Redesigning Ratings: Assessing the Discriminatory Power of Credit Scores under Censoring

Holger Kraft, Gerald Kroisandt, Marlene Müller*

Fraunhofer Institut für Techno- und Wirtschaftsmathematik (ITWM)

This version: June 29, 2004

ABSTRACT: In the light of Basel II, redesigning rating systems has been becoming an important issue for banks and other financial institutions. The available data base for this task typically contains only the accepted credit applicants and is thus censored. To evaluate existing and alternative rating systems, we would actually need the full data base of all past credit applicants. In this paper we discuss how to assess the performance of credit ratings under the assumption that for credit data only a part of the defaults and non-defaults is observed. The paper investigates criteria that are based on the difference of the score distributions under default and non-default such as the accuracy ratio. We show how to estimate bounds for these criteria in the usual situation that the bank storages only data of the accepted credit applicants.

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.fraunhofer.de

1 Introduction

A bank which wishes to decide whether a credit applicant will obtain a credit or not has to assess if the applicant will be able to redeem the credit. Among other criteria, the bank needs 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 equity and liquidity ratios or factors 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, banks are forced to assess the quality of their rating systems and to optimize them with respect to the above-mentioned requirements. The available data base for this task typically contains only the accepted credit applicants (the debitors of the bank). Data entries for the rejected credit applicants do often not exist. This leads to a non-representative data base which gives biased estimates of all relevant parameters if this censoring is not appropriately handled. To evaluate a 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 possibility 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 (2004) 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 (PDs). 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 the discriminatory power of credit scores given that only a part of the defaults and non-defaults is observed.
- It is strongly emphasized that censoring leads to biased estimates for any kind of performance measure for rating systems.
- We derive lower and upper bounds for criteria that are based on the difference of the score distributions under default and non-default.

The paper is organized as follows: In Section 2 we discuss how to define the discriminatory power of a credit score. We introduce two criteria that are simple to illustrate and measure the difference between the score distributions of defaulted and non-defaulted loans. Section 3 discusses the consequences of censoring. In our context, censoring means that we assume to have default or non-default information only for a restricted set of applicants. In Section 4 we show how to find lower and upper bounds for the proposed criteria under very weak assumptions. Section 5 illustrates the obtained bounds by a real data example. Finally, Section 6 summarizes our results.

2 Discriminatory Power of a Score

To assess the default risk of an credit applicant a bank usually identifies several indicators $X_1, \dots, X_p$ such as debt equity ratio or return on investments which are then aggregated to a single value, the rating of the credit applicant. We will call the mechanism that is used to aggregate the factors *score* function. Formally, the score function S is a real valued function which maps the indicator variables $X_1, \dots, X_p$ into the single value $S(X_1, \dots, X_p)$. For the sake of brevity we will denote the random variable $S(X_1, \dots, X_p)$ by S.

To formalize the default event for debitor i, we introduce the random variable $Y^{(i)}$ that can take on only the two values 1 (default) and 0 (non-default). The corresponding score $S^{(i)}$ aims to reflect the risk of this default event. In the following we will study the relation between $S^{(i)}$ and $Y^{(i)}$ for a credit portfolio consisting of $i = 1, \dots, n$ debitors. In practice, a bank typically calculates the score values $S^{(i)}$ at time t and observes the default events $Y^{(i)}$ at some future time point (for example $t + 1$). For the redesign of a rating system the bank is able to check if its historical ratings (produced by the score function) have led to reliable predictions of the default events.

An important criterion to assess the quality of the score function is its *discriminatory power*, i.e. its ability to separate the good from the bad applicants. More precisely, the bank can use its ex-post observation of the defaults to split the sample of all debitors into two subsamples: the non-defaulted and the defaulted debitors. If we denote the score cumulative distribution functions of these two samples by F_0 and F_1, discriminatory power then means that the locations of the probability masses of these distributions are as different as possible.

From now on and without loss of generality high score values stand for high default risk and low score values stand for low default risk. It is important to note that comparing ratings with respect to their discriminatory power is only meaningful if these ratings imply a reasonable ordering of the debitors. In this context reasonable means that the default probability of a good rated debitor is lower than for bad rated debitor. Formally this property means that F_1 stochastically dominates F_0. For this reason we can exclude all those ratings which do not satisfy this minimal requirement. Obviously no practitioner would take ratings into consideration that do not exhibit this form of stochastic dominance.

From a formal point of view we need to define how the distance between F_0 and F_1 should be measured. We will consider the following two approaches:

- the maximum distance measured by
$$T = \max_s \{F_0(s) - F_1(s)\},$$
- the average distance measured by the area under curve of the receiver operating characteristic (ROC) curve
$$AUC = 1 - \int F_1(s)\, dF_0(s)$$
which will shown to be equivalent to the accuracy ratio AR derived from the Lorenz curve.

In the remaining part of this section we will describe some properties of these measures. Instead of analyzing T one can also consider $R = 1 - T$ which is given by

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

If F_0 and F_1 have densities f_0 and f_1 possessing only one point of intersection, then R has the intuitive interpretation as the overlapping region of these densities. This is illustrated in Figure 1. It is obvious that for distributions F_0, F_1 on completely different supports (perfect separation), the value of T is 1. If both distributions are identical (no separation) then T equals 0. In all other cases T will take on values between 0 and 1.

In practice we have observations $S^{(i)}$ for the scores and $Y^{(i)}$ for the default events. Estimates of the cumulative distribution functions F_0, F_1 can be easily found by the empirical distribution functions

$$\widehat{F}_j(s) = \frac{\sum_i I(S^{(i)} \le s, Y^{(i)} = j)}{\sum_i I(Y^{(i)} = j)}, \quad j = 0, 1\,, \tag{2}$$

where $I(\bullet)$ denotes the indicator function. We wish to remark that

$$\widehat{T} = 1 - \widehat{R} = \max_s \left\{ \widehat{F}_0(s) - \widehat{F}_1(s) \right\}$$

also serves as the test statistic in the Kolmogorov-Smirnov test which checks the hypothesis of stochastic dominance of F_1 over F_0. Applying a Kolmogorov-Smirnov test, this minimal requirement can also be tested formally. However, our goal is to identify a score function leading to a maximal possible value of $\widehat{T}$.

