Assessing the Discriminatory Power of Credit Scores under Censoring

Holger Kraft, Gerald Kroisandt, Marlene Müller*

Fraunhofer Institut für Techno- und Wirtschaftsmathematik (ITWM)

This version: August 27, 2003

Abstract: We discuss how to assess the performance of 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 same approach is used for the accuracy ratio as an alternative measure for discriminatory power.

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

JEL Classification: G21

*Corresponding author: Marlene Müller, Fraunhofer ITWM, Postfach 3049, D–67653 Kaiserslautern, Germany, Phone: +49 631 205 4189, email: marlene.mueller@itwm.fhg.de

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, such as 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. A particular problem of credit scoring is that defaults and non-defaults are only observed for a subsample of applicants. 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).

Consequently, in practice banks are forced to regularly redesign their rating systems. The available data base for this task typically contains only the accepted credit applicants. Data entries for the rejected credit applicants do often not exist. This leads to a non-representative data base which may give biased estimates of all relevant parameters if this censoring is not appropriately handled. To evaluate the new rating system, e.g. by comparing it with the existing rating system, we would actually need the full data base of all past credit applicants. One opportunity would be to introduce a model which allows us to extrapolate on the data for the rejected applicants (Greene, 1998; Feelders, 2000). For instance, Ash and Meester (2002) and Crook and Banasik (2002) report that such bias corrections typically have a smaller effect than necessary.

Therefore, we use an approach which avoids any specification of a model for the rejected

applicants. We perform a worst case analysis to derive lower and upper bounds for criteria used to evaluate rating systems. More precisely, we consider measures for the discriminatory power of a rating system and especially its corresponding credit scores (numerical values that reflect ratings of the credit applicants). We introduce a criterion that is based on the discrepancy between the score distributions of defaulted and non-defaulted credit applicants. As another criterion of interest we study the accuracy ratio (computed from Gini coefficients, see for example Keenan and Sobehart, 1999) which compares the score distribution of defaulted applicants with that of all credit applicants.

In a different context, Horowitz and Manski (1998) consider a similar censoring problem, namely survey nonresponse. They derive bounds for the regression function, which in our case would correspond to default probabilities. In contrast to their analysis, our focus lies on performance measures for credit scores. For these measures we can exploit the fact that the probabilities of default and non-default cannot vary independently.

To summarize, the main contributions of the paper are the following:

- We discuss how to evaluate credit scores given that only a part of the defaults and non-defaults is observed.
- We strongly emphasize that censoring leads to biased estimates for any kind of performance measure for rating systems.
- We introduce a criterion that is based on the difference of the score distributions under default and non-default. It is then demonstrated how to estimate lower and upper bounds for this criterion.
- We also derive lower and upper bounds for the accuracy ratio as an alternative measure for discriminatory power.

The paper is organized as follows: In Section 2 we discuss how to define the discriminatory power of a credit score. We introduce a criterion that is simple to illustrate and measures the difference between the score distributions of defaulted and non-defaulted loans. Section 3 discusses the consequences of censoring for this criterion. In our context, censoring means that we assume to have default or non-default information only for a restricted set of applicants. To keep things simple we first study these consequences for

normally distributed scores. In Section 4 we consider the nonparametric case and show how to find lower and upper bounds for the proposed criterion under very weak assumptions. Section 5 extends our approach to lower and upper bounds for the Gini coefficient and the accuracy ratio (AR). Section 6 illustrates the obtained bounds by a real data example. Finally, Section 7 summarizes our results.

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 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 a 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.

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. Let f_0, f_1 denote the probability densities of $S|Y = 0$ and $S|Y = 1$, and 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

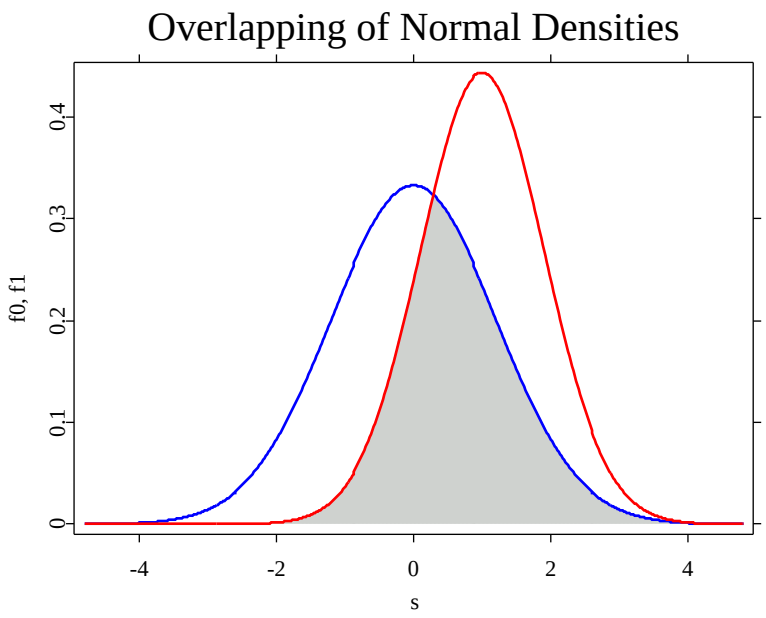

Figure 1: Overlapping area U for two normal densities

(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 discriminant 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 case of more than one intersection point and 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 for 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 exists in the monotone case. As for the normal case, we assume that f_1 is located to the 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 in this case the distribution of

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

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

3 Credit Scoring & Censoring

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 can be described as *censoring* or *sample selection*.

Obviously, this data sampling will result in biased estimates of all relevant parameters due to the non-representative data base. In particular, we want to stress that this bias can be positive or negative. In the sequel, we will see that formally this comes from the fact that we are working with conditional probabilities. The problem of bias correction in this case has been mainly studied by using (regression) models that extrapolate on the unobserved data and with a focus on the estimating regression coefficients and PDs. In the econometric literature, bivariate regression models for sample selection (Heckman, 1979) are well-known. For example, Greene (1998) and Boyes, Hoffman and Low (1989) use a bivariate probit model for credit data. In the statistical literature, this bias correction technique is know as *reject inference*. Feelders (2000) and Crook and Banasik (2002) are relevant references here.

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 $c = 2$) cannot be observed but needs in fact to be considered for a correct assessment of the performance of the score.

