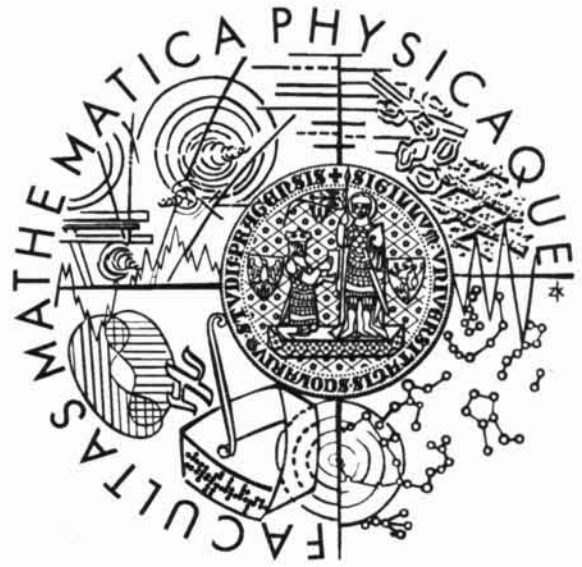
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Tomáš Magyar

Modelování pravděpodobností úvěrového selhání v rámci zobecněných lineárních modelů

Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: RNDr. Aleš Slabý, PhD.

Studijní program: Matematika
Studijní obor: Pravděpodobnost, matematická statistika a ekonometrie

Rád bych poděkoval svému vedoucímu diplomové práce Aleši Slabému, za obětavou pomoc a cenné připomínky, kterými přispěl ke vzniku této práce. Práce pod jeho vedením byla radostí a inspirací pro další profesní rozvoj.

Dále děkuji Mirce za její trpělivost a především svým rodičům za laskavou všestrannou podporu, kterou mi během studia poskytovali.

Prohlašuji, že jsem svou diplomovou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce.

V Praze dne 10. srpna 2005 Tomáš Magyar

Contents

Název práce: Modelování pravděpodobností úvěrového selhání v rámci zobecněných lineárních modelů
Autor: Tomáš Magyar
Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: RNDr. Aleš Slabý, PhD.
e-mail vedoucího: ales_slaby@kb.cz
Abstrakt: Statistické metody pro odhad kreditního rizika protistran se staly standardním nástrojem bank a finančních institucí při posuzování kreditního zdraví stávajících i budoucích klientů. Tato práce se zabývá statistickými metodami vývoje modelů pro odhad pravděpodobnosti úvěrového selhání protistran. První část práce je soustředěna na budování statistického modelu v kontextu zobecněných lineárních modelů. V druhé části se zabýváme testováním výkonnosti modelů a jejich validací. Výkonnost modelů je posuzována na základě ROC křivek a statistik z nich odvozených. Hlavním cílem této práce byl vývoj metodologie, jejíž použití vede k vývoji vhodných statistických modelů pro odhad pravděpodobnosti selhání protistran.
Klíčová slova: skóringový model, zobecněný lineární model, logistická regrese, validace, ROC křivka, AUC statistika

Title: Modelling of the probability of default in the context of Generalized Linear Models
Author: Tomáš Magyar
Department: Department of Probability and Mathematical Statistics
Supervisor: RNDr. Aleš Slabý, PhD.
Supervisor's e-mail address: ales_slaby@kb.cz
Abstract: Statistical methods for assessment of the counterparty's credit risk characteristics became standard tools of banks and other financial institutions, when estimating whether an applicant for credit will pay back his liabilities. This thesis presents the statistical methodology of credit scoring model development. The first part of the thesis focuses on statistical techniques employed in the model development process. The model development process is based on the class of models referred to as generalized linear models. The second part describes statistical methods of testing model performance and model validation. We investigate the model performance on the ground of receiver operating characteristics curves and the related summary statistic area under the receiver operating characteristics. The main purpose of the thesis was to develop a methodology, whose application leads to reasonable statistical models for assessment and quantification of counterparty's risk probability.
Keywords: scoring model, generalized linear model, logistic regression, validation, ROC curve, AUC statistic

Chapter I

Introduction

1. Legal framework and practice

The ongoing development of contemporary risk management methods and the increased use of innovative financial products have brought about substantial changes in the business environment faced by credit institutions nowadays. The Basel Committee on Banking Supervision, established in the end of 1974, represents an institution which formulates broad supervisory standards and guidelines and recommends statements of best practice in expectation that individual authorities (central-banks) will take steps to implement them to their own national systems. In 1988, the Committee decided to introduce a capital measurement system commonly referred to as the Basel Capital Accord. In June 1999, the Committee issued a proposal for a New Capital Adequacy Framework to replace the 1988 Accord. Following extensive interaction with banks the revised framework was issued on June 26, 2004 under the name Basel II Capital Accord. The Basel II Capital Accord is legally underpinned by the Capital Adequacy Directive (12/2000), issued by the European parliament and the Council. The new Basel II capital Accord demands a lot of attention both from regulators and regulated subjects. Among various innovations a new internal rating based approach (IRB), determining the capital requirements in the area of credit risk, was proposed.

One of the Committee's goals in setting forward an IRB approach is to align more precisely capital requirements with the intrinsic amount of credit risk to which banks are exposed. The orientation of the IRB approach is consistent with the framework currently being used by many banks with well-developed risk management systems to assess internally both their credit risk profile and their capital adequacy.

The Committee believes that such an approach, which relies heavily upon bank's internal quantitative and qualitative assessment of its counterparties and exposures, can better secure key objectives consistent with wider risk management practice.

In order to comply with the recommendations of the Basel II Capital Accord, each bank is required to estimate its set of probabilities of default (PD) related to its lending policy in each specific portfolio segment. To be more specific, bank's internal measures of credit risk are based on assessments of risk characteristics of both the borrower and the transaction. Most banks orient their borrower rating methodologies and risk management practices to the risk of borrower's default. The PD of the borrower or a group of borrowers is the central concept on which the IRB approach is built. The PD of the borrower does not, however, provide the complete picture of the potential credit loss. Banks also seek to measure how much they will lose, should the borrower default on an obligation. This is contingent upon two elements.

First, the loss is contingent upon the amount to which the bank was exposed to the borrower at the time of default, commonly expressed as Exposure at Default (EAD). Second, the magnitude of likely loss on the exposure referred to as the

Loss Given Default (LGD), which is expressed as a percentage of the exposure. For the sake of completeness, note that the IRB approach also takes into account the effective maturity (M) of exposures. These components (PD, EAD, LGD, M) form the basic inputs to the IRB approach, thus they must be assessed and estimated accurately, starting with the basic quantity, which is the PD. For this reason banks apply several sophisticated statistical methods for classifying their potential clients into certain rating categories and estimating the probabilities of default in these categories. We refer to this prescribed estimation process as to the process of *credit scoring*. In our context the process of credit scoring refers to statistical methods used to develop a statistical model for estimating and predicting the probability that a loan applicant or an existing obligor will default or become deliquent. The final scoring tool is called the credit scoring model. To build a credit scoring model, statisticians analyze historical data on the performance of provided loans to determine which of the borrower's characteristics are useful in predicting whether the loan performed well.

To be more precise the bank has several information related to the creditworthiness of its potential clients or obligors. This information might be encoded in several characteristics which depend on the actual commercial area the scoring model is build for.

2. The general model

From the mathematical point of view the basic setup can be expressed in the following way: we have n observations $\boldsymbol{x}_i^\top = (x_{i1}, \ldots, x_{im}), i = 1, \ldots, n$, of a random vector in $\boldsymbol{X}$ in $\mathbb{R}^m$. That is, there are m explanatory (independent) variables $X_1, \ldots, X_m$ referred to as predictors. Further we have a dependent (response) random variable Y. The observations can be expressed as a row vector $(y_i, \boldsymbol{x}_i^\top), i = 1, \ldots, n$. Thus, x_{ij} is the value of the j-th predictor $j = 1, \ldots, m$ of the i-th customer $i = 1, \ldots, n$. Similarly y_i is the value of realization of random variable Y. It has two values coded by 1 and 0 (default and non-default), respectively. The actual data set can be expressed in the matrix form as

$$\mathcal{X} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,m} \\ \vdots & & \vdots \\ x_{n,1} & \cdots & x_{n,m} \end{pmatrix} \quad \boldsymbol{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Statisticians make a sample of its past debtors and analyze their characteristics. Practically, the selected sample is divided into two subsamples. The first, called the *train sample*, is used for model development. The second, called the *test sample*, is used for model testing and validation. The scoring model is generated from this input data through various approaches and then it is applied to new clients in order to estimate their expected probability of default. Obviously, the purpose of credit scoring is highly practical, it provides the bank with better knowledge of its clients. It effectively helps to manage and reduce credit risk. In other words, statistical techniques used for credit scoring are based on the idea of discrimination between several subgroups in the underlying population of bank clients. The goal is

to develop statistical method which is sufficiently sensitive, quick in computation, transparent and easy to interpret.

This thesis introduces the background knowledge of credit scoring in the context of generalized linear models (GLM), in particular logistic regression. The statistical aspect of credit scoring methodology are discussed. Emphasis is put on statistical techniques and statistical computing employed in credit scoring model development and validation procedures. The thesis resolves statistical issues connected to the second and third stage of the credit scoring process illustrated in Figure 1.1.

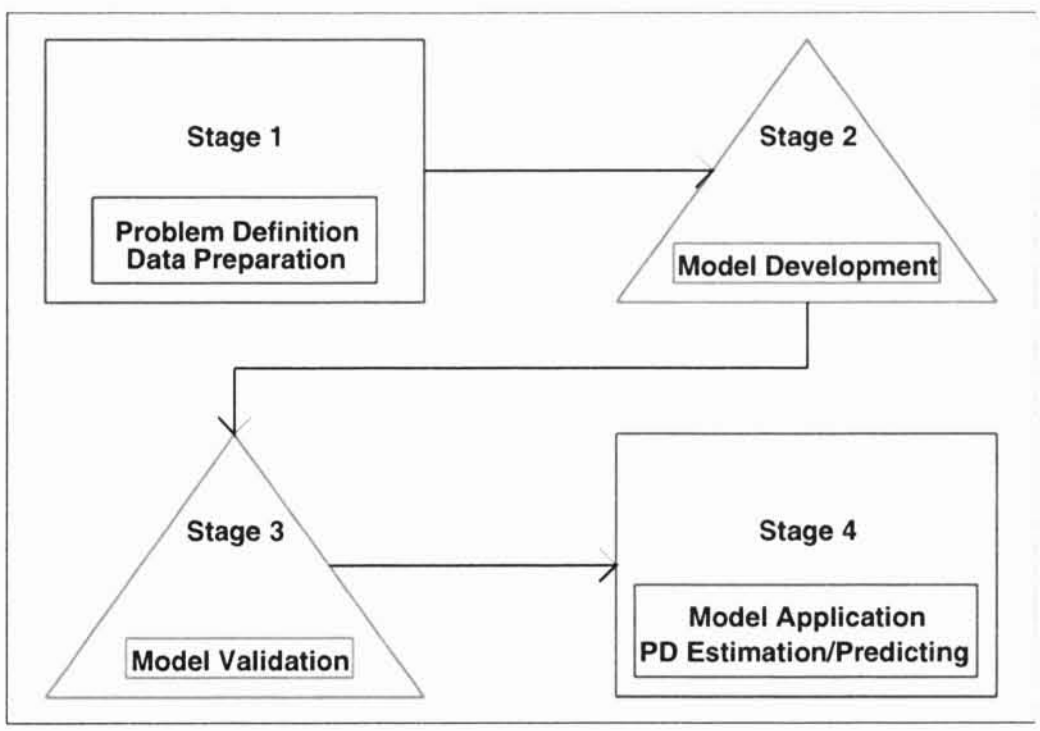


Figure 1.1. *The process of credit scoring.*

3. Data sources

Data sets employed in the above presented stages of the credit scoring process stem from a Czech bank. A special database has been developed to store and process detailed quantitative and qualitative data. The data was collected by means of electronic forms filled by the branch network. However, additional information is confidential and therefore names of all variables used in analytical examples, as well as in illustrative figures have been removed. With regard to internal confidentiality and data privacy protection we do not present final results, that is, concrete models. Instead we present statistical methodology, whose application leads to reasonable statistical models in similar situations we faced. For the sake of completeness, note that we examined the segment of Corporate/SME firms and the considered default horizon was one year.

CHAPTER II

MODEL DEVELOPMENT TECHNIQUES

1. Single - factor analysis

A common characteristic of credit scoring models based on information from financial statements, is a large number of independent variables that can be used in the model development phase. It is not so complicated to define a huge amount of financial ratios, combining all the useful information contained in the financial statements of a company in very different ways to assess its credit worthiness. The way this information is employed to build the model is crucial in determining the capability and robustness of the final model in predicting default. Actually, some of the financial ratios that can be derived, might be useful to predict default, but others might not be related to the default variable at all. Furthermore, some of the ratios can take extremely high or low values for some clients, without serving any information for default prediction purposes. These facts highlight the importance of variable selection and transformation processes that are performed during the single-factor analysis phase.

Single-factor analysis is the first step in statistical part of building a credit scoring model. The aim of the single-factor analysis is to prepare a reasonable set of default predictors that can be used later in multi-factor analysis. Given a large amount of possible predictors, it is important to reduce this list to predictors that enter the final model selection process. In order to understand the reason why predictors are treated separately, we should be aware of the fact that modelling database contain raw data.

There are several problems that has to be solved within the single-factor analysis before any multi factor analysis can be performed.

Statisticians distinguish two types of predictor variables. Namely the *categorial* and *continuous* predictors. The nature of the predictor variables is different and so are the problems that statisticians need to solve. Dealing with categorial predictors involves the following issues.

- Order of predictor categories (in terms of expected default frequency) need not be completely clear beforehand. Definitely, we should always assess what is the position of missing values (NA) within this order. It is convenient to assign categories certain numeric levels beforehand to optimize statistical performance of the predictor in multi-factor analysis.
- There may be strong dependence between categorial predictors, but their vague and subjective definition can hide it. Thorough investigation has to be undertaken to uncover possible dependence.
- Predictive power of categorial predictors can be rather volatile in time because of subjective nature of assignment of obligors to categories. Regular validation of particular categorial predictors is necessary.

Dealing with continuous predictors involves the following issues.

- Continuous predictors have to be tested for outliers because outliers can significantly disturb predictive power of single predictor as well as its contribution to the multi-factor model.
- Certain transformation or truncation of the continuous predictor may be desirable to optimize its statistical performance in multi-factor analysis.
- Strong dependence among continuous predictors is typical in credit modelling. Usually, one has a group of predictors describing the same or very similar things in slightly different manner at hand. Within such a group only one or two predictors are convenient to be considered in multi-factor analysis. Inclusion of the whole group is counterproductive.
- Treatment of missing values.

1.1. Categorial predictor analysis

Procedures concerning categorial predictors are relatively easy to get along with, because the single-factor analysis of categorial predictors is simpler than the one of continuous predictors and the results are easier to interpret. Nevertheless the following things have to be done. We have to asses the proper order of predictor categories. As soon as it is done, joining of non-significant categories is carried out. Afterwards we might proceed with reducing the predictor set.

1.1.1. Ordering predictor categories. Assessing the proper order of predictor categories including the position of NA is primary problem to get along with. We solve this task by assigning all categories certain numeric levels, respecting the following property: The larger is the numeric level the larger is the corresponding expected probability of default. Beside proper ordering, we require that the levels should translate the predictor into a linear world of logistic regression model in order to make it optimized for further use in multi-factor analysis. The level assignment is carried out by fitting a one-dimensional logistic additive model for each categorial predictor X as follows

$$\log\left\{\frac{p}{1-p}\right\} = f(X). \tag{2.1}$$

Model 2.1 is a special case of the general multi factor model A.8 described in Appendix A, where we set $m = 1$ and $\alpha = 0$. In this application it is useful to think of the generalized additive model as of a method for estimating the appropriate metameter in which to measure the variables. It follows that this way we ensure both the proper ordering and the translation into the linear world.

Thus, we assume that some predictor X can attain S abstract values $\mathcal{C}_1, \ldots, \mathcal{C}_S$ and refer to them as categories. In view of 2.1 and Appendix A we seek for an assignment f such that

$$f(\mathcal{C}_s) = L_s, \quad s = 1, \ldots, S. \tag{2.2}$$

where $L_1, \ldots, L_S$ are arbitrary real numbers. These numbers are estimated via a so called *local-scoring algorithm*, computational details are given in Hastie and

Tibshirani (1997). In order to work this well data sample size has to be significantly larger than S. Figure 2.1 displays a possible level assignment carried out on real data.

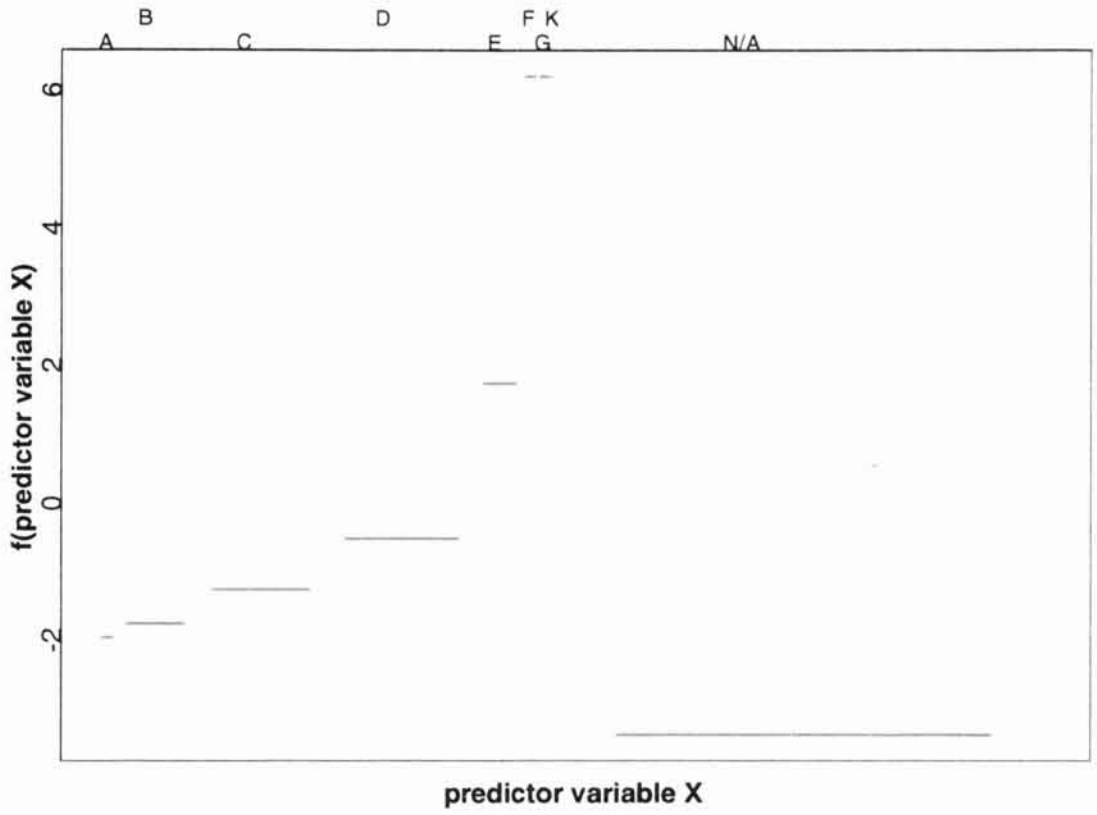


Figure 2.1. *A possible level assignment carried out on qualitative predictor variable X, which has 9 levels including the category of NA. The length of each line represents to the proportion of observations having the corresponding level outlined above this line on the top horizontal axis.*

Since the above described approach is quite general, we might want to specify the assumptions concerning function f from 2.1 when choosing the method of estimating and quantifying the appropriate ordering and significance of predictor categories. In this sense the univariate generalized additive model is equivalent to fitting an univariate general linear model of the form

$$\log\left\{\frac{p}{1-p}\right\} = \sum_{s=1}^{S} L_s I\{X = \mathcal{C}_s\} \tag{2.3}$$

where $I\{X = C_s\}$ is the indicator of the fact that predictor X assumes category $\mathcal{C}_s$. These two methods are intended to be equivalent from the the point of view of proper ordering of predictor categories, nevertheless the fitting procedures related to the two model types are iterative and might differ in implementation in different software packages. As a result the corresponding estimates might slightly differ in absolute values as well. Because X can always assume just one of the categories the univariate model 2.3 can be viewed as a special S-dimensional (multivariate) model.

An important features of this model is that regressors $I\{X = C_1\}, \ldots, I\{X = C_S\}$ are perfectly orthogonal. Moreover, if we add a constant to the model we

obtain a singular regression matrix, thus another benefit is the absence of the intercept.

Assume we fitted model 2.3 and obtained estimates $\widehat{L}_1, \ldots, \widehat{L}_S$ of coefficients $L_1, \ldots, L_S$, respectively. Let $\widehat{L}_{(1)}, \ldots, \widehat{L}_{(S)}$ be estimates $\widehat{L}_1, \ldots, \widehat{L}_S$ ordered in ascending order. This ordering provides ordering of obligors by their estimated probabilities of default and suggests order of categories $\mathcal{C}_1, \ldots, \mathcal{C}_S$ in these terms. Finally, denote by $\mathcal{C}_{(1)}, \ldots, \mathcal{C}_{(S)}$ the suggested order of categories $\mathcal{C}_1, \ldots, \mathcal{C}_S$.

Once we obtain optimal level assignment for all categorical predictors we test, whether two subsequent categories differ significantly or not, the corresponding coefficients are compared. Afterwards, it might be useful to joint the non-significant categories together.

1.1.2. Jointing predictor categories. Realize that, with the absence of the intercept, tests whether $L_s = 0$ does not have much meaning. Moreover, $L_s = 0$ does not mean that category $\mathcal{C}_s$ is not statistically significant. Here, significance should rather be considered in terms of the relative number of observations of category $\mathcal{C}_s$ and by mutual comparison of levels L_s. This comparison is crucial for jointing categories.

