Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

On the three-way equivalence of AUC in credit scoring with tied scores

Guoping Zeng & Edward Zeng

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
2019, VOL. 48, NO. 7, 1635–1650
https://doi.org/10.1080/03610926.2018.1435814

Taylor & Francis
Taylor & Francis Group

Check for updates

On the three-way equivalence of AUC in credit scoring with tied scores

Guoping Zeng and Edward Zeng

Liberty High School, Oak Shores Dr, Plano, TX, USA

ABSTRACT

In credit scoring, it is well known that AUC (the area under curve) can be calculated geometrically, by the probability of a correct ranking of a good and bad pair, and by the Wilcoxon Rank-Sum statistic. This three-way equivalence was first present by Hanley and McNeil in 1982 without considering tied scores and without giving analytical proofs. In this paper, we extend the three-way equivalence to the case with tied scores and provide analytic proofs for the three-way equivalence.

ARTICLE HISTORY

Received 25 August 2017
Accepted 26 January 2018

KEYWORDS

Concordant pair; Discordant pair; Receiver operating characteristic (ROC); Tied pair; The area under curve (AUC); Wilcoxon rank-sum statistic.

1. Introduction

Recently, credit scoring models have been extensively used to evaluate the defaulting probability of credit applicants in order to determine whether or not to grant them credit (Siddiqi 2006; Refaat 2011; Zeng 2013; Zeng 2014a; Zeng 2014b; Zeng 2015; Zeng 2016a; Zeng 2016b; Zeng 2017). Kolmogorov-Smirnov statistic (KS), the receiver operating characteristic (ROC) curve, the area under the ROC curve (AUC) and Gini coefficient are often used to measure the performance or quality of a credit scoring model. The 4 performance measures, KS, ROC curve, AUC and Gini coefficient, are all based on cumulative distribution functions.

KS is the maximum difference between the cumulative proportion of bads and the cumulative proportion of goods among all scores. It measures the predictive ability of the model to separate goods from bads.

ROC curve shows the predictive power of a credit scoring model, i.e., the ability to identify goods and bads. To construct the ROC curve, for each score, the cumulative proportion of goods is plotted on the x-axis and the cumulative proportion of bads is plotted on the y-axis, though some users do swap the axes around. A perfect scoring model would follow the left vertical and top horizontal axes, accepting 100% of goods before accepting any of the bads. A random scoring model which accepts or rejects applicants randomly would follow the diagonal line $y = x$. In general, the higher the ROC curve above the diagonal line, the better the model is.

AUC, also known as the c-statistic, is a quantitative measure of ROC. It measures the predictive accuracy of a scoring model. The larger AUC, the more accurate the model is. A perfect scoring model has an AUC value of 1. A random model has an AUC value of 0.5. A worthless scoring model has an AUC value of 0.5 or less. In general, a useful model has an AUC value between 0.5 and 1.

CONTACT Guoping Zeng guopingtx@yahoo.com Liberty High School, 4522 Oak Shores Dr., Plano, TX 75024, USA.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lsta.

Gini coefficient is another quantitative measure of ROC. It is defined as the ratio of the area between the ROC curve and the diagonal to the area under the diagonal. So, Gini coefficient is 2 times the area between the ROC curve and the diagonal. Mathematically, it is 1 less than 2 times AUC. It measures the predictive power of the model. The higher Gini coefficient, the more predictive the model is. A perfect scoring model will have a Gini of 1, a random model will have an AUC value of 0.

KS is a stand-alone measure, whereas ROC curve, AUC and Gini coefficient are closely related.

As pointed out by Finlay (2012), KS measures the performance of a model through a single score, whereas AUC and Gini coefficient each measures the performance across the range of the model's scores. In this sense, AUC and Gini coefficient are better than KS.

AUC can be calculated in 3 ways. The most natural way is to find the area geometrically. Note that the ROC curve consists of piecewise linear horizontal and vertical sections if there are no ties of scores. In this case, AUC is the sum of areas of rectangles (Thomas 2009). In case of score ties where there are goods and bads with the same score, the ROC curve will have the stepwise form and AUC is the sum of areas of rectangles and trapezoids. In general, AUC can be found by the trapezoidal rule, that is, by forming trapezoids using the observed points as corners, computing the areas of these trapezoids and then adding them up (Gönen 2006; Denis et al. 2010; Refaat 2011; Pandey et al. 2016). This may take quite an effort when there are many different scores in the model.

AUC can be found by the probability of a correct ranking of a (good, bad) pair, that is, the probability the good has a higher score than the bad in a randomly selected pair of (good, bad). AUC can be also found by the Wilcoxon Rank-Sum statistic. The three-way equivalence for AUC was first presented in (Hanley and McNeil 1982) and has since been cited by many others. It was then presented in Gönen (2006), Pandey et al. (2016) and Rasouliyan and Miller (2012) as outputs of proc logistic in SAS.