Let $\widetilde{S} = (S|\mathcal{A})$ and $\widetilde{Y} = (Y|\mathcal{A})$ denote the observed parts of the score and the group

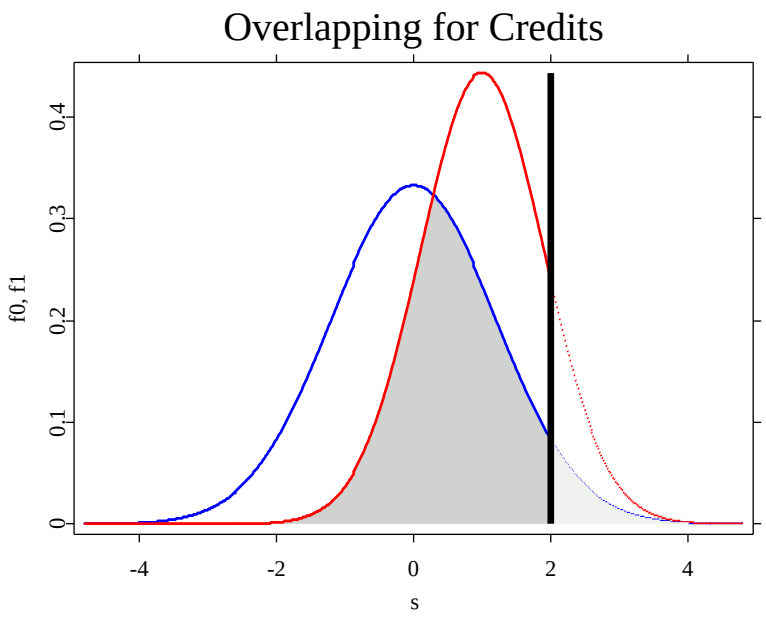

Figure 2: Truncated overlapping area for credit data

indicator. Hence, we have only observations for the censored score $\widetilde{S}_j = (\widetilde{S}|\widetilde{Y} = j)$, $j = 0, 1$, while we are interested in the non-censored score $S_j = (S|Y = j)$. Under the assumption (6), the relation between $\widetilde{S}_j$ and S_j is given by

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

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

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

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

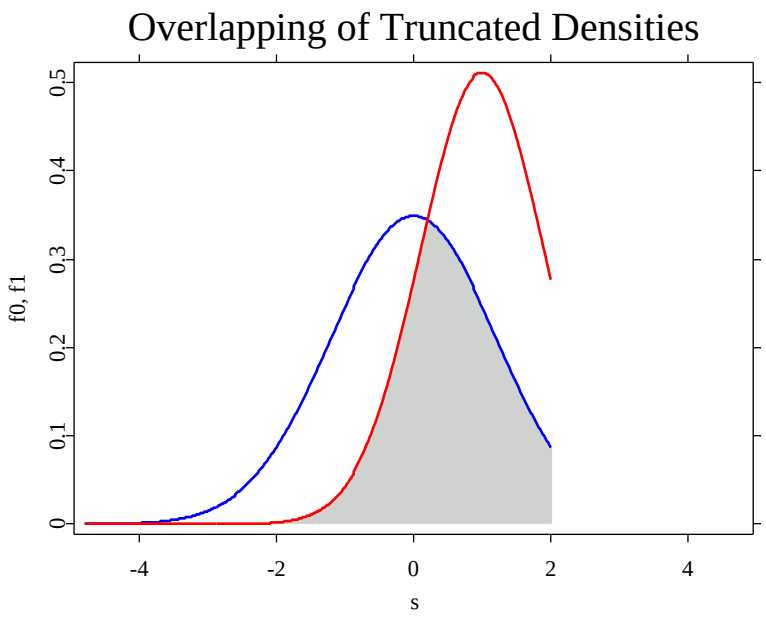

Figure 3: Observed overlapping area $\widetilde{U}$

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 region from incompletely observed data is usually different (biased) from the objective overlapping region U.

The difference in $T = 1-U$ and $\widetilde{T} = 1-\widetilde{U}$ can be considerably important as the following 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 1.44. 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 a censored sample of $n = 491$ observations.

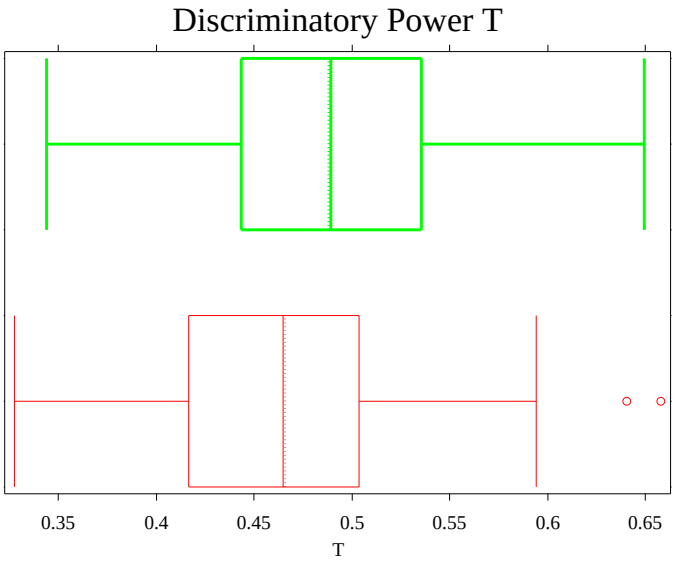

Figure 4: Difference in T (upper boxplot) and $\widetilde{T}$ (lower boxplot)

In Figure 4 boxplots for the realized distributions of the estimated $\widetilde{T}$ and T are displayed. The graphic shows that in our simulated example $\widetilde{T}$ is typically smaller than T. A closer inspection of the data shows that in 91 cases $\widehat{\widetilde{T}} < \widehat{T}$ and in 9 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 underestimation). Recall that this comes from the fact that the acceptance condition $\mathcal{A}$ leads to conditional probabilities.

Before explaining a very general approach to approximate T from $\widetilde{T}$ in the following section, let us first discuss the solution under the assumption that the type of the score distribution is known. In this case a correction for $\widetilde{T}$ can be easily calculated. Let us outline this idea for the example of normal distributions: Here the moments of $\widetilde{S}_j$ can be calculated (see for example 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, $\alpha_j = (c - \mu_j)/\sigma_j$, and $\lambda(\alpha) = -\varphi(\alpha)/\Phi(\alpha)$ denoting the inverse Mills ratio (using φ and Φ for the normal density and cumulative distribution functions). 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 approach 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 $\log(S_j) \sim N(\mu_j, \sigma_j^2)$. Since the logarithm is monotone $F_j(s) = P(S_j \leq s) = P(\log(S_j) \leq \log(s))$. 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 transformation such as a Box–Cox transformation.