Let us introduce a set of orthogonal vectors $\boldsymbol{D_k}, k = 1, \ldots, S-1$, each of length S having the following property. The k-th element of vector $\boldsymbol{D_k}$ equals 1, while the $k+1$-th element equals -1, the other elements of this vector equal 0. We refer to them as *contrasts*. Further, denote the vector of the ordered estimates of coefficients $L_1, \ldots, L_S$ by $\widehat{\boldsymbol{L}}_{()}$, thus $\widehat{\boldsymbol{L}}_{()}^{\top} = (\widehat{L}_{(1)}, \ldots, \widehat{L}_{(S)})$. For an arbitrary k we have that $\boldsymbol{D_k}^{\top}\widehat{\boldsymbol{L}}_{()} = \widehat{L}_{(k+1)} - \widehat{L}_{(k)}$. Employing the introduces notation, we are able to describe the problem of testing, whether two subsequent categories differ significantly, with a sequence of null hypotheses as follows

$$\begin{aligned}
&\mathrm{H}_{0,1}\colon\ \boldsymbol{D_1}^{\top}\widehat{\boldsymbol{L}}_{()} = \widehat{L}_{(2)} - \widehat{L}_{(1)} = 0\\
&\vdots\\
&\mathrm{H}_{0,k}\colon\ \boldsymbol{D_k}^{\top}\widehat{\boldsymbol{L}}_{()} = \widehat{L}_{(k+1)} - \widehat{L}_{(k)} = 0\\
&\vdots\\
&\mathrm{H}_{0,S-1}\colon\ \boldsymbol{D_{S-1}}^{\top}\widehat{\boldsymbol{L}}_{()} = \widehat{L}_{(S)} - \widehat{L}_{(S-1)} = 0
\end{aligned}$$

Recall that in ordinary linear regression we are used to employ the t-statistic to check whether two coefficients are significantly different. In the context of generalized linear models, in particular logistic regression, this statistic does not follow t-distribution. Nevertheless, by Wald's (1943) results for maximum likelihood estimators (ML), having enough data one can safely use asymptotic approximation by the standard normal distribution, for the test statistics T_k, corresponding to the k-th hypothesis, which reads

$$T_k = \frac{\widehat{L}_{(k+1)} - \widehat{L}_{(k)}}{\sigma\big(\widehat{L}_{(k+1)} - \widehat{L}_{(k)}\big)} = \frac{\widehat{L}_{(k+1)} - \widehat{L}_{(k)}}{\sqrt{\sigma^2\big(\widehat{L}_{(k+1)}\big) + \sigma^2\big((\widehat{L}_{(k)}\big)}}, \quad k = 1, \ldots, S-1, \tag{2.4}$$

where σ denotes the standard error. Equation 2.4 holds because all coefficient estimates are uncorrelated since linear regressors $I\{X = L_1\}, \ldots, I\{X = L_S\}$ are orthogonal. Realize that T_k is large if the difference $\widehat{L}_{(k+1)} - \widehat{L}_{(k)}$ is large and if standard error $\sigma(\widehat{L}_{(k+1)} - \widehat{L}_{(k)})$ is small. The standard error is small if particular standard errors $\sigma(\widehat{L}_{(k+1)})$ and $\sigma(\widehat{L}_{(k+1)})$ are small, thus, if estimated coefficients are not inaccurate. In our case, inaccuracy is mainly caused by small number of observations in the considered category.

Note that because of the ascending ordering of coefficients it should be enough to test the null hypotheses against a one-sided alternatives, however, we discuss also the case of two-sided alternative. With respect to the large-sample normality of ML estimators, we compare T_k with the appropriate critical values of standard normal distribution to obtain test results of one- or two-sided alternatives. Equivalently, for the two-sided alternative, admitting that T_k^2 has asymptotically the χ^2 distribution with one degree of freedom, critical values of $\chi^2(1)$ distribution can be employed.

Set the confidence level to be α, $0 < \alpha < 1$. If $T_k > \Phi^{-1}(1-\alpha)$, we reject the k-th null hypothesis and categories $\mathcal{C}_k$ and $\mathcal{C}_{k+1}$ are considered to be significantly different and no jointing is committed. On the other hand, $T_k \leq \Phi^{-1}(1-\alpha)$ does not necessarily imply that categories $\mathcal{C}_k$ and $\mathcal{C}_{k+1}$ should be joined as it is outlined in the following paragraph.

It often happens that a sequence of non-significant differences is encountered. Let us consider three (ordered) categories $\mathcal{C}_{(k)}$, $\mathcal{C}_{(k+1)}$ and $\mathcal{C}_{(k+2)}$ whose corresponding subsequent estimated coefficients $\widehat{L}_{(k)}$, $\widehat{L}_{(k+1)}$ and $\widehat{L}_{(k+2)}$ do not differ significantly. If $\widehat{L}_{(k)}$ and $\widehat{L}_{(k+2)}$ differ significantly we cannot reason jointing all three categories in one even if the neighbouring couples do not significantly differ. In these cases we have to choose which couple should be joined. However, the new joint category need not be significantly different from the left one after jointing because the new level coefficient of the joint category shifts towards the coefficient of the left category. Such situations might occur if the statistic T_k for $\mathcal{C}_{(k)}$ and $\mathcal{C}_{(k+2)}$ is not highly above quantile $\Phi^{-1}(1-\alpha)$.

The above prescribed situations suggest, that beside statistics 2.4 it might be useful to calculate $T_k^{(2)}$ statistics for the second neighbours, namely,

$$T_k^{(2)} = \frac{\widehat{L}_{(k+2)} - \widehat{L}_{(k)}}{\sqrt{\sigma^2\big(\widehat{L}_{(k+2)}\big) + \sigma^2\big(\widehat{L}_{(k)}\big)}}, \quad k = 1, \ldots, S-2, \tag{2.5}$$

Statistics 2.4 and 2.5 are usually sufficient for jointing categories, but we could compute statistics $T_k^{(m)}$ for $m > 2$ if necessary.

In practical applications expert opinion should also be taken into account, especially in those situations, when statistically suggested order of categories is different from expert expectation, and in the same time, when two categories which does not differ significantly must not be joined at any case.

1.1.3. Treatment of missing values. Missing values of categorial predictors are treated as a category referred to as NA. The level assignment and jointing

described above fully applies to this category. Generally, it might happen that for certain reason the NA category improves the predictive power, or its order among the other categories is somehow suspicious. These cases require special treatment if any. Usually it is recommended to exclude these predictors from the modeling database, that's why we do not examine this issue further.

1.1.4. Reduction of the predictor set. Referring to our previous comments, we outline again that the aim of the single-factor analysis is to prepare a reasonable set of default predictors that can be used later in multi-factor analysis. Because of the fact that at the beginning we are given a large amount of possible predictors, it is important to reduce this list of predictors in order to obtain the most reasonable ones, which can be finally used in the multi-factor analysis. When assessing the appropriate criteria for reduction of the predictors set, two kind of characteristics are taken into account. First, we would like to evaluate the discriminative ability of each predictor with respect to separation between defaulting and non-defaulting obligors. Discriminative characteristics, namely, the *receiver operating characteristics curve* and the related area under this curve are employed to exclude predictors with none or low discriminative power. Second, we desire to choose predictors that are not probability dependent in a significant way. The reason for this is obvious, strong dependence among predictors causes serious problems within the multivariate analysis as the estimates of coefficients are inaccurate and they can also have different signs than expected. It is advised to choose only one, at most two, representatives out of group of strongly dependent predictors.

Discriminative ability. The assessment of the discriminative ability of a specific predictor is accomplished by using the receiver operating characteristics (ROC curve) and computing its appropriate summary statistics, the area under the ROC curve – AUC. In this context ROC analysis represents an overall measures for assessing the amount of information included in the underlying predictor regarding its ability to discriminate between good and bad cases. Because of the fact that the major of Chapter III is dedicated to these issues we do not expand this problematic here. Instead, we focus on the dependence structure that determines the final set of predictors used in multi factor analysis.

Dependence structure. The dependence structure is a fundamental issue of each statistical analysis. No model can be contemplated without making some assumptions about dependence structure of elements involved. In credit scoring models we are strongly interested in dependence among predictors. Dependence structure that joins marginal distributions to the joint distribution is fully described by the copula function. Nevertheless, simpler tools which characterize grade of dependence instead of the copula are often used. These tools are usually called measures of association. Note, that there is a specific subgroup of measures of association which are called measures of dependence. There are several properties of these measures that are important for practical application.

Assume we have two random variables X_1 and X_2 whose dependence is measured by a measure of dependence $\delta(X_1, X_2)$. Realize that according to Nelsen (1998) and Slabý (2004), the following properties are required in any case:

(1) The measure of dependence is symmetric: $\delta(X_1, X_2) = \delta(X_2, X_1)$.
(2) The measure of dependence is bounded by 0 and 1: $0 \leq \delta(X_1, X_2) \leq 1$.
(3) The measure of dependence distinguish independence: $\delta(X_1, X_2) = 0$ if and only if X_1 and X_2 are (probability) independent.
(4) The measure of dependence distinguish certain kind of perfect dependence:
   - For continuous X_1 and X_2, $\delta(X_1, X_2) = 1$ if and only if each of X_1 and X_2 is a strictly monotonous function of the other.
   - For categorial X_1 and X_2 with S_1 and S_2 categories, say $S_1 \leq S_2$, $\delta(X_1, X_2) = 1$ if and only if there is a one-to-one correspondence between categories of X_1 and an S_1-element subset of categories of X_2 where each of possible $S_2 - S_1$ left categories of X_2 occurs merely with one of the categories of X_1.
(5) The measure of dependence is invariant under certain transformations:
   - For continuous X_1 and X_2, $\delta(g(X_1), h(X_2)) = \delta(X_1, X_2)$ whenever g and h are strictly monotonous functions.
   - For categorial X_1 and X_3, $\delta\big(g(X_1), h(X_2)\big) = \delta(X_1, X_2)$ whenever in this case $g(X_1)$ and $h(X_2)$ arise by permutation of categories of X_1 and X_2, respectively.

Measures of association are less restricted than measures of dependence since they are designed to measure special types of dependence. For example, equivalence in point (3) of the above prescribed list does not hold, an implication is satisfactory. Similarly, the lower bound in point (2) could be different too, it often equals to -1.

Further, we introduce a measures of dependence that we derived using the standard Pearson χ^2 statistics.

Assume again that we have two categorial predictors X_1 and X_2 which attain S_1 and S_2 abstract values (categories) $\mathcal{C}_{11}, \ldots, \mathcal{C}_{1S_1}$ and $\mathcal{C}_{21}, \ldots, \mathcal{C}_{2S_2}$, respectively. Denote

$$p_{ij} = P(X_1 = \mathcal{C}_{1i}, X_2 = \mathcal{C}_{2j}) \quad i = 1, \ldots, S_1, \quad j = 1, \ldots, S_2$$

the joint probabilities of attaining categories $\mathcal{C}_{1i}$, $\mathcal{C}_{1j}$ and

$$p_{i\cdot} = P(X_1 = \mathcal{C}_{1i}) = \sum_{j=1}^{S_2} p_{ij},$$

$$p_{\cdot j} = P(X_2 = \mathcal{C}_{2j}) = \sum_{i=1}^{S_1} p_{ij},$$

the marginal probabilities. The matrix (p_{ij}) is usually called as the *matrix of probabilities*. By definition, predictors X_1 and X_2 are (probability) independent if $p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$.

Conversely, for $S_1 = S_2$ predictors are perfectly dependent when categories $\mathcal{C}_{11}, \ldots, \mathcal{C}_{1S_1}$ form faithful pairs with categories $\mathcal{C}_{21}, \ldots, \mathcal{C}_{2S_2}$, that is, these pairs

always occur together. Formally speaking it means that for any fixed i or j we have just one $p_{ij} > 0$, and $p_{ij} = p_{i\cdot}$ or $p_{ij} = p_{\cdot j}$, respectively.

If numbers of categories differ, for example $S_2 > S_1$, then the notion of perfect dependence is slightly modified. In such a case there is an $S_1 \times S_1$-square sub matrix of (p_{ij}), which satisfies the above conditions while the columns left contain just one $p_{ij} > 0$. Informally speaking, all categories $\mathcal{C}_{11}, \ldots, \mathcal{C}_{1S_1}$ have faithful partners among categories $\mathcal{C}_{21}, \ldots, \mathcal{C}_{2S_2}$.

Now assume that we have n simultaneous observations of predictors X_1 and X_2. Let n_{ij} be the number of cases when X_1 assumes category $\mathcal{C}_{1i}$ and X_2 assumes category $\mathcal{C}_{2j}$. Further let $n_{i\cdot}$ be the number of cases when X_1 assumes category $\mathcal{C}_{1i}$ and $n_{\cdot j}$ the number of cases when X_2 assumes category $\mathcal{C}_{2j}$. Thus, we have

$$n_{i\cdot} = \sum_{j=1}^{S_2} n_{ij},$$

$$n_{\cdot j} = \sum_{i=1}^{S_1} n_{ij}, \tag{2.6}$$

$$n = \sum_{i=1}^{S_1} n_{i\cdot} = \sum_{j=1}^{S_2} n_{\cdot j}.$$

In statistical literature the matrix (n_{ij}) is called the *contingency table.*

The chi-square statistic χ^2 defined as

$$\chi^2 = \sum_{i=1}^{S_1} \sum_{j=1}^{S_2} \frac{\left(n_{ij} - \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot} \cdot n_{\cdot j}}{n}} \tag{2.7}$$

is the standard tool for testing independence in contingency tables. Formula 2.7 is actually not suitable for computation, so we try to derive an equivalent form which used to be employed while computation:

$$\begin{aligned}
\chi^2 &= \sum_{i=1}^{S_1} \sum_{j=1}^{S_2} \frac{\left(n_{ij} - \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot} \cdot n_{\cdot j}}{n}} = \frac{1}{n} \sum_{i=1}^{S_1} \sum_{j=1}^{S_2} \frac{(n_{ij} n^2 - 2n n_{ij} n_{i\cdot} n_{\cdot j} + n_{i\cdot}^2 n_{\cdot j}^2)}{n_{i\cdot} n_{\cdot j}} \\
&= n \sum_{i=1}^{S_1} \sum_{j=1}^{S_2} \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} - 2 \sum_{i=1}^{S_1} \sum_{j=1}^{S_2} n_{ij} + \frac{1}{n} \sum_{i=1}^{S_1} \sum_{j=1}^{S_2} n_{i\cdot} n_{\cdot j} \\
&= n \sum_{i=1}^{S_1} \sum_{j=1}^{S_2} \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} - 2n + n \\
&= n \sum_{i=1}^{S_1} \sum_{j=1}^{S_2} \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} - n. \qquad (2.8)
\end{aligned}$$

Note that large values of the χ^2 statistic suggest that hypothesis of independence does not hold. In practice, asymptotic critical values are usually used in tests as χ^2 statistic 2.7 has asymptotically $\chi^2_{(S_1-1)(S_2-1)}$ distribution with $(S_1-1)(S_2-1)$

degrees of freedom. It is known that this asymptotic approximation is plausible only in cases where $n_{i\cdot}n_{\cdot j}/n > 5$ for all combinations of i and j.

Now we employ the described χ^2 statistics to derive a theoretical measure of dependence that has similar properties to those we have listed, Slabý (2004). We follow those prescribed properties and we show that a theoretical measure of dependence between X_1 and X_2 defined as

$$\chi^2_T = \frac{1}{\min(S_1, S_2) - 1}\left(\sum_{i=1}^{S_1}\sum_{j=1}^{S_2}\frac{p_{ij}^2}{p_{i\cdot}\cdot p_{\cdot j}} - 1\right) \tag{2.9}$$

satisfies them. It is clear that χ^2_T is symmetric. Further we show that χ^2_T distinguishes perfect dependence if $\chi^2_T = 1$ and perfect independence if $\chi^2_T = 0$. In other words we need to prove that χ^2_T is bounded and the introduced standardization ensures that lies between 0 and 1.

First we consider that our two categorical predictors X_1, X_2 can assume S_1 and S_2 categories, where $S_1 = S_2$ and that predictors are perfectly dependent. Thus we are looking for a constant K such that the following holds

$$K\left(n\sum_{i=1}^{S_1}\sum_{j=1}^{S_2}\frac{p_{ij}^2}{p_{i\cdot}p_{\cdot j}} - n\right) = 1 \implies K = \frac{1}{\underbrace{n\left(\sum_{i=1}^{S_1}\sum_{j=1}^{S_2}\frac{p_{ij}^2}{p_{i\cdot}p_{\cdot j}} - 1\right)}_{T_1}} \tag{2.10}$$

$$\begin{aligned} T_1 &= \sum_{i=1}^{S_1}\left(\frac{p_{i1}^2}{p_{i\cdot}p_{\cdot 1}} + \cdots + \frac{p_{iS_2}^2}{p_{i\cdot}p_{\cdot S_2}}\right) - 1 = \underbrace{\left(\frac{p_{11}^2}{p_{1\cdot}p_{\cdot 1}} + \cdots + \frac{p_{1S_2}^2}{p_{1\cdot}p_{\cdot S_2}}\right)}_{=1} \\ &+ \underbrace{\left(\frac{p_{21}^2}{p_{2\cdot}p_{\cdot 1}} + \cdots + \frac{p_{2S_2}^2}{p_{2\cdot}p_{\cdot S_2}}\right)}_{=1} + \cdots + \underbrace{\left(\frac{p_{S_1 1}^2}{p_{S_1\cdot}p_{\cdot 1}} + \cdots + \frac{p_{S_1 S_2}^2}{p_{S_1\cdot}p_{\cdot S_2}}\right)}_{=1} - 1 \end{aligned}$$

$$= S_1 - 1 \Rightarrow K = \frac{1}{n(S_1 - 1)} \tag{2.11}$$

Second we consider that our two categorical predictors X_1, X_2 can assume S_1 and S_2 categories, where $S_1 < S_2$ and that predictors are perfectly dependent, in this case for the formula labelled as T_1 we find that

$$\begin{aligned} T_1 &= \sum_{i=1}^{S_1}\sum_{j=1}^{S_2}\frac{p_{ij}^2}{p_{i\cdot}p_{\cdot j}} - 1 = \sum_{i=1}^{S_1}\sum_{j=1}^{S_1+(S_2-S_1)}\frac{p_{ij}^2}{p_{i\cdot}p_{\cdot j}} - 1 \\ &= \sum_{i=1}^{S_1}\left(\frac{p_{i1}^2}{p_{i\cdot}p_{\cdot 1}} + \cdots + \frac{p_{iS_1}^2}{p_{i\cdot}p_{\cdot S_1}}\frac{p_{i(S_1+1)}^2}{p_{i\cdot}p_{\cdot(S_1+1)}} + \cdots + \frac{p_{i(S_1+(S_2-S_1))}^2}{p_{i\cdot}p_{\cdot(S_1+(S_2-S_1))}}\right) - 1 \end{aligned}$$

$$
\begin{aligned}
(2.12) \quad &= \underbrace{\left(\frac{p_{11}^2}{p_{1\cdot}p_{\cdot 1}} + \cdots + \frac{p_{1S_1}^2}{p_{1\cdot}p_{\cdot S_1}} \frac{p_{1(S_1+1)}^2}{p_{1\cdot}p_{\cdot(S_1+1)}} + \cdots + \frac{p_{1(S_1+(S_2-S_1))}^2}{p_{1\cdot}p_{\cdot(S_1+(S_2-S_1))}} \right)}_{=1} + \cdots \\
&\cdots + \underbrace{\left(\frac{p_{S_1 1}^2}{p_{S_1\cdot}p_{\cdot 1}} + \cdots + \frac{p_{S_1 S_1}^2}{p_{S_1\cdot}p_{\cdot S_1}} \frac{p_{S_1(S_1+1)}^2}{p_{S_1\cdot}p_{\cdot(S_1+1)}} + \cdots + \frac{p_{S_1(S_1+(S_2-S_1))}^2}{p_{S_1\cdot}p_{\cdot(S_1+(S_2-S_1))}} \right)}_{=1} - 1 \\
(2.13) \quad &= S_1 - 1.
\end{aligned}
$$

The terms in the above prescribed equation sum to one, thanks to the definition of the perfect dependence in the case where $S_1 < S_2$.

This can be easily presented on the next example. Consider two categorical predictor variables X_1 and X_2 such that X_1 assume three categories coded by integers 1 to 3, thus $S_1 = 3$. X_2 assumes six categories coded by integers 1 to 6, thus $S_2 = 6$. The variables has the following form

$$
\begin{aligned}
X_1 &= (1, 1, 2, 3, 3, 2, 1, 1, 2, 2, 3) \\
X_2 &= (1, 1, 5, 4, 3, 2, 1, 1, 2, 2, 6)
\end{aligned}
$$

The corresponding contingency table and the matrix of probabilities have the following form

3	0	0	1	1	0	1
2	0	3	0	0	1	0
1	4	0	0	0	0	0
Categories of X_1/X_2	**1**	**2**	**3**	**4**	**5**	**6**

Table 2.1. *Contingency table related to categorical predictors X_1, X_2.*

3	0	0	1/11	1/11	0	1/11
2	0	3/11	0	0	1/11	0
1	4/11	0	0	0	0	0
Categories of X_1/X_2	**1**	**2**	**3**	**4**	**5**	**6**

Table 2.2. *The matrix of probabilities related to categorical predictors X_1, X_2. The square sub-matrix outline with blue colour satisfies the condition that for each i or j there is only one $p_{ij} > 0$, and $p_{ij} = p_{i\cdot}$ or $p_{ij} = p_{\cdot j}$, respectively. The columns left contain just one $p_{ij} > 0$.*

In this illustrative example we find that the term labelled as T_1 has the following form

$$\begin{aligned} T_1 &= \underbrace{\left(\frac{(4/11)^2}{(4/11)(4/11)}\right)}_{=1} + \underbrace{\left(\frac{(3/11)^2}{(4/11)(3/11)} + \frac{(1/11)^2}{(4/11)(1/11)}\right)}_{=1} \\ &+ \underbrace{\left(\frac{(1/11)^2}{(3/11)(1/11)} + \frac{(1/11)^2}{(3/11)(1/11)} + \frac{(1/11)^2}{(3/11)(1/11)}\right)}_{=1} - 1 \\ &= S_1 - 1. \end{aligned}$$

The above presented example should provide clear insight how the terms in equation 2.12 sum to one, under the assumption of the perfect dependence in the case of $S_1 < S_2$.

