Uses and Misuses of Measures for Credit Rating Accuracy

Version: April 28, 2003

Alfred Hamerle[1], Robert Rauhmeier[2], Daniel Rösch[3]

University of Regensburg

Key Words: Credit Rating, Basel II, Performance Measurement, CAP, ROC, Accuracy Ratio, Power Curve, Gini

---

[1] *Prof. Dr. Alfred Hamerle,* Department of Statistics, Faculty of Business and Economics, University of Regensburg, 93040 Regensburg, Germany, Phone: +49-941-943-2588, Fax : +49-941-943-4936, Email: Alfred.Hamerle@wiwi.uni-regensburg.de

[2] *Robert Rauhmeier*, Department of Statistics, Faculty of Business and Economics, University of Regensburg, 93040 Regensburg, Germany, Phone: +49-941-943-2751, Fax : +49-941-943-4936 Email: Robert.Rauhmeier@wiwi.uni-regensburg.de

[3] *Dr. Daniel Rösch,* Department of Statistics, Faculty of Business and Economics, University of Regensburg, 93040 Regensburg, Germany, Phone: +49-941-943-2752, Fax : +49-941-943-4936 Email: Daniel.Roesch@wiwi.uni-regensburg.de

---

Uses and Misuses of Measures for Credit Rating Accuracy

Abstract

The New Basel Capital Accord will allow the determination of banks' regulatory capital requirements due to default probabilities which are estimated and forecasted from internal ratings. External ratings from rating agencies play fundamental roles in capital and credit markets. Discriminatory power of internal and external ratings is a key requirement for the soundness of a rating system in general and for the acceptation of a bank's internal rating systems under Basel II. Statistics such as the area under a receiver operating characteristic or the accuracy ratio, are widely used in practice as measures for the performance. This note shows that such measures should only be interpreted with caution. Firstly, the outcomes of the measures depend not only on the discrimination power of the rating system but mainly on the structure of the portfolio under consideration. Thus, the absolute values achieved do not measure the performance of a rating system solely. Secondly, comparisons of the outcomes between different portfolios, different time periods or both may be misleading. As a positive result we show that the value achieved by a rating system which predicts all default probabilities correctly can not be beaten.

1 The Problem

Credit ratings from external rating agencies are widely used in practice as key indicators for a borrower's inherent credit risk. Within the New Basel Capital Accord banks will be allowed to determine their regulatory capital requirements due to default probabilities which are estimated and forecasted from internal ratings (IRB Approach). Before its approval by the supervisory authority a bank has to show that its rating system meets the requirements described in §237 ff of the Consultative Document.

One key requirement for a sound rating system in general and for an internal rating in particular is to "demonstrate an ability to differentiate risk, have predictive and discriminatory power […] and ensure that ratings are designed to distinguish risk rather than to minimise regulatory capital requirements"[1]. So far there are no explicit instructions, but in practice many banks and rating agencies use so-called Gini curves to indicate the power of their ratings to discriminate between "good" and "bad" credits, respective non-defaulters and defaulters[2]. There are several other termini for this methodology of performance measuring. For example Sobehart et al. (2000, 2001) use Cumulative Accuracy Profiles (CAPs) and Receiver Operating Characteristics (ROCs), while Liebig/Nyberg (1999) refer to them as Power Curves[3]. Also one-dimensional measures are derived from these graphical illustrations, such as Accuracy Ratio (AR) and Area Under a Receiver Operating Characteristic (AUROC), in particular when two or more rating systems are compared. Rating systems which perfectly discriminate between defaulters and non-defaulters have an AR of 100%.

---

[1] See Basel Committee on Banking Supervision (2001), §264.

[2] See Basel Committee on Banking Supervision (2000a), p. 38. and Blochwitz et al. (2000).

[3] See Basel Committee in Banking Supervision (2000b), p. 121.

That means to each borrower who does not default a better rating has been assigned than to each defaulter. On the other extreme a completely so-called “non-informative” rating system has an AR of 0%. Therefore one might be tempted to postulate cut-offs such as “an AR of x%” for ratings to qualify for the IRB Approach or to assess their discriminative power in general.

Within this context the purpose of the present paper is to clarify some fundamental interpretations on applications and frontiers of such measures. We interpret the measures within the context of Basel II, in that we assume that each borrower exhibits an (unknown) default probability. Firstly, we show that the AR is identical to the well established Somers’D (1962), known for a long time as a measure of association between two ordinal variables. Then we show that outcomes of these performance measures do not measure the discrimination ability of a rating system solely. Rather they are mainly functions of the underlying default probabilities of the borrowers in the portfolio under consideration. Secondly, it is then straightforward to indicate when a comparison of rating systems makes sense and when it does not. Thirdly, as a positive result we demonstrate that a rater who predicts the PDs of the borrowers correctly is expected to achieve the best possible value for the performance measures. These findings are presented in section three. In section four a simple example is given for undesirable results of the misuse of accuracy measures regarding the approval of the IRB Approach in practice. Section five concludes.