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

$$\mathcal{A} = \{S^\star \leq c\}. \tag{11}$$

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

$$E(S_j|S_j^\star < c) = \mu_j + \rho_j\sigma_j\lambda(\alpha_j), \tag{12}$$

$$Var(S_j|S_j^\star < c) = (\sigma_j)^2[1 - \rho_j^2\lambda(\alpha_j)\{\lambda(\alpha_j) - \alpha_j\}], \tag{13}$$

(cf. Greene, 1993, Theorem 22.4). Here we denote $\alpha_j = (c - \mu_j^\star)/\sigma_j^\star$ while λ stands for the inverse Mills ratio as before. In addition we have the cumulative distribution function of $S_j^\star$ given by

$$\widetilde{F}_j^\star(x) = \Phi_2\left(\frac{x - \mu_j^\star}{\sigma_j^\star}, \frac{c - \mu_j}{\sigma_j}, \rho_j\right)\left\{\Phi\left(\frac{c - \mu_j}{\sigma_j}\right)\right\}^{-1}, \tag{14}$$

using the notation Φ_2 for the bivariate normal cumulative distribution function. 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 (12)–(14) 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 such 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 T 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 often not available.

A possible remedy to this problem is the calculation of lower and upper bounds for the discriminatory power. The approach which we apply is inspired by Horowitz and Manski (1998). 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

$$T = \max_s \{F_0(s) - F_1(s)\}$$

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 $\widetilde{S}_j = (\widetilde{S}_j|\widetilde{Y} = j) = (S|Y = j, \mathcal{A})$. To derive upper and lower bounds for T we have to relate the unobservable $F_j(s)$ to the observable function $\widetilde{F}_j(s)$. The following lemma shows this relation. For the sake of clarity we have collected all more complex derivations in the appendix.

Lemma 4.1

Using the notation $\alpha_j = P(\mathcal{A}|Y = j)$, *we have*

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

To apply this lemma for calculating bounds for T, we now need bounds for α_j. These follow from

$$P(Y = j, \mathcal{A}) \leq P(Y = j) \leq P(Y = j, \mathcal{A}) + P(\overline{\mathcal{A}}). \tag{15}$$

which is a consequence of $P(Y = j) = P(Y = j, \mathcal{A}) + P(Y = j, \overline{\mathcal{A}})$, where $\overline{\mathcal{A}}$ stands for the complement of $\mathcal{A}$. Since

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

it follows by (15) that

$$\alpha_j \in [\alpha_j^{low}, 1], \quad \text{where } \alpha_j^{low} = \frac{P(\widetilde{Y} = j)P(\mathcal{A})}{P(\widetilde{Y} = j)P(\mathcal{A}) + P(\overline{\mathcal{A}})}\,. \tag{16}$$

Lemma 4.1 together with (16) yields upper and lower bounds for T. We summarize this result in the following proposition:

Proposition 4.2

Bounds for T *are given by*

$$\max_s \left[\alpha_0^{low}\,\widetilde{F}_0(s) + \alpha_1^{low}\,\{1 - \widetilde{F}_1(s)\}\right] - 1$$
$$\leq T \leq 1 - \min_s \left[\alpha_0^{low}\{1 - \widetilde{F}_0(s)\} + \alpha_1^{low}\,\widetilde{F}_1(s)\right]\,.$$

We want to stress that in the special case of no censoring (i.e., if all credit applicants were accepted for a loan and we observe their default or non-default) we have $P(\overline{\mathcal{A}}) = 0$ and $\alpha_0^{low} = \alpha_1^{low} = 1$. As a consequence, the inequality of Proposition 4.2 reduces to $T = \max\{F_0(s) - F_1(s)\}$ which is exactly the definition for the uncensored case.

Let us further remark that the bounds in Proposition 4.2 are quite useful but not the optimal ones. In contrast to Horowitz and Manski (1998), we can exploit that $P(Y = 0)$

and $P(Y = 1)$ are complements and cannot vary independently. Using this fact, we can derive improved bounds which are summarized in Proposition 4.3. It turns out, however, that in Monte–Carlo simulations the improvement by Proposition 4.3 is very modest. We refer here to the simulation example which is presented later on.

Proposition 4.3

Improved bounds for T are given by

$$\max_s \left[\frac{\beta_0}{p_s^{up}} \widetilde{F}_0(s) + \frac{\beta_1}{1 - p_s^{up}} \{1 - \widetilde{F}_1(s)\} \right] - 1$$
$$\leq T \leq 1 - \min_s \left[\frac{\beta_0}{p_s^{low}} \{1 - \widetilde{F}_0(s)\} + \frac{\beta_1}{1 - p_s^{low}} \widetilde{F}_1(s) \right],$$

where $\beta_j = P(Y = j, \mathcal{A})$ and the functions p_s^{low} and p_s^{up} are defined as in (22) and (25) in the appendix.

To apply these bounds to empirical data we need to estimate all unknown quantities in Proposition 4.2 or 4.3. This is possible because we know the total number of scored credit applicants N. More precisely: For the observed scores under default and non-default we know their empirical distribution functions $\widehat{\widetilde{F}}_j$ which can be obtained analogously to (5). To estimate α_j^{low}, β_j, p_0^{low}, and p_0^{up} we consider the probabilities of $\{\widetilde{Y} = j\} = \{Y = j|\mathcal{A}\}$, $\mathcal{A}$, and $\overline{\mathcal{A}}$, which can be approximated by their observed relative frequencies

$$\widehat{P}(\widetilde{Y} = j) = \frac{n_j}{n}, \quad \widehat{P}(\mathcal{A}) = \frac{n}{N}, \quad \widehat{P}(\overline{\mathcal{A}}) = \frac{N - n}{N}. \tag{17}$$

Here n_0 denotes the number of observed non-defaults and similarly n_1 denotes for the number of observed defaults. As before, n stands for the sample size of the observed credits (i.e., $n = n_0 + n_1$). This provides the estimates

$$\widehat{\alpha}_j^{low} = \frac{n_j}{n_j + N - n}, \quad \widehat{\beta}_j = \frac{n_j}{N}. \tag{18}$$

Estimates for p_0^{low} and p_0^{up} can be found by plugging $\widehat{\beta}_j$, $\widehat{P}(\overline{\mathcal{A}})$ and $\widehat{\widetilde{F}}_j(s)$ into (22)–(23) and (25)–(26).