Realize that if we reverse the role of S_1 and S_2, thus if we assume that $S_2 < S_1$, the result in equation 2.13 would be $S_2 - 1$. Finally we are able to conclude that the standardization constant that brings the χ^2 statistics to $[0, 1]$ has the following form

$$K = \frac{1}{n\big(\min(S_1, S_2) - 1\big)} \tag{2.14}$$

For the sake of completeness we show that if predictors X_1 and X_2 are probability independent, statistics χ^2_T equals zero. Thus under the assumption of independence $p_{ij} = p_{i\cdot}p_{\cdot j}$ we have

$$\chi^2_T = \frac{1}{\min(S_1, S_2) - 1} \underbrace{\left(\sum_{i=1}^{S_1}\sum_{j=1}^{S_2} \frac{p_{ij}^2}{p_{i\cdot} \cdot p_{\cdot j}} - 1\right)}_{T_2} = 0 \tag{2.15}$$

thus 2.15 equals zero if T_2 equals zero under the assumption of independence. This can be shown as follows

$$\begin{aligned} T_2 = \sum_{i=1}^{S_1}\sum_{j=1}^{S_2} \frac{p_{ij}^2}{p_{i\cdot}p_{\cdot j}} - 1 &= 0 \\ \sum_{i=1}^{S_1}\sum_{j=1}^{S_2} \frac{p_{ij}^2}{p_{i\cdot}p_{\cdot j}} &= 1, \qquad p_{ij} = p_{i\cdot}p_{\cdot j}, \end{aligned}$$

the later statement is true because

$$\begin{aligned} \sum_{i=1}^{S_1}\sum_{j=1}^{S_2} \frac{p_{ij}^2}{p_{i\cdot}p_{\cdot j}} &= \sum_{i=1}^{S_1}\sum_{j=1}^{S_2} \frac{(p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}} \\ &= \sum_{i=1}^{S_1}\sum_{j=1}^{S_2} p_{i\cdot}p_{\cdot j} \\ &= 1. \end{aligned}$$

For the theoretical measure 2.9 it holds that,

$$0 \leq \chi_T^2 \leq 1$$

and statements about $\chi_T^2 = 1$ and $\chi_T^2 = 0$ are exact. Indeed, $\chi_T^2 = 1$ if and only if X_1 and X_2 are perfectly dependent while $\chi_T^2 = 0$ if and only if X_1 and X_2 are independent.

Finally, we outline that the sample counter party to theoretical measure 2.9 is defined as

$$\chi_S^2 = \frac{1}{\min(S_1, S_2) - 1} \left(\sum_{i=1}^{S_1} \sum_{j=1}^{S_2} \frac{n_{ij}^2}{n_{i\cdot} \cdot n_{\cdot j}} - 1 \right). \tag{2.16}$$

Using the above described measure of dependence, precisely it's sample version 2.16 for determining potential dependencies among predictors and using the AUC statistic while assessing the discriminative ability of the predictors, we are able to reduce the long list of categorial predictors.

1.2. Continuous predictor analysis

The single-factor analysis in case of continuous predictors consists of slightly different steps than it was described above, obviously due to different nature of the predictors. Because of potential difficulties that might occure, it is useful to perform the reduction of the continuous predictor set, before any other analysis is done. Being aware of the facts, that further analysis involves setting boundaries (*cutoff points*) which define the range of reasonable predictor values while excluding potential outliers, next assessing necessary transformation of predictors. We should perform the reduction of the predictor set using such measures that are invariant with respect to monotone transformations and which are robust against outliers. At this stage we employ the Spearman correlation coefficient and the discriminative statistics discussed in detail in Chapter III. These fulfill the above outlined properties and thus in this way some predictors can be discarded even before any finer analysis is done, just on the ground of Spearman coefficient matrix, discriminative power statistics.

The ongoing part of the single-factor analysis regarding continuous predictors is performed only on the ground of the reduced predictor set. This involves the following steps.

In contrast with categorial predictors, continuous predictors have to be tested for outliers. The goal of the outlier analysis is to check continuous predictor data for outstanding cases which we may have better excluded before multi-factor modeling.

Further we need to check for the relationship between a specific predictor and the default status. We expect that a reasonable predictor variable has a monotone relationship with respect to the default probability. In order to obtain notion or a possible shape of this relationship we employ a non-parametric smoothing technique.

The next step involves check on linearity assumptions. Due to the fact that we are intent to employ a logistic regression model which implies a linear relationship

between the log odd and the input predictor variables.

1.2.1. Reduction of the predictor set. When assessing the appropriate criteria for reduction of the predictors set, again two kind of characteristics are taken into account. First, we would like to evaluate the discriminative ability of each predictor with respect to separation between defaulting and non-defaulting obligors. Second, we would desire to choose predictors that are not probability dependent in a significant way. The reasons for this are the same a before, strong dependence among predictors causes serious problems within the multivariate analysis as: the estimates of coefficients are inaccurate and the could also have different signs than expected. Thus it is advised to choose only one, at most two representatives out of group of strongly dependent predictors.

Discriminative ability. The assessment of the discriminative ability of a specific predictor is again accomplished by performing the ROC analysis and computing the appropriate summary statistics as it was done in the categorical case. In this context the ROC curve represents again an overall measures for assessing the amount of information included in the underlying predictor regarding its ability to discriminate between good and bad cases. Because of the fact that the major of Chapter III is dedicated to these issues we again do not expand this problematic here.

Dependence structure. Continuous predictors comprise various financial criteria. Financial criteria are likely to be highly dependent. That is why the grade of dependence is emphasized as a key feature when selecting continuous predictors.

In the first step, we classify financial criteria in groups from an economical point of view, such as liquidity, activity, turnover, solvency etc. Criteria within these groups are typically very strongly dependent and hence only one or two predictors from each group are plausible to select at most. Moreover, one can find a lot of strongly dependent couples of financial criteria, each belonging to a different subject group. Note that the strong dependence among continuous predictors is caused by the fact that financial criteria are composed of relatively small number of aggregate items of financial reports. It means that the number of carefully selected financial criteria is typically not larger than the number of the aggregate items. In other words the amount of information gained from financial criteria cannot be larger than the number of the aggregate items.

In credit scoring models, continuous predictors are supposed to be concordant or discordant. The concordance/discordance is a typical representative of a measure of association, a slightly relaxed measure of dependence Nelsen (1998). Two points (x_1, y_1) and (x_2, y_2) are concordant if $(x_1 - x_2)(y_1 - y_2) > 0$ whereas they are discordant if $(x_1 - x_2)(y_1 - y_2) < 0$. The more realizations of random vector (X, Y) are concordant the more concordant is conceived the random vector itself. The same applies to discordance, but understand that concordance and discordance compensate each other similarly to the positive and negative linear correlation. Thus in the case of credit scoring models a certain measure of concordance should

be used to measure the grade of dependence between continuous predictors. We employ the Spearman coefficient.

Assume that we have two continuous predictors X_1 and X_2, that is, they follow distributions with certain continuous distribution functions. Assume that we possess two simultaneous samples of values of the predictors, namely, $(x_{11}, \ldots, x_{1n})$ and $(x_{21}, \ldots, x_{2n})$. Let $R_{11}, \ldots, R_{1n}$ and $R_{21}, \ldots, R_{2n}$ be the corresponding ranks of $(x_{11}, \ldots, x_{1n})$ and $(x_{21}, \ldots, x_{2n})$, respectively. Thus, $R_{sk} = m$ if X_{sk} is the m-th largest value in marginal sample $(x_{s1}, \ldots, x_{sn}), s = 1, 2$. Since predictors are continuous the ranks are well defined with probability one. The Spearman coefficient is defined as the sample correlation coefficient calculated from ranks $R_{11}, \ldots, R_{1n}$ and $R_{21}, \ldots, R_{2n}$

$$\rho_S = \frac{\sum_{i=1}^{n} R_{1i}R_{2i} - n\overline{R}_1\overline{R}_2}{\sqrt{\left(\sum_{i=1}^{n} R_{1i}^2 - n\overline{R}_1^2\right)\left(\sum_{i=1}^{n} R_{2i}^2 - n\overline{R}_2^2\right)}}, \tag{2.17}$$

where $\overline{R}_j = n^{-1}\sum_{i=1}^{n} R_{ji}, j = 1, 2$. Note that sometimes the Spearman coefficient is called the Spearman correlation coefficient obviously because of its definition. However, this term is unfortunate since the Spearman coefficient do not measure the linear correlation between X_1 and X_2 but a kind of more general association, namely the above prescribed concordance. It holds that the Spearman coefficient can be rewritten as follows

$$\rho_S = 1 - \frac{6}{n(n+1)(n-1)} \sum_{k=1}^{n} \left(R_{1k} - R_{2k}\right)^2. \tag{2.18}$$

The later equation is suitable for computer implementation and can be derived from 2.17 by substituting the following terms

$$\begin{aligned}
\overline{R}_j &= \frac{1}{n}\sum_{i=1}^{n} R_{1i} = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n+1}{2} \qquad j = 1, 2, \\
\sum_{i=1}^{n} R_{1i}^2 &= \sum_{i=1}^{n} R_{2i}^2 = \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}, \\
\sum_{i=1}^{n} R_{1i}R_{2i} &= \frac{1}{2}\left(\sum_{i=1}^{n} R_{1i}^2 + \sum_{i=1}^{n} R_{2i}^2\right) - \frac{1}{2}\sum_{i=1}^{n}(R_{1i} - R_{2i})^2, \\
&= \frac{n(n+1)(2n+1)}{6} - \frac{1}{2}\sum_{i=1}^{n}(R_{1i} - R_{2i})^2.
\end{aligned}$$

Simplifying equation 2.17 after substitution is straightforward thus we do not give the calculus here.

Realize that the Spearman coefficient does not change its value under strictly increasing transformations applied to variables X_1 and X_2 whose mutual concordance is measured, because such transformation does not influence ranks. Under strictly decreasing transformations it just reverts its sign. Note that because the Spearman coefficient is based on ranks, it is also robust against outliers. Further

note that

$$-1 \leq \rho_S \leq 1 \tag{2.19}$$

where $\rho_S = 1$ means that variables X_1 and X_2 are perfectly concordant, $\rho_S = -1$ means that variables X_1 and X_2 are perfectly discordant, and $\rho_S = 0$ suggests that X_1 and X_2 are neither concordant neither discordant. In other words two random variables are perfectly concordant if they are strictly increasing transformations of each other. In turn, two random variables are perfectly discordant if they are strictly decreasing transformations of each other.

The critical values $\rho_S^*(\alpha)$ are tabulated, however for $n > 30$ the asymptotic normality of the Spearman coefficient is employed. We compute the critical values at level α, $0 < \alpha < 1$ as follows

$$\rho_S^*(\alpha) = \frac{\Phi^{-1}(\frac{\alpha}{2})}{\sqrt{n-1}}, \tag{2.20}$$

thus the hypothesis of independence is rejected in case $|(\rho_S)| \geq \rho_S^*(\alpha)$.

As prescribed above the Spearman coefficient is robust against outliers. It has the advantage that we can calculate the matrix of Spearman coefficients before any outlier analysis is carried out. In this way some predictors can be discarded even before any single-factor analysis, just on the ground of Spearman coefficient matrix and discriminative power statistics. It can save quite a time because the single-factor analysis of continuous predictors is quite complicated and elaborate. Further note that at any case the Spearman coefficients must be calculated from data before cutoffs and missing values replacement. One simple reason is that the Spearman coefficient rests on ranks which are not well defined for samples with ties. Certain modifications can be used for equal observations, but there are other reasons not to do it. Replacement of missing values can artificially change the value of the Spearman coefficient. On the other hand, cutoffs can significantly decrease the value of the Spearman coefficient because all diversity beyond cutoff points is shrunk in one point.

1.2.2. Assessment of outliers. Single-factor analysis of continuous predictors involves a check of the tail distribution and outliers. Firstly, wrong input data can be detected, secondly, a check of tails is useful for assessment of cutoff points and suggestion of appropriate transformation. Wrong input data involves both wrong figures and obligors which does not belong to the modelled group. For example, financial institutions typically exhibit huge balance sheet sums and tiny ratios of equity in contrast with non-financial corporates. Hence the first check of outliers can be also viewed as data validation. The simple check involves plotting the data, or more precisely the left and right tails, to see whether there are not a few observations which are one or several times larger or smaller than their neighbours. An illustrative example is provided in Figure 2.2. Such obligors are then checked for validity. Outliers are discarded depending on the result of the validity check and on feasibility of data recovery.

We are ahead to estimate the 5% and 95% quantiles $u_{0.05}$ and $u_{0.95}$ of the

underlying distribution of the actual predictor,i.e.

$$\begin{aligned} u_{0.05} &= \inf\{x : Q(x) \geq 0.05\}, \\ u_{0.95} &= \inf\{x : Q(x) \geq 0.95\}, \end{aligned}$$

where $Q(x)$ denotes the unknown distribution function of the considered predictor.

As we do not assume any particular distribution to be followed by the predictor we estimate the appropriate quantiles according to the following scheme. Assume that we observe n realizations of the underlying predictor variable X, that is $x_1, \ldots, x_n$. We order the sample $x_1, \ldots, x_n$ obtaining the ordered sample $x_{(1)}, \ldots, x_{(n)}$. The appropriate quantiles are than estimated as

$$\begin{aligned} \hat{u}_{0.05} &= x_{(\lfloor (n+1)0.05 \rfloor)}, \\ \hat{u}_{0.95} &= x_{(\lfloor (n+1)0.95 \rfloor)}, \end{aligned}$$

where $\lfloor \quad \rfloor$ denotes the lower interger part. Obligors corresponding to values smaller than the estimated $\hat{u}_{0.05}$ quantile and larger than the estimated $\hat{u}_{0.95}$ quantile are selected as possible problematic cases. Proceeding this way for all continuous predictors we obtain a set of suspicious obligors (in the sense of their data).

Realize that at this stage no regression outliers analysis is worth doing because later possible transformations of predictors can change everything in drastic way. Regression outlier analysis is postponed to multi-factor analysis when final models are put through regression diagnostic procedures.

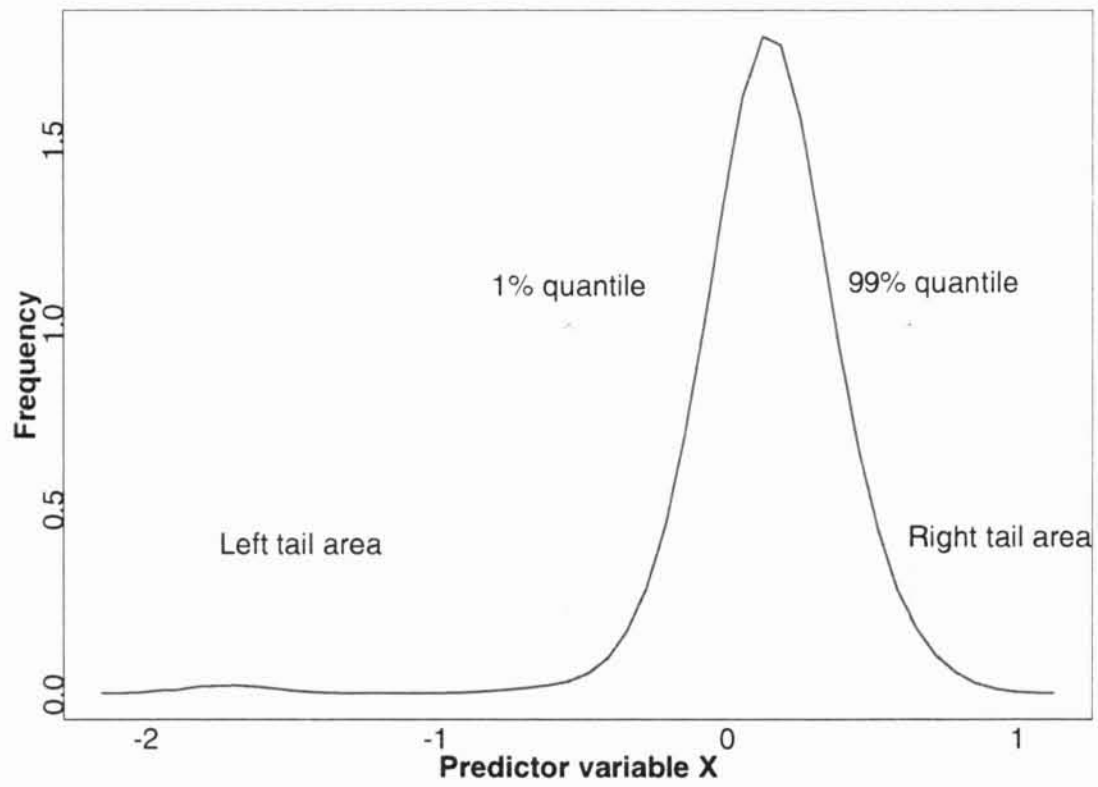


Figure 2.2. *An illustrative example of the possible distribution of a single predictor variable. The suspicious values are expected to be found in one of the tail areas of the distribution.*

1.2.3. Treatment of missing values. In single-factor analysis of continuous predictors, missing values are not substituted in the first step as it is not desirable

to dim down the real discriminative ability of the predictor by any substitutions. As soon as the discriminative ability of the predictor with missing observations discarded is checked one can proceed with a check of what happens if various substitutions are used. There are natural cases when missing values appear and substitutions are relevant. Typically, some financial criteria are not always well-defined because of division by zero or, for instance, because they have a good meaning only for positive values of input. In such cases it is usually logical and well reasoned by experts to assign an ultimate upper or lower value. These levels are typically put equal to cutoff points described below. Generally, we can observe anything–decrease, increase as well as stagnation of the discriminative ability after substitutions. Substantial decrease is unfortunate, of course, and suggests that problems may be encountered in subsequent multi-factor analysis where substitution of missing values can be needed because of combining many predictors. On the other hand, substantial increase is suspicious and has to be investigated carefully. It can suggest that missing values correspond systematically to certain kind of obligors.

1.2.4. Testing monotonicity and suggestion of transformations. At this stage of the single-factor analysis we are ahead to get notion about the univariate regression dependence between the considered continuous predictor variables and the default probability. We assume that the predictor set is already reduced using the discriminatory power statistics as described above, so here it is not expected that there would be any predictor that would have a constant (no) regression relationship with respect to the default probability. Note that such predictor would perform bad also on the ground of the discriminatory ability, thus it would be excluded earlier. However, if a certain predictor has a reasonable predictive power in term of the discriminatory statistics, it does not have to be clear, whether this predictor has a monotone regression relationship with respect to the default probability. Realize that the assumption of a monotone relationship is obvious since it is desired that a reasonable predictor does not have for example a decreasing regression relationship with respect to the default probability in the certain range of its values and a increasing relationship in another (disjunct to the the first one) range of its values. The test on the monotonicity assumption is performed by employing a *smoothing technique.*

Having observed quantity X, the expected value of Y is given by the regression function. It is of great interest to have some knowledge about this relationship. If n data points $\{(x_i, y_i)\}_{i=1}^n$ have been collected, the regression relationship can be modeled

$$y_i = m(x_i) + \epsilon_i \qquad i = 1, \ldots, n,$$

with the unknown regression function m and observation errors ϵ_i. The aim of a regression analysis is to produce a reasonable estimate $\hat{m}(x)$ to the unknown response function $m(x)$. By reducing the observational errors it allows interpretation to concentrate on important details of the mean dependence of Y on X. This curve approximation procedure is commonly called smoothing. Especially in this section we have performed a nonparametric smoothing approach, which offers

a flexible tool in analyzing unknown regression relationships. The term nonparametric refers to the non-prespecified functional form of the regression curve. The function estimates $\hat{m}(x)$ are often called *smooths* (when smoothness is assumed). We assume that the reader is familiar with the basic ideas about non-parametric regression techniques. However we outline that in our context we deal with regression estimates $\hat{m}(x)$ that can be expressed in the following form

$$\hat{m}(x) = \frac{1}{n}\sum_{i=1}^{n} W_{ni}(x) y_i, \tag{2.21}$$

where $\{W_{ni}(x)\}_{i=1}^{n}$ denotes a sequence of weights which may depend on the whole vector $(x_1, \ldots, x_n)$. Thus a local averaging procedure of the form 2.21 can be viewed as the basic idea of smoothing. The amount of averaging is controlled by the weight sequence $W_{ni}(x), i = 1, \ldots, n$, which is tuned by a *smoothing parameter.*

For our purposes we employ the so called *supersmoother* proposed by Friedman (1984) which is based on local linear $k-NN$ (k nearest neighbor estimates) fits in a variable neighborhood of the estimation point x. Local cross-validation is applied to estimate the optimal span as a function of the predictor variable. A great advantage of the $k - NN$ estimate is that its computation can be updated quite easily when x runs along the sorted array of values of predictor X. The algorithm requires essentially $O(n)$ operations to compute the smooth at all x_i. It is therefore highly computationally efficient thanks to the following recursive approach. In case we suppose that the data have been pre-sorted, so that $x_i \leq x_{i+1}, i = 1, \ldots, n-1$. Then if the estimate has already been computed at some point x_i, the smooth at x_{i+1} can be recursively determined as

$$\hat{m}_k(x_{i+1}) = \hat{m}_k(x_i) + \frac{1}{k}\left(y_{i+[k/2]+1} - y_{i-[k/2]}\right), \tag{2.22}$$

where k denote the number of nearest neighbors corresponding to x_i and x_{i+1} and $[k/2] = \sup\{i : i \leq k/2\}$. We do not expand the technical details of this smoothing method here. We refer the interested reader for example to Härdle (1994).

As it was prescribed above we employed this smoothig technique in order get insight about the regression relationship between our dependent variable Y having two possible outcomes (default coded as 1, non-default coded as 0) and the explanatory (predictor) variable X. A simple look at a scatter plot of observations x_i versus y_i, $i = 1, \ldots, n$ does not always suffice to establish an interpretable regression relationship as shown in Figure 2.3.

On the other hand, if we smooth the values x_i against y_i using the Friedman's adaptive supersmoother, we are able to produce very clear and reasonable estimates of the regression relationship between the two considered variables. This is illustrated in Figure 2.4.

However, we do not observe such a nice behaviour every time. There are plenty of cases when the monotonicity assumption is violated. An illustrative example is provided in Figure 2.5. Those predictor variables where the monotonicity assumptions are violated has to be checked for validity. There might occure some cases when the non-monotone behaviour has a reasonable economic interpretation, however these predictors are not suitable for later multifactor modelling.