The second measure for the discriminatory power introduced above is the the area under the ROC curve (Hand and Henley; 1997; Engelmann et al.; 2003; Sobehart and Keenan;

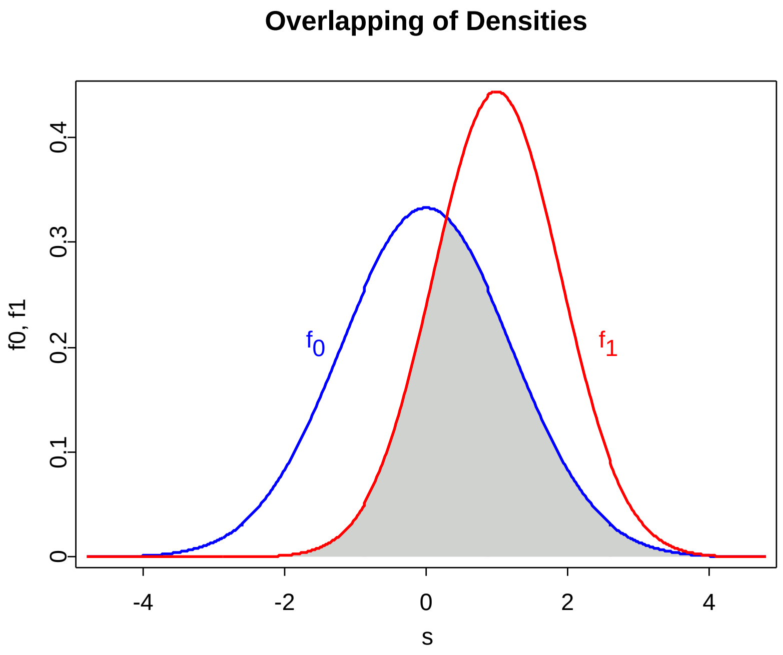

Figure 1: Overlapping area R

2001). The ROC curve is obtained by plotting cumulated percentages of the non-defaulted and defaulted debitors on the horizontal and vertical scales, respectively. If we sort the percentages of applicants from "bad" to "good" scores, this means to graph

$$1 - F_0(s) \quad \text{vs.} \quad 1 - F_1(s) \qquad \text{for all } s \in (+\infty, -\infty).$$

Figure 2 shows how the area under curve (AUC) is then calculated:

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

An optimal score would exactly separate defaults and non-defaults. The corresponding optimal ROC leads to an area under curve that fills the complete unit square so that AUC equals 1 in this case. The worst score is the one which does not contain any information about defaults and non-defaults, i.e. it randomly assigns score values to credit applicants. The corresponding ROC curve is thus identical to the diagonal and its AUC equals $\frac{1}{2}$.

The AUC is directly related to the accuracy ratio AR which is based on the Lorenz curve and its Gini coefficient (Keenan and Sobehart; 1999; Engelmann et al.; 2003). In contrast to the ROC curve the Lorenz curve plots

$$1 - F(s) \quad \text{vs.} \quad 1 - F_1(s) \qquad \text{for all } s \in (+\infty, -\infty)$$

where F denotes the cumulative distribution function of S, the score values of all debitors. Figure 3 shows the principle of the Lorenz curve. The Lorenz curve is also known as the power curve or the cumulative accuracy profile (CAP).

The optimal Lorenz curve (for a score that exactly separate defaults and non-defaults) reaches the vertical axis 100% at a horizontal percentage of $P(Y = 1)$, the probability