The following Monte Carlo simulation illustrates the effect of the estimated bounds. We use the previously simulated data set. Figure 5 shows estimates for T, $\widetilde{T}$ and the estimated

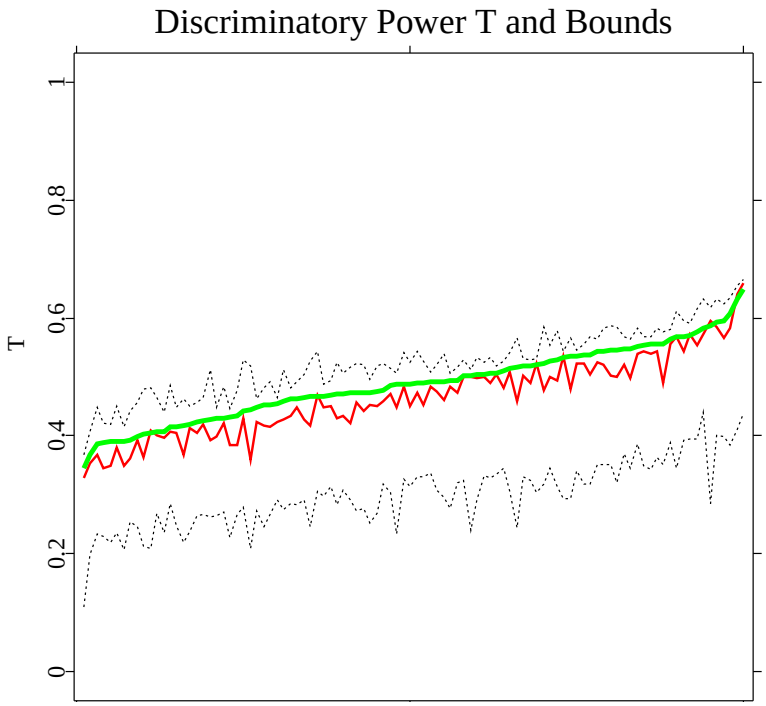

Figure 5: Estimated T (thick solid), $\widetilde{T}$ (solid) and bounds (dashed)

upper and lower bounds according to Proposition 4.3 for all 100 simulated data sets. To simplify the comparison all simulated values are sorted by the estimated values of T. The bounds according to Proposition 4.2 are only slightly wider, such that we omit them here.

Note that in practice the estimation of $\widehat{T}$ could not have been carried out because data on rejected applicants are usually not collected. However, due to our simulation experiment, we have the opportunity to estimate both $\widetilde{T}$ and T. The simulation analysis shows in particular, that T might be smaller or larger than $\widetilde{T}$. A closer inspection of the simulated data reveals that $\widetilde{T}$ tends to be larger than T if $P(\widetilde{Y} = 1) \approx P(Y = 1)$. In other words, $\widetilde{T}$ tends to overestimate T if the censoring condition does reject a too small number of defaults. In the afore-mentioned 9 cases using $\widetilde{T}$ instead of T would have led to a too optimistic value for the discriminatory power of the score. The upper and lower bounds, however, indicate a correctly specified range for $\widehat{T}$.

We see that the lower bound in Figure 5 seems to be quite far away from both estimated T and $\widetilde{T}$. This is a consequence of the fact that our bounds do not require any information about the structure of the censoring condition $\mathcal{A}$. A narrower lower bound could be

calculated if additional information on $\mathcal{A}$ is available. A trivial example is $\mathcal{A} = \{S \leq c\}$, assuming that we would want to evaluate the score S which is at same time used for the rejection of credit applicants. A more realistic example is $\mathcal{A} = \{S^\star \leq c\}$ as discussed in Section 3. We have seen, however, that to exploit the fact that the acceptance condition is of the form $\{S^\star \leq c\}$, the relation between S and $S^\star$ must be known more precisely.

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 (Keenan and Sobehart, 1999; Engelmann, Hayden and Tasche, 2003). In the case of censored data, the accuracy ratio computed from the observed part of the data will be biased as well. As in the case of T we will now derive bounds for AR if the distribution of the score is unknown.

Let us first introduce the relevant terms for the non-censored case. The Lorenz curve aims to visualize scores by means of comparing the distributions of S and $S_1 = (S|Y = 1)$. Figure 6 shows the principle of the Lorenz curve. On the horizontal and vertical scales, the percentages of applicants are sorted from "bad" to "good" scores. The Lorenz curve is also known as the power curve or the cumulative accuracy profile (CAP). A related curve is the receiver operating characteristic (ROC) curve (Hand and Henley, 1997; Engelmann et al., 2003; Sobehart and Keenan, 2001) which compares the distributions of S_0 and S_1.

To be consistent with our previous notation, let V denote the negative score, i.e., $V = -S$. The Lorenz curve of S is then defined by the coordinates

