we can reject the null hypothesis and can assume that DDS is not normally distributed with the specified parameters.

The two-sample Kolmogorov-Smirnov test to compare the distribution of all BUs within ING with those of all Bus in the research data set (See Section 4.2.2) is done by the same command *ks.gof:*

> ks.gof(BUsample, Overallsample)
ks = 0.129, p-value = 0.6784

We have a p-value that is much larger than 0.05 and we thus cannot reject the null hypothesis and have to conclude that the two samples are not significantly different but we also have no prove that they are equal.

Dependency tests

These tests use the following hypotheses:

$$H_0 : X_i \text{ en } Y_i \text{ are independent}$$
$$H_1 : X_i \text{ en } Y_i \text{ are dependent}$$

The Spearman rank correlation test is a distribution-free dependency test and is used for detecting dependencies between pairs of continuous variables and pairs of a continuous and a categorical variable. The test uses rank numbers R and S of the ordered values of X and Y. The test statistic is:

$$l = \frac{\sum_{i=1}^{n} (r_i - \bar{r})(s_i - \bar{s})}{[\sum_{i=1}^{n} (r_i - \bar{r})^2 \sum_{i=1}^{n} (s_i - \bar{s})^2]^{\frac{1}{2}}}$$

The null hypothesis is rejected for values of *l* close to –1 or 1

We show an example of this test using the continuous variables BP and DP in the following Splus command.

> cor.test (BP, DP, method ="s")
rho = 0.612, p-value = 0
alternative hypothesis: true rho is not equal to 0

The p-value of zero is obviously smaller than 0.05 and we thus reject the H_0 and can thus assume that BP and DP are dependent.

A test for dependencies between a pair of categorical variables is based on a contingency table. We look at a general form of a contingency table based on [5]. We have two factors A en B that will be tested against each other.

Factor A (levels: I = 1, .. , k)	Factor B (levels: j = 1, .. , r)					
	B_1	.	B_j	.	B_r	Row Total
A_1	N_{11}	.	.	.	N_{1r}	$N_{1.}$
.	.	.	.	.	.	.
A_i	.	.	N_{ij}	.	.	$N_{i.}$
.	.	.	.	.	.	.
A_k	N_{1k}	.	.	.	N_{kr}	$N_{k.}$
Column total	$N_{.1}$	.	$N_{.j}$	.	$N_{.r}$	n = N..

- A_i and B_j are the different levels of the factor.
- N_{ij} is the actual number of observations that meet both levels of the factors A_i and B_j.
- $N_{.j}$ stands for the column total, in other words the number of observations that meets factor B_j.
- $N_{i.}$stands for the row total, in other word the total of projects that meets factor A_i.
- n = N.. stands for the total size of the sample of observations.

We can use these contingency tables to investigate the relation between each pair of factors. Each cell of this contingency table has a probability p_{ij}. These probabilities have to fulfil the following condition.

$$\sum_{i=1}^{k}\sum_{j=1}^{r} p_{ij} = 1;$$

The dependence of this data can be tested against the following null hypothesis of independence.

$$H_0 : p_{ij} = p_{i.} * p_{.j}, \quad i = 1,\cdots\cdots,k, \quad j = 1,\cdots\cdots r.$$
$$H_1 : p_{ij} \neq p_{i.} * p_{.j}, \quad i = 1,\cdots\cdots,k, \quad j = 1,\cdots\cdots r.$$

Our next job is to test the null hypothesis and the difficulty is the absence of the exact (expected) probabilities. We can solve this problem by using a so-called maximum likelihood estimator (MLE) in order to calculate these probabilities. In our case this is the MLE based on the distribution under the null hypothesis. Our estimated p_{ij} should thus be equal to the probability of factor A ($p_{i.}$) * the probability of factor B ($p_{.j.}$).