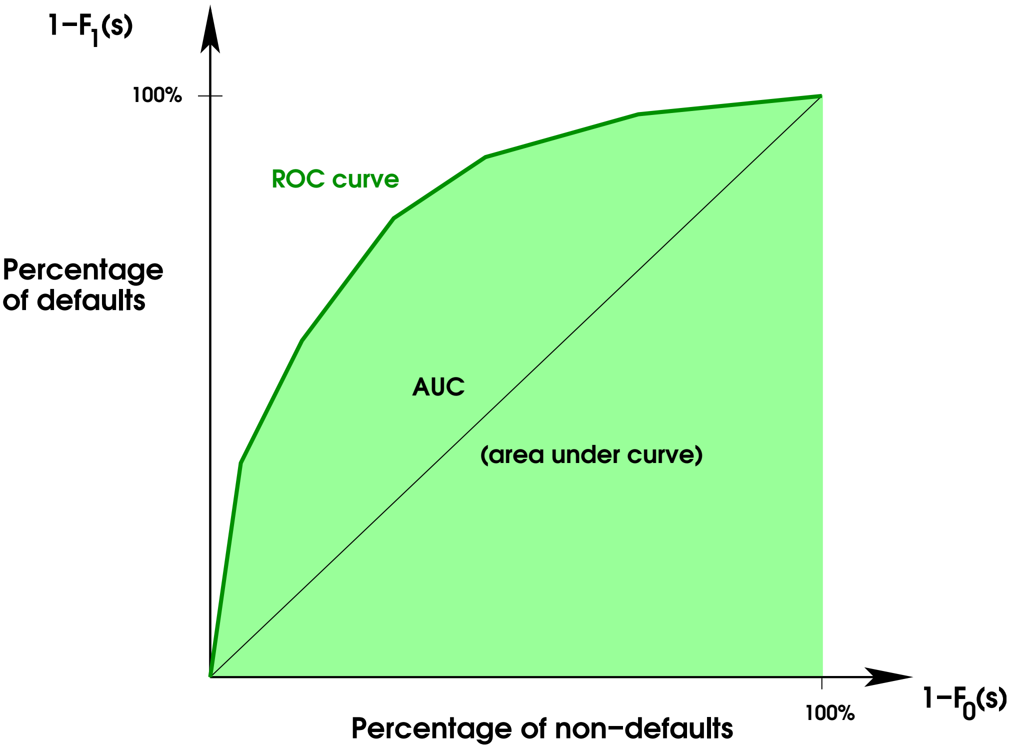

Figure 2: ROC curve for credit scores

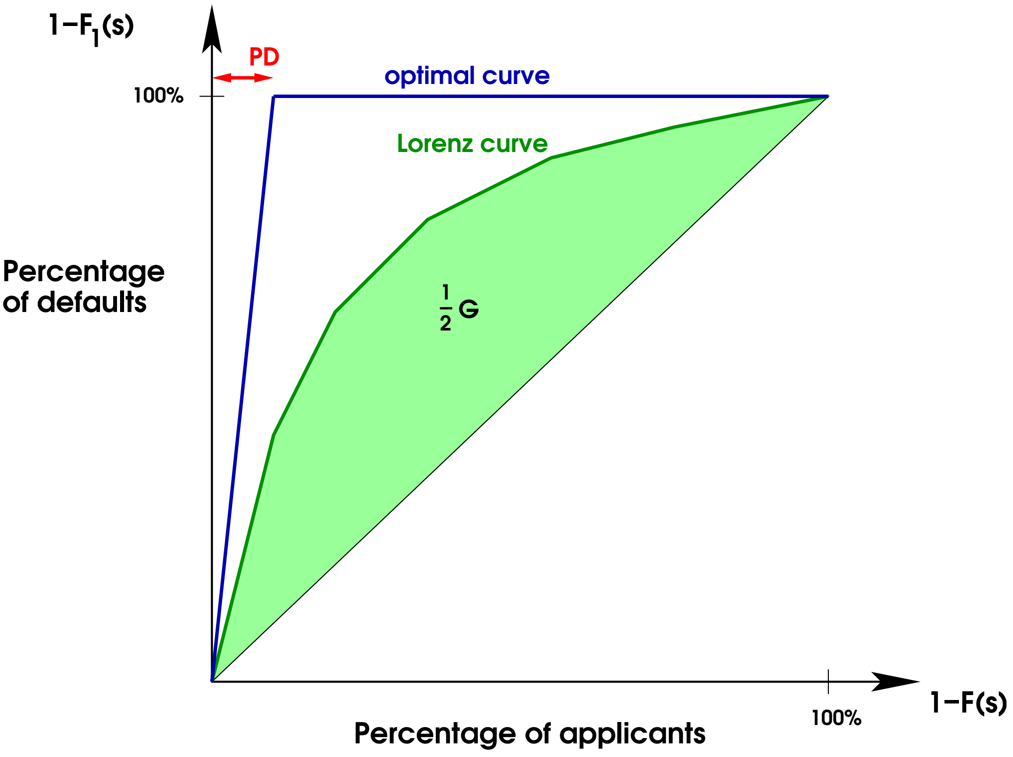

Figure 3: Lorenz curve for credit scores

of default. The worst Lorenz curve is identical to the diagonal since $F_1 = F$ in that case. A quantitative measure for the performance of a score is then based on the area between the Lorenz curve and the diagonal. The *Gini coefficient* G equals twice this area, i.e.

$$G = 2\int_{+\infty}^{-\infty} (1 - F_1)(s)\, d(1 - F)(s) - 1 = 1 - 2\int_{-\infty}^{+\infty} F_1(s)\, dF(s)\,. \tag{3}$$

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

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

The AR is a positive linear transformation of AUC which can be shown as follows: By the Bayes theorem we have

$$F = P(Y = 0)\, F_0 + P(Y = 1)\, F_1.$$

This provides

$$\begin{aligned} G &= 1 - 2\int F_1\, dF = 1 - 2\int F_1\, d\left\{P(Y = 0)F_0 + P(Y = 1)F_1\right\} \\ &= 1 - 2\,P(Y = 0)\int F_1\, dF_0 - 2\,P(Y = 1)\int F_1\, dF_1 = P(Y = 0)\,(2\,AUC - 1), \end{aligned}$$

where the relations $\int F_1\, dF_0 = 1 - AUC$, $P(Y = 1) = 1 - P(Y = 0)$ and $\int F_1\, dF_1 = \frac{1}{2}$ are used. We therefore obtain

$$AR = \frac{G}{P(Y = 0)} = 2\,AUC - 1.$$

In practice the integrals in AUC and AR are estimated by numeric integration of $\widehat{F}_1$ over $\widehat{F}_0$ or $\widehat{F}$ where $\widehat{F}$ denotes the empirical distribution function

$$\widehat{F}(s) = \frac{\sum_i I(S^{(i)} \leq s)}{n}. \tag{4}$$

The AUC and hence the accuracy ratio AR are related to the Wilcoxon rank sum test and its equivalent, the Mann–Whitney U test. Both are classical nonparametric tests to check if two distributions are identical, in our case F_0 and F_1. In its simplest form, the U test is derived for continuous score distributions. Denote by S_0 the score given non-default and by S_1 the score given default. Let $S_0^{(i)}$ and $S_1^{(k)}$ be two sample observations from the credit portfolio. The U test statistic then counts the number of pairs where $S_0^{(i)} < S_1^{(k)}$ holds, i.e. $\widehat{U} = \#\{S_0^{(i)} < S_1^{(k)}\}$ over all i, k. If the score function perfectly separates defaults and non-defaults we should obviously obtain $\widehat{U} = n_0 \cdot n_1$. If there is no relation between score and default event, then $S_0^{(i)} < S_1^{(k)}$ happens with probability $\frac{1}{2}$ such that $\widehat{U} \approx 0.5(n_0 \cdot n_1)$. Consequently, a rescaled version of the test statistics, $\widehat{U}/(n_0 \cdot n_1)$ is an estimate for

$$P\{S_0 < S_1\} = 1 - \int F_1(s)\, dF_0(s) = AUC = \frac{AR + 1}{2}.$$

In the case of discrete score distributions this equation involves an additional term for $P(S_0 = S_1)$. The U test is another way to formally test our minimal requirement of stochastic dominance of F_1 over F_0. Again we wish to stress that our aim is to find a score function leading to the maximal possible value of AUC or AR.

