STATISTICAL ASPECTS OF SETTING UP A CREDIT RATING SYSTEM

Beatriz Clavero Rasero

Beim Fachbereich Mathematik
der Technischen Universität Kaiserslautern
zur Verleihung des akademischen Grades
Doktor der Naturwissenschaften
(Doctor rerum naturalium, Dr. rer. nat.)
genehmigte Dissertation

1. Gutachter: Prof. Dr. Ralf Korn
2. Gutachter: Priv. Doz. Dr. Marlene Müller

Vollzug der Promotion: 19.09.2006

First of all, I would like to say that I feel very fortunate in having been accompanied by such great people during my thesis. I would like to thank Prof. Dr. Ralf Korn for the interesting subject. He, together with Dr. Marlene Müller and Dr. Gerald Kroisandt were always guiding me with useful advices and dedication. I am also very thankful to Stefan Lorenz and the rest of my colleagues at the department of Financial Mathematics in Fraunhofer ITWM for their support. Special thanks also to Eva Barrena and Jan Hauth for their precious help. And at last but not least, my family and friends, who took care of me.

Contents

1 Introduction

The new international capital standard for credit institutions (“Basel II”) allows banks to use internal rating systems in order to determine the risk weights that are relevant for the calculation of capital charge. Therefore, it is necessary to develop a system that enfolds the main practices and methods existing in the context of credit rating. The aim of this thesis is to give a suggestion of setting up a credit rating system, where the main techniques used in practice are analyzed, presenting some alternatives and considering the problems that can arise from a statistical point of view. Finally, we will set up some guidelines on how to accomplish the challenge of credit scoring.

1.1 Background

The credit institutions offer their customers a variety of financial services and bring together investors and credit receivers. As a result, financial institutions face a multitude of risks such as credit, market and operational risks. These risks have to be covered with sufficient quantity of own capital.

The current regulations for credit risks are the result of a recommendation of the Committee on Banking Supervision in Basel, (“Basel II”), which started with a discussion 1988 that has been successfully concluded by the compromise formula of July 10, 2002. The amendment of the Basle commission is supposed to substitute the current flat rate equity capital securities of 8% of the standard risk-weighted credit positions (“Basel I”) at the end of 2006, including equity capital securities that depend on the credit risk.

The judgement of the quality of a credit with respect to the probability of default is called **credit rating**. A method based on a multi-dimensional criterion seems to be natural, due to the numerous effects that can influence this rating. However, owing to governmental rules, the tendency is that typically one-dimensional criteria will be required in the future as a measure for the credit worthiness or for the quality of a credit.

1.2 Approach

The problem as described above can be resolved via transformation of a multi-dimensional data set into a one-dimensional one while keeping some monotonicity properties and also

keeping the loss of information (due to the loss of dimensionality) at a minimum level. The following scheme will help us to understand better the process of credit rating:

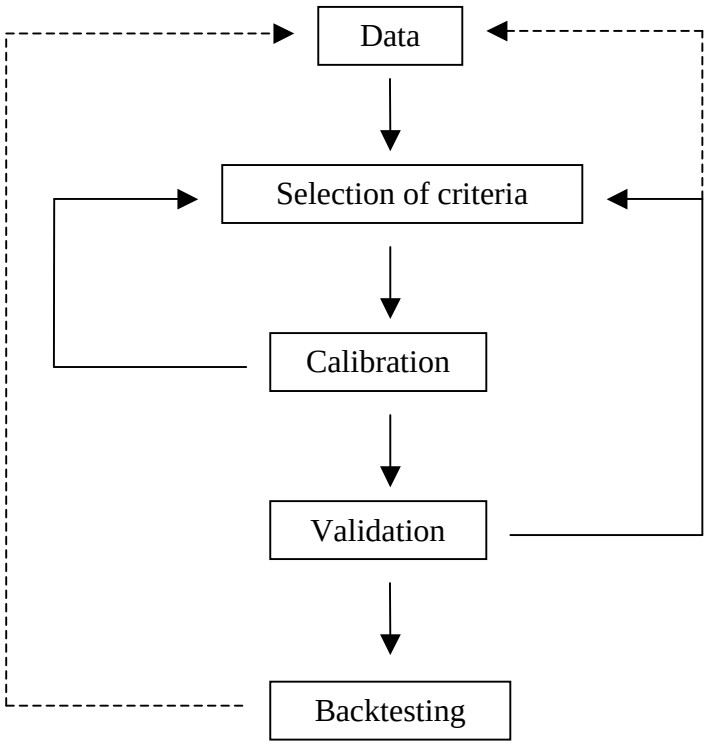


Figure 1.1: Concept of a credit rating system

The steps that are required for a method to evaluate credits are very briefly described as follows:

- Selection of rating criteria.
- Conception of ratings: identification and choice of influence factors with respect to their ability to distinguish between default and non-default.
- Estimation of scores: determination of the optimal weighting of the relevant influent factors with the help of econometric models (logit, probit, and extensions).
- Allocation of the scores in rating classes.

- Estimation of the default probabilities: binary choice models, panel models for the estimation of data covering several years; neural networks, nonparametric and semiparametric methods for the estimation of non-monotonous effects.

- Evaluation and comparison of scores and ratings with respect to their discriminatory power.

- Evaluation of a rating system: validation and backtesting using historical data.

These steps are developed along this thesis, which is organized in the following sections:

Section 2 is devoted to the study of the discriminatory power of credit ratings. There, we will study the different techniques that are used in practice, like the overlapping area, accuracy ratio, Lorenz curve or Gini index. In addition, the criterion for reduction in impurity is presented as another option to assess the discriminatory power of a score. We found out that the entropy-based criterion is also a valid discriminatory power measure, since hypothesis can be tested—contrary to the misconception in the Basel Committee on Banking Supervision (2005), that there are no applicable tests for the entropy-based measures. Further, we will review other measures, e.g. the misclassification rate, questioning their suitability for credit rating. To complete, we do a comparison of the most appropriate measures, i.e. overlapping area, accuracy ratio and the entropy-based criterion, evaluating their pros and cons in diverse situations.

In **Section 3** we offer an overview of the different methods for estimating the default probability and some aspects related, like the generation of rating classes. The well-known logit and probit models will be discussed in relation to other estimation methods, e.g. nonparametric and semiparametric, neural networks, panel data models and CART.

Section 4 introduces some tests for the validation and backtesting of PDs. As there is no single method that suits for alls situations in practice, a combination of the different techniques should be favoured. The binomial test and the chi-square test can be applied under the assumption of independence of default events, although default events are in fact correlated. The one factor threshold model of Basel II and the derived tests observe this correlation. In order to determine the adequacy of a forecasted default probability for time series, we may use the normal test or the extended traffic lights approach.

This work concludes with a summary in **Section 5,** where we recommend a standard procedure for credit scoring.

Finally, in the **Appendix** it can be found a summary of the literature about rating criteria.

2 Discriminatory power of credit ratings

2.1 Introduction

Suppose we deal with the following classification problem: we consider random variables $X_1,\ldots,X_p$ and $Y \in \{0,1\}$ as a group indicator. A score $S = S(X_1,...,X_p) \in \mathbb{R}$, used to rate applicants for a loan, is an aggregation of the variables $X_1,\ldots,X_p$ into a single number. Each individual variable can also be regarded as a score, although here we will only refer to the relation between the random variables S and Y.

In the following we will agree that a reasonable score function should assign higher score values to credit applicants who have higher probabilities of default. Formally, it means that the distribution of defaults dominates stochastically over the non-defaults. A basic feature of a credit score function is consequently the efficiency to separate the two groups of observations according to $Y = 1$ (default) and $Y = 0$ (non-default). A measure for the discriminatory power can thus be used as a performance measure for a credit score.

A measure of discriminatory power is not required to quantify the goodness of fit of the estimated probabilities of default (see section 3) to the real PDs. Suppose that we consider as our score the probability of default estimated by the model. Then, for any (strictly) monotone transformation of S, the discriminatory power would not change, although the output of such a transformation has nothing to do with the original range of values.

We will describe here the measures that are being used in the credit rating practice. The overlapping area criterion and its associated Kolmogorov-Smirnov test are introduced in section 2.2. In section 2.3 we will depict the well-known accuracy ratio, which is related to the Wilcoxon-Mann-Whitney U test. An alternative discriminatory power measure is given by the standardized maximal distance (2.4.4) defined in section 2.4. Furthermore, we will see in section 2.4.5, that the entropy-based criterion for reduction in impurity (2.4.10) is linked to the test for homogeneity in 2×2 contingency tables.

On the other hand, there are also measures (section 2.5) that are not appropriate for assessing the discriminatory power of a score, i.e. misclassification rate (2.5.2), although they can be found in the literature or used in practice.

At the end of this section, we will compare the most suitable measures, analysing their behaviour under specific circumstances. An important feature of these measures is that they actually remain constant under (strictly) monotone transformation of the score.

As it is easier to see graphically how some measures separate the data, we will picture the kernel density estimates of non-defaults (blue) and defaults (red) with a Gaussian smoothing kernel for *score1* (2.1.2), *score2* (2.1.3) in the following example, and other simulated examples.

Example 2.1

We built up two scores from a sample of private loans we got from Fahrmeir & Hamerle (1984), having on account some variables like personal characteristics, credit characteristics and credit history. The sample size is 1000; 300 of them are defaults. For the calibration and validation of the model, we chose at random two samples with 80% and 20% of the data and default rates 0.3075 and 0.27, respectively. In practice, the proportion of defaults is normally lower. The variables used are listed in the following table:

Variable	Specification 1	Specification 2
Bank account	×	
Duration of the credit	×	
Payment previous credits	×	
Amount of credit	×	
Savings	×	×
Time current job	×	×
Monthly payment % income	×	
Marital status, sex	×	×
Properties	×	
Age	×	
Previous credits	×	

Table 2.1: Variables selected for each specification

The scores were determined by a logit model (see section 3.2.2), such that, given the vector of realizations of the explanatory variables, i.e. $x = (1, x_1, ..., x_{p-1})^T$, the estimated probability of default is given:

$$P(Y=1 \mid X=x)=\frac{1}{1+\exp\left(-\beta^{\mathsf{T}}x\right)}=\frac{\exp\left(\beta^{\mathsf{T}}x\right)}{1+\exp\left(\beta^{\mathsf{T}}x\right)}, \tag{2.1.1}$$

β being the vector of regression parameters. For every specification, the scores $S=\beta^{\mathsf{T}}X$ were calculated:

$$\begin{aligned} \textit{score1} = {} & 1.389\,{*}{*} - 0.574\,{*}{*}{*}{*}\,\text{account} + 0.031\,{*}{*}{*}{*}\,\text{duration} - 0.537\,{*}{*}{*}{*}\,\text{pay} \\ & + 5.024e-05\,\text{amount} - 0.213\,{*}{*}{*}{*}\,\text{savings} - 0.160\,{*}{*}\,\text{time} - 0.256\,{*}{*}{*}\,\text{month} \\ & - 0.241\,{*}\,\text{status} + 0.114\,\text{properties} - 0.009\,\text{age} + 0.325\,{*}\,\text{prev\_credits} \end{aligned} \tag{2.1.2}$$

$$\textit{score2} = 0.826\,{*}{*} - 0.258\,{*}{*}{*}{*}\,\text{savings} - 0.177\,{*}{*}{*}\,\text{time} - 0.208\,{*}\,\text{status} \tag{2.1.3}$$

*, **, ***, **** denote significant coefficients at the 10%, 5%, 1%, 0.1% level, respectively.

2.2 Overlapping area

In this section we will show the construction of the overlapping area criterion—which goes back to the work of Kraft, Kroisandt & Müller (2004)—and its associated Kolmogorov-Smirnov test. The distance between defaults and non-defaults scores can be simply assessed by calculating the overlapping area of their respective densities. Let us denote by F_0, F_1 the cumulative distribution functions of $S \mid Y=0$ and $S \mid Y=1$. We will first observe the case of two normal distributions having one point of intersection; the conditional densities f_0 and f_1 are easy to visualize and to compute:

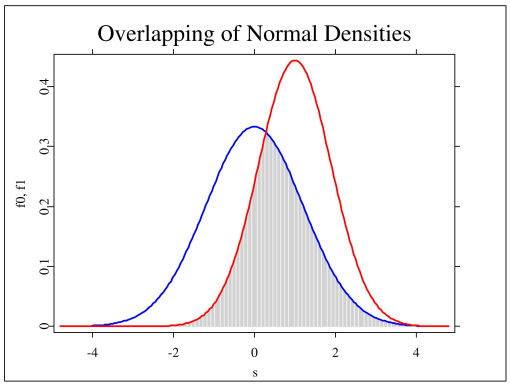


Figure 2.1: Overlapping area of two normal densities

The horizontal coordinate of this intersection is denoted by $s \in \mathbb{R}$, which is the separation threshold such that all score values $S > s$ are predicted as default.

Let us assume a normal distribution means that both densities f_0 and f_1 are determined by their expectations μ_0, μ_1 and standard deviations σ_0, σ_1. Without loss of generality we can suppose from now on that $\mu_1 > \mu_0$. Then the region of overlapping O for both densities can be calculated as

$$O = F_1(s) + 1 - F_0(s). \tag{2.2.1}$$

There is exactly one point of intersection if both standard deviations are identical ($\sigma_0 = \sigma_1$) for the normal case, which is given by

$$s = \frac{\mu_0 + \mu_1}{2}.$$

On the other hand, if they are different ($\sigma_0 \neq \sigma_1$), then there may be one or two points of intersection (as in quadratic discriminate analysis) and the horizontal coordinates are given by $f_0(s) = f_1(s)$, i.e. as solutions of the quadratic equation

$$s^2(\sigma_1^2 - \sigma_0^2) + 2s(\mu_1\sigma_0^2 - \mu_0\sigma_1^2) + \mu_0^2\sigma_1^2 - \mu_1^2\sigma_0^2 + 2\sigma_1^2\sigma_0^2(\log(\sigma_0) - \sigma_0^2\log(\sigma_1)) = 0.$$

If we do not make distributional assumption for S, then the definition of O can be easily generalized to the nonparametric case

$$O = \int \min\{f_0(s), f_1(s)\} ds,$$

which allows any number of intersection points.

Assume only one optimal point of intersection. For a positive monotone relationship between the score S and the default probability, the overlapping area is defined by

$$O_{pos} = \min_s \{F_1(s) + 1 - F_0(s)\}. \tag{2.2.2}$$

Alternatively, for a negative monotone influence of the score values S on Y, we set

$$O_{neg} = \min_s \{F_0(s) + 1 - F_1(s)\}. \tag{2.2.3}$$

For a monotone relationship it obviously holds

$$O_{mon} = \min\{O_{pos}, O_{neg}\},$$

It is clear that for perfectly separated distributions the region of overlapping O is zero. If both distributions are identical, then $O=1$. A measure of the discriminatory power is then given by

$$T = 1 - O_{mon} = \max_s \left| F_0(s) - F_1(s) \right|. \tag{2.2.4}$$

The discriminatory power indicator T takes on values in the interval [0, 1], where $T=1$ stands for perfect separation and $T=0$ means no separation. The positive- and negative-monotone versions of T are settled

$$T_{pos} = 1 - O_{pos} = \max_s \{F_0(s) - F_1(s)\} \tag{2.2.5}$$

$$T_{neg} = 1 - O_{neg} = \max_s \{F_1(s) - F_0(s)\} \tag{2.2.6}$$

In the monotone case, T can be estimated by nonparametric estimates of the cumulative distribution functions F_0, F_1, i.e. the empirical distribution functions. Under the assumption of normality, O (and hence T) can be computed by their empirical moments $\widehat{\mu_0}$, $\widehat{\mu_1}$, $\widehat{\sigma_0}$, and $\widehat{\sigma_1}$. Under more general assumptions on the distributions, O and T can be calculated for example by nonparametric estimates of the densities, like histograms or kernel density estimators. For more information on this topic, see Härdle (1991) or Silverman (1986).

2.2.1 Kolmogorov-Smirnov test

We remark that the measure T is related to the Kolmogorov-Smirnov test statistics. If we consider the hypotheses:

	H_0	H_1	*Test statistic*	*Reject condition*
(1)	$F_1(x) = F_0(x)$	$F_0(x) > F_1(x)$	$\widehat{T}_{pos} = \max_s \left\{ \widehat{F}_0(s) - \widehat{F}_1(s) \right\}$	$\widehat{T}_{pos} > \Delta_{n_1,n_0;1-\alpha}$
(2)	$F_1(x) = F_0(x)$	$F_0(x) < F_1(x)$	$\widehat{T}_{neg} = \max_s \left\{ \widehat{F}_1(s) - \widehat{F}_0(s) \right\}$	$\widehat{T}_{neg} > \Delta_{n_1,n_0;1-\alpha}$
(3)	$F_1(x) = F_0(x)$	$F_1(x) \neq F_0(x)$	$\widehat{T} = \max_s \left\| \widehat{F}_0(s) - \widehat{F}_1(s) \right\|$	$\widehat{T} > \Delta_{n_1,n_0;1-\alpha/2}$

then we can use the test statistics in (1) and (2) to check for the stochastic dominance of F_1 over F_0 and vice versa. The critical values are given in Miller (1956), as follows:

$$\Delta_{n_1,n_0;1-\alpha} = \Delta_{q;1-\alpha} \quad \text{with} \quad q = \left\lfloor \frac{n_0 \cdot n_1}{n_0 + n_1} \right\rfloor,$$

where n_0 and n_1 are the number of non-defaults and defaults. For n_1 decreasing the critical values are increasing, but it does not imply that the value of the test statistic increases. This means that, given a test statistic value, the null hypothesis will be more difficult to reject if the default rates are lower.

Example 2.2

Given the Fahrmeir et al. (1984) dataset for the scores (2.1.2) and (2.1.3) and a confidence level $1-\alpha$, for the validation sample we obtain the following results:

	$\widehat{T}_{pos}$	$1-\alpha$	$\Delta_{39;1-\alpha}$	s
score1	0.555	0.995	0.255	-1.253
score2	0.254	0.95	0.191	-1.801

Table 2.2: Kolmogorov-Smirnov Test Statistic values

The two scores are significant, since $\widehat{T}_{pos}$ is larger than the critical value. Therefore we can reject the null hypothesis in favour of the alternative and conclude that F_1 dominates stochastically over F_0 for both of them and *score1* has more discriminatory power than *score2*.

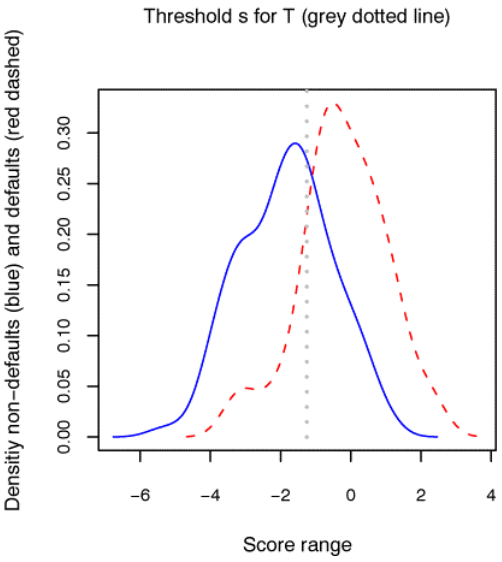


Figure 2.2: *score1*, $\widehat{T} = 0.555$

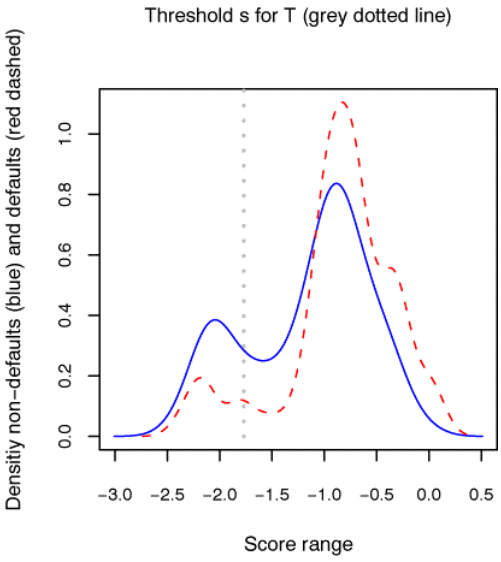


Figure 2.3: *score2*, $\widehat{T} = 0.254$

In order to illustrate how the densities should look like, we pictured the density estimates of defaults and non-defaults and the threshold s for *score1* (Figure

2.2) and *score2* (Figure 2.3), respectively. We must remark that these pictures do not correspond to the integrals of the estimated distribution functions, but they are the kernel density estimates of defaults and non-defaults.

2.3 Accuracy ratio

Another commonly used measure for the performance of a score is the accuracy ratio AR, based on the Lorenz curve and its Gini coefficient. The Lorenz curve (Figure 2.4), also known as selection curve, plots the distribution of the score S against the defaults distribution $S \mid Y = 1$, and thus we can compare different credit scores graphically. For the cumulated probabilities, the percentages of applicants are arranged from "bad" to "good" scores (highest to lowest). Variants of the Lorenz curve are the receiver operating characteristic (ROC) curve (see Hand & Henley, 1997) and the performance curve (see Gourieroux & Jasiak, 2001). A generalization of the accuracy ratio may be interpreted by the Somers' D, which is also a conditional version of the Kendall's τ, both of them rank order statistics; for more information, see Basel Committee on Banking Supervision (2005) and Lienert (1973).

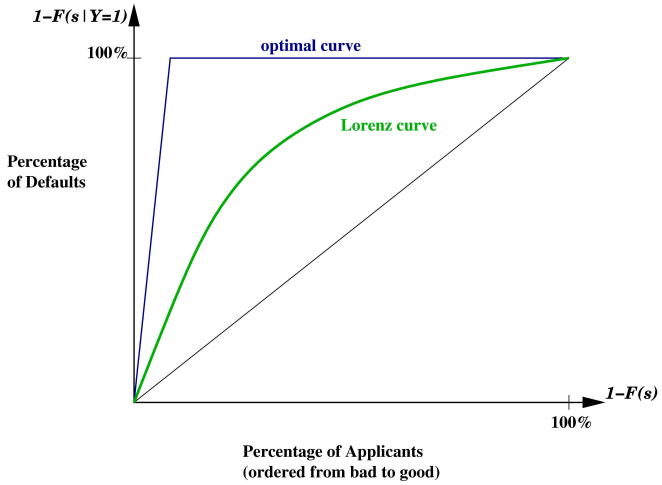


Figure 2.4: Lorenz curve for credit scoring

2.3.1 Lorenz curve

In order to operate with cumulative distribution functions we denote the negative score by

$$R = -S\,.$$

The Lorenz curve of the score S is then given by the coordinates

$$\{L_1(r), L_2(r)\} = \{P(R < r), P(R < r \mid Y = 1)\}, \quad r \in (-\infty, \infty).$$

As $P(R < r) = 1 - F(s)$, this is equivalent to

$$\mathcal{L}(s) = \{L_1(s), L_2(s)\} = \{1 - F(s), 1 - F_1(s)\}, \quad s \in (-\infty, \infty).$$

The Lorenz curve can be estimated by means of the empirical cumulative distribution functions.

The optimal Lorenz curve corresponds to a score that perfectly separates defaults and non-defaults. It reaches the vertical 100% at a horizontal percentage of $P(Y = 1)$, the probability of default, and is given by

$$\mathcal{L}_{opt}(s) = \{1 - F(s), g(1 - F(s))\}, \quad s \in (-\infty, \infty), \text{ being}$$

$$g(x) = \begin{cases} \dfrac{x}{P(Y=1)} & 0 < x \le P(Y = 1) \\ 1 & P(Y = 1) < x \le 1 \end{cases}$$

A Lorenz curve identical to the diagonal corresponds to a score that orders the credit applicants totally randomly. Thus, Lorenz curves can also be used to compare different score distributions: better scores are closer to the optimal Lorenz curve and worse scores are closer to the diagonal.

2.3.2 Gini coefficient

Now we need a quantitative measure for the performance of a score, which is based on the area between Lorenz curve and the diagonal. The Gini coefficient G denotes twice this area:

$$G = 2\int_{+\infty}^{-\infty} \{1 - F_1(s)\}\, d\{1 - F(s)\} - 1 = 1 - 2\int_{-\infty}^{+\infty} F_1(s)\, dF(s). \qquad (2.3.1)$$

This integral can be estimated by numeric integration of $\widehat{F}_1$ over the range of $\widehat{F}$.

Proposition 2.3

For the optimal Lorenz curve, we have that the optimal Gini coefficient is given by:

$$G_{opt} = P(Y = 0) = 1 - P(Y = 1).$$

Proof:

The optimal Gini coefficient is twice the area of the triangle between the optimal Lorenz curve and the diagonal. This is the same as to calculate the area of a parallelogram = base * height. In this case, the base is $P(Y=0)$ and the height $1 = P(Y=0) + P(Y=1)$.

□

2.3.3 Accuracy ratio

To compare different scores, we use their accuracy ratios, which are given by the relation of the Gini coefficient of each score to the Gini coefficient of the optimal Lorenz curve. The accuracy ratio is defined as

$$AR = \frac{G}{G_{opt}} = \frac{G}{1 - P(Y=1)},$$

and therefore empirically given by the estimates of both Gini coefficients. The value of AR lies between 0 and 1 if the Lorenz curve is really concave, i.e. if there is a positive-monotone relationship between S and Y (higher score values correspond to higher default probability). Negative values are possible if the relation is negative monotone; in that case we would change the sign of the score, in order to obtain a positive value of the accuracy ratio.

Example 2.4

In order to estimate the accuracy ratios for the scores (2.1.2) and (2.1.3) in the validation sample, we first pictured their Lorenz curves (see Figure 2.5 and 2.6). Then we calculated their respective Gini coefficients, and AR:

	$\widehat{G}$	$\widehat{P}(Y=0)$	$\widehat{AR}$
score1	0.463	0.73	0.635
score2	0.223	0.73	0.306

Table 2.3: Gini coefficients and accuracy ratios

We can conclude from these results, that the first score has more discriminatory power than the second one and that there is a positive monotone relationship between the scores and Y.

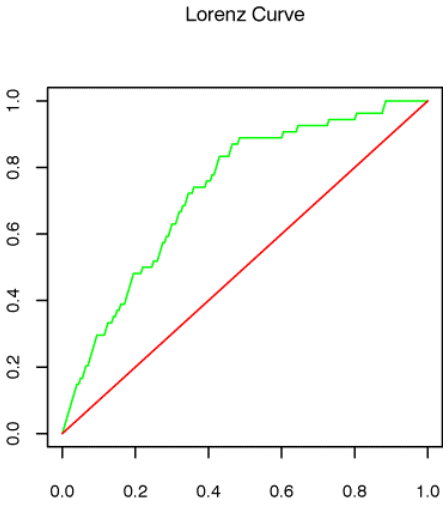


Figure 2.5: *score1*

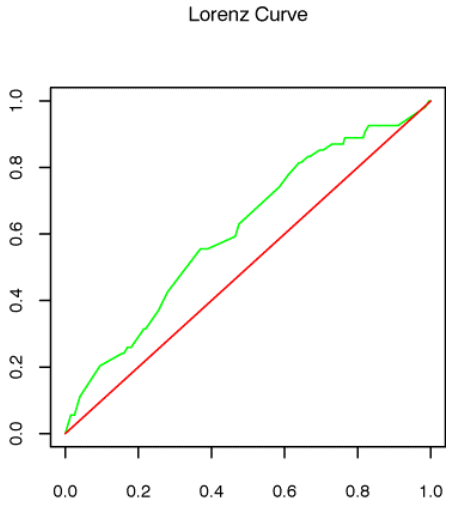


Figure 2.6: *score2*

2.3.4 ROC curve

A graphical alternative to the Lorenz curve is the receiver operating characteristic (ROC) curve. The curve is set by the coordinates:

$$\mathcal{R}(s) = \{1 - F_0(s), 1 - F_1(s)\}. \tag{2.3.2}$$

Contrary to the Lorenz curve, the ROC compares the score distribution of the non-defaults versus that of the defaults. The resulting graph resembles that of $\mathcal{L}(s)$ as the number of defaults is typically small and therefore we have $F \approx F_0$.

The optimal ROC curve corresponds to a score that exactly separates defaults and non-defaults and it is determined by the points (0, 0), (0, 1) and (1, 1).

2.3.5 AUC (area under curve)

In order to quantify the deviation between F_0 and F_1, we use the so-called area under curve (AUC):

$$AUC = \int_{+\infty}^{-\infty} \{1 - F_1(s)\}\, d\{1 - F_0(s)\} = 1 - \int_{-\infty}^{+\infty} F_1(s)\, dF_0(s), \tag{2.3.3}$$

taking values between 0, for the shortest deviation, and 1 for the largest deviation. It is important to note that:

Proposition 2.5

The AUC and the accuracy ratio are linearly related as follows:

$$AR = 2AUC - 1. \tag{2.3.4}$$

Proof:

Recall the definition of the Gini coefficient G (2.3.1). Thus,

$$\begin{aligned}\frac{1-G}{2} &= \int_{-\infty}^{+\infty} F_1(s)\,dF(s) = \int_{-\infty}^{+\infty} F_1(s)\,d\left\{P(Y=0)F_0(s) + P(Y=1)F_1(s)\right\} \\ &= P(Y=0)\int_{-\infty}^{+\infty} F_1(s)\,dF_0(s) + P(Y=1)\int_{-\infty}^{+\infty} F_1(s)dF_1(s) \\ &= P(Y=0)\cdot(1-AUC) + P(Y=1)\cdot\frac{1}{2} = \frac{P(Y=0)}{2} - P(Y=0)\cdot AUC + \frac{1}{2}.\end{aligned}$$

We obtain $G = 2\cdot AUC\cdot P(Y=0) - P(Y=0)$, plugging this into $AR = G / P(Y=0)$ completes the proof.

□

Therefore, using AUC and AR to rank a set of different score functions will lead to identical conclusions. With the help of relationship (2.3.4) we can demonstrate the following proposition.

Proposition 2.6

Let F_0, F_1 be cumulative distribution functions having the same expectation $\mu \in \mathbb{R}$. Suppose that they are point-symmetric about $\left(\mu, F_j(\mu)\right)$ for $j = 0,1$. Then we obtain:

$$AR = 0.$$

Proof:

The condition on symmetry is equivalent to: $F_j(\mu + s) + F_j(\mu - s) = 2F_j(\mu)$, $\forall s \in \mathbb{R}$ and $j = 0,1$, being $F_j(\mu) = 1/2$.

Let us denote the random variable $\widetilde{S} = S - \mu$ and $\widetilde{F}_j(s) = P\left(\widetilde{S} \le s \mid Y = j\right)$. It is easy to see that $E\left(\widetilde{S}\right) = E\left(\widetilde{S} \mid Y = j\right) = 0$, $\widetilde{F}_j(s) = 1 - \widetilde{F}_j(-s)$ and thus $\widetilde{F}_j(0) = 1/2$ for $j = 0,1$.

We will calculate AUC for the transformed variable $\widetilde{S}$. By definition (2.3.3):

$AUC = 1 - \int_{-\infty}^{\infty} F_1(s)\,dF_0(s) = 1 - \int_{-\infty}^{\infty} \widetilde{F}_1(s)\,d\widetilde{F}_0(s)$, being

$$\int_{-\infty}^{\infty} \widetilde{F}_1(s)\, d\widetilde{F}_0(s) = \int_{-\infty}^{0} \left(1 - \widetilde{F}_1(-s)\right) d\widetilde{F}_0(s) + \int_{0}^{\infty} \widetilde{F}_1(s)\, d\widetilde{F}_0(s)$$

$$= \int_{-\infty}^{0} 1\, d\widetilde{F}_0(s) - \int_{-\infty}^{0} \widetilde{F}_1(-s)\, d\widetilde{F}_0(s) + \int_{0}^{\infty} \widetilde{F}_1(s)\, d\widetilde{F}_0(s)$$

$$= \widetilde{F}_0(0) - 0 = \frac{1}{2}, \text{ therefore}$$

$$AUC = 1 - \int_{-\infty}^{\infty} \widetilde{F}_1(s)\, d\widetilde{F}_0(s) = \frac{1}{2}.$$

By the relationship (2.3.4), we obtain $AR = 2AUC - 1 = 0$.

□

Remark 2.7

If for a score S we have the conditional density distribution functions f_0, f_1 (or probability mass functions, in case S is discrete), and they are even functions with respect to a common expectation μ, then $AR = 0$.

2.3.6 Wilcoxon-Mann-Withney U test

Some classical nonparametric tests to check whether two distributions are identical are the Wilcoxon rank sum test and its equivalent, the Mann-Whitney U test. In its simplest form, the U test is derived for continuous score distributions. Denote s_{jo} all observed scores of non-defaults and s_{i1} all observed scores of defaults. The U test statistic is given by

$$U = \#\left\{s_{i1} > s_{j0}\right\} \text{ over all } i, j.$$

For perfectly separated defaults and non-defaults, we obtain $U = n_0 \cdot n_1$. If S and Y are not related at all, then the event $s_{i0} > s_{j1}$ occurs with probability $1/2$, such that $U \approx 1/2 \cdot (n_0 \cdot n_1)$. Consequently, a rescaled version of the test statistics, $U/(n_0 \cdot n_1)$ is an estimate for

$$\widetilde{U} = P\left\{(S \mid Y = 1) > (S \mid Y = 0)\right\} = \int \left\{1 - F_1(s)\right\} d\left\{1 - F_0(s)\right\} = AUC,$$

and therefore using (2.3.4),

$$U = \left(\frac{\widehat{AR} + 1}{2}\right) \cdot n_0 \cdot n_1. \tag{2.3.5}$$

The relation between $\widetilde{U}$ and AUC will remain valid if the score distributions are not continuous. However, it could happen that for any score values we observed both defaults and non-defaults, i.e. tied observations. In that case a corrected version of the U statistic must be used (add $1/2$ if $s_{i0} = s_{j1}$) to estimate

$$P\{(S \mid Y = 0) > (S \mid Y = 1)\} + \frac{1}{2} P\{(S \mid Y = 0) = (S \mid Y = 1)\}.$$

It is demonstrated (see Lehmann, 1975, p. 365) that, under the hypothesis $F_1(x) = F_0(x)$, for large n_0, n_1, U is approximately normally distributed. We consider the hypotheses:

	H_0	H_1	*Test Statistic*	*Reject condition*
(1)	$F_1(x) = F_0(x)$	$F_0(x) > F_1(x)$	U	$U > k_{n_1,n_0;1-\alpha}$
(2)	$F_1(x) = F_0(x)$	$F_0(x) < F_1(x)$	U	$U < n_0 \cdot n_1 - k_{n_1,n_0;1-\alpha}$

being the critical value

$$k_{n_0,n_1,1-\alpha} = \frac{n_0 \cdot n_1}{2} + u_{1-\alpha} \cdot \sqrt{\frac{1}{12} \cdot n_0 \cdot n_1 \cdot (n_0 + n_1 + 1)}. \qquad (2.3.6)$$

The critical values and the test statistic decrease for n_1 decreasing. Hence, we cannot say that the null hypothesis will be more difficult to reject for lower default rates.

