Kraft, Holger; Kroisandt, Gerald; Müller, Marlene

Working Paper

Assessing the discriminatory power of credit scores

SFB 373 Discussion Paper, No. 2002,67

Provided in Cooperation with:
Collaborative Research Center 373: Quantification and Simulation of Economic Processes, Humboldt University Berlin

Suggested Citation: Kraft, Holger; Kroisandt, Gerald; Müller, Marlene (2002) : Assessing the discriminatory power of credit scores, SFB 373 Discussion Paper, No. 2002,67, Humboldt University of Berlin, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes, Berlin,
https://nbn-resolving.de/urn:nbn:de:kobv:11-10049306

This Version is available at:
https://hdl.handle.net/10419/65361

Assessing the Discriminatory Power of Credit Scores

Holger Kraft[1], Gerald Kroisandt[1], Marlene Müller[1,2]

[1] Fraunhofer Institut für Techno- und Wirtschaftsmathematik (ITWM)
Gottlieb-Daimler-Str. 49, 67663 Kaiserslautern, Germany

[2] Humboldt-Universität zu Berlin, Institut für Statistik & Ökonometrie
Spandauer Str. 1, 10178 Berlin, Germany

August 15, 2002

We discuss how to assess the performance for credit scores under the assumption that for credit data only a part of the defaults and non-defaults is observed. The paper introduces a criterion that is based on the difference of the score distributions under default and non-default. We show how to estimate bounds for this criterion, the Gini coefficient and the accuracy ratio.

Keywords:
credit rating, credit score, discriminatory power, sample selection, Gini coefficient, accuracy ratio

JEL Classification: G21

1 Introduction

A bank which wants to decide whether a credit applicant will get a credit or not has to assess if the applicant will be able to redeem the credit. Among other criteria, the bank requires an estimate of the probability that the applicant will default prior to the maturity of the credit. At this step, a rating of the applicant is a valuable decision support. The idea of a rating system is to identify criteria which separate the "good" from the "bad" creditors, as for example liquidity ratios or ratios concerning the capital structure of a firm. In a more formal sense a rating corresponds to a guess of the default probability of the credit. Obviously, the question arises how a bank can identify a sufficient number of selective criteria and, especially, what selectivity and discriminatory power means in this context. In the following sections we try to make a first step to a rigorous treatment of this subject which is rarely addressed in literature.

Apart from the theoretical attractiveness this issue is of highly practical importance. This is due to the fact that the Basel Committee on Banking Supervision is working on a New Capital Accord (Basel II) where default risk adjusted capital requirements shall be established. In this context ratings and the design of ratings play an important role. Clearly, the committee wants the banks to identify factors which "have an ability to differentiate risk [and] have predictive and discriminatory power" (Banking Committee on Banking Supervision, 2001, p. 50). Unfortunately, they do not give any formal definition of "predictive" or "discriminatory power".

The paper is organized as follows: In Section 2 we discuss how to measure discriminatory power of a score (a numerical value that reflects the rating of a credit applicant). We introduce a criterion that is based on the difference between the distributions of the score conditioned on default or non-default and is simple to compute. Section 3 discusses the consequences of the typical censoring in credit data due to the fact that not all credit applicants are accepted. This implies that we do have default or non-default information only for a restricted set of applicants. To keep things simple we first discuss discriminatory power using a parametric setting. In Section 4 we consider the nonparametric case and show how to find lower and upper bounds for the proposed criterion. Finally, Section 5 extends our approach to lower and upper bounds for the Gini coefficient and the accuracy ratio (AR).

2 Discriminatory Power of a Score

Let us start with the following classification problem: Consider random variables $X_1, \ldots, X_p$ and a group indicator $Y \in \{0, 1\}$. A score S (used to rate applicants

for a loan) is an aggregation of the variables $X_1, \ldots, X_p$ into a single number. Hence, we can consider any real valued function $S(X_1, \ldots, X_p)$ to be a score. For the sake of brevity we will use S to denote the random variable $S(X_1, \ldots, X_p)$. In the following we will only study the relation between S and Y.

There exists a variety of criteria to assess the quality of a score. A reasonable score function for credit rating should assign higher score values to credit applicants who have higher *probabilities of default* (PDs). Therefore the capability to separate the two groups of observations corresponding to $Y = 1$ (default) and $Y = 0$ (non-default) is a basic feature of a credit score function. A measure for the *discriminatory power* can consequently be used as a performance measure for a credit score.

A straightforward approach to assess discriminatory power is the comparison of the conditional distributions of S given default or non-default. We will first focus on the "difference" of these two conditional distributions. The methodology that is derived here can however be used for other measures of performance as well.

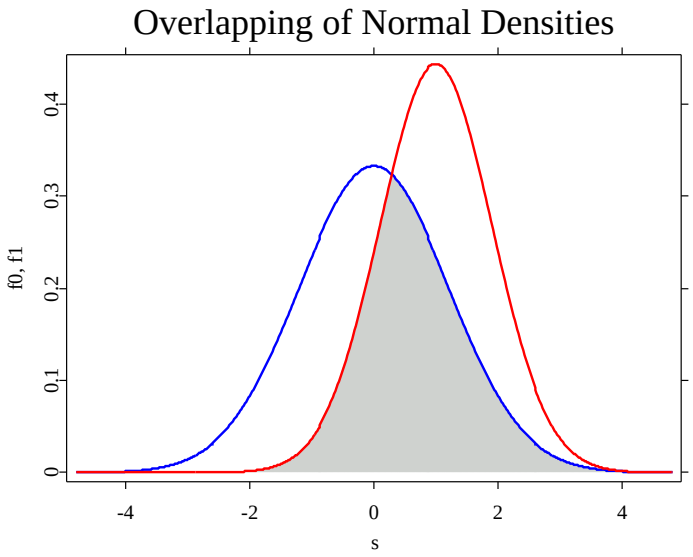

Figure 1: Overlapping area U for two normal densities

In the case of a normal distribution the conditional densities of S given $Y = j$, $j = 0, 1$ are easy to visualize and to compute. Denote f_0, f_1 the probability densities of $S|Y = 0$ and $S|Y = 1$, further F_0, F_1 their cumulative distribution functions. Consider first the special case that f_0 and f_1 have exactly one point of intersection, cf. Figure 1. (A condition for this property will be given in a moment.) Let s be the horizontal coordinate of this intersection. Assuming 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. We suppose (w.l.o.g.) in the

following that $\mu_1 > \mu_0$. Then the region of overlapping U for the two densities can be calculated as