However, ties of scores are not considered in Hanley and McNeil (1982). Moreover, there are no analytical proofs for the three-way equivalence for AUC in Hanley and McNeil (1982). All the subsequent works (Bradley 1997; Hand and Robert 2001; Mason and Graham 2002; Cortes and Mohri 2003; Fawcett 2006; Gönen 2006; Clemencon, Vayatis, and Depecker 2009; Denis et al. 2010; Pandey et al. 2016; Rasouliyan and Miller 2012; Kraus 2014) simply cited and inherited the three-way equivalence without proving it analytically.

In this paper, we fill the two important gaps in the literature relating to the three-way equivalence of AUC. First, we extend the three-way equivalence to tied scores. Secondly, we provide analytical proofs for the three-way equivalence. The rest of the paper is organized as follows. In Section 2, we review some basic concepts, namely, ROC curve and AUC, probability of correct ranking of a (good, bad) pair, and Wilcoxon Rank-sum statistic. In Section 3, we analytically prove the three-way equivalence of AUC. In Section 4, a numerical example is provided to demonstrate the three-way equivalence and to illustrate the proof in Section 3. SAS code and R code to calculate AUC are provided in the Appendix. Finally, the paper is concluded in Section 5.

2. Preliminaries

2.1. ROC curve and AUC

From now on, we assume that a sample of n records have been scored by a credit scoring model. We also assume that the binary dependent variable y has n_B bads ($y = 1$) and n_G goods

($y = 0$) so that $n_B + n_G = n$. We assume each sample has a score and a status: good (with $y = 0$) or bad (with $y = 1$).

The ROC curve is constructed by plotting the cumulative distribution of the conditional scores of the goods and the bads, $F(s|G)$ and $F(s|B)$, against one another: $F(s|G) = \Pr\{score \le s|G)$ on the x-axis versus $F(s|B) = \Pr\{score \le s|B)$ on the y-axis, though some users do swap the axes around as pointed by Thomas (2009).

For the numerical purpose, let us assume there are a total of K different scores among the n records. Sort K scores in ascending order and denote them by $S_1, S_2, \ldots, S_K$ where $S_1 < S_2 < \cdots < S_K$. Let $n_g(s)$ and $n_b(s)$ be the number of goods and bads with score s. Let $p_G(s) = \frac{n_g(s)}{n_G}$ and $p_B(s) = \frac{n_b(s)}{n_B}$, then

$$F\,(s|G) = \sum_{S_i \le s} p_G(S_i), \tag{2.1}$$

$$F\,(s|B) = \sum_{S_i \le s} p_B\,(S_i). \tag{2.2}$$

In particular, $F(S_K|G) = 1$ and $F(S_K|B) = 1$.

Let $x_i = F(S_i|G) = \frac{\sum_{k=1}^{i} n_g(S_k)}{n_G}$ and $y_i = F(S_i|B) = \frac{\sum_{k=1}^{i} n_b(S_k)}{n_B}$ for $i = 1, 2, \ldots, K$, then $x_1 \le x_2 \le \cdots \le x_K$ and $y_1 \le y_2 \le \cdots \le y_{K.}$ ROC is thus a broken-line curve to connect $(K + 1)$ points:

$$(x_0, y_0), (x_1, y_1), (x_2, y_2), \ldots, (x_K, y_K)$$

where $(x_0, y_0) = (0, 0)$ and $(x_K, y_K) = (1, 1)$.

The area under the ROC curve is the sum of areas of up to K triangles, rectangles or trapezoids: $(x_i, 0), (x_i, y_i), (x_{i+1}, y_{i+1}), (x_{i+1}, 0)$, $\;i = 0, 1, \ldots, K-1$. Note that if $x_i = x_{i+1}$ for some i, that is, score S_{i+1} has no goods, then $(x_i, 0), (x_i, y_i), (x_{i+1}, y_{i+1}), (x_{i+1}, 0)$ are collinear. In this case, we continue to add $(x_{i+2}, y_{i+2}), (x_{i+3}, y_{i+3}), \ldots$, until a nonzero area is obtained.

Example 2.1. Consider a sample of 13 observations, consisting of five bads with scores 150, 190, 200, 250, 260, and eight goods with scores 150, 180, 200, 205, 230, 260, 280, 300. The ROC curve will go through points O = (0, 0), A = (0.125, 0.2), B = (0.25, 0.2), C = (0.25, 0.4), D = (0.375, 0.6), E = (0.5, 0.6), F = (0.625, 0.6), G = (0.625, 0.8), H = (0.75, 1), I = (0.875, 1), J = (1, 1).

This gives the ROC curve in Fig. 1. The area under the ROC curve is the sum of one triangle, five rectangles and two trapezoids. Note that only the first one could be a triangle, that is, when the smallest score has at least one good and one bad.