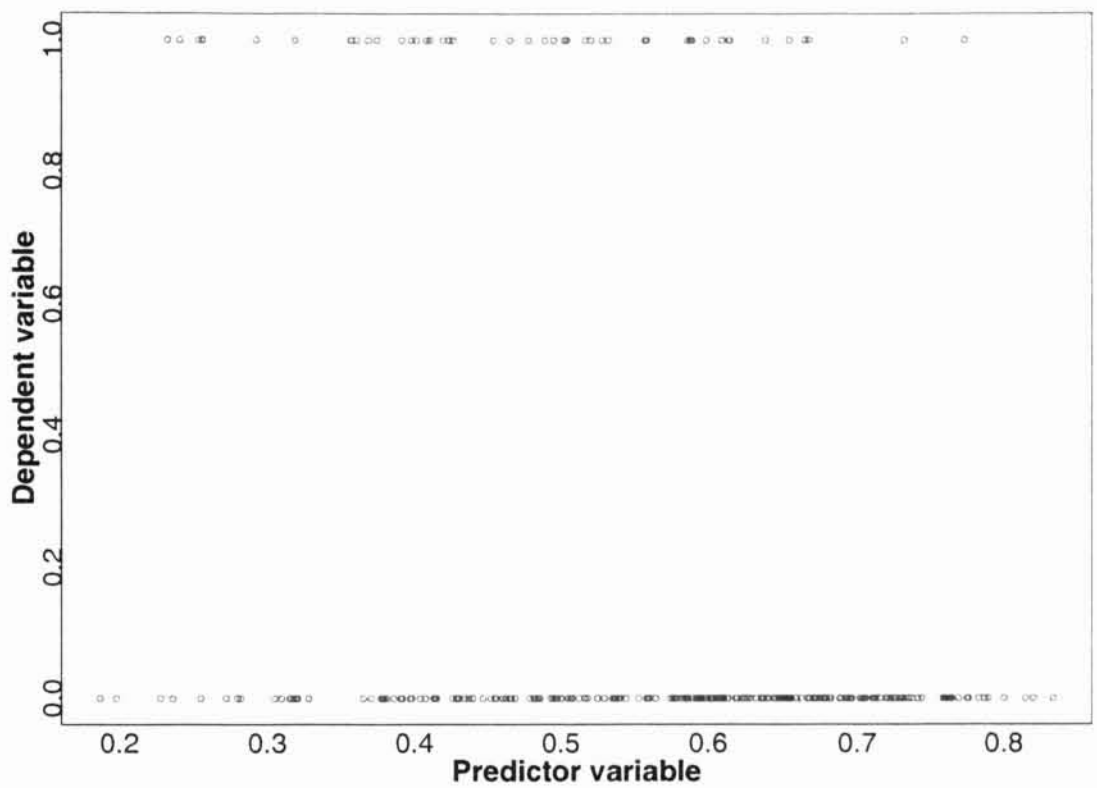


Figure 2.3. *A scatter plot of x_i versus y_i, $i = 1, \ldots, n$. Clearly it does not provide any useful information about the relationship between the two underlying variables.*

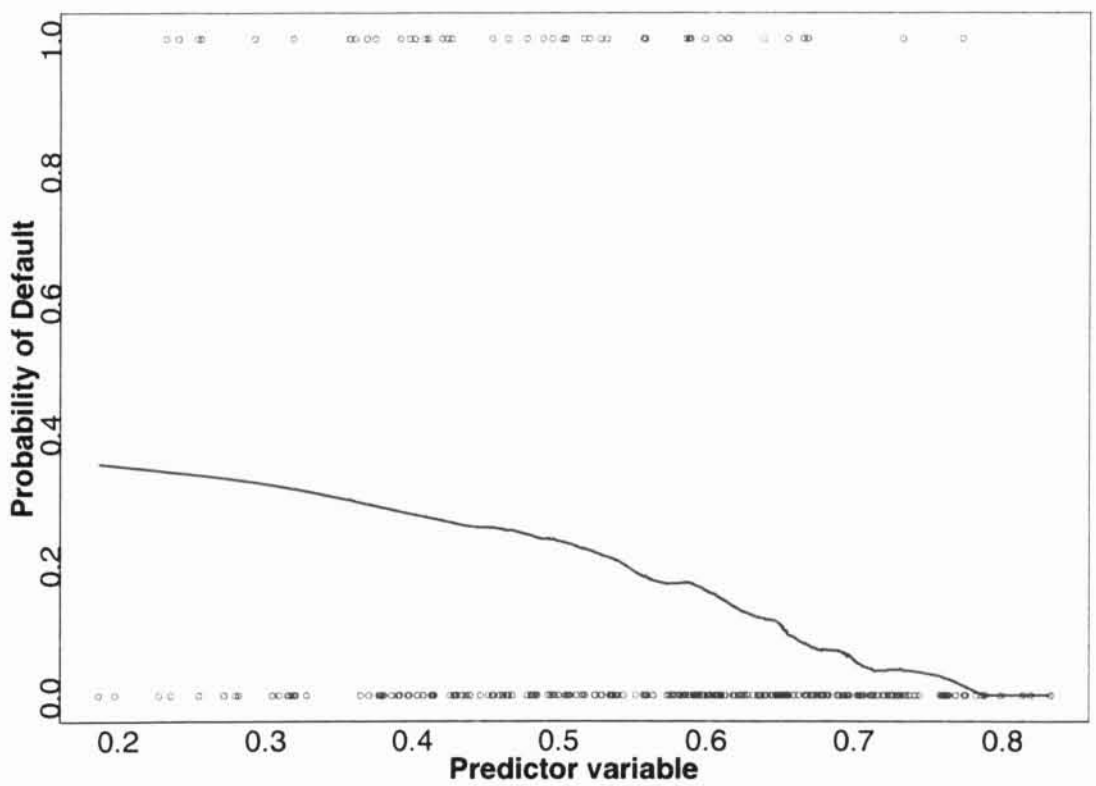


Figure 2.4. *A scatter plot of x_i versus y_i, $i = 1, \ldots, n$ and the appropriate smooth. A clear monotone/decreasing regression relationship is observed.*

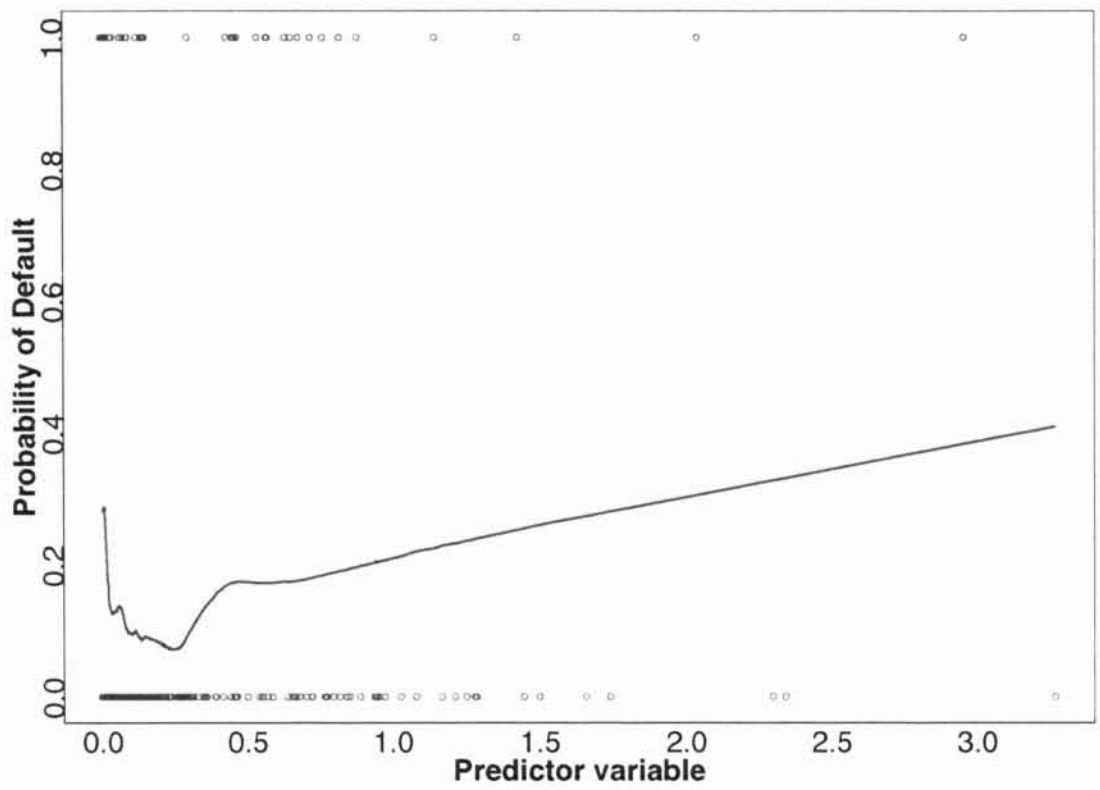


Figure 2.5. *A scatter plot of x_i versus y_i and the appropriate smooth. A clear violation of the monotonicity assumption.*

Once we have tested the predictor variables for monotonicity we further desire to test another stronger assumption, i.e. the test of linearity assumption. With respect to the ongoing multi-factor analysis it is reasonable to check whether the assumptions of the underlying logistic regression model apply to the data. With reference to Appendix A the logistic regression model assumes the following relationship

$$\pi(\boldsymbol{x}) = \frac{\exp\left\{\alpha + \sum_{j=1}^{m} x_j\beta_j\right\}}{1 + \exp\left\{\alpha + \sum_{j=1}^{m} x_j\beta_j\right\}} \tag{2.23}$$

where $\pi(\boldsymbol{x}) = P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = 1 - P(Y = 0|\boldsymbol{X} = \boldsymbol{x})$. Equation 2.23 implies a linear relationship between the log odd and the input explanatory variables:

$$\log\left(\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}\right) = \alpha + \sum_{j=1}^{m} x_j\beta_j \tag{2.24}$$

In case the above described linear relationship does not hold it is recommended to suggest certain transformations of predictor variables.

Namely, for each predictor we look for an appropriate function f in an univariate model

$$\log\left\{\frac{\pi(X)}{1 - \pi(X)}\right\} = f(X). \tag{2.25}$$

Since f is non-parametric in nature we employ the smoothing technique again.

In the first stage we use the Friedman supersmoother to get notion about possible shape of f for values of the predictor within data range. Smoothing can

be done without excluding any obligors as outliers because smoothers focus on data locally and hence outliers do not disturb the final estimate unless in the very tail area. After we have determined the shape of the non-parametric smooth we try to give a parametric transformation which is close to the fitted smooth. The closeness is determined according to a technique for comparing parametric and non-parametric curves. However, we do not expand the related problematic here. The choice of the parametric transformation can be taken within a predefined set of parametric functions such as $f(x) = b\log(x - a)$. Also, there is a possibility to select a group of good candidates for transformation f and make the decision later based on practical arguments. The final choice of the transformation need not be strictly technical a broad space can be given to expert arguments.

2. Multi factor analysis

As soon as a reasonable set of predictors is settled within single-factor analysis, building of multivariate models can be launched. All the models that we consider within the process of multi factor analysis stems from the class of *generalized linear models*. The first part of this section provides a background information about the genesis of generalize linear models and explains why they are useful in our situation. Further paragraphs contain the information about statistical properties of this specific class of regression models. Finally, model selection techniques and techniques of assessing the goodness of fit of a model as well as model diagnostics methods are described.

The generalized linear regression model is a generalization of the usual linear regression model, so it is important to outline the limitations of the standard linear model and why we would like to generalize it. In practical applications it is quite common that the relationship between the response and the predictor variables is not linear. The response variables could be bounded, such as categorical response variables as in our situation, or the variance is non-constant, it could be expressed as the function of the means. Thus in these cases, the assumptions concerning the standard linear regression model does not hold.

General linear models are a generalization of linear regression models. Specifically, the predictor effects are assumed to be linear in the parameters, but the distribution of the response, as well as the *link* between the predictors and this distribution, can be quite general. A general linear model also consist of a *random component*, a *systematic component* and a additional *link function*, linking the two components.

The response variable Y represents the random component and it is assumed to have exponential family density 2.26

$$f(y;\theta;\phi) = \exp\Big\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\Big\}, \tag{2.26}$$

where $b(\cdot)$ is a smoothly differentiable function to the second order, $a(\phi)$ and $c(y,\phi)$ are functions such that $a(\phi) > 0$ and $c(y,\phi)$ does not depend on θ. Further note that ϕ is called the dispersion parameter. The parameter θ depends on values $\boldsymbol{x}^\top = (x_1,\ldots,x_m)$ of explanatory variables and on the vector of coefficients $\boldsymbol{\beta}$

trough the linear predictor $\eta = \alpha + \boldsymbol{x}^\top \boldsymbol{\beta}$. The linear predictor η represents the systematic component. Further there is a monotonic differentiable link function g such that $\eta = g(\mu)$, i.e.

$$g(\mu) = \alpha + \sum_{j=1}^{m} x_j \beta_j, \tag{2.27}$$

where $\mu = E(Y|\boldsymbol{X} = \boldsymbol{x})$. Note that the mean μ is related to the θ by $\mu = b'(\theta)$ and that a link function for which $g(\mu) = \theta$ is called the *canonical link.*

Many useful models fall into this class, including the Logistic regression model for binary data, that we employ.

2.1. The Logistic regression model

The logistic regression model assumes that we have a binary response variable Y having alternative (Bernoulli) distribution Alt(π). Thus variable Y has two possible outcomes $Y = 1$ indicating that the obligor is a defaulter and $Y = 0$ indicating that he is a non-defaulter. The mean μ in this situation is equal to $\mu = E(Y|\boldsymbol{X} = \boldsymbol{x}) = P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \pi$. We denote this probability as $\pi(\boldsymbol{x})$ reflecting its dependence on values $\boldsymbol{x}^\top = (x_1, \ldots, x_m)$ of predictors. In case of logistic regression the link function $g(\cdot)$ introduced above has the form 2.28

$$g(\pi(\boldsymbol{x})) = \log \frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}. \tag{2.28}$$

With this notation the logistic regression model takes the form

$$\log \frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})} = \alpha + \sum_{j=1}^{m} x_j \beta_j. \tag{2.29}$$

or equivalently

$$\pi(\boldsymbol{x}) = \frac{\exp(\alpha + \sum_{j=1}^{m} x_j \beta_j)}{1 + \exp(\alpha + \sum_{j=1}^{m} x_j \beta_j)}. \tag{2.30}$$

The term on the left side of equation 2.29 is called the *logit* or the *log odds.*

The corresponding conditional Bernoulli density of Y can be then written as

$$f(y, \pi(\boldsymbol{x})) = P(Y = y \mid \boldsymbol{X} = \boldsymbol{x}) = \big(\pi(\boldsymbol{x})\big)^y \big(1 - \pi(\boldsymbol{x})\big)^{(1-y)}, \quad y = 0, 1. \tag{2.31}$$

Fundamental model fitting techniques seek for estimates of α and $\beta_1, \ldots, \beta_m$ which maximize the conditional log-likelihood implied by Bernoulli distribution 2.31. The conditional log-likelihood is introduced below.

Assume we have n independent subjects. With n independent subjects we threat n binary responses $(y_1, \ldots, y_n)$ of random variable Y. Further $\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top$ are the corresponding settings i.e. $\boldsymbol{x}_i^\top = (x_{i1}, \ldots, x_{im})$ denotes the setting i of

values of m predictor variables, $i = 1, \ldots, n$. The likelihood function of the n independent subjects is then equal to the product of marginal Bernoulli densities 2.31.

$$L_n(\alpha, \boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\pi(\boldsymbol{x_i}, \alpha, \boldsymbol{\beta})\right)^{y_i} \left(1 - \pi(\boldsymbol{x_i}, \alpha, \boldsymbol{\beta})\right)^{(1-y_i)}. \tag{2.32}$$

Above we denote $\pi(\boldsymbol{x_i}) = \pi(\boldsymbol{x_i}, \alpha, \boldsymbol{\beta})$ in order to outline to which term the α and $\boldsymbol{\beta}$ relate. The corresponding conditional joint log-likelihood, denote it as $l(\alpha, \boldsymbol{\beta})$ is equal to the sum of the corresponding marginal conditional log-likelihoods. The formula for the joint conditional log-likelihood reads

$$l(\alpha, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i \log\left(\pi(\boldsymbol{x}_i, \alpha, \boldsymbol{\beta})\right) + \left(1 - y_i\right) \log\left(1 - \pi(\boldsymbol{x}_i, \alpha, \boldsymbol{\beta})\right)\right). \tag{2.33}$$

Estimates $\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_m$ of α and $\beta_1, \ldots, \beta_m$ obtained by maximization of 2.33 are called *maximum likelihood estimates*. The estimating procedure is done by the Newton-Raphson algorithm, which is general-purpose iterative method for solving nonlinear equations. Computational details are given for example in Agresti (2002).

Realize that maximum likelihood estimates are parameter values under which the data observed have the highest probability of occurrence. Note the parameter values that maximizes the log likelihood function 2.33 also maximizes the underlying likelihood function, however it is simpler to maximize the log likelihood since it is a sum rather than a product of terms. Finally, for further purpose denote $SE(\hat{\boldsymbol{\beta}})$ the standard error of a multivariate parameter $\hat{\boldsymbol{\beta}}$ and let $\mathrm{cov}(\hat{\boldsymbol{\beta}})$ denote the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$.

2.2. Parameter interpretation in the logistic regression model

A reasonable interpretation of the regression parameters is a key feature of understanding the magnitude of the estimated effects. The interpretation of parameters in the logistic regression model is based on the following simple calculations. Exponentiating equation 2.29, the logistic model can be equivalently written in terms of odds of the positive response as

$$\frac{\pi(\boldsymbol{x_i})}{1 - \pi(\boldsymbol{x_i})} = \exp\left(\alpha + \sum_{j=1}^{m} x_{ij}\beta_j\right), \tag{2.34}$$

thus the probability of the positive response is

$$\pi(\boldsymbol{x_i}) = \frac{\exp\left(\alpha + \sum_{j=1}^{m} x_{ij}\beta_j\right)}{1 + \exp\left(\alpha + \sum_{j=1}^{m} x_{ij}\beta_j\right)}. \tag{2.35}$$

Equation 2.34 provides a basic interpretation for the magnitude of β's. Assuming that predictors $X_1, \ldots, X_m$ are functionally uncorrelated, we can say that the odds increase multiplicatively by $\exp(\beta_j)$ for every one unit increase in X_j, holding all other predictors values fixed. In other words we can say that $\exp(\beta_j)$ is the

odds ratio, thus the odds at $X = x_j + 1$ divided by the odds at $X = x_j$.

Further we can state that the sign of the a certain parameter β determines whether $\pi(x)$ is increasing or decreasing as x increases. The rate of climb or descent is determined by $|\beta|$. In case a specific $\beta_j = 0$ than random variable Y is independent of variables X_j.

2.3. Strategies in model selection

Model selection for logistic regression faces the same issues as in the case of standard linear regression. The selection process becomes harder as the number of explanatory variables increases, because of the rapid increase in possible effects and interactions. There are two competing goals: The model should be complex enough to fit the data well on one hand. On the other hand, it should be simple to interpret, smoothing rather than over fitting the data.

There exist many model selection procedures, no one of which is always the best. Caution is required for any generalized linear model building process. A model with several predictors may suffer from multi-collinearity among predictors making it seem that no one variable is important when all the other are in the model. A variable may seem to have a little effect because it overlaps considerably with other predictors in the model, itself being predicted well by another predictors. Deleting such a redundant predictor can be helpful, for instance to reduce standard errors of other estimated effects.

The common model building procedures are described in the next paragraph of this section. However, realize that no matter which model building strategy we choose there is a common feature on which they are based. This relates to the statistics which we use when evaluating significance of the predictor variables.

The standard tool for testing the significance of predictors generalized linear regression is the regression z-statistic. Formally speaking, for the logistic regression model of the form A.10 significance test of predictor variable X_j focuses on testing the hypothesis

$$\mathrm{H}_0 : \ \beta_j = 0, \tag{2.36}$$

which speaks for the independence of the response variable Y on the predictor variable X_j. According to Wald's asymptotic results for maximum-likelihood estimators, parameter estimators in logistic regression models have large sample normal distribution. Based on these results, the significance tests of the null hypothesis 2.36 has the following form. With nonnull standard error $SE(\hat{\beta})$ of $\hat{\beta}$ the test statistics

$$z = \frac{\hat{\beta}}{SE(\hat{\beta})} \tag{2.37}$$

has an approximate standard normal distribution under H_0. Thus z^2 has a chi-squared null distribution with one degree of freedom. This type of statistics is called the *Wald statistics*. The multivariate extension for the Wald test of $\mathrm{H}_0 : \boldsymbol{\beta} = \boldsymbol{\beta_0}$ has test statistics

$$W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})^\top \left[\mathrm{cov}(\hat{\boldsymbol{\beta}})\right]^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}). \tag{2.38}$$

The asymptotic normal distribution of $\hat{\beta}$ implies the asymptotic chi-squared distribution for W. The degrees of freedom equal to the rank of the asymptotic covariance matrix $\text{cov}(\hat{\boldsymbol{\beta}})$ of $\hat{\boldsymbol{\beta}}$.Thus generally if $|z| > \Phi^{-1}(\alpha)$, where Φ is the distribution function of the standard normal distribution the predictor X_j is considered not to be statistically significant within the current model at the confidence level α.

These test are incorporated within the automatic model building procedures describe in the next paragraph.

2.4. Stepwise procedures

In case when we face a relatively big amount of candidate predictor variables an algorithmic method for searching among models can be informative if we use the results cautiously.

The first possible approach is the forward selection procedure. Forward selection adds variables sequentially until further additions do not improve the fit. At each stage it selects a variable giving the greatest improvement in the fit. The minimum P- value for testing the significance of the variable in the model is a sensible criterion, a complement to this is the reduction in deviance. A stepwise variation of this procedure retests, at each stage the variables added at previous stages to see if they are still significant.

Backward elimination begins with a complex model an sequentially removes the predictor variables. At each stage, it selects a variable for which its removal has the least damaging effects on the model. In other words, the largest P-value. The process stops when any further deletion leads to a significantly poorer fit. With either approach, for qualitative (categorical) predictor variables with more than two categories, the process should consider the entire variable at any stage rather than just one of its dummies. Otherwise, the result depends on the coding.

We prefer the the backward elimination over forward selection, feeling it safer to delete variables from an overly complex model than to add variables to an overly simple one. Forward selection can stop prematurely because a particular test in the sequence has a lower power. Realize that neither strategy must lead to a reasonable model, thus the algorithmic/automatic variable selection procedures has to be used with caution. Finally note that statistical significance should not be the sole criterion for the inclusion of predictor variables in the model. Credit expert usually suggest also certain predictor variables which should be at least tested for significance even if they are not included in the proposed short list containing variables for multi factor modelling.

2.5. Assessing the Goodness of fit

Let us denote the fitted values of a particular logistic regression model as $\hat{\pi}(\boldsymbol{x}_1, \hat{\alpha}, \hat{\boldsymbol{\beta}}), \ldots, \hat{\pi}(\boldsymbol{x}_n, \hat{\alpha}, \hat{\boldsymbol{\beta}})$, for simplicity we will write $\hat{\pi}_1, \ldots, \hat{\pi}_n$. Further denote $l(\boldsymbol{\pi}; \boldsymbol{y})$ the log-likelihood function expressed in terms of means $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$. Let $l(\hat{\boldsymbol{\pi}}; \boldsymbol{y})$ denote the maximum of the log likelihood for the considered model. Realize hat for all possible model, the maximum log likelihood is $l(\boldsymbol{y}; \boldsymbol{y})$. This

occurs for the most complex model, having a separate parameter for each observation and thus a perfect fit $\hat{\boldsymbol{\pi}} = \boldsymbol{y}$. Such a model is called the *saturated model*. This model is not useful, since it does not provide data reduction. However, it serves as a baseline for comparison with other model fits.