2 The Model for the Default Process

We assume a discrete time hazard rate process for the random default event. The default event occurs if the return Y_{nt} on firm *n*'s assets at time *t* hits some threshold α_{nt}, given that the firm did not default before *t* (n=1,…,N_t, t=1,…,T). Let Y_{nt}^* be the indicator variable with

$$Y_{nt}^* = \begin{cases} 1 & \text{borrower } n \text{ defaults at time } t \\ 0 & \text{else} \end{cases}$$

then

$$Y_{nt} \le \alpha_{nt} \quad \Leftrightarrow \quad Y_{nt}^* = 1 \qquad (* 1).$$

Furthermore let λ_{nt} denote the probability of default (PD), given that the firm has survived until $t-1$, i.e.

$$\lambda_{nt} = P\left(Y_{nt}^* = 1 \,\middle|\, Y_{nt-1}^* = 0 \right) = P\left(Y_{nt} \le \alpha_{nt} \,\middle|\, Y_{nt-1}^* = 0 \right) \qquad (* 2).$$

Thus, the realization of the random variable "default of firm *n* at time *t*" is governed by its PD if a firm has survived until $t-1$. Therefore λ_{nt} is the true but unknown probability of default of firm *n* at time *t*. As Hilden et al. (1978, p. 240) noted: "…[λ_{nt}] is an elusive concept. However, given that such probabilities are thought to be conceptually well defined, there can be no disagreement that they are the unknown parameters which the as-

signed probabilities serve to estimate“. This view is also consistent with the New Basel Accord. Though a rating system itself may be designed to measure relative risk (i.e. a ranking order), estimates for PDs which measure absolute risk are input quantities for the determination of economic and regulatory capital requirements. Or in the words of the Basel Committee (2000b), p. 121: “In practice, we are not able to classify firms into ‘will default’ and ‘will not default’ categories, we can only hope to estimate probabilities of default. Therefore, testing the performance of a default model means to investigate its ability to discriminate between different levels of default risk.”

3 Properties of Performance Measures

Assume a rater who assigns ratings to all N_t borrowers under consideration in a bank or a rating agency due to his information before time t (“out-of-time”). These ratings may be ordinal rankings, metric scores or PD forecasts. Next the index “t” is skipped for convenience. Then the borrowers are ordered due to their ratings which can be assumed as ordinal numbers $R_1 < R_2 < ... < R_K$ $(K \le N)$ in ascending order of their default risk. The default event is a dichotomy, so as described in Agresti (1984) the two groups, the defaulters and the non-defaulters in the subsequent year are compared and the conditional distributions of the ordered labels can be displayed in a $2 \times K$ table.

[*** Insert Table 1 about here ***]

π_{rj} denote the joint probabilities of a borrower falling into category *r* and exhibiting rating R_j (r=0,1; j=1,…,K), π_{+j} denote the percentages of borrowers in the portfolio who exhibit rating R_j (j=1,…,K), and $\bar{\lambda} = \frac{1}{N}\sum_{n=1}^{N}\lambda_i$ is simply the average default probability.

In practice Cumulative Accuracy Profiles (CAPs) are often used to get a visual, qualitative assessment of the performance of rating systems (e.g. see Sobehart et al., 2000). The CAP is a plot of the fraction of the default rate (Ordinate) which is captured by the according fraction of borrowers (Abscissa). Borrowers are ordered in descending order of their default risk (starting with the riskiest) by their assigned Rating R_j. Exhibit 1 shows an illustrative CAP (dashed line) and an "ideal line" (dotted line), which would result if all defaulters would be arranged primarily to the non-defaulters. A so-called non-informative rating system would result in a CAP which is identical to the 45°-diagonal. Thus, the closer the empirical CAP is to the "ideal line" the better is the rater's ability assessed to separate defaulters from non-defaulters.

[*** Insert Exhibit 1 about here ***]

In order to condense the inherent information of the CAP into a one-dimensional measure the Accuracy Ratio (AR) is calculated as the ratio of areas: $AR = \frac{A}{A+B}$. An illustration of

a CAP and AR is displayed in Exhibit 1. In our framework the AR could simply be computed after some geometrical considerations as

$$AR = \frac{1}{0.5\left(1-\bar{\lambda}\right)}\left[0.5\pi_{+K}\frac{\pi_{1K}}{\bar{\lambda}} + \pi_{+K-1}\frac{\pi_{1K}}{\bar{\lambda}} + 0.5\pi_{+K-1}\frac{\pi_{1K-1}}{\bar{\lambda}} + \right.$$
$$\left. + \pi_{+K-2}\frac{\sum_{j=K-1}^{K}\pi_{1j}}{\bar{\lambda}} + 0.5\pi_{+K-2}\frac{\pi_{1K-2}}{\bar{\lambda}} + \ldots + \pi_{+1}\frac{\sum_{j=2}^{K}\pi_{1j}}{\bar{\lambda}} + 0.5\pi_{+1}\frac{\pi_{11}}{\bar{\lambda}} - 0.5\right] \quad (* 3).$$