Under the H_0 we thus use the so-called Chi-squared (X^2) test statistic. This test statistic has approximately a X^2 –distribution with (k-1)(r-1) degrees of freedom (df).

$$X^2 = \sum_{i=1}^{k}\sum_{j=1}^{r}\frac{(N_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}}, \text{ with MLE: } \hat{p}_{ij} = \frac{N_{i.}N_{.j}}{n^2}.$$

This approximation is reasonable when the following rule of thumb is satisfied:

$$E_{H0}N_{ij} = n * p_{i.} * p_{.j} = \begin{cases} > 1 & \textit{for all cells } (i,j) \\ > 5 & \textit{for at least } 80\% \textit{ of the cells} \end{cases}$$

We now have to compare the test statistic with the X^2–distribution with (k-1)(r-1) for a certain level of unreliability (α = 0.05 is the common value of this level). We can reject the null hypothesis for higher values than the X^2– distribution and then can conclude that the variables are certainly not independent, but are likely to be dependent on each other.

We illustrate the use of the Chi-squared test using an example. We investigate the possible dependency between the type of Executive Centre (EC) and the project source (IO). Table B2.1 depicts the observed number of projects for each combination of the levels of EC and IO.

Executive Centres (EC)	Project source (IO)		
	In house	Outsourced	Row total
Americas	These numbers are confidential		
Asia/Pacific			
Europe			
Column total			

Table B2.1: Contingency table with actual # of projects in levels of EC and IO

In this example the rows consist of all four levels of EC and the columns depict the levels of IO (the project is either developed in-house or outsourced).
The numbers in the cell stand for the actual number of projects, which meet the belonging levels of both variables. Table B2.2 provides the expected numbers of projects assuming these two variables are independent.

EC level (EC)	Project source (IO)		
	In house	Outsourced	Row total
Americas	These numbers are confidential		
Asia/Pacific			
Europe			
Column total			

Table B2.2: Contingency table with expected # of projects in levels of EC and IO

We indeed see in Table B2.2 that none of the cells has an expected number smaller than five. We can thus use the X^2–test by using Splus since more than 80% of the cells has thus an expected value of five or more.

> chisq.test (EC, IO)
X-square = 17.6165, df = 2, p-value = 0.0001

Our calculated p-value is smaller than 0.05 and we reject the hypothesis that the type of EC has influence on whether a project was developed in-house or was largely outsourced. We thus have to assume that both variables are dependent on each other.

We also consider a pair of variables for which we cannot apply the X^2–test.
Tables B2.3 and B2.4 depict the actual and expected number of this pair of variables.

Executive Centres (EC)	CMM level			
	1	2	3	Row total
Americas		These numbers are confidential		
Asia/Pacific				
Europe				
Column total				

Table B2.3: Contingency table with actual # of projects in levels of EC and CMM

Executive Centres (EC)	CMM level			
	1	2	3	Row total
Americas		These numbers are confidential		
Asia/Pacific				
Europe				
Column total				

Table B2.4: Contingency table with expected # of projects in levels of EC and CMM

We now only see 66.7% (3/9) of the cells in Table B2.4 with an expected value larger than five. The X^2–test is not suitable and we use the Fisher exact test. This Fisher exact test calculates the probabilities of all similar tables (with the same row and column totals) of a certain contingency table. This probability is calculated with the formula depicted below:

$$p = \frac{\prod_{i=1}^{k} N_{i.}! \prod_{j=1}^{r} N_{.j}!}{N_{..} \prod_{i=1}^{k} \prod_{j=1}^{r} N_{ij}!}$$

The Fisher test thus calculates the probabilities on all possible similar tables. The probability of the tested contingency table is taken as upper limit and all probabilities that are smaller or equal than this limit are summed up. This summation leads to the exact estimated p-value. In Splus this test is simply the following command:

> fisher.test(ec, cmm)
p-value = 0.0062

We can thus assume that CMM and EC are dependent (0.0062 < 0.05). The interested reader can find more information on this Fisher test in [15].

B3: Logistic Regression

The first part provides a general description of logistic regression. The second part focuses more on model building and the third part on the quality issues of the model.