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

which is equivalent to

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

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

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

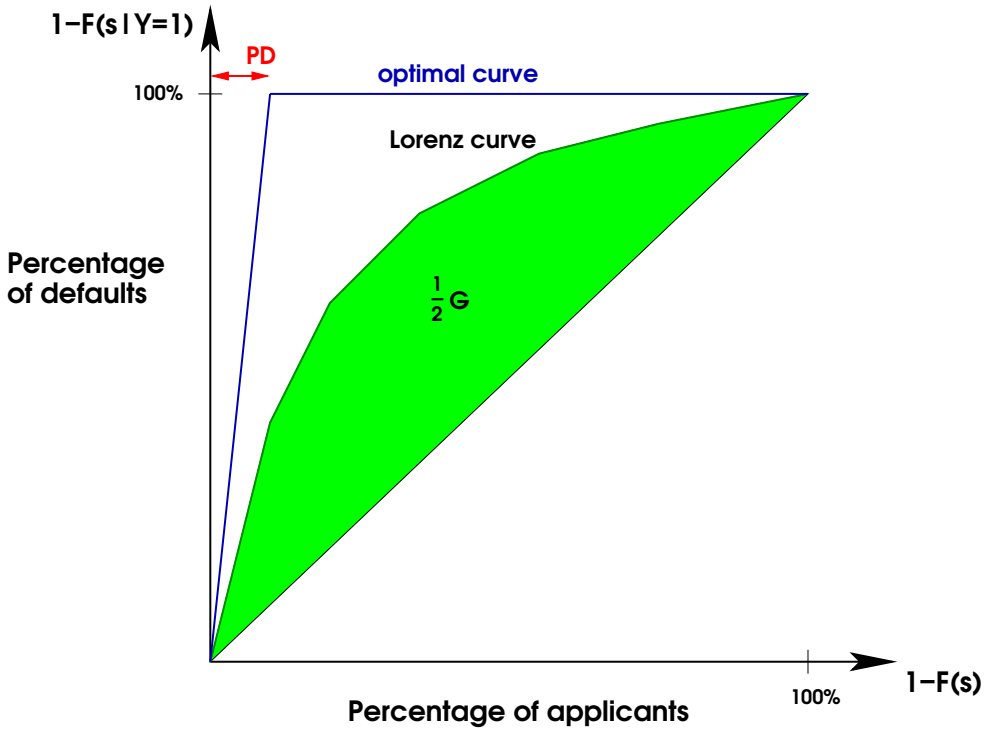

Figure 6: Lorenz curve for credit scores

large. The best (optimal) score does exactly separate defaults and non-defaults. The corresponding optimal Lorenz curve reaches the vertical 100% at a horizontal percentage of $P(Y = 1)$, the probability of default. The worst score is one that does not contain any information about defaults and non-defaults, i.e., assigns randomly score values to credit applicants. The corresponding Lorenz curve is thus (since $F_1(s) = F(s)$ in that case) identical to the diagonal.

A typical Lorenz curve is located between the optimal curve and the diagonal (cf. Figure 6). Since better scores should be more close to the optimal curve, Lorenz curves can be applied to compare different score functions. 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.,

$$G = 2 \int (1 - F_1)(s)\, d(1 - F)(s) - 1 = 1 - 2 \int F_1(s)\, dF(s)\,. \tag{19}$$

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 defined by

$$AR = \frac{G}{G_{opt}} = \frac{G}{P(Y=0)} \in [-1, 1].$$

Note that negative values of AR do only occur if the score is unreasonably defined, for example if low score values correspond with high PDs.

We now turn to the censored case. Here we observe $\widetilde{G} = (G|\mathcal{A})$ and $P(\widetilde{Y} = 0) = P(Y = 0|\mathcal{A})$, such that we can calculate only

$$\widetilde{AR} = \frac{\widetilde{G}}{P(\widetilde{Y}=0)}$$

instead of AR. Thus, analogously to $\widetilde{T}$, the Gini coefficient $\widetilde{G}$ and the accuracy ratio $\widetilde{AR}$ are biased. We will now show how to obtain upper and lower bounds for G as well as AR.

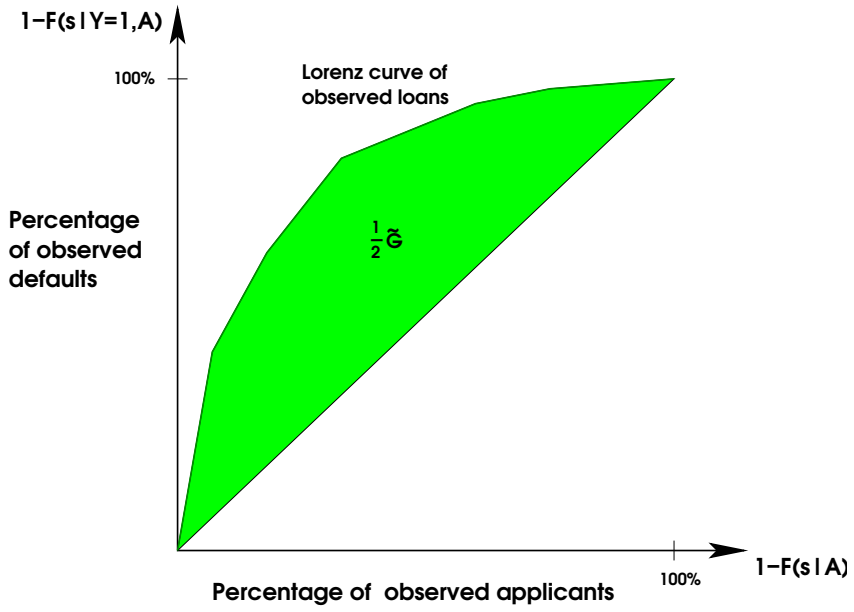

Figure 7: Lorenz curve under censoring

Suppose that the Lorenz curve for the observed loans looks as in Figure 7. To obtain lower and upper bounds for the Lorenz curve of all credit applicants, we consider two extreme cases for the unobserved part: (a) all unobserved loans default and (b) all unobserved loans do not default. These assignments lead to the curves in Figure 8.