Another measure which is often used is the Area Under a Receiver Operating Curve (AUROC). As shown for example in Agresti (1984) or in Engelmann et al. (2002) the statistics AUROC and AR are equivalent with respect to their information content. The following relation holds

$$AR = 2\left(AUROC - 0.5\right) \quad (* 4).$$

If one has a sample of rated borrowers and realized defaults the calculated CAP, the AR and AUROC from the empirical data are outcomes of random variables. As shown in Engelmann et al. (2002) a sample U-statistic due to Mann-Whitney is equivalent to the sample AUROC and is an unbiased estimator for the population AUROC. Thus in principle, using the sample data, confidence intervals for the expectation can be computed and tests can be conducted. As we show below these expectations exhibit some special properties.

The above measures can be alternatively expressed within well-known measures of association, see e.g. Agresti (1984). Let X_0 and X_1 be the column numbers of the rating of borrowers selected randomly from the non-defaulters and the defaulters, independently from each other. Within the default-mode framework it is interesting to check if X_1 tends to be larger than X_0. Then AR can be written as

$$\begin{aligned} AR &= P(X_1 > X_0) - P(X_0 > X_1) \\ &= \sum\sum_{i<j} \frac{\pi_{0i}}{\pi_{0+}} \cdot \frac{\pi_{1j}}{\pi_{1+}} - \sum\sum_{i>j} \frac{\pi_{0i}}{\pi_{0+}} \cdot \frac{\pi_{1j}}{\pi_{1+}} \end{aligned} \qquad (* 5)$$

Since the denominator of the sum in (* 5) is simply $\bar{\lambda}(1-\bar{\lambda})$ the AR can be rewritten as (see the Appendix for details)

$$\begin{aligned} AR &= \frac{1}{\bar{\lambda}(1-\bar{\lambda})} \left(\sum\sum_{i<j} \pi_{0i} \cdot \pi_{1j} - \sum\sum_{i>j} \pi_{0i} \cdot \pi_{1j} \right) \\ &= \frac{1}{1-\left[\bar{\lambda}^2 + (1-\bar{\lambda})^2\right]} \left(2\sum\sum_{i<j} \pi_{0i} \cdot \pi_{1j} - 2\sum\sum_{i>j} \pi_{0i} \cdot \pi_{1j} \right) \\ &= Somers'\, D \end{aligned} \qquad (* 6)$$

which is known as *Somers' D* due to Somers (1962) if the rows are interpreted as ordered variables. The first term in the brackets of the nominator in (* 6) is usually called the probability of concordance, the second term is called the probability of discordance.

Proposition 1:

If each borrower possesses a default probability, AUROC and AR depend on the true underlying PDs of the borrowers in the portfolio under consideration.

To see this, define the random variable $\widetilde{R}$ as the column number of the rating of a borrower selected randomly from the whole distribution, write the joint probabilities in (* 5) and (* 6) as

$$\begin{aligned}\pi_{rj} &= P\left(D = r, \widetilde{R} = j\right) \\ &= P\left(D = r \,\middle|\, \widetilde{R} = j\right) \cdot P\left(\widetilde{R} = j\right)\end{aligned} \qquad (* 7)$$

and insert these expressions into (* 6). The AR then becomes

$$\begin{aligned}AR = \frac{1}{\bar{\lambda}\left(1-\bar{\lambda}\right)}\Bigg(&\sum_{i<j}\sum P\left(D=0 \,\middle|\, \widetilde{R}=i\right)\cdot P\left(\widetilde{R}=i\right)\cdot P\left(D=1 \,\middle|\, \widetilde{R}=j\right)\cdot P\left(\widetilde{R}=j\right) \\ &- \sum_{i>j}\sum P\left(D=0 \,\middle|\, \widetilde{R}=i\right)\cdot P\left(\widetilde{R}=i\right)\cdot P\left(D=1 \,\middle|\, \widetilde{R}=j\right)\cdot P\left(\widetilde{R}=j\right)\Bigg)\end{aligned} \qquad (* 8)$$

where $\bar{\lambda}$ is the a priori average default probability. $P\left(\widetilde{R}=j\right)$ is simply the percentage of borrowers who are assigned to rating R_j (j=1,…,K) and $P\left(D=1 \,\middle|\, \widetilde{R}=j\right)$ is the average of the true default probabilities of all borrowers who are assigned to label R_j (j=1,…,K). ■