3 Credit Scoring under Censoring

A particular problem with credit data in practice is that we usually observe defaults and non-defaults only for a subsample of applicants. This is so because 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*. Note that condition $\mathcal{A}$ formalizes the criterion for granting credit applied by the bank under consideration.

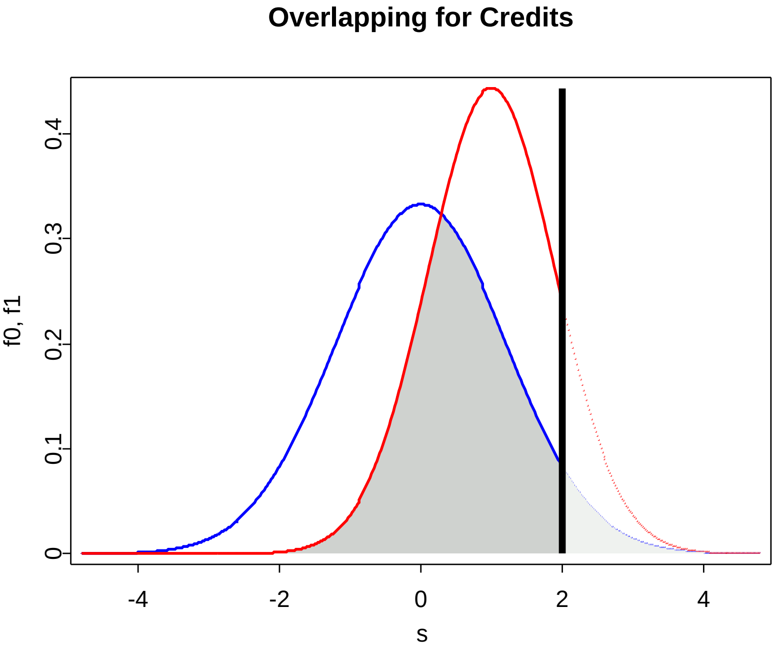

Figure 4: Truncated overlapping area for credit data

Obviously, this data sampling will result in biased estimates of all relevant parameters due to the non-representative data base. We wish 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 et al. (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 (2004) are references here.

To illustrate the effect of censoring (or sample selection) for estimating T, assume again that we have two continuous densities f_0, f_1. Assume further for a moment that the censoring condition has the intuitive form

$$\mathcal{A} = \{S \leq c\}\,, \tag{5}$$

where c is a threshold such that credit applicants are accepted for a loan if their score S is smaller than c. Figure 4 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 $\widetilde{S}$ and $\widetilde{Y}$ denote the observed parts of the score and the group indicator. Hence, we have only observations for the censored default scores $\widetilde{S}_1$ and non-default scores $\widetilde{S}_0$, while we are interested in the non-censored scores S_0 and S_1. Under the assumption (5), the relation between $\widetilde{S}_j$ and S_j $(j = 0, 1)$ is given by

$$P(\widetilde{S}_j \le s) \;=\; P(\widetilde{S} \le s|\widetilde{Y} = j) = \frac{P(\widetilde{S} \le s, \widetilde{Y} = j)}{P(\widetilde{Y} = j)} = \frac{P(S \le s, Y = j|\mathcal{A})}{P(Y = j|\mathcal{A})}.$$

From $\mathcal{A} = \{S \le c\}$ and $P(S_j \le s) = P(S \le s, Y = j)/P(Y = j)$ it follows that

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

which then shows

$$\widetilde{F}_j(s) = \frac{F_j(s)}{F_j(c)} \qquad \text{for } s \le c\,. \tag{6}$$

Here $\widetilde{F}_j$ denotes the cumulative distribution functions of $\widetilde{S}_j$. Under the assumption that S_j has a continuous distribution, (6) results in an equivalent rescaling of the densities by $F_j(c)$. AskFigure 5 illustrates this case, note the difference to Figure 4 on the vertical scale.

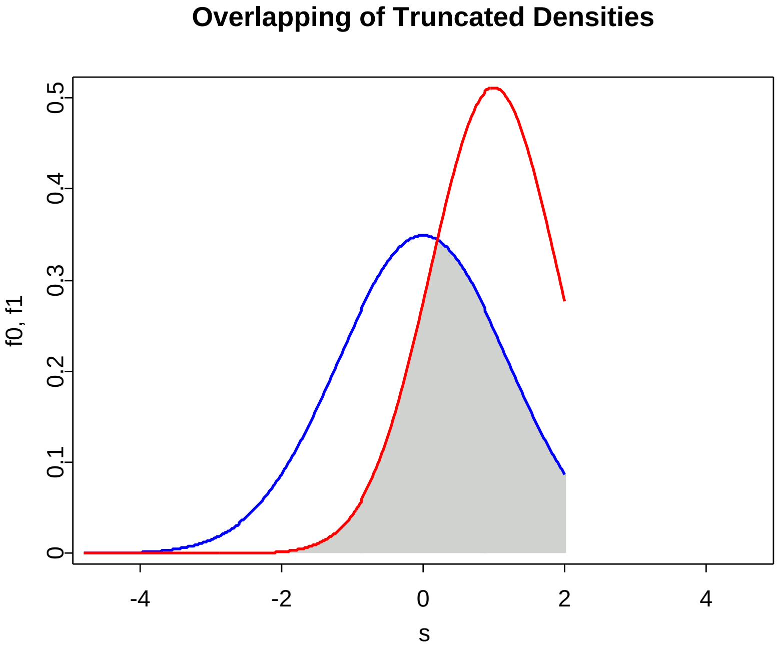

Figure 5: Observed overlapping area $\widetilde{R}$

We will now analyse the relation between $\widetilde{R}$ and R, the regions of overlapping for the censored (observed) and the non-censored (partially unobserved) sample. Computing $\widetilde{R}$ in the same way as R and using (6), would hence give

$$\widetilde{R} = \min_s \left\{ \widetilde{F}_1(s) + 1 - \widetilde{F}_0(s) \right\} = \min_s \left\{ \frac{F_1(s)}{F_1(c)} + 1 - \frac{F_0(s)}{F_0(c)} \right\}\,. \tag{7}$$