Example 2.8

Let us test now the hypothesis (1) for the validation sample, with $n_0 = 146$ and $n_1 = 54$. For a level of significance $\alpha = 0.005$, we get the critical value $k_{146,54,0.995} = 4879.562$. Being $U = 6446.999$ for *score1* and $U = 5148.499$ for *score2*, both larger than the critical value and highly significant. We can reject for the two of them the null hypothesis and conclude that F_1 dominates stochastically over F_0.

2.4 Impurity functions

We consider now the fact that the discriminatory power of a score S is closely related to the heterogeneity, or impurity, of the distribution of Y conditioned to this score. Actually, a good discriminating score will separate the defaults and non-defaults in preferably heterogeneous classes. The impurity of a set should be largest, when all events are equally likely and

smallest when only one event happens. In order to formulate the criterion for reduction in impurity, we will first define an impurity function. The best general references here are Fahrmeir, Hamerle & Tutz (1996) and Breiman, Friedman, Olshen & Stone (1984).

Definition 2.9

A function ϕ *defined on the simplex*

$$S_g = \left\{\pi = (\pi_1,\ldots,\pi_g)^t, \pi_i \geq 0, \sum\nolimits_i \pi_i = 1\right\} \text{ by } \phi : S_g \to \mathbb{R},$$

is called impurity function, if holds

(i) $\arg\min_{\pi \in S_g} \phi(\pi) \in \left\{(1,0,\ldots 0)^t, (0,1,\ldots 0)^t, \ldots, (0,0,\ldots 1)^t\right\}$,

(ii) $\phi(\pi)$ *is a symmetric function of* π*, i.e. it remains invariant with regard to any permutation of* $\pi_1,\ldots,\pi_g$.

(iii) $\phi(\pi)$ *is concave.*

Remark 2.10

Our definition is more restricting than the definition of impurity function given in Breiman et al. (1984), or Fahrmeir et al. (1996). It differs in the third condition, as they write:

(iii)' $\phi\left(\left(\frac{1}{g},\ldots,\frac{1}{g}\right)\right) = \max_{\pi \in S_g} \phi(\pi)$.

From the conditions on symmetry (ii) *and concavity* (iii) *follows* $(iii)'$*, but from the definition given by these authors, it does not follow concavity. We will see later that this condition* (iii) *is necessary for the proof of Proposition 2.12.*

Remark 2.11

We can express also the impurity function as $\phi(\pi) = \phi(\pi_1,\ldots,\pi_{g-1})$*, since* $\pi_g = 1 - \sum_{i=1}^{g-1} \pi_i$.

According to Definition 2.9, so as to Breiman et al. (1984) or Fahrmeir et al. (1996), some impurity functions are:

1. Misclassification rate: $\phi(\pi) = 1 - \max_i \pi_i$

2. Gini index: $\phi(\pi) = \sum_{i \neq j} \pi_i \pi_j$
3. Entropy: $\phi(\pi) = -\sum_i \pi_i \log \pi_i$

In the following we will study the case $g = 2$, for Y has only two possible outcomes: default, with probability $\pi_1 = P(Y = 1)$ and non-default, with probability $\pi_0 = 1 - \pi_1 = P(Y = 0)$. It suffices to write the impurity function in terms of only one of these probabilities, i.e. $\phi(\pi) = \phi(1 - \pi_1, \pi_1) = \phi(\pi_1)$. Figure 2.7 shows the shapes of this impurity functions for $g = 2$.

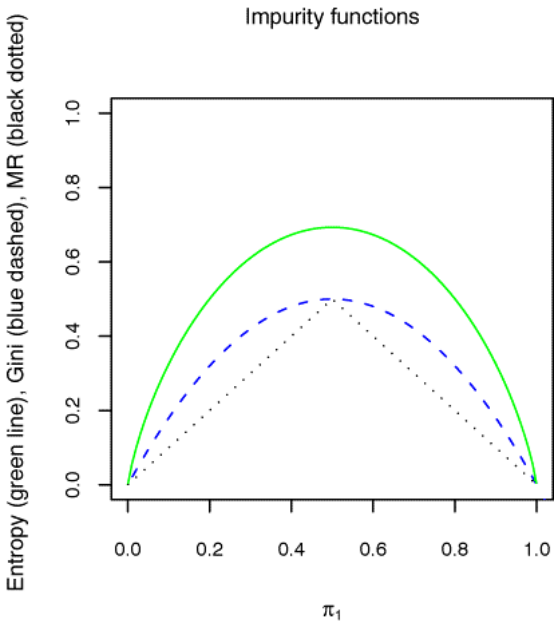


Figure 2.7: Entropy, Gini index and misclassification rate for $g = 2$.

We can observe how the impurity functions are concave, whereas entropy and Gini index are strictly concave. They reach their maximum at $\pi_1 = \pi_0 = 1 - \pi_1 = 1/2$, their minimum, at $\pi_1 = 1$ and $\pi_1 = 0$, and they are symmetric in π. They all are monotonically increasing for $\pi_1 \leq 1/2$, and decreasing for $\pi_1 \geq 1/2$.

The impurity of a subset $S_s = \{S \leq s\}$ of the score S for some $s \in \mathbb{R}$ can be defined as

$$\mathrm{i}(S_s) = \phi(1 - P(1 \mid S_s), P(1 \mid S_s)) = \phi(P(1 \mid S_s)). \tag{2.4.1}$$

$\mathrm{i}(S_s)$ is maximal, if $P(0 \mid S_s) = P(1 \mid S_s) = 1/2$ and it is minimal, i.e. $\mathrm{i}(S_s) = 0$, if $P(0 \mid S_s) = 1$ or $P(1 \mid S_s) = 1$.

The goal of splitting up the observations into subsets as heterogeneous as possible can be fastened by maximizing the following distance:

$$d_s(S \le s, S > s) := \Delta \mathrm{i}\left(S; S_s, \overline{S}_s\right) = \mathrm{i}(S) - P(S_s)\mathrm{i}(S_s) - P(\overline{S}_s)\mathrm{i}\left(\overline{S}_s\right), \qquad (2.4.2)$$

being $P(S_s) = P(S \le s)$ and $P(\overline{S}_s) = 1 - P(S_s) = P(S > s)$.

Proposition 2.12

For every partition of S at s,

$\Delta \mathrm{i}\left(S; S_s, \overline{S}_s\right) \ge 0$,

with equality if $\mathrm{i}(S_s) = \mathrm{i}\left(\overline{S}_s\right) = \mathrm{i}(S)$.

Proof:

By the concavity of ϕ,

$$\mathrm{i}(S_s)P(S_s) + \mathrm{i}\left(\overline{S}_s\right)P\left(\overline{S}_s\right) = \phi\left(P(1 \mid S_s)\right)P(S_s) + \phi\left(P\left(1 \mid \overline{S}_s\right)\right)P\left(\overline{S}_s\right)$$
$$\le \phi\left(P(1 \mid S_s)P(S_s) + P\left(1 \mid \overline{S}_s\right)P\left(\overline{S}_s\right)\right)$$

Now by the theorem of total probability, we have that:

$$P(1 \mid S_s)P(S_s) + P\left(1 \mid \overline{S}_s\right)P\left(\overline{S}_s\right) = P(1 \mid S) = \pi_1$$

Therefore,

$\mathrm{i}(S_s)P(S_s) + \mathrm{i}\left(\overline{S}_s\right)P\left(\overline{S}_s\right) \le \phi\left(P(1 \mid S)\right) = \mathrm{i}(S)$ and

$$\Delta \mathrm{i}\left(S; S_s, \overline{S}_s\right) = \mathrm{i}(S) - P(S \le s)\mathrm{i}(S_s) - \left(1 - P(S \le s)\right)\mathrm{i}\left(\overline{S}_s\right) \ge 0$$

With equality holding, if $\mathrm{i}(S_s) = \mathrm{i}\left(\overline{S}_s\right) = \mathrm{i}(S)$.

□

The discriminatory power of the score S is therefore given by the best partition, if we allow only one splitting point:

$$d = \max_s d_s(S \le s, S > s). \qquad (2.4.3)$$

Proposition 2.13

The maximal distance d is bounded:

$0 \le d \le d_{opt} = \mathrm{i}(S)$

Proof:

1. By (2.4.3) and Proposition 2.12, it follows immediately that $d \ge 0$. If the split offers no decrease in impurity, then we have $d = 0$.

2. In case of a perfect separation, then $\mathrm{i}(S_s) = \mathrm{i}(\overline{S}_s) = 0$, and the optimal $d_{opt} = \mathrm{i}(S)$.

□

In order to compare different scores, we use the standardized maximal distance:

$$D = \frac{d}{d_{opt}}, \tag{2.4.4}$$

which ranges from 0 to 1.

In the next section, we will set the standardized maximal distance for the misclassification rate (D_{mr}), and note its inadequacy for credit rating. Therefore, we will introduce another class of impurity functions (section 2.4.2), where the misclassification rate is not included. Further, we will determine the criterion for reduction in impurity with the Gini index (D_g) and entropy (D_e) as impurity functions. Moreover, we show in section 2.4.5 that the test for homogeneity in 2×2 contingency tables is related to D_e. At the end, we do a detailed comparison of D_g and D_e.

2.4.1 Misclassification rate

A relatively intuitive criterion is to choose that split which most reduces the misclassification rate. If, for a given S, we assign a posterior observation $\widehat{Y} = j$, such that maximizes $P(j \mid S)$, then the misclassification rate is given by minimizing this probability, as follows:

$$\mathrm{i}(S) = 1 - \max_{j=0,1} P(j \mid S) = \min_{j=0,1} P(j \mid S) = \min\big(P(0 \mid S), P(1 \mid S)\big) \tag{2.4.5}$$

The best split for this criterion is therefore given by substituting this formula in (2.4.3). Finally, we get the standardized D in (2.4.4):

$$D_{mr} = 1 - \min_{s}\left(\frac{P(S_s)\min\left(P(0 \mid S_s), P(1 \mid S_s)\right) + P(\overline{S}_s)\min\left(P(0 \mid \overline{S}_s), P(1 \mid \overline{S}_s)\right)}{\min\left(\pi_0, \pi_1\right)}\right) \qquad (2.4.6)$$

In this section we defined the misclassification rate as an impurity function of Y given S. But we also find in the literature the misclassification rate as a measure of discriminatory power when referred to the distribution of the score S given the default Y. We will pay attention to that definition in section 2.5.1, and we will show that, under some assumptions, both criteria are linearly related.

Still, in spite of its attractiveness, choosing the misclassification rate as splitting criterion has some serious defects—one of them is easy to see with the following example:

Example 2.14

If we have a score with equal priors, for example $n_1 = n_0 = 600$, then consider two possible splits (see Figure 2.8). For the first split, there are 400 observations misclassified, 200 of them are defaults and 200 non-defaults and we obtain $d_1\left(S_s, \overline{S}_s\right) = 1/6$. The second split misclassifies also 400 observations, all of them are non-defaults and we get also $d_2\left(S_s, \overline{S}_s\right) = 1/6$. Even though both splits are equally rated, we find the second split more appropriate, since one of the subsets ($\overline{S}_s$) is totally pure and therefore for further partitions of the score range we only need to consider the complementary (S_s).

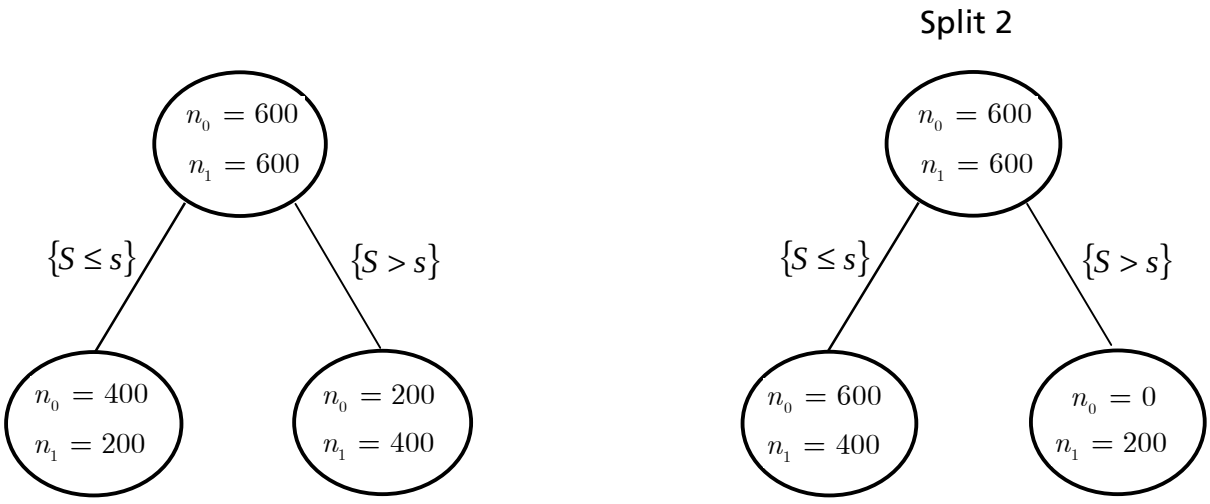


Figure 2.8: Two different splits with equal priors

Another defect of the misclassification rate is the degeneracy, which happens when for all splits in S there is no single or small number of best partitions. It occurs often in credit rating and we explain it in the next proposition.

Proposition 2.15

Let us suppose that

$\pi_1 < 1/2$, $P(1 \mid S_s) < 1/2$ and $P(1 \mid \overline{S}_s) < 1/2 \quad \forall S_s = \{S \leq s\}$.

Then it holds

$\Delta \mathrm{i}(S; S_s, \overline{S}_s) = 0 \quad \forall S_s$.

Proof:

By the definitions (2.4.2) and (2.4.5), and applying the theorem of total probability it is easy to see:

$$\Delta \mathrm{i}(S; S_s, \overline{S}_s) = \pi_1 - P(S_s)P(1 \mid S_s) - P(\overline{S}_s)P(1 \mid \overline{S}_s) = 0 .$$

□

In practice, we have that for a low probability of default, which is the normal case in credit scoring; the best split is normally given by misclassifying most of the defaults. It does not happen if we choose the entropy or the Gini index as impurity functions, as we can see in the examples of Section 2.4.6.

2.4.2 Another class of impurity functions

Our purpose in this section is to introduce another class of impurity functions that do not have the defects of the misclassification rate. We will introduce the condition of strictly concavity of ϕ due to two main reasons:

- It is necessary to avoid degeneracy (see the proof of Proposition 2.18).
- In a situation similar to Example 2.14 the impurity function should favour the second split.

Definition 2.16

A class C *of impurity functions* $\phi(\pi_1)$, $0 \le \pi_1 \le 1$, *with continuous second derivatives on* $0 \le \pi_1 \le 1$, *is defined as the class that satisfies*

(i) $\phi(1) = \phi(0) = 0$,

(ii) $\phi(\pi_1) = \phi(1-\pi_1)$,

(iii) $\phi"(\pi_1) < 0, \;\; 0 < \pi_1 < 1$, *i.e.* $\phi(\pi_1)$ *is strictly concave.*

The Gini index belongs to class C, since

$\phi_g(\pi_1) = 2\pi_1(1-\pi_1)$, $\phi_g"(\pi_1) = -4 < 0$,

and also does the entropy:

$\phi_e(\pi_1) = -\pi_1 \log \pi_1 - (1-\pi_1)\log(1-\pi_1)$, $\phi_e"(\pi_1) = -1/\pi_1(1-\pi_1) < 0$.

Since $\phi_e"(\pi_1) < \phi_g"(\pi_1)$ for $0 < \pi_1 < 1/2$, we have that the entropy increases faster than the Gini index as π_1 increases. For $1/2 < \pi_1 < 1$ we have $\phi_e"(\pi_1) > \phi_g"(\pi_1)$, which means that the entropy decreases faster than the Gini index as π_1 increases (see Figure 2.7).

Example 2.17

In Example 2.14 we obtained for the misclassification rate:

$d_2\left(S_s, \overline{S}_s\right) - d_1\left(S_s, \overline{S}_s\right) = 0$.

If we choose the Gini index or the entropy as impurity functions, we have:

$$d_1\left(S_s, \overline{S}_s\right) = \phi\left(\frac{1}{2}\right) - \frac{1}{2}\phi\left(\frac{1}{3}\right) - \frac{1}{2}\phi\left(\frac{2}{3}\right) \text{ and } d_2\left(S_s, \overline{S}_s\right) = \phi\left(\frac{1}{2}\right) - \frac{5}{6}\phi\left(\frac{2}{5}\right) - \frac{1}{6}\phi(1),$$

but the difference in this case is $d_2\left(S_s, \overline{S}_s\right) - d_1\left(S_s, \overline{S}_s\right) > 0$, which means that the second split is preferred to the first one.

Actually,

$$d_2\left(S_s,\overline{S}_s\right)-d_1\left(S_s,\overline{S}_s\right)=-\frac{5}{6}\phi\left(\frac{2}{5}\right)+\frac{1}{2}\phi\left(\frac{1}{3}\right)+\frac{1}{2}\phi\left(\frac{2}{3}\right)>0$$
$$\Leftrightarrow \frac{5}{6}\phi\left(\frac{2}{5}\right)<\frac{1}{2}\phi\left(\frac{1}{3}\right)+\frac{1}{2}\phi\left(\frac{2}{3}\right)\Leftrightarrow (1)<(2)$$

Because of the symmetry of the impurity functions we get:

$$(2)=\frac{1}{2}\phi\left(\frac{1}{3}\right)+\frac{1}{2}\phi\left(\frac{2}{3}\right)=\phi\left(\frac{1}{3}\right).$$

And because of strictly concavity:

$$(1)=\frac{5}{6}\phi\left(\frac{2}{5}\right)=\frac{5}{6}\phi\left(\frac{2}{5}\right)+\frac{1}{6}\phi(0)<\phi\left(\frac{5}{6}\cdot\frac{2}{5}+\frac{1}{6}\cdot 0\right)=\phi\left(\frac{1}{3}\right)=(2).$$

The problem of the misclassification rate is that $\phi_{mr}(\pi_1)=\min(\pi_1,1-\pi_1)=\pi_1$ decreases only linearly in π_1. Therefore we had to require that, as π_1 increases, $\phi(\pi_1)$ decreases faster than linearly, which means that the impurity function should be strictly concave. Hence, if $\pi_1'' > \pi_1'$, we want $\phi(\pi_1'')$ be less than the point on the tangent line at π_1'' (see Figure 2.9).

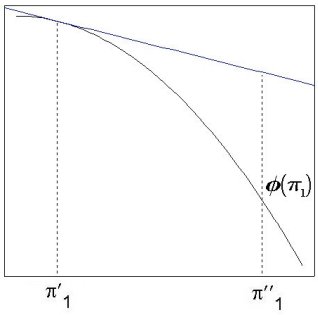


Figure 2.9: A strictly concave impurity function $\phi(\pi_1)$

The following proposition states that strictly concave impurity functions never lead to degeneracy as the misclassification rate does.

Proposition 2.18

Let $\phi(\pi_1)$ be a strictly concave function on $0\le\pi_1\le 1$. Then for every partition of S at s,

$\Delta i\left(S;S_s,\overline{S}_s\right)\ge 0$,

with equality if, and only if, $P(1\mid S_s)=P\left(1\mid\overline{S}_s\right)=\pi_1$.

Proof:

By the strict concavity of ϕ,

$$\begin{aligned} \mathrm{i}(S_s)P(S_s) + \mathrm{i}(\overline{S}_s)P(\overline{S}_s) &= \phi\big(P(1 \mid S_s)\big)P(S_s) + \phi\Big(P\big(1 \mid \overline{S}_s\big)\Big)P(\overline{S}_s) \\ &\leq \phi\Big(P(1 \mid S_s)P(S_s) + P\big(1 \mid \overline{S}_s\big)P(\overline{S}_s)\Big) = \mathrm{i}(S) \end{aligned}$$

with equality holding if, and only if, $P(1 \mid S_s) = P\big(1 \mid \overline{S}_s\big) = \pi_1$.

□

2.4.3 Gini index

In section 2.3.2 the Gini coefficient was formulated in terms of the distribution of the score S conditioned to Y, but there is no obvious relation with the Gini index defined here. Indeed, there are many representations of the Gini coefficient as a measure of discriminatory power (or concentration), and also of the Gini index as a measure of impurity (or heterogeneity); Gini proposed some of them, but there are also some versions introduced by other authors (see Piesch, 1975). The Gini index as impurity function belonging to class C of a score gives us information about the heterogeneity of $Y \mid S$ and it is given by

$$\mathrm{i}(S) = \sum_{i \neq j} P(i \mid S)P(j \mid S) = 1 - \sum_{j=0,1} P(j \mid S)^2 = 2P(0 \mid S)P(1 \mid S) \qquad (2.4.7)$$

By substituting in (2.4.3) and (2.4.4), we get the maximal distance function and D, as follows

$$D_g = 1 - \min_s \left\{ \frac{P(S_s)P(0 \mid S_s)P(1 \mid S_s) + P(\overline{S}_s)P(0 \mid \overline{S}_s)P(1 \mid \overline{S}_s)}{\pi_0 \pi_1} \right\} \qquad (2.4.8)$$

The Gini index is simple and has two interesting interpretations:

(1) It is twice the variance of the default variable conditioned to S: $\mathrm{i}(S) = 2\,\mathrm{var}(Y \mid S)$.

(2) If, given S, we choose a rule of classification that assigns $\widehat{Y} = i$ to a posterior observation selected at random with probability $P(i \mid S)$, then the probability that this observation is actually $Y = j$ is $P(j \mid S)$. Hence, the probability of misclassification under this rule is the Gini index: $\sum_{i \neq j} P(i \mid S)P(j \mid S)$.

2.4.4 Entropy

Information entropy is normally regarded in the theory of communication as a summary measure of the "information uncertainty" that a probability distribution represents, but it can also be seen as a measure for the heterogeneity, or the impurity of that distribution. The conditional entropy of a score belongs to class C and it is settled:

$$\mathrm{i}(S) = -\sum_{j=0,1} P(j \mid S)\log(P(j \mid S)). \tag{2.4.9}$$

By substituting in (2.4.3), we get the maximal distance function and finally we get the standardized D from (2.4.4):

$$D_e = 1 - \min_s \left\{ \frac{P(S_s)\left(P_{S_s}(0)\log P_{S_s}(0) + P_{S_s}(1)\log P_{S_s}(1)\right) + P(\overline{S}_s)\left(P_{\overline{S}_s}(0)\log P_{\overline{S}_s}(0) + P_{\overline{S}_s}(1)\log P_{\overline{S}_s}(1)\right)}{\pi_0\log(\pi_0) + \pi_1\log(\pi_1)} \right\},$$

being $P_S(j) = P\left(Y = j \mid S\right)$.

(2.4.10)

This criterion has been paid no attention, due to the misconception in the Basel on Banking Supervision (2005), where they argue that the entropy-based measures have a limited use for validation, since there are no applicable statistical tests for comparisons. However, we found in Tutz (2000), that the deviance of the test for homogeneity in 2×2 contingency tables is linearly related to the estimate of the distance (2.4.2) with the entropy as impurity function. Therefore, a test can be constructed and we will see it in the following section.

2.4.5 Test for homogenity in 2x2 contingency tables

Let us consider a partition of S in S_s and $\overline{S}_s$, such that $S = S_s \cup \overline{S}_s$. Then we can build up the following 2×2 contingency table for non-defaults and defaults:

		Y		
S		0	1	
	S_s	$n_0\left(S_s\right)$	$n_1\left(S_s\right)$	$n\left(S_s\right)$
	$\overline{S}_s$	$n_0\left(\overline{S}_s\right)$	$n_1\left(\overline{S}_s\right)$	$n\left(\overline{S}_s\right)$
		n_0	n_1	n

In order to test the hypothesis:

$$H_0 : P(1 \mid S_s) = P(1 \mid \overline{S}_s), \quad H_1 : P(1 \mid S_s) \neq P(1 \mid \overline{S}_s), \tag{2.4.11}$$

we can apply the deviance criterion:

$$Dev\left(S; S_s, \overline{S}_s\right) = Dev(S) - \left(Dev\left(S_s\right) + Dev\left(\overline{S}_s\right)\right) \overset{(a)}{\sim} {\chi_1}^2, \tag{2.4.12}$$

being the respective deviances:

$$Dev(S) = -2\left(n_1 \log\left(\frac{n_1}{n}\right) + n_0 \log\left(\frac{n_0}{n}\right)\right),$$

$$Dev\left(S_s\right) = -2\left(n_1\left(S_s\right)\log\left(\frac{n_1\left(S_s\right)}{n\left(S_s\right)}\right) + n_0\left(S_s\right)\log\left(\frac{n_0\left(S_s\right)}{n\left(S_s\right)}\right)\right) \text{ and}$$

$$Dev\left(\overline{S}_s\right) = -2\left(n_1\left(\overline{S}_s\right)\log\left(\frac{n_1\left(\overline{S}_s\right)}{n\left(\overline{S}_s\right)}\right) + n_0\left(\overline{S}_s\right)\log\left(\frac{n_0\left(\overline{S}_s\right)}{n\left(\overline{S}_s\right)}\right)\right).$$

If we use the entropy (2.4.9) as impurity function, then it is easy to see the following relationship:

$$\widehat{\Delta \mathrm{i}}\left(S; S_s, \overline{S}_s\right) = \frac{1}{2n} Dev\left(S; S_s, \overline{S}_s\right) = \frac{1}{2n}\left(Dev(S) - \left(Dev\left(S_s\right) + Dev\left(\overline{S}_s\right)\right)\right).$$

For $\widehat{d_e} = \max_s \widehat{\Delta \mathrm{i}}\left(S; S_s, \overline{S}_s\right)$, it also holds:

$$Dev\left(S; S_s, \overline{S}_s\right) = 2n\widehat{d_e} = 2n\left(\widehat{D_e} \cdot \widehat{d_{opt}}\right), \tag{2.4.13}$$

which we adopt as our test statistic. We will reject H_0 for a level of significance α, if $Dev\left(S; S_s, \overline{S}_s\right) > \chi^2_{1;1-\alpha}$.

2.4.6 Examples

As we did for the accuracy ratio and the overlapping area, here we are comparing the results given by the diverse impurity functions for *score1* and *score2* at the validation sample obtained at random from the Fahrmeir et al. (1984) dataset. We also tested these results for the entropy. The misclassification rate was included in order to illustrate how sometimes it can lead to good results, if the proportion of defaults is not low.

Example 2.19

For *score1*, we get:

- Misclassification rate: $\widehat{D_{mr}}$ = 0.240 and the split point s_{mr} = -0.053.
- Gini index: $\widehat{D_g}$ = 0.245 and s_g = -1.130.
- Entropy: $\widehat{D_e}$ = 0.227 and s_e = -1.287.

We can observe that s, which is the value of the score that maximizes the distance function, is close for the Gini index and the entropy and close to the one given by T. The values of D vary slightly, and they are neither too close to 0, nor to 1, which means the score is not extremely bad or good discriminating.

For this sample we have n = 200. If we consider the entropy as impurity function, then $\widehat{D_e}$ = 0.227, $\widehat{d_{opt}}$ = 0.583, and subtituting in (2.4.13) we obtain $Dev\left(S;S_s,\overline{S}_s\right)$ = 53.084. The critical value for a level of significance α = 0.005 is given by $\chi^2_{1;0.995}$ = 7.88. Our test statistic is highly significant, since $Dev\left(S;S_s,\overline{S}_s\right)$ = 53.084 > 7.88. Therefore, we reject the null hypothesis in (2.4.11), i.e. H_0 : $P(1 \mid S_s) = P(1 \mid \overline{S}_s)$. Figure 2.10 shows the estimates for the density functions of non-defaults and defaults and the split points.

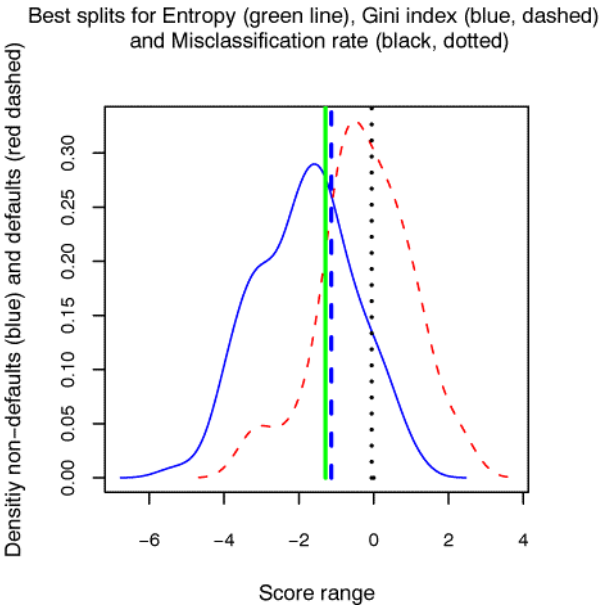


Figure 2.10: different split points for *score1*

Example 2.20

For *score2,* we obtain:

- Misclassification rate: $\widehat{D_{mr}}$ = 0.166 and s_{mr} = -0.335.
- Gini index: $\widehat{D_g}$ = 0.117 and s_g = -0.335.
- Entropy: $\widehat{D_e}$ = 0.087 and s_e = -0.335.

From these results we can conclude that *score1* has more discriminatory power than *score2*. The split points are equal for every impurity function (see Figure 2.11).

If we consider the entropy as impurity function, we have again $\widehat{d_{opt}}$ = 0.583 and $\chi^2_{1;0.995}$ = 7.88 for a level of significance α = 0.005. Then we obtain $\widehat{D_e}$ = 0.087, and subtituting in (2.4.13), we get $Dev\left(S;S_s,\overline{S}_s\right)$ = 20.475, which is also highly significant, since 20.475 > 7.88. Here again we will reject the null hypothesis in favour of the alternative in (2.4.11).

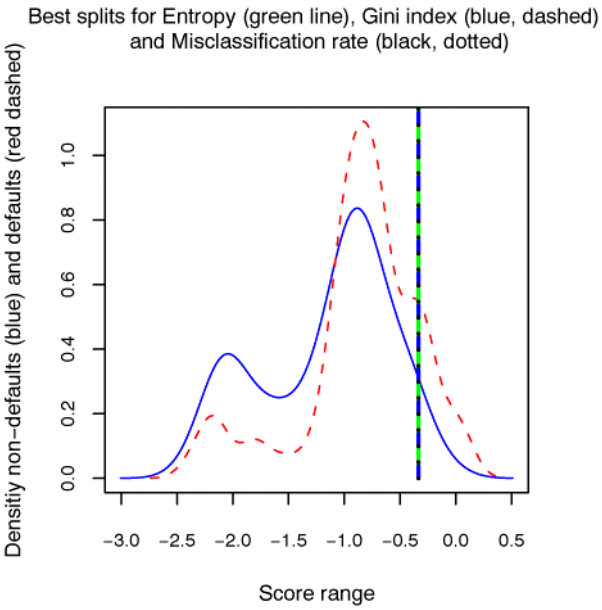


Figure 2.11: the same split points for *score2*

2.4.7 Comparison of entropy and Gini index

Having seen the defects of the misclassification rate, it remains only to compare the Gini index and the entropy as splitting criteria for the standardized maximal distance D. This comparison cannot be done generally, but only under some assumptions. With this purpose we will study in this section four cases. The first of these is linked to a proposition. Examples

and simulations follow the other three, since it is not possible to compare both measures theoretically, and we will explain the reasons.

For ease of calculations, we will consider in the following for the entropy s_p, solution of $d = \max_s d_s(S \le s, S > s)$ and for the Gini index the optimal split point s_q.

Case 1

We will study a case where the default variable Y and the score S are independent.

Proposition 2.21

Assume we have a random variable S and Y is a Bernoulli such that:

$P(Y=1 \mid S=s) = P(Y=1) = \pi_1,\ \forall s \in \mathbb{R}$.

Then, it holds: $D_e = D_g = 0$.

Proof:

We must calculate the following conditioned probabilities:

$$P(Y=1 \mid S \le s_i) = \frac{P(Y=1, S \le s_i)}{P(S \le s_i)} = \frac{P(Y=1)P(S \le s_i)}{P(S \le s_i)} = \pi_1, \text{ for } i = p,q.$$

Similarly we get:

$$P(Y=0 \mid S \le s_i) = 1 - P(Y=1 \mid S \le s_i) = 1 - \pi_1,$$
$$P(Y=1 \mid S > s_i) = P(Y=1 \mid S > s_i) = \pi_1 \text{ and}$$
$$P(Y=0 \mid S > s_i) = P(Y=0 \mid S > s_i) = 1 - P(Y=1 \mid S > s_i) = 1 - \pi_1, \text{ for } i = p,q.$$

1. Let us first prove $D_e = 0$.

The impurity functions are:

$$\mathrm{i}(S \le s_p) = \mathrm{i}(S > s_p) = \mathrm{i}(S) = \pi_1 \log(\pi_1) + (1-\pi_1)\log(1-\pi_1)$$

Therefore,

$$D_e = 1 - \frac{\left(P\left(S \le s_p\right) + P\left(S > s_p\right)\right)\left(\pi_1 \log\left(\pi_1\right) + \left(1-\pi_1\right)\log\left(1-\pi_1\right)\right)}{\pi_1 \log\left(\pi_1\right) + \left(1-\pi_1\right)\log\left(1-\pi_1\right)} = 0\,.$$

2. And then for the Gini index $D_g = 0$.

Here we have:

$$\mathrm{i}\left(S \le s_q\right) = \mathrm{i}\left(S > s_q\right) = \mathrm{i}(S) = 2\pi_1\left(1-\pi_1\right)$$

and we get:

$$D_g = 1 - \frac{P\left(S \le s_q\right)2\pi_1\left(1-\pi_1\right) + P\left(S > s_q\right)2\pi_1\left(1-\pi_1\right)}{2\pi_1\left(1-\pi_1\right)} = 0\,.$$

An alternative proof can be accomplished with the help of Proposition 2.18.