General theory

We have a set of observations (X_1,Y_1) …, (X_n,Y_n) with the binary response variable Y_i (with thus possible values of 0 or 1) and a number corresponding explanatory variables $X_i = (x_{i1},\ldots,x_{im})$. With logistic regression we study the dependence of the expected response $E(Y_I)$ on the set of covariates (read: explanatory variables). We introduce the generalized linear model (GLM) since logistic regression is a special case of this GLM.

$$h(p_i) = h(E(Y_i)) = X_i\beta \qquad with \quad p_i = P(Y_i = 1 \mid X_i) = p(X_i) = E(Y_i)$$

h is called the link function. To ensure that the probabilities are always between zero and one, the logit link function is used for *h*, which maps (0,1) onto R (real numbers).

$$Logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = X_i\beta \Rightarrow p_i = \frac{e^{X_i\beta}}{1+e^{X_i\beta}}$$

The variance function for the response variable of the logistic regression model is defined as:

$$Var(Y_i) = \phi\frac{p_i}{1-p_i} \qquad with\ \phi\ as\ the\ nuisance\ or\ scale\ parameter$$

In general ϕ is assumed to be one for a logistic regression problem. This assumption will result in underestimating of the standard errors of the parameter estimates when ϕ is smaller or greater than one (under- or over-dispersion). Y_i is said to be over (under)-dispersed if the sampling variance of a response variable Y_i is significantly greater (smaller) than the sampling variance expected by probability distribution (which is in this case the binomial distribution). In [11] the use of the quasi family for quasi-likelihood estimation is proposed as solution for the problem of over- or under-dispersion. The reader who is interested in the exact theory behind quasi-likelihood estimation is referred to McCullagh and Nelder [12].

The coefficients $\beta = (\beta_0, \beta_1, \ldots\ldots, \beta_m)$ of the logistic regression equation need to be estimated. A GLM uses the maximum likelihood estimates. These values are obtained by maximising the log-likelihood $l(\beta)$ with respect to β.

$$l(\beta) = \sum_i [Y_i \log(p_i) + (1-Y_i)\log(1-p_i)]$$

The statistical package Splus which we used to analyse the data maximised $l(\beta)$ numerically by solving score equations. This method is known as *Fisher Scoring* proposed by Nelder and Wedderburn [13]. The method comes down to an *iterative reweighted least-squares (IRLS)* procedure that iteratively obtains updated estimates of βs based on prior estimates of β. We show the principle of the method using the *core* equation that calculated the working response vector z^0 from [13].

$$z_i^0 = \left(Y_i - \mu_i^0\right)h'\left(\mu_i^0\right) + x_i^T \beta^0 \quad with\ for\ example \quad \mu_i^0 = E_{\beta_0} Y_i$$

This working response is reweighted at every iterative step, which leads to a new estimate of β. This updating is repeated until β_n - β_{n-1} is small enough. A more extensive overview can be found in [12,13].

Model building

An important measure in model building is the *deviance*, which measures the difference between a model and the largest possible or saturated model. A saturated model has included one parameter for each observation and thus fits perfectly. In the logistic regression case the deviance is defined in [12] as the scaled log-likelihood ratio statistic:

$$D = 2\phi\left[l\left(\tilde{\theta},\phi\right) - l\left(\hat{\theta},\phi\right)\right] \qquad with \quad \tilde{\theta} : the\ parameter\ values\ of\ the\ saturated\ MODEL$$
$$\hat{\theta} : the\ parameter\ values\ of\ our\ estimated\ MODEL$$

In the logistic regression case the log-likelihood of the saturated model is zero, which leads to a simple formula for the deviance:

$$D = -2\phi\, l\left(\hat{\theta},\phi\right)$$

The deviance is used in the likelihood ratio test, which tests the null hypothesis that the smaller of two *nested* models is adequate. Two models are nested if the simpler model is a part of the more complex model. The difference in deviance between both models is tested with a Chi-square statistic. This test can be used to manually build a model by testing each more complex model against the previous model.

A standard method that is used in most statistical packages is the stepwise model building function. This function adds (or deletes) stepwise variables to (from) a certain model. The stepwise model building method in Splus uses the Akaike Information Criterion (AIC) to stop the iterative step function. This criterion is defined in [11]:

AIC = Deviance (D) + 2*scale (ϕ)*df.resid

The reader should take into account that the residual degrees of freedom (df.resid) are simply the number of parameters in our model and that scale represents the dispersion parameter. The AIC could thus as well be written (with n the number of parameters in the model) as:

$$AIC = -2l\left(\hat{\theta},\phi\right) + 2\phi n$$

The stepwise function stops when the AIC is not getting bigger. Like the likelihood ratio-test, this AIC only includes variables to make the model more complex if these variables significantly decrease the deviance.