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

The difference in $T = 1 - R$ and $\widetilde{T} = 1 - \widetilde{R}$ can be considerably important as the following Monte Carlo simulation shows. We have simulated 250 data sets, each of $N = 500$ applicants for a credit. The scores $S^{(i)}$ of these applicants are generated only once and come from a normal distribution with expectation -2.5 and variance 1.44. The simulated PDs are obtained from a Logit model, i.e. we define

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

and generate the $Y^{(i)}$ as Bernoulli random variables with probability parameter $p(S^{(i)})$. We reject the 2% worst scored credit applicants. This leads to a credit portfolio consisting of $n = 490$ debitors.

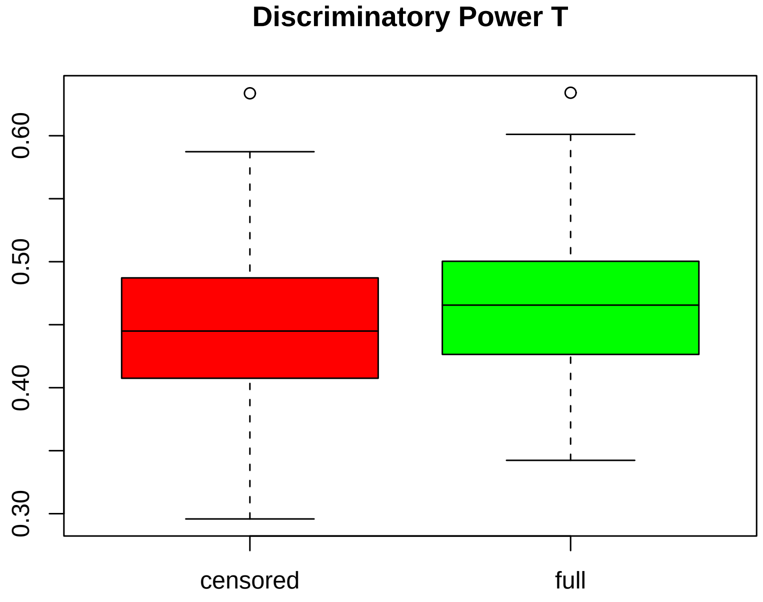

Figure 6: Difference in T (upper boxplot) and $\widetilde{T}$ (lower boxplot)

In Figure 6 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 (in 235 out of the 250 cases). A similar boxplot could be shown for $\widetilde{AR}$ and AR. So using $\widetilde{T}$ and $\widetilde{AR}$ instead of T and AR can mislead in both directions (over- and underestimation) if the bank wishes to assess the performance of its rating system. Recall that formally this comes from the fact that the acceptance condition $\mathcal{A}$ leads to conditional probabilities.

4 Inequalities for the Nonparametric Case

As we have seen in Section 3, the computation of T and AR from $\widetilde{S}_j$ requires specific assumptions on the distributions of S_j and their relations to the censoring condition $\mathcal{A}$. In the general case in which

the relation between $\mathcal{A}$ and S is completely unknown

there is no possibility to estimate these distributions beyond $\mathcal{A}$. As pointed out, this is a relevant problem for a bank redesigning its rating system since data on rejected applicants are usually not available.

A possible remedy to this problem is the calculation of lower and upper bounds on the measures for discriminatory power. Our approach is inspired by Horowitz and Manski (1998) who consider a similar censoring problem in survey nonresponse. 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 and AR. We emphasize that N typically does not contain applicants who are rejected without being rated.

4.1 Maximum Distance Measure T

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 function $F_j(s)$ 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

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

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

$$P(Y = j, \mathcal{A}) \le P(Y = j) \le P(Y = j, \mathcal{A}) + P(\overline{\mathcal{A}}). \tag{8}$$

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 (8) that

$$\alpha_j \in [\alpha_j^{low}, 1], \qquad \text{where } \alpha_j^{low} = \frac{P(\widetilde{Y} = j)P(\mathcal{A})}{P(\widetilde{Y} = j)P(\mathcal{A}) + P(\overline{\mathcal{A}})} \,. \tag{9}$$

Lemma 4.1 together with (9) 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 wish 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 the probabilities of default and non-default 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 presented in Section 5.

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 (14) and (17) 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 (2). 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

$$\widehat{P}(\widetilde{Y} = j) = \frac{n_j}{n}, \quad \widehat{P}(\mathcal{A}) = \frac{n}{N}, \quad \widehat{P}(\overline{\mathcal{A}}) = \frac{N-n}{N}. \tag{10}$$

Here n_0 denotes the number of observed non-defaults and similarly n_1 denotes 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

$$\widehat{\alpha}_j^{low} = \frac{n_j}{n_j + N - n}, \quad \widehat{\beta}_j = \frac{n_j}{N}. \tag{11}$$

Estimates for p_0^{low} and p_0^{up} can be found by substituting $\widehat{\beta}_j$, $\widehat{P}(\overline{\mathcal{A}})$ and $\widehat{\widetilde{F}}_j(s)$ into (14)–(15) and (17)–(18).