□

As it is obvious in this case, there is no difference between both measures of impurity. The score has no discriminatory power at all, and thus $D_e = D_g = 0$.

Case 2

We want to see here how the standardized maximal distance D and the split points vary as we choose the Gini index or the entropy as splitting criteria for a given score with a probability of default that increases as the score increases.

Proposition 2.22

Let us suppose that the probability of default increases as the score takes higher values and we have a score S, which is discrete uniformly distributed, being:

$P\left(S = s_j\right) = \frac{1}{n}$, $\forall j \in \{1,\ldots,n\}$ and

$P\left(Y = 1 \mid S = s_j\right) = c \cdot f\left(s_j\right)$, with

$c = \frac{P(Y=1)}{P\left(S = s_j\right)\sum_{j=1}^{n} f\left(s_j\right)}$ a constant, for $j = 1,\ldots,n$, and $f\left(s_j\right)$ monotonically increasing,

such that

$$P(Y=1)=\sum_{j=1}^{n}P\left(Y=1\mid S=s_j\right)P\left(S=s_j\right)=\sum_{j=1}^{n}P\left(Y=1\mid s_j\right)\frac{1}{n}=\frac{1}{n}\sum_{j=1}^{n}cf\left(s_j\right).$$

Further, assume the optimal split point s_p for the entropy and s_q for the Gini index, and call:

$$x=\frac{\sum_{j=1}^{p}P\left(Y=1\mid S=s_j\right)}{\sum_{j=1}^{n}P\left(Y=1\mid S=s_j\right)},\qquad \tilde{x}=\frac{\sum_{j=1}^{q}P\left(Y=1\mid S=s_j\right)}{\sum_{j=1}^{n}P\left(Y=1\mid S=s_j\right)}.$$

Then we get the following expressions for the standardized maximal distance:

1. For the entropy

$$D_e=1-\frac{\pi_1 x\log\left(\frac{x(n-p)}{(1-x)p}\right)+\left(\frac{p}{n}-\pi_1 x\right)\left(\log\left(1-\pi_1 x\frac{n}{p}\right)-\log\left(1-\pi_1(1-x)\frac{n}{n-p}\right)\right)}{\pi_1\log(\pi_1)+(1-\pi_1)\log(1-\pi_1)}$$
$$-\frac{(1-\pi_1)\log\left(1-\pi_1(1-x)\frac{n}{n-p}\right)+\pi_1\log\left(\pi_1(1-x)\frac{n}{n-p}\right)}{\pi_1\log(\pi_1)+(1-\pi_1)\log(1-\pi_1)}$$

We need the restriction $0\le\pi_1<\min\left\{\frac{p}{xn},\ 1,\ \frac{n-p}{(1-x)n}\right\}$, which is equivalent to:

$$\begin{cases}0\le\pi_1<\dfrac{p}{xn} & if\ \ p\le xn\\ 0\le\pi_1<\dfrac{n-p}{(1-x)n} & otherwise\end{cases}$$

so that the logarithms exist.

2. For the Gini index

$$D_g=1-\frac{1-\pi_1\frac{n}{(n-q)}\left(\frac{n}{q}\tilde{x}^2+1-2\tilde{x}\right)}{1-\pi_1},\ \text{for}\ 0\le\pi_1<1.$$

Proof:

We need to calculate the conditioned probabilities, for $i=p,q$:

$$P(Y=1\mid S\le s_i)=\frac{P(Y=1,S\le s_i)}{P(S\le s_i)}=\frac{P(S\le s_i\mid Y=1)P(Y=1)}{P(S\le s_i)}$$
$$=\frac{\sum_{j=1}^{i}P(Y=1\mid s_j)P(S=s_j)}{i/n}=\frac{\sum_{j=1}^{i}P(Y=1\mid s_j)\frac{1}{n}}{i/n}=\frac{1}{i}\sum_{j=1}^{i}P(Y=1\mid s_j),$$

and the probability of its complementary

$$P(Y=0\mid S\le s_i)=1-P(Y=1\mid S\le s_i)=1-\frac{1}{i}\sum_{j=1}^{i}P(Y=1\mid s_j).$$

Similarly, we obtain

$$P(Y=1\mid S>s_i)=\frac{1}{n-i}\sum_{j=i+1}^{n}P(Y=1\mid s_j)\text{ and}$$
$$P(Y=0\mid S>s_i)=1-\frac{1}{n-i}\sum_{j=i+1}^{n}P(Y=1\mid s_j).$$

1. Then we have for the entropy:

$$\mathrm{i}(S\le s_p)=$$
$$=\frac{1}{p}\sum_{j=1}^{p}P(1\mid s_j)\log\left(\frac{1}{p}\sum_{j=1}^{p}P(1\mid s_j)\right)+\left(1-\frac{1}{p}\sum_{j=1}^{p}P(1\mid s_j)\right)\log\left(1-\frac{1}{p}\sum_{j=1}^{p}P(1\mid s_j)\right)$$
$$=\frac{1}{p}\frac{\sum_{j=1}^{p}P(1\mid s_j)}{\frac{1}{n}\sum_{j=1}^{n}P(1\mid s_j)}\frac{1}{n}\sum_{j=1}^{n}P(1\mid s_j)\log\left(\frac{1}{p}\frac{\sum_{j=1}^{p}P(1\mid s_j)}{\frac{1}{n}\sum_{j=1}^{n}P(1\mid s_j)}\frac{1}{n}\sum_{j=1}^{n}P(1\mid s_j)\right)$$
$$+\log\left(1-\frac{1}{p}\sum_{j=1}^{p}P(1\mid s_j)\right)-\frac{1}{p}\sum_{j=1}^{p}P(1\mid s_j)\log\left(1-\frac{1}{p}\sum_{j=1}^{p}P(1\mid s_j)\right)$$
$$=\frac{n}{p}x\pi_1\log\left(\frac{n}{p}x\pi_1\right)+\log\left(1-\frac{n}{p}x\pi_1\right)-\frac{n}{p}x\pi_1\log\left(1-\frac{n}{p}x\pi_1\right).$$

And

$$\mathrm{i}(S>s_p)=$$

$$
\begin{aligned}
&= \frac{1}{n-p}\sum_{j=p+1}^{n} P\left(1 \mid s_j\right)\log\left(\frac{1}{n-p}\sum_{j=p+1}^{n} P\left(1 \mid s_j\right)\right) \\
&\qquad + \left(1-\frac{1}{n-p}\sum_{j=p+1}^{n} P\left(1 \mid s_j\right)\right)\log\left(1-\frac{1}{n-p}\sum_{j=p+1}^{n} P\left(1 \mid s_j\right)\right) \\
&= \frac{1}{n-p}\left(n\pi_1 - \sum_{j=1}^{p} P\left(1 \mid s_j\right)\right)\log\left(\frac{1}{n-p}\left(n\pi_1 - \sum_{j=1}^{p} P\left(1 \mid s_j\right)\right)\right) \\
&\qquad + \left(1-\frac{1}{n-p}\left(n\pi_1 - \sum_{j=1}^{p} P\left(1 \mid s_j\right)\right)\right)\log\left(1-\frac{1}{n-p}\left(n\pi_1 - \sum_{j=1}^{p} P\left(1 \mid s_j\right)\right)\right) \\
&= \frac{1}{n-p}\left(n\pi_1 - n\pi_1 x\right)\log\left(\frac{1}{n-p}\left(n\pi_1 - n\pi_1 x\right)\right) \\
&\qquad + \left(1-\frac{1}{n-p}\left(n\pi_1 - n\pi_1 x\right)\right)\log\left(1-\frac{1}{n-p}\left(n\pi_1 - n\pi_1 x\right)\right) \\
&= \frac{n}{n-p}\pi_1\left(1-x\right)\log\left(\frac{n}{n-p}\pi_1\left(1-x\right)\right) + \left(1-\frac{n}{n-p}\pi_1\left(1-x\right)\right)\log\left(1-\frac{n}{n-p}\pi_1\left(1-x\right)\right)
\end{aligned}
$$

We must calculate

$$
\begin{aligned}
&\frac{p}{n}\mathrm{i}\left(S \le s_p\right) + \frac{n-p}{n}\mathrm{i}\left(S > s_p\right) = \\
&\quad = x\pi_1 \log\left(\frac{n}{p}x\pi_1\right) + \frac{p}{n}\log\left(1-\frac{n}{p}x\pi_1\right) - x\pi_1\log\left(1-\frac{n}{p}x\pi_1\right) \\
&\qquad + \pi_1\left(1-x\right)\log\left(\frac{n}{n-p}\pi_1\left(1-x\right)\right) + \left(\frac{n-p}{n} - \pi_1\left(1-x\right)\right)\log\left(1-\frac{n}{n-p}\pi_1\left(1-x\right)\right) \\
&\quad = x\pi_1 \log\left(\frac{n}{p}x\pi_1\right) + \frac{p}{n}\log\left(1-\frac{n}{p}x\pi_1\right) - x\pi_1\log\left(1-\frac{n}{p}x\pi_1\right) \\
&\qquad\qquad + \pi_1\log\left(\frac{n}{n-p}\pi_1\left(1-x\right)\right) - x\pi_1\log\left(\frac{n}{n-p}\pi_1\left(1-x\right)\right) \\
&\qquad\qquad + \log\left(1-\frac{n}{n-p}\pi_1\left(1-x\right)\right) - \frac{p}{n}\log\left(1-\frac{n}{n-p}\pi_1\left(1-x\right)\right) \\
&\qquad\qquad - \pi_1\log\left(1-\frac{n}{n-p}\pi_1\left(1-x\right)\right) + x\pi_1\log\left(1-\frac{n}{n-p}\pi_1\left(1-x\right)\right) \\
&\quad = \pi_1 x\log\left(\frac{x\left(n-p\right)}{\left(1-x\right)p}\right) + \left(\frac{p}{n} - \pi_1 x\right)\left(\log\left(1-\pi_1 x\frac{n}{p}\right) - \log\left(1-\pi_1\left(1-x\right)\frac{n}{n-p}\right)\right) \\
&\qquad\qquad + \left(1-\pi_1\right)\log\left(1-\pi_1\left(1-x\right)\frac{n}{n-p}\right) + \pi_1\log\left(\pi_1\left(1-x\right)\frac{n}{n-p}\right)
\end{aligned}
$$

By substituting it into (2.4.10) we get the above-mentioned expression of D_e.

2. For the Gini index we need to calculate

$$\frac{q}{n}P\left(Y=1\mid S\le s_q\right)P\left(Y=0\mid S\le s_q\right)+\frac{n-q}{n}P\left(Y=1\mid S>s_q\right)P\left(Y=0\mid S>s_q\right)=$$

$$=\frac{q}{n}\left(\frac{1}{q}\sum_{j=1}^{q}P\left(Y=1\mid s_j\right)-\frac{1}{q}\left(\sum_{j=1}^{q}P\left(Y=1\mid s_j\right)\right)^2\right)$$

$$+\frac{n-q}{n}\left(\frac{1}{n-q}\sum_{j=q+1}^{n}P\left(Y=1\mid s_j\right)-\frac{1}{n-q}\left(\sum_{j=q+1}^{n}P\left(Y=1\mid s_j\right)\right)^2\right)$$

$$=\frac{1}{n}\sum_{j=1}^{n}P\left(Y=1\mid s_j\right)-\frac{1}{n}\left(\frac{1}{q}\left(\sum_{j=1}^{q}P\left(Y=1\mid s_j\right)\right)^2+\frac{1}{n-q}\left(\sum_{j=q+1}^{n}P\left(Y=1\mid s_j\right)\right)^2\right)$$

$$=(1)+(2)$$

In order to simplify, we write

$$\frac{1}{n-q}\left(\sum_{j=q+1}^{n}P\left(Y=1\mid s_j\right)\right)^2=\frac{1}{n-q}\left(\frac{n}{n}\sum_{j=1}^{n}P\left(Y=1\mid s_j\right)-\sum_{j=1}^{q}P\left(Y=1\mid s_j\right)\right)^2$$

$$=\frac{1}{n-q}\left(n\pi_1-\sum_{j=1}^{q}P\left(Y=1\mid s_j\right)\right)^2=\frac{1}{n-q}\left(\left(n\pi_1\right)^2+\left(\sum_{j=1}^{q}P\left(Y=1\mid s_j\right)\right)^2\right)$$

$$-\frac{1}{n-q}2n\pi_1\sum_{j=1}^{q}P\left(Y=1\mid s_j\right)$$

We have $(1)=\pi_1$ and

$$(2)=-\frac{1}{n}\left(\frac{1}{q}\left(\sum_{j=1}^{q}P\left(Y=1\mid s_j\right)\right)^2+\frac{1}{n-q}\left(\sum_{j=q+1}^{n}P\left(Y=1\mid s_j\right)\right)^2\right)$$

$$=-\frac{1}{n}\left(\left(\sum_{j=1}^{q}P\left(Y=1\mid s_j\right)\right)^2\left(\frac{n-q+q}{(n-q)q}\right)\right)+\frac{1}{n}\frac{1}{n-q}\left(n\pi_1\right)^2-2\pi_1\frac{1}{n-q}\sum_{j=1}^{q}P\left(Y=1\mid s_j\right)$$

$$=-\frac{1}{n-q}\left(\frac{1}{q}\left(\sum_{j=1}^{q}P\left(Y=1\mid s_j\right)\right)^2+n\pi_1^{\,2}-2\pi_1\sum_{j=1}^{q}P\left(Y=1\mid s_j\right)\right)$$

Therefore, we obtain

$$D_g = 1 - \frac{(1)+(2)}{\pi_1(1-\pi_1)}$$

$$= 1 - \frac{\pi_1 - \frac{1}{n-q}\left(\frac{1}{q}\left(\sum_{j=1}^{q} P\left(Y=1 \mid s_j\right)\right)^2 + n\pi_1{}^2 - 2\pi_1 \sum_{j=1}^{q} P\left(Y=1 \mid s_j\right)\right)}{\pi_1(1-\pi_1)}$$

$$= 1 - \frac{1 - \frac{1}{n-q}\left(\frac{1}{q}\pi_1\left(\frac{\sum_{j=1}^{q} P\left(Y=1 \mid s_j\right)}{\pi_1}\right)^2 + n\pi_1 - 2\pi_1\left(\frac{\sum_{j=1}^{q} P\left(Y=1 \mid s_j\right)}{\pi_1}\right)\right)}{1-\pi_1}$$

$$= 1 - \frac{1 - \frac{1}{n-q}\left(\frac{n^2}{q}\pi_1\left(\frac{\sum_{j=1}^{q} P\left(Y=1 \mid s_j\right)}{\sum_{j=1}^{n} P\left(Y=1 \mid s_j\right)}\right)^2 + n\pi_1 - 2\pi_1 n\left(\frac{\sum_{j=1}^{q} P\left(Y=1 \mid s_j\right)}{\sum_{j=1}^{n} P\left(Y=1 \mid s_j\right)}\right)\right)}{1-\pi_1}$$

$$= 1 - \frac{1 - \frac{1}{n-q}\left(\frac{n^2}{q}\pi_1\tilde{x}^2 + n\pi_1 - 2\pi_1 n\tilde{x}\right)}{1-\pi_1}$$

□

Proposition 2.23

We can write also

$D_g = 1 - \frac{1-\pi_1 k}{1-\pi_1}$, for $0 \le \pi_1 < 1$,

being $k = \frac{n}{(n-q)}\left(\frac{n}{q}\tilde{x}^2 + 1 - 2\tilde{x}\right) \ge 1$.

Proof:

We want to prove here that $k \ge 1$:

$$k = \frac{n}{(n-q)}\left(\frac{n}{q}\tilde{x}^2 + 1 - 2\tilde{x}\right) \ge 1 \quad \Leftrightarrow \quad \frac{n}{q}\tilde{x}^2 - 2\tilde{x} + 1 - \frac{n-q}{n} \ge 0$$

$$\frac{n}{q}\tilde{x}^2 - 2\tilde{x} + 1 - \frac{n-q}{n} = \frac{n}{q}\tilde{x}^2 - 2\tilde{x} + \frac{q}{n} = \tilde{x}^2 - 2\frac{q}{n}\tilde{x} + \left(\frac{q}{n}\right)^2 = \left(\tilde{x} - \frac{q}{n}\right)^2 \ge 0$$

□

In order to compare both standardized maximal distances, we calculate the first and second derivatives with respect to the default probability π_1. For D_g they are:

1. $\frac{dD_g}{d\pi_1} = \frac{k-1}{(1-\pi_1)^2} \geq 0 \quad \forall \pi_1 : 0 \leq \pi_1 < 1,$

   since $k \geq 1$, which means that D_g increases as π_1 increases.

2. $\frac{d^2 D_g}{d\pi_1^2} = \frac{2(k-1)}{(1-\pi_1)^3} \geq 0 \quad \forall \pi_1 : 0 \leq \pi_1 < 1.$

   Together with the fact $0 \leq D_g \leq 1$, we have that D_g as function of π_1 is convex.

The first and second derivatives of D_e with respect to π_1 are given by:

1. $$\frac{dD_e}{d\pi_1} = \frac{1}{n\left((\pi_1 - 1)\ln(1-\pi_1) - \pi_1 \ln \pi_1\right)^2}$$
$$\cdot \left(\ln \pi_1 \left((n-p) \ln\left(1 - \frac{n\pi_1(1-x)}{n-p}\right) + p \ln\left(1 - \frac{n\pi_1 x}{p}\right)\right)\right.$$
$$+ \ln(1-\pi_1)\left((p-nx)\ln\left(1 - \frac{n\pi_1(1-x)}{n-p}\right)\right.$$
$$\left.\left. - n\left(\ln\left(\frac{n\pi_1(1-x)}{n-p}\right) + x \ln\left(\frac{(n-p)x}{p(1-x)}\right)\right) + (nx-p)\ln\left(1 - \frac{n\pi_1 x}{p}\right)\right)\right)$$

2. $$\frac{d^2 D_e}{d\pi_1^2} = \frac{1}{\left(-(-1+\pi_1)\ln(-1+\pi_1) + \pi_1 \ln(\pi_1)\right)^3}$$
$$\cdot \left(\frac{\left(np - p^2 + n^2 \pi_1 (x-1)x\right)\left(-1 + \pi_1 \ln(1-\pi_1) - \pi_1 \ln \pi_1\right)^2}{\pi_1\left(-p + n\left(1 + \pi_1(-1+x)\right)\right)\left(-p + n\pi_1 x\right)}\right.$$
$$+ 2\left(\ln(1-\pi_1) - \ln(\pi_1)\right)\left((-1+\pi_1)\ln(1-\pi_1) - \pi_1 \ln(\pi_1)\right)$$
$$\cdot \left((-1+x)\ln\left(1 + \frac{n\pi_1(-1+x)}{n-p}\right) + \ln\left(-\frac{n\pi_1(-1+x)}{n-p}\right)\right.$$
$$\left.\left. + x\left(\ln\left(\frac{(-n+p)x}{p(-1+x)}\right) - \ln\left(1 - \frac{n\pi_1 x}{p}\right)\right)\right)\right.$$
$$+ n\pi_1 \ln\left(-\frac{n\pi_1(-1+x)}{n-p}\right) + n\pi_1 x \ln\left(\frac{(-n+p)x}{p(-1+x)}\right) + \ln\left(1 - \frac{n\pi_1 x}{p}\right)\left(p - n\pi_1 x\right)\Bigg)$$
$$- \frac{\left(2\left(\ln(1-\pi_1) - \ln \pi_1\right)^2\right)}{n}\left((n - n\pi_1 - p + n\pi_1 x)\ln\left(1 + \frac{n\pi_1(-1+x)}{n-p}\right)\right)$$

$$+\frac{1}{n(-1+\pi_1)\pi_1}\left(\left((-1+\pi_1)\ln(1-\pi_1)-\pi_1\ln\pi_1\right)\left((n-n\pi_1-p-n\pi_1 x)\ln\left(1-\frac{n\pi_1 x}{p}\right)\right.\right.$$
$$\left.\left.+n\pi_1\ln\left(-\frac{n\pi_1(-1+x)}{n-p}\right)+n\pi_1 x\ln\left(\frac{(-n+p)x}{p(-1+x)}\right)+\ln\left(1-\frac{n\pi_1 x}{p}\right)(p-n\pi_1 x)\right)\right)$$

But from these derivatives we cannot guess if the function is increasing or decreasing, because there is no solution in π_1 for the equation $D_e = 0$, neither for

$\frac{dD_e}{d\pi_1} = 0$, nor $\frac{d^2 D_e}{d\pi_1^{\,2}} = 0$.

We would like to know, as we did for the Gini index, if the first or second derivatives are always positive or negative, but it is not possible here. For example, given $n = 700$, $p = 100$ and $x = 0.1$, then for two different probabilities of default:

- $\pi_1 = 0.6$, the second derivative is positive $\frac{d^2 D_e}{d\pi_1^{\,2}} = 0.446$,
- and for $\pi_1 = 0.05$, it is negative $\frac{d^2 D_e}{d\pi_1^{\,2}} = -0.077$.

In this case, depending on the value of π_1, D_e can be convex or concave.

We would like to compare theoretically the standardized maximal distance for the entropy and the Gini index. However, this comparison has to be numerical, because, as there is no solution in π_1 for the equation $D_e = 0$, there is also no solution for $D_e - D_g = 0$. Therefore, we illustrate how both $D's$ can be compared graphically in the following example.

Example 2.24

We choose a sample of $n = 700$ score values, having both entropy and Gini index the same optimal split point at $p = q = 100$, and thus $x = \tilde{x} = 0.3$. The standardized maximal distances D_e and D_g are depicted in Figure 2.12 (in the x-axis are represented the probabilities of default for $0 \leq \pi_1 < 0.476 = p/xn$). Here, it would be better to use the entropy as splitting criterion, since for lower probabilities of default, which is the case in the credit rating practice, it gives higher values of D.

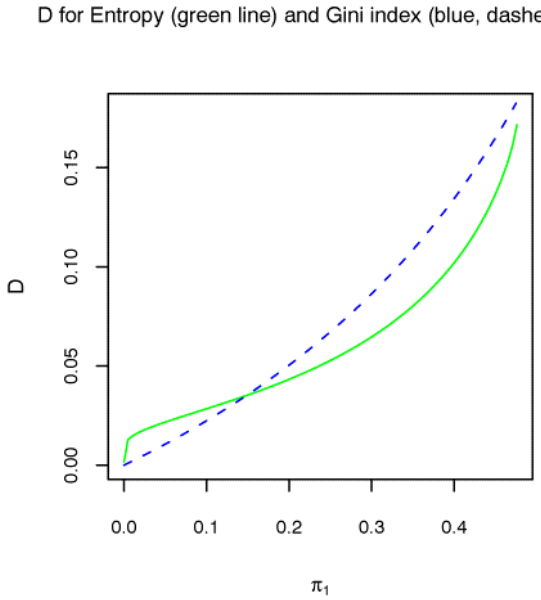


Figure 2.12: D_e and D_g for $n = 700$, $p = 100$, $x = 0.3$

Simulation 2.25

We also simulated a sample of $n = 500$ random values of a discrete uniformly distributed score $S \sim U(1,5)$. Then we considered the default variables:

$Y_1^k,\ldots,Y_n^k$, *i.i.d.*, $Y_i^k \sim Be\left(\frac{k}{500}\right)$, for $i,k = 1,\ldots,500$, with $P\left(Y_i^k = 1\right) = \frac{k}{500}$,

and call: $Y^k = \frac{1}{n}\sum_{i=1}^{n} Y_i^k$ "Proportion of defaults in the sample", being

$Y^k \sim Bi\left(n, \frac{k}{500}\right)$, such that $P\left(Y^k = 1\right) = E\left(Y^k\right) = \frac{k}{500}$.

We would like that the probability of default increases as the score takes higher values:

$Y_i^k \mid s_j \sim Be\left(w_k / 1 + \exp\left(-s_j\right)\right)$, with $w_k = \frac{k/500}{\frac{1}{n}\sum_{j=1}^{n}\frac{1}{1+\exp(-s_j)}}$, for $i,j,k = 1,\ldots,500$,

such that

$$P\left(Y_i^k = 1\right) = \sum_{j=1}^{n} P\left(Y_i^k = 1 \mid s_j\right) P\left(s_j\right) = \frac{1}{n}\sum_{j=1}^{n}\frac{w_k}{1+\exp\left(-s_j\right)} = \frac{k}{500}.$$

Then, for every realisation s_j, $j = 1,\ldots,n$, of the score S, we choose $Y_j^k \mid s_j$ at random from a Bernoulli distribution with $P\left(Y_j^k = 1 \mid s_j\right) = w_k / 1 + \exp\left(-s_j\right)$. Finally, we calculated D for the score S and for every random sample $Y_1^k,\ldots,Y_n^k$, $k = 1,\ldots,500$.

We picture first in Figure 2.13 the different split points obtained from using the entropy and the Gini Index as splitting criteria. In the x-axis are represented the estimated probabilities of default $0 < \widehat{P}\left(Y^k = 1\right) < 1$.

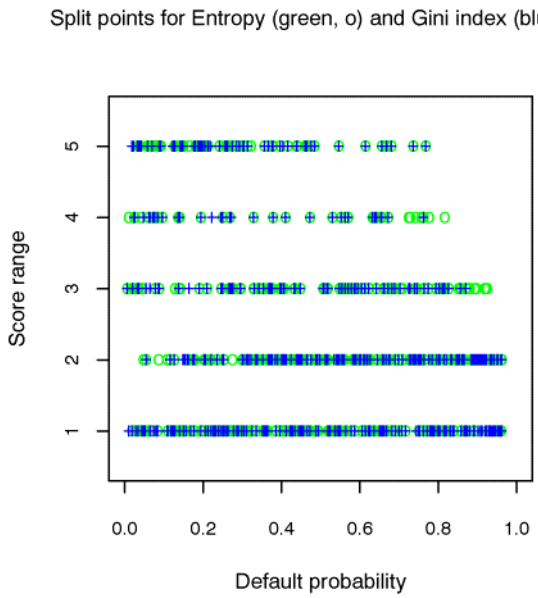


Figure 2.13: Split points for the simulation in case 2

In this picture and in the following table we can observe that most of the optimal split points coincide (82.4%). For the noncoincident, the entropy raises higher split points for higher PDs ($s_p > s_q$). When it comes to lower probabilities of default, we have that the split points obtained using Gini index are higher than those obtained using entropy ($s_q > s_p$). This means that the entropy criterion is a little more conservative than the Gini index, as we simulated probabilities of default increasing with the score values. The splits differ more from each other as the probability of default increases or decreases.

PD	$s_q = s_p$	$s_q > s_p$	$s_p > s_q$
(0,0.2]	62	23	11
(0.2,0.4]	89	10	3
(0.4,0.6]	96	3	0
(0.6,0.8]	93	0	11
(0.8,1]	72	0	27
Total	412 (82.4%)	36 (7.2%)	52 (10.4%)

Table 2.4: Split points for the simulation in case 2

In Figure 2.14, we represent the different estimates of D. It decreases for the Gini index as the probability of default decreases. For the entropy, D decreases and increases again for low $\widehat{P}\left(Y^k = 1\right)$. As regards very low probabilities of default, D decreases slightly more rapidly for the Gini index, being very close to 0. However, D is more stable with respect to the probability of default than the Gini index. Hence, we have no relevant evidence of a better criterion in this case.

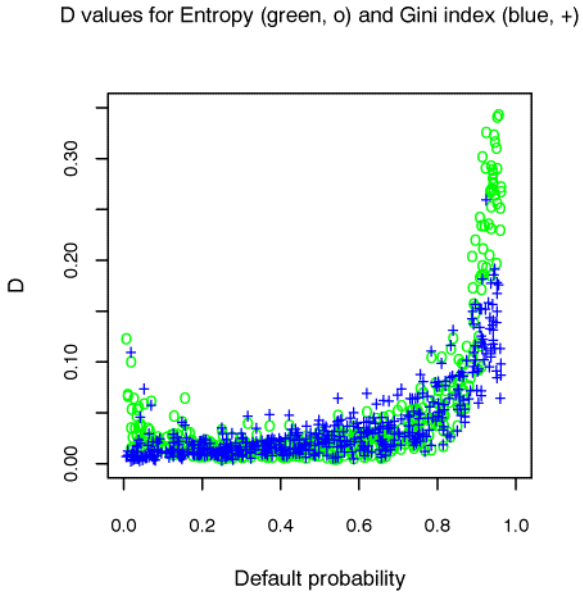


Figure 2.14: D values for the simulation in case 2

Case 3:

Here we want to see how the standardized maximal distance D and the split points vary as we choose the Gini index or the entropy as splitting criteria for a score that is normally distributed given default and non-default with different means, such that, as the difference between means increases, the overlapping area decreases and thus we can say that the score discriminates better.

Proposition 2.26

If we have a score with conditional distributions

$S \mid Y = 0 \sim N(0,1)$ and

$S \mid Y = 1 \sim N(\mu,1),\ \mu \geq 0,$

then the standardized maximal distance function can be written:

1. For the entropy, we assume that s_p is the optimal split point and we obtain

$$
\begin{aligned}
D_e = 1 - & \frac{1}{\pi_1 \log(\pi_1) + (1-\pi_1)\log(1-\pi_1)} \cdot \\
& \cdot \left(\pi_1 \Phi(s_p - \mu) \log\left(\frac{\pi_1 \Phi(s_p - \mu)}{\Phi(s_p - \mu)\pi_1 + \Phi(s_p)(1-\pi_1)} \right) \right. \\
& + (1-\pi_1)\Phi(s_p) \log\left(\frac{(1-\pi_1)\Phi(s_p)}{\Phi(s_p - \mu)\pi_1 + \Phi(s_p)(1-\pi_1)} \right) \\
& + \pi_1 \left(1 - \Phi(s_p - \mu)\right) \log\left(\frac{\pi_1 \left(1 - \Phi(s_p - \mu)\right)}{\left(1 - \Phi(s_p - \mu)\right)\pi_1 + \left(1 - \Phi(s_p)\right)(1-\pi_1)} \right) \\
& \left. + (1-\pi_1)\left(1 - \Phi(s_p)\right) \log\left(\frac{(1-\pi_1)\left(1 - \Phi(s_p)\right)}{\left(1 - \Phi(s_p - \mu)\right)\pi_1 + \left(1 - \Phi(s_p)\right)(1-\pi_1)} \right) \right)
\end{aligned}
$$

2. For the Gini index we assume that s_q is the optimal split point and we obtain

$$
D_g = 1 - \left(\frac{\Phi(s_q - \mu) \cdot \Phi(s_q)}{\pi_1 \Phi(s_q - \mu) + (1-\pi_1)\Phi(s_q)} + \frac{\left(1 - \Phi(s_q - \mu)\right) \cdot \left(1 - \Phi(s_q)\right)}{\pi_1 \left(1 - \Phi(s_q - \mu)\right) + (1-\pi_1)\left(1 - \Phi(s_q)\right)} \right),
$$

$\Phi(\cdot)$ being the distribution function of a normally distributed random variable with mean zero and variance one.

Proof:

For $j = p, q$, the probabilities are:

$$
\begin{aligned}
P(S \le s_j) &= P(S \le s_j \mid Y = 1) P(Y = 1) + P(S \le s_j \mid Y = 0) P(Y = 0) \\
&= \Phi(s_j - \mu)\pi_1 + \Phi(s_j)(1-\pi_1) \\
P(S > s_j) &= P(S > s_j \mid Y = 1) P(Y = 1) + P(S > s_j \mid Y = 0) P(Y = 0) \\
&= \left(1 - \Phi(s_j - \mu)\right)\pi_1 + \left(1 - \Phi(s_j)\right)(1-\pi_1)
\end{aligned}
$$

And the conditioned probabilities:

$$P\left(Y=1 \mid S \le s_j\right)=\frac{P\left(S \le s_j \mid Y=1\right)P(Y=1)}{P\left(S \le s_j\right)}=\frac{\pi_1\Phi\left(s_j-\mu\right)}{P\left(S \le s_j\right)}$$

$$P\left(Y=0 \mid S \le s_j\right)=\frac{P\left(S \le s_j \mid Y=0\right)P(Y=0)}{P\left(S \le s_j\right)}=\frac{\left(1-\pi_1\right)\Phi\left(s_j\right)}{P\left(S \le s_j\right)}$$

$$P\left(Y=1 \mid S > s_j\right)=\frac{P\left(S > s_j \mid Y=1\right)P(Y=1)}{P\left(S > s_j\right)}=\frac{\pi_1\left(1-\Phi\left(s_j-\mu\right)\right)}{P\left(S > s_j\right)}$$

$$P\left(Y=0 \mid S > s_j\right)=\frac{P\left(S > s_j \mid Y=0\right)P(Y=0)}{P\left(S > s_j\right)}=\frac{\left(1-\pi_1\right)\left(1-\Phi\left(s_j\right)\right)}{P\left(S > s_j\right)}$$

1. Then we have for the entropy the following impurity functions:

$$\mathrm{i}\left(S \le s_p\right)=\frac{\pi_1\Phi\left(s_p-\mu\right)}{P\left(S \le s_p\right)}\log\left(\frac{\pi_1\Phi\left(s_p-\mu\right)}{P\left(S \le s_p\right)}\right)+\frac{\left(1-\pi_1\right)\Phi\left(s_p\right)}{P\left(S \le s_p\right)}\log\left(\frac{\left(1-\pi_1\right)\Phi\left(s_p\right)}{P\left(S \le s_p\right)}\right)$$

$$\mathrm{i}\left(S > s_p\right)=\frac{\pi_1\left(1-\Phi\left(s_p-\mu\right)\right)}{P\left(S > s_p\right)}\log\left(\frac{\pi_1\left(1-\Phi\left(s_p-\mu\right)\right)}{P\left(S > s_p\right)}\right)$$
$$+\frac{\left(1-\pi_1\right)\left(1-\Phi\left(s_p\right)\right)}{P\left(S > s_p\right)}\log\left(\frac{\left(1-\pi_1\right)\left(1-\Phi\left(s_p\right)\right)}{P\left(S > s_p\right)}\right)$$

By substituting into (2.4.10) we obtain the above-mentioned expression of D_e.