The *deviance* of the logistic regression model is defined as

$$D(\boldsymbol{y}; \hat{\boldsymbol{\pi}}) = -2\frac{\text{maximum likelihood for model}}{\text{maximum likelihood for saturated model}} \tag{2.39}$$

$$= -2[l(\hat{\boldsymbol{\pi}}; \boldsymbol{y}) - l(\boldsymbol{y}; \boldsymbol{y})].$$

what describes the lack of the fit. It is the likelihood ratio statistics for testing the null hypothesis that the model holds against the alternative that a more complex model holds.

The deviance function is most directly useful not as an absolute measure of goodness-of-fit but for comparing two nested models. Consider two models M_0 and M_1 with fitted values $\hat{\boldsymbol{\pi}}_{\mathbf{0}} = (\hat{\pi}_{01}, \ldots, \hat{\pi}_{0n})$ and $\hat{\boldsymbol{\pi}}_{\mathbf{1}} = (\hat{\pi}_{11}, \ldots, \hat{\pi}_{1n})$. Further let M_0 be a special case of M_1, thus M_0 is nested within M_1. Since M_0 is simpler than M_1, a smaller set of parameter values satisfies M_0 than it satisfies M_1. Thus maximizing the log likelihood over a smaller space cannot yield a larger maximum. So, we have that $l(\hat{\boldsymbol{\pi}}_{\mathbf{0}}; \boldsymbol{y}) \leq l(\hat{\boldsymbol{\pi}}_{\mathbf{1}}; \boldsymbol{y})$ and from 2.40 it follows that

$$D(\boldsymbol{y}; \hat{\boldsymbol{\pi}}_{\mathbf{1}}) \leq D(\boldsymbol{y}; \hat{\boldsymbol{\pi}}_{\mathbf{0}}) \tag{2.40}$$

Thus simpler models have larger deviances. Assuming that model M_1 holds, the *likelihood-ratio test* of the hypothesis that M_0 holds uses the test statistics

$$-2[l(\hat{\boldsymbol{\pi}}_{\mathbf{0}}; \boldsymbol{y}) - l(\hat{\boldsymbol{\pi}}_{\mathbf{1}}; \boldsymbol{y})] = -2[l(\hat{\boldsymbol{\pi}}_{\mathbf{0}}; \boldsymbol{y}) - l(\boldsymbol{y}; \boldsymbol{y})] - \{-2[l(\hat{\boldsymbol{\pi}}_{\mathbf{1}}; \boldsymbol{y}) - l(\boldsymbol{y}; \boldsymbol{y})]\}$$

$$= D(\boldsymbol{y}; \hat{\boldsymbol{\pi}}_{\mathbf{0}}) - D(\boldsymbol{y}; \hat{\boldsymbol{\pi}}_{\mathbf{1}})$$

We observe that the likelihood ration statistics is the difference between the deviances. Obviously, this statistics is large when M_0 fits the data poorly compared to M_1. Under certain regularity conditions, this difference has asymptotically a chi-square null distribution with degrees of freedom equal to the difference between the number of parameters in the two models.

The deviance function serves a great purpose when comparing several logistic regression models. Realize that in case models M_0 and M_1 differ only in one predictor variable. The deviance function tests the significance of this individual predictor variable, although generally it allows testing the significance of a group of predictor variables.

2.6. Logistic regression diagnostics

In the previous paragraphs we introduced statistics for checking the model fit in a global sense. After selecting a model candidate we move to a more detailed analysis of the models quality. Specifically, we describe the basis properties of the analysis of residuals in the context of generalized linear models. The analysis of residuals is used to carry out regression outlier analysis and influential analysis.

For continuous predictors, graphical methods are also very common to use.

Here we concentrate only on one type of residuals although there are more types which could be employed. On the background the deviance function we can define *deviance residuals* in the following way. Realize that if the deviance is a measure of discrepancy of the model, than each unit contributes a quantity d_i, so that the deviance equals $D = \sum_{i=1}^{n} d_i$. The deviance residuals are defined as

$$r_D = \text{sign}(y_i - \hat{\pi}_i)\sqrt{d_i}, \tag{2.41}$$

where

$$d_i = 2\left(y_i \log\left(\frac{y_i}{\hat{\pi}_i}\right) + (1 - y_i)\log\left(\frac{1 - y_i}{1 - \hat{\pi}_i}\right)\right). \tag{2.42}$$

Using the deviance residuals may help us identify whether the are observation for which the model fits poorly. Whenever a residual indicates that the model fit the data poorly in the appropriate region, it can be informative to delete the observation and refit the model to the remaining ones. Note that this is equivalent to adding a parameter to the model for that observation, in order to provide a perfect fir for it. Residual analysis is a important step when assessing observation which could possible influence the parameter estimates (thus the whole fit) in a undesirable way, however we do not expand these issues here.

2.7. Final comments

This chapter was focused on the statistical methodology related to the model development process. First we have discussed single-factor analysis in order to determine a reasonable set of predictor variables, which were later used in multi-factor modeling. The result of the multi-factor analysis is a proposal of model candidates. Note that there is no best model, thus in practical application several model candidates are developed.

For each obligor, the models produce a score value $s = \boldsymbol{x}^\top \hat{\boldsymbol{\beta}}$, which is afterwards transformed to the obligors estimated default probability. These two outcomes are considered to be equivalent.

In the next chapter we discuss statistical methods which enable the validation of the proposed models. Based on the results obtained from these methods the final best feasible model is chosen. For the purpose of validation procedures we divide the estimated score values produced by the models into two groups corresponding to defaulting and non-defaulting obligors. Although the suggested logistic model produces continuous score values, the next chapter covers a discrete case as well in order to provide general validation framework.

Chapter III

Model Validation and Benchmarking Techniques

In the previous chapter we have presented the methodology we have developed while building credit scoring models for the environment of corporate and semi-corporate firms. However, a model without sufficient validation can only stay a hypothesis. Without adequate objective validation criteria and processes, the benefits of implementing and using these scoring models cannot be fully realized. This makes reliable validation techniques crucial at this point. Such testing also gives the user confidence that the model is stable and has not been overfitted. Model evaluation techniques and methods are necessary tools to aid searching and choosing of the appropriate model.

1. Criteria of model validation

In the situation when we are ahead to choose a specific model as our final scoring tool, one of the criteria to be considered, is the classification accuracy - *the number of obligors classified correctly among all the obligors.* This is a simple and natural quantity, which can be measured quantitatively and the model builder could possibly be interested in. Commonly an equivalent criterion, the misclassification error rate, is considered.

$$\text{Error rate} = \frac{\text{The number of incorrectly classified cases}}{\text{The total number of cases}}.$$

The error rate of the model being evaluated should be in the range between the zero error rate (of a perfect model that classify all cases correctly) and a random model's error rate (random assignment of rating scores).

Nevertheless the classification accuracy (equivalently the misclassification error rate) is not the only criterion that counts. There are also other aspects to be considered with respect to the practical application of the model. The issue of the computational time might be as well a key criterion. The required time for both the training and the applying of the classification model should be considered. With the drifting of the population, most models fail to be stable in long time run. Thus the time needed to train a model is therefore important since model should be regularly revised. Finally we mention the transparency and the interpretability of the considered model. An important attribute of the models is that there is a transparent relationship between input variables and the output so that one can see the impact of each input variable on the output. In practice, model builders try to come up with the best trade-off between these criteria.

2. Methods of model validation

In the case of default risk models, validation involves examining the goodness of the model according to the following basic stream. The *power* of the model, which can be expressed as how well a model discriminates between defaulting and non-defaulting obligors. In case we have two models that produce ratings of *good*

and *bad*, the more powerful is the model that has a higher percentage of defaults (lower percentage of non-defaults) in its bad category and a higher percentage of non-defaults (and a lower percentage of defaults) in its good category. In practice, measuring this quantity can be challenging.

2.1. ROC Graphs and Power Statistic

2.1.1. ROC Graphs and their generation. Assume that we generally have n obligors and let us consider a model that assigns each obligor a score s out of a specific set T. T might be either a finite set of r discrete values $\{s_1, \ldots, s_r\}$; $s_1 < \ldots < s_r$, or it might be a continuous interval, say $[0, 1]$. In general assume that the score values are ranging from worst to best. It means that a high score indicates a low default probability. Further we introduce random variables S_D and S_{ND}. The random variables S_D and S_{ND} follow the score distribution of the defaulters and the non-defaulters, respectively. According to this, we will distinguish two cases, the case when S_D and S_{ND} have continuous distribution and the case when they have finitely discrete distributions. A random variable having a finite discrete distribution is meant to be a variable that can equal only a finite number of values with nonzero probability. A possible distribution of rating scores for defaulting and non-defaulting obligors, that were assigned by a specific model is illustrated in Figure 3.1. Note that for a perfect scoring model, the distributions of defaulters and non-defaulters are separate. Naturally for real world scoring models a perfect discrimination is not possible, thus both distributions overlap each other.

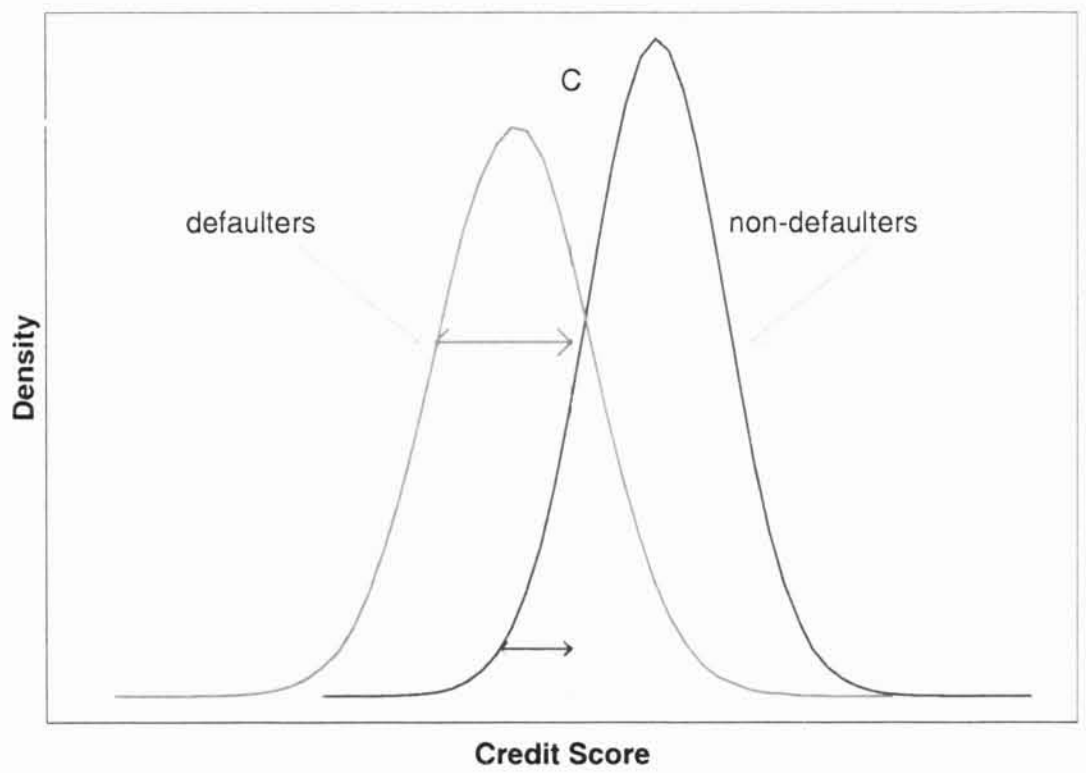


Figure 3.1. *The overlapping distribution of credit scores. The cut-off point C represents a potential boundary for classifying obligors.*

Suppose that our aim is to decide (given a set of credit scores assigned by the model, with properties as prescribed above) which obligor will default during a certain period of time and which will survive. In order to come up with

a reasonable decision and classify certain obligors as potential defaulters and potential non-defaulters it is useful to introduce a cut-off point C as in Figure 3.1. According to the cut-off point C each obligor with credit score lower than C is classified as a potential defaulter and each obligor with credit score higher than C as a non-defaulter. Thus four possible outcomes are possible. A common means of representing these outcomes is summarized in a *confusion matrix* as in Table 3.1.

		Actual	
		Default	Non-default
Model	below C	True Positives (correct prediction)	False Positives (type II error)
	above C	False Negatives (type I error)	True Negatives (correct prediction)

Table 3.1. *A confusion matrix describing the four possible outcomes of the decision problem introduced above in the text.*

Note that in our case, when the model produces credit scores (probabilities transformed to credit scores) instead of just let say two separate prediction classes, a specific confusion matrix is only valid for a certain model cut-off point. The cells in the confusion matrix represent the number of so called *true positives* (TP), *true negatives* (TN), *false positives* (FP) and *false negatives* (FN). The term TP indicates a predicted default that really occurs, a TN is predicted non-default that really occurs, a FP is a predicted default that does not occur and a FN is a predicted non-default in the case the company defaults at last. Note that the numbers on the major diagonal represents correct decisions, and the numbers on the off diagonal represents mistakes – the confusion between classes. The cell in which the number of FNs is present quantifies the *type I error* and the one with the number of FPs the *type II error*. There are several metrics that can be calculated from the confusion matrix and further used as indicators of model performance.

We define the *true positive rate* or equivalently called as the *hit rate (HR)* as

$$HR(C) = P(S_D \leq C), \tag{3.1}$$

which can be estimated as

$$\widehat{P}(S_D \leq C) = \frac{TP(C)}{n_D}, \tag{3.2}$$

where $TP(C)$ is the number of defaulters predicted correctly according to the cut-off value C and n_D is the total number of defaulters.

We define the *false positive rate* also called as the *false alarm rate (FR)* as

$$FR(C) = P(S_{ND} \leq C), \tag{3.3}$$

which can be estimated as

$$\widehat{P}(S_{ND} \leq C) = \frac{FP(C)}{n_{ND}}, \tag{3.4}$$

where $FP(C)$ is the number of non-defaulters classified incorrectly as defaulters according to cut-off value C and n_{ND} is the total number of non-defaulters. Note that it holds $HR(C) = 1 - P(\text{type I error})$ and $FR(C) = P(\text{type II error})$.

For graphical interpretation in Figure 3.1, we can express the true positive rate as the area on the left from the cut-off value C between the red and the blue line and the false positive rate as the lower area on the same side under the blue line.

A useful measure used in the validation process is the *Receiever Operating characteristic curve – ROC curve*. ROC curves represent a more general analysis of the confusion matrix providing information about the performance of a model at any admissible cut-off point. The ROC graphs are constructed in the following manner.

Continuous S_D ***and*** S_{ND}**.** Let us suppose that we are in the situation when S_D and S_{ND} follow continuous distributions, note that this case includes also the situation when we use the logistic regression model as our scoring tool. Now let us take an arbitrary cut-off value C and consider a point whose horizontal coordinate is $P(S_{ND} \leq C)$ and its vertical coordinate is $P(S_D \leq C)$. Denote this point by $I(C)$. Because of the fact that these two coordinates are represented in terms of probabilities, the point $I(C)$ lies always within a unit square graph. Imagine that for all possible cut-off values C ranging from $-\infty$ to ∞, there is a point $I(C)$ plotted on this graph. When C equals $-\infty$, $I(C)$ corresponds to the point having coordinates $(0, 0)$. As C is raised $I(C)$ generates a continuous curve that reaches the point $(1, 1)$ when C equals ∞. So when S_D and S_{ND} are assumed to be continuous in general it holds, that varying the cut-off values from $-\infty$ to ∞ and drawing a curve across the ROC space would produce the theoretical, continuous ROC curve. This curve is running from $(0, 0)$ to $(1, 1)$, as described in Figure 3.2 on the left.

Discrete S_D ***and*** S_{ND}**.** In this case we assume that S_D and S_{ND} follow a discrete distribution. Consider the set T, that is a set of r discrete values $\{s_1, \ldots, s_r\}$ which are ordered in the following way $-\infty < s_1 < \ldots < s_r < \infty$ as mentioned above. For the values $s_i, i = 1, \ldots, r$, it holds that either $P(S_D = s_i) > 0$ or $P(S_{ND} = s_i) > 0$, which means that each of the values included in set T could be assumed with positive probability. Finally put $s_0 = -\infty$ and $s_{r+1} = \infty$. Just like in the continuous case, consider an arbitrary cut-off value C and the corresponding point $I(C)$ in the ROC space with coordinates $P(S_{ND} \leq C)$ and $P(S_D \leq C)$. It holds that $I(s_0) \neq I(s_1) \neq \cdots \neq I(s_r)$ and $I(s_r) = I(s_{r+1})$. Next, note that for a cut-off value C which meets condition $s_i \leq C < s_{i+1}$, we put $I(C) = I(s_i), i = 0, \ldots, r$. It means, that no matter the fact that C could obtain infinite number of values, the corresponding ROC graph would only consist of $r+1$ distinct values $I(s_0), I(s_1), \ldots, I(s_r)$. Note that the points $I(s_0)$ and $I(s_r)$ equal $(0, 0)$ and $(1, 1)$, respectively. The graphical interpretation is presented again in Figure 3.2 on the right.

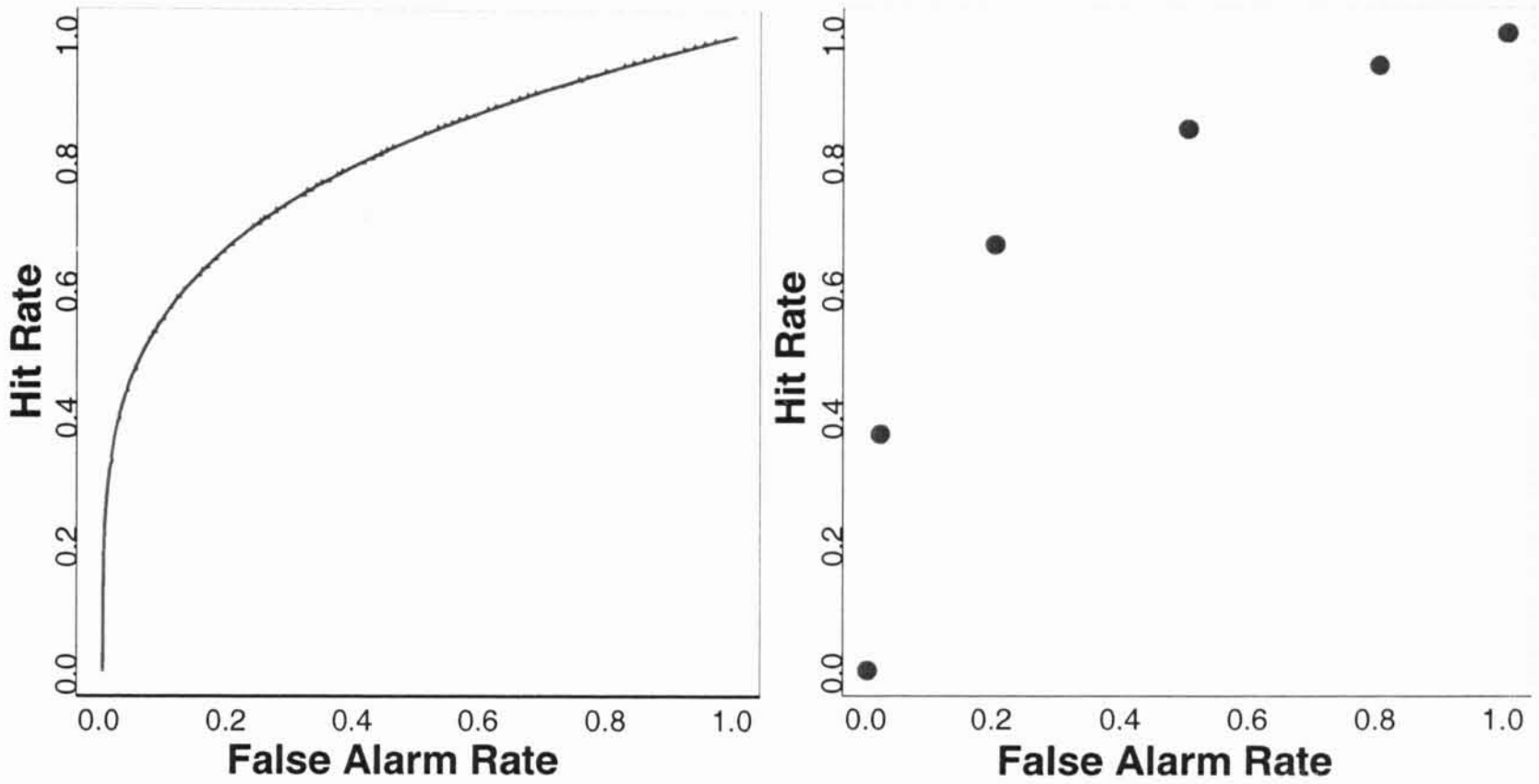


Figure 3.2. *Receiver Operating Characteristics. Left: The continuous case. Right: The discrete case.*

According to our notation $P(S_D \leq C) = HR(C)$, $P(S_{ND} \leq C) = FR(C)$, we can see that ROC graphs are two dimensional graphs that plot the *hit rate* on the vertical axis against the *false alarm rate* on the horizontal axis and as such, illustrate a relative trade-off between true positives and false positives.

So far, we have outlined the generation process by which the ROC graphs are produced, in real life situations we are not able to produce a continuous (theoretical) ROC curve. This is the reason why we would like to have a reasonable estimate of the theoretical equivalent. Thus it is useful to define a *sample ROC graph*. Suppose that we have N_D and N_{ND} observations of random variables S_D and S_{ND}, respectively. Recall that in 3.2, 3.4 we have denoted $\hat{P}(S_D \leq C)$ and $\hat{P}(S_{ND} \leq C)$ the proportion of N_D and N_{ND} observations respectively, that are less than or equal to the cut-off value C. Consider a point $\hat{I}(C)$ in the ROC space, having coordinates $\hat{P}(S_D \leq C)$ and $\hat{P}(S_{ND} \leq C)$. We define the *sample ROC graph* as a graph consisting of points $\hat{I}(C)$ for all C in the actual range of credit scores. The coordinates $\hat{P}(S_D \leq C)$ and $\hat{P}(S_{ND} \leq C)$ are unbiased estimates of $P(S_D \leq C)$ and $P(S_{ND} \leq C)$ which are actually the theoretical coordinates of $I(C)$. In this sense we can consider the sample ROC graph as an unbiased estimate of its theoretical equivalent. Note that in practical applications we are only able to get a finite sample of points in the ROC space, namely the sample ROC graph. To obtain a curve, we connect these points by linear interpolation.