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

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

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

For different standard deviations ($\sigma_0 \neq \sigma_1$), there may be one or two points of intersection (as in quadratic discriminance analysis) and the horizontal coordinates are determined by $f_0(s) = f_1(s)$ i.e. as solutions of the quadratic equation

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

The definition of U can be easily generalized to the nonparametric case when no distributional assumption for S is made:

$$U = \int \min\{f_0(s), f_1(s)\}\, ds\,. \tag{2}$$

This definition allows any number of intersection points of f_0 and f_1. Alternatively, assuming a monotone relationship between the score S and the default probability, a variant of the definition can be given by

$$U = \min_s \{F_1(s) + 1 - F_0(s)\}\ . \tag{3}$$

This definition is based on the idea that only one optimal intersection point should exist in this case. As for the normal case, we assume that f_1 is right of f_0. An analogous definition could be formulated for a monotone decreasing relationship.

It is obvious that for densities f_0, f_1 on completely different supports (perfect separation) the region of overlapping U is zero. If both densities are identical (no separation) then U equals one. In all other cases U will take on values between 0 and 1. An indicator of discriminatory power is now given by

$$T = 1 - U. \tag{4}$$

As U, the discriminatory power indicator T takes on values in the interval $[0, 1]$.

In practice we have observations $S^{(i)}$ for the scores and $Y^{(i)}$ for the groups (defaults and non-defaults in credit scoring). Under the assumption of a normal distribution U (and hence T) can be computed using the empirical moments $\widehat{\mu}_0$, $\widehat{\mu}_1$, $\widehat{\sigma}_0$, and $\widehat{\sigma}_1$.

Under more general assumptions on the distribution, U and T can be computed for example by nonparametric estimates of the densities (histograms, kernel density estimators). In the monotone case it is sufficient to have nonparametric

estimates of the cumulative distribution functions F_0, F_1. Those estimates can be easily found by the empirical distribution functions

$$\widehat{F}_j(s) = \frac{\sum_i I(S^{(i)} \leq s, Y^{(i)} = j)}{\sum_i I(Y^{(i)} = j)}, \quad j = 0, 1\,. \tag{5}$$

We remark that the distribution of T is related to the Kolmogorov-Smirnov test statistics, which checks the hypothesis $F_0 = F_1$. Hence, this test can be applied to find out if the score influences the PD at all.

3 Credit Scoring & Unobservable Areas

Consider now a sample of n credit applicants, for which a set of variables is given (e.g. age of the applicant, amount and duration of the loan, income etc.). As above we assume that a real valued score S is calculated from these variables at time $t = 0$ and the default ($Y = 1$) or non-default ($Y = 1$) is observed at time $t = 1$.

The particular problem of credit scoring is that we observe defaults and non-defaults only for a subsample of applicants. In more detail, this means that the bank computes scores for N applicants but only n of them ($n < N$) are accepted for a loan. Hence, default and non-default observations are preselected by a condition, which we denote by $\mathcal{A}$. This type of sample preselection is usually described as *censoring* or *sample selection*.

The problem of sample selection has been mainly studied in the (econometric) literature with a focus on the estimating regression coefficients and PDs. Greene (1998) for example uses a Heckman two-step procedure (see Heckman, 1979) for estimating probit and count data models for credit data. Gourieroux and Jasiak (2001, Ch. 7) consider maximum-likelihood probit and a Bayesian approach whereas a rather general approach is introduced by Horowitz and Manski (1998).

We consider the problem of estimating discriminatory power in the censored case under very general distributional assumptions. This will lead to upper and lower bounds for the performance criteria rather than classical point estimates.

To illustrate the effect of censoring (or sample selection) for estimating U and T assume again that both densities f_0, f_1 have exactly one intersection point. Assume also that the censoring condition is

$$\mathcal{A} = \{S \leq c\}\,, \tag{6}$$

where c is a threshold such that no credit applicants are accepted for a loan when their score S is larger than c. Figure 2 shows this modified situation in comparison to Figure 1. The distribution right to the black line (here $c = 2$)

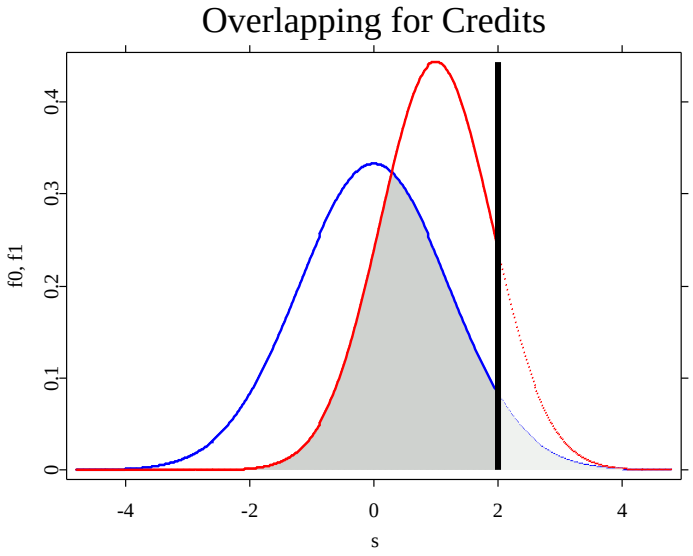

Figure 2: Truncated overlapping area for credit data

cannot be observed but needs in fact to be considered for a correct assessment of the performance of the score.