2. And for the Gini index:

$$\mathrm{i}\left(S \le s_q\right)=2\frac{\pi_1\Phi\left(s_q-\mu\right)\left(1-\pi_1\right)\Phi\left(s_q\right)}{P\left(S \le s_q\right)^2},$$

$$\mathrm{i}\left(S > s_q\right)=2\frac{\pi_1\left(1-\Phi\left(s_q-\mu\right)\right)\left(1-\pi_1\right)\left(1-\Phi\left(s_q\right)\right)}{P\left(S > s_q\right)^2}$$

By substituting in (2.4.8) we get the last expression of the standardized maximal distance function D_g.

□

For both Gini index and entropy the first and second derivatives of D with respect to π_1 and with respect to μ are very complicated and they are in terms of the normal cumulative

distribution function, which can only be computed numerically or otherwise approximated. Therefore, we can study and compare D_e and D_g only numerically.

Example 2.27

We can see (Figure 2.15) how the graphics look like for different probabilities of default, having the same split point $s_p = s_q = 1$ and $\mu = 3$. In this case D_e is more stable with respect to the probability of default than D_g, i.e. D_e is higher for extreme values of the default probability. Since having very low probabilities of default is the normal case in credit scoring, we can affirm that, in this case, D_e is a better measure of the discriminatory power than D_g.

In Figure 2.16 we can appreciate how D looks like for different means μ, for a given probability of default $\pi_1 = 0.1$ and $s_p = s_q = 3$ in the following picture. We have for both Gini index and entropy that D increases as μ increases, in a very similar way.

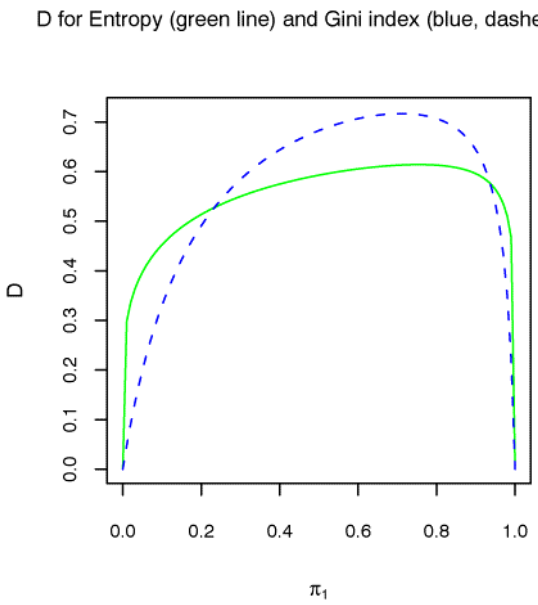


Figure 2.15: D_e and D_g for $s = 1$, $\mu = 3$

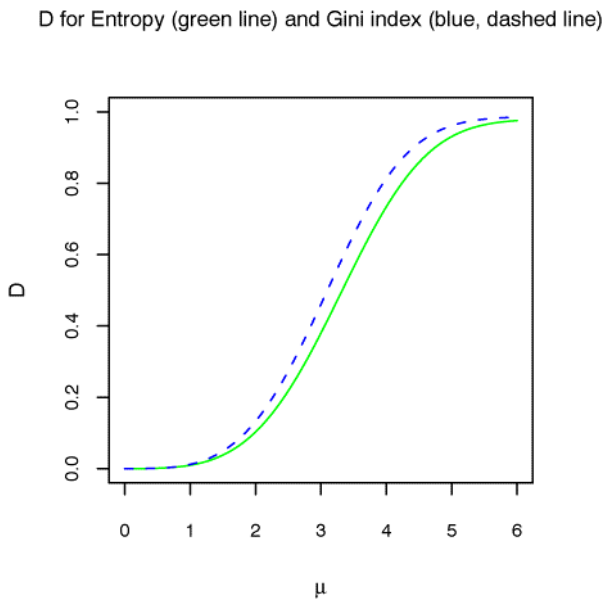


Figure 2.16: D_e and D_g for, $\pi_1 = 0.1$, $s = 3$

Simulation 2.28

We also simulated 300 samples of $n = 1000$ random values of a score with $P(Y = 1) = 0.05$, such that

$S \mid Y = 0 \sim N(0,1)$ and

$S \mid Y = 1 \sim N(\mu_k, 1)$, $\mu_k = 6k/300$, $k = 1, \ldots, 300$.

For $\mu_0 = 0$, there is a total overlapping of the densities of default and non-default, and the score is not discriminating at all. As μ_k increases, the overlapping area decreases; therefore the score has more discriminatory power. For $\mu_{300} = 6$ the score discriminates almost perfectly.

Figure 2.17 pictures the different split points obtained from using the entropy and the Gini index as splitting criteria for a given probability of default as the means vary. In the x-axis there are represented the mean estimates: $0 \leq \hat{\mu}_k \leq 6$, $k = 1, \ldots, 300$. For both criteria we have that the split points are higher as μ_k increases, which is reasonable. The results are listed in the table below.

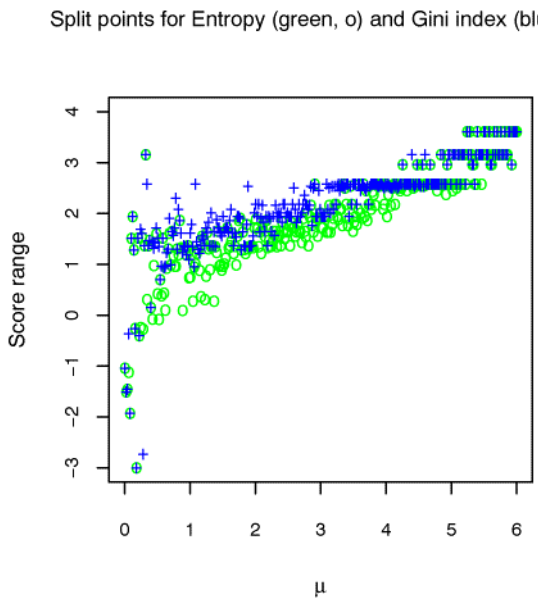


Figure 2.17: Split points for simulation in case 3, for $P(Y = 1) = 0.05$

μ_k	$s_q = s_p$	$s_q > s_p$	$s_p > s_q$
[0,2)	43	56	1
(2,4]	23	77	0
(4,6]	78	22	0
Total	144 (48%)	155 (51.6%)	1 (0.3%)

Table 2.5: Split points for simulation in case 3, for $P(Y = 1) = 0.05$

The table shows that almost half of the split points coincide and for the non-coincident, those obtained using the Gini index are higher than if we use the

entropy. This means that the entropy is a little more conservative than the Gini index, since we did the simulations for $\mu_k \geq 0$ and thus the distribution of $S \mid Y = 1$ lays on the right of (or overlaps) the distribution of $S \mid Y = 0$.

Figure 2.18 pictures the split points for $\mu = 3$ as the probability of default changes. The split points coincide for almost 75% of the scores and they are similarly spread over the score range. They are consequently higher for lower probabilities of default.

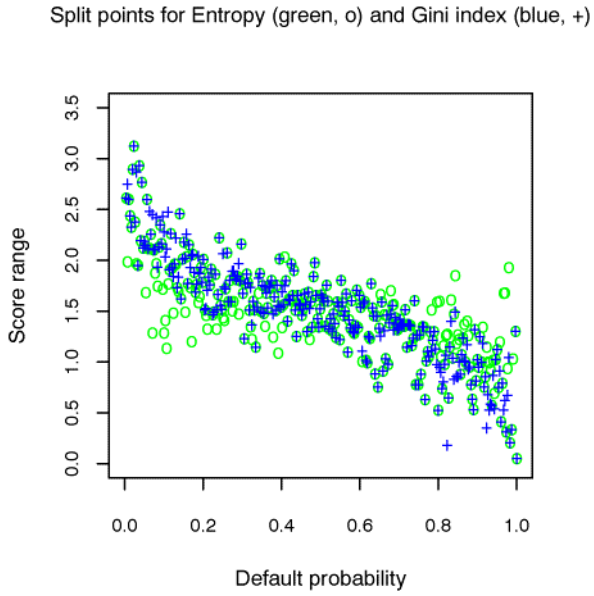


Figure 2.18: Split points for the simulation in case 3, for $\mu = 3$

Figure 2.19 shows the different estimates of D. We appreciate also here, that the difference between applying the Gini index or the entropy as impurity functions is not relevant. For both criteria we have that D increases as μ_k increases.

Finally, in Figure 2.20 are depicted the estimates of D for $\mu = 3$. The results obtained from applying both measures are similar, but the entropy gives slightly better results for extreme values of the probability of default, since D for the entropy does not vary as much as for the Gini index as the probability of default varies.

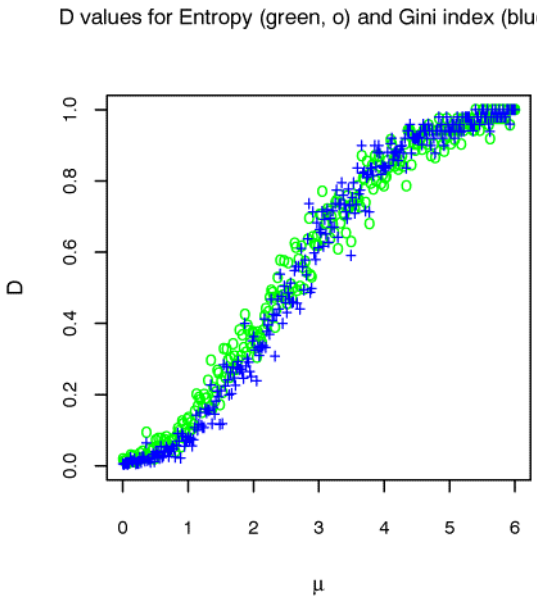


Figure 2.19: case 3, for $P(Y=1)=0.05$

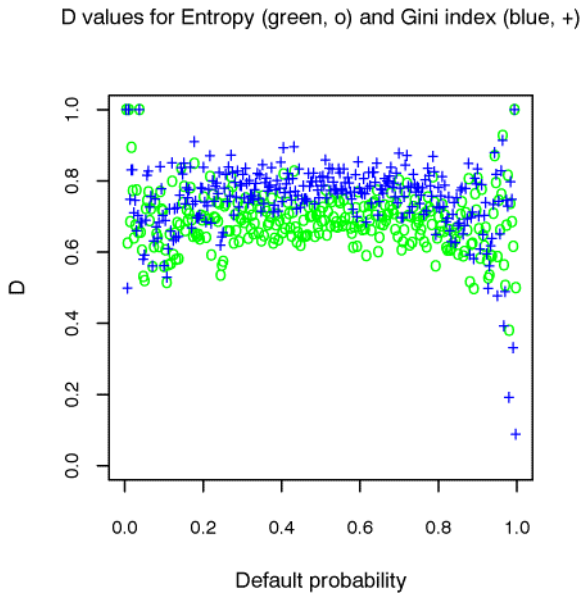


Figure 2.20: case 3, for $\mu=3$

Case 4:

We consider now the distributions conditioned to default and non-default both having the same means but different standard deviations, such that, as the difference between standard deviations increases, the overlapping area decreases and therefore the score has more discriminatory power.

Proposition 2.29

Assume a score with conditional distributions:

$S \mid Y=0 \sim N(0,1)$ and

$S \mid Y=1 \sim N(0,\sigma),\ 0<\sigma\leq 1,$

Then the standardized maximal distance function can be written:

1. For the entropy

$$D_e = 1 - \frac{\pi_1 \Phi\left(\frac{s_p}{\sigma}\right) \log\left(\frac{\pi_1 \Phi\left(\frac{s_p}{\sigma}\right)}{P\left(S \le s_p\right)}\right) + \left(1-\pi_1\right)\Phi\left(s_p\right)\log\left(\frac{\left(1-\pi_1\right)\Phi\left(s_p\right)}{P\left(S \le s_p\right)}\right)}{\pi_1 \log\left(\pi_1\right) + \left(1-\pi_1\right)\log\left(1-\pi_1\right)}$$

$$- \frac{\pi_1\left(1-\Phi\left(\frac{s_p}{\sigma}\right)\right)\log\left(\frac{\pi_1\left(1-\Phi\left(\frac{s_p}{\sigma}\right)\right)}{P\left(S > s_p\right)}\right) + \left(1-\pi_1\right)\left(1-\Phi\left(s_p\right)\right)\log\left(\frac{\left(1-\pi_1\right)\left(1-\Phi\left(s_p\right)\right)}{P\left(S > s_p\right)}\right)}{\pi_1 \log\left(\pi_1\right) + \left(1-\pi_1\right)\log\left(1-\pi_1\right)}$$

2. For the Gini index

$$D_g = 1 - \left(\frac{\Phi\left(\frac{s_q}{\sigma}\right)\cdot\Phi\left(s_q\right)}{\pi_1\Phi\left(\frac{s_q}{\sigma}\right) + \left(1-\pi_1\right)\Phi\left(s_q\right)} + \frac{\left(1-\Phi\left(\frac{s_q}{\sigma}\right)\right)\cdot\left(1-\Phi\left(s_q\right)\right)}{\pi_1\left(1-\Phi\left(\frac{s_q}{\sigma}\right)\right) + \left(1-\pi_1\right)\left(1-\Phi\left(s_q\right)\right)} \right).$$

Proof:

In this case, we get the following probabilities for $j = p, q$:

$$P\left(S \le s_j\right) = \pi_1\Phi\left(\frac{s_j}{\sigma}\right) + \left(1-\pi_1\right)\Phi\left(s_j\right),$$

$$P\left(S > s_j\right) = \pi_1\left(1-\Phi\left(\frac{s_j}{\sigma}\right)\right) + \left(1-\pi_1\right)\left(1-\Phi\left(s_j\right)\right),$$

$$P\left(Y=1 \mid S \le s_j\right) = \frac{\pi_1\Phi\left(\frac{s_j}{\sigma}\right)}{P\left(S \le s_j\right)},\ P\left(Y=0 \mid S \le s_j\right) = \frac{\left(1-\pi_1\right)\Phi\left(s_j\right)}{P\left(S \le s_j\right)},$$

$$P\left(Y=1 \mid S > s_j\right) = \frac{\pi_1\left(1-\Phi\left(\frac{s_j}{\sigma}\right)\right)}{1-P\left(S \le s_j\right)},\ P\left(Y=0 \mid S > s_j\right) = \frac{\left(1-\pi_1\right)\left(1-\Phi\left(s_j\right)\right)}{1-P\left(S \le s_j\right)}.$$

1. Then we get the following impurity functions, for the entropy

$$\mathrm{i}\left(S \le s_p\right) = \frac{\pi_1\Phi\left(\frac{s_p}{\sigma}\right)}{P\left(S \le s_p\right)}\log\left(\frac{\pi_1\Phi\left(\frac{s_p}{\sigma}\right)}{P\left(S \le s_p\right)}\right) + \frac{\left(1-\pi_1\right)\Phi\left(s_p\right)}{P\left(S \le s_p\right)}\log\left(\frac{\left(1-\pi_1\right)\Phi\left(s_p\right)}{P\left(S \le s_p\right)}\right)$$

$$\mathrm{i}\left(S > s_p\right) = \frac{\pi_1\left(1-\Phi\left(\frac{s_p}{\sigma}\right)\right)}{1-P\left(S \le s_p\right)} \log\left(\frac{\pi_1\left(1-\Phi\left(\frac{s_p}{\sigma}\right)\right)}{1-P\left(S \le s_p\right)}\right) + \frac{\left(1-\pi_1\right)\left(1-\Phi\left(s_p\right)\right)}{1-P\left(S \le s_p\right)} \log\left(\frac{\left(1-\pi_1\right)\left(1-\Phi\left(s_p\right)\right)}{1-P\left(S \le s_p\right)}\right)$$

2. And for the Gini index

$$\mathrm{i}\left(S \le s_q\right) = 2\frac{\pi_1\Phi\left(s_q\right)\left(1-\pi_1\right)\Phi\left(\frac{s_q}{\sigma}\right)}{P\left(S \le s_q\right)^2}, \; \mathrm{i}\left(S > s_q\right) = 2\frac{\pi_1\left(1-\Phi\left(s_q\right)\right)\left(1-\pi_1\right)\left(1-\Phi\left(\frac{s_q}{\sigma}\right)\right)}{\left(1-P\left(S \le s_q\right)\right)^2}.$$

By substituting these impurity functions in (2.4.10) and (2.4.8), respectively, we get the above defined D_e and D_g.

□

The first and second derivatives of D with respect to π_1 and with respect to σ for Gini index and entropy are again very complicated and in terms of the normal cumulative distribution function, which has no explict form. Hence, we can study and compare them only numerically.

Example 2.30

Look at Figure 2.21 for the same optimal split point $s_p = s_q = 0.5$ and $\sigma = 1/2$. We can see how the graphics D look like for different probabilities of default. Here again, D_e is higher than D_g for extreme values of the probability of default. In this case we can say that, using the entropy as impurity function, is a better way to measure the discriminatory power of the score.

We can also see (Figure 2.22) how, for a given probability of default $\pi_1 = 0.1$ and $s_p = s_q = 0.5$, D_e and D_g look like for different standard deviations σ. Here, there is a clear difference between the Gini index and the entropy. For the second one we get higher values of D as σ decreases, which means that D_e increases as the overlapping area decreases more rapidly than D_g and therefore using the entropy gives better results for measuring the discriminatory power.

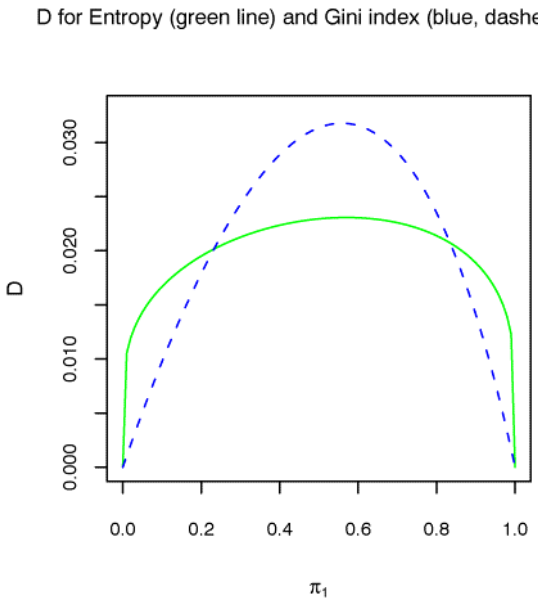


Figure 2.21: $s = 0.5$, $\sigma = 1/2$

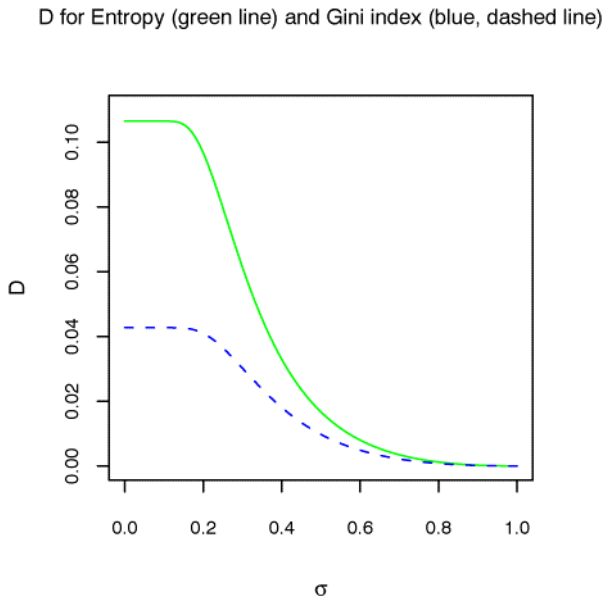


Figure 2.22: $\pi_1 = 0.1$ and $s_p = 0.5$

Simulation 2.31

Again, we simulate 300 samples of n= 1000 random values of a score with default probability $P(Y=1)$= 0.05, but now we consider both distributions of defaults and non-defaults having the same mean and different standard deviations, such that:

$S \mid Y = 0 \sim N(0,1)$ and

$S \mid Y = 1 \sim N(0,\sigma_k)$, $\sigma_k = k/300$, $k = 1,\ldots,300$.

As σ_k increases, the overlapping area increases; therefore the score has less discriminatory power. For σ_{300}= 1, there is a total overlapping of the densities of default and non-default.

In Figure 2.23 there are represented the split points obtained from using the entropy and the Gini Index. In the x-axis are the estimated standard deviations: $0 < \hat{\sigma}_k \leq 1$, $k = 1,\ldots,300$. The split points coincide for 60% of the scores. As σ_k increases, the discriminatory power of the score decreases and the split points deviate more from each other, and from the mean μ= 0.

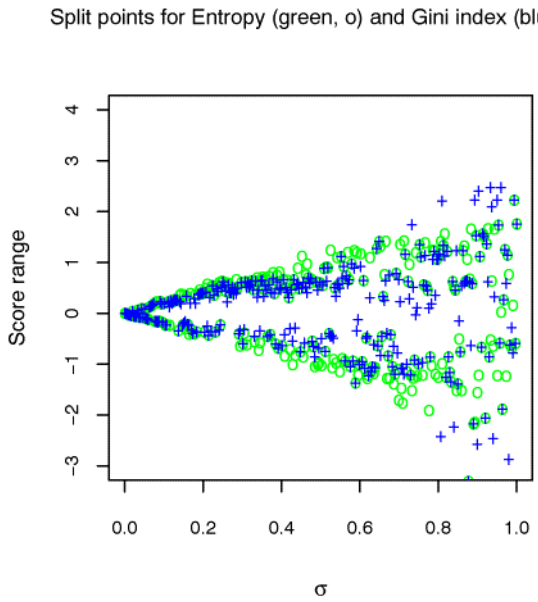


Figure 2.23: Split points for the simulation in case 4, $P(Y=1)=0.05$

Figure 2.24 pictures the split points for σ = 0.5 as the default probability varies. The split points coincide for 72% of the scores and they are similarly disseminated over the range of the score.

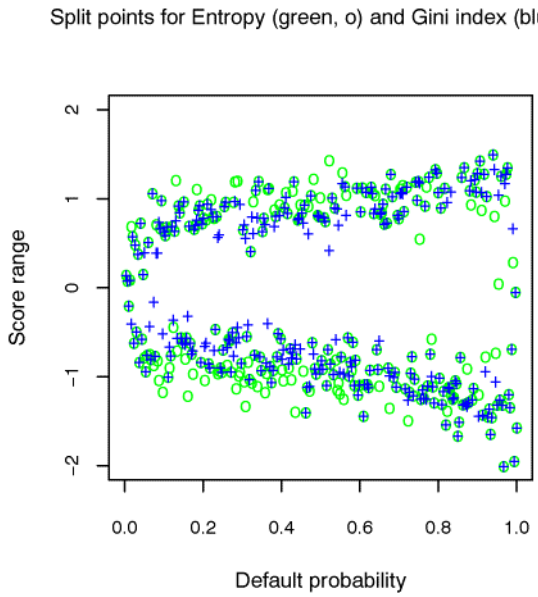


Figure 2.24: Split points for the simulation in case 4, $\sigma = 0.5$

Figure 2.25 pictures the different estimates of D. For both criteria we have that D decreases as σ_k increases, but we obtain higher values of D when we use the entropy as splitting criterion. Therefore, we can conclude here that using the entropy in D gives better results than using the Gini index, because it is more sensitive to the discriminatory power of the score.

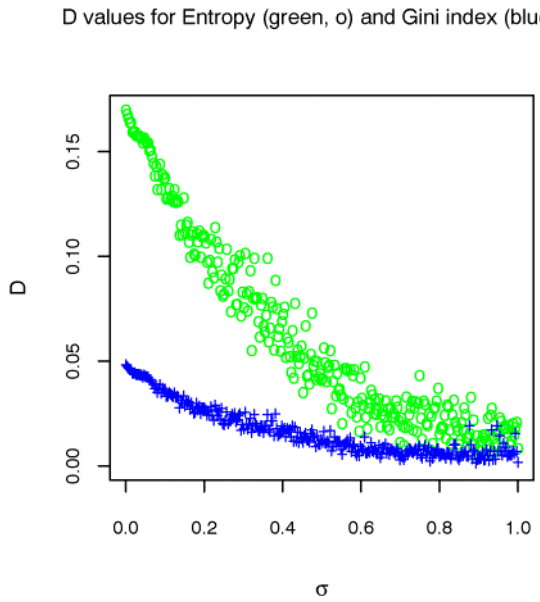


Figure 2.25: D values for the simulation in case 4, $P(Y=1)=0.05$

Figure 2.26 shows the estimates of D for $\sigma = 0.5$. The D values for the entropy are more stable with respect to the probability of default than for the Gini index and thus higher for lower probabilities of default. We can state again that using the entropy as impurity function in D gives better results than using the Gini index.

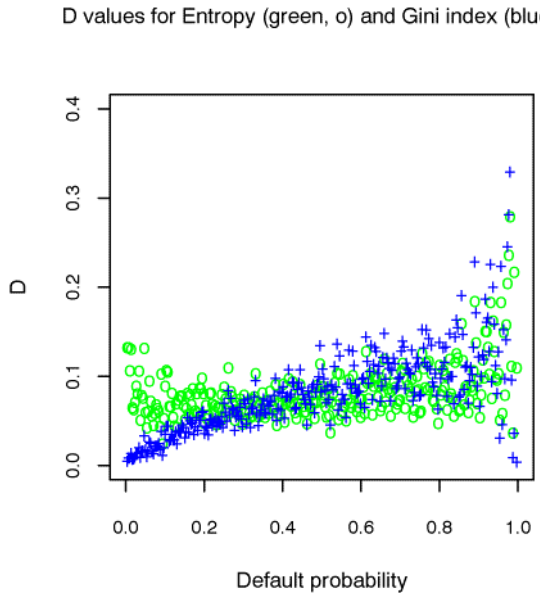


Figure 2.26: D values for the simulation in case 4, $\sigma = 0.5$

Conclusion

In the simulations and examples of this section, we have that the entropy compared to the Gini index gives better results for lower probabilities of default, which is the case in credit

scoring; it is often more stable with respect to the probability of default; and it is more sensitive to the discriminatory power of the score. Most of the optimal split points are coincident for both criteria. For the noncoincident, the split points obtained with entropy are frequently more conservative. Because of these reasons, and together with the fact that we can apply the test for homogeneity (see section 2.4.5) to the partition yielded by D_e, we recommend the use of the entropy as splitting criterion criterion in D.

2.5 Other measures

In this section we describe the disadvantages of the misclassification rate and the coefficient of correlation for measuring the discriminatory power of a score. Although these measures are sometimes used in practice, we do not find them appropriate for of credit rating, and we will explain the reasons.

2.5.1 Misclassification rate

If a score S and a separation threshold $s \in \mathbb{R}$ are given, such that all score values $S > s$ are predicted as defaults, then the discriminatory power of the score can be also estimated by the misclassification rate. Here the following scheme:

		$\widehat{Y}$		
		0	1	
Y	0		$\alpha_{01}(s)$	
	1	$\alpha_{10}(s)$		
				1

The misclassification rates are defined by

$$\alpha_{10}(s) = P(S \le s, Y = 1) = P(\widehat{Y} = 0, Y = 1) \quad (1.\ \text{type})$$
$$\alpha_{01}(s) = P(S > s, Y = 0) = P(\widehat{Y} = 1, Y = 0) \quad (2.\ \text{type})$$

such that, for a fixed sample size n, if $\alpha_{10}(s)$ increases then $\alpha_{01}(s)$ decreases, and vice versa. We can calculate them by

$$\alpha_{10}(s) = P(S \le s, Y = 1) = F_1(s)P(Y = 1)$$
$$\alpha_{01}(s) = P(S > s, Y = 0) = \{1 - F_0(s)\}\, P(Y = 0)$$

The aggregate misclassification rate

$$\alpha(s) = \alpha_{10}(s) + \alpha_{01}(s) = F_1(s)P(Y=1) + \{1 - F_0(s)\}P(Y=0)$$

presents a weighted version of the overlapping area and can be estimated by the use of $\widehat{F}_1(s)$, $\widehat{F}_0(s)$, n_1/n, n_0/n. An optimal misclassification rate can be defined as

$$\alpha = \min_s \alpha(s) \tag{2.5.1}$$

Also here it can again be distinguished between positive-monotone ($\widehat{Y}=1$ if $S > s$) and negative-monotone ($\widehat{Y}=1$ if $S \le s$) for the definition.

We already have seen in section 2.4.1, that the misclassification rate can be expressed as an impurity function of the default variable Y given a score S. Further, we spoke about its defects as splitting criterion. Also the aggregate misclassification rate defined in this section has some deficiencies and we will put some examples later. Now we show how, under some assumptions, both criteria D_{mr} and α are related.

Proposition 2.32

Assume $P(Y=1) < 1/2$ and $\exists\, s^*$, being the solution of

$$\max_s d_s(S \le s, S > s)$$
$$s.t. \quad P(Y=1 \mid S > s) > P(Y=0 \mid S > s)$$

Then we get that both criteria D_{mr} and α are linearly related:

$$D_{mr} = 1 - \alpha / P(Y=1). \tag{2.5.2}$$

Proof:

As we have assumed $P(Y=1) < 1/2$, then by the definition of impurity function for the misclassification rate (2.4.5), we get $\mathrm{i}(S) = \min\big(P(Y=1), 1 - P(Y=1)\big) = P(Y=1)$.

The impurity functions for $S_s = \{S \le s\}$ and $\overline{S}_s = \{S > s\}$ are given by

$$\mathrm{i}(S_s) = \min\big(P(1 \mid S_s), P(0 \mid S_s)\big) = \min\big(P(Y=1, S \le s)/P(S_s), P(Y=0, S_s)/P(S_s)\big)$$
$$= \min\big(\alpha_{10}(s)/P(S_s), P(Y=0, S \le s)/P(S_s)\big) = \frac{1}{P(S_s)} \min\big(\alpha_{10}(s), P(Y=0, S \le s)\big)$$

and

$$\mathrm{i}\left(\overline{S}_s\right) = \min\left(P\left(1 \mid \overline{S}_s\right), P\left(0 \mid \overline{S}_s\right)\right) = \min\left(P\left(Y=1, \overline{S}_s\right) / P\left(\overline{S}_s\right), P(Y=0, S>s) / P\left(\overline{S}_s\right)\right)$$
$$= \min\left(P(Y=1, S>s) / P\left(\overline{S}_s\right), \alpha_{01}(s) / P\left(\overline{S}_s\right)\right) = \frac{1}{P\left(\overline{S}_s\right)} \min\left(P(Y=1, S>s), \alpha_{01}(s)\right)$$

Then by the definition (2.4.2) we get:

$$d = \max_s d_s\left(S_s, \overline{S}_s\right) = \max_s \left(\mathrm{i}(S) - P\left(S_s\right)\mathrm{i}\left(S_s\right) - P\left(\overline{S}_s\right)\mathrm{i}\left(\overline{S}_s\right)\right)$$
$$= P(Y=1) + \max_s \left\{-\min\left(\alpha_{10}(s), P(Y=0, S \le s)\right) - \min\left(P(Y=1, S>s), \alpha_{01}(s)\right)\right\}$$
$$= P(Y=1) + \max_s \left\{-\left(\alpha_{10}(s) + \alpha_{01}(s)\right)\right\} = P(Y=1) - \min_s \alpha(s) = P(Y=1) - \alpha$$

And therefore, by substituting in (2.4.4) for the misclassification rate:

$$D_{mr} = \frac{d}{d_{opt}} = \frac{P(Y=1) - \alpha}{P(Y=1)},$$

we get the above mentioned linear relationship.

□

Remark 2.33

This proposition can be extended to the case where the scores are ordered from "bad to good", i.e. higher probabilities of default correspond to lower score values, and the misclassification rate is defined such that all score values $S \le s$ *are predicted as defaults. Then, under the restriction* $P\left(Y=1 \mid S \le s\right) > P\left(Y=0 \mid S \le s\right)$ *we get the same linear relationship (2.5.2). If otherwise, we had* $P(Y=1) > 1/2$*, then we would obtain similarly under other restrictions:* $D = 1 - \alpha / P(Y=0)$*.*

The aggregate misclassification rate is a sensible measure of the discriminatory power of a score, if the size of defaults is not too low, and we can see it in the following example.

Example 2.34

So, if we test it by *score2* for the validation sample from the Fahrmeir et al. (1984) dataset, where the proportion of defaults is 27%, we get the estimate $\hat{\alpha}$ = 0.225 and the threshold s = -0.335 (see Figure 2.27), which is equal to the threshold given by the measure D for every impurity function.

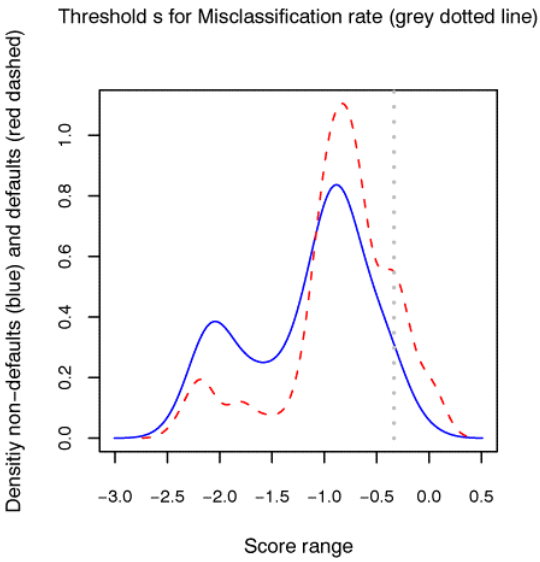


Figure 2.27: *score2*, $\widehat{\alpha} = 0.225$, $s = -0.335$

But this is generally not the case in credit scoring, where the proportion of defaults is clearly lower. Here, the optimal misclassification rate is normally achieved when all defaults are misclassified, which does not make this measure a reasonable criterion.