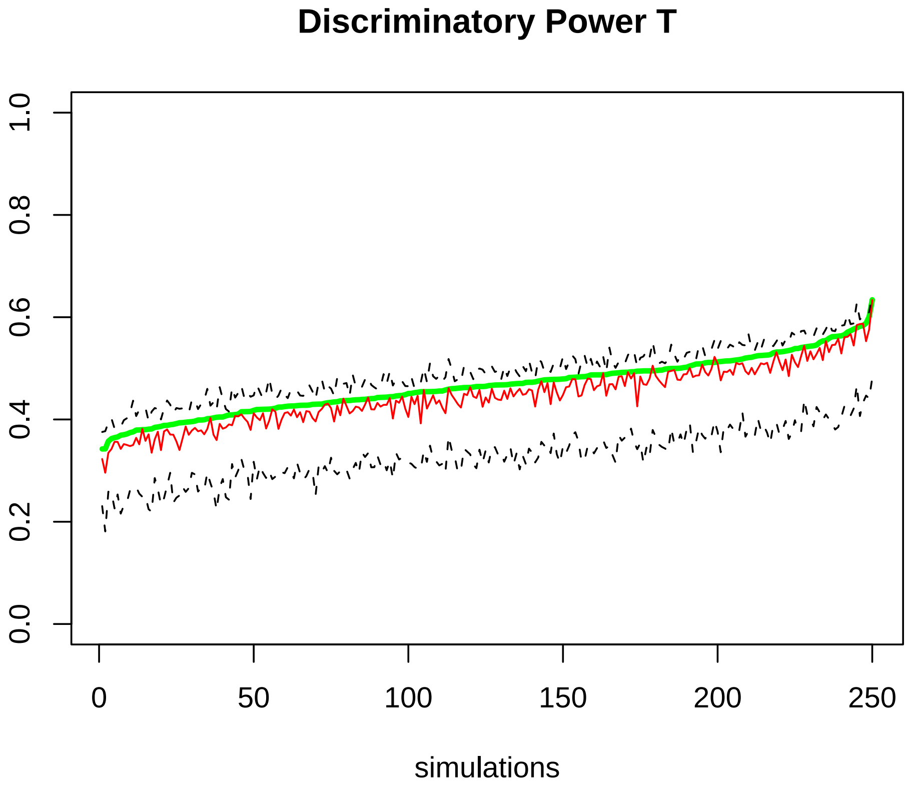

Figure 7: Estimated T (smooth solid line), $\widetilde{T}$ (solid) and bounds (dashed)

The following Monte Carlo simulation illustrates the effect of the estimated bounds. We use the previously simulated data set. Figure 7 shows estimates for T, $\widetilde{T}$ and the estimated upper and lower bounds according to Proposition 4.3 for all 250 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 so 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 these 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, lead to a correctly specified range for $\widehat{T}$.

We see that the lower bound in Figure 7 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, for example if $\mathcal{A}$ is determined by a different score $S^\star$ and we know the dependence structure between S and $S^\star$.

4.2 Accuracy Ratio AR

We consider the censored accuracy ratio

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

where $\widetilde{G}$ denotes the censored Gini coefficient and $P(\widetilde{Y} = 0) = P(Y = 0|\mathcal{A})$. Thus, analogously to $\widetilde{T}$, the Gini coefficient $\widetilde{G}$ and the accuracy ratio $\widetilde{AR}$ are biased. We will show how to obtain upper and lower bounds for both G and AR.

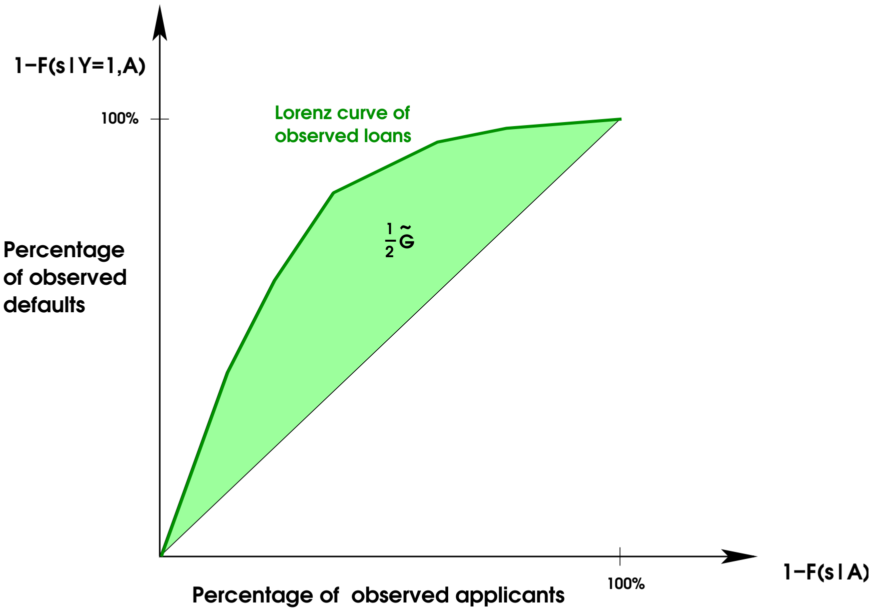
1−F(s | Y=1,A)
100%
Lorenz curve of observed loans
$\frac{1}{2}\tilde{G}$
Percentage of observed defaults
100%
1−F(s | A)
Percentage of observed applicants

Figure 8: Lorenz curve under censoring

Suppose that the Lorenz curve for the observed loans looks as in Figure 8. To obtain lower and upper bounds for the Lorenz curve of all credit applicants, we consider two extreme cases for the unobserved part: the unobserved loans possess either the lowest or the best ratings. These assignments lead to the Lorenz curves in Figure 9.

Hence, lower and upper bounds for G (and subsequently for AR) can be derived by calculating the areas under the curves in Figure 9. The resulting inequality for AR is summarized by the following proposition. As before we refer to the appendix for a detailed proof.

Proposition 4.4
Bounds for AR are given by

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

where again $\beta_j = P(Y=j,\mathcal{A})$ and

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

Let us remark that in the special case if all credit applicants are accepted, it holds $p_0^\star = \beta_0$ and $1 - p_0^\star = \beta_1$. Hence, the upper and lower bounds for the Lorenz curve as well for Gini coefficient and accuracy ratio coincide with their respective values in the non-censored case.