Denote $\widetilde{S} = (S|\mathcal{A})$ und $\widetilde{Y} = (Y|\mathcal{A})$ the observed part of the score and the group variable. Hence, we have only observations for $\widetilde{S}_j = (\widetilde{S}|\widetilde{Y} = j)$, $j = 0, 1$ while we are interested in $S_j = (S|Y = j)$. Under the assumption (6), the relation between $\widetilde{S}_j$ and S_j is given by

$$\begin{aligned} P(\widetilde{S}_j \le s) &= \frac{P(\widetilde{S} \le s, \widetilde{Y} = j)}{P(\widetilde{Y} = j)} = \frac{P(S \le s, Y = j|\mathcal{A})}{P(Y = j|\mathcal{A})} \\ &= \frac{P(S \le s, Y = j)}{P(S \le c, Y = j)} \qquad \text{if } s \le c. \end{aligned}$$

Since $P(S_j \le s) = P(S \le s|Y = j) = P(S \le s, Y = j)/P(Y = j)$ it follows that

$$P(\widetilde{S}_j \le s) = \frac{P(S_j \le s)\, P(Y = j)}{P(S \le c, Y = j)} = \frac{P(S_j \le s)}{P(S_j \le c)},$$

which shows

$$\widetilde{F}_j(s) = \frac{F_j(s)}{F_j(c)}. \tag{7}$$

Here $\widetilde{F}_j$ denotes the cumulative distribution functions of $\widetilde{S}_j$. Under the assumption that S_j has a continuous distribution, (7) results in an equivalent rescaling of the densities by $F_j(c)$. These densities and their region of overlapping $\widetilde{U}$ for

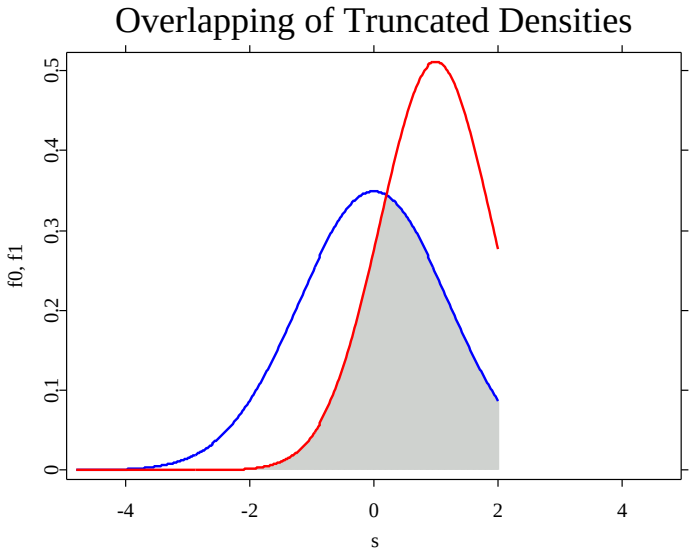

Figure 3: Observed overlapping area $\widetilde{U}$

the normal case are shown in Figure 3. Note the difference to Figure 2 on the vertical scale, since $\widetilde{f}_j(s) \geq f_j(s)$.

We will now examine the difference between $\widetilde{U}$ and U, the regions of overlapping for the censored (observed) and the non-censored (partially unobserved) sample. In the following we will consider the monotone version of the overlapping region:

$$U = \min_s \{F_1(s) + 1 - F_0(s)\}.$$

Computing the overlapping region $\widetilde{U}$ in the same way and using (7), would hence give

$$\widetilde{U} = \min_s \left\{ \widetilde{F}_1(s) + 1 - \widetilde{F}_0(s) \right\} = \min_s \left\{ \frac{F_1(s)}{F_1(c)} + 1 - \frac{F_0(s)}{F_0(c)} \right\}. \tag{8}$$

This shows that the naive calculation of the overlapping from incompletely observed data is usually different (biased) from the objective overlapping region U.

The difference in $\widetilde{U}$ and U (or T and $\widetilde{T}$) can be considerably important as a small Monte Carlo simulation shows. We have simulated 100 data sets, each of $N = 500$ observations. The scores $S^{(i)}$ are generated only once and come from a normal distribution with expectation -3 and variance 2.25. The simulated PDs are obtained from a Logit model, i.e.

$$p(s) = \frac{1}{1 + \exp(-s)}$$

and the $Y^{(i)}$ are Bernoulli random variables with probability parameter $p(S^{(i)})$. The threshold is chosen as $c = -0.5$, this gives here $n = 483$.

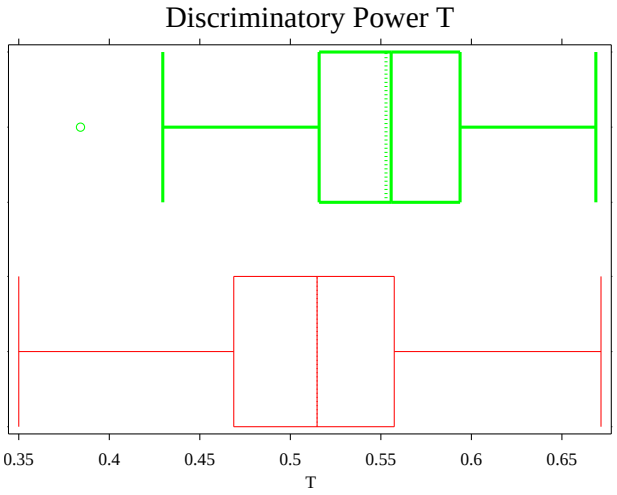

Figure 4: Boxplots for T (upper) and $\widetilde{T}$ (lower)

Figure 4 shows boxplots for the realized distributions of the estimated $\widetilde{T} = 1 - \widetilde{U}$ (lower boxplot) and $T = 1 - U$ (upper boxplot). The graphic shows that in our simulated example $\widetilde{T}$ is typically smaller than T. In particular, both mean and median of the 100 estimated Ts are as large as the upper quartile of the estimated $\widetilde{T}$s. A closer inspection of the data shows that in 95 cases $\widehat{\widetilde{T}} < \widehat{T}$ and in 5 cases $\widehat{\widetilde{T}} > \widehat{T}$. So using $\widetilde{T}$ at the place of T can mislead in assessing the performance of the score in both directions (over- and under-estimation).