Example 2.35

`We picture below the estimated densities for a simulated sample with proportion of defaults 6%:`

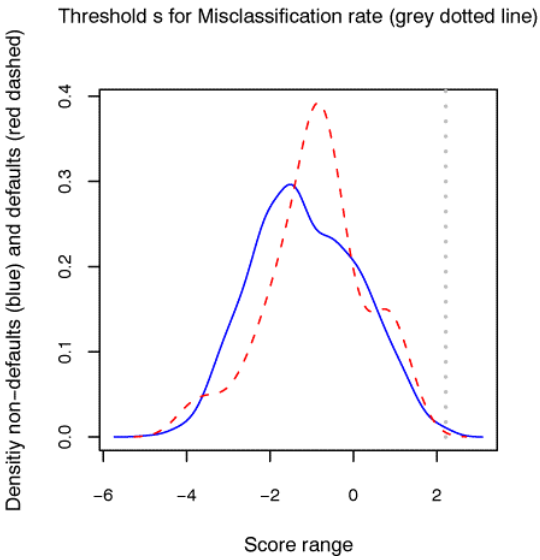


Figure 2.28: 6% defaults, $\widehat{\alpha} = 0.06$, $s = 2.215$

The estimate of the misclassification rate is $\hat{\alpha} = 0.06$, which corresponds exactly to the frequency of defaults in the dataset, since the threshold is at the highest value of the score, namely $s = 2.215$, and consequently all defaults are misclassified. We must remark also here that there are represented the kernel density estimates of defaults and non-defaults. This is the reason why the estimates of the densities do not lie completely on the left of the threshold.

2.5.2 Correlation coefficient

The coefficient of correlation is indeed a measure of association, but sometimes it is presented in the literature for sociologists (see for instance Bortz & Döring, 1995) as a measure of discriminatory power. The correlation between two random variables S and Y, denoting a score and a default variable respectively, with variances $\mathrm{var}(S)$ and $\mathrm{var}(Y)$ existing and positive, is defined as:

$$\rho_{S,Y} = \frac{\mathrm{cov}(S,Y)}{\sqrt{\mathrm{var}(S)} \cdot \sqrt{\mathrm{var}(Y)}}. \tag{2.5.3}$$

Remark 2.36

From the Chauchy-Schwarz inequality follows: $-1 \leq \rho_{S,Y} \leq 1$.

We will now introduce another expression for the coefficient of correlation.

Proposition 2.37

Let S, Y be two random variables, with $Y \sim Be(\pi_1)$, such that $0 < \pi_1 < 1$. Denote $E(S \mid Y = 0) = \mu_0$ and $E(S \mid Y = 1) = \mu_1$, such that $\mu_1 \geq \mu_0 \geq 0$ (w.l.o.g.), and σ_0, σ_1 denote the respective standard deviations of S conditioned to Y. The correlation coefficient is then given by

$$\rho_{S,Y} = \frac{\pi_1(1-\pi_1)(\mu_1 - \mu_0)}{\sqrt{\pi_1 {\sigma_1}^2 + (1-\pi_1){\sigma_0}^2 + \pi_1(1-\pi_1)(\mu_1 - \mu_0)^2} \cdot \sqrt{\pi_1(1-\pi_1)}} \tag{2.5.4}$$

Proof:

For the numerator, we have that:

$$\mathrm{cov}(S,Y) = \mathrm{cov}(Y, E(S \mid Y)) = E\big(Y \cdot E(S \mid Y)\big) - E(Y) \cdot E\big(E(S \mid Y)\big),$$

being $E\big(Y \cdot E\big(S \mid Y\big)\big) = 1 \cdot E\big(S \mid Y = 1\big) \cdot \pi_1 = \mu_1\pi_1$, $E(Y) = \pi_1$ and

$$E\big(E\big(S \mid Y\big)\big) = E\big(S \mid Y = 1\big) \cdot \pi_1 + E\big(S \mid Y = 0\big) \cdot (1 - \pi_1) = \mu_1\pi_1 + \mu_0(1 - \pi_1).$$

By substituting, then

$$\operatorname{cov}(S,Y) = \mu_1\pi_1 - \pi_1(\mu_1\pi_1 + \mu_0(1 - \pi_1)) = \mu_1\pi_1(1 - \pi_1) - \mu_0\pi_1(1 - \pi_1).$$

The denominator is given by multiplying $\sqrt{\operatorname{var}(S)} \cdot \sqrt{\operatorname{var}(Y)}$, such that:

$\operatorname{var}(Y) = \pi_1(1 - \pi_1)$ and

$\operatorname{var}(S) = E\big(S^2\big) - E(S)^2 = E\big(E\big(S^2 \mid Y\big)\big) - E\big(E\big(S \mid Y\big)\big)^2$, where

$$E\big(E\big(S^2 \mid Y\big)\big) = E\Big(\operatorname{var}\big(S \mid Y\big) + E\big(S \mid Y\big)^2\Big) = (\sigma_1^{\,2} + \mu_1^{\,2})\pi_1 + (\sigma_0^{\,2} + \mu_0^{\,2})(1 - \pi_1),$$

$$E\big(E\big(S \mid Y\big)\big)^2 = (\mu_1\pi_1 + \mu_0(1 - \pi_1))^2 = \mu_1^{\,2}\pi_1^{\,2} + \mu_0^{\,2}(1 - \pi_1)^2 + 2\mu_1\mu_0\pi_1(1 - \pi_1)$$

and therefore

$$\begin{aligned}\operatorname{var}\big(S\big) &= \pi_1\sigma_1^{\,2} + (1 - \pi_1)\sigma_0^{\,2} + \pi_1\mu_1^{\,2} + (1 - \pi_1)\mu_0^{\,2} - \pi_1^{\,2}\mu_1^{\,2} - (1 - \pi_1)^2\mu_0^{\,2} - 2\pi_1(1 - \pi_1)\mu_1\mu_0 \\ &= \pi_1\sigma_1^{\,2} + (1 - \pi_1)\sigma_0^{\,2} + \pi_1(1 - \pi_1)(\mu_1 - \mu_0)^2.\end{aligned}$$

□

Remark 2.38

Under the assumptions of Proposition 2.37, the coefficient of correlation reaches its maximal value, $\rho_{S,Y} = 1$, *if and only if* $\sigma_0 = \sigma_1 = 0$, *and it is minimal at* $\rho_{S,Y} = 0$, *for* $\mu_1 = \mu_0$.

A reasonable selectivity measure would assign its maximal value to a perfect separation between the distributions of the score conditioned to default and non-default, and its minimal value for a total overlapping of the distributions. However, these requirements are not fulfilled by the coefficient of correlation.

In practice we will have scores taking more than two values; this implies that the conditional standard deviations will be different from 0. In this case the coefficient of correlation is not a good measure of the discriminatory power of a score, because it does not reach its maximum.

Example 2.39

Suppose we have two perfectly separated uniform distributions $S \mid Y = 0 \sim U(-1,0)$ and $S \mid Y = 1 \sim U(0,1)$, with $P(Y = 1) = 1/2$ (see Figure 2.29), and $\sigma_0 = \sigma_1 = 1/12$. Then $\rho_{S,Y} = \sqrt{3}/2 < 1$.

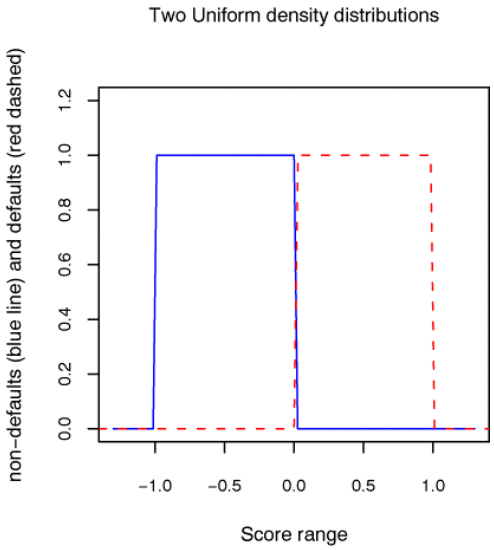


Figure 2.29: Perfectly separated densities, $\rho_{S,Y} = \sqrt{3}/2 < 1$.

On the other hand, if we have two distributions with equal means, for example $S \mid Y = 0 \sim N(0, 3/2)$ and $S \mid Y = 1 \sim N(0, 1/2)$ (see Figure 2.30), then we obtain $\rho_{S,Y} = 0$, although there is not a total overlapping of the areas for defaults and non-defaults.

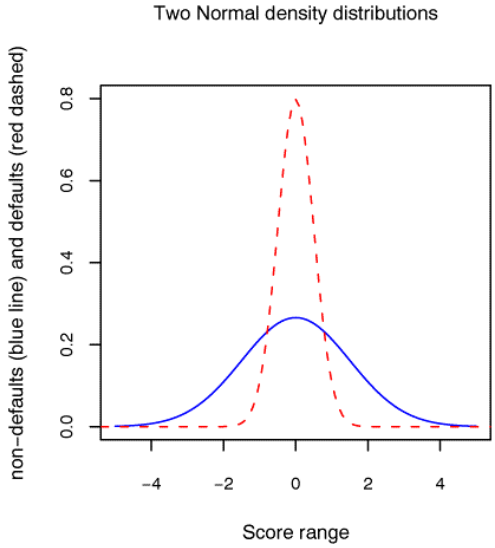


Figure 2.30: $\mu_1 = \mu_0 = 0$, $\rho_{S,Y} = 0$

2.6 Comparison of measures of discriminatory power

Along Section 2, we analyzed different measures of discriminatory power in the context of credit rating. Here, we devote us to the comparison between the overlapping area criterion T, the accuracy ratio AR, and the standardized maximal distance D with the entropy as impurity function. The other measures are discarded because of their inadequacy, or in the case of the Gini index, because the entropy gives better results. A further advantage of the entropy compared to the Gini index is that hypotheses can be tested (see section 2.4.7).

These three measures cannot be compared in their absolute values. T gives the maximum distance between the ROC curve and the diagonal, whereas the accuracy ratio is related to the average of the difference between ROC and diagonal. However, in spite of their differences, these measures lead many times to the same conclusions, as we saw in sections 2.2, 2.3 and 2.4. There we calculated and tested T, AR and D_e for the sample scores (2.1.2) and (2.1.3). For all of them, *score1* was preferred to *score2*, and we were able to reject in any case the null hypothesis for their respective tests. However, H_0 would be more difficult to reject if we had lower default rates for the Kolmogorov-Smirnov test. Further, T and D_e lifted close optimal thresholds for the first score. We could also see that AR depends on the probability of default. So, for various scores having the same Lorenz curve and Gini coefficient but different PDs, AR would raise different values.

This section includes five cases for the purpose of illustrating the behaviour of T, AR and D_e. Since an overall theoretical comparison is not possible, we will have to make some assumptions of the distribution of the score conditioned to the default variable, and specify the values of their parameters. In the first two cases, we will point out extreme circumstances where T, AR and D_e attain their maxima and minima. The third and fourth cases present situations where the use of the accuracy ratio should be avoided. On the other hand, the three measures perform properly in the last case.

Case 1:

The score lacks totally of discriminatory power if it is independent of the default variable. A suitable measure should therefore lift zero under these circumstances.

Proposition 2.40

For S and Y independent random variables, it holds:

$$T = AR = D_e = 0\,.$$

Proof:

As for independence we have $F_j(s) = F(s)$, for $j = 0,1$, then is easy to see:

1. By definition (2.2.5), $T_{pos} = \max_s \{F_0(s) - F_1(s)\} = \max_s \{F(s) - F(s)\} = T_{neg} = 0\,.$
2. Applying (2.3.3) and (2.3.4), we get:

$$AUC = 1 - \int_{-\infty}^{\infty} F_1(s)\,dF_0(s) = 1 - \left(\frac{F(\infty)^2}{2} - 0\right) = \frac{1}{2} \text{ and } AR = 2AUC - 1 = 0\,.$$

3. We already saw in Proposition 2.21, that under the assumption of independence, $D_e = 0\,.$

□

Case 2:

It is also required to study the case of a score that separates perfectly defaults from non-defaults under monotonicity. In this case, an appropriate measure of discriminatory power should raise one.

Proposition 2.41

Given two random variables S and $Y \sim Be(\pi_1)$, and a threshold $s^* \in \mathbb{R}$, such that $P(Y = 0 \mid S \le s^*) = 1$ and $P(Y = 1 \mid S > s^*) = 1$, it holds:

$$T = AR = D_e = 1.$$

Proof:

Applying Bayes' theorem and the Law of total probability, it follows that $F_0(s^*) = 1$ and $F_1(s^*) = 0\,.$

1. By the definition (2.2.5), $T = T_{pos} = \max_s \{F_0(s) - F_1(s)\} = F_0(s^*) - F_1(s^*) = 1\,.$

2. The Lorenz curve given by a perfectly discriminating score is optimal and therefore the Gini coefficient is also optimal (see Proposition 2.3) and $AR = G / G_{opt} = 1$.

3. We obtain the following impurity functions: $\mathrm{i}(S \leq s^*) = \mathrm{i}(S > s^*) = 0$. According to Proposition 2.13 we have $d = d_{opt} = \mathrm{i}(S)$, and by (2.4.4), $D_e = d / d_{opt} = 1$.

□

Remark 2.42

If the defaults were classified to the left of the threshold s^*, *i.e.* $P(0 \mid S \leq s^*) = 0$ *and,* $P(1 \mid S \leq s^*) = 1$, *then we would have* $T = T_{neg} = AR = D_e = 1$.

Case 3:

Now we will contemplate the case where the distributions of default and non-default are perfectly separated but there is not a monotone relationship between the score and the default variable. Under these circumstances, the measures T, AR and D_e are not optimal. Further, if we assume the conditions of Proposition 2.6, we get an even worse result for the accuracy ratio, which lifts zero. In this situation we will always prefer T and D_e, as we can see in the following example.

Example 2.43

In Figure 2.31 are represented the conditional densities f_0 and f_1 of a score, having the same expectation $\mu = 0$. They are even functions, i.e. symmetric with respect to the ordinate axis. The optimal threshold s_1 coincides for both T_{pos} and D_e (or s_2, if we consider T_{neg}). For a simulated sample with $n = 1000$ and $n_1 = 100$, we obtain $T_{pos} = 0.5$ and $D_e = 0.197$, and the split point s= -0.498 (or s= 0.498). In this case T will be preferred to D_e, since it is more sensitive to the discriminatory power of the score. Graphically, is also easy to see that $AR = 0$, since the area between the Lorenz curve (Figure 2.32) and the diagonal will sum up zero.

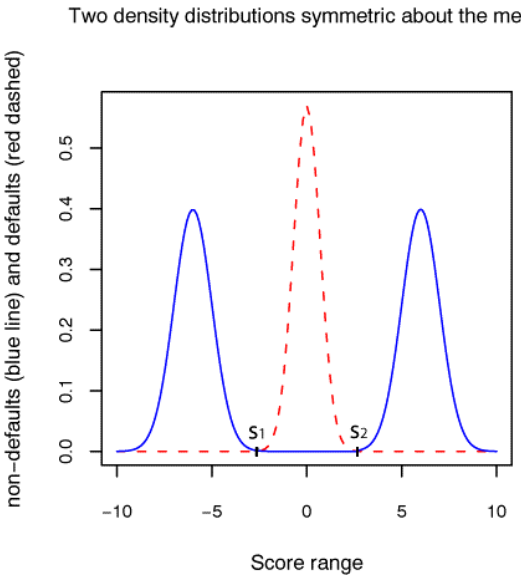


Figure 2.31: Symmetric densities w.r.t. $\mu = 0$

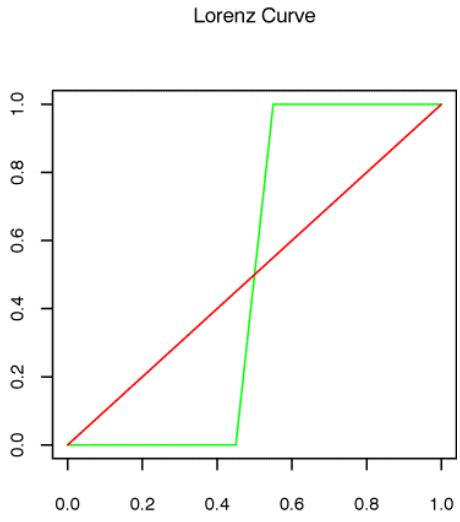


Figure 2.32: Lorenz curve, 10% defaults

Case 4:

We would like to see how these measures behave if the score conditioned to the default variable is normally distributed with different means and standard deviations for default and non-default. The expressions for T, AR and D_e are given in Appendix C, Proposition C.1. However, it is not possible to compare these expressions theoretically.

Thus, we will have to set some values for the parameters of the distribution functions. First, we will study the case of both distributions having the same mean and different standard deviations, i.e. $S \mid Y = 0 \sim N(0,1)$ and $S \mid Y = 1 \sim N(0,\sigma^2)$, $(0 < \sigma \leq 1)$.

As σ increases, the overlapping area also increases, both distributions being totally overlapped at $\sigma = 1$. The score is less discriminating as the standard deviation increases. In the same way, an adequate measure of discriminatory power should be decreasing with respect to σ. Let us see now how the measures behave:

1. We have that T_{pos} is strictly convex and decreasing in $\sigma \in (0,1)$, since:

$$\frac{dT_{pos}}{d\sigma} = -\frac{\sigma^{-\frac{\sigma^2}{-1+\sigma^2}}\sqrt{\log\sigma}}{\sqrt{\pi}\sqrt{-1+\sigma^2}} < 0 \quad \forall\ 0 < \sigma < 1,$$

$$\frac{d^2T_{pos}}{\left(d\sigma\right)^2}=\frac{\sigma^{-\frac{1-2\sigma^2}{-1+\sigma^2}}\left(-\left(-1+\sigma^2\right)^2+4\sigma^2\left(-1+\sigma^2-\log\sigma\right)\log\sigma\right)}{2\sqrt{\pi}\left(-1+\sigma^2\right)^{5/2}\sqrt{\log\sigma}}>0 \quad \forall\ 0<\sigma<1.$$

And by definition, $T_{pos}=0$ for $\sigma=1$. T_{pos} is hence an adequate discriminatory power measure.

2. The accuracy ratio is disappointing in this case, since we always have $AR=0$ (see Proposition 2.6), without concerning the size of the overlapping area.

3. As we already saw in case 4 of section 2.4.7, the first and second derivatives of D_e with respect to the probability of default π_1 and σ are very complicated and in terms of the normal cumulative distribution function, which does not have an explicit form. There is also no explicit form for the optimal split point s. But we know from the simulations in case 4 of section 2.4.7, that D_e decreases as the standard deviation decreases (see Figure 2.25), and it does not decrease apparently for lower probabilities of default (see Figure 2.26). Further, we saw that the optimal split points deviate more from the mean as σ increases (see Figure 2.23). These reasons depict D_e as a favourable measure of discriminatory power.

To summarize, the accuracy ratio in this case is not sensitive to the discriminatory power of the score, contrary to T_{pos} and D_e. However, we cannot assert that T_{pos} will always be better than D_e or vice versa. The comparison between them can only be accomplished numerically, and it is illustrated in the following example.

Example 2.44

Let $S\mid Y=0\sim N(0,1)$ and $S\mid Y=1\sim N\left(0,\sigma^2\right)$, $(0<\sigma\leq 1)$. For a given probability of default π_1= 0.05 see Figure 1.1. According to this picture, we can say that for both measures the score is less discriminating for higher standard deviations. T_{pos} raises higher values and is more sensitive to the discriminatory power of the score than D_e. In this case, we would prefer T_{pos}.

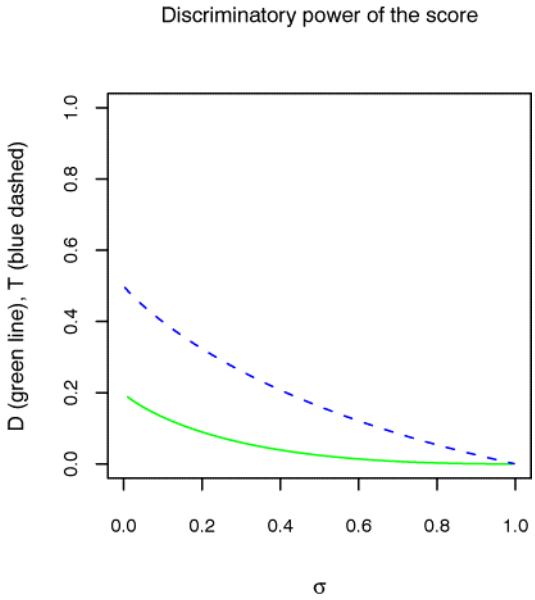


Figure 2.33: T_{pos} and D_e for $\pi_1 = 0.05$

Case 5:

We want to see now how T, AR and D_e behave for equal standard deviations. We already saw in Section 2.2, that the point of intersection for equal variances is: $s = (\mu_0 + \mu_1)/2$. Here again, it is not possible to do a theoretical comparison of the expressions given in Appendix C, Proposition C.2, with respect to μ. Therefore, we will study the case of having $\sigma = 1$, i.e. $S \mid Y = 0 \sim N(0,1)$ and $S \mid Y = 1 \sim N(\mu,1)$, with $\mu \geq 0$, such that, as μ increases the overlapping area decreases and thus the discriminatory power of the score increases.

1. We have that T_{pos} is strictly increasing and concave for $\mu > 0$, since:

$$\frac{dT_{pos}}{d\mu} = \frac{e^{-\frac{\mu^2}{8}}}{\sqrt{2\pi}} < 0 \quad \forall\ \mu,$$

$$\frac{d^2T_{pos}}{(d\mu)^2} = -\frac{e^{-\frac{\mu^2}{8}}\mu}{4\sqrt{2\pi}} < 0 \quad \forall\ \mu > 0.$$

   And by definition, $T_{pos} = 0$ for $\mu = 1$.

2. There are no explicit forms for the first and second derivatives of the accuracy ratio. But, if we calculate AR numerically, we can observe that it is increasing and concave in μ.

3. We already studied D_e in case 3 of section 2.4.7. We can observe that, for a low probability of default, the optimal split points are increasing for increasing expectation (see Figure 2.17). The values of D_e are also increasing in μ (see Figure 2.19) and they do not vary much as π_1 varies.

In this case the three measures seem to be appropriate. The following example pictures a numerical comparison.

Example 2.45

If we have $S \mid Y = 0 \sim N(0,1)$, $S \mid Y = 1 \sim N(\mu,1)$, ($\mu \geq 0$) and $\pi_1 = 0.05$, we can see in the following picture that T_{pos}, AR and D_e increase as μ increases. D_e is more conservative (or less sensitive) with respect to the discriminatory power than the other measures, since it increases slower. Therefore, our order of preference would be in this case AR, T_{pos}, and D_e.

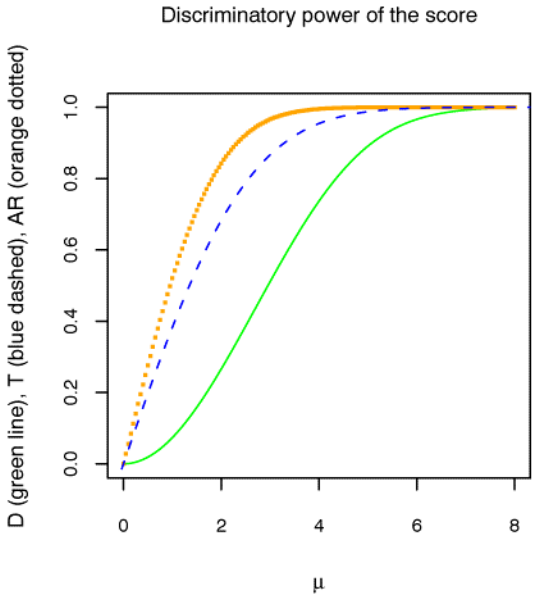


Figure 2.34: T_{pos}, AR and D_e for $\pi_1 = 0.05$

3 Estimation of default probabilities

3.1 Introduction

Estimating default probabilities for individual obligors is the first step when assessing the credit exposure and potential losses faced by an investor or financial institution. The PDs can be fixed a priori, and then every loan must be adequately assigned to a rating class. They can be determined by default rates from former years or they will be calculated from individual probabilities of default, that are determined by statistical scoring systems, by systems that aggregate experts' knowledge or by the combination of both. However, this estimation could be challenging due to limitations on data availability.

Our aim here is to offer an overview of several techniques existing for estimating default probabilities. In section 3.2, we will introduce binary choice models, including the well-known logit and probit, the calibration of the models and parameters and their significance. Further estimation methods and aspects of the calibration of PDs are integrated in the next sections—3.3 and 3.4. There, in Example 3.2, were generated for *score1* (2.1.2) the rating classes that will be used later for the validation tests presented in section 4.2. The last section summarizes the different points of view regarding the estimation of the probability of default.

3.2 Binary choice models

We are interested in knowing how credit worthiness depends on observable individual characteristics, like duration and amount of the credit, savings, purpose of the loan, etc. Binary choice models are regression models intended to estimate the functional relation between the binary variable Y (default indicator), and a vector of explanatory variables $X = (X_0, ..., X_{p-1})^\top \in \mathbb{R}^p$.

Suppose that we know the true score, given by the latent variable

$$Y^* = \beta^\top X + \varepsilon\,, \tag{3.2.1}$$

β being the parameter vector that assigns a weight β_j to the jth explanatory variable and ε an error term. We observe a default if the score Y^* is positive:

$$Y = \begin{cases} 1 & if\ Y^* > 0, \\ 0 & if\ Y^* \leq 0. \end{cases}$$

The regression function is given by the expectation of the response variable Y conditioned to the vector of independent variables X (for a constant $X_0 = 1$):

$$E(Y \mid X) = P(Y = 1 \mid X) = \Psi\left(\beta_0 + X_1\beta_1 + \ldots + X_{p-1}\beta_{p-1}\right) = \Psi\left(\beta^{\mathsf{T}} X\right). \quad (3.2.2)$$

The function Ψ is chosen as a cumulative distribution function. The normal and the logistic distribution functions, giving rise to the probit and logit models, respectively, are most commonly used.

3.2.1 MLE (Maximum Likelihood Estimator)

The method used to estimate the vector of parameters β is the maximum likelihood. In order to apply this method, we will suppose that the sample of n observed independent realizations follows a known distribution. The probability of occurrence of a realization x_i is $\psi(x_i, \beta)$, for $i = 1, \ldots, n$. The joint distribution can be calculated here as the product of the separate probabilities:

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n) = \psi(x_1, \ldots, x_n, \beta) = \prod_{i=1}^{n} \psi(x_i, \beta) = L(\beta \mid \mathbf{X}). \quad (3.2.3)$$

The likelihood function is denoted by $L(\beta \mid \mathbf{X})$. $\mathbf{X}$ is the matrix given by the rows x_i^{T}, of dimension $n \times p$ and is called the regression, or design matrix. For ease of calculations, we determine the logarithm of the likelihood function:

$$\log L(\beta \mid \mathbf{X}) = \sum_{i=1}^{n} \log \psi(x_i, \beta). \quad (3.2.4)$$

This representation is called the log-likelihood function. Since in this case the response variable takes only two states, we can rewrite L and $\log L$:

$$L(\beta \mid \mathbf{X}) = \prod_{\{i:y_i=0\}} \left(1 - \Psi\left(\beta^{\mathsf{T}} x_i\right)\right) \prod_{\{i:y_i=1\}} \Psi\left(\beta^{\mathsf{T}} x_i\right) = \prod_{i=1}^{n} \left(\Psi\left(\beta^{\mathsf{T}} x_i\right)\right)^{y_i} \left(1 - \Psi\left(\beta^{\mathsf{T}} x_i\right)\right)^{1-y_i} \quad (3.2.5)$$

which is a member of the exponential family of distributions (see Dobson, 1990). And thus:

$$\log L(\beta \mid \mathbf{X}) = \sum_{i=1}^{n} \left(y_i \log \Psi\left(\beta^{\mathsf{T}} x_i\right) + (1 - y_i) \log\left(1 - \Psi\left(\beta^{\mathsf{T}} x_i\right)\right)\right). \quad (3.2.6)$$

The vector of parameter estimates $\hat{\beta}$ that maximizes the likelihood function also does for the log-likelihood function, since the logarithmic function is monotonic. Having an unrestricted parameter space, and the likelihood function belonging to the exponential family, we can obtain the maximum-likelihood estimator (MLE) of β uniquely by solving the equations:

$$\nabla(\beta) = \frac{\partial \log L(\beta)}{\partial \beta} = 0, \tag{3.2.7}$$

as can be seen in Cox & Hinkley (1974).

For the special case in (3.2.6) the first derivative can be calculated:

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^{n} \left(\frac{y_i}{\Psi\left(\beta^{\mathsf{T}} x_i\right)} \frac{d\Psi\left(\beta^{\mathsf{T}} x_i\right)}{d\beta^{\mathsf{T}} x_i} - (1 - y_i) \frac{1}{1 - \Psi\left(\beta^{\mathsf{T}} x_i\right)} \frac{d\Psi\left(\beta^{\mathsf{T}} x_i\right)}{d\beta^{\mathsf{T}} x_i} \right) x_i .$$

In general (3.2.7) is a nonlinear system of equations. Hence, an iterative solution has to be computed. We can use the Newton-Raphson algorithm, which determines the optimal $\hat{\beta}$ with the following iteration steps:

$$\hat{\beta}^{new} = \hat{\beta}^{old} - \mathbf{H}\left(\hat{\beta}^{old}\right)^{-1} \nabla\left(\hat{\beta}^{old}\right),$$

being $\mathbf{H}(\beta) = \partial^2 \log L(\beta) / \partial\beta\partial\beta^{\mathsf{T}}$ the Hessian matrix. A variant of this method is the Fisher scoring algorithm, which replaces the Hessian by its expectation (see Tutz, 2000).

Under relatively general conditions, the maximum-likelihood estimator has the following interesting properties (see Theil, 1979):

- Consistency, i.e. $\hat{\beta}$ converges in probability to the real β.
- It is asymptotically normally distributed, i.e. $\hat{\beta} \overset{(a)}{\sim} N\left(\beta, \left\{\mathbf{I}(\beta)\right\}^{-1}\right)$.
- The MLE is asymptotically efficient.

The variance of the above mentioned normal distribution is the inverse of the Fisher information matrix, i.e.

$$\left\{\mathbf{I}(\beta)\right\}^{-1} = \left\{ -E\left[\frac{\partial^2 \log L(\beta)}{\partial\beta\partial\beta^{\mathsf{T}}} \right] \right\}^{-1} \tag{3.2.8}$$

In order to estimate the variance of the MLE, in (3.2.8) the parameter β can be substituted by $\widehat{\beta}$. However, this procedure is not allways feasible, since the expectation of the second derivative of the log-likelihood function is very difficult to calculate, except for the logit and probit models (see Greene, 1993). Therefore, other estimators were proposed. One of them is given by:

$$\left\{\hat{\mathbf{I}}\left(\widehat{\beta}\right)\right\}^{-1} = \left\{-\frac{\partial^2 \log L\left(\widehat{\beta}\right)}{\partial\hat{\beta}\partial\hat{\beta}^{\mathsf{T}}}\right\}^{-1} \tag{3.2.9}$$

For this estimator we do not have to determine the expectation, but still the second derivatives. Another estimator that requires the calculation of the first partial derivatives only is called OPG (outer product of gradients) or BHHH-Estimator (Berndt, Hall, Hall and Hausman):

$$\left\{\hat{\mathbf{I}}\left(\widehat{\beta}\right)\right\}^{-1} = \left\{\sum_{i=1}^{n}\hat{\mathbf{g}}_i\hat{\mathbf{g}}_i^{\mathsf{T}}\right\}^{-1} = \left\{\widehat{\mathbf{G}}^{\mathsf{T}}\widehat{\mathbf{G}}\right\}^{-1}, \tag{3.2.10}$$

being

$$\hat{\mathbf{g}}_i = \frac{\partial \log \psi\left(x_i, \widehat{\beta}\right)}{\partial\hat{\beta}}$$

and $\widehat{\mathbf{G}} = \left(\hat{\mathbf{g}}_1, \hat{\mathbf{g}}_2, \ldots, \hat{\mathbf{g}}_n\right)^{\mathsf{T}}$. We recommend this last estimator, since it avoids difficult computations (see Gourieroux & Monfort, 1995).

3.2.2 Logit

The probability that an event occurs is for the logit model:

$$P\left(Y = 1 \mid X = x\right) = \frac{\exp\left(\beta^{\mathsf{T}}x\right)}{1 + \exp\left(\beta^{\mathsf{T}}x\right)} = \Lambda\left(\beta^{\mathsf{T}}x\right), \tag{3.2.11}$$

which is strictly monotone increasing in $\beta^{\mathsf{T}}x$. So, we will ensure that the probability of default is a monotone function of the score $S = \beta^{\mathsf{T}}X$. The first derivative of the sample log-likelihood function for the logit model is:

$$\frac{\partial \log L}{\partial\beta} = \sum_{i=1}^{n}\left(y_i - \Lambda\left(\beta^{\mathsf{T}}x\right)_i\right)x_i$$

The $(k \times k)$ matrix of second derivatives for the log-likelihood function:

$$\frac{\partial^2 \log L}{\partial \beta \partial \beta^{\mathsf{T}}} = -\sum_{i=1}^{n} \Lambda\left(\beta^{\mathsf{T}} x\right)_i \left(1 - \Lambda\left(\beta^{\mathsf{T}} x\right)_i\right) x_i x_i^{\mathsf{T}}$$

3.2.3 Probit

The probability that an event occurs is for the probit model:

$$P\left(Y = 1 \mid X = x\right) = \int_{-\infty}^{\beta^{\mathsf{T}} x} \phi(t)\, dt = \Phi\left(\beta^{\mathsf{T}} x\right), \tag{3.2.12}$$

such that the default probability is modelled as a strictly monotone increasing function of the score. The first derivative of the sample log-likelihood function for the probit model is:

$$\frac{\partial \log L}{\partial \beta} = \sum_{i:y_i=0} \frac{-\phi\left(\beta^{\mathsf{T}} x_i\right)}{1 - \Phi\left(\beta^{\mathsf{T}} x_i\right)} x_i + \sum_{i:y_i=1} \frac{\phi\left(\beta^{\mathsf{T}} x_i\right)}{\Phi\left(\beta^{\mathsf{T}} x_i\right)} x_i = \sum_{i=1}^{n} \left(\frac{\left(2y_i - 1\right)\phi\left(\left(2y_i - 1\right)\beta^{\mathsf{T}} x_i\right)}{\Phi\left(\left(2y_i - 1\right)\beta^{\mathsf{T}} x_i\right)} \right) x_i$$

And the $(k \times k)$ matrix of second derivatives:

$$\frac{\partial^2 \log L}{\partial \beta \partial \beta^{\mathsf{T}}} = \sum_{i=1}^{n} -\lambda_i \left(\lambda_i + \beta^{\mathsf{T}} x_i\right) x_i x_i^{\mathsf{T}},$$

denoting

$$\lambda_i = \frac{\left(2y_i - 1\right)\phi\left(\left(2y_i - 1\right)\beta^{\mathsf{T}} x_i\right)}{\Phi\left(\left(2y_i - 1\right)\beta^{\mathsf{T}} x_i\right)}.$$

3.2.4 PD estimation

In order to estimate the probability of default, we must first calculate the asymptotical variance of the MLE. For this purpose we can use the BHHH-estimator defined in (3.2.10), being for the logit model $\mathbf{g}_i = y_i - \Lambda\left(\beta^{\mathsf{T}} x\right)_i$ and for the probit $\mathbf{g}_i = \lambda_i$.