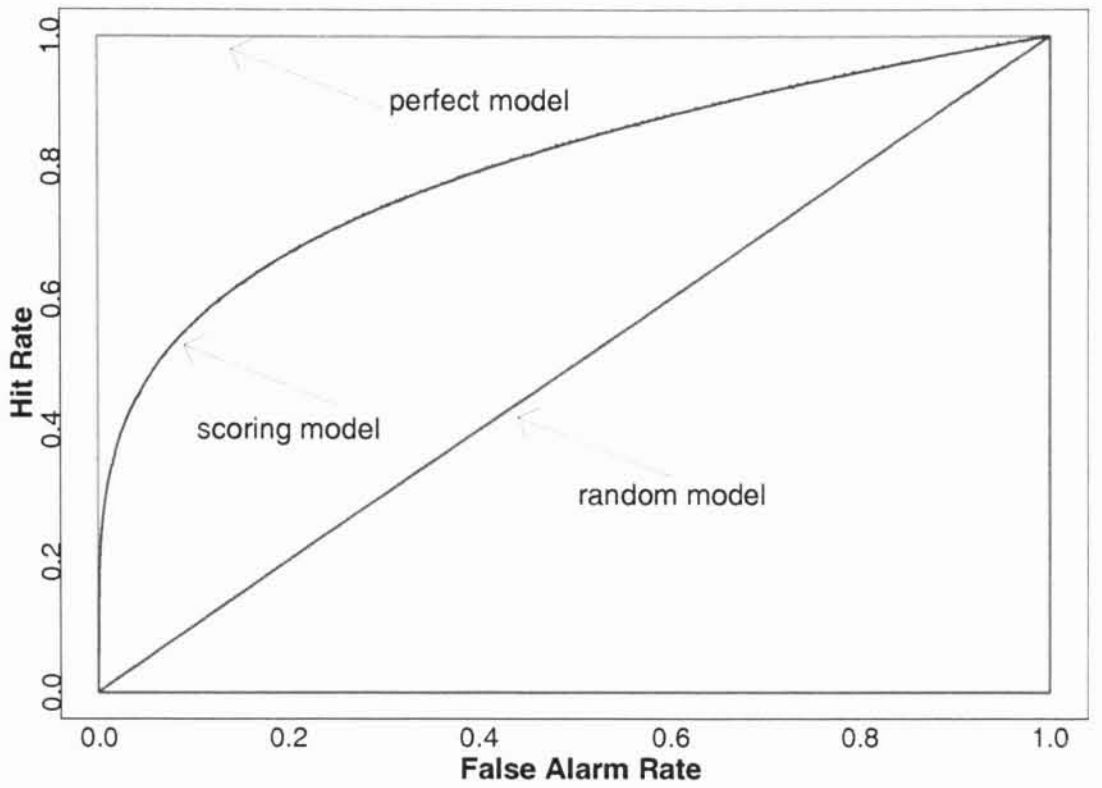


Figure 3.3. *Possible Receiver operating characteristics curves.*

Remark that several points in ROC graphs deserve further attention. Points $(0,0)$ and $(1,1)$ are always present in the ROC space since for $C < min_{s \in T}(s)$ we get $TP(C) = FP(C) = 0$ similarly for $C \geq max_{s \in T}(s)$ we get $TP(C) = n_D$ and $FP(C) = n_{ND}$. The point $(0,0)$ describes the case of having no positive classifications, in such a situation the decision rule produces no false positive errors but also provides no true positive cases. The opposite case describes granting positive classification absolutely.

Figure 3.3 outlines the possible shapes of ROC curves. The line labeled as the *scoring model*, represent the performance of a reasonable model under evaluation, the diagonal line labeled as *random model*, describes the state of zero information that means random assignment of credit score. For example if the model guesses the positive class half the time, it can be expected to get half of the positives and half of the negatives correctly, thus it does not separate the classes at all. Finally the curve rising vertically from $(0,0)$ to $(0,1)$ and then horizontally to $(1,1)$ represents the performance of the *perfect model* which orders all bad cases before good cases. It scores 100% bad cases into the default class and 0% good cases into the default class according to a specific cut-off value.

Finally, assume that we are validating a reasonable scoring model, in sense that lower values of rating scores assigned by the model indicate higher probability of default. We say that random variable X is stochastically smaller than random variable Y if for every constant C it holds, $P(X \leq C) \geq P(Y \leq C)$. Following up, random variables X and Y are stochastically comparable if X is stochastically smaller than Y or reversed. In our context it means that applying a reasonable scoring model, leads us to the conclusion that variables S_D and S_{ND} should be stochastically comparable, furthermore S_D should be stochastically smaller S_{ND},

which finally means, we await that the ROC curve of a reasonable scoring model should be concave.

2.1.2. Power Statistic. In order to have a summary statistics beside the complex ROC graph, one may wish to come up with a single scalar value representing the expected performance of the actual model. In the context of credit scoring model validation, such statistics are called the *power statistics*. An effective method for summarizing the ROC graphs is to calculate the *area under the ROC curve* as described in Figure 3.4. We denote this statistics as AUC.

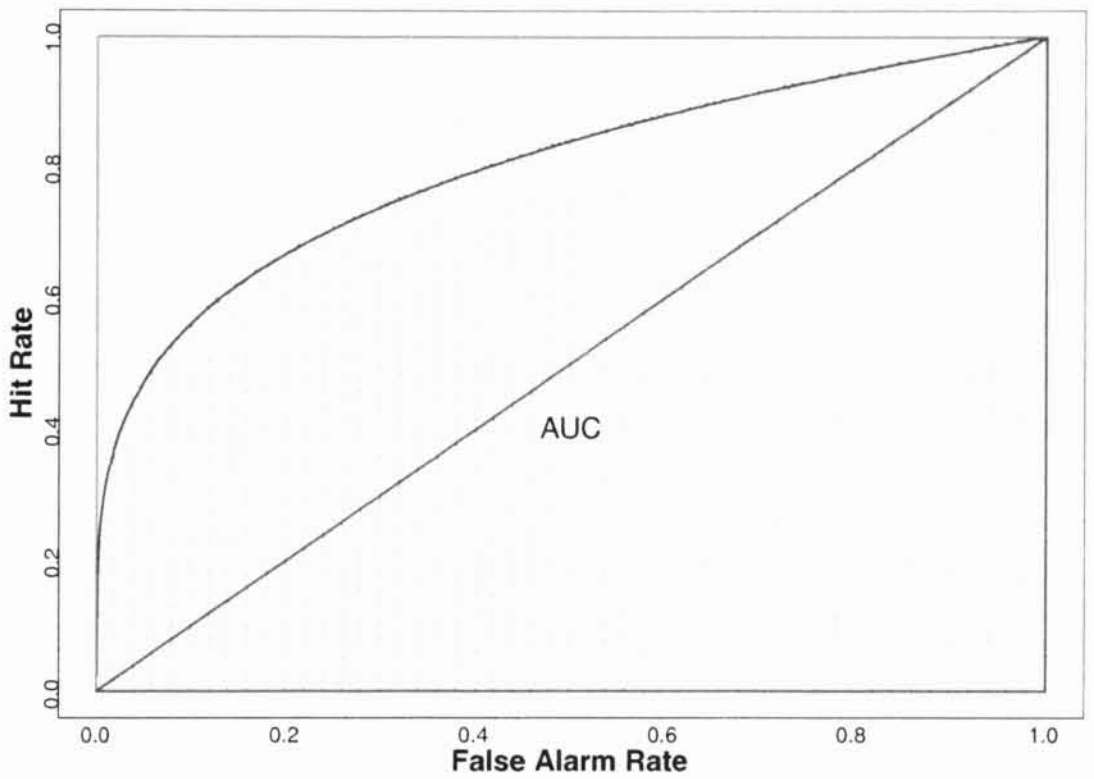


Figure 3.4. *Graphical interpretation of the area under the receiver operating characteristics*

Because of the fact that the AUC is a part of a unit square its value ranges between 0 and 1. With respect to the description of the random models ROC curve, it holds than the random models expected performance expressed in terms of AUC yields the value of 0.5. Thus the value 0.5 corresponds to a model with no discriminative power, on the other hand the perfect model has the AUC equal to 1. According to these rules we can state that any model its ROC curve appears in the lower right triangle of the ROC space has a value of AUC lower than 0.5, thus it could be possibly worse than random guessing. In this case we should be aware of the fact that the ROC space is symmetrical about the diagonal (random model's ROC curve). According to this, if we reverse the classification rule of the model, its true positives become false positives and vice versa. Such case should not occur during the model validation process because it would mean that we completely misinterpreted the meaning of the model, but is likely to occur while employing ROC analysis and its summary statistics during single factor analysis.

The following paragraphs contain a description of statistical properties of the AUC statistics, as well as a convenient interpretation of AUC in context of credit

scoring model validation. Let us assume that we have two obligors, one drawn from the distribution of defaulters and the other one from the distribution of non-defaulters. In this situation we obtain two credit scores that correspond to the realization of random variables S_D and S_{ND}. If we have to decide, using these values, which of the two obligors is a defaulter, we obviously state that the defaulter is the obligor with the lower credit score. In case both values are the same, we can decide at random. Thus the probability that our decision was correct equals $P(S_D < S_{ND}) + \frac{1}{2}P(S_D = S_{ND})$. We will examine how this probability relates to the AUC statistic.

Continuous S_D ***and*** S_{ND}**.** Let us suppose that we are again in the situation when S_D and S_{ND} follow a continuous distribution. In this case we are able to calculate the AUC statistics as follows

$$\begin{aligned} AUC &= \int_0^1 P(S_D \leq C)\, dP(S_{ND} \leq C) \\ &= \int_{\{S_D \leq S_{ND}\}} dP = P(S_D \leq S_{ND}). \end{aligned} \tag{3.5}$$

Because of the fact that S_D and S_{ND} are continuous, we have that $P(S_D = S_{ND})$ is zero, and that $AUC = P(S_D \leq S_{ND}) = P(S_D < S_{ND}) = P(\text{correct decision})$.

Discrete S_D ***and*** S_{ND}**.** In this case we assume that S_D and S_{ND} follow discrete distributions, and they obtain values $\{s_1, \ldots, s_r\}$ from the set T. We have shown that in this situation the corresponding ROC graph consist only of a finite number of points. So, the appropriate AUC statistics can not be calculated as in the continuous case. Thus we define the AUC statistics in the discrete case to be the area under the discrete ROC graph, where we connect the distinct points in the ROC space by linear interpolation. Figure 3.5 sketches a possibility, how to calculate the discussed area.

We can see that the AUC statistics can be expressed as a sum of certain trapezoids U_i. Specifying the proportions of these trapezoids, we have that the height of the i-th trapezoid U_i is equal to

$$P(S_{ND} \leq s_i) - P(S_{ND} \leq s_{i-1}) = P(S_{ND} = s_i)$$

and the lengths of corresponding edges are $P(S_D \leq s_i)$ and $P(S_D \leq s_{i-1})$. According to this, the volume of the i-th trapezoid is

$$U_i = P(S_{ND} = s_i)\left[\frac{1}{2}P(S_D \leq s_i) + \frac{1}{2}P(S_D \leq s_{i-1})\right].$$

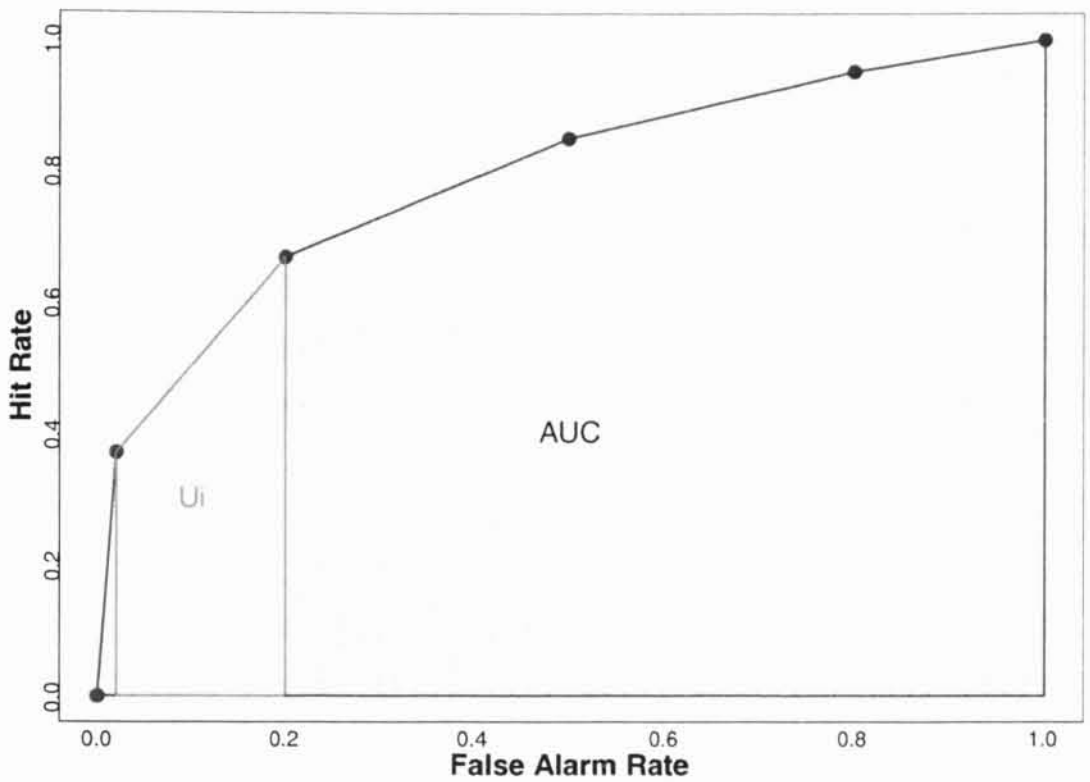


Figure 3.5. *The AUC statistics in the discrete case is defined as the area under the linearly interpolated discrete ROC graph. Trapezoids U_i provide a clear, and intuitive suggestion about computational matters.*

With respect to the fact that AUC is the sum of r trapezoids as described above we have that

$$\begin{aligned} AUC &= \sum_{i=1}^{r} \left[\frac{1}{2} P\left(S_D \leq s_i\right) + \frac{1}{2} P\left(S_D \leq s_{i-1}\right) \right] P\left(S_{ND} = s_i\right) \\ &= \sum_{i=1}^{r} \left[P\left(S_D \leq s_{i-1}\right) + \frac{1}{2} P\left(S_D = s_i\right) \right] P\left(S_{ND} = s_i\right) \qquad (3.6) \\ &= \sum_{i=1}^{r} P\left(S_D \leq s_{i-1}\right) P\left(S_{ND} = s_i\right) + \frac{1}{2} \sum_{i=1}^{r} P\left(S_D = s_i\right) P\left(S_{ND} = s_i\right) \\ &= P\left(S_D < S_{ND}\right) + \frac{1}{2} P\left(S_D = S_{ND}\right). \end{aligned}$$

Recall that in the continuous case we have that $P(S_D = S_{ND}) = 0$. This means that equation

$$P\left(S_D < S_{ND}\right) + \frac{1}{2} P\left(S_D = S_{ND}\right) \qquad (3.7)$$

holds for both the discrete and the continuous case. From this point, till the end of this subsection, we will consider both the continuous version 3.5 and the discrete version 3.6 and denote them jointly by AUC.

Now from these equations we are able to deduce two possible interpretation. First, we can state that the AUC statistics may be interpreted as the probability that the variable S_D yield a smaller credit score than variable S_{ND}. Second, recall the above described situation in which we have two randomly drawn obligors. One

from the distribution of non-defaulters and one from the distribution of defaulters and we have to decide which obligor belongs to each distribution. Intuitively we proclaim the obligor with the lower score as the one from the default distribution. When the scores are equal we can decide by chance. In this case and with respect to 3.6 we may interpret the AUC statistics as the probability that our decision is correct. Hence the AUC statistic is not just some quantity that gains certain possible values according to which we are able to deduce some properties about our scoring model, but it has a clear, stand alone interpretation. The key task is to determine the way in which we would like to employ this statistics. Again, we have two possibilities.

- First, AUC can be used as a measure of the size of the difference between two populations. In this sense AUC measures the extent to which the distribution of S_D lies below the distribution of S_{ND}. The ultimate values of AUC could be, in this case, interpreted as follows. The highest value, $AUC = 1$, could be attained if and only if the distribution of defaulter lies entirely below the distribution of non-defaulters, non of them overlapping each other. The smallest value, $AUC = 0$, could be obtained in the opposite case. Finally, if the two distributions are identical AUC equals 0.5. Thus, the closer the AUC is to zero or one, the larger the difference between the two sample populations, whereas the closer the AUC is to 0.5 the smaller the difference between them.

- Second and more importantly, according to our purpose, AUC can be used as a measure of discrimination accuracy. In this sense the AUC measures the extent how accurately a given model discriminates between defaulters and non-defaulters, and that is our main task. Recall Figure 3.3. Thus when $AUC = 1$, it means that the model discriminates the two sample populations perfectly. In other words it means that theoretically there exists a critical score/cut-off below which all the defaulters scores are, and above which all the non-defaulters scores are. So, when the AUC is close to 1, the actual model classifies obligors almost perfectly, reversely if the AUC is only a little above 0.5 then the model assigns the appropriate scores almost randomly.

2.1.3. Estimates and Confidence Intervals concerning AUC. There exist several possibilities how to estimate the AUC. Naturally, we can decide to employ a parametric or a nonparametric procedure. The latest could be proclaimed to be more popular. Mainly because of its relative simplicity, but moreover because of the fact that it doesn't require any distributional assumptions concerning variables S_D and S_{ND}.

Let us begin with the following idea. Assume that we have sampled n_D and n_{ND} observations of S_D and S_{ND}, respectively. So, we have $n_D \times n_{ND}$ possibilities of pairing these observations. Recall the definition of the sample ROC graph and similarly as in that definition denote $\hat{P}(S_D < S_{ND})$, $\hat{P}(S_D = S_{ND})$ and

$\hat{P}(S_D \neq S_{ND})$ the proportions of $n_D \times n_{ND}$ observations for which $S_D < S_{ND}$, $S_D = S_{ND}$ and $S_D \neq S_{ND}$. Again, it is clear that these proportions are unbiased estimates of their theoretical equivalents. The area under the Sample ROC graph denoted by $\widehat{AUC}$ can be calculated by the trapezoid rule as in 3.6 that is as follows

$$\widehat{AUC} = \hat{P}(S_D < S_{ND}) + \frac{1}{2}\hat{P}(S_D = S_{ND}).$$

Next for a randomly drawn defaulter with score s_D from S_D distribution and a non-defaulter with score s_{ND} from S_{ND} distribution we define the variable $v_{D,ND}$ as

$$v_{D,ND} = \begin{cases} 1, & \text{if} \quad s_D < s_{ND}, \\ \frac{1}{2}, & \text{if} \quad s_D = s_{ND}, \\ 0, & \text{if} \quad s_D > s_{ND}. \end{cases} \tag{3.8}$$

Now recall that the Man-Whitney (1947) U statistic is defined as the total number of pairs for which $S_D < S_{ND}$. According to this definition we can see that quantities AUC, $\widehat{AUC}$ and the Man-Whitney U statistic are closely related. In case we define the alternative Man-Whitney statistics $\hat{U}$ as

$$\hat{U} = \frac{1}{n_D n_{ND}} \sum_{(D,ND)} v_{D,ND} \ ,$$

where the sum is over all possible pairs of defaulters and non-defaulters, we observe that $\widehat{AUC}$ equals $\hat{U}$. Further we observe that $\hat{U}$ is an unbiased estimator of $P(S_D < S_{ND}) + \frac{1}{2}P(S_D = S_{ND})$, which means that $\hat{U}$ is an unbiased estimator of the theoretical AUC, and since $\hat{U}$ equals $\widehat{AUC}$ we have an alternative prove that $\widehat{AUC}$ is an unbiased estimate of the theoretical AUC. That is

$$AUC = \mathsf{E}(\hat{U}) = \mathsf{E}(\widehat{AUC}) = P(S_D < S_{ND}) + \frac{1}{2}P(S_D = S_{ND}).$$

According to these statements we are able to utilize statistical properties of the Man-Whitney statistic to predict statistical properties of the AUC statistic.

The variance of $\widehat{AUC}$**.** At the first place, we are going to lay emphasis on the estimation process of variance of the $\widehat{AUC}(= \hat{U})$ statistic. As outlined above certain results about the variance of Man-Whitney statistic are employed. There are several possibilities deriving formulas for the variance of this statistics under the assumption that S_D and S_{ND} are continuous. According to Bamber (1975) and his reference to Noether (1967) it is possible to relax this assumption. Thus for the variance $\sigma^2_{\widehat{AUC}}$ of $\widehat{AUC}$ we employ Bamber's formula

$$\begin{aligned} \sigma^2_{\widehat{AUC}} &= \frac{1}{4n_D n_{ND}}\Big[P_{D\neq ND} + (n_D - 1)P_{D,D,ND} \\ &+ (n_{ND} - 1)P_{ND,ND,D} - 4(n_D + n_{ND} - 1)(AUC - \frac{1}{2})^2\Big], \end{aligned} \tag{3.9}$$

where $P_{D\neq ND}$, $P_{D,D,ND}$ and $P_{ND,ND,D}$ are defined as follows. Assume we have sampled two independent observations from the S_{ND} distribution, these are $S_{ND,1}$

$S_{ND,2}$ and one independent observation from S_D distribution $S_{D,1}$. According to this $P_{ND,ND,D}$ is define as

$$\begin{aligned} P_{ND,ND,D} &= P(S_{ND,1}, S_{ND,2} < S_{D,1}) + P(S_{D,1} < S_{ND,1}, S_{ND,2}) \\ &- P(S_{ND,1} < S_{D,1} < S_{ND,2}) - P(S_{ND,2} < S_{D,1} < S_{ND,1}), \end{aligned}$$

$P_{D,D,ND}$ is define similarly by reversing the role of the independent observations sampled from the two populations as described above. For completeness we define $P_{D\neq ND}$ as $P(S_D \neq S_{ND})$. It is clear that for practical applications we would like to have a reasonable estimate of these probabilities and thus a estimate for the discussed variance $\sigma^2_{\widehat{AUC}}$. Similarly as in previous paragraphs consider triples as following $(S_{ND,1}, S_{ND,2}, S_{D,1})$ with independent $S_{ND,1}$ and $S_{ND,2}$. The total number of these triples is $n_{ND}(n_{ND}-1)n_D$. We denote the proportion of these triples for which $S_{ND,1}$, $S_{ND,2}$ are less than equal to $S_{D,1}$ by $\hat{P}(S_{ND,1}, S_{ND,2} < S_{D,1})$. $\hat{P}(S_{ND,1} < S_{D,1} < S_{ND,2})$ and $\hat{P}(S_{ND,2} < S_{D,1} < S_{ND,1})$ also represent the proportions in appropriate cases. So we have that

$$\begin{aligned} \hat{P}_{ND,ND,D} &= \hat{P}(S_{ND,1}, S_{ND,2} < S_{D,1}) + \hat{P}(S_{D,1} < S_{ND,1}, S_{ND,2}) \\ &- 2\hat{P}(S_{ND,1} < S_{D,1} < S_{ND,2}), \end{aligned}$$

is an unbiased estimate of $P_{ND,ND,D}$. From the computational point of view we can obtain the estimate $\hat{P}_{ND,ND,D}$ in the consecutive way. First of all, rank order the combined vector of S_D and S_{ND} score values. For each defaulters score value S_D, evaluate the number of non-defaulters score values S_{ND} that are less than S_D, and the number of score values that are greater and denote these values by a_D and b_D, respectively. $\hat{P}_{ND,ND,D}$ can be rewritten as

$$\hat{P}_{ND,ND,D} = \frac{1}{n_{ND}(n_{ND}-1)n_D} \sum_{(D)} \Big[a_D(a_D-1) + b_D(b_D-1) - 2a_D b_D \Big],$$

where we sum runs over all defaulters. Analogously, we can define the estimate of $P_{D,D,ND}$.