Under the assumption that the types of the distributions of S_j are known a correction for $\widetilde{U}$ can be easily calculated. Let us outline this for the example of normal distributions: Here the moments of $\widetilde{S}_j$, $j = 0, 1$, can be calculated (Greene, 1993, Theorem 22.2) by

$$E(S_j | S_j \leq c) = \mu_j + \sigma_j \lambda(\alpha_j), \tag{9}$$

$$Var(S_j | S_j \leq c) = \sigma_j^2 [1 - \lambda(\alpha_j)\{\lambda(\alpha_j) - \alpha_j\}], \tag{10}$$

with μ_j and σ_j denoting the moments of the unconditional distributions,

$$\alpha_j = \frac{c - \mu_j}{\sigma_j} \quad \text{and} \quad \lambda(\alpha) = \frac{-\phi(\alpha)}{\Phi(\alpha)}$$

denoting the inverse Mills ratio. The expectations $\mu_j = E(S_j)$ and variances $\sigma_j^2 = Var(S_j)$ can hence be calculated from the credit data using the empirical moments of $\widetilde{S}_j$ and by solving the system of equations (9)–(10). Estimates of f_j and F_j are then obtained by plugging $\widehat{\mu}_j$, $\widehat{\sigma}_j$ into the density and cumulative distribution function of the normal distribution.

We remark that this idea can be generalized to any monotone transformation of the normal distribution. For example, many variables used for credit scoring have a skewed distribution. This typically transfers to scores which are linearly weighted sums of these variables. The log-normal distribution, which can model such a skewed score, has a direct relation to the normal distribution: Assume S_j is log-normal with parameters μ_j, σ_j, then for the log-score

$$\log(S_j) \sim N(\mu_j, \sigma_j^2). \tag{11}$$

Since the logarithm is monotone

$$F_j(s) = P(S_j \leq s) = P(\log(S_j) \leq \log(s)). \tag{12}$$

The computation for log-normal scores is therefore completely determined by the normal case. An even wider class of distributions is covered by using any monotone distribution as e.g a Box–Cox transformation.

A correction of $\widetilde{U}$ is also possible if the censoring is determined by another score function $S^\star$, i.e.

$$\mathcal{A} = \{S^\star \leq c\}. \tag{13}$$

This is a more realistic assumption since in practice $S^\star$ can be considered as the score function from a previous credit rating system. If the credit rating system is redesigned, the performance of the new score function S needs to be assessed. Under the very restrictive assumption of a joint normal distribution of S_j and $S_j^\star$ with moments μ_j, σ_j, $\mu_j^\star$, $\sigma_j^\star$ and correlation ρ_j it is known that

$$E(S_j|S_j^\star < c) = \mu_j + \rho_j\sigma_j\lambda(\alpha_j), \tag{14}$$

$$Var(S_j|S_j^\star < c) = (\sigma_j)^2[1 - \rho_j^2\lambda(\alpha_j)\{\lambda(\alpha_j) - \alpha_j\}], \tag{15}$$

see e.g. Greene (1993, Theorem 22.4). Here

$$\alpha_j = \frac{c - \mu_j^\star}{\sigma_j^\star}$$

and λ denotes the inverse Mills ratio as before. In addition we have

$$\widetilde{F}_j^\star(x) = \Phi_2\left(\frac{x - \mu_j^\star}{\sigma_j^\star}, \frac{c - \mu_j}{\sigma_j}, \rho_j\right)\left\{\Phi\left(\frac{c - \mu_j}{\sigma_j}\right)\right\}^{-1}. \tag{16}$$

The moments of $S_j^\star$ could be estimated from equations analogous to (9)–(10). With these estimates for $\mu_j^\star$, $\sigma_j^\star$, the system of equations (14)–(16) could be used to find estimates of the unconditional moments μ_j, σ_j and ρ_j.

This technique could again be generalized to monotone transformations as the logarithm or the Box-Cox transformation. However, apart from the restrictive distributional assumptions this approach requires that observations for both score functions $S^\star$ and S given $\mathcal{A} = \{S^\star \leq c\}$ are available.

4 Inequalities for the Nonparametric Case

As we have seen in Section 3, the computation of U from $\widetilde{S}_j$ requires specific assumptions on the distributions of S_j and their relations to the censoring condition $\mathcal{A}$. In the case of completely unknown distributions there is no possibility to estimate these distributions beyond $\mathcal{A}$. This is a relevant problem when a bank redesigns its credit rating system, since data on rejected applicants are normally not available.

A possible remedy to this problem is the calculation of upper and lower bounds for the discriminatory power T. The general assumption throughout this section is that we know the percentage of rejected loans, i.e. the full number of credit applicants. Denote this number of all credits (accepted or rejected) by N. Under the assumption that the percentages of both rejected applicants and defaults are small, relatively narrow bounds can be found for T. We want to stress that N typically does not contain applicants who are rejected without being rated.

Recall that the computation of U requires the cumulative distribution functions $F_j(s)$ of $S_j = (S|Y = j)$. However, we only observe $\widetilde{F}_j(s)$, the cumulative distribution function of $\widetilde{S}_j = (S|Y = j, \mathcal{A})$. Therefore we consider now the relation between $F_j(s)$ and $\widetilde{F}_j(s)$ in this general case. We have

$$\begin{aligned} F_j(s) &= P(S \le s|Y = j) \\ &= P(S \le s, \mathcal{A}|Y = j) + P(S \le s, \overline{\mathcal{A}}|Y = j) \\ &= P(S \le s|\mathcal{A}, Y = j)P(\mathcal{A}|Y = j) + P(S \le s, \overline{\mathcal{A}}|Y = j), \end{aligned}$$

hence

$$F_j(s) = \widetilde{F}_j(s)\, P(\mathcal{A}|Y = j) + P(S \le s, \overline{\mathcal{A}}|Y = j) \tag{17}$$

where $\overline{\mathcal{A}}$ denotes the complement of $\mathcal{A}$. We find an upper bound for $F_j(s)$ by using that $\{S \le s\} \cap \overline{\mathcal{A}} \subseteq \overline{\mathcal{A}}$ in the second term of (17), i.e.

$$\begin{aligned} F_j(s) &\le \widetilde{F}_j(s)P(\mathcal{A}|Y = j) + P(\overline{\mathcal{A}}|Y = j) \\ &= 1 - P(\mathcal{A}|Y = j)\{1 - \widetilde{F}_j(s)\}. \end{aligned} \tag{18}$$

A lower bound for $F_j(s)$ is given by omitting the second term of (17) completely, such that

$$F_j(s) \ge \widetilde{F}_j(s)P(\mathcal{A}|Y = j). \tag{19}$$

Both inequalities (18) and (19) involve $P(\mathcal{A}|Y = j)$ which can not be directly estimated, since the distribution of Y in $\mathcal{A}$ is unknown. However, we can describe the range of $P(\mathcal{A}|Y = j)$.