Having estimated the vector of coefficients of the regression model and its variance-covariance matrix, we can estimate the probability of default and its variance. For both logit and probit models we have:

$$\widehat{P}\left(Y = 1 \mid x\right) = \widehat{\Psi\left(\beta^{\mathsf{T}} x\right)} = \Psi\left(\hat{\beta}^{\mathsf{T}} x\right) \tag{3.2.13}$$

The asymptotical variance of this forecast is given by:

$$Var^{asy}\left(\Psi\left(\hat{\beta}^{\mathsf{T}}x\right)\right)=\left(\frac{\partial\Psi\left(\hat{\beta}^{\mathsf{T}}x\right)}{\partial\hat{\beta}}\right)^{\mathsf{T}}\cdot\left\{\hat{\mathbf{I}}\left(\hat{\beta}\right)\right\}^{-1}\cdot\left(\frac{\partial\Psi\left(\hat{\beta}^{\mathsf{T}}x\right)}{\partial\hat{\beta}}\right) \tag{3.2.14}$$

Together with

$$\frac{\partial\Psi\left(\hat{\beta}^{\mathsf{T}}x\right)}{\partial\hat{\beta}}=\frac{d\Psi\left(\hat{\beta}^{\mathsf{T}}x\right)}{d\left(\hat{\beta}^{\mathsf{T}}x\right)}\frac{\partial\left(\hat{\beta}^{\mathsf{T}}x\right)}{\partial\hat{\beta}}=\psi\left(\hat{\beta}^{\mathsf{T}}x\right)\cdot x\,, \tag{3.2.15}$$

ψ denoting the density distribution function. Equality (3.2.14) can be expressed like:

$$Var^{asy}\left(\Psi\left(\hat{\beta}^{\mathsf{T}}x\right)\right)=\left(\psi\left(\hat{\beta}^{\mathsf{T}}x\right)\right)^{2}\cdot x^{\mathsf{T}}\cdot\left\{\hat{\mathbf{I}}\left(\hat{\beta}\right)\right\}^{-1}\cdot x \tag{3.2.16}$$

3.2.5 Significance of the model and parameters, optimal weighting

The simplest method to test the significance of a parameter, i.e. $H_0:\beta_j=0$, is to use the asymptotical normal distribution of the MLE. For more involved restrictions, of the type $H_0:\mathbf{R}\beta=\mathbf{q}$ we can use the Wald test:

$$W=\left(\mathbf{R}\hat{\beta}-\mathbf{q}\right)^{\mathsf{T}}\left\{\mathbf{R}\cdot\left\{\hat{\mathbf{I}}\left(\hat{\beta}\right)\right\}^{-1}\cdot\mathbf{R}^{\mathsf{T}}\right\}^{-1}\left(\mathbf{R}\hat{\beta}-\mathbf{q}\right),$$

being $W\overset{(a)}{\sim}\chi_r$, with $r=rank\,\mathbf{R}$ equal to the number of restrictions being tested.

In order to assess the adequacy of the model M for describing a set of data, we can compare the likelihood under the fitted model with the likelihood under the saturated model, which is the model with number of parameters equal to the total number of observations, n. The maximum likelihood achievable in a saturated model is attained at $\widetilde{\Psi}_i=y_i$, denoting $\Psi_i=\Psi\left(\beta^{\mathsf{T}}x_i\right)$. The deviance of the model M measures the discrepancy of the fit and it is defined as twice the difference between the maximum achievable log-likelihood of the saturated model and that attained by the fitted model, as follows:

$$\begin{aligned}Dev_M\left(y;\widehat{\Psi}\right)&=2\log L\left(\widetilde{\Psi};y\right)-2\log L\left(\widehat{\Psi};y\right)\\&=2\sum_{i=1}^{n}\left\{y_i\log\left(y_i/\widehat{\Psi}_i\right)+\left(1-y_i\right)\log\left(1-y_i/1-\widehat{\Psi}_i\right)\right\}.\end{aligned} \tag{3.2.17}$$

The deviance function is most directly useful not as an absolute measure of goodness-of-fit, but for the comparison between two nested models. Let $M_0 \subset M$ be a submodel with $p_0 < p$ regression parameters and consider testing M_0 within M. Then, we have the test given by

$$Dev_{M_0} - Dev_M \overset{(a)}{\sim} \chi^2_{p-p_0}, \tag{3.2.18}$$

which is identical to the likelihood-ratio statistic for testing M_0 against M.

A similar model selection procedure for non-nested models can be based on Akaike's information criterion (AIC):

$$AIC = -2\log L\left(\widehat{\Psi}; y\right) + 2p. \tag{3.2.19}$$

See Gourieroux (2000) or Greene (1993) for more information on this subject.

3.2.6 Logit versus Probit

It is not possible to compare directly both models, since they have different variance parameters, namely 1 for the normal and $\pi^2/3$ for the standard logistic distribution function. Thus, we must rescale any of the parameters in order to compare both distributions. Figure 3.1 plots the standard logistic cumulative distribution function against the cdf of $N\left(0, \pi^2/3\right)$. There is not a large difference between these cumulative distribution functions in the left graph, but we are interested in lower probabilities of default, which is the normal case in credit rating. Therefore we plot the right graph, and appreciate here that the logistic cdf vanishes to zero at a lower rate. This explains the fact that the logit model handles slightly better the case of extreme observations.

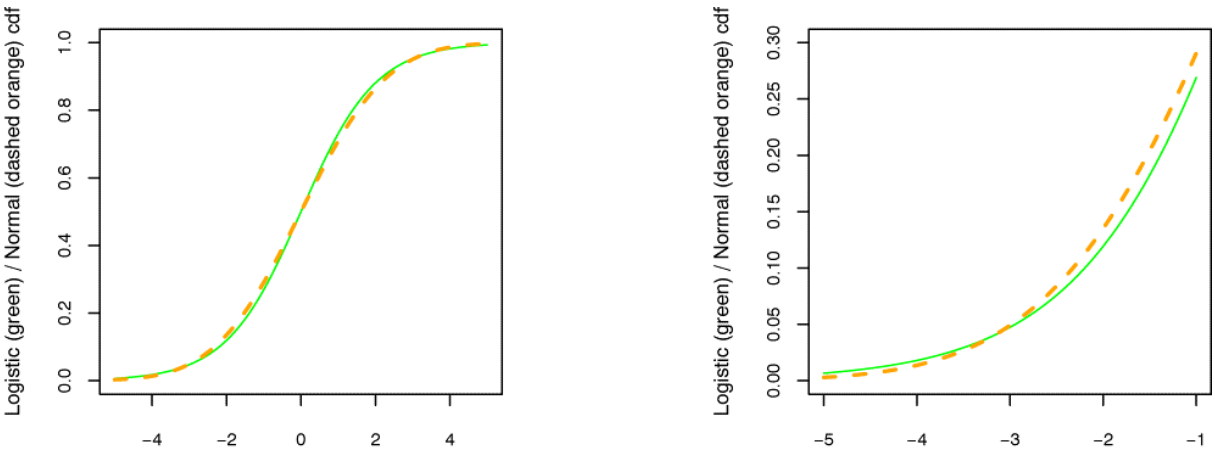


Figure 3.1: Logit vs. rescaled probit, left: on the range [-5,5], right: on [-5,-1].

Example 3.1

In Section 2.1 we already introduced *score1* (2.1.2) and *score2* (2.1.3), fitted by a logit model for the calibration sample. Let us denote their respective models by M (Table 3.1) and M_0 (Table 3.2), being $M_0 \subset M$. The coefficients were estimated with the help of the Fisher Scoring algorithm.

Variable	Coefficient	Std. Error	Z value	p-value
constant	1.389e+00	5.804e-01	2.393	0.016 **
account	-5.467e-01	7.628e-02	-7.167	7.69e-13 ****
duration	3.143e-02	9.521e-03	3.301	9.64e-04 ****
pay	-5.374e-01	1.010e-01	-5.323	1.02e-07 ****
amount	5.024e-05	4.368e-05	1.150	0.250
savings	-2.135e-01	6.279e-02	-3.401	6.71e-04 ****
time	-1.605e-01	7.658e-02	-2.096	0.036 **
month	2.560e-01	8.926e-02	2.868	4.126e-03 ***
status	-2.417e-01	1.246e-01	-1.940	0.052 *
properties	1.142e-01	9.135e-02	1.250	0.211
age	-9.964e-03	8.539e-03	-1.167	0.243
prev_credits	3.256e-01	1.748e-01	1.862	0.062 *
Signif. codes: '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 ' ' 1				
Deviance: 781.69, AIC: 805.69				
Number of Fisher Scoring iterations: 5				

Table 3.1: logit model M for *score1*

Variable	Coefficient	Std. Error	Z value	p-value
constant	0.826	0.358	2.304	0.021 **
savings	-0.258	0.056	-4.554	5.25e-06 ****
time	-0.177	0.065	-2.693	7.09e-03 ***
status	-0.208	0.109	-1.900	0.057 *
Signif. codes: '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 ' ' 1				
Deviance: 948.78, AIC: 956.78				
Number of Fisher Scoring iterations: 4				

Table 3.2: logit model M_0 for *score2*

In order to compare the two nested models, we calculate the likelihood ratio statistic, i.e. $Dev_{M_0} - Dev_M$=948.78–781.69= 167.09, which is highly significant,

since $Dev_{M_0} - Dev_M$ = 167.09 > $\chi^2_{8,0.9999}$ = 31.828. This means that *score1* fits better than *score2*.

Now we use a probit instead of a logit model to fit M' (see Table 3.3) and M'_0 (Table 3.4), for the same vector of explanatory variables $X \in \mathbb{R}^p$ in *score1* and $X \in \mathbb{R}^{p_0}$ in *score2*, being $M'_0 \subset M'$:

Variable	Coefficient (original)	(rescaled)	Std. Error	Z value	p-value
constant	7.598e-01	1.378e+00	3.374-01	2.252	0.024 **
account	-3.232e-01	-5.863e-01	4.353e-02	-7.425	1.13e-13 ****
duration	1.857e-02	3.369e-02	5.616e-03	3.308	9.41e-04 ****
pay	-3.159e-01	-5.729e-01	5.768e-02	-5.476	4.35e-08 ****
amount	2.795e-05	5.068e-05	2.585e-05	1.081	0.279
savings	-1.211e-01	-2.214e-01	3.557e-02	-3.414	6.41e-04 ****
time	-9.445e-02	-1.713e-01	4.467e-02	-2.115	0.034 **
month	1.509e-01	2.736e-01	5.194e-02	2.904	3.68e-03 ***
status	-1.340e-01	-2.430e-01	7.274e-02	-1.842	0.065 *
properties	7.161e-02	1.298e-01	5.315e-02	1.347	0.177
age	-5.556e-03	-1.007e-02	4.984e-03	-1.123	0.261
prev_credits	2.009e-01	3.644e-01	1.011e-02	1.987	0.046 **
Signif. codes: ‘****’ 0.001 ‘***’ 0.01 ‘**’ 0.05 ‘*’ 0.1 ‘ ’ 1					
Deviance: 781.73, AIC: 805.73					
Number of Fisher Scoring iterations: 5					

Table 3.3: probit model M' for *score1*

Variable	Coefficient (original)	(rescaled)	Std. Error	Z value	p-value
constant	0.486	0.882	0.217	2.240	0.02511 **
savings	-0.151	-0.275	0.032	-4.677	2.92e-06 ****
time	-1.108	-0.197	0.035	-2.780	6.02e-03 ***
status	-0.124	-0.226	0.059	-2.347	0.059 *
Signif. codes: ‘****’ 0.001 ‘***’ 0.01 ‘**’ 0.05 ‘*’ 0.1 ‘ ’ 1					
Deviance: 948.75,.AIC: 956.75					
Number of Fisher Scoring iterations: 4					

Table 3.4: probit model M'_0 for *score2*

Again, we calculate the statistic: $Dev_{M'_0} - Dev_{M'}$ = 167.02, which is also highly significant, since $Dev_{M'_0} - Dev_{M'} > \chi^2_{8,0.9999}$ = 31.828. Also here we prefer the larger model M' against M'_0.

The significance of the regression parameters and deviances are very similar for both logit and probit models. The deviances and difference of deviances indicate that the logit model fits slightly better than probit. As we want to compare the estimated coefficients from the probit to those of the logit model, we multiplied the probit coefficients in Table 3.3 and Table 3.4 by $\pi/\sqrt{3}$. The resulting rescaled coefficients are very similar to those for the logit model.

3.3 Further estimation methods

As the credit industry and large loan portfolios grow, the industry is developing more accurate credit scoring models. Even a fraction of a percent in credit scoring accuracy is an achievement. This is giving rise to the investigation of estimation methods like neural networks, that also inlcude nonparametric and semiparametric statistical methods.

3.3.1 Neural networks

As an alternative to linear discriminant analysis and regression models, neural networks have been analyzed more exhaustively in the last years since they represent the relationship between independent and dependent variables in a more flexible way. However, neural networks present some cons, as they are like a black box when it comes to interpret the resulting network. Moreover, calculating default probabilities with the help of neural networks is possible only to a limited extent and it requires considerable extra effort. Some empirical studies on this topic were accomplished by West (2000) or Barniv (1997).

3.3.2 Nonparametric and semiparametric methods

In the last section we studied the logit and probit models, which are special cases of the generalized linear model (GLM, see McCullagh & Nelder, 1983). For this class of nonlinear regression models, there is a variety of non- and semiparametric extensions. For example, in the nonparametric case we may estimate a single index model (SIM)

$$E(Y \mid X) = \widetilde{\Psi}\left(\beta^{\mathsf{T}} X\right),$$

where $\widetilde{\Psi}(\bullet)$ denotes an unknown smooth link function. Thus, this model overcomes restrictive assumptions of the functional form of the regression function. However, its interpretation may be difficult. If the number of regressors is large, SIM yields inaccurate estimates.

An example of semiparametric GLM is the generalized partial linear model (GPLM), which combines a linear and a nonparametric function. It is specified

$$E(Y \mid X) = \Psi\left(\beta^{\intercal} X_1 + m(X_2)\right),$$

$\Psi(\bullet)$ being a known parametric link function, $m(\bullet)$ an unknown smooth (possibly multidimensional) function and $X^{\intercal} = (X_1, X_2)^{\intercal} \in \mathbb{R}^p$. This kind of model keeps the easy interpretability of the parametric models and retains some of the flexibility of the nonparametric models.

Specific choices of the logit model can be found in Müller & Härdle (2002). For a thorough treatment of this topic we refer the reader to Härdle, Müller, Sperlich & Werwatz (2004).

3.4 Further aspects

3.4.1 Regression models for binary dependent variables and panel data

Panel models are also called models for clustered longitudinal data in statistics. They deal with the type of credit data that results from repeated measurements on the same individuals (loans) at different time points. Standard references for econometric panel data analyses are Arellano (2003) and Hsiao (1990).

An observation y_{it} has thus a transversal dimension ($i = 1,\ldots,n$) and a longitudinal dimension ($t = 1,\ldots,T$). In credit rating we have different time points i.e. $t = 1,\ldots,T_i$, for every loan. In this case, we speak about “unbalanced panel”, with $\sum_{i=1}^{n} T_i$ observations altogether.

The convenience of estimating methods that have on account the data structure explicitly is that they model the individual heterogeneity. By the study of an individual observation through the time, its individual characteristics can be differentiated from others. We also have to take into account that observations from the same individual are correlated.

In this section we will describe more carefully the fixed-effects-logit model and the random-effects-probit model. The modelling of the random-effects-logit model is analogue to that of the probit approach. The variant of the fixed-effects approach for the probit model is more problematic in its estimation (see Greene, 1993).

The general approach of the probit-model is:

$$\begin{aligned} y_{it}^{*} &= \beta^{\mathsf{T}} x_{it} + \varepsilon_{it} \\ y_{it} &= \begin{cases} 1 & if\ y_{it}^{*} > 0 \\ 0 & otherwise \end{cases} \end{aligned} \tag{3.4.1}$$

Here we assume that the error terms ε_{it} are independent standard normally distributed. The index i, $(i = 1,\ldots,n)$ describes the cross sectional dimension and t, $(t = 1,\ldots,T_i)$ the temporal dimension.

For the random-effects-model, the approach (3.4.1) is given by the following description of the error terms:

$$\varepsilon_{it} = v_{it} + u_i\,. \tag{3.4.2}$$

Both components are independent from each other and normally distributed with null expectation. The variance of ε_{it} (in which the variance of v_{it} is standardized to one) and the correlation between ε_{it} and ε_{is} are:

$$Var(\varepsilon_{it}) = \sigma_v^{\,2} + \sigma_u^{\,2} = 1 + \sigma_u^{\,2} \tag{3.4.3}$$

$$Corr(\varepsilon_{it}, \varepsilon_{is}) = \frac{\sigma_u^{\,2}}{1 + \sigma_u^{\,2}} =: \rho \tag{3.4.4}$$

The value of ρ represents the proportion of individual effects in the overall variance. The existence of individual effects can be studied by tests of significance of ρ. The likelihood function can be maximized after some transformations by means of numerical procedures.

The fixed-effects-logit model models the probability of occurrence of the interesting events, as follows:

$$P(y_{it} = 1) = \frac{\exp(\alpha_i + \beta^{\mathsf{T}} x_{it})}{1 + \exp(\alpha_i + \beta^{\mathsf{T}} x_{it})}. \tag{3.4.5}$$

The significance of individual effects can be tested with a Hausman-test. Under the hypothesis that there are no individual effects $\left(H_0 : \alpha_i = \alpha \ \forall i\right)$, is the conventional logit model appropriate (with the parameter estimates $\widehat{\beta}$). If the null hypothesis is rejected, then the fixed-effects-logit approach (with $\widehat{\beta}_{FE}$) is appropriate from a statistical point of view. The test statistic is given by:

$$\left(\widehat{\beta}_{FE} - \widehat{\beta}\right)^{\mathrm{T}} \left(Var\left(\widehat{\beta}_{FE}\right) - Var\left(\widehat{\beta}\right)\right)^{-1} \left(\widehat{\beta}_{FE} - \widehat{\beta}\right). \tag{3.4.6}$$

This follows a χ^2 distribution with p degrees of freedom, p being the amount of explanatory variables.

3.4.2 Classification and regression trees (CART)

Classification and regression trees present an alternative to fitting classical regression models. CART is a rule for predicting the behaviour of the response of interest from the values of its predictor variables. Classification trees apply when the response is categorical—in our case a credit default indicator—and regression trees when the response is continuous. The tree is constructed by recursively partitioning the learning sample of data into increasingly homogeneous subsets. The resulting subsets are heterogeneous among each other. To decide about this, we can use the entropy or the Gini index (see section 2.4) as impurity functions for the splitting criterion.

Assume we have a learning sample $A \subseteq \mathbb{R}^p$, containing the values of p predictor variables $X_1, \ldots, X_p$, and a default indicator Y. The Classification tree is generated as follows:

- First, we will choose that variable which discriminates the most between default and non-default. This variable and the split point obtained lift a partition of the initial learning sample $A = A_1$ in the subsets A_2, A_3, such that $A_2 \cap A_3 = \varnothing$ and $A_2 \cup A_3 = A_1$.

- Then we choose that variable which, starting from the first or the second subset has most discriminatory power. Thus, only one of the two possible subsets, e.g. A_2, will be split again in A_4, A_5 (being $A_4 \cap A_5 = \varnothing$ and $A_4 \cup A_5 = A_2$).

- By successively splitting the subsets, we will obtain a tree T. Choosing the maximal distance function (2.4.3) at every step minimizes the impurity of T. Denote $\widetilde{\mathrm{T}}$ the index set for the final nodes, the impurity of the tree is defined by:

$$\mathrm{i}(\mathrm{T}) = \sum_{\tau \in \tilde{\mathrm{T}}} \mathrm{i}(A_\tau) P(A_\tau). \tag{3.4.7}$$

For a thorough description of the CART algorithm, we refer the reader to Breiman et al. (1984) and Fahrmeir et al. (1996).

Classification trees can better model “non-monotonous” effects and interactions between the explanatory variables. However, as we normally deal with many explanatory variables in the credit-scoring context, the final interpretation of the tree may be complicated for our preference. We can also use CART for generating classes of a score, as we will do in the next section. There we will see that in practice, the algorithm tends to create pure or almost pure subsets of very little size.

3.4.3 Generation of rating classes

In the third consultative paper of the Basel Committee on Banking Supervision (2003) it is suggested that banks should have a minimum of 7 rating grades for non-defaulted borrowers and 1 grade for the defaulted ones; and they should be reasonably distributed across these grades, with no excessive concentrations.

Apart from this recommendation, there is in the literature neither consensus on the number of rating grades for the partition, nor a unique method to accomplish it. In some papers, i.e. Carey & Hrycay (2001), they use rating schemes of 5 and 10 rating classes; the “Oesterreichische Nationalbank” uses a fine and a coarse scale, the coarse scale containing 6 rating grades, with grade 6 denoting default; the rating agency Standard & Poor’s uses a scheme with 17 non-defaulted classes plus 1 class for defaults or its shortened version, with 7 plus 1 class for defaulted loans.

In Bemmann (2005) (about: Basel Committee [2000c, p. 23f]) it is summarized that from 30 rating agencies investigated, 22 used letter combinations, 6 used *numerical* marks and 2 used probabilities of default for expressing their ratings. From the letter-ratings 16 are conform to the S&P notation. However, most of the banks (ca. 85%) use *numerical* rating class notations.

About the generation of rating classes, it is being practised that the score is classified following some rule with respect to the default probability, e.g. doubled mean PDs per class; or it can be divided in intervals of a given size:

Example 3.2

We divided *score1* (2.1.2) for the calibration sample in 5 classes of equal size. For every class we calculated the mean probabilities of default (PD) asserted by the logit model. The column "percent" contains the percentages in the validation set of every class interval. We can observe here that the default rates for the validation sample are very close to the PDs for every class, but for the second:

Rating	score range	PD	percent (validation set)	default rate (validation set)
1	-4.848 - -2.192	0.061	0.295	0.067
2	-2.192 - -1.507	0.137	0.16	0.062
3	-1.507 - -0.791	0.243	0.19	0.236
4	-0.791 - 0.087	0.413	0.18	0.472
5	0.087 - 2.322	0.681	0.175	0.628

Table 3.5: rating classes for *score1*

We can also use CART (see 3.4.2) to generate classes of a score. In this case, instead of having p variables we consider the score S, which we split successively. See the following example:

Example 3.3

We will illustrate here the application of CART with two scores for a given number of classes (=5). Also here, we divided the score for the calibration sample, computing the mean probabilities of default (PD) asserted by the model. Then we calculated the percentages of every class interval and the default rates for the validation sample.

Rating	score range	PD	percent (validation sample)	default rate (validation sample)
1	-4.848 - -2.769	0.039	0.225	0.066
2	-2.769 - -0.399	0.197	0.52	0.221
3	-0.399 - 0.437	0.497	0.14	0.428
4	0.437 - 1.831	0.724	0.1	0.65
5	1.831 - 2.215	0.913	0.015	1

Table 3.6: rating classes for *score1* generated with CART

In Table 3.6 are represented the results for *score1* (2.1.2). Since we are interested in knowing if the classes are homogeneous, we will calculate the impurity of the resulting tree, using (3.4.7) with the entropy (2.4.9) as impurity function. In this case, $\mathrm{i}(T)$= 0.473 is close to the impurity given by the partition in Table 3.5, $\mathrm{i}(T')$= 0.454. If compared to the score $\mathrm{i}(S)$= 0.583, we can appreciate the diminution in impurity.

In the following Table 3.7 we present the results for another simulated sample, with 30% of defaults. Now we get that the resulting tree obtained with CART and entropy as impurity function has $\mathrm{i}(T)$ = 0.511. This is less than the overall impurity of the score, i.e. $\mathrm{i}(S)$= 0.610, and also less than the impurity we would obtain if we made a simple partition of the score in increasing order and intervals of equal length, i.e. $\mathrm{i}(T')$= 0.568. We rated the classes according to the percentage of default rates. In this case, it was better to make the classification with CART, since it contemplates the fact that the default rates are not increasing for increasing values of the score.

Rating	score range	percent (validation sample)	default rate
3	-7.661 - -5.359	0.185	0.075
5	-5.358 - -3.916	0.515	0.469
2	-3.914 - -3.851	0.018	0.055
1	-3.851 - -3.552	0.091	0.000
4	-3.552 - -1.093	0.191	0.225

Table 3.7: rating classes generated with CART

For both scores, we get that the impurity of the partition made with CART is close to the impurity obtained by a simple partition of the score in equal length intervals or even diminished, which speaks in favour of CART. The problem is, as we already mentioned in section 3.4.2, that CART tends to “isolate” little pure subsets of defaults or non-defaults, as for example, class 5 in Table 3.6, or classes 1, 2 and 3 in Table 3.7.

3.5 Summing up

The methods presented in the last sections can be applied for the purpose of quantifying credit risk. Now we will consider some aspects of the monotonicity and prediction of the default probability, and the parameter estimation in practice.

Binary choice models (3.2.2) can be estimated using statistical software[1], without a big effort if the number of defaults is sufficient. The observations of one year, i.e. a cross sectional dataset, will suffice. Moreover, the results produced by such models can be interpreted directly as predicted default probabilities. Among all the binary choice models, the logit model (3.2.11) is definitely the current standard, both in its practical application by regulators and in the academics literature.

However, variables that are identical for all borrowers, e.g. macro variables, will not be taken into account by the binary choice models. The parameters of these variables cannot be estimated, until there are observations on hand for all credit users during many years. The data structure together with the methods of section 3.4.1 can be applied in order to detect differences between the individual debtors. The problem of regression models for binary dependent variables and panel data was that historical data is not always available. This will not be the case after Basel II, since data has to be collected for at least 5 years for credit risk. Panel data will therefore play a more important role in the future.

Back to the binary choice models, we may say that their most important features are simplicity and the fact that the probability of default is modelled as a strictly monotone increasing function of the score. This does not hold in general for other estimation methods, as nonparametric and semiparametric methods, neural networks or CART. In case of monotonicity, we can efficiently apply the measures T, AR and D (described in sections, 2.2, 2.3 and 2.4), in order to assess the discriminatory power of the score.

[1] The methods described in section 3.2 were implemented with R 2.1.1 (www.r-project.org).

4 Validation and backtesting of PDs

4.1 Introduction

The rating classes of a rating system are normally constructed on the basis of probabilities of default that refer to one-year time horizons. This assignment can be accomplished in different ways, as we saw in the last section. In practice, the estimated probabilities of default will differ from the default rates that are afterwards observed. A problem arises when these deviations do not occur at random, but systematically. The question here is how the PDs suggested by the rating system can be reviewed with the updated default rates. A collection of studies on the topic of validation can be found in Basel Committee on Banking Supervision (2005).

For a rating system with R rating classes, let $0 < \pi_r < 1$ denote the default probability asserted by the rating system, $0 < p_r < 1$ the (unknown) actual PD and $0 < \hat{p}_r < 1$ the observed default rate, i.e. the proportion of defaulted borrowers of a total of n_r borrowers in the rating class r. We can differentiate between one-sided and two-sided test formulations. The one-sided is characterised through the hypotheses:

$$H_0 : p_1 = \pi_1, \ldots, p_R = \pi_R, \quad H_1 : \exists r \in \{1, \ldots, R\} \; with \; p_r > \pi_r, \tag{4.1.1}$$

and it is conform to the perception of the Banking Supervision, which is concerned about the fact that the risk should not be underestimated. The two-sided test formulation is depicted by:

$$H_0 : p_1 = \pi_1, \ldots, p_R = \pi_R, \quad H_1 : \exists r \in \{1, \ldots, R\} \; with \; p_r \neq \pi_r, \tag{4.1.2}$$

this can be identified with the perception of a risk controller, who is interested in as exact as possible estimation.

Under the assumption of stochastically independent default events, these two test formulations concern standard problems that can be addressed by the binomial test or the chi-square test, as we see in section 4.2.

The main problem is that credit defaults are *not* stochastically independent. As an option, we will assume in section 4.3 the dependence structure of the IRBA (Internal Ratings-Based Approach). By this dependence structure, the default rate does not converge to the associated probability of default, but to a non-degenerate probability distribution on the interval [0, 1].

In order to assess the quality of time-varying PD forecasts, we consider two approaches in section 4.4: normal test and traffic lights approach.

4.2 Tests based on the independence assumption

The construction of tests under the assumption of independent default events is based on the well-known facts:

1. The number of defaults in a rating class r with n_r credits and default probability p_r is binomially distributed:

$$n_{1r} \sim Bin(n_r, p_r), \tag{4.2.1}$$

being $n_{1r} = n_r \hat{p}_r$.

2. For the default rate $\hat{p}_r$ and $n_r \to \infty$ holds the strong law of large numbers,

$$\hat{p}_r \xrightarrow{a.s.} p_r, \tag{4.2.2}$$

and the central limit theorem

$$\sqrt{n_r}\,\frac{\hat{p}_r - p_r}{\sqrt{p_r(1-p_r)}} \xrightarrow{D} N(0,1). \tag{4.2.3}$$

3. For R rating classes holds

$$\sum_{r=1}^{R} n_r \frac{(\hat{p}_r - p_r)^2}{p_r(1-p_r)} \xrightarrow{D} \chi^2(R), \tag{4.2.4}$$

where $\chi^2(R)$ denotes a chi-square distribution with R degrees of freedom.

4.2.1 Binomial test

If we want to test if the probability of default of a rating category is correct against the alternative hypothesis that it is underestimated, then we can use the one-sided binomial test:

$$H_0 : p_r = \pi_r, \quad H_1 : p_r > \pi_r$$

for each rating category $r = 1,\ldots,R$.

The null hypothesis for a given level of significance α is rejected if the observed number of defaults n_{1r} is greater than a critical value k^*, given by:

$$k^* = \min\left\{k \mid \sum_{i=k}^{n_r} \binom{n_r}{i} \pi_r (1-\pi_r) \leq \alpha\right\}, \tag{4.2.5}$$

n_r being the total number of loans. For larger values of n_r, the calculation of (4.2.5) is very costly. Here we can make use of (4.2.3), i.e. the binomial distribution converges to the normal distribution as the number of trials increases. Therefore, the critical value $k_{1-\alpha}$ can be approximated as follows:

$$k^* \approx \Phi^{-1}(1-\alpha)\sqrt{n_r \pi_r (1-\pi_r)} + n_r \pi_r, \tag{4.2.6}$$

$\Phi^{-1}(\cdot)$ being the inverse function of a standard normal distribution. Put in terms of the default rate if preferred, we reject the null hypothesis if the observed default probability $\hat{p}_r$ is greater than $p_{1-\alpha}$:

$$p_{1-\alpha} \approx \Phi^{-1}(1-\alpha)\sqrt{\frac{\pi_r (1-\pi_r)}{n_r}} + \pi_r . \tag{4.2.7}$$

For the two-sided test:

$$H_0 : p_r = \pi_r, \quad H_1 : p_r \neq \pi_r,$$

we have that the critical region for $\hat{p}_r$ and an asymptotical level of significance α is given by: $\left[0, p_{\alpha/2}\right) \cup \left(p_{1-\alpha/2}, 1\right]$.

In both test formulations, the null hypothesis will be more difficult to reject for a lower number of loans in the rating class r, since $p_{1-\alpha}$ increases as n_r decreases.

Example 4.1

For *score1* (2.1.2) we apply here the one-sided binomial test to the rating class 3 defined in Table 3.5. For a level of significance α = 0.005 we have the default rate in the validation sample $\hat{p}_3$ = 0.236, which is not greater than $p_{0.995} \approx$ 0.422. Therefore, we have not enough statistical evidence to reject the null hypothesis against the alternative (that the probability of default asserted by the model, in this case π_3 = 0.243, is underestimated). In order to compare numerically the

binomial test with the test defined in section 4.3.1, which takes into account default correlation, we calculated the p-values for the rating classes 3 and 4. The results are listed in Table 4.1 of section 4.3.3.

4.2.2 Chi-square test

Now we want to test if the probabilities of default are correct for every rating category simultaneously:

$$H_0 : p_1 = \pi_1, \ldots, p_R = \pi_R, \quad H_1 : \exists r \in \{1, \ldots, R\} \; with \; p_r \neq \pi_r .$$

The chi-square test statistic is derived from the original and most known Pearson's chi-square statistic (see D'Agostino & Stephens, 1986) and is given by:

$$t_R = \sum_{r=1}^{R} n_r \frac{\left(\hat{p}_r - \pi_r\right)^2}{\pi_r \left(1 - \pi_r\right)}, \tag{4.2.8}$$

which fulfils (4.2.4) under H_0, when $n_r \to \infty$ simultaneously for all $r = 1, \ldots, R$, if all default events are independent within categories and between categories.