Up to this point, $\widehat{AUC}$, $\hat{P}(S_D \neq S_{ND})$, $\hat{P}_{ND,ND,D}$ and $\hat{P}_{D,D,ND}$ were unbiased estimates of their theoretical equivalents. Bamber (1975) outlined, that before substituting them into equation 3.9 we have to be aware that the expected value of $(\widehat{AUC} - \frac{1}{2})^2$ is $(AUC - \frac{1}{2})^2 + \sigma_{\widehat{AUC}^2}$. This bias could be corrected with multiplying equation 3.9 by $n_D n_{ND}/(n_D - 1)(n_{ND} - 1)$. In this way we obtain a unbiased estimate $\hat{\sigma}^2_{\widehat{AUC}}$ of $\sigma^2_{\widehat{AUC}}$

$$\begin{aligned} \hat{\sigma}^2_{\widehat{AUC}} &= \frac{1}{4(n_D-1)(n_{ND}-1)} \Big[\hat{P}_{D\neq ND} + (n_D-1)\hat{P}_{D,D,ND} \\ &+ (n_{ND}-1)\hat{P}_{ND,ND,D} - 4(n_D + n_{ND} - 1)(\widehat{AUC} - \frac{1}{2})^2 \Big]. \end{aligned} \tag{3.10}$$

Confidence interval for the AUC ***statistics.*** It is known that if n_D, n_{ND} are held in constant ration, then if $n_D, n_{ND} \to \infty$, statistics $(AUC - \widehat{AUC})/\hat{\sigma}^2_{\widehat{AUC}}$ is asymptotically normally distributed with mean zero and standard deviation one Bamber (1975). According to this statement and because of the fact that sample

sizes employed in credit modelling are very large, we are able to compute confidence intervals for AUC. Thus the asymptotic confidence interval at level $1-\alpha, \alpha \in (0,1)$ has the form

$$\left[\widehat{AUC} - \hat{\sigma}_{\widehat{AUC}}\Phi^{-1}\left(1-\frac{\alpha}{2}\right), \widehat{AUC} + \hat{\sigma}_{\widehat{AUC}}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right]$$

where Φ^{-1} represents the quantile function of the standard normal distribution.

Comparing the areas under ROC graphs. A common situation that we usually face while performing model building and validation is, to choose the best suiting model for our problem under consideration. We usually build more than just one scoring model and at this stage of the validation process we would like to choose the most suitable one, namely, the most powerful in terms of the AUC statistics. In this context the primary application of ROC analysis is the comparison of different scoring models applied to to the same data. In the simplest case we are interested in comparing the discrimination ability of two scoring models A and B. Denote the corresponding areas under the ROC curves by AUC_A and AUC_B. Formally speaking, we would like to test the hypothesis

$$\mathrm{H}_2\colon\ AUC_A = AUC_B,$$

against the alternative

$$\mathrm{A}_2\colon\ AUC_A \neq AUC_B$$

Comparing the appropriate estimated values $\widehat{AUC}_A$ and $\widehat{AUC}_B$ and choosing the final model only upon these quantities would not be appropriate from the statistical point of view. To be precise and to derive a reasonable test on the difference between two AUC statistics, namely AUC_A and AUC_B, we have to calculate the variances $\hat{\sigma}^2_{\widehat{AUC}_A}$ and $\hat{\sigma}^2_{\widehat{AUC}_B}$ of estimators $\widehat{AUC}_A$ and $\widehat{AUC}_B$. Because of the fact that two models for which we have estimated the AUC statistics might be correlated, we also need to compute the covariance between the two estimators $\widehat{AUC}_A$, $\widehat{AUC}_B$. With respect to previous notation and according to Delong et al. (1988) and Engelmann, Hayden, Tasche (2002) for the estimate of $\sigma^2_{\widehat{AUC}_A,\widehat{AUC}_B}$ we find that

$$\begin{aligned}\hat{\sigma}^2_{\widehat{AUC}_A,\widehat{AUC}_B} &= \frac{1}{4(n_D-1)(n_{ND}-1)}\Big[\hat{P}^{AB}_{D,D,ND,ND} \\ &+ (n_D-1)\hat{P}^{AB}_{D,D,ND} + (n_{ND}-1)\hat{P}^{AB}_{ND,ND,D} \\ &- 4(n_D+n_{ND}-1)(\widehat{AUC}_A - \frac{1}{2})(\widehat{AUC}_B - \frac{1}{2})\Big]. \qquad (3.11)\end{aligned}$$

where $\hat{P}^{AB}_{D,D,ND,ND}$, $\hat{P}^{AB}_{D,D,ND}$ and $\hat{P}^{AB}_{ND,ND,D}$ are estimators for probabilities $P^{AB}_{D,D,ND,ND}$, $P^{AB}_{D,D,ND}$, $P^{AB}_{ND,ND,D}$ defined as follows

$$\begin{aligned}
P^{AB}_{D,D,ND,ND} &= P(S^A_D > S^A_{ND}, S^B_D > S^A_{ND}) + P(S^A_D < S^A_{ND}, S^B_D < S^B_{ND}) \\
&- P(S^A_D > S^A_{ND}, S^B_D < S^B_{ND}) - P(S^A_D < S^A_{ND}, S^B_D > S^B_{ND}), \\
P^{AB}_{D,D,ND} &= P(S^A_{D,1} > S^A_{ND}, S^B_{D,2} > S^B_{ND}) + P(S^A_{D,1} < S^A_{ND}, S^B_{D,2} < S^B_{ND}) \\
&- P(S^A_{D,1} > S^A_{ND}, S^B_{D,2} < S^B_{ND}) - P(S^A_{D,1} < S^A_{ND}, S^B_{D,2} > S^B_{ND}), \\
P^{AB}_{ND,ND,D} &= P(S^A_D > S^A_{ND,1}, S^B_D > S^B_{ND,2}) + P(S^A_D < S^A_{ND,1}, S^B_D < S^B_{ND,2}) \\
&- P(S^A_D > S^A_{ND,1}, S^B_D < S^B_{ND,2}) - P(S^A_D < S^A_{ND,1}, S^B_D > S^B_{ND,2}).
\end{aligned}$$

Quantities S^A_D, S^B_D, $S^A_{D,1}, S^B_{D,1}$ and $S^A_{D,2}, S^B_{D,2}$ are observations independently drawn from the distribution of defaulters. Similarly the ones labeled by ND are observations independently drawn from the distribution of non-defaulters. Finally the testing procedure constitutes of evaluating statistics M defined as

$$M = \frac{(\widehat{AUC}_A - \widehat{AUC}_B)^2}{\hat{\sigma}^2_{\widehat{AUC}_A} + \hat{\sigma}^2_{\widehat{AUC}_A} - 2\hat{\sigma}^2_{\widehat{AUC}_A,\widehat{AUC}_B}}, \tag{3.12}$$

which is asymptotically χ^2-distributed with one degree of freedom. The appropriate critical values are calculated from the $\chi^2(1)$ distribution given confidence level $\alpha \in (0,1)$.

Finally we provide an alternative method for estimation of the joint covariance of estimators $\widehat{AUC}_A$, $\widehat{AUC}_B$ which might be quite useful for computer implementation. In the first place we define quantities $V(s_D)$, $V(s_{ND})$ as placements of scores s_D and s_{ND} in the distributions of S_D and S_{ND}, respectively. It means that $V(s_D)$ is the the placement of score value s_D in the distribution of S_{ND} and $V(s_{ND})$ is the placement of score value s_{ND} in the distribution of S_D. In other words $V(s_D)$ is the fraction of S_{ND} scores that exceed it, similarly $V(s_{ND})$ is the fraction of S_D scores that it exceeds. So we evaluate these two quantities as

$$V(s_{D,j}) = \frac{1}{n_{ND}} \sum_{k=1}^{n_{ND}} v_{(D,j),(ND,k)}, \quad j = 1, \ldots, n_D,$$

and

$$V(s_{ND,k}) = \frac{1}{n_D} \sum_{j=1}^{n_D} v_{(D,j),(ND,k)}, \quad k = 1, \ldots, n_{ND}.$$

Note that the nonparametric estimate of the AUC statistic denoted $\widehat{AUC}$ is the average of placement values in both cases

$$\widehat{AUC} = \frac{\sum_{k=1}^{n_{ND}} V(s_{ND,k})}{n_{ND}} = \frac{\sum_{j=1}^{n_D} V(s_{D,j})}{n_D}$$

According to Delong et al. (1988) and with reference to the method of structural components Sen (1960), the joint covariance is computed as the sum of scaled

covariances of placement values for defaulting and non-defaulting scores. Thus, we have

$$\begin{aligned}\hat{\sigma}^2_{\widehat{AUC}_A,\widehat{AUC}_B} &= \frac{\sum_{j=1}^{n_D}\left[V(s^A_{D,j})-\widehat{AUC}_A\right]\left[V(s^B_{D,j})-\widehat{AUC}_B\right]}{n_D(n_D-1)} \\ &+ \frac{\sum_{k=1}^{n_{ND}}\left[V(s^A_{ND,j})-\widehat{AUC}_A\right]\left[V(s^B_{ND,k})-\widehat{AUC}_B\right]}{n_{ND}(n_{ND}-1)},\end{aligned}$$

letters A and B indicate the calculation of the respective placements values in the appropriate model.

2.2. Cumulative accuracy profile

Another concept which is currently popular in practise for evaluation the discriminative power of scoring models and which is similar to the receiver operating characteristics is the *cumulative accuracy profile (CAP) curve*. In this subsection we focus on the genesis of the CAP graph its interpretation. Further we derive an analytical relationship between the summary statistic related to the CAP curve, the *accuracy ratio (AR)* and the AUC statistic. This relationship demonstrates how statistical properties of the AUC statistic can be used in determining statistical properties of the AR summary statistic.

In order to comply with our notation assume again that we generally have n obligors and let us consider a model that assigns each obligor a score s out of a specific set T. T might be either a finite set of r discrete values $\{s_1,\ldots,s_r\}$; $s_1 < \ldots < s_r$, or it might be a continuous interval, say $[0,1]$. In general assume that the score values are ranging from worst to best. It means that a high score indicates a low default probability. Further we introduce random variables S_T and we consider again the random variables S_D and S_{ND}. The random variable S_T follow the score distribution of all obligors and recall that S_D and S_{ND} follow the score distribution of defaulters and non-defaulters, respectively. The cumulative accuracy profile is defined as the graph consisting of points, whose horizontal coordinate is defined as $P(S_T \leq C)$ and vertical coordinate is defined as $P(S_D \leq C)$, where C runs across the finite set of discrete values $\{s_1,\ldots,s_r\}$ or across the range of possible score values. Realize that in case T is a continuous interval, i.e. the model produces continuous score values, we have that $r = n$, because the probability that two different obligors obtain the same score value equals zero. Further note, that varying the score values smoothly from $C = min_{s\in T}(s)$ to $C = max_{s\in T}(s)$ and computing the appropriate coordinates produces the theoretical CAP curve. However, in practise we are able to assess again only a finite number of points, so the curve is obtained again by linear interpolation between them.

A perfect model assigns the lowest score values to the defaulting obligors. In this case the CAP curve is increasing linearly and staying at one. For a random model without any discriminative power, we have that the fraction k of all debtors with the lowest rating scores contains k percent of all defaulters, thus this is the case when $P(S_T \leq C) = P(S_D \leq C), C \in T$. Based on this interpretation, one can

also conceive of a perfect model which gives all defaults worse scores than non-defaults, and a random or uninformative model, which excludes defaults at the same rate as non-defaults. Reasonable scoring models are somewhere in between these two extremes.

The discriminative ability of the scoring model can be again summarized by a single number, the accuracy ratio AR. It is defined as the ratio of the area AS between the CAP curve of the scoring model under evaluation and the CAP curve of the random model, compared to the area AP between the CAP curve of the perfect model and the CAP curve of the random model. Formally speaking, it follows for AR that

$$AR = \frac{AS}{AP}. \tag{3.13}$$

In these sense the scoring model is the better the closer the AR is to one. Graphical interpretation of the later statements is provided in Figure 3.13

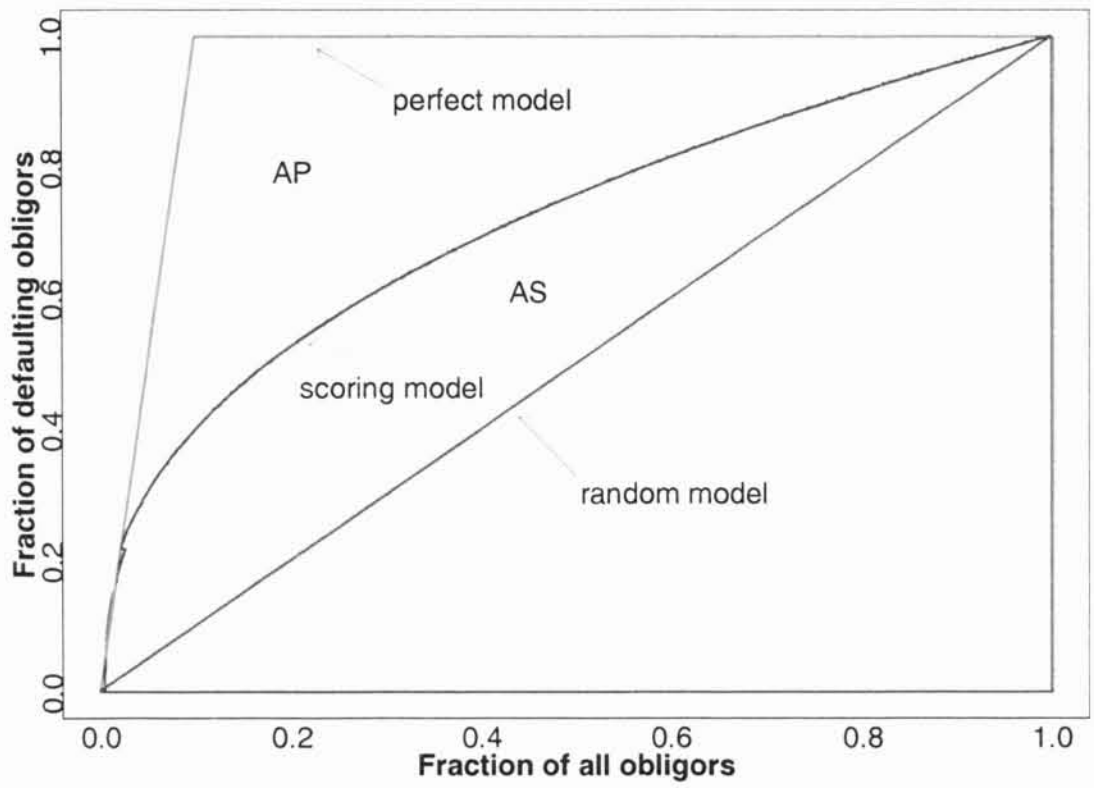


Figure 3.6. *Cumulative accuracy profile. The blue line shows the performance of a model under evaluation, it depicts the percentage of defaulting obligors identified by the model at different percentages of the total number of obligors. The diagonal line represents the state of random assignment of score values. The accuracy ratio is defined as the ratio of area AS and AP.*

Further we are going to illustrate that the statistical properties of ROC curves and the related summary statistics AUC are also applicable to the CAP curve and its summary statistic AR. The key relationship between the two performance measures is determined by the following formula

$$AR = 2AUC - 1. \tag{3.14}$$

The above presented relationship 3.14 can be derived as follows. Assume that the number of defaulting and non-defaulting obigors is n_D and n_{ND}, respectively. Obviously for the total number of obligors n, we have $n = n_D + n_{ND}$. For AP we find that

$$\begin{aligned} AP &= \frac{n_{ND}}{n_{ND}+n_D} + \frac{n_D}{2(n_{ND}+n_D)} - \frac{1}{2} \\ &= \frac{1}{2}\frac{n_{ND}}{n_{ND}+n_D}. \end{aligned} \tag{3.15}$$

In order to compute the AS we need to express the cumulative distribution function $P(S_T \leq C)$. In terms of S_D and S_{ND} the cumulative distribution function $P(S_T \leq C)$ can be expressed as

$$P(S_T \leq C) = \frac{n_D}{n_{ND}+n_D} P(S_D \leq C) + \frac{n_{ND}}{n_{ND}+n_D} P(S_{ND} \leq C), \tag{3.16}$$

where $n_D/(n_{ND}+n_D)$ is the prior default probability of all obligors and $n_{ND}/(n_{ND}+n_D)$ equals one minus this probability. Employing expression 3.16 for AS we find that

$$\begin{aligned} AS &= \int_0^1 P(S_D \leq C) dP(S_T \leq C) - \frac{1}{2} \\ &= \frac{n_D}{n_{ND}+n_D} \int_0^1 P(S_D \leq C) dP(S_D \leq C) \\ &+ \frac{n_{ND}}{n_{ND}+n_D} \int_0^1 P(S_D \leq C) dP(S_{ND} \leq C) - \frac{1}{2} \\ &= \frac{n_D \frac{1}{2} + n_{ND} AUC}{n_{ND}+n_D} - \frac{1}{2} = \frac{n_{ND}(AUC - \frac{1}{2})}{n_{ND}+n_D}. \end{aligned} \tag{3.17}$$

Substituing 3.15 and 3.17 into equation 3.13 we find that

$$AR = \frac{AS}{AP} = \frac{n_{ND}(AUC - \frac{1}{2})}{\frac{1}{2} n_{ND}} = 2AUC - 1. \tag{3.18}$$

This means that the accuracy ratio can be computed directly from the area under the ROC curve and vice versa. Thus the statistical properties of the accuracy ratio can be also derived from the statistical properties of the AUC statistics.

We conclude this paragraph providing comparison of the interpretation of the two concepts described up to this point. We have shown that the above described performance statistics are equivalent in the sense of equation 3.18, however the curves answer slightly different questions.

- ROC curves answer the question: What percentage of non-defaulters would a model have to exclude to exclude a specific percentage of defaulters?
- CAP curves answer the question: What percentage of an entire portfolio would a model have to exclude to avoid a specific percentage of defaulters?

Although CAP curves are the representation typically used in practice by finance professionals and business people, the research concerning statistical properties

was traditionally performed in the context of ROC curves, primarily because of their practical application also in medical research communities.

2.3. The Kolmogorov-Smirnov statistic

An important measure that is also commonly used in practise for credit scoring model comparison and validation is the Kolmogorov-Smirnov statistic. Notice that both ROC and CAP curves (and the related summary statistics AUC and AR), as well as the Kolmogorov-Smirnov statistic measure the models ability to discriminate between defaulting and non-defaulting obligors. In other words these discriminative measures address model's capability of proper ordering of obligors in terms of probability of default. Since the obligors probability of default is determined by the corresponding score value produced by the model, we expect a proper ordering in terms of score values. If we assume that higher score values indicate lower default probability, we expect that the distribution of defaulting obligors is shifted to the left from the distribution of non-defaulting obligors. Denote the distribution functions of defaulting and non-defaulting obligors by $F(x)$ and $G(x)$, respectively. As outlined before, a reasonable scoring model should clearly separate the defaulting and non-defaulting cases, thus the corresponding distributions should be clearly shifted from each other.

The Kolmogorov-Smirnov statistic is used to test the hypothesis that the distribution function of score values of defaulting and non-defaulting obligors are identical against a general alternative that they are different.

$$\mathrm{H}_3\colon\ F(x) = G(x),$$

$$\mathrm{A}_3\colon\ F(x) \neq G(x).$$

The Kolmogorov-Smirnov statistic is related to the supremum distance between distribution functions. The Kolmogorov-Smirnov distance between two distribution functions F and G, is defined as

$$KS_{dist} = \sup_{-\infty<x<+\infty} |F(x) - G(x)|\,. \tag{3.19}$$

With respect to general properties of distribution functions, it is clear that

$$0 \leq KS_{dist} \leq 1.$$

In case $KS_{dist} = 0$ then $F(x) = G(x)$ for all x and so, both distributions are identical. In case $KS_{dist} = 1$ then $F(x) = 0$ and $G(x) = 1$ or vice versa for some x. It implies that that either $F(x) \geq G(x)$ or $F(x) \leq G(x)$ for all x. Recall that if certain random variables X and Y possess distribution functions F and G, respectively, and $F \geq G$ then X is stochastically smaller than Y. Indeed, $F(x) \geq G(x)$ means $P(X \leq x) \geq P(Y \leq x)$.