We start with a first approximation. Let us introduce the notation

$$\alpha_j = P(\mathcal{A}|Y = j),$$

such that (18) and (19) can be written as

$$\alpha_j \widetilde{F}_j(s) \leq F_j(s) \leq 1 - \alpha_j + \alpha_j \widetilde{F}_j(s). \tag{20}$$

From $P(Y = j) = P(\mathcal{A}, Y = j) + P(\overline{\mathcal{A}}, Y = j)$ we conclude that

$$P(\mathcal{A}, Y = j) \leq P(Y = j) \leq P(\mathcal{A}, Y = j) + P(\overline{\mathcal{A}}). \tag{21}$$

Thus from

$$\alpha_j = P(\mathcal{A}|Y = j) = \frac{P(Y = j|\mathcal{A})P(\mathcal{A})}{P(Y = j)} = \frac{P(\widetilde{Y} = j)P(\mathcal{A})}{P(Y = j)}$$

it follows that

$$\alpha_j^\star \leq \alpha_j \leq 1, \qquad \text{where } \alpha_j^\star = \frac{P(\widetilde{Y} = j)P(\mathcal{A})}{P(\widetilde{Y} = j)P(\mathcal{A}) + P(\overline{\mathcal{A}})}\,. \tag{22}$$

Equation (20) together with (22) yields

$$\begin{aligned} &\alpha_1^\star\, \widetilde{F}_1(s) + \alpha_0^\star\{1 - \widetilde{F}_0(s)\} \\ &\qquad \leq F_1(s) + 1 - F_0(s) \leq 2 - \alpha_1^\star\,\{1 - \widetilde{F}_1(s)\} - \alpha_0^\star\,\widetilde{F}_0(s)\,. \end{aligned} \tag{23}$$

As a consequence we obtain upper and lower bounds for the discriminatory power indicator $T = 1 - U = 1 - \min\{F_1(s) + 1 - F_0(s)\}$ which are given by

$$\begin{aligned} &1 - \min_s \left[2 - \alpha_1^\star\,\{1 - \widetilde{F}_1(s)\} - \alpha_0^\star\,\widetilde{F}_0(s)\right] \\ &\qquad \leq T \leq 1 - \min_s \left[\alpha_1^\star\,\widetilde{F}_1(s) + \alpha_0^\star\{1 - \widetilde{F}_0(s)\}\right]\,. \end{aligned} \tag{24}$$

We want to stress that in the special case where all credit applicants are accepted we have $\mathcal{A} = \Omega$ and $\alpha_0 = \alpha_1 = 1$. As a consequence (24) reduces to

$$T = 1 - \min_s \{F_1(s) + 1 - \ F_0(s)\}\ ,$$

which is exactly the definition introduced in Section 2.

More sophisticated bounds for $F_1(s) + 1 - F_0(s)$ can be obtained as follows. We use the additional abbreviations

$$\beta_j = P(\mathcal{A}, Y = j), \quad p_j = P(Y = j),$$

such that

$$\alpha_j = \frac{\beta_j}{p_j}.$$

Consider the lower bound first. From (18) and (19) we have

$$\begin{aligned} F_1(s) + 1 - F_0(s) &\geq \alpha_1 \widetilde{F}_1(s) + \alpha_0\{1 - \widetilde{F}_0(s)\} \\ &= \frac{\beta_1}{1-p_0}\widetilde{F}_1(s) + \frac{\beta_0}{p_0}\{1 - \widetilde{F}_0(s)\} \end{aligned} \tag{25}$$

In the last term every probability can be estimated from the observed data except for p_0. Hence, for given s the last term has to be minimized with respect to p_0. For this minimization one has to consider the three cases $\beta_1\widetilde{F}_1(s) = \beta_0\{1-\widetilde{F}_0(s)\}$, $\beta_1\widetilde{F}_1(s) > \beta_0\{1-\widetilde{F}_0(s)\}$, and $\beta_1\widetilde{F}_1(s) < \beta_0\{1-\widetilde{F}_0(s)\}$, which all lead to the same result:

$$p_0^\star = \begin{cases} \beta_0 & \text{if } \gamma_s < \beta_0, \\ \beta_0 + P(\overline{\mathcal{A}}) & \text{if } \gamma_s > \beta_0 + P(\overline{\mathcal{A}}), \\ \gamma_s, & \text{otherwise,} \end{cases} \tag{26}$$

and

$$\gamma_s = \frac{\sqrt{\beta_0\{1 - \widetilde{F}_0(s)\}}}{\sqrt{\beta_0\{1 - \widetilde{F}_0(s)\}} + \sqrt{\beta_1\widetilde{F}_1(s)}}\,. \tag{27}$$

The upper and lower thresholds in (26) are consequences of the bounds in (21).