We will reject the null hypothesis for an asymptotical level of significance α, if t_R is greater than the $(1-\alpha)$-quantile of a χ^2 distribution with R degrees of freedom. For a lower number of loans in every rating class, the null hypothesis will be more difficult to reject.

Example 4.2

Now we test simultaneously for every rating class $r = 1, \ldots, 5$ of *score1* in Table 3.5 if the probabilities of default asserted by our logit model coincide with the real ones—in this case the default rates in the validation sample. For an asymptotical level of significance α = 0.005, we get that our statistic T_5 = 2.510 is not greater than $\chi^2_{5,0.995}$ = 16.75 and thus we cannot reject the hypotheses that the probabilities of default predicted by the model coincide with the real probabilities of default. We can observe the p-values in Table 4.2 of section 4.3.3 for a numerical comparison of this test with the test described in section 4.3.2.

4.3 The one factor threshold model of Basel II

As we have seen in section 4.2, the construction of the binomial and the chi-square tests is very simple and intuitive. However, from empirical studies it is known that default events are slightly correlated. Typical values for default correlation are around 0.005 to 0.03. Although these numbers may seem small, applying both tests under the assumption of correlated defaults makes the mathematical framework more complex.

For the modelling of the dependence structure of the Bernoulli distributed default variables

$$Y_{ri} \sim Ber(p_r) = Bin(1, p_r), \quad i = 1,\ldots,n_r, \quad r = 1,\ldots,R$$

for a given probability of default p_r we use continuous variables B_{ri} (*financial well-beings*). They stand for changes in the asset value or in the ability to pay.

For a given threshold γ_r, the default variable is defined

$$Y_{ri} = \begin{cases} 1 & if\ B_{ri} \le \gamma_r, \\ 0 & else. \end{cases} \tag{4.3.1}$$

The dependence structure of the financial well-beings is modelled

$$B_{ri} = \sqrt{\rho_r}\, Z + \sqrt{1-\rho_r}\, \varepsilon_{ri}, \tag{4.3.2}$$

where ρ_r denotes the asset correlation. B_{ri} depends on a systematic factor Z common to all debtors and a factor ε_{ri} that is specific to the debtor. Further assumptions are:

$$B_{ri} \sim N(0,1),\ Z \sim N(0,1),\ \varepsilon_{ri} \sim N(0,1),\ Cov(Z, \varepsilon_{ri}) = 0,\ Cov(\varepsilon_{ri}, \varepsilon_{sj}) = 0,$$

$$\text{and } Corr(B_{ri}, B_{sj}) = \sqrt{\rho_r \rho_s} =: \rho_{rs} \text{ for all } i = 1,\ldots,n_r,\ j = 1,\ldots,n_s,\ r,s = 1,\ldots,R.$$

Assume all loans are in the same rating category having the same threshold $\gamma_r = \Phi^{-1}(p_r)$, thus having the same probability of default and assume also that the asset correlation is the same for all pairs of loans. Then we have the properties:

$$E(Y_{ri}) = p_r = \Phi(\gamma_r),\ Var(Y_{ri}) = p_r(1-p_r) \text{ and}$$

$$Corr(Y_{ri}, Y_{sj}) = \frac{\Phi_2\left(\Phi^{-1}(p_r), \Phi^{-1}(p_s); \rho_{rs}\right) - p_r p_s}{\sqrt{p_r(1-p_r)\, p_s(1-p_s)}} =: \delta_{rs}, \text{ with } i \neq j \text{ in case that } r = s,$$

$\Phi_2(\cdot,\cdot;\rho_{rs})$ denoting the bivariate standard normal distribution function with correlation ρ_{rs} and δ_{rs} the default correlation.

The joint and marginal distributions of the default rates $\hat{p}_1,\ldots,\hat{p}_R$, can be calculated as:

$$P(n_1\hat{p}_1 = k_1,\ldots,n_R\hat{p}_R = k_R) = \int_{-\infty}^{\infty}\prod_{r=1}^{R}\binom{n_r}{k_r}p_{r|z}{}^{k_r}\left(1-p_{r|z}\right)^{n_r-k_r}d\Phi(z), \tag{4.3.3}$$

$$P(n_r\hat{p}_r = k_r) = \int_{-\infty}^{\infty}\binom{n_r}{k_r}p_{r|z}{}^{k_r}\left(1-p_{r|z}\right)^{n_r-k_r}d\Phi(z), \tag{4.3.4}$$

for $k_r = 1,\ldots,n_r$, $r = 1,\ldots,R$. $\Phi(\cdot)$ denotes the standard normal distribution function and

$$p_{r|z} := P(Y_{ri} = 1 \mid Z = z) = \Phi\left(\frac{\Phi^{-1}(p_r) - \sqrt{\rho_r}\,z}{\sqrt{1-\rho_r}}\right).$$

Due to the complexity of these formulas, it is difficult to develop exact tests for finite sample sizes. However, for sufficient many observations the tests' construction can rely on the asymptotical distribution of adequate test statistics. This work goes back to Huschens (2004), where the tests of the following subsections were developed.

4.3.1 Tests for one probability of default

The variance of the default rate $\hat{p}_r$ is given by

$$Var(\hat{p}_r) = \frac{p_r(1-p_r)}{n_r} + \frac{n_r-1}{n_r}\delta_r p_r(1-p_r)$$

and the asymptotical variance

$$\lim_{n_r\to\infty} Var(\hat{p}_r) = \delta_r p_r(1-p_r) = \Phi_2\left(\Phi^{-1}(p_r),\Phi^{-1}(p_r);\rho_r\right) - p_r{}^2. \tag{4.3.5}$$

The asymptotical distribution of a default rate $\hat{p}_r$ for $n_r \to \infty$ is given at Vasicek (2002):

$$\hat{p}_r \xrightarrow{D} p_r(Z) := \Phi\left(\frac{\Phi^{-1}(p_r) - \sqrt{\rho_r}Z}{\sqrt{1-\rho_r}}\right). \tag{4.3.6}$$

The respective cumulative distribution function of the random variable $p_r(Z)$ for $\rho_r > 0$ is the so-called Vasicek distribution:

$$F_r(x) := P\left(p_r(Z) \le x\right) = \Phi\left(\frac{\sqrt{1-\rho_r}\,\Phi^{-1}(x) - \Phi^{-1}(p_r)}{\sqrt{\rho_r}}\right). \tag{4.3.7}$$

By means of a convenient transformation, we get an asymptotically normally distributed random variable from the asymptotical distribution of the default rate. For $\rho_r > 0$ and $n_r \to \infty$ holds

$$A_r := \frac{\sqrt{1-\rho_r}\,\Phi^{-1}(\hat{p}_r) - \Phi^{-1}(p_r)}{\sqrt{\rho_r}} \xrightarrow{D} N(0,1), \quad r = 1,\ldots,R \tag{4.3.8}$$

- A suitable test statistic for the hypotheses

  $$H_0 : p_r = \pi_r, \quad H_1 : p_r > \pi_r$$

  and for a given $\rho_r > 0$ is

  $$\Lambda_r = \frac{\sqrt{1-\rho_r}\,\Phi^{-1}(\hat{p}_r) - \Phi^{-1}(\pi_r)}{\sqrt{\rho_r}}. \tag{4.3.9}$$

  Under the null hypothesis holds: $\lim_{n_r \to \infty} P\left(\Lambda_r \le \Phi^{-1}(\alpha)\right) = \alpha, \quad 0 < \alpha < 1$.

  The critical region for Λ_r with the asymptotical level of significance α is therefore given by: $\left(\Phi^{-1}(1-\alpha), \infty\right)$.

- For the two-sided test formulation

  $$H_0 : p_r = \pi_r, \quad H_1 : p_r \neq \pi_r$$

  we have the critical region for Λ_r: $\left(-\infty, \Phi^{-1}(\alpha/2)\right) \cup \left(\Phi^{-1}(1-\alpha/2), \infty\right)$.

Example 4.3

We want to apply here the one-sided test to the rating class 3 of *score1* (2.1.2) in Table 3.5, assuming two different default correlations. We have the probability of default asserted by the model π_3 = 0.243 and the default rate $\hat{p}_3$ = 0.236 for the validation sample. For ρ_3 = 0.005, the test statistic is Λ_3 = -0.293, which is not greater than the quantile of the normal distribution function $\Phi^{-1}(0.995)$ = 2.575. If we assume a higher correlation, i.e. ρ_3 = 0.03, then we obtain Λ_3 = -0.067. From

these results we can conclude that, in both cases, we have not enough statistical evidence to reject the null hypothesis.

4.3.2 Simultaneous tests for multiple probabilities of default

We have that for $\min_{r=1,..,R} n_r \to \infty$ holds

$$(\hat{p}_1,\ldots,\hat{p}_R) \xrightarrow{D} (p_1(Z),\ldots,p_R(Z)), \text{ with } Z \sim N(0,1)$$

and together with the assumption $\min_{r=1}^{R} \rho_r > 0$, also holds

$$(A_1,\ldots,A_R) \xrightarrow{D} (Z,\ldots,Z).$$

This implies that

$$\max_{r=1,..,R} A_r \xrightarrow{D} N(0,1) \text{ and } \frac{1}{R}\sum_{i=1}^{R} {A_r}^2 \xrightarrow{D} \chi^2(1).$$

- The one-sided test for

  $$H_0 : p_1 = \pi_1,\ldots,p_R = \pi_R, \quad H_1 : \exists r \in \{1,\ldots,R\} \; with \; p_r > \pi_r$$

  can be designed with the following test statistic

  $$\Lambda_{max} = \max_{r=1,..,R} \Lambda_r . \tag{4.3.10}$$

  The null hypothesis will be rejected in favour of the alternative H_1, with an asymptotical level of significance α, if $\Lambda_{max} > \Phi^{-1}(1-\alpha)$.

- Analogously, the two-sided test

  $$H_0 : p_1 = \pi_1,\ldots,p_R = \pi_R, \quad H_1 : \exists r \in \{1,\ldots,R\} \; with \; p_r \neq \pi_r$$

  can be based on the test statistic

  $$\Lambda = \frac{1}{R}\sum_{r=1}^{R} {\Lambda_r}^2 . \tag{4.3.11}$$

  In this case, for an asymptotical level of significance α, H_0 will be rejected if Λ is greater than the $(1-\alpha)$-quantile of a χ^2 distribution with one degree of freedom.

Example 4.4

Now we will use the one-sided simultaneous test for every rating class of *score1* (2.1.2) in Table 3.5. Assume the same correlations as in the last example, being constant for every rating class. First consider $\rho_r = 0.005$ for $r = 1,\ldots,5$. The test statistic $\Lambda_{max} = 2.118$ is not greater than the critical value $\Phi^{-1}(0.995) = 2.575$. The p-value= 0.017 is greater than the level of significance $\alpha = 0.005$. For $\rho_r = 0.03$ we get $\Lambda_{max} = 0.869 < 2.575$ and a p-value= $0.192 > 0.005$. Thus, in both cases we have not enough statistical evidence to reject the null hypothesis. However, if we chose $\alpha = 0.05$, then we would reject H_0 for the lowest correlation ($\rho_r = 0.005$), since the p-value= $0.017 < 0.05$.

4.3.3 A numerical comparison of the tests

We can compare the tests under the one factor threshold model of Basel II with the binomial and chi-square tests defined in section 4.2 by observing their respective p-values. So, for a certain default correlation, we can see if there are differences between these tests in the decision whether to reject or not the null hypothesis. We will illustrate this section with the p-values obtained by applying these tests for *score1* (2.1.2) in Table 3.5, assuming two different default correlations, $\rho_r = 0.005$ and $\rho_r = 0.03$, being constant for every rating class. Logically, the p-values given by the binomial and the chi-square tests will remain constant for different default correlations, as these tests rely on the independence assumption.

In the following table we applied the one-sided versions of the binomial test and the test for one probability of default for the rating classes 3 and 4:

ρ_r	p-value			
	binomial, $r = 3$	Λ_r, $r = 3$	binomial, $r = 4$	Λ_r, $r = 4$
0.005	0.603	0.615	0.236	0.017
0.03	0.603	0.526	0.236	0.192

Table 4.1: p-values for the one-sided tests of sections 4.2.1 and 4.3.1

For class $r = 3$, we arrive at the same conclusion for both tests and different default correlations, i.e. we lack of statistical evidence to reject H_0. On the other hand, the p-values differ more from each other for the rating class 4, being clearly higher for the binomial test. The decision here will depend on the significance level we choose. For $\alpha \geq 0.05$ and $\rho_r =$

0.005, we will reject H_0 for the test defined in section 4.3.1, but not for the binomial test. If we assume a higher correlation ρ_r = 0.03, then for both tests we have not enough evidence to reject the null hypothesis.

The p-values for the chi-square and the two-sided simultaneous test for multiple probabilities of default are depicted in the next table. The difference between them is obvious. Suppose we have a low default correlation ρ_r = 0.005, constant for every rating class r = 1,…,5. Then we can reject the null hypothesis for the test defined in section 4.3.2., but not for the chi-square test. In case ρ_r = 0.03, we have not enough statistical evidence to reject H_0 for both tests.

ρ_r	p-value	
	chi-square	Λ
0.005	0.774	4.087e-12
0.03	0.774	0.218

Table 4.2: p-values for the tests of sections 4.2.2 and 4.3.2

Thus, we can affirm that in this case, it is easier to reject the null hypothesis of exact forecasts for the tests based on the one factor threshold model of Basel II, especially for a low default correlation.

4.4 Validation of PDs for short time series

As in credit risk, defaults are collected generally only once per year, a comparison between the forecasts and the actual PDs can be made rarely. There exists an error that results from neglecting correlation in time and between assets for the validation methods of this section. For the empirical analysis of this error, we refer the reader to Blochwitz, Hohl, Tasche & Wehn (2004). They introduced a model that can be seen as extension of the Vasicek one factor threshold model of Basel II into the time dimension. In order to create close to reality time series of annual default rates for the simulation study, the following assumptions should be done:

1. A fixed portfolio is observed in years $t = 1,\ldots,T$

2. At any time t the number of borrowers in the portfolio is a deterministic number n_t, which is known a priori.

3. The variable C_t expresses the change in the general economic conditions from time $t-1$ to time t. The larger C_t, the better are the economic conditions.

4. $C = (C_1,...,C_T)^{\top} \sim N(\mu,\Sigma)$, being $C_t \sim N(0,1)$ for all $t = 1,\ldots,T$ and

$$\Sigma = \begin{pmatrix} 1 & r_{12} & \cdots\cdots & r_{1T} \\ r_{21} & 1 & \cdots\cdots & r_{2T} \\ \vdots & & \ddots & \vdots \\ r_{T1} & \cdots\cdots & r_{TT-1} & 1 \end{pmatrix} \tag{4.4.1}$$

   Denoting $r_{st} = \vartheta^{|s-t|}$ for some properly chosen $\vartheta \in [0,1]$.

5. The number of defaults n_{1t} at time t conditioned on C_t are independent, identically distributed random variables $n_{1t} \mid C_t \sim Bi\left(n_t, p_t\left(C_t\right)\right)$, being

$$p_t\left(C_t\right) = \Phi\left(\frac{\Phi^{-1}\left(p_t\right) - \sqrt{\rho_t}C_t}{\sqrt{1-\rho_t}}\right) \tag{4.4.2}$$

   and ρ_t represent the correlations of the changes of obligor's asset values from time $t-1$ to time t. The annual percentage default rates $\hat{p}_t$ will therefore be calculated as $\hat{p}_t = n_{1t} / n_t$.

4.4.1 Normal test

The normal test is a multi-period test of correctness of a default probability forecast for a single rating category. The assumptions needed for the test are that default events in different years are independent and the variance of the default rates is constant over time. Cross-sectional dependence is admissible.

Let $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_T$ be independent random variables with means $p_1, p_2, \ldots, p_T$ and common variance $\sigma^2 > 0$. Then by the central limit theorem we have:

$$\frac{\sum_{t=1}^{T}\left(\hat{p}_t - p_t\right)}{\sqrt{T}\sigma} \xrightarrow{D} N(0,1), \tag{4.4.3}$$

for T tending towards ∞. The rate of convergence is generally quite high. Thus, the approximation of the standardized sum to the standard normal distribution seems reasonable even for small values of T (e.g. $T = 5$).

Two different estimators of the assumed common variance are given by:

$$\hat{\sigma}_0^{\,2} = \frac{1}{T-1}\sum_{t=1}^{T}\left(\hat{p}_t - \pi_t\right)^2 \text{ and} \tag{4.4.4}$$

$$\hat{\sigma}^2 = \frac{1}{T-1}\left(\sum_{t=1}^{T}\left(\hat{p}_t - \pi_t\right)^2 - \frac{1}{T}\left(\sum_{t=1}^{T}\left(\hat{p}_t - \pi_t\right)\right)^2\right), \tag{4.4.5}$$

where π_t denotes the forecasted probability of default for the year t. Under the hypothesis of exact forecasts, both estimates are unbiased. In case of mismatches, both are biased, but the second estimator (4.4.5) reduces considerably the bias.

We will test the hypotheses

$$H_0 : p_1 = \pi_1, \ldots, p_T = \pi_T, \quad H_1 : \exists t \in \{1, \ldots, T\} \; with \; p_t > \pi_t$$

with the following test statistic:

$$\Lambda_N = \frac{\sum_{t=1}^{T}\left(\hat{p}_t - \pi_t\right)}{\sqrt{T}\,\hat{\sigma}}. \tag{4.4.6}$$

H_0 is rejected for an asymptotical level of significance α if $\Lambda_N > \Phi^{-1}(1-\alpha)$.

4.4.2 Extended traffic lights approach

The traffic lights approach can be considered as an efficient tool for identifying dubious credit portfolios or rating grades. It is based on individual trigger levels for default probabilities, as such thresholds that should not be exceeded by an ex-post default rate for a given rating class and its respective ex-ante PD. This approach is rather a graphical visualization of the observed default rate in relation to the forecasted default probability than a statistical test. So, rating grades that have a reddish colour are assumed to underestimate the credit risk; rating grades with a rather green colour are supposed to be conservative enough and the rest (basically yellow) should be treated in-between. Tasche (2003) presents a method for calculating the critical values that avoids simulations but requires explicit specification of asset correlations.

In addition, the extended traffic lights approach can be regarded as a multi-period backtesting tool for a single rating category under the assumptions of independent default events in a

rating class and in time. In contrast to the normal test, the traffic lights test does not assume constant or nearly constant variance of the default rates over time. For more information on the traffic lights approach, we refer the reader to Blochwitz, Hohl & Wehn (2005).

By the central limit theorem, the distribution of the standardized default rate can be approximated to the standard normal distribution as long as $n_t p_t$ is not too small[2]:

$$\frac{n_{1t} - n_t p_t}{\sqrt{n_t p_t (1 - p_t)}} = \frac{\hat{p}_t - p_t}{\sqrt{\frac{p_t (1 - p_t)}{n_t}}} \xrightarrow{D} N(0,1),$$

Define the probabilities q_g, q_y, q_o and q_r[3]—which correspond to the colours green, yellow, orange and red—with $q_g > q_y > q_o > q_r$ and $q_g + q_y + q_o + q_r = 1$, and the mapping

$$M(\hat{p}_t) = \begin{cases} g, & \hat{p}_t \le \Phi^{-1}(q_g) \\ y, & \Phi^{-1}(q_g) < \hat{p}_t \le \Phi^{-1}(q_y) \\ o, & \Phi^{-1}(q_y) < \hat{p}_t \le \Phi^{-1}(q_r) \\ r, & \Phi^{-1}(q_r) < \hat{p}_t \end{cases} \tag{4.4.7}$$

Since the annual numbers of default are assumed to be independent, the vector A counting the appearances of colour $c \in \{g, y, o, r\}$ in the sequence $M(\hat{p}_1), \ldots, M(\hat{p}_T)$ is approximately multinomially distributed with

$$P[A = (a_g, a_y, a_o, a_r)] = \frac{T!}{a_g! a_y! a_o! a_r!} q_g^{a_g} q_y^{a_y} q_o^{a_o} q_r^{a_r}, \tag{4.4.8}$$

for every quadruple of non-negative integers such that $a_g + a_y + a_o + a_r = T$. Each quadruple for any time series can be labelled uniquely by means of an order function:

$$\Lambda_T = \Lambda_T(a_g, a_y, a_o, a_r) = w_g a_g + w_y a_y + w_o a_o + w_r a_r, \tag{4.4.9}$$

such that $w_g > w_y > w_o > w_r$. In the existing literature for simulation studies (see Blochwitz et al., 2005), we found that for $T \le 9$, vectors of weights such as $w^{\intercal} = (q_g, q_y, q_o, q_r)$ or $w^{\intercal} = (1000, 100, 10, 1)$ were used and they turned out to be appropriate.

---

[2] Dinges & Rost (1982) suggest the rule of thumb: $n_t p_t (1 - p_t) > 9$.

[3] In the simulation study by Blochwitz et al. (2004), they choose q_g ,= 0.5, q_y = 0.3, q_o = 0.15 and q_r = 0.05.

The test hypotheses can be formulated as follows:

$$H_0 : p_1 = \pi_1, \ldots, p_T = \pi_T, \quad H_1 : \exists t \in \{1, \ldots, T\} \; with \; p_t > \pi_t .$$

H_0 will be rejected against the alternative for an asymptotical level of significance α, if $\Lambda_T \leq v_{1-\alpha}$, being $v_{1-\alpha}$ the greatest number v such that $P(\Lambda_T \leq v) < \alpha$.

4.4.3 Normal vs. traffic lights

We have seen that both the normal and the traffic lights tests are asymptotic, with respect to the length of the time series and the portfolio size, respectively. Consequently, even under complete independence in time and in the portfolio, the observed type I[4] errors might be lower than the nominal error level of the test. If the type I error agrees with the nominal level of the test, the next question is for which test the type II[5] error is lower, i.e. which test is more powerful.

Some simulation studies, like Blochwitz et al. (2004), are intended to solve these questions. There we can realise that both methodologies are robust against the violation of the assumption of independence in time in their designs, being the normal test slightly more robust. On the other hand, the traffic lights test seems to be generally more powerful than the normal test, in particular for short time series. Therefore, simultaneous applications of the tests should be favoured.

4.5 Summary

Along this section, we did an overview of the statistical methods existing in the academic literature for assessing the estimation quality of the default probabilities. However, these methods display shortcomings in practice.

When independence of default events is assumed, the binomial test (section 4.2.1) can be applied in order to test the accuracy of a one period default probability forecast, for only one rating category at a time. We can check several categories simultaneously by applying the chi-square test (section 4.2.2). The problem of these tests is that they do not estimate correctly the true type I error, since in reality default events are correlated.

[4] The probability of erroneously rejecting the null hypothesis.
[5] The probability of not rejecting the null hypothesis if specific alternatives are true.

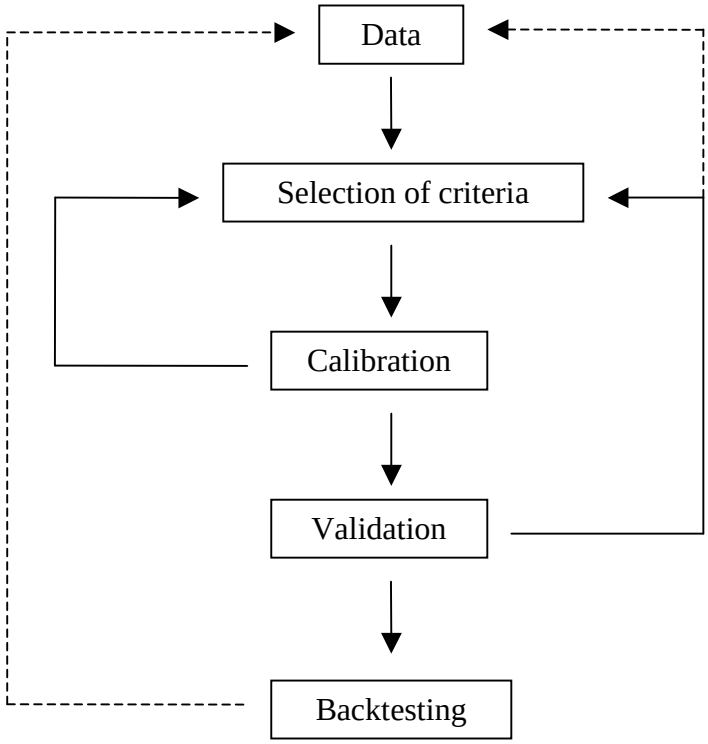
In order to deal with the problem of correlation, we presented the tests of section 4.3. These tests are based on the one factor threshold model of Basel II, which is intended to model the dependence structure of the default variables. As these tests are asymptotic, as well as the rest of the tests of this section, it is also expected that the actual type I errors are lower than the nominal error level of the test. But there are no simulation studies that can show to which extent this would happen.

Furthermore, it is also required to validate the PD estimates for time series. By availability of historical data, this can be attained by applying the normal test and the extended traffic lights approach, which are described in section 4.4. The traffic lights test has more power than the normal test, which is in contrast mildly more robust against the violation of the independence assumption.

Therefore, it should be emphasized, that there is no method to fit all situations that might occur in the validation process. Depending on the specific circumstances, the combination of different techniques will be the most appropriate way to address the validation exercise.

5 Guidelines for credit rating

As the purpose of this thesis is to study the statistical aspects of developing a credit rating system, we accomplished in the previous sections an elaborate overview of the different methodologies that are used in practice, analysing their pros and cons and, in some cases, proposing alternative measures. A layout of the process of credit rating is given by the scheme (already pictured in section 1. Introduction, Figure 1.1):



So, the first step of the process will be the selection of rating criteria from our dataset. In Appendix B we summarize the recommendations of the Basel Committee on Banking Supervision (2001) and the literature existing on this topic. In addition, we may select among all the factors, the most relevant with respect to their capacity to distinguish between default and non-default, i.e. by assessing their discriminatory power.

The different measures of discriminatory power that are applied in practice were described in section 2. The overlapping area criterion T (section 2.2) and the accuracy ratio AR (section 2.3)—linked to the Kolmogorov-Smirnov and the Mann-Whitney U tests, respectively—

seem to be in general appropriate discriminatory power measures. On the other hand, we could see also that there are other measures that do not suit for credit rating, e.g. the misclassification rate.

Moreover, we found out that the entropy-based criterion for reduction in impurity D_e (2.4.10) is also valid for assessing the discriminatory power. This criterion has been paid no attention, because of the misconception in the Basel Committee on Banking Supervision (2005), where they argue that are no tests applicable for those entropy-based measures. However, in section 2.4.5 we showed that it is related to the test for homogeneity in 2×2 contingency tables and hypotheses can hence be tested.

Thus, T, AR and D_e turn out to be suitable discriminatory power measures for the purpose of credit scoring. In the comparison of section 2.6 we could see that there is not a measure that performs best for every situation. It is remarkable, that the accuracy ratio performs in general worse than the other measures if there is no monotonicity.

After having selected the rating criteria, we proceed to the calibration of the model. We described the different methods on hand in section 3. The well-known logit model (3.2.11), and the probit model (3.2.12) belong to the binary choice models (section 3.2). If there are enough defaults in our dataset with the observations of one year, they are easy to estimate. By availability of historical data, we may apply panel models (see section 3.4.1). That way, one can consider time dependent macro variables that cannot be taken into account by binary choice models.

The binary choice models are widespread in credit rating because they have an easy interpretation and the probability of default is modelled as a strictly monotone increasing function of the score. This does not apply in general for other estimation methods, like neural networks (section 3.3.1), nonparametric and semiparametric methods (section 3.3.2) or CART (section 3.4.2).

In the case of logit or probit models, we can test if a model fits better than another by the difference of their deviances (3.2.18) or with Akaike's information criterion (3.2.19). If the model is not significant, we may discard variables that are not relevant and/or select additional explanatory variables. To continue with the process of validation, we divide our score in rating classes (section 3.4.3) and calculate for every class the mean probability of default asserted by the model.

An overview of the different methods for the validation and backtesting of PDs was given in section 4. We must remark, that there is no method to fit all situations. So, under the assumption of independence of default events, the binomial test (section 4.2.1) and the chi-square test (section 4.2.2) can be applied. But these tests underestimate the true type I error, since default events are in fact correlated. The tests of section 4.3 are based on the one factor threshold model of Basel II, which models the dependence structure of the default variables.

In order to validate the default probabilities for time series by availability of historical data, we can apply the normal test (section 4.4.2) and the extended traffic lights approach (section 4.4.2). The normal test is slightly more robust against the violation of the assumption of independence, and the traffic lights test is more powerful.

If the forecasted probabilities of default are significantly different of the default rates in the validation sample, we may choose alternative rating criteria or review the data, and calibrate the model again.

Once our model is validated, it can be backtested with real default rates. Obviously, the model can also turn outdated, so that the original forecasted PDs will not coincide with the real default rates. In that case, we should return to the dataset, complement it with new data and start from the beginning.

Appendix

A Notation

Y	default variable ($Y = 1$ for default, $Y = 0$ otherwise).
S	score $S = S(X_1,...,X_p) \in \mathbb{R}$.
$X = (X_1,\ldots,X_p)^\intercal$	vector of explanatory variables.
β	parameter vector, normally of weights.
$S \mid Y = 0$	score S conditioned to the default variable $Y = 0$.
$S \mid Y = 1$	score S conditioned to the default variable $Y = 1$.
$f_0,\ f_1$	probability density functions of $S \mid Y = j,\ j = 0,1$.
$F_0,\ F_1$	cumulative distribution functions of $S \mid Y = j,\ j = 0,1$.
$\lfloor \mathrm{x} \rfloor$	gives back the greatest integer smaller or equal to $x \in \mathbb{R}$.
PD	probability of default

B Selecting rating criteria

The Basel Committee on Banking Supervision (2001) issued a second round of consultative documents proposing changes to the capital requirements for banks. The core of the Internal Ratings Based (IRB) approach is to meaningfully differentiate borrowers based on risk. Banks should therefore take all relevant information into account in assigning ratings to a borrower. This information should be current. The methodologies and data used in assigning ratings should be clearly specified and documented. As a minimum, a bank should look at each of the following factors for each borrower:

- Historical and projected capacity to generate cash to repay its debts and support other cash requirements, such as capital expenditures;
- Capital structure and the likelihood that unforeseen circumstances could exhaust its capital cushion and result in insolvency;
- Quality of earnings, that is, the degree to which its revenue and cash flow emanate from core business operations as opposed to unique and non-recurring sources;
- Quality and timelines of information about the borrower, including the availability of audited financial statements, the applicable accounting standards and its conformity with the standards;
- Degree of operating leverage and the resulting impact that demand variability would have on its profitability and cash flow;
- Financial flexibility resulting from its access to the debt and equity markets to gain additional resources;
- Depth, skill and prudence of management and its ability to effectively respond to changing conditions and deploy resources;
- Position within the industry and the future prospects; and
- Risk characteristics of the country it is operating in, and the impact on the borrower's ability to repay, (including transfer risk) where the borrower is located in another country and may not be able to obtain foreign currency to service its debt obligations

In this summary, we collected some rating criteria to be found in the existing literature. For private companies, there will be required information about their volume of sales, legal form, financial state and profitability. For private consumers and mortgages, financial as well as personal information will be considered.

B.1 Private companies

B.1.1 Kaiser & Szczesny (2001)

The dataset was raised from The Centre for Financial Studies, CFS. It consists of 260 credit records of medium-size companies from 1992 to 1998. For the empirical analysis there were used the following variables:

- **Default:** dummy variable, which takes the value 1 if there were problems with the fulfilment of the contract and 0 otherwise.
- **Default_3**: a value of 0 indicates no problems; a value of 1 indicates some problems, but still no total failure of the credit, and a value of 2 stands for severe problems.
- **ln(Sales Volume):** this variable represents the size of the company on the basis of the sales volume, which are transformed with the natural logarithm.
- **ln(Sales Volume)^2:** in order to consider possible non linear influences of the business size, squared logarithmic conversions were taken up for the estimations.
- **Equity Ratio:** equity ratio, computed as the quotient made of own capital funds and total assets.
- **Cash flow:** dynamic cash flow, expressed as the quotient from cash flow and net debts.
- **Assets Coverage Degree:** the quotient from medium- and long-term liabilities and medium- and long-term assets.
- **Restricted Liability:** dummy variable, which takes the value 1, if the firm is only limited reliable, otherwise takes the value 0.
- **1992, 1993,..., 1998:** dummy variables, which indicate in which year was originated the observation, whereby the year 1992 is taken in estimation as reference.

- **Manufacturing:** binary variable, which labels companies from the sector of the manufacturing industry.
- **Construction:** binary variable for the construction industry.
- **Retail:** binary variable for the retail market.
- **Other:** binary variable for other firms, which mainly proceed form the sectors service, transport and logistics.

B.1.2 Khandani, Lozano & Carty (2001)

The goal of RiskCalc™ Germany is to provide a probability of default for private firms in Germany, with annual turnover of more than € 0.5 m. However, due to the very different nature of some firms, they eliminated from their analysis small companies, financial institutions, public institutions, real estate companies and affiliates.

For the model, they considered companies as having defaulted, if they entered or undergone bankruptcy, debt compositions proceedings, debt moratorium or cheque or bill protest. And used nine factors, which fall within the following broad categories: leverage/gearing, profitability, debt coverage, growth, activity and productivity.

Leverage/Gearing Ratios

- **Equity ratio**: (Equity – Intangible assets) / (Total assets – Intangible assets – Cash & Equivalents – Land & Buildings)
- **Net indebtedness** = (Current liabilities – Cash & Equivalents) / Total assets
- **Liabilities structure** = (Trade liabilities + Notes payable + Bank liabilities) / (Liabilities – Advances).