A simple example is provided with a portfolio which consists of borrowers who exhibit one of two PDs. Let there be 500 borrowers with a PD of $\lambda_1 = 1\%$ each and 500 borrowers with a PD of $\lambda_2 = 5\%$ each and assume that the rater rates the borrowers due to their true PDs. Then the above probabilities are

$$\bar{\lambda} = \frac{1}{N}\sum_{i=1}^{N}\lambda_i = 0.03,$$

$$1-\bar{\lambda} = 0.97,$$

$$P\left(\tilde{R}=1\right) = 0.5,$$

$$P\left(\tilde{R}=2\right) = 0.5$$

$$P\left(D=1\middle|\tilde{R}=1\right) = 0.01,\ P\left(D=0\middle|\tilde{R}=1\right) = 0.99$$

$$P\left(D=1\middle|\tilde{R}=2\right) = 0.05,\ P\left(D=0\middle|\tilde{R}=2\right) = 0.95.$$

The AR is calculated as

$$\begin{aligned} AR = & \frac{P\left(D=0\middle|\tilde{R}=1\right)\cdot P\left(\tilde{R}=1\right)}{1-\bar{\lambda}} \cdot \frac{P\left(D=1\middle|\tilde{R}=2\right)\cdot P\left(\tilde{R}=2\right)}{\bar{\lambda}} \\ & - \frac{P\left(D=1\middle|\tilde{R}=1\right)\cdot P\left(\tilde{R}=1\right)}{\bar{\lambda}} \cdot \frac{P\left(D=0\middle|\tilde{R}=2\right)\cdot P\left(\tilde{R}=2\right)}{1-\bar{\lambda}} \\ = & \frac{0.99\cdot 0.5}{0.97}\cdot\frac{0.05\cdot 0.5}{0.03} - \frac{0.01\cdot 0.5}{0.03}\cdot\frac{0.95\cdot 0.5}{0.97} \approx 0.344 \end{aligned}$$

and AUROC $\approx$ 0.672. Exhibit 2 contains the values of AR for varying values of λ_1 (PD1) and λ_2 (PD2), each between 0.01% and 99% with $\lambda_1 \leq \lambda_2$. Note that the values of AR strongly depend on the difference between the PDs and can take nearly any value between 0 and 1 – although it is always assumed that the rater knows all PDs and assigns them correctly! Moreover, in the case that all PDs in the portfolio are equal, the AR is always 0.

[*** Insert Exhibit 2 about here ***]

Two remarks on proposition 1 should be noted:

- A rater's attainable discrimination power is predetermined by the structure of the portfolio.

The reason for this lies in the fact that AR and AUROC are measures for the association between ordinal responses. Although a rating may be ordinal, the true default probabilities of the borrowers are metric and determine the outcome. A rater's AR can only move between certain limits which are functions of properties of the portfolio which he rates rather than functions of his personal discrimination ability solely.

- In addition, the measures do not indicate the riskiness of the portfolio.

A bank with a rather homogenous portfolio of high or middle quality loans may exhibit a much lower measure than a bank with a high risk portfolio with higher dispersion of default probabilities. We will provide an example for this in section 4.

Proposition 2:

In general, the discrimination abilities of two raters who rate different portfolios at the same time or the same portfolio in different periods, or both, can not be compared by AUROC and AR.

To see this note that AR and AUROC are functions of the individual PDs of all borrowers in a portfolio. If this portfolio is compared with another portfolio which differs from the first in at least one PD then AR and AUROC will also differ in general. The same is true for the comparison between different time periods. Even if the same borrowers are in the portfolio, but if at least one default probability of a borrower changes, AR and AUROC will also change despite the same discrimination ability. ■

While comparisons across portfolios and across time do not seem meaningful, we now ask for the upper limit of the outcome within the same portfolio at the same point in time for different ranking orders. This is summarized in proposition 3.

Proposition 3:

The AUROC and AR for given PDs within a bank's portfolio which is achieved by a rater who knows all PDs and assigns them correctly can not be beaten.

Assume that a rater arranges all borrowers due to their true PDs. In this case the AR can be transformed into the Gini coefficient which is known from standard statistic text books. The proof is given in the Appendix. There it is shown that

$$\begin{aligned} AR &= \frac{1}{1-\bar{\lambda}} \sum_{i=1}^{K} \left(x_{i-1}\ y_i - x_i\ y_{i-1} \right) \\ &= \frac{1}{1-\bar{\lambda}} \cdot Gini \end{aligned} \qquad (* 9)$$

where $x_i = \sum_{l=1}^{i} \pi_{+l}$ is the cumulative share of borrowers and $y_i = \frac{1}{\bar{\lambda}} \sum_{l=1}^{i} \pi_{1l}$ is the cumulative proportion of the average default probability. Note that the trait is the default probability. See for example Lee (1997) for this definition of a Gini coefficient.

Note that AR equals $\left(1-\bar{\lambda}\right)^{-1} \cdot \text{Gini}$ by definition only if all borrowers are correctly ranked according to their default probabilities.