To derive an upper bound of $F_1(s) + 1 - F_0(s)$ we conclude from (18) and (19)

$$\begin{aligned} F_1(s) + 1 - F_0(s) &\leq 2 - \alpha_1\{1 - \widetilde{F}_1(s)\} - \alpha_0\widetilde{F}_0(s) \\ &= 2 - \frac{\beta_1}{1-p_0}\{1 - \widetilde{F}_1(s)\} - \frac{\beta_0}{p_0}\widetilde{F}_0(s). \end{aligned} \tag{28}$$

Maximization of the last term with respect to p_0 leads to a similar result as before:

$$p_0^\bullet = \begin{cases} \beta_0 & \text{if } \delta_s < \beta_0, \\ \beta_0 + P(\overline{\mathcal{A}}) & \text{if } \delta_s > \beta_0 + P(\overline{\mathcal{A}}), \\ \delta_s, & \text{otherwise,} \end{cases} \tag{29}$$

and

$$\delta_s = \frac{\sqrt{\beta_0\widetilde{F}_0(s)}}{\sqrt{\beta_0\widetilde{F}_0(s)} + \sqrt{\beta_1\{1 - \widetilde{F}_1(s)\}}}\,. \tag{30}$$

Combining the results we obtain

$$\begin{aligned} &\frac{\beta_1}{1-p_0^\star}\,\widetilde{F}_1(s) + \frac{\beta_0}{p_0^\star}\{1 - \widetilde{F}_0(s)\} \\ &\qquad \leq F_1(s) + 1 - F_0(s) \leq 2 - \frac{\beta_1}{1-p_0^\bullet}\,\{1 - \widetilde{F}_1(s)\} - \frac{\beta_0}{p_0^\bullet}\,\widetilde{F}_0(s) \end{aligned} \tag{31}$$

and as in (24)

$$1 - \min_s \left[2 - \frac{\beta_1}{1-p_0^\bullet} \{1 - \widetilde{F}_1(s)\} - \frac{\beta_0}{p_0^\bullet} \widetilde{F}_0(s) \right] \\ \leq T \leq 1 - \min_s \left[\frac{\beta_1}{1-p_0^\star} \widetilde{F}_1(s) + \frac{\beta_0}{p_0^\star} \{1 - \widetilde{F}_0(s)\} \right] . \tag{32}$$

All quantities in the inequalities (24) and (32) can be estimated. For the observed scores under default and non-default we have their empirical distribution functions as in (5). To estimate $\alpha_j^\star$, β_j, $p_0^\star$ and $p_0^\bullet$ we consider the probabilities of the events $\{\widetilde{Y} = j\}$, $\mathcal{A}$ and $\overline{\mathcal{A}}$ which can be estimated by their observed relative frequencies

$$\widehat{P}(\widetilde{Y} = j) = \frac{n_j}{n}\,, \quad \widehat{P}(\mathcal{A}) = \frac{n}{N}\,, \quad \widehat{P}(\overline{\mathcal{A}}) = \frac{N-n}{N}\,. \tag{33}$$

Here n_0 denotes the number of observed non-defaults ($Y^{(i)} = 0$) and n_1 the number of observed defaults ($Y^{(i)} = 1$). As before we use n for the sample size of the observed credits (i.e. $n = n_0 + n_1$), and N for the number of all the credits. This gives the estimates

$$\widehat{\alpha}_j^\star = \frac{n_j}{n_j + N - n}, \quad \widehat{\beta}_j = \frac{n_j}{N}. \tag{34}$$

Estimates for $p_0^\star$ and $p_0^\bullet$ can be found by plugging $\widehat{\beta}_j$, $\widehat{P}(\overline{\mathcal{A}})$ and $\widehat{\widetilde{F}}_j(s)$ into (27) and (30).

As before we use a Monte Carlo simulation to illustrate the effect of these estimated bounds. The construction of the simulated data set is as above with one modification: We use scores $S^{(i)}$ with a variance of 1.44. This yields a value of $n = 491$ for the sample size of the observable scores. We find $\widehat{T} > \widehat{\widetilde{T}}$ in 91 cases and $\widehat{T} < \widehat{\widetilde{T}}$ in 9 cases.

Figure 5 shows estimates for T (thick solid line), $\widetilde{T}$ (thin solid line) and the estimated upper and lower bounds according to (32) for all 100 simulated data sets (sorted by the estimated Ts). The bounds according to (24) are wider but of very similar size, such that we omit them here. Recall that in practice the estimation of $\widehat{T}$ could not have been carried out, this is only possible here for simulated data. The simulation shows in particular, that in the mentioned 9 cases $\widetilde{T}$ as a replacement of T would have led to a too large value for the discriminatory power of the score. The upper and lower bounds however (which cover both $\widetilde{T}$ and T) indicate a correctly specified range for $\widehat{T}$.

We remark that the lower bound in Figure 5 seems to be quite far away from both $\widehat{T}$ and $\widehat{\widetilde{T}}$. This is a consequence of the fact that this bound does not require

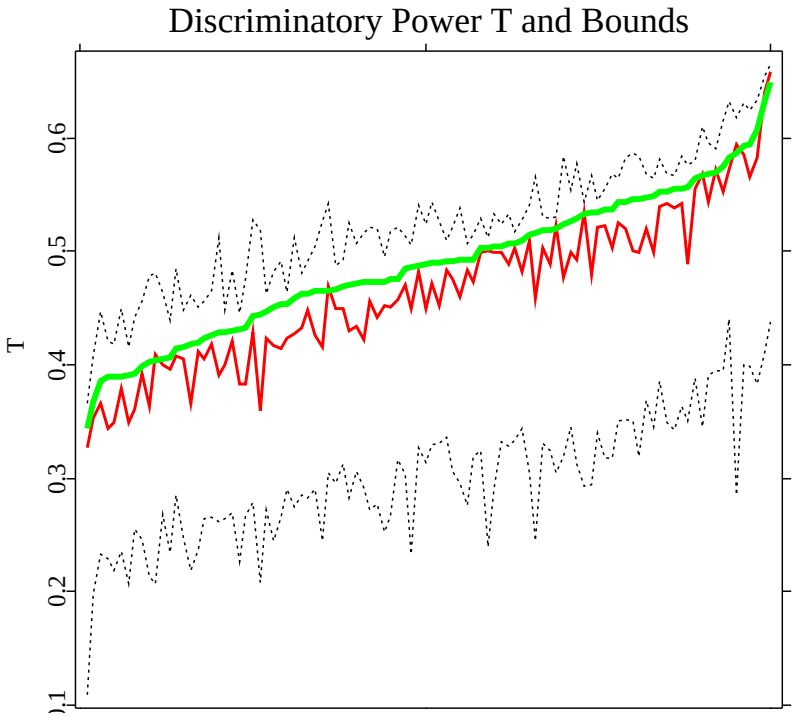

Figure 5: Estimated T (thick solid), $\widetilde{T}$ (solid) and bounds (dashed)

any information about the structure of the censoring condition $\mathcal{A}$. This bound could be considerably improved if additional information as e.g. $\mathcal{A} = \{S \leq c\}$ is used.