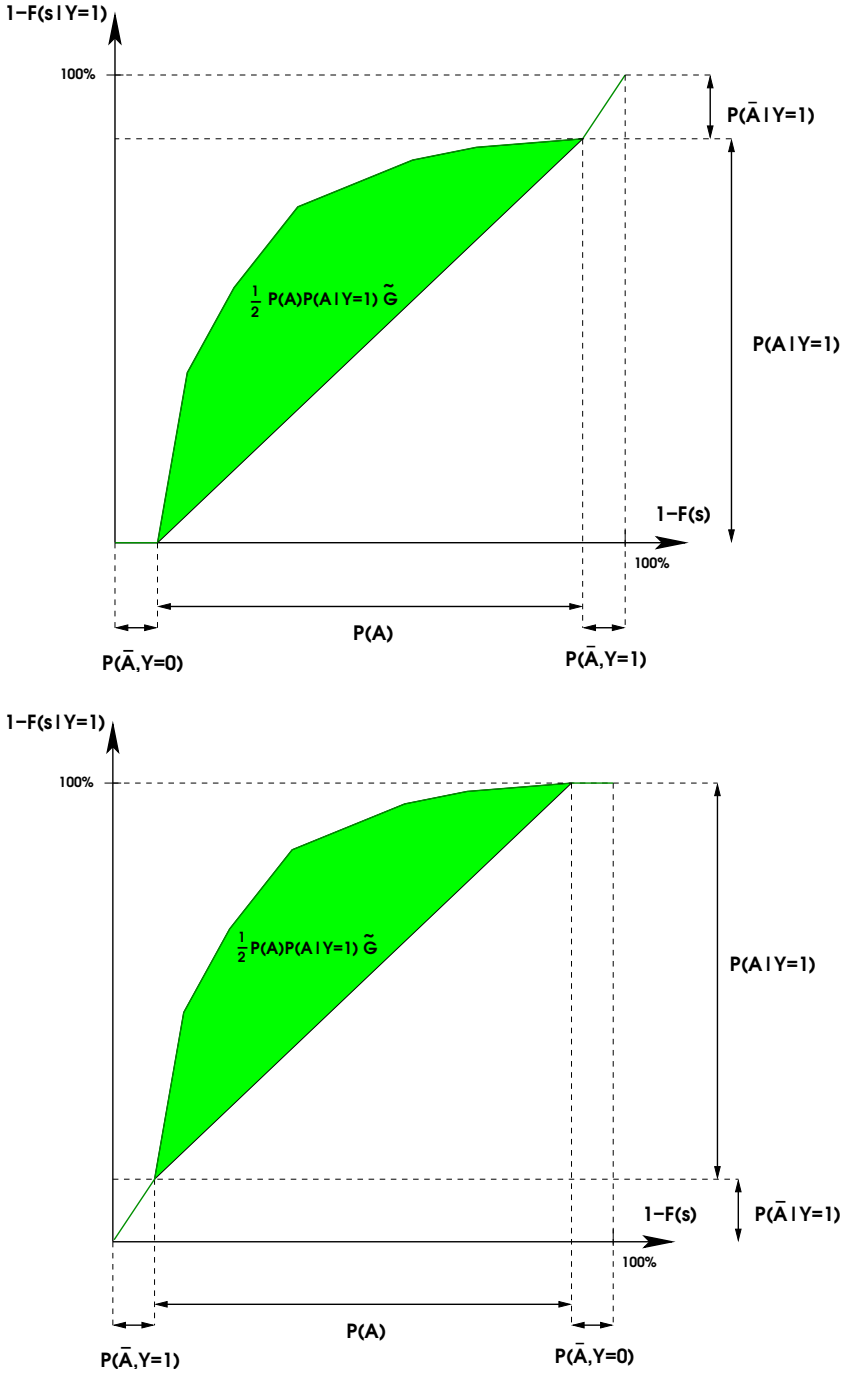

Figure 8: Lorenz curve under censoring

Hence, lower and upper bounds for G (and subsequently for AR) can be derived by calculating the areas under the curves in Figure 8. The resulting inequality for AR is summarized by the following proposition. As before we refer to the appendix for the detailed proof.

Proposition 5.1

Bounds for AR are given by

$$\left(\widetilde{AR}+1\right)\frac{\beta_0\beta_1}{p_0^\star(1-p_0^\star)}-1\leq AR\leq\left(\widetilde{AR}-1\right)\frac{\beta_0\beta_1}{p_0^\star(1-p_0^\star)}+1$$

where

$$p_0^\star=\begin{cases}\beta_0 & \text{if } \beta_0>\frac{1}{2}\,,\\ \frac{1}{2} & \text{if } \beta_0\leq\frac{1}{2}\leq\beta_0+P(\overline{\mathcal{A}})\,,\\ \beta_0+P(\overline{\mathcal{A}}) & \text{if } \beta_0+P(\overline{\mathcal{A}})<\frac{1}{2}\,.\end{cases}$$

Let us remark that in the special case if all credit applicants are accepted, it holds $P(\overline{\mathcal{A}})=0$ and therefore $\beta_0=\frac{1}{2}$. Hence, the upper and lower bounds for the Lorenz curve as well for Gini coefficient and accuracy ratio coincide with their respective values in the non-censored case.

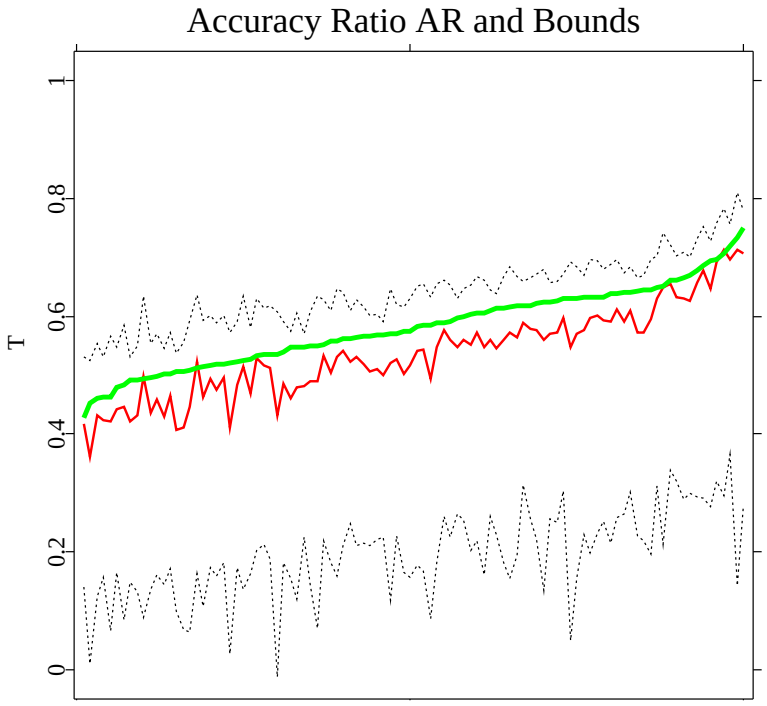

Figure 9: Estimated AR (thick solid), $\widetilde{AR}$ (solid) and bounds (dashed)

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

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

To illustrate the result of Proposition 5.1, we reuse the data from the Monte Carlo simulation in Section 4. Figure 9 shows the estimated AR and $\widetilde{AR}$ as well as the estimated upper and lower bounds according to 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 $\widetilde{AR}$ instead 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 $\widehat{AR}$. The remarks on the simulation in Section 4 apply here as well. We see, however, that the estimated bounds are wider (relative to the values of $\widetilde{AR}$ and AR) and that the lower bound may be negative. Thus, often only the upper bound has a useful interpretation.

6 Application

Let us now consider a brief illustration on real data. We use the credit data from Fahrmeir and Tutz (1994) which are publicly available[1]. The data set comprises 1000 observations of private loans. One of the variables is credit history. We will now try to assess discriminatory power under the assumption that customers with a negative credit history (those which showed a "hesitant payment of previous credits") would not have granted a loan and that their default or non-default would not have been observed. This means we use a sample of $n = 960$ observed customers whereas the sample size of all applicants is equal to $N = 1000$.