Model quality measures

In this part we explain measures that represent the goodness of fit of the model and also consider the quality of the estimates of the individual model coefficients.

A measure that is in line with the well-known R^2 for linear regression is the Darlington LRFC statistic from [4]. This statistic compares the log-likelihood of the null model (with all coefficients of the variables zero) with those of the current model.

$$LRFC = \frac{\exp\left[\left(l(\tilde{\theta},\phi) - l(0,\phi)\right)/n\right] - 1}{\exp\left[-l(0)/n\right] - 1}$$

The McFadden *Pseudo* R^2 is a similar qualitative measure. The original definition from [23] defines the null model as the model with intercept, which is the same null model as for the LRFC statistic.

$$McFadden\, R^2 = 1 - \frac{-2l(\tilde{\theta},\phi)}{-2l(0,\phi)}$$

The quality of the estimates β can be assessed by a confidence interval. An approximate $(1-\alpha)*100\%$ confidence interval for β_i:

$$\hat{\beta}_i \pm t_{(n-p-1);(1-\alpha/2)} * st.error\left(\hat{\beta}_i\right)$$

Remember that the n stands for the total number of observations, p represents the number of included variables (without the intercept) in the model and *st.error* stands for the observed error of the estimate.

The intervals with 0 included indicate that the influence of the coefficient is not that significant in the model. The coefficient can either be larger or smaller than zero, which means either a positive or negative influence on the response. The variables with this type of coefficients are not important. The intervals, which do not include zero, thus depict the statistical significant variables.

Predictive quality measure

The accuracy of the predictions of the model is the main concern in order to determine the predictive quality. The following measure determines the predictive quality of the model.

The mean minus log-likelihood error (MML) is used to assess the quality of the p-value itself. The value zero of MML stands for the perfect model and a big negative value stands for the worst model. This MML should ideally be calculated based on a set of projects that are not used in the model-building phase. If the model predicts on a set that is equal to the training set then MML cannot be interpreted as a predictive quality measure, but only as a goodness of fit metric. The MML definition from [3] is:

$$MML = \frac{1}{n}\sum_i \left[Y_i \log(p_i) + (1 - Y_i)\log(1 - p_i)\right]$$

Appendix C: Research plots and results

C1: Exploratory analysis plots

We depict in this appendix plots that are very helpful to visualize the shape of the possible distribution and outliers. We have used these plots in our exploratory data analysis (Chapter 4.3.2). We have examined four plots for each continuous variable.

Histogram
displays the frequency of a variable in certain value classes, which gives us a rough indication of the distribution.

Box plot
represents a graphical sketch of the numerical statistics. The solid box depicts the data between the 1st and the 3rd quartile (the inter-quartile range), displaying 50% of the data. The white line within the box stands for the median. The so-called whiskers embody the boundaries of the box plot; data points outside these limits are often considered as outliers.

Density plot
depicts a smooth estimate of the distribution or density. This estimate is based on sub parts of the values of variables. We have used two times the inter-quartile range as cut off points, which leads to a fairly smooth estimate.

QQ plot
tests the points of the variable against the normal distribution. If this plot indicates a straight line, then we have an indication of a normal distribution.

The rest of this section contains confidential information.

C2: Logistic modelling results

This section is not published due to confidential information.

C3: Quality measures of logistic models

Classification statistics

This section summarises numerical statistics that depict the classification quality of our logistic models. We refer to Chapter 4.3.4 for an extensive explanation on these statistics. Tables C3.1 and C3.2 show us that more complex models perform better and budget and functionality models perform as best. We also see that the budget models are the most stable models. These models perform equal or even better on subsets than on all data according to Tables C3.3 and C3.4 (See bold values).