5 Gini coefficient and Accuracy Ratio

An alternative and frequently used measure for the performance of a score is the accuracy ratio AR which is based on the Lorenz curve and its Gini coefficient. In the case of censored data, the accuracy ratio computed from the observed part of the data is biased as well. As for T we can estimate bounds for the AR if the distribution of the score is unknown.

Let us first introduce the relevant terms. The Lorenz curve visualizes scores by means of comparing the distributions of S_1 and S. Figure 6 shows the principle of the Lorenz curve. On the horizontal and vertical scales, the percentages of applicants are sorted from high to low scores. The Lorenz curve is also known as selection curve. Variants of the Lorenz curve are the receiver operating characteristic (ROC) curve (Hand and Henley, 1997) and the performance curve (Gourieroux and Jasiak, 2001, Ch. 4).

To operate with cumulative distribution functions denote the negative score by

$$V = -S.$$

The Lorenz curve of S is then defined by the coordinates

$$\{L_1(v), L_2(v)\} = \{P(V < v), P(V < v|Y = 1)\} \ , \quad v \in (-\infty, \infty) \, .$$

Since $P(V < v) = 1 - F(-v)$, this is equivalent to

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

A estimate of the Lorenz curve can be computed by means of the empirical cumulative distribution functions $\widehat{F}$ and $\widehat{F}_1$.

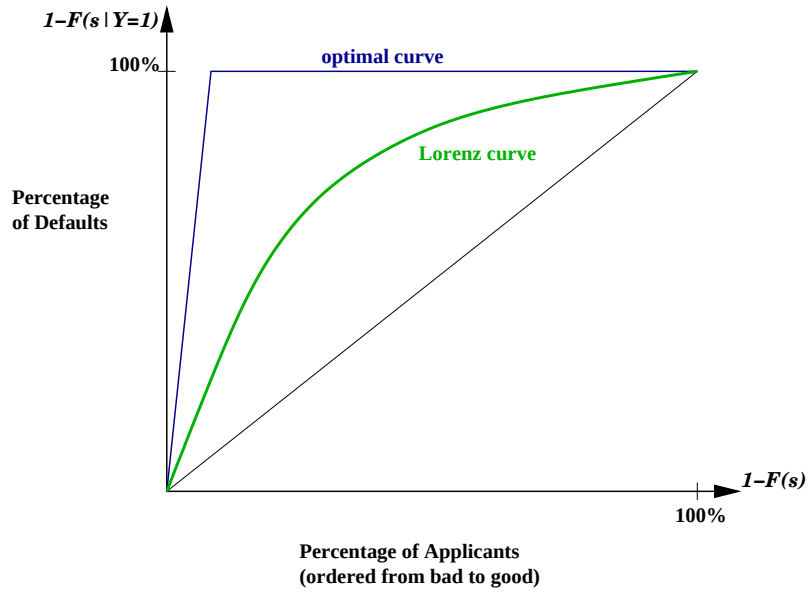

Figure 6: Lorenz curve for Credit Scoring

Recall that scores should assign higher score values to credit applicants with higher PDs. Such a credit score is obviously good if all vertical coordinates of the Lorenz curve are large. The best (optimal) Lorenz curve corresponds to a score that exactly separates defaults and non-defaults. This optimal curve reaches the vertical 100% at a horizontal percentage of $P(Y = 1)$, the probability of default. A random assignment of credit applicants to score values corresponds to a Lorenz curve identical to the diagonal.

Lorenz curves can also be used to compare different score functions. Better scores are more close to the optimal Lorenz curve. A quantitative measure for

the performance of a score is based on the area between the Lorenz curve and the diagonal. The *Gini coefficient* G denotes twice this area, i.e.

$$G = 2\int_0^1 \{1 - F_1(H(z))\}\, dz - 1 = 1 - 2\int_0^1 F_1(H(z))\, dz \tag{35}$$

where H is the inverse of $1 - F$. In practice the latter integral is estimated by numeric integration of $\widehat{F}_1$ over the range of $\widehat{F}$.

To compare different scores, their *accuracy ratios* AR are defined by relating the Gini coefficient of each score to the Gini coefficient of the optimal Lorenz curve. The accuracy ratio is hence defined as

$$AR = \frac{G}{G_{opt}} = \frac{G}{P(Y=0)}\,.$$

In the censored case we would compute $\widetilde{G}$ and $\widetilde{AR}$ instead of G and AR. Note that as for $\widetilde{T}$ and T the Gini coefficients and accuracy ratios are biased. We will now show how to obtain upper and lower bounds for G and AR in this censored case, i.e. if observations for $\mathcal{A}$ are not available. As before let $\widetilde{S}_1$, $\widetilde{S}$ denote the observed scores and $\widetilde{F}_1$, $\widetilde{F}$ their cumulative distribution functions. We use (18) and (19) for $\widetilde{F}_1$ and derive similar inequalities for $\widetilde{F}$ using the same ideas we used for $\widetilde{F}_j$. Consider first

$$\widetilde{F}(s) \;=\; \frac{P(S \le s, \mathcal{A})}{P(\mathcal{A})} \le \frac{P(S \le s)}{P(\mathcal{A})} = \frac{F(s)}{P(\mathcal{A})}\,.$$