Another notation of the Gini coefficient in individual form is

$$Gini = 1 + \frac{1}{N} - \frac{2}{N^2\ \bar{\lambda}} \left(N\ \lambda_{(1)} + (N-1)\, \lambda_{(2)} + \ldots + \lambda_{(N)} \right) \qquad (* 10)$$

where $\lambda_{(1)}, \ldots, \lambda_{(N)}$ are the ordered default probabilities from the lowest to the highest. If any two of the borrowers are ordered incorrectly, it can be easily seen that the expression

(* 10) becomes smaller since the term in brackets becomes larger. Hence, any deviation from the correct ordering of the default probabilities diminishes expression (* 10) and thus AR. Furthermore (* 10) is no longer a Gini coefficient. ■

In practice, statistical tests can be employed which compare the discrimination power of two rating systems for given sample data, see DeLong et al. (1988) or Engelmann et. al. (2002). In statistical terms, this is a test on the equality of two (population) AUROCs. Given two rating systems A and B, one tests if the (expected) AUROC of A is different from the (expected) AUROC of B. The null hypothesis is

$$H_0 : AUROC_A - AUROC_B = 0 \qquad (* 11)$$

against the alternative of inequality.

If this kind of test is conducted between different portfolios, or different time periods, or both, then proposition 2 holds and the null is generally false by construction whether or not the discrimination power is equally good. Thus, the discrimination power can not be assessed by the test result. Only if it is guaranteed that the null is true when the discrimination power is the same for both rating systems, a meaningful test can be provided. This is in general only the case if it is carried out within the same underlying portfolio and time period.

4 Practical Impacts and Consequences for Approval of the IRB Approach

In this section some practical impacts and consequences of the preceding theoretical considerations are mentioned. Although no precise guidelines for the approval of the IRB Approach are determined by the supervisors, one might have in mind postulations such as "a rating system has to attain at least an AUROC of – say for example – 65%". This requirement could be senseless, misleading and even result in converse actions as it is demonstrated in the following simple example.

Consider a bank A with 1000 obligors, 500 with PD of 1%, the other 500 with a PD of 2% each. Bank A could be suggested as a "bank with medium quality obligors in its portfolio". Suppose bank A's rating system orders all borrowers according to their true PD (that is, a rating system which orders all borrowers correctly due to their inherent default risk). Therefore an $\text{AUROC}_{\text{A}} \approx 0.585$ is calculated.

Another bank B with a "low quality portfolio", for example 500 obligors with a PD of 2.5% and 500 obligors with a PD of 20% each, achieves an $\text{AUROC}_{\text{B}} \approx 0.719$, if bank B's rating system also ranked the obligors perfect according to their PD's.

Now think of a non perfect rating system applied by bank B: 75 obligors with PD 2.5% are classified to rating R_2 instead of the "correct" rating R_1, and on the other hand 75 obligors with a PD of 20% are falsely rated into R_1 instead of R_2. Thus, altogether 150 out of

1000 Obligors are falsely rated. This non perfect rating system achieves an AUROC$^{f}_{B} \approx 0.653$.

Note that AUROC$^{f}_{B}$ > AUROC$_{A}$. This surplus is due to the structure of the portfolios, and is not an indicator for the quality of the rating system!

Bank A fails the minimum requirement hurdle of an AUROC of 0.65, even though it uses a perfect rating system. Bank B achieves an AUROC which exceeds the target of 0.65, even though applying the non perfect rating system.[4] Therefore bank B gets the approval to use the IRB Approach, and bank A does not, albeit bank B should be better off in discriminating since the PDs in bank B's portfolio differ more (2.5% and 20%), whereas in bank A's portfolio there is very little discrepancy between the PDs (1% vs. 2%).

The situation may become even more perverse if bank A with its perfect rating system attempts to pass the 0.65 hurdle. Then bank A may accommodate obligors with higher PDs which implicitly goes along with reducing the quality of the portfolio, shifting the default rate, and increasing risk. Starting from the existing portfolio, bank A could take in another 500 Obligors with PDs of 4% for example, in order to achieve the threshold of 0.65 for the AUROC.

---

[4] In our example the portfolios of banks A and B distinguish in their overall default rate. Examples can be constructed where the default rate of bank A's and bank B's portfolio are equal and all our statements hold as well.

5 Conclusion

The present paper provides some guidelines for uses of measures for the discriminatory power of credit rating systems. Three main statements were made:

→ The outcomes of the performance measures strongly depend on the structure of the true default probabilities in the underlying portfolio. Thus, the measures AR and AUROC are not able to separate properties of the rating system from properties of the rated portfolio. However, this is a fundamental necessary assumption for the construction of measures designated to judge rating systems. As a consequence, their magnitudes are not interpretable regarding the discriminatory power of the rating system.

→ It follows that rating systems generally cannot be compared across time and across portfolios. Moreover, the construction of confidence intervals and tests for the expected values of the measures applied to different portfolios is not much more than a mathematical exercise, but without significant value for practice.

→ The highest measure is expected to be earned by a rating system which assesses all true PDs correctly.

As a positive result one can conclude that comparisons of ratings at the same point in time within one portfolio can be conducted. Then standard tests can be employed using the methodology described in DeLong et al. (1988) or applied in Engelmann et. al. (2002).