Classification statistics

	Model1	Model2	Model3	Model4	Random
Budget	53.8%	60.1%	69.6%	71.7%	30.9%
Duration	58.5%	58.5%	58.5%	80.9%	41.2%
Functionality	60.4%	64.0%	69.9%	87.9%	27.3%
Overall	78.2%	78.2%	81.4%	91.0%	66.1%

Table C3.1: Eleven-point average recall of risk models on all data

	Model1	Model2	Model3	Model4
Budget	1.74	1.94	2.25	2.32
Duration	1.42	1.42	1.42	1.96
Functionality	2.21	2.34	2.56	3.22
Overall	1.18	1.18	1.23	1.38

Table C3.2: Relative lift factor (recall / random) of risk models on all data

	All data	Outlier Set	BP set	DP set	PP set	BS set
Budget model1	53.80%	50.10%	**59.60%**	52.60%	**56.30%**	50.10%
Budget model2	60.10%	58.10%	56.50%	50.40%	**61.10%**	48.00%
Budget model3	69.60%	66.00%	59.40%	56.40%	58.90%	49.90%
Budget Random	30.90%	29.94%	30.00%	26.77%	30.77%	26.23%
Func. model1	60.40%	58.20%	53.30%	49.61%	**62.00%**	60.70%
Func. model2	64.00%	60.90%	54.00%	50.20%	62.10%	58.40%
Func. model3	69.90%	66.70%	59.20%	59.40%	63.60%	66.00%
Func. Random	27.30%	27.39%	26.15%	25.20%	28.46%	27.05%

Table C3.3: Eleven-point average recall of the budget and func. models on subsets

	All data	Outlier Set	BP set	DP set	PP set	BS set
Budget model1	1.74	1.67	**1.99**	**1.96**	**1.83**	**1.91**
Budget model2	1.94	**1.94**	1.88	1.88	**1.99**	1.83
Budget model3	2.25	2.20	1.98	2.11	1.91	1.90
Func. model1	2.21	2.12	2.04	1.97	2.18	**2.24**
Func. model2	2.34	2.22	2.06	1.99	2.18	2.16
Func. model3	2.56	2.44	2.26	2.36	2.23	2.44

Table C3.4: Relative lift factors of the budget and func. models on subsets

Lift charts

We can visualise the classification quality of our models with so-called lift charts. We again refer to Chapter 4.3.4 for a general explanation on these lift charts.