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

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

To illustrate the result of Proposition 4.4, we use the data from the Monte Carlo simulation in Section 4 again. Figure 10 shows the estimated AR and $\widetilde{AR}$ as well as the estimated upper and lower bounds according to all 250 simulated data sets (sorted by the estimated ARs). We find $\widehat{AR} > \widehat{\widetilde{AR}}$ in 237 (out of the 250) 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 provide 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, only the upper bound has a useful interpretation.

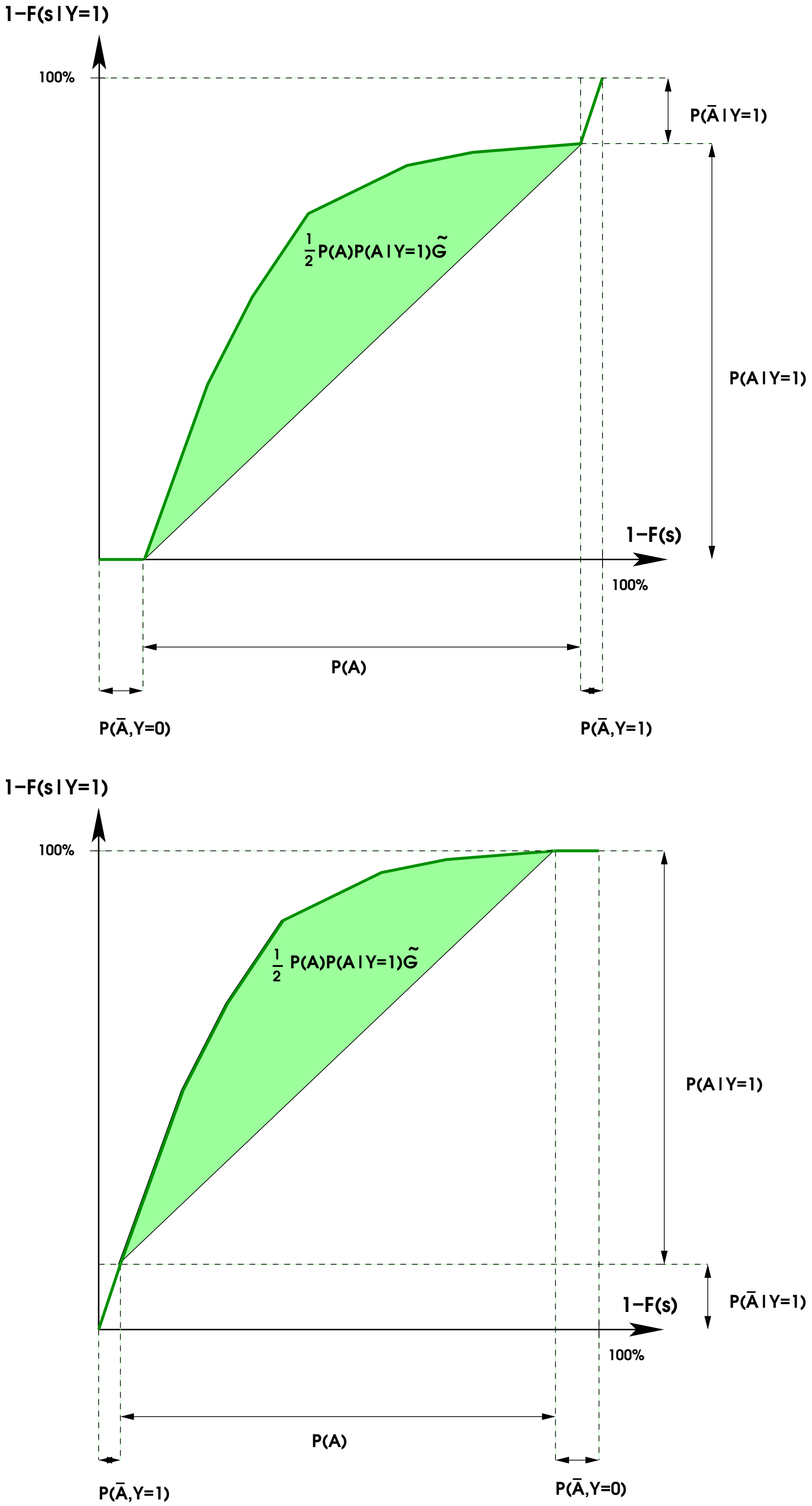

Figure 9: Lorenz curves under censoring

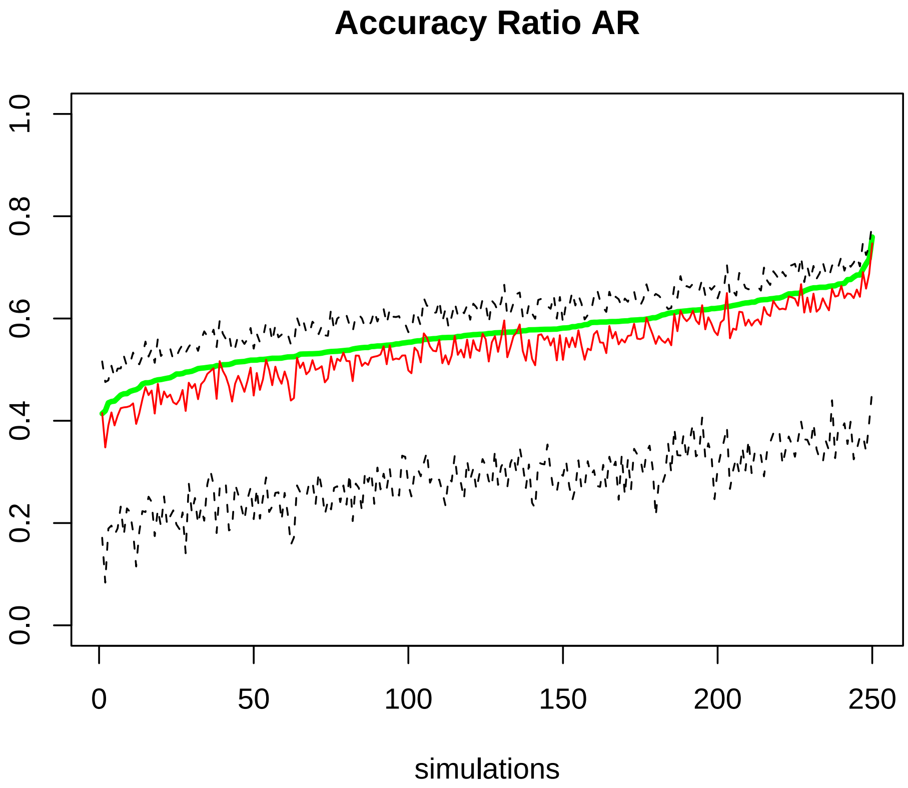

Figure 10: Estimated AR (smooth solid line), $\widetilde{AR}$ (solid) and bounds (dashed)

5 Application

Let us now consider a brief illustration on real data. We use the 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 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 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 7 and 10 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] See 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.

6 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. Corrected calculations of the criteria are possible if details about the rejected credit applicants are known. However, often no precise information about these applicants is available. For this case the paper demonstrates how to assess discriminatory power by computing lower and upper bounds of such criteria. The calculation of bounds is possible under the weak assumption that only the percentage of rejected credits is known.

Appendix: Proofs

Proof of Lemma 4.1

We have

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

hence

$$F_j(s) = \widetilde{F}_j(s) \, P(\mathcal{A} | Y = j) + P(S \leq s, \overline{\mathcal{A}} | Y = j). \tag{12}$$

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 (12), i.e.

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

A lower bound for $F_j(s)$ is given by omitting the second term of (12) 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 (9). □

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 R and later on transfer them into bounds for T.

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

$$\begin{aligned} F_1(s) + 1 - F_0(s) &\geq \alpha_1 \widetilde{F}_1(s) + \alpha_0 \{1 - \widetilde{F}_0(s)\} \\ &= \frac{\beta_1}{1 - p} \widetilde{F}_1(s) + \frac{\beta_0}{p} \{1 - \widetilde{F}_0(s)\} \end{aligned} \tag{13}$$

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:

$$p_s^{low} = \begin{cases} \beta_0 & \text{if } \gamma_s < \beta_0, \\ \beta_0 + P(\overline{\mathcal{A}}) & \text{if } \gamma_s > \beta_0 + P(\overline{\mathcal{A}}), \\ \gamma_s, & \text{otherwise,} \end{cases} \tag{14}$$