Also we have

$$\begin{aligned} F(s) &= \widetilde{F}(s)\, P(\mathcal{A}) + P(\widetilde{S} \le s | \overline{\mathcal{A}})\, P(\overline{\mathcal{A}}) \\ &= \widetilde{F}(s)\, P(\mathcal{A}) + P(\{\widetilde{S} \le s\} \cap \mathcal{A}) \\ &\le \widetilde{F}(s)\, P(\mathcal{A}) + P(\overline{\mathcal{A}}) = 1 - P(\mathcal{A})\,\{1 - \widetilde{F}(s)\}. \end{aligned}$$

Together this gives

$$\widetilde{F}(s)\, P(\mathcal{A}) \le F(s) \le 1 - P(\mathcal{A})\,\{1 - \widetilde{F}(s)\}\,. \tag{36}$$

Using this together with (20) for F_1, we find lower bounds

$$\{\widehat{L}_1^\star(s), \widehat{L}_2^\star(s)\} = \left[P(\mathcal{A})\,\left\{1 - \widehat{\widetilde{F}}(s)\right\}, \alpha_1^\star\left\{1 - \widehat{\widetilde{F}}_1(s)\right\}\right]$$

and upper bounds

$$\{\widehat{L}_1^\bullet(s), \widehat{L}_2^\bullet(s)\} = \left\{1 - P(\mathcal{A})\,\widehat{\widetilde{F}}(s), 1 - \alpha_1^\star\,\widehat{\widetilde{F}}_1(s)\right\}$$

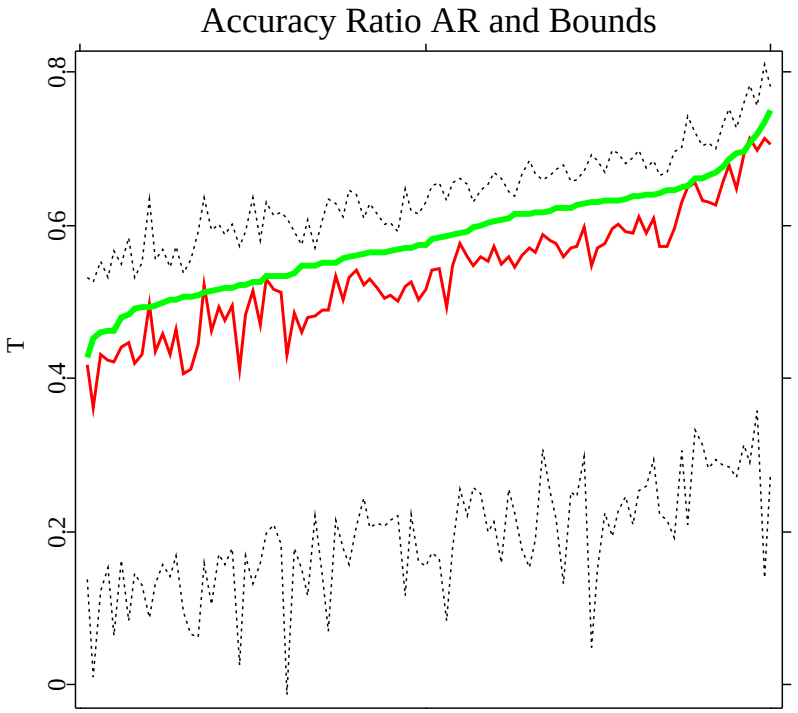

Figure 7: Estimated AR (thick solid), $\widetilde{AR}$ (solid) and bounds (dashed)

for the Lorenz curve. In practice we use the estimates $\widehat{\alpha}_1^\star$, $\widehat{\widetilde{F}}_1(s)$, $\widehat{P}(\mathcal{A})$ from Section 4 and

$$\widehat{\widetilde{F}}(s) = \frac{\sum_i I(S^{(i)} \le s)}{n}\,.$$

The upper and lower bounds for the Lorenz curve obviously lead to upper and lower bounds $\widehat{G}^\bullet$ and $\widehat{G}^\star$ for the Gini coefficient since integration preserves monotonicity. For the accuracy ratio AR we need the additional estimate for $P(Y=0)$. As we discussed before, a point estimate of $P(Y=0)$ is not available. However (21) motivates upper and lower estimates

$$\frac{n_0}{N} \le \widehat{P}(Y=0) \le \frac{N-n_1}{N}\,.$$

Hence, bounds for the estimated accuracy ratio can be found from

$$\frac{N}{N-n_1}\,\widehat{G}^\star \le \widehat{\widetilde{AR}} \le \frac{N}{n_0}\,\widehat{G}^\bullet\,. \tag{37}$$

As we have seen for T, in the special case that all credit applicants are accepted, it holds $\mathcal{A} = \Omega$ and $\alpha_0 = \alpha_1 = 1$. Hence, the upper and lower bounds for the

Lorenz curve as well for Gini coefficient and accuracy ratio coincide with their respective values in this fully observed case.

As an illustration, we use the data from the Monte Carlo simulation in Section 4. Figure 7 shows the estimated AR (thick solid line) and $\widetilde{AR}$ (thin solid line) as well as the estimated upper and lower bounds according for all 100 simulated data sets (sorted by the estimated ARs). We find $\widehat{AR} > \widehat{\widetilde{AR}}$ in 97 cases and $\widehat{AR} < \widehat{\widetilde{AR}}$ in 3 cases. As for T we can conclude that using $\widetilde{AR}$ as a replacement of AR would have led to too large or small values for the discriminatory power of the score, whereas the upper and lower bounds indicate a correctly specified range for $\widehat{AR}$. We also remark that the resulting plot in Figure 7 is very similar to that for T in Figure 5.

References

Banking Committee on Banking Supervision (2001). *The New Basel Capital Accord*, Bank for International Settlements.

Gourieroux, C. and Jasiak, J. (2001). *Financial Econometrics: Problems, Models, and Methods*, Princeton University Press.

Greene, W. H. (1993). *Econometric Analysis*, 2 edn, Prentice Hall.

Greene, W. H. (1998). Sample selection in credit-scoring models, *Japan an the World Economy* **10**: 299–316.

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

Heckman, J. (1979). Sample selection bias as a specification error, *Econometrica* **47**: 153–161.

Horowitz, J. L. and Manski, C. F. (1998). Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputations, *Journal of Econometrics* **84**: 37–58.