We estimate two different Logit specifications. The corresponding variables are listed in Table 1. The first specification uses more personal and credit information but is not a superset of the second specification. We compare the scores estimated by a Logit model in both specifications with respect to T and AR. The resulting criteria on the observed data as well as the estimated lower and upper bounds are shown in Table 2.

We recognize that as in Figures 5 and 9 the intervals for AR are clearly wider. Consequently, information that we get out of the interval estimates is more precise in the case of T. In particular, we observe a negative lower bound for AR in specification 2. However, in this special example the intervals for AR do not have an intersection. This means that

[1] http://www.stat.uni-muenchen.de/service/datenarchiv/kredit/kredit.html

Variable	Specification 1	Specification 2
previous loans (1 for OK, 0 for unknown)	×	
employed (1 for more than one year, 0 otherwise)		×
duration of the loan (discretized with dummies for 10–12, 13–18, 19– 24 and more than 24 months)	×	
amount of the loan (+ amount squared)	×	×
age of the borrower (+ age squared)	×	
interaction term for amount and age	×	
savings (1 for more than 1000 DM, 0 otherwise)	×	
foreigner (1 if yes, 0 otherwise)	×	
purpose (1 if loan is used to buy a car, 0 otherwise)	×	
house owner (1 if yes, 0 otherwise)	×	

Table 1: Variables for score estimation

Estimated criterion	Specification 1	Specification 2
$\widetilde{T}$	0.292	0.159
maximal range of T	[0.222,0.349]	[0.108,0.235]
$\widetilde{AR}$	0.419	0.125
maximal range of AR	[0.238,0.492]	[-0.018,0.236]

Table 2: Discriminatory power of the scores

specification 1 is definitely better than specification 2. Here, the unobserved data cannot improve the accuracy ratio for specification 2 over that for specification 1.

7 Conclusions

The discriminatory power of a credit score can be estimated by comparing the score distributions of the defaults with that of the non-defaults or the full sample. We consider two possible criteria in this paper: The maximal difference of the cumulative score distribution functions for non-defaults and defaults

$$T = \max_s \{F_0(s) - F_1(s)\}$$

and the accuracy ratio

$$AR = \frac{1 - 2\int F_1(s)\, dF(s)}{P(Y = 0)}$$

As we have seen, a censored sample can lead to considerable bias when using the criteria to evaluate the score with respect to discriminatory power. Our simulations show that the bias might be positive or negative, i.e., there is no simple rule to take account for this bias.

A corrected calculation of the criteria is possible if details on the acceptance or rejection of the credit applicants are known. However, often no precise information about rejected clients is available. For this case the paper offers the possibility to assess discriminatory power by computing lower and upper bounds of such criteria. The calculation of bounds is possible under the very weak assumption that only the percentage of rejected credits is known.

Appendix: Proofs

Proof of Lemma 4.1

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

$$F_j(s) = \widetilde{F}_j(s)\, P(\mathcal{A} | Y = j) + P(S \le s, \overline{\mathcal{A}} | Y = j). \tag{20}$$

We find an upper bound for $F_j(s)$ by using that $\{S \leq s\} \cap \overline{\mathcal{A}} \subseteq \overline{\mathcal{A}}$ in the second term of (20), i.e.,

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

A lower bound for $F_j(s)$ is given by omitting the second term of (20) completely, such that

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

□

Proof of Proposition 4.2

The result follows directly by combining Lemma 4.1 and the bounds in (16). □

Proof of Proposition 4.3

We introduce the additional abbreviations $\beta_j = P(Y=j,\mathcal{A})$ and

$$p = P(Y=0),$$

such that $\alpha_0 = \beta_0/p$ and $\alpha_1 = \beta_1/(1-p)$. We will first consider bounds for U and later on transfer them into bounds for T.

Consider the lower bound for U first. From the proof of Lemma 4.1 we see

$$\begin{aligned} F_1(s) + 1 - F_0(s) &\geq \alpha_1\widetilde{F}_1(s) + \alpha_0\{1-\widetilde{F}_0(s)\} \\ &= \frac{\beta_1}{1-p}\widetilde{F}_1(s) + \frac{\beta_0}{p}\{1-\widetilde{F}_0(s)\} \end{aligned} \tag{21}$$

In the last term each of the probabilities can be estimated from the observed data except for p. Hence, for given s the last term has to be minimized with respect to p. 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 optimum:

$$p_s^{low} = \begin{cases} \beta_0 & \text{if } \gamma_s < \beta_0, \\ \beta_0 + P(\overline{\mathcal{A}}) & \text{if } \gamma_s > \beta_0 + P(\overline{\mathcal{A}}), \\ \gamma_s, & \text{otherwise,} \end{cases} \tag{22}$$

and

$$\gamma_s = \frac{\sqrt{\beta_0\{1-\widetilde{F}_0(s)\}}}{\sqrt{\beta_0\{1-\widetilde{F}_0(s)\}}+\sqrt{\beta_1\widetilde{F}_1(s)}}. \tag{23}$$

The upper and lower thresholds in (22) are consequences of the bounds in (15).

To derive the upper bound of U we have again from the proof of Lemma 4.1

$$\begin{aligned} F_1(s)+1-F_0(s) &\leq 2-\alpha_1\{1-\widetilde{F}_1(s)\}-\alpha_0\widetilde{F}_0(s) \\ &= 2-\frac{\beta_1}{1-p}\{1-\widetilde{F}_1(s)\}-\frac{\beta_0}{p}\widetilde{F}_0(s). \end{aligned} \tag{24}$$

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

$$p_s^{up} = \begin{cases} \beta_0 & \text{if } \delta_s < \beta_0, \\ \beta_0 + P(\overline{\mathcal{A}}) & \text{if } \delta_s > \beta_0 + P(\overline{\mathcal{A}}), \\ \delta_s, & \text{otherwise,} \end{cases} \tag{25}$$

and

$$\delta_s = \frac{\sqrt{\beta_0\widetilde{F}_0(s)}}{\sqrt{\beta_0\widetilde{F}_0(s)}+\sqrt{\beta_1\{1-\widetilde{F}_1(s)\}}}. \tag{26}$$