Profitability

- **EBITD** = (Net profit + Interest expenses + Income taxes + Depreciation) / Total assets
- **Profit on Sales** = Ordinary profit / Sales

Debt coverage

- **Debt coverage** = Cash Flow / (Liabilities – Advances)

Growth

- **Sales** Growth Sales(t) / Sales(t-1)

Activity

- **Trade creditors ratio** = ((Notes payable + Trade liabilities) *360) / Sales

Productivity

- **Personnel expenses on sales** = Personnel expenses / Sales

B.1.3 Standard & Poor's (2005)

- **EBIT** interest Coverage = Earnings from continuing operations* before interest and taxes / Gross interest incurred before subtracting (1) capitalized interest and (2) interest income

- **EBITDA** interest coverage = Earnings from continuing operations* before interest, taxes, depreciation and amortization / Gross interest incurred before subtracting (1) capitalized interest and (2) interest income

- **Funds from operations / total debt** = Net income from continuing operations plus depreciation, amortization, deferred income taxes, and other non-cash items / Long-term debt** plus current maturities, commercial paper, and other short-term borrowings

- **Free operating cash flow / total debt** = Funds from operations minus capital expenditures, minus (plus) the increase (decrease) in working capital (excluding changes in cash, marketable securities, and short term debt) / Long-term debt** plus current maturities, commercial paper, and other short-term borrowings

- **Return on capital** = EBIT / Average of beginning of year capital, including short-term debt, current maturities, long-term debt**, non-current deferred taxes, and equity

- **Operating income / sales** = Sales minus cost of goods manufactured (before depreciation and amortization), selling, general and administrative, and research and development costs / Sales

- **Long-term debt / capital** = Long-term debt** / Long-term debt + shareholders' equity (including preferred stock) plus minority interest

- **Total debt / capital** = Long-term debt** plus current maturities, commercial paper, and other short-term borrowings / Long-term debt plus current maturities, commercial paper, and other short-term borrowings + shareholders´ equity (including preferred stock) plus minority interest

* Including interest income and equity earnings; excluding nonrecurring items.

** Including amount for operating lease debt equivalent.

B.2 Private consumers

B.2.1 Giese (2002)

When we speak about retail customers, according to the definition of Basel II it concerns exclusively private consumers, because Basel II allows the local modulators a certain margin in order to rank likewise small firms among the retail portfolio. The rating due to the personal data of the company's owner is here more meaningful than consulting financial ratios. Typical data for a rating within the retail sector are:

- **Personal Data:** age, sex, yearly income, civil status, number of renting members in the household, living years at the current/previous address, residential property, etc.

- **Occupation**: kind of job, professional years, years in current/previous conditions of employment, number of employees (if executive/autonomous), etc.

- **Credit**: kind of credit, size of the credit, running time, frequency of the repayments, presence/value of collaterals, etc.

- **Past behaviour**: number and size of credits in the past, late/failed repayments with previous credits, etc.

B.2.2 Barron & Staten (2003)

The credit-reporting environment varies widely around the globe. The limits on the reporting of consumer payment histories are typically government-imposed (perhaps as a result of concerns about consumer privacy) or the result of the reluctance of incumbent lenders to share valuable customer information with potential competitors.

Historically, credit reporting in most countries began with the sharing of so-called “negative” information (delinquencies, charge-offs, bankruptcies, etc.) on borrowers. Only gradually and recently has information about the *successful* handling of accounts (prior and current) been contributed to the data repository. They also demonstrate in their paper how the availability of such “positive” data can substantially boost the effectiveness of scoring models and expand credit availability to consumers.

- **Outstanding Debt and Types of Credit**
    - Total number of open, paid or closed trades
    - No open, paid or closed trades
    - Number of trades open with a balance greater or equal to zero
    - No trades open with a balance greater than or equal to zero
    - Number of trades opened in the last 6 months
    - No trades opened in the last 6 months
    - Number of trades opened in the last 12 months
    - No trades opened in the last 12 months
    - Proportion of open trades that is revolving
    - Proportion of open trades that is finance instalment
    - Proportion of open trades that is real state/property
    - Zero balance on open trades

  - Average balance across all open trades
  - Average balance across open revolving trades
  - Proportion of debt that is revolving
  - Proportion of debt that is finance instalment
  - Proportion of debt that is real state/property
  - Bankcard balance/limit ratio on all open trades reported in last 6 months
  - Bankcard balance/limit ratio on all open trades reported in last 12 months

- **Length of credit history**
  - Age, in months, oldest trade
  - Age, in months, of most recently open trade
  - Age, in months, of most recently open trade = 9999
  - Average age, in months, of all trades
  - Ratio of number of open trades reported, last 12 months to age of oldest trade

- **New Applications For Credit (Inquiries)**
  - Total number of inquiries made for credit purposes
  - No inquiries made for credit purposes
  - Total number of bankcard inquiries made for credit purposes
  - No bankcard inquiries made for credit purposes
  - Months since most recent inquiry for credit purposes was made
  - Months since most recent bankcard inquiry for credit purposes was made
  - Total number of inquiries for credit purposes made, last 6 months

- Proportion of inquiries to open trades, last 6 months
- Total number of inquiries for credit purposes made, last 12 months
- Proportion of inquiries to open trades, last 12 months

- **Late Payments, Delinquencies and Bankruptcies**
  - Proportion of all trades never delinquent/ derogatory
  - Proportion of all trades that have never been delinquent, last 12 months
  - Positive number of trades ever 60+ days delinquent or derogatory
  - Number of trades ever 60+ days delinquent or derogator
  - Proportion of trades ever 60+ days delinquent or derogatory
  - Positive number of trades ever derogatory, including collection, charge-off, etc.
  - Number of trades ever derogatory
  - Proportion of trades ever derogatory
  - Positive number of bankruptcy tradelines ever
  - Total number of bankruptcy tradelines ever (only available for all)
  - Proportion of trades ever bankruptcy tradelines
  - Months since most recent tradeline bankruptcy
  - Worst status ever (including current) on a trade
  - Worst ever status on trades reported, last 12 months
  - Worst present status on an open trade
  - Worst status ever (including current) on a bankcard trade
  - Worst ever status on bankcard trades reported, last 12 months

- Worst present status on an open bankcard trade
- Months since most recent 30-180 day delinquency on any trade
- Not ever delinquent or derogatory on any trade
- Months since most recent 90+ delinquency or derogatory, any trade
- Not ever 90+ days delinquency or derogatory item on any trade

B.2.3 Jacobson & Roszbach (2003)

This paper uses a large data set with Swedish consumer credit data that contains extensive financial and personal information on both rejected and approved applicants at a major Swedish lending institution between September 1994 and August 1995.

The variables that have been selected for the estimation of the empirical model are:

- AGE: **age** of applicant
- MALE: dummy, takes value 1 if applicant is **male**.
- DIVORCE: dummy, takes value 1 if applicant is **divorced**.
- HOUSE: dummy, takes value 1 if applicant **owns** a (possible mortgaged) **house**.
- BIGCITY: dummy, takes value 1 if applicant lives in one of the three **greater metropolitan** areas around Göteborg, Malmö and Stockholm.
- NRQUEST: number of **requests** for **information** on the **applicant** that the credit agency received during the last 36 months
- ENTREPR: dummy, takes value 1 if applicant has taxable income form a registered **business**.
- INCOME: **annual income** from wages, relative to preceding year, as reported to Swedish tax authorities in 1993 or 1994 (depending on granting date) (in SEK 1000)
- DIFINC: **change** in **annual income** from wages, relative to preceding year, as reported to Swedish tax authorities (in SEK 1000)

- CAPINC: dummy, tales value 1 if applicant has **taxable income** from **capital**.
- BALINC: **ratio** of **total collateral-free credit facilities** actually utilized and **INCOME**, expressed as percentage. This variable is defined as: DUMMY{income>0} * (BALANCE/INCOME).
- ZEROLIM: dummy, takes value 1 if applicant has **no collateral-free loans outstanding**.
- LIMIT: total **amount** of **collateral-free credit facilities already outstanding** (in 1000 SEK)
- NRLOANS: **number** of collateral-free loans already outstanding
- LIMUTIL: percentage of **LIMIT** that is **actually being utilized**.
- LOANSIZE: **amount of credit granted** (in 1000 SEK)
- COAPPLIC: dummy, takes value 1 if applicant has a **guarantor**.

B.2.4 Dionne, Artís & Guillén (1996)

The data come from a sample of clients that had been granted credit by a Spanish bank. The sample was taken in May 1989. Even though this paper is limited to the study of the probabilities of default for those clients who had already a credit, the authors think that the methodology could also be considered for other applications in this field, including the granting decision when data on refused clients are available.

- Y: **number of non-payments.**
- YDUM: 1 if the **number of non-payments** is equal to or **greater** than four. 0 otherwise.
- DT6: 1 if total **contract duration** of return period is more than four years. 0 otherwise (reference group).
- DUREEA: **Number** of **months** form the beginning of the contract at the sampling date.

- AGE1: 1 if the age group is **18-24 years**. 0 otherwise.
- AGE2: 1 if the age group is **25-39 years**. 0 otherwise.
- AGE3: 1 if the age group is **40 years or more**. 0 otherwise.
- DESTIN: 1 if the credit is used to purchase a good with collateral. 0 otherwise.
- ETUI1: 1 if the client has **not completed primary education**. 0 otherwise.
- ETUI2: 1 if the client has **completed primary education**. 0 otherwise.
- ETUI3: 1 if the client has completed **higher education**. 0 otherwise.
- ETUI4: 1 if the client has a **university** degree. 0 otherwise.
- RECSAL: 1 if the client **receives** the **salary through** the **bank**. 0 otherwise.
- M1: 1 if **married**, **non-owner**, salary **under $3,000**. 0 otherwise.
- M2: 1 if **married**, **non-owner**, salary **higher than** (equal to) **$3,000**. 0 otherwise.
- M3: 1 if **married**, **owner**, salary **under $3,000**. 0 otherwise.
- M4: 1 if **married**, **owner**, salary **higher than** (equal to) **$3,000**. 0 otherwise.
- NM1: 1 if **not married**, **non-owner**. 0 otherwise.
- NM2: 1 if **not married**, **owner**. 0 otherwise.
- CENTRE: 1 if the credit is granted by a **store**. 0 otherwise.
- RESID: 1 if resident in the city for **at least four years**. 0 otherwise.
- Z1: 1 if south Spain (Andalucía, Canarias, Castilla-La Mancha, Extremadura, Murcia). 0 otherwise.
- Z2: 1 if north (Aragon, Asturias, Cantabria, Castilla-León, Galicia, Navarra, País Vasco). 0 otherwise.
- Z3: 1 if east (Baleares, Catalunya, Valencia). 0 otherwise.

- Z4: 1 if centre (Madrid). 0 otherwise.

B.2.5 Hand & Henley (1997)

Hand and Henley indicate in its paper as a data example a table, which represents the kind of characteristics of a credit rating for private customers

- **Time at the present address:** 0-1, 1-2, 3-4, 5 + years
- **Home status:** owner, tenant, other
- **Postcode:** band A, B, C, D, E
- **Telephone:** yes, no
- **Applicant's annual income**
- **Credit card:** yes, no
- **Type of bank account:** cheque and/or savings, none
- **Age:** 18-25, 26-40, 41-55, 55 + years
- **Country Court judgments:** number
- **Type of occupation:** coded
- **Purpose of loan:** coded
- **Marital status:** married, divorced, single, widow, other
- **Time with bank:** years
- **Time with employer:** years

B.3 Mortgages

B.3.1 Fair Isaac & Company, Inc. (n.d.)

FICO scores were developed by Fair Isaac & Company, Inc. and it might be the most commonly used method to value mortgages. They only consider the information contained in a person's credit file. This information can be grouped into five categories as outlined below:

- **Payment History**
    - Account payment information on specific types of accounts (credit cards, retail accounts, instalment loans, finance company accounts, mortgage, etc.)
    - Presence of adverse public records (bankruptcy, judgements, suits, liens, wage attachments, etc.), collection items, and/or delinquency (past due items)
    - Severity of delinquency (how long past due)
    - Amount past due on delinquent accounts or collection items
    - Time since (recency of) past due items (delinquency), adverse public records (if any), or collection items (if any)
    - Number of past due items on file
    - Number of accounts paid as agreed
- **Amounts Owed**
    - Amount owing on accounts
    - Amount owing on specific types of accounts
    - Amount owing on specific accounts
    - Lack of a specific type of balance, in some cases
    - Number of accounts with balances
    - Proportion of credit lines used (proportion of balances to total credit limits on certain types of revolving accounts)

- Proportion of instalment loan amounts still owing (proportion of balance to original loan amount on certain types of instalment loans)

- **Length of Credit History**
  - Time since accounts opened
  - Time since accounts opened, by specific type of account
  - Time since account activity
- **New Credit**
  - Number of recently opened accounts, and proportion of accounts that are recently opened, by type of account
  - Number of recent credit inquiries
  - Time since recent account opening(s), by type of account
  - Time since credit inquiry(s)
  - Re-establishment of positive credit history following past payment problems
- **Types of Credit Used**
  - Number of (presence, prevalence, and recent information on) various types of accounts (credit cards, retail accounts, instalment loans, mortgage, consumer finance accounts, etc.)

B.3.2 Lam & Da Silva (1999)

Based on a research of Singapore's economy, lending and property markets, ad property laws, Standard & Poor's has established its preliminary rating criteria for Singapore residential mortgage securitization. This criterion is adapted solely for private mass residential properties, rather than luxury residential properties.

From research into parameters for mortgage underwriting and property characteristics, a benchmark pool as shown below was devised. The benchmark pool serves as a yardstick by

which the risk in any given mortgage pool may be measured. The benchmark pool is assumed to incur a specific maximum amount of credit losses appropriate for a rating category.

- **Pool size:** minimum of 300 loans.
- **Loan size:** maximum of S$600,000.
- **Loan-to-Value Ratio (LTV):** maximum of 80%.
- **Debt Servicing-to-Income Ratio:** total monthly debt repayment obligations may not exceed 30% of gross monthly income.
- **Loan type:** level pay, fully amortizing, variable and fixed rates.
- **Loan term:** maximum of 30 years.
- **Loan seasoning:** minimum of six payments made.
- **Loan performance:** not delinquent at the time of transfer and clean delinquency record over the previous 12 months.
- **Loan purpose:** purchase or refinance without equity release
- **Security:** first registered mortgage over property.
- **Land type and title:** Freehold land or crown leaseholds.
- **Property type:** Condominiums and apartments.
- **Property age:** less than 10 years old at the time of mortgage origination.
- **Occupancy status:** owner-occupied.
- **Geographic location/concentration:** properties located in reasonable proximity to mass transit and good schools are viewed more favourably.
- **Concentration limits by region (%)**
- **Borrower employment:** salaried employee or professional.
- **Borrower residency:** Singapore citizens or permanent resident individuals.

- **Property insurance:** fully insured for at least the replacement value of the property against fire.
- **Property valuation:** performed by a registered chartered surveyor.
- **Mortgage originators:** prudent and experienced mortgage lending banks.
- **Mortgage assignment:** mortgages assigned to special-purpose vehicle issuer by legal assignment.

C Some useful propositions

In this small section we will introduce some statements that are useful for the comparison of discriminatory power measures in section 2.6.

Proposition C.1

Let $S \mid Y = 0 \sim N(0,1)$ and $S \mid Y = 1 \sim N\left(\mu, \sigma^2\right)$, such that $\mu \geq 0$. Then we obtain the following expressions (in order to simplify, we will denote $z_1 = (s - \mu)/\sigma$):

1. The overlapping area criterion is given by:

   $$T_{pos} = \max_s \left\{\Phi(s) - \Phi(z_1)\right\},$$

   being the score value that maximizes T_{pos}:

   $$s = \frac{-\mu - \sqrt{\mu^2\sigma^2 - 2\sigma^2 \log(\sigma) + 2\sigma^2 \log(\sigma)}}{-1 + \sigma^2}, \text{ if } 0 < \sigma < 1 \text{ and}$$

   $$s = \frac{-\mu + \sqrt{\mu^2\sigma^2 - 2\sigma^2 \log(\sigma) + 2\sigma^2 \log(\sigma)}}{-1 + \sigma^2}, \text{ if } \sigma > 1.$$

2. For the accuracy ratio:

   $$AR = 1 - 2\int_{-\infty}^{\infty} \Phi(z_1) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}s^2} ds,$$

[6]we have the inequality:

$$2\min_{s}\left\{\Phi(s)-\Phi(z_1)\right\} \le AR \le 2T_{pos}$$

3. Given an optimal split point s, for the standardized maximal distance holds:

$$\begin{aligned} D_e = 1 - \frac{1}{\pi_1\log(\pi_1)+(1-\pi_1)\log(1-\pi_1)} \cdot \\ \cdot\left(\Phi(z_1)\pi_1\log\left(\frac{\Phi(z_1)\pi_1}{\Phi(z_1)\pi_1+\Phi(s)(1-\pi_1)}\right)\right. \\ +\Phi(s)(1-\pi_1)\log\left(\frac{\Phi(s)(1-\pi_1)}{\Phi(z_1)\pi_1+\Phi(s)(1-\pi_1)}\right) \\ +\left(1-\Phi(z_1)\right)\pi_1\log\left(\frac{\left(1-\Phi(z_1)\right)\pi_1}{\left(1-\Phi(z_1)\right)\pi_1+\left(1-\Phi(s)\right)(1-\pi_1)}\right) \\ \left.+\left(1-\Phi(s)\right)(1-\pi_1)\log\left(\frac{\left(1-\Phi(s)\right)(1-\pi_1)}{\left(1-\Phi(z_1)\right)\pi_1+\left(1-\Phi(s)\right)(1-\pi_1)}\right)\right). \end{aligned}$$

Proof:

1. The expression of T_{pos} follows from the definition (2.2.5). The optimal s is obtained by solving:

$$\frac{dT_{pos}}{ds} = \frac{e^{-\frac{s^2}{2}}}{\sqrt{2\pi}} - \frac{e^{-\frac{(s-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} = 0$$

2. We will calculate first the left side of the inequality and then the right side. Using the definition of AUC (2.3.3) and the relationship between AUC and AR (2.3.4):

   I. We have $AUC = 1 - \int_{-\infty}^{\infty} F_1(s)\,dF_0(s)$, being

$$\begin{aligned} \int_{-\infty}^{\infty} F_1(s)\,dF_0(s) &= \int_{-\infty}^{\infty}\left(F_1(s)-F_0(s)\right)dF_0(s) + \int_{-\infty}^{\infty} F_0(s)\,dF_0(s) \\ &\le \max_{s}\left\{F_1(s)-F_0(s)\right\}.\left(F_0(\infty)-F_0(-\infty)\right) + \left(\frac{\left(F_0(\infty)\right)^2}{2} - \frac{\left(F_0(-\infty)\right)^2}{2}\right) \end{aligned}$$

[6] This integral can only be computed numerically or otherwise approximated.

$$= \max_s \{F_1(s) - F_0(s)\} + \frac{1}{2}$$

$$AUC = 1 - \int_{-\infty}^{\infty} F_1(s)\, dF_0(s) \geq \frac{1}{2} - \max_s \{F_1(s) - F_0(s)\}$$

$$AR = 2AUC - 1 \geq 2\left(\frac{1}{2} - \max_s \{F_1(s) - F_0(s)\}\right) - 1 = 2\min_s \{F_0(s) - F_1(s)\}$$

II. Similarly, we obtain: $\int_{-\infty}^{\infty} F_1(s)\, dF_0(s) \geq \min_s \{F_1(s) - F_0(s)\} + \frac{1}{2}$

$$AUC \leq \frac{1}{2} - \min_s \{F_1(s) - F_0(s)\} = \frac{1}{2} + \max_s \{F_0(s) - F_1(s)\}$$

$$AR = 2AUC - 1 \leq 2\left(\frac{1}{2} + \max_s \{F_0(s) - F_1(s)\}\right) - 1 = 2\max_s \{F_0(s) - F_1(s)\}$$

3. Is a generalization of the expression given in Proposition 2.26 for D_e, if we have different standard deviations 1 and σ.

□

Proposition C.2

Given $S \mid Y = 0 \sim N(0, \sigma^2)$ and $S \mid Y = 1 \sim N(\mu, \sigma^2)$, such that $\mu \geq 0$. Then we can state (for ease of notation, we write $z = \mu / 2\sigma$, $z_0 = s/\sigma$, $z_1 = (s - \mu)/\sigma$):

1. For the overlapping area criterion:

   $$T_{pos} = 2\Phi(z) - 1$$

2. For the accuracy ratio we get:

   $$AR = 1 - 2\int_{-\infty}^{\infty} \Phi(z_1) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}z_0^2} ds,$$

   and the following inequalities

   $$AR \geq \max\left\{2\min_s \{\Phi(s) - \Phi(z_1)\},\ 2\Phi(z)^2 - 1\right\} \text{ and}$$

   $$AR \leq \min\left\{4\Phi(z) - 2,\ 4\Phi(z) - 2\Phi(z)^2 - 1\right\}$$

3. For the standardized maximal distance, having an optimal split point s:

$$D_e = 1 - \frac{1}{\pi_1 \log(\pi_1) + (1-\pi_1)\log(1-\pi_1)} \cdot$$
$$\cdot \left(\Phi(z_1)\pi_1 \log\left(\frac{\Phi(z_1)\pi_1}{\Phi(z_1)\pi_1 + \Phi(z_0)(1-\pi_1)} \right) \right.$$
$$+ \Phi(z_0)(1-\pi_1)\log\left(\frac{\Phi(z_0)(1-\pi_1)}{\Phi(z_1)\pi_1 + \Phi(z_0)(1-\pi_1)} \right)$$
$$+ \left(1-\Phi(z_1)\right)\pi_1 \log\left(\frac{\left(1-\Phi(z_1)\right)\pi_1}{\left(1-\Phi(z_1)\right)\pi_1 + \left(1-\Phi(z_0)\right)(1-\pi_1)} \right)$$
$$\left. + \left(1-\Phi(z_0)\right)(1-\pi_1)\log\left(\frac{\left(1-\Phi(z_0)\right)(1-\pi_1)}{\left(1-\Phi(z_1)\right)\pi_1 + \left(1-\Phi(z_0)\right)(1-\pi_1)} \right) \right).$$

Proof:

1. Note that for a normal distribution function, we have: $\Phi(z) = 1 - \Phi(-z)$. As for equal standard deviations, we have the optimal $s = \mu/2$, then by the definition (2.2.5) of T_{pos}:

$$T_{pos} = \max_s \left\{ F_0(s) - F_1(s) \right\} = F_0\left(\frac{\mu}{2}\right) - F_1\left(\frac{\mu}{2}\right)$$
$$= \Phi\left(\frac{\mu/2 - 0}{\sigma} \right) - \Phi\left(\frac{\mu/2 - \mu}{\sigma} \right) = \Phi\left(\frac{\mu}{2\sigma} \right) - \Phi\left(\frac{-\mu}{2\sigma} \right) = \Phi(z) - \left(1 - \Phi(z)\right).$$

2. By the definition of AUC (2.3.3) and the relationship (2.3.4), we proceed as follows:

   I. For the left side, we have

$$\int_{-\infty}^{\infty} F_1(s)\, dF_0(s) = \int_{-\infty}^{\frac{\mu}{2}} F_1(s)\, dF_0(s) + \int_{\frac{\mu}{2}}^{\infty} F_1(s)\, dF_0(s)$$
$$\le F_1\left(\frac{\mu}{2}\right) \int_{-\infty}^{\frac{\mu}{2}} f_0(s)\, ds + F_1(\infty) \int_{\frac{\mu}{2}}^{\infty} f_0(s)\, ds$$
$$= \Phi\left(\frac{-\mu}{2\sigma} \right) \left(F_0\left(\frac{\mu}{2}\right) - F_0(-\infty) \right) + 1 \cdot \left(F_0(\infty) - F_0\left(\frac{\mu}{2}\right) \right)$$
$$= \Phi\left(\frac{-\mu}{2\sigma} \right) \left(\Phi\left(\frac{\mu}{2\sigma} \right) - 0 \right) + 1 - \Phi\left(\frac{\mu}{2\sigma} \right) = \Phi(z)\left(\Phi(-z) - 1 \right) + 1 = 1 - \Phi(z)^2$$

   Thus, $AR = 2AUC - 1 \ge 2\left(1 - 1 + \Phi(z)^2\right) - 1 = 2\Phi(z)^2 - 1.$

Moreover, from Proposition C.1, we got

$$AR \geq 2 \min_{s} \left\{ \Phi(s) - \Phi(z_1) \right\}.$$

II. And for the right side,

$$\int_{-\infty}^{\frac{\mu}{2}} F_1(s) f_0(s) ds + \int_{\frac{\mu}{2}}^{\infty} F_1(s) f_0(s) ds \geq F_1\left(\frac{\mu}{2}\right) \int_{\frac{\mu}{2}}^{\infty} f_0(s) ds$$

$$= \Phi\left(\frac{-\mu}{2\sigma}\right)\left(F_0(\infty) - F_0\left(\frac{\mu}{2}\right)\right) = \left(1 - \Phi(z)\right)^2$$

Then we get

$AUC \leq 1 - \left(1 - \Phi(z)\right)^2 = 2\Phi(z) - \Phi(-z)^2$, and

$AR = 2AUC - 1 \leq 2\left(2\Phi(z) - \Phi(z)^2\right) - 1.$

And we know from Proposition C.1, that $AR \leq 2T_{pos} = 2\left(2\Phi(z) - 1\right)$.

3. It derives from the expression of D_e given in Proposition C.1 for equal standard deviations σ.

□

References

[1] Arellano, M. (2003). *Panel Data Econometrics*. Oxford: University Press.

[2] Barniv, R. (1997). Predicting the Outcome following Bankruptcy Filing: a Three-State Classification using Neural Networks. *International Journal of Intelligent Systems in Accounting, Finance and Management*, *6*, 177_194.

[3] Barron, J.M., & Staten, M.E. (2003). The Value of Comprehensive Credit Reports: Lessons from the U.S. Experience. In M. J. Miller (Ed.), *Credit Reporting Systems and the International Economy*. Cambridge: MIT Press, forthcoming.

[4] Basel Committee on Banking Supervision (2001, January). *The New Basel Capital Accord: Second Consultative Paper*. Basel, Switzerland: Bank for International Settlements. Retrieved March 3, 2003, from: http://www.bis.org/bcbs/bcbscp2.htm

[5] Basel Committee on Banking Supervision (2003, April). *The New Basel Capital Accord: Third Consultative Paper*. Basel, Switzerland: Bank for International Settlements. Retrieved September 17, 2005, from: http://www.bis.org/bcbs/bcbscp3.htm

[6] Basel Committee on Banking Supervision (2005, February). *Studies on the Validation of Internal Rating Systems: Working Paper No. 14*. Basel, Switzerland: Bank for International Settlements. Retrieved May 06, 2005, from: http://www.bis.org/publ/bcbs_wp14.htm

[7] Bemmann, M. (2005). Improving the Comparability of Insolvency Predictions. *Dresden Discussion Paper in Economics, 08 (05)*. Retrieved September 17, 2005, from DefaultRisk.com, website: http://www.defaultrisk.com/pp_score_52.htm

[8] Blochwitz, S., Hohl, S., Tasche, D. & Wehn, C.S. (2004, December). Validating Default Probabilities on Short Time Series, *2*. Retrieved September 09, 2005, from Federal Reserve Bank of Chicago, website: http://www.chicagofed.org/banking_information/capital_and_ market_risk_insights.cfm

[9] Blochwitz, S., Hohl, S. & Wehn, C.S. (2005, May). Reconsidering Ratings. *Wilmott Magazine*, p. 60-69.

[10] Bortz, J., & Döring, N. (1995). *Forschungsmethoden und Evaluation*. Berlin: Springer (in German).

[11] Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.

[12] Carey, M., & Hrycay, M. (2001). Parameterizing Credit Risk Models with Rating Data. *Journal of Banking and Finance, 25 (1)*, 197-270.

[13] Cox, D.R., & Hinkley, D.V (1974). *Theoretical Statistics*. London: Chapman and Hall.

[14] Cox, D.R., & Snell, E.J. (1989). *Analysis of Binary Data*. London: Chapman and Hall.

[15] D'Agostino, R.B., & Stephens, M.A. (1986). *Goodness-of-Fit Techniques*. New York: Dekker.

[16] Dinges, H., & Rost, H. (1982). *Prinzipien der Stochastik*. Stuttgart: Teubner (in German).

[17] Dionne, G., Artís, M., & Guillén, M. (1996). Count Data Models for a Credit Scoring System. *Journal of Empirical Finance, 3*, 303-325.

[18] Dobson, A. (1990). *An Introduction to Generalized Linear Models*. London: Chapman and Hall.

[19] Fahrmeir, L., & Hamerle, A. (1984). *Multivariate statistische Verfahren*. Berlin: Walter de Gruyter (in German).

[20] Fahrmeir, L., Hamerle, A., & Tutz, G. (1996). *Multivariate Statistische Verfahren*. Berlin: Walter de Gruyter (in German).

[21] Fair Isaac & Company, Inc. (n.d.). *What's in Your Score*. Retrieved January 11, 2006, from myFICO, website: http://www.myfico.com/CreditEducation/WhatsInYourScore.aspx

[22] Giese, G. (2002). Einführung von Internen Rating-Verfahren unter Basel II. *Der Schweizer Treuhänder, 76* (*H. 9)*, 803-810 (in German).

[23] Greene, W.H. (1993). *Econometric Analysis*. New Jersey: Prentice-Hall.

[24] Gourieroux, C. (2000). *Econometric of Qualitative Dependent Variables*. Cambridge: University Press.

[25] Gourieroux, C., & Jasiak, J. (2001). *Econometric Analysis of Individual Risks*. Retrieved January 15, 2006, from http://dept.econ.yorku.ca/~jasiakj

[26] Gourieroux, C., & Monfort, A. (1995). *Statistics and Econometric Models*. Cambridge: University Press.

[27] Hand, D.J. (2001). Modelling Consumer Credit Risk. *IMA Journal of Management Mathematics, 12,* 11139-155.

[28] Hand, D.J., & Henley, W.E. (1997). Statistical Classification Methods in Consumer Credit Scoring: A Review. *Journal of the Royal Statistical Society, A 160 (3),* 523-541.

[29] Härdle, W. (1991). *Smoothing Techniques: With implementation in S.* New York: Springer Verlag.

[30] Härdle, W., Müller, M., Sperlich, S., & Werwatz, A. (2004). *Nonparametric and Semiparametric Models.* Berlin: Springer.

[31] Henking, A. (2004, May 21). Simultane Validierung von Ausfallwahrscheinlichkeiten. *Dresdner Beiträge zu Quantitativen Verfahren, 38 (04)* (in German). Retrieved September, 12, 2005 from http://www.tu-dresden.de/wwqvs/f-db.htm

[32] Hsiao, C. (1990). *Analysis of Panel Data*. Econometric Society Monographs No. 11. London: Chapman and Hall.

[33] Huschens, S. (2004). Backtesting von Ausfallwahrscheinlichkeiten. *Dresdner Beiträge zu Quantitativen Verfahren, 40 (04)* (in German). Retrieved September, 12, 2005 from http:// www.tu-dresden.de/wwqvs/f-db.htm

[34] Jacobson, T., & Roszbach, K. (2003). Bank Lending Policy, Credit Scoring and Value at Risk. *Journal of Banking & Finance, 27,* 615-633.

[35] Kaiser, U., & Szczesny, A. (2001). Einfache Ökonometrische Verfahren für die Kreditrisikomessung: Logit- und Probit- Modelle. *University of Frankfurt Working Paper Series Finance & Accounting, 61* (in German).

[36] Khandani, B., Lozano, M., & Carty, L. (2001, November). *Moody´s RiskCalc™ for Private Companies: The German Model*. Retrieved March 3, 2003, from RISKCALC online http://riskcalc.moodysrms.com/us/research/crm/720431.pdf

[37] Kraft, H., Kroisandt, G., & Müller, M., (2004, June 29). *Redisigning Ratings: Assesing the Discriminatory Power of Credit Scores under Censoring*. Retrieved January 10, 2006, from Fraunhofer ITWM, website: http://www.itwm.fhg.de/fm/projects/rating/uKKM.pdf

[38] Lam, D., & Da Silva, M.L. (1999, June). *Preliminary Rating Criteria for Singapore Residential Mortgage-Backed Securities*. Retrieved March 3, 2003, from Standard and Poor´s: http://www2.standardandpoors.com/servlet/Satellite?pagename=sp/Page/FixedIncomeRatingsCriteriaPg&r=1&l=EN&b=2&s=21&ig=2&ft=24

[39] Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden-Day, Inc.

[40] Lienert, G.A. (1973). *Verteilungsfreiemethoden in der Biostatistik*. Meisenheim am Glan: Verlag Anton Hain (in German).

[41] McCullagh, P., & Nelder, J.A. (1983). *Generalized Linear Models*. London: Chapman and Hall.

[42] Miller, L.H. (1956). Table of the Percentage Points of the Kolmogorov Statistics. *Journal of the American Statistical Association*, 51, 113-115.

[43] Müller, M., & Härdle, W. (2002). Exploring Credit Data. In: Bol, G., Nakhaeizadeh, G., Rachev, S. T., Ridder, T., & Vollmer, K.-H. (eds.). *Credit Risk - Measurement, Evaluation and Management*. (Proceedings Ökonometrie-Workshop: Kreditrisiko - Messung, Bewertung und Management Universität Karlsruhe). Physica-Verlag.

[44] Piesch, W. (1975). *Statistische Konzentrationsmaße*. Tübingen: J. C. B. Mohr (in German).

[45] Shannon, C., & Weaver, W. (1969). *The Mathematical Theory of Communication*. Illinois: The University of Illinois Press.

[46] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

[47] Standard & Poor´s (2005). *Corporate Ratings Criteria*. Retrieved January 11, 2006, from Standard & Poor's, website: http://www2.standardandpoors.com/spf/pdf/fixedincome/CorporateRatings2005.pdf

[48] Tasche, D. (2003, May 2). A Traffic Lights Aproach to PD Validation. Retrieved September 15, 2005, from arXiv.org, website: http://arxiv.org/PS_cache/cond-mat/pdf/0305/0305038.pdf

[49] Theil, H. (1979). *Principles of Econometrics*. New York: Wiley.

[50] Tutz, G. (2000). *Die Analyse Kategorialer Daten*. München: Oldenbourg (in German).

[51] Vasicek, O. (2002). Loan Portfolio Value. *Risk*, 160-162.

[52] West, D. (2000). Neural Network Credit Scoring Models. *Computers & Operations Research*, *27*, 1131-1152.