References

Agresti, A, 1984, Analysis of Ordinal Categorical Data, New York et al.

Basel Committee on Banking Supervision, 2000a, Range of Practice in Banks' Internal Ratings Systems, Basel, January 2000

Basel Committee on Banking Supervision, 2000b, Credit Ratings and Complementary Sources of Quality Information, August 2000

Basel Committee on Banking Supervision, 2001, The New Basel Capital Accord, Consultative Document, January 2001

Blochwitz, S, Liebig, T, Nyberg, M, 2000, Benchmarking Deutsche Bundesbank's Default Risk Model, The KMV Private Firm Model and Common Financial Ratios for German Corporations, Working Paper, Deutsche Bundesbank

DeLong, ER, DeLong DM, Clarke-Pearson, DL, 1988, Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach, Biometrics, 44, 837-845

Engelmann, B, Hayden, E, Tasche, D, 2003, Testing for Rating Accuracy, Risk, 16, January, 82-86

Hilden, J, Habbema, JDF, Bjerregaard, B, 1978, The Measurement of Performance in Probabilistic Diagnosis: III. Methods Based on Continuous Functions of the Diagnostic Probabilities, Methods of Information in Medicine, 17, 238-246

Lee, WC, 1999, Probabilistic Analysis of Global Performances of Diagnostic Tests: Interpreting the Lorenz Curve-based Summary Measures, Statistics in Medicine, 18, 455-471

Liebig, T, Nyberg, M, 1999, Testing Results of Credit Monitor (KMV) for listed German Companies, Deutsche Bundesbank

NN, 2002, Validation of Internal and External Rating Methods – An Overview, Deutsche Bundesbank, mimeo

Sobehart, JR, Keenan, SC, 2001, Measuring Default Accurately, Credit Risk Special Report, Risk, 14, March, 31-33

Sobehart, JR, Keenan, SC, Stein, RM, 2000, Benchmarking Quantitative Default Risk Models: A Validation Methodology, Moody's Rating Methodology

Somers, RH, 1962, A New Asymmetric Measure of Association for Ordinal Variables, American Sociological Review, 27, 799-811

Stein, RM, 2002, Benchmarking Default Prediction Models: Pitfalls, Moody's KMV, Technical Report #020305, New York

Appendix

We show that *AR* can be written as *Somers' D.* Starting with (* 3) it results

$$AR = \frac{1}{0{,}5\left(1-\overline{\lambda}\right)}\left[0{,}5\pi_{+K}\frac{\pi_{1K}}{\overline{\lambda}} + \pi_{+K-1}\frac{\pi_{1K}}{\overline{\lambda}} + 0{,}5\pi_{+K-1}\frac{\pi_{1K-1}}{\overline{\lambda}} + \right.$$

$$\left. + \pi_{+K-2}\frac{\sum_{j=K-1}^{K}\pi_{1j}}{\overline{\lambda}} + 0{,}5\pi_{+K-2}\frac{\pi_{1K-2}}{\overline{\lambda}} + \ldots + \pi_{+1}\frac{\sum_{j=2}^{K}\pi_{1j}}{\overline{\lambda}} + 0{,}5\pi_{+1}\frac{\pi_{11}}{\overline{\lambda}} - 0{,}5\right] =$$

$$= \frac{1}{\overline{\lambda}\left(1-\overline{\lambda}\right)}\left[\pi_{+1}\pi_{11} + 2\pi_{+1}\sum_{j=2}^{K}\pi_{1j} + \pi_{+2}\pi_{12} + 2\pi_{+2}\sum_{j=3}^{K}\pi_{1j} + \ldots + \right.$$

$$\left. + \pi_{+K-1}\pi_{1K-1} + 2\pi_{+K-1}\pi_{1K} + \pi_{+K}\pi_{1K} - \overline{\lambda}\right]$$

(* A1)

Note that $\pi_{+j} = \pi_{0j} + \pi_{1j}$. Replacing π_{+j} in (* A1) and rearranging the terms yields to

$$AR = \frac{1}{\overline{\lambda}\left(1-\overline{\lambda}\right)}\left[\sum\sum_{i<j}\pi_{0i}\pi_{1j} + \sum\sum_{i<j}\pi_{0i}\pi_{1j} + 2\sum\sum_{i<j}\pi_{1i}\pi_{1j} + \sum_{i}\pi_{0i}\pi_{1i} + \sum_{i}\pi_{1i}\pi_{1i} - \overline{\lambda}\right] =$$

$$= \frac{1}{\overline{\lambda}\left(1-\overline{\lambda}\right)}\left[\sum\sum_{i<j}\pi_{0i}\pi_{1j} + \sum\sum_{i\le j}\pi_{0i}\pi_{1j} + \sum\sum_{i\le j}\pi_{1i}\pi_{1j} + \sum\sum_{i<j}\pi_{1i}\pi_{1j} - \overline{\lambda}\right]$$

(* A2)