In the context of credit scoring model validation, the Kolmogorov-Smirnov statistic is understood to be the sample version of Kolmogorov-Smirnov distance between empirical distribution functions corresponding to default and non-default samples. To be more precise we denote by $F_{n_D}(x)$ and $G_{n_{ND}}(x)$ the empirical distribution functions of credit scores corresponding to n_D defaulting and n_{ND}

non-defaulting obligors.

$$KS = \sup_{-\infty<x<+\infty} |F_{n_D}(x) - G_{n_{ND}}(x)| . \tag{3.20}$$

Because of the fact that empirical distribution functions have bounded supports, supremum can be replaced by maximum in 3.20

$$KS = \max_{-\infty<x<+\infty} |F_{n_D}(x) - G_{n_{ND}}(x)| . \tag{3.21}$$

Let us consider the following characteristics

$$KS_{sup} = \sup_{-\infty<x<+\infty} \big(F(x) - G(x)\big) \tag{3.22}$$

and

$$KS_{inf} = \inf_{-\infty<x<+\infty} \big(F(x) - G(x)\big) \tag{3.23}$$

Thus it holds

$$0 \leq KS_{sup} \leq 1$$

and

$$-1 \leq KS_{inf} \leq 0.$$

Note that if $KS_{sup} = 0$ then $F(x) \leq G(x)$ for all x while if $KS_{inf} = 0$ $F(x) \geq G(x)$ for all x. Also the opposite ultimate situation when $KS_{sup} = 1$ and $KS_{inf} = -1$ distinguish the two cases described above in connection with $KS_{dist} = 1$. Finally we have that

$$KS_{dist} = \max\left(KS_{sup}, KS_{inf}\right) .$$

Again, we are able to obtain the empirical versions of 3.22 and 3.23, similarly supremum and infimum can be replaced by maximum and minimum, respectively. These two empirical characteristics provide better notion about possible order relationship between $F(x)$ and $G(x)$ because once we do not have either $F(x) \leq G(x)$ or $F(x) \geq G(x)$ for all x then the empirical version of KS_{dist} does not measure the grade of stochastic ordering between the corresponding samples. In the context of credit scoring model validation it means that we do not recognize the situation when the model is good only for some regions of the predictor vector. That is why the whole graph of $F_{n_D}(x) - G_{n_{ND}}(x)$ should accompany the figure of the Kolmogorov-Smirnov statistic.

It is necessary to outline that formula 3.21 is not a convenient solution for computational matters concerning Kolmogorov-Smirnov statistics since it requires too many computational operations. That's why we employ a different expression introduced below,which shows that Kolmogorov-Smirnov statistic can be viewed as a rank statistic see Antoch, Vorlíčková (1992), on condition that $F(x)$ and $G(x)$ are continuous. Moreover, there is a modification for the discrete case. Recall that empirical distribution functions $F_{n_D}(x)$ and $G_{n_{ND}}(x)$ which correspond to

obligors' scores $X_1, \ldots, X_{n_D}$, and non-defaulting obligors' scores $X_{n_D+1}, \ldots, X_n$ where $n = n_D + n_{ND}$ can be expressed as

$$F_{n_D}(x) = \frac{1}{n_D} \sum_{i=1}^{n_D} u(x - X_i)$$

and

$$G_{n_{ND}}(x) = \frac{1}{n_{ND}} \sum_{i=n_D+1}^{n} u(x - X_i)$$

where $u(x) = 1, x \geq 0, u(x) = 0, x < 0$. The sums count the number of defaulted obligors' scores or not defaulted obligors' scores which are smaller than or equal to x, respectively.

Further let random variables $U_1, \ldots, U_{n_D+n_{ND}}$ be the order statistics of the joint default and non-default sample $X_1, \ldots, X_{n_D}, X_{n_D+1}, \ldots, X_{n_D+n_{ND}}$, thus, U_k is the k-th largest value within the joint sample. We define variables $Z_1, \ldots, Z_{n_D+n_{ND}}$ as indicators of the fact that $U_1, \ldots, U_{n_D+n_{ND}}$ belong to the default sample. It means that $Z_k = 1$ if U_k corresponds to some of $X_1, \ldots, X_{n_D}$ and $Z_k = 0$ if U_k corresponds to some of $X_{n_D+1}, \ldots, X_n$. The Kolmogorov-Smirnov statistic can be then expressed as

$$KS = \max_{1 \leq k \leq n_D+n_{ND}} \left| \frac{n_D + n_{ND}}{n_D n_{ND}} \sum_{i=1}^{k} Z_i - \frac{k}{n_{ND}} \right|. \tag{3.24}$$

The derivation of 3.24 works as follows

$$\begin{aligned}
KS &= \max_{-\infty<x<+\infty} |F_{n_D}(x) - G_{n_{ND}}(x)| \\
&= \max_{-\infty<x<+\infty} \left| \frac{1}{n_D} \sum_{i=1}^{n_D} u(x - X_i) - \frac{1}{n_{ND}} \sum_{i=n_D+1}^{n} u(x - X_i) \right| \\
&= \max_{-\infty<x<+\infty} \left| \frac{1}{n_D n_{ND}} \left[n_{ND} \sum_{i=1}^{n_D} u(x - X_i) - n_D \sum_{i=n_D+1}^{n} u(x - X_i) \right] \right| \\
&= \max_{-\infty<x<+\infty} \left| \frac{1}{n_D n_{ND}} \left[(n - n_D) \sum_{i=1}^{n_D} u(x - X_i) - n_D \sum_{i=n_D+1}^{n} u(x - X_i) \right] \right| \\
&= \max_{-\infty<x<+\infty} \left| \frac{1}{n_D n_{ND}} \left[n \sum_{i=1}^{n_D} u(x - X_i) - n_D \sum_{i=1}^{n} u(x - X_i) \right] \right|
\end{aligned}$$

Since the empirical distribution functions are defined for a finite number of points, we can take the maximum over a finite set of these values $X_k, k \in \{1, \ldots, n_D + n_{ND}\}$. Because we take the maximum over all possible values without loss of generality we can assume that we take it over an order set of values $X_{(1)} \leq X_{(k)} \leq X_{(n_D+n_{ND})}$

$$KS = \max_{X_{(1)}\leq X_{(k)}\leq X_{(n_D+n_{ND})}} \left| \frac{1}{n_D n_{ND}} \left[n \sum_{i=1}^{n_D} u(X_{(k)} - X_i) - n_D \sum_{i=1}^{n} u(X_{(k)} - X_i) \right] \right|$$
$$= \max_{X_{(1)}\leq X_{(k)}\leq X_{(n_D+n_{ND})}} \left| \frac{n}{n_D n_{ND}} \sum_{i=1}^{n_D} u(X_{(k)} - X_i) - \frac{1}{n_{ND}} \sum_{i=1}^{n} u(X_{(k)} - X_i) \right|$$

The first sum in the latest equation equals the number of defaulting score values which are less or equal than the the $k-$th ordered statistic. The second term is the order of the $k-$th ordered statistic, which is k. So we find that

$$KS = \max_{1\leq k\leq n_D+n_{ND}} \left| \frac{n_D + n_{ND}}{n_D n_{ND}} \sum_{i=1}^{k} Z_i - \frac{k}{n_{ND}} \right|. \tag{3.25}$$

Formula 3.24 works well if F and G are continuous, then variables Z_k are well defined with probability 1, because ties, that is, observations of the same value, have zero probability. Precisely it means that $P(X_i = X_j) = 0$ for arbitrary i and j.

Finally realize that in credit modelling we can easily observe ties, for instance if we check categorial predictors or if we deal with categorial models. If equal observations can occur we have to adjust formula 3.24 so that ties are gathered in one case. To do so let $\tilde{U}_1 < \ \ldots < \tilde{U}_j$ be distinct values of realizations of order statistics $U_1, \ldots, U_{n_D+n_{ND}}$, thus, $j \leq n_{ND}+n_D$. Instead of Z_k consider $\tilde{Z}_k$ equal to the number of X_i such that $X_i = \tilde{U}_k$, $k = 1, \ldots, j$, and define K_k as the number of all observations (X_i and Y_i) which are less then or equal to $\tilde{U}_k$, $k = 1, \ldots, j$. The adjustment of formula 3.24 can be written as

$$KS = \max_{1\leq k\leq j} \left| \frac{n_D + n_{ND}}{n_D n_{ND}} \sum_{i=1}^{k} \tilde{Z}_i - \frac{K_k}{n_{ND}} \right|. \tag{3.26}$$

Formula 3.26 is computationally more efficient formula for calculating Kolmogorov-Smirnov statistics. As it was outlined above it is useful to accompany the value of the Kolmogorov-Smirnov statistics with a graph obtained by plotting

$$\tilde{K}_k \quad \text{against} \quad \tilde{V}_k = \frac{n_D + n_{ND}}{n_D n_{ND}} \sum_{i=1}^{k} \tilde{Z}_i - \frac{K_k}{n}, \quad k = 0, \ldots, j. \tag{3.27}$$

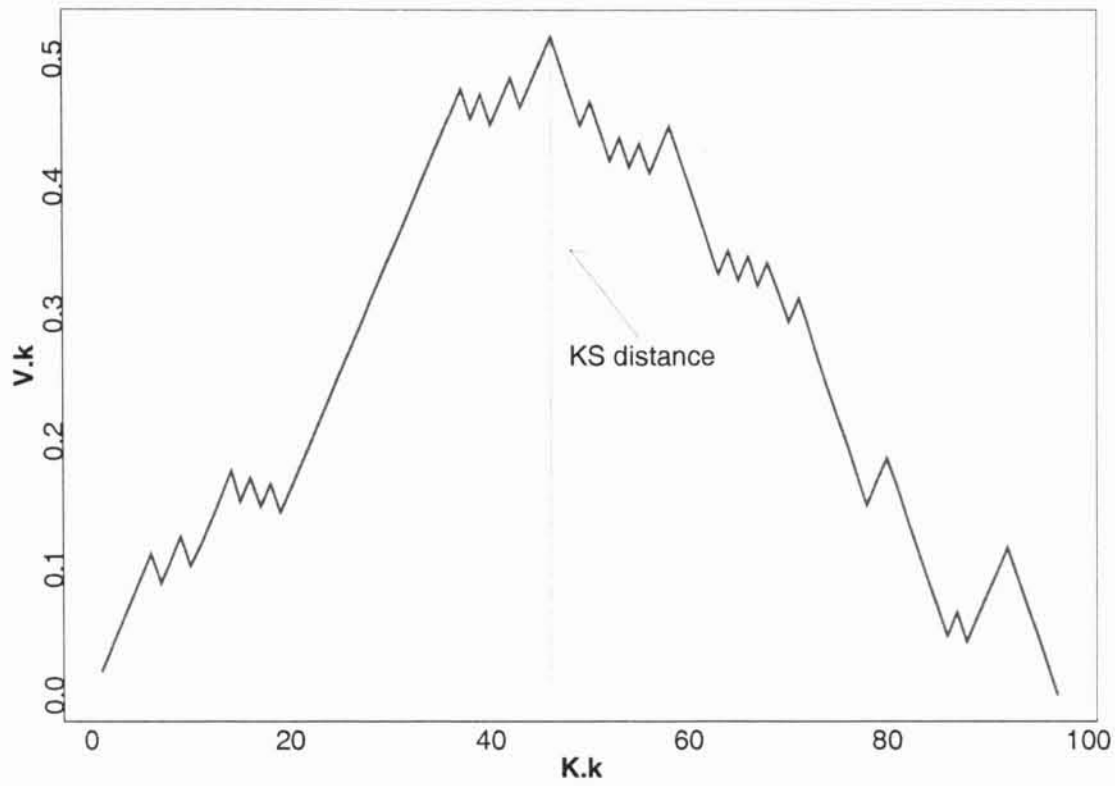


Figure 3.7. *The figure depicts the distance between empirical distribution functions of the defaulting and non-defaulting obligors. The maximum of these distances is the the sample version of the Kolmogorov-Smirnov statistics.*

Chapter IV

Summary

The present technical literature on corporate credit risk modelling is scarce. Several academical papers deal with this issue, however most of them focus only on certain area of the overall problem. The academic research in this area is substantially limited by the unavailability of public data. In contrast with similar papers on credit scoring this thesis intend to provide a self-contained concept of the statistical methodology related to the overall problem, instead of an empirical study. The main purpose of the thesis was to develop the statistical methodology whose application leads to reasonable credit scoring models.

The first part of the document proposes statistical methods that should be employed within the process of determining a reasonable set of predictor variables which could be later used in multivariate modelling. Here we propose an adjustment of the standard Pearson χ^2. We show that this adjustment leads to a reasonable sample measure of dependence.

The second part of the thesis is focused on model validation techniques. Measures of model performance are described and their statistical properties are investigated. Emphasis is put on the concept of the receiver operating characteristics and the related summary statistics, the area under the receiver operating characteristics. Further the relationship of the receiver characteristics analysis and the cumulative accuracy profile analysis in described. Statistical properties of the later concepts are reviewed and the computational aspect are discussed.

Finally, we focus on a different concept of a model performance measure. Namely, the Kolmogorov-Smirnov statistics. We derive alternative formula for the computation of the Kolmogorov-Smirnov statistics which shows that the Kolmogorov-Smirnov statistics can be viewed as a rank statistics and which is suitable for computer implementation.

Appendix A

Regression Models

1. Introduction of the model classes

First we restrict our attention to the standard multiple regression problem. We have n observations of a response (dependent) variable Y, denoted by $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ measured at n design vectors $\boldsymbol{x}_i^\top = (x_{i1}, \ldots, x_{im})$. The points $\boldsymbol{x}_i^\top$ may be chosen in advance, or may be themselves measurements of random variables vector $\boldsymbol{X} = (X_{i1}, \ldots, X_{im})$. We do not distinguish these two cases.

Our goal is to model dependence of Y on $X_1, \ldots, X_m$. There could exist several reasons we would like to do this. The first is a *description*. We want a model to describe the dependent variable on the predictors so that we could better understand the proces that produces Y. On the second place we are also interested in *inference*. We want to asses the relative contribution of each of the predictors in explaining Y. Finally, we could be interested in *prediction*. We wish to predict Y for some set of values obtained from $X_1, \ldots, X_m$.

For all these purposes, the standard tool for the applied statistician is the *multiple linear regression* model:

$$Y = \alpha + X_1\beta_1 + \cdots + X_m\beta_m + \epsilon, \tag{A.1}$$

where $\mathsf{E}(\epsilon) = 0$, $\mathrm{var}(\epsilon) = \sigma^2$ and $\alpha, \beta_1, \ldots,$ $beta_m$ are parameters whose values are unknown and have to be estimated from the data. Fitting linear regression models is performed by employing the standard *least squares* optimization procedure. In general, if we denote the conditional expectation of Y by μ, then the systematic part of the model can be expressed

$$\mathsf{E}(Y|\boldsymbol{X}) = \mu = \alpha + \sum_{j=1}^{m} X_j\beta_j, \tag{A.2}$$

Further specification of the model involves the stronger assumption for the random part of the model. Namely, that errors $\epsilon_i, i = 1, \ldots, n$, follow the Normal distribution with mean zero and constant variance σ^2. The least squares optimization procedure

$$\min_{\alpha,\boldsymbol{\beta}} \left(Y - \alpha + \sum_{j=1}^{m} X_j\beta_j\right)^2, \tag{A.3}$$

yields estimates $\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_m$ of $\alpha, \beta_1, \ldots, \beta_m$ which minimizes A.3. As outlined above, this model makes a strong assumption about dependence structure of μ on $X_1, \ldots, X_m$, that the dependence is linear in each of the predictors. In case, this assumption holds, the linear regression model is very useful and convenient. It provides a simple description of the data, summarizes the additive contribution of each predictor with a single coefficient. Thus it is easy to interpret and finally, provides a simple method for predicting new observations.

Because of the strong assumptions concerning the standard multivariate regression model described above, it is applicable only in situations when they are satisfied. The *Generalized linear regression model* is a generalization of the usual

linear regression model, so it is important to outline the limitations of the standard linear model and why we would like to generalize it. In practical applications it is quite common that the relationship between the response and the predictor variables is not linear. The response variables could be bounded, such as categorical response variables, or the variance is not constant – it could be expressed as the function of the means. Thus, in these cases, assumptions concerning the standard model does not hold.

General linear models are a generalization of linear regression models. Specifically, the predictor effects are assumed to be linear in the parameters, but the distribution of the response, as well as the *link* between the predictors and this distribution, can be quite general. A general linear model also consist of a *random component*, a *systematic component* and an additional *link function*, linking the two components. The response variable Y represents the random component of the model, it assumes to have exponential family density

$$f(y;\theta;\phi) = \exp\Big\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\Big\}, \tag{A.4}$$

wwhere $b(\cdot)$ is a smoothly differentiable function to the second order, $a(\phi)$ and $c(y,\phi)$ are functions such that $a(\phi) > 0$ and $c(y,\phi)$ does not depend on θ. Further note that ϕ is called the dispersion parameter. The parameter θ depends on values $\boldsymbol{x}^\top = (x_1,\ldots,x_m)$ of explanatory variables and on the vector of coefficients $\boldsymbol{\beta}$ trough the linear predictor $\eta = \alpha + \boldsymbol{x}^\top\boldsymbol{\beta}$. The linear predictor η represents the systematic component. Further there is a monotonic differentiable link function g such that $\eta = g(\mu)$, i.e.

$$g(\mu) = \alpha + \sum_{j=1}^{m} x_j\beta_j, \tag{A.5}$$

where $\mu = E(Y|\boldsymbol{X} = \boldsymbol{x})$. Note that the mean μ is related to the θ by $\mu = b'(\theta)$ and that a link function for which $g(\mu) = \theta$ is called the *canonical link*. Note that formally we can write

$$g(\mu) = \alpha + \sum_{j=1}^{m} X_j\beta_j. \tag{A.6}$$

Many useful models fall into this class, including the linear logistic regression model for binary data, that we employ and discuss in the section dedicated to multivariate analysis, where we further expand the theoretical background.

A further extension of the class of generalized linear models, is called *Generalized additive models*. They extend generalized linear models by replacing the linear form $\alpha + \sum_{j=1}^{m} X_j\beta_j$ with an additive form $\alpha + \sum_{j=1}^{m} f_j(X_j)$, where $f_j, j = 1,\ldots,m$ are arbitrary univariate functions, one for each predictor. As outlined above, linear models have an important feature that made them so popular for statistical inference. The are additive in predictor effects. Generalized additive models retain this important feature. Thus, they are also additive in predictor effects, however this time on the transformed scale given by the link function. Specifically, we assume that response Y has a distribution of the form A.4, with the conditional

expectation $\mu = \mathsf{E}(Y|X_1, \ldots, X_m)$ linked to the predictors as

$$g(\mu) = \alpha + \sum_{j=1}^{m} f_j(X_j). \tag{A.7}$$

2. Logistic models

The logistic regression model assumes that we have a binary response variable Y having alternative (Bernoulli) distribution $\mathrm{Alt}(\pi)$. Thus variable Y has two possible outcomes $Y = 1$ indicating that the obligor is a defaulter and $Y = 0$ indicating that he is a non-defaulter. The mean μ in this situation is equal to $\mu = E(Y|\boldsymbol{X} = \boldsymbol{x}) = P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \pi$. We denote this probability as $\pi(\boldsymbol{x})$ reflecting its dependence on values $\boldsymbol{x}^\top = (x_1, \ldots, x_m)$ of predictors.

2.1. Additive models

The *logistic additive model* assumes that the relation between p and the predictors has the form

$$\log\left\{\frac{\pi(\boldsymbol{X})}{1 - \pi(\boldsymbol{X})}\right\} = \alpha + \sum_{j=1}^{m} f_j(X_j) \tag{A.8}$$

where the link function on the left hand side is called the *logit* and where $\boldsymbol{X} = (X_1, \ldots, X_m)$. The additive predictor on the right hand side is determined by additive constant α and by arbitrary univariate functions $f_1, \ldots, f_m$ translating predictors $X_1, \ldots, X_m$ into a linear world.

By A.8 we have

$$\pi(\boldsymbol{X}) = \frac{\exp\left\{\alpha + \sum_{j=1}^{m} f_j(X_j)\right\}}{1 + \exp\left\{\alpha + \sum_{j=1}^{m} f_j(X_j)\right\}} \tag{A.9}$$

2.2. Linear model

The *logistic linear model* assumes that the relation between p and the predictors has the form

$$\log\left(\frac{\pi(\boldsymbol{X})}{1 - \pi(\boldsymbol{X})}\right) = \alpha + \sum_{j=1}^{m} X_j \beta_j. \tag{A.10}$$

where the link fuction on the left hand side is called the *logit* and where $\boldsymbol{X} = (X_1, \ldots, X_m)$.

The relationship between the above presented models is straightforward. The logistic additive model is a generalization of the logistic regression model as the additive predictor is replaced by a linear one, i.e. $f_j(X_j) = \beta_j X_j$.

Bibliography

[1] Agresti, A. (2002): *Categorical data analysis.* John Wiley and Sons. Inc., Hoboken, New Jersey.

[2] Anděl, J. (1985): *Matematická statistika.* SNTL/ALFA. Praha.

[3] Antoch, J. and Vorlíčková, D. (1992): *Vybrané metody statistické analýzy dat.* Academia, Praha.

[4] Bamber, D. (1975): The area above the ordinal dominance graph and the area below the receiver operating characteristics. *Journal of Mathematical Psychology* **12**, 387–415.

[5] Basel Committee on Banking Supervision (2004): *Internal convergence of capital measurement and capital standards: A rewised framework.* Bank for International Settlements, Basel.

[6] Basel Committee on Banking Supervision (1999): *A New Capital Adequacy Framework.* Bank for International Settlements, Basel.

[7] DeLong, E. R. and DeLong, D. M. (1988): Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845.

[8] Commission of the European communities (2000): *Credit Risk Directive (2000/12/EC)*, Basel.

[9] Fawcett, T. (2003): *ROC Graphs: Notes and Practical Considerations for Data Mining Researchers.* Intelligent Enterprise Technologies Laboratory, HP Laboratories Palo Alto.

[10] Green, P. J. and Silverman, B.W. (1995): *Nonparametric regression and generalized linear models.* Chapman and Hall, Florida.

[11] Hand, D. J. and Henley, W. E. (1997): Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society* **A 160**, 523–541.

[12] Hanley, J. A. and McNeil, B. J. (1982): The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36.

[13] Härdle, W. (1994): *Applied nonparametric regression: ebook.* Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät, Berlin.

[14] Hastie, T.J. and Tibshirani, R.J. (1997): *Generalized additive models.* Chapman and Hall, London.

[15] Hayden, E. (2002): *Modeling an accounting-based rating system for austrian firms: dissertation thesis.* Universität Wien, Fakultät für Wirtschaftswissenschaften und Informatik, Wien.

[16] Hoeffding, W. (1948): A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics* **19**, 293–325.

[17] McCullagh, P. and Nelder, J.A. (1989): *Generalized linear models (2nd ed.).* Chapman and Hall, London.

[18] Nelsen, R. B. (1998): *An Introduction to Copulas.* Springer.

[19] Slabý, A. and Magyar, T. (2004): *A Technical Guide to Rating Model Development.* ČSOB a.s., Praha.

[20] Sobehart, J. R., Keenan, S. C., and Stein, R. M. (2000): *Benchmarking Quantitative Default Risk Models.* Moody's KMV, New York.

[21] Stein,R. M. (2002): *Benchmarking default prediction models.* Moody's KMV, New York.

[22] Tasche, D., Hayden, E., and Engelman, B. (2002): *Measuring the discriminative power of rating systems.* Deutsche Bundesbank.

[23] Zvára K. (1989): *Regresní analýza.* Academia, Praha.