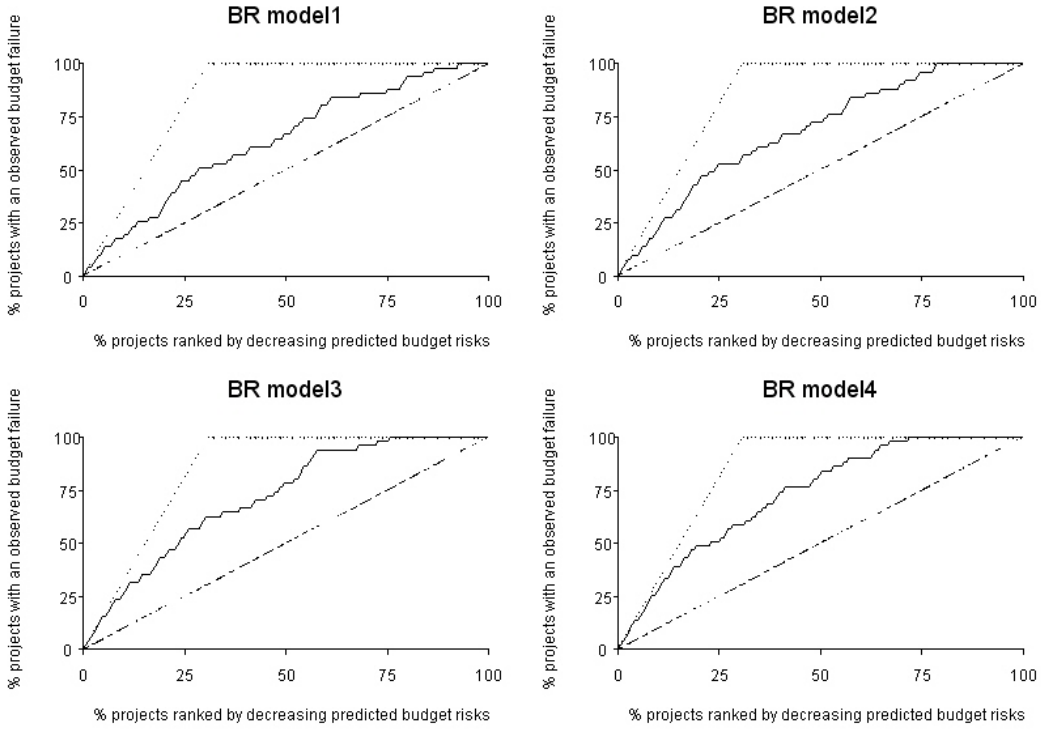


Figure C3.5: Lift charts of different budget risk models

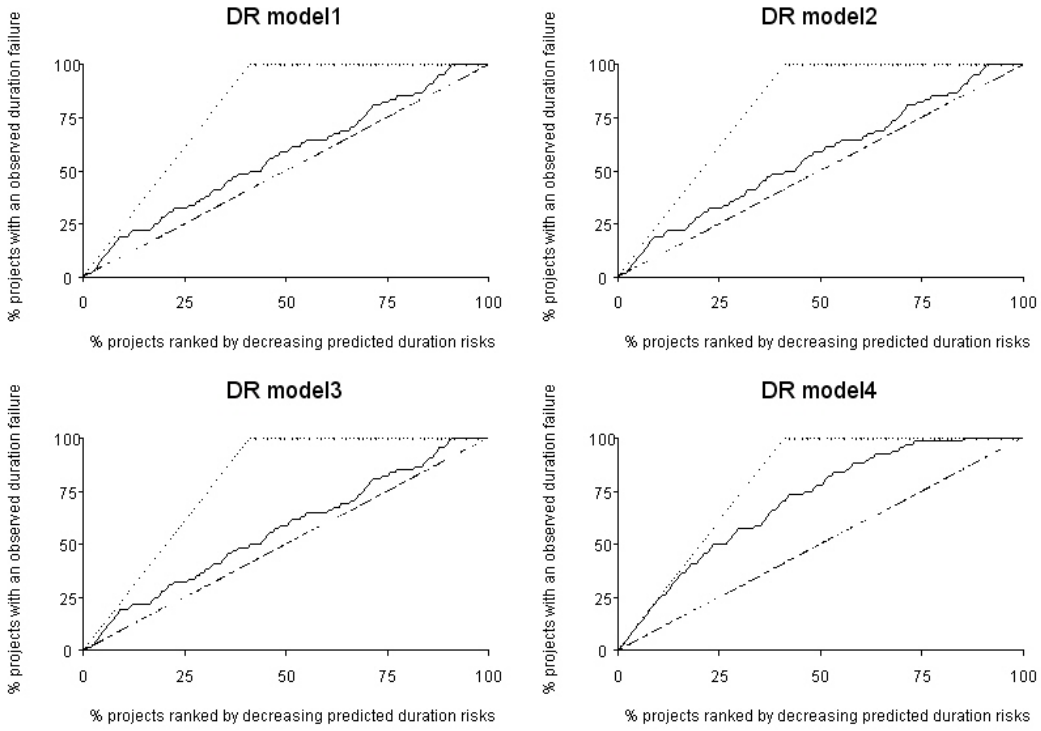


Figure C3.6: Lift charts of different duration risk models