Combining the results we obtain

$$\begin{aligned} &\frac{\beta_1}{1-p_s^{low}}\widetilde{F}_1(s)+\frac{\beta_0}{p_s^{low}}\{1-\widetilde{F}_0(s)\} \\ &\qquad \leq F_1(s)+1-F_0(s) \leq 2-\frac{\beta_1}{1-p_s^{up}}\{1-\widetilde{F}_1(s)\}-\frac{\beta_0}{p_s^{up}}\widetilde{F}_0(s) \end{aligned} \tag{27}$$

such that by using $T = 1-U$ the statement is proved. □

Proof of Proposition 5.1

We recall the notation $\beta_j = P(Y=j,\mathcal{A})$, which allows us to write $\beta_0+\beta_1$ instead of $P(\mathcal{A})$. Additionally, we introduce the notation

$$p_j = P(Y=j).$$

Obviously these probabilities are related by $p_0+p_1=1$. We can now express the following terms using p_j and β_j.

$$P(Y=j|\mathcal{A}) = \frac{\beta_j}{\beta_0+\beta_1}$$

$$
\begin{aligned}
\widetilde{AR} &= \frac{\beta_0+\beta_1}{\beta_0}\,\widetilde{G} \quad \Longleftrightarrow \quad \widetilde{G} = \frac{\beta_0}{\beta_0+\beta_1}\,\widetilde{AR} \\
P(\overline{\mathcal{A}}, Y=j) &= p_j - \beta_j \\
P(\overline{\mathcal{A}}|Y=j) &= \frac{p_j-\beta_j}{p_j}
\end{aligned}
$$

Consider first the lower bound for AR. From the first plot of Figure 8 we see that the lower bound for G (twice the area under the curve minus 1) equals

$$
\begin{aligned}
G^{low} &= P(\mathcal{A})P(\mathcal{A}|Y=1)\widetilde{G} + P(\mathcal{A})P(\mathcal{A}|Y=1) \\
&\qquad + P(\overline{\mathcal{A}}, Y=1)\left\{1 - P(\mathcal{A}|Y=1)\right\} - 1 \\
&= (\beta_0+\beta_1)\,\frac{\beta_1}{p_1}\,(\widetilde{G}+1) - (p_1-\beta_1)\left(1+\frac{\beta_1}{p_1}\right) - 1\,.
\end{aligned}
$$

Using the relation between $\widetilde{G}$ and $\widetilde{AR}$ leads to

$$
G^{low} = \frac{1}{p_1}\left\{(\widetilde{AR}+1)\,\beta_0\beta_1 - p_0p_1\right\}.
$$

Thus we obtain

$$
AR^{low} = \frac{G^{low}}{p_0} = \left(\widetilde{AR}+1\right)\frac{\beta_0\beta_1}{p_0p_1} - 1\,. \tag{28}
$$

We now use the same approach for the upper bound of AR. From the second plot in Figure 8 we calculate as an upper bound for G

$$
\begin{aligned}
G^{up} &= P(\overline{\mathcal{A}}|Y=1)P(\overline{\mathcal{A}}, Y=1) + P(\mathcal{A})P(\mathcal{A}|Y=1)\widetilde{G} \\
&\qquad + P(A)\left\{P(\mathcal{A}|Y=1)+1\right\} + P(\overline{\mathcal{A}}, Y=0) - 1 \\
&= \frac{p_1-\beta_1}{p_1}\,(p_1-\beta_1) + (\beta_0+\beta_1)\,\frac{\beta_1}{p_1}\,\widetilde{G} \\
&\qquad + (\beta_0+\beta_1)\left(\frac{p_1-\beta_1}{p_1}+1\right) + 2(p_0-\beta_0) - 1 \\
&= \frac{1}{p_1}\left\{\beta_0\beta_1(\widetilde{AR}-1) + p_0p_1\right\}.
\end{aligned}
$$

This results in

$$
AR^{up} = \frac{G^{up}}{p_0} = \left(\widetilde{AR}-1\right)\frac{\beta_0\beta_1}{p_0p_1} + 1\,. \tag{29}
$$

Hence, we obtain together with (28)

$$
\left(\widetilde{AR}+1\right)\frac{\beta_0\beta_1}{p_0p_1} - 1 \le AR \le \left(\widetilde{AR}-1\right)\frac{\beta_0\beta_1}{p_0p_1} + 1\,. \tag{30}
$$

To achieve the minimal value for the lower and the maximal value for the upper bound, it is obvious that $p_0 p_1 = p_0(1 - p_0)$ must be maximal. It is important to note that p_0 cannot vary freely since from (15) we have

$$\beta_0 \leq p_0 \leq \beta_0 + P(\overline{\mathcal{A}})\,.$$

As a consequence, we have to distinguish three cases:

(1) $\beta_0 \leq \frac{1}{2} \leq \beta_0 + P(\overline{\mathcal{A}})$
In that case, the value that maximizes $p_0(1 - p_0)$ is $p_0^\star = \frac{1}{2}$ (as if p_0 could take on all values between 0 and 1).

(2) $\frac{1}{2} < \beta_0$
Here, the optimal value is $p_0^\star = \beta_0$.

(3) $\frac{1}{2} > \beta_0 + P(\overline{\mathcal{A}})$
Here, the optimal value is $p_0^\star = \beta_0 + P(\overline{\mathcal{A}})$.

□

References

Ash, D. and Meester, S. (2002). Best practices in reject inferencing, *Conference presentation*, Credit Risk Modelling and Decisioning Conference, Wharton Financial Institutions Center, Philadelphia.

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

Boyes, W. J., Hoffman, D. L. and Low, S. A. (1989). Measuring default accurately, *Journal of Econometrics* **40**: 3–14.

Crook, J. and Banasik, J. (2002). Does reject inference really improve the performance of application scoring models?, *Working paper*, Credit Research Centre, The School of Management, University of Edinburgh.

Engelmann, B., Hayden, E. and Tasche, D. (2003). Testing rating accuracy, *Risk* **16**: 82–86.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer.

Feelders, A. J. (2000). Credit scoring and reject inference with mixture models, *International Journal of Intelligent Systems in Accountings, Finance & Management* **9**: 1–8.

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

Greene, W. H. (1998). Sample selection in credit-scoring models, *Japan and 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.

Keenan, S. C. and Sobehart, J. R. (1999). Performance measures for credit risk models, *Research report # 1-10-10-99*, Risk Management Services, Moody's Investors Service.

Sobehart, J. R. and Keenan, S. C. (2001). Measuring default accurately, *Risk* **14**: 31–33.