and

$$\gamma_s = \frac{\sqrt{\beta_0\{1-\widetilde{F}_0(s)\}}}{\sqrt{\beta_0\{1-\widetilde{F}_0(s)\}} + \sqrt{\beta_1\widetilde{F}_1(s)}}\,. \tag{15}$$

The upper and lower thresholds in (14) are consequences of the bounds in (8).

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

$$\begin{aligned} F_1(s) + 1 - F_0(s) \;\; &\leq \;\; 2 - \alpha_1\{1-\widetilde{F}_1(s)\} - \alpha_0\widetilde{F}_0(s) \\ &= 2 - \frac{\beta_1}{1-p}\{1-\widetilde{F}_1(s)\} - \frac{\beta_0}{p}\widetilde{F}_0(s). \end{aligned} \tag{16}$$

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

$$p_s^{up} = \begin{cases} \beta_0 & \text{if } \delta_s < \beta_0, \\ \beta_0 + P(\overline{\mathcal{A}}) & \text{if } \delta_s > \beta_0 + P(\overline{\mathcal{A}}), \\ \delta_s, & \text{otherwise,} \end{cases} \tag{17}$$

and

$$\delta_s = \frac{\sqrt{\beta_0\widetilde{F}_0(s)}}{\sqrt{\beta_0\widetilde{F}_0(s)} + \sqrt{\beta_1\{1-\widetilde{F}_1(s)\}}}\,. \tag{18}$$

Combining the results we obtain

$$\begin{aligned} &\frac{\beta_1}{1-p_s^{low}}\,\widetilde{F}_1(s) + \frac{\beta_0}{p_s^{low}}\{1-\widetilde{F}_0(s)\} \\ &\qquad \leq F_1(s) + 1 - F_0(s) \leq 2 - \frac{\beta_1}{1-p_s^{up}}\,\{1-\widetilde{F}_1(s)\} - \frac{\beta_0}{p_s^{up}}\,\widetilde{F}_0(s) \end{aligned} \tag{19}$$

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

Proof of Proposition 4.4

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.

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

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

Consider first the lower bound for AR. From the first plot of Figure 9 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

$$AR^{low} = \frac{G^{low}}{p_0} = \left(\widetilde{AR}+1\right)\frac{\beta_0\beta_1}{p_0p_1} - 1\,. \tag{20}$$

We now use the same approach for the upper bound of AR. From the second plot in Figure 9 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

$$AR^{up} = \frac{G^{up}}{p_0} = \left(\widetilde{AR}-1\right)\frac{\beta_0\beta_1}{p_0p_1} + 1\,. \tag{21}$$

Hence, we obtain together with (20)

$$\left(\widetilde{AR}+1\right)\frac{\beta_0\beta_1}{p_0p_1} - 1 \leq AR \leq \left(\widetilde{AR}-1\right)\frac{\beta_0\beta_1}{p_0p_1} + 1\,. \tag{22}$$

To achieve the minimal value for the lower and the maximal value for the upper bound, it is obvious that $p_0p_1 = p_0(1-p_0)$ must be maximal. It is important to note that p_0 cannot vary freely since from (8) 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. (2004). Does reject inference really improve the performance of application scoring models?, *Journal of Banking and Finance* **28**(4): 857–874.

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. (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.