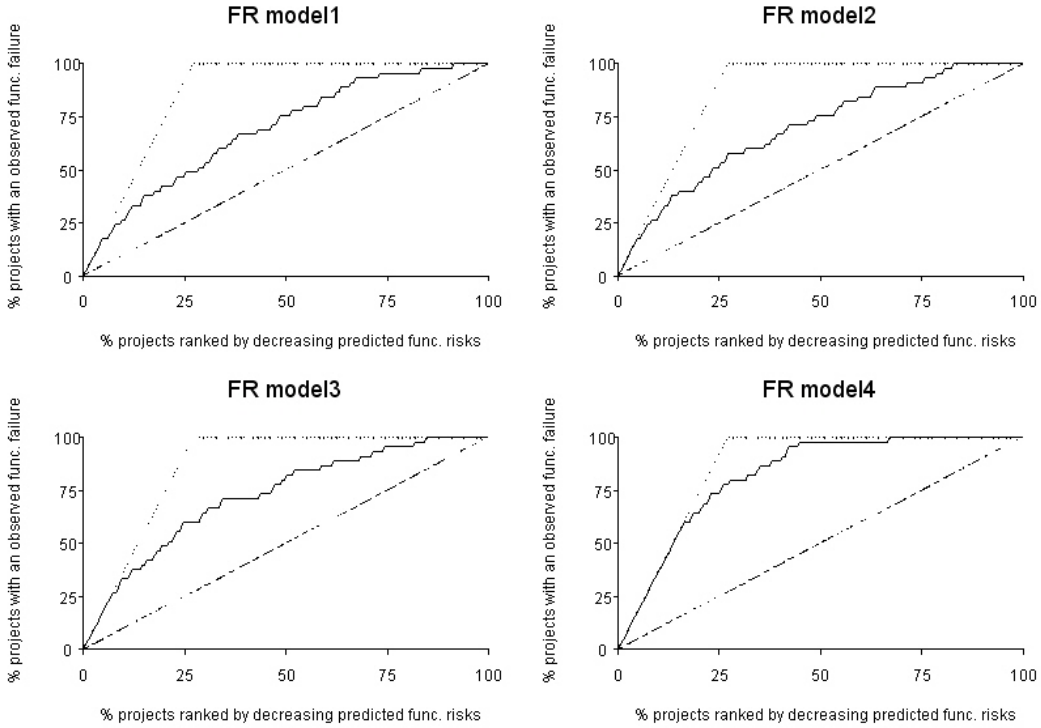


Figure C3.7: Lift charts of different functionality risk models

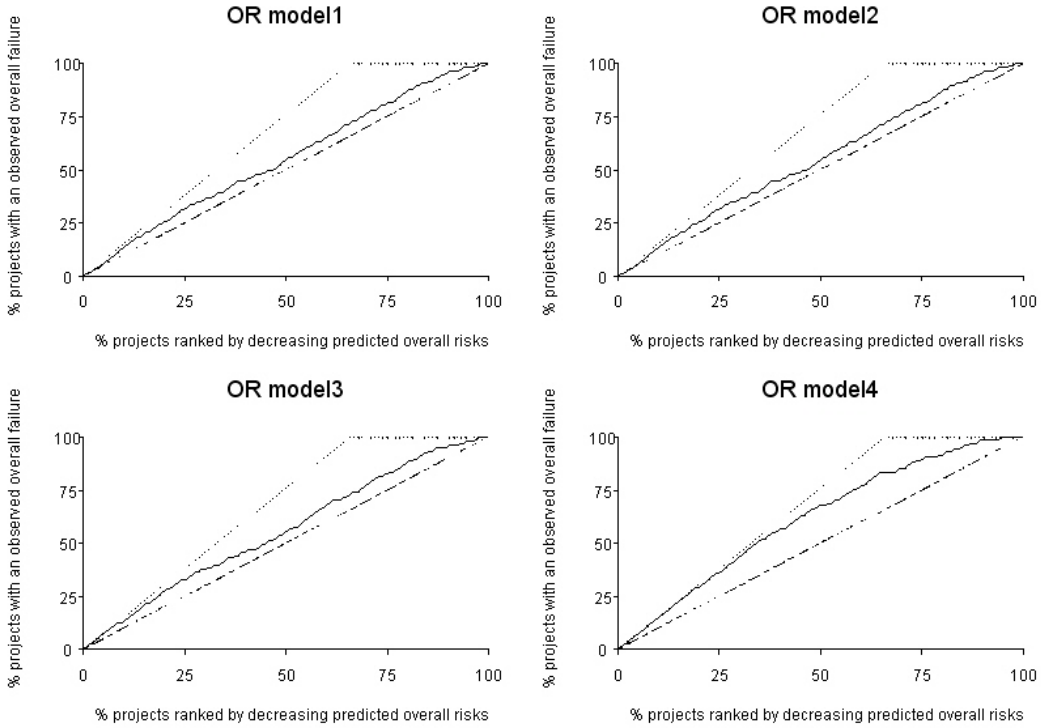


Figure C3.8: Lift charts of different overall risk models