Now transform the last four expressions in (* A2) with some algebra:

$$\sum_{i \le j}\sum \pi_{0i}\pi_{1j} + \sum_{i \le j}\sum \pi_{1i}\pi_{1j} + \sum_{i < j}\sum \pi_{1i}\pi_{1j} - \bar{\lambda} =$$

$$= \sum_{i \le j}\sum \pi_{0i}\pi_{1j} + \sum_{i \le j}\sum (\pi_{+i} - \pi_{0i})\pi_{1j} + \sum_{i < j}\sum \pi_{1i}(\pi_{+j} - \pi_{0j}) - \sum_{i} \pi_{1i} =$$

$$= \sum_{i \le j}\sum \pi_{+i}\pi_{1j} + \sum_{i < j}\sum \pi_{1i}\pi_{+j} - \sum_{i < j}\sum \pi_{1i}\pi_{0j} - \sum_{i} \pi_{1i} =$$

$$= \pi_{+1}(\pi_{11} + \pi_{12} + \ldots + \pi_{1K}) + \pi_{+2}(\pi_{12} + \ldots + \pi_{1K}) + \ldots + \pi_{+K}\pi_{1K} +$$

$$+ \pi_{11}(\pi_{+2} + \ldots + \pi_{+K}) + \pi_{12}(\pi_{+3} + \ldots + \pi_{+K}) + \ldots + \pi_{1K-1}\pi_{+K} -$$

$$- [\pi_{11}(\pi_{02} + \pi_{03} + \ldots + \pi_{0K}) + \pi_{12}(\pi_{03} + \ldots + \pi_{0K}) + \ldots + \pi_{1K-1}\pi_{0K}] -$$

$$- \pi_{11} - \pi_{12} - \pi_{13} - \ldots - \pi_{1K} =$$

$$= \pi_{11}(\pi_{+1} + \pi_{+2} + \ldots + \pi_{+K}) + \pi_{12}(\pi_{+1} + \pi_{+2} + \ldots + \pi_{+K}) + \ldots +$$

$$+ \pi_{1K}(\pi_{+1} + \pi_{+2} + \ldots + \pi_{+K}) - \pi_{11} - \pi_{12} - \ldots - \pi_{1K} -$$

$$- [\pi_{0K}(\pi_{11} + \pi_{12} + \ldots + \pi_{1K-1}) + \pi_{0K-1}(\pi_{11} + \pi_{12} + \ldots + \pi_{1K-2}) + \ldots + \pi_{02}\pi_{01}] =$$

$$= -\sum_{i > j}\sum \pi_{0i}\pi_{1j}$$

(* A3)

Putting this result together with (* A2), we get

$$AR = \frac{1}{\bar{\lambda}(1 - \bar{\lambda})}\left[\sum_{i < j}\sum \pi_{0i}\pi_{1j} - \sum_{i > j}\sum \pi_{0i}\pi_{1j}\right] \qquad (* 6)$$

which is known as the formula for *Somers' D* (Agresti, 1984, p. 167).

We show that *AR* can be written in terms of *Gini*.

Start with:

$$AR = \frac{1}{\left(1-\overline{\lambda}\right)}\left(\sum\sum_{i<j}\frac{\pi_{0i}\cdot\pi_{1j}}{\overline{\lambda}} - \sum\sum_{i>j}\frac{\pi_{0i}\cdot\pi_{1j}}{\overline{\lambda}}\right) \qquad (* 6)$$

For convenience substituting $\frac{1}{\overline{\lambda}}\sum_{l=1}^{i}\pi_{1l} = y_i$ in (* 6) results in

$$\begin{aligned} AR &= \frac{1}{\left(1-\overline{\lambda}\right)}\left(\sum_{i=1}^{K-1}\pi_{0i}\left(1-y_i\right) - \sum_{i=2}^{K}\pi_{0i}y_{i-1}\right) = \\ &= \frac{1}{\left(1-\overline{\lambda}\right)}\left(\sum_{i=1}^{K-1}\pi_{0i} - \sum_{i=1}^{K-1}\pi_{0i}y_i - \sum_{i=2}^{K}\pi_{0i}y_{i-1}\right) \end{aligned} \qquad (* A4)$$

Furthermore rewrite $\sum_{l=1}^{i}\pi_{+l} = x_i$ and note

$$\pi_{0i} = \left(x_i - x_{i-1}\right) - \overline{\lambda}\left(y_i - y_{i-1}\right) \qquad (* A5)$$

Therefore

$$\begin{aligned} \pi_{0i} + \pi_{0i+1} &= \left(x_i - x_{i-1}\right) - \overline{\lambda}\left(y_i - y_{i-1}\right) + \left(x_{i+1} - x_i\right) - \overline{\lambda}\left(y_{i+1} - y_i\right) = \\ &= \left(x_{i+1} - x_{i-1}\right) - \overline{\lambda}\left(y_{i+1} - y_{i-1}\right) \end{aligned} \qquad (* A6)$$

and the first term in (* A4) can be simplified to

$$\sum_{i=1}^{K-1}\pi_{0i} = \left(x_{K-1} - x_0\right) - \bar{\lambda}\left(y_{K-1} - y_0\right) \qquad \text{(* A7)}$$

Now turn to the second and third term. Again using (* A5) leads to

$$\begin{aligned} &- \sum_{i=1}^{K-1}\pi_{0i}y_i - \sum_{i=2}^{K}\pi_{0i}y_{i-1} = \\ &= -\sum_{i=1}^{K-1}\left[\left(x_i - x_{i-1}\right) - \bar{\lambda}\left(y_i - y_{i-1}\right)\right]y_i - \sum_{i=2}^{K}\left[\left(x_i - x_{i-1}\right) - \bar{\lambda}\left(y_i - y_{i-1}\right)\right]y_{i-1} \end{aligned} \qquad \text{(* A8)}$$

Rewriting (* A8) extensively reveals that many elements compensate each other. The remaining is given by

$$\left[\left(x_0y_1 + x_1y_2 + \ldots + x_{K-2}y_{K-1}\right) - \left(x_2y_1 + x_3y_2 + \ldots + x_Ky_{K-1}\right)\right] - \bar{\lambda}y_1y_0 + \bar{\lambda}y_Ky_{K-1} \qquad \text{(* A9)}$$

Putting (* A7) and (* A9) together we get

$$\begin{aligned} AR = \frac{1}{\left(1-\bar{\lambda}\right)}\Big[&x_{K-1} - x_0 - \bar{\lambda}\left(y_{K-1} - y_0\right) + \\ &+ x_0y_1 + x_1y_2 + \ldots + x_{K-2}y_{K-1} - \left(x_2y_1 + x_3y_2 + \ldots + x_Ky_{K-1}\right) - \\ &- \bar{\lambda}y_1y_0 + \bar{\lambda}y_Ky_{K-1}\Big] \end{aligned} \qquad \text{(* A10)}$$

Note that $x_0 = y_0 = 0$ and $x_K = y_K = 1$. Therefore we can rewrite (* A10) in a very short form as

$$AR = \frac{1}{\left(1-\bar{\lambda}\right)}\left[\sum_{i=1}^{K}\left(x_{i-1}y_i - x_i y_{i-1}\right)\right] = \frac{1}{\left(1-\bar{\lambda}\right)} \cdot Gini \qquad (* 9)$$

Tables

Table 1: Contingency Table with notation for joint and marginal probabilities

	R_1	…	R_K	
$D=0$	π_{01}	…	π_{0K}	$\pi_{0+}=1-\overline{\lambda}$
$D=1$	π_{11}	…	π_{1K}	$\pi_{1+}=\overline{\lambda}$
	π_{+1}	…	π_{+K}	1

Exhibits

Exhibit 1:

Illustration of a Cumulative Accuracy Profile (dashed line), "ideal" CAP (dotted line) and Accuracy Ratio

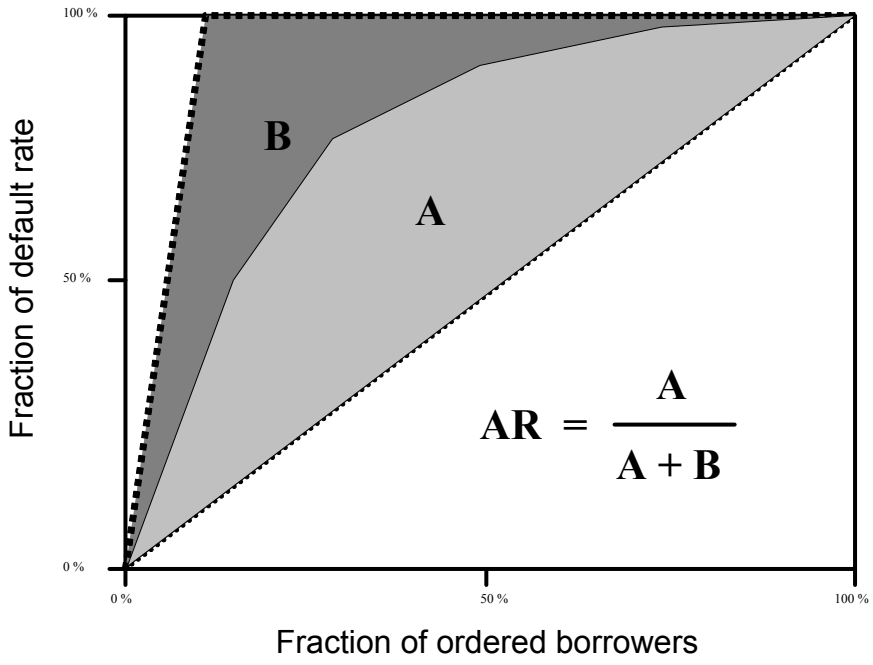

Exhibit 2:

Values of AR for varying default probabilities; portfolio consists of two groups of borrowers with 500 borrowers each