2.2. Probability of correct ranking of a (good, bad) pair

Consider a sample of n records with K distinct scores. If, in a randomly selected pair of (good, bad), the good has a higher score than the bad, this pair is said to be a concordant pair. If, in a randomly selected pair of (good, bad), the good has a lower score than the bad, this pair is said to be a discordant pair. If, in a randomly selected pair of (good, bad), the good has the same score as the bad, this pair is said to be a tied pair.

To count the concordant pairs, let us first define Status-Score table.

Definition 2.1. The Status-Score table is a table with 2 rows such that

(1) Columns in the second row contain scores $S_1, S_2, \ldots, S_K$ one by one from left to right;

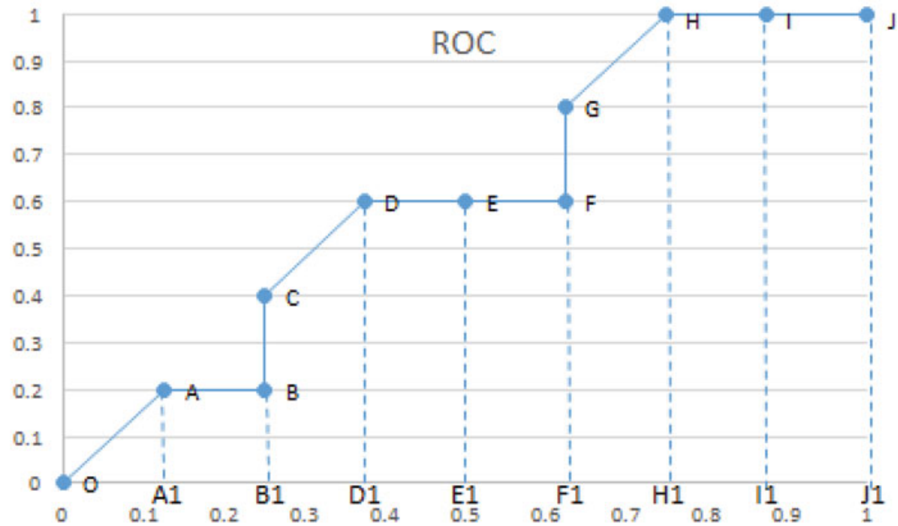

Figure 1. ROC curve.

(2) Columns in the first row contain all the statuses corresponding to the score in the second row;
(3) Neighboring columns can be combined into one column if their scores all have bad statuses and no good status.

The following 3 observations can be made immediately from the above definition.

Remark 2.2. The following holds regarding the Status-Score table:
(1) The total number of statuses on the first row is n.
(2) One score can only appear in one cell on the second row.
(3) If one cell on the second row has more than one score, the corresponding cell on the first row will have all bad statuses.

Example 2.3. Let us revisit Example 2.1. Its Status-Score table is at shown in Table 1. We see that there are 10 distinct scores $S_1 = 150$, $S_2 = 180$, $S_3 = 190$, $S_4 = 200$, $S_5 = 205$, $S_6 = 230$, $S_7 = 250$, $S_8 = 260$, $S_9 = 280$, $S_{10} = 300$, and 3 tied scores 150, 200 and 260, each with 2 statuses.

Let us define some basic notation regarding concordant pairs and tied pairs.

Notation 2.4. Denote
(1) the number of bads whose scores are lower than S_i by $C(S_i)$;
(2) the number of tied pairs with score S_i by $T(S_i)$;
(3) the number of concordant pairs by n_C;
(4) the number of tied pairs by n_T.

Table 1. Status-Score table for Example 2.1.

Status	0, 1	0	1	0, 1	0	0	1	0, 1	0	0
Score	150	180	190	200	205	230	250	260	280	300

Clearly, $C(S_1) = 0$. If $i > 1$, then

$$C\left(S_i\right) = \sum_{k=1}^{i-1} n_b(S_k). \tag{2.4}$$

Lemma 2.5. *The number of the concordant pairs is*

$$n_C = \sum_{\{i=1,\ldots,K|n_g(S_i)\neq 0\}} n_g\left(S_i\right) C(S_i). \tag{2.5}$$

Proof. To count the number of concordant pairs, we begin with good statuses. We first count the concordant pairs with respect to each good status and then sum up the concordant pairs with respect to all the good statuses. □

For a good status G_j in the first row of the Status-Score table, we may count all the concordant pairs with respect to G_j by counting bads in the cells according to their positions with G_j: the same cell; the cells on right; and the cells on left.

The bads, if only, in the same cell as G_j will not make a concordant pair with G_j, since they have the same score as G_j. The bads, if only, in the right cells will not make a concordant pair with G_j either since they have larger scores than G_j. All the bads, if only, in the left cells will make a concordant pair with G_j, since they have smaller scores than G_j.

Assume G_j is corresponding to score S_i in the Status-Score table. Since $C(S_i)$ is the number of bads on the left of S_i in the Status-Score table, the number of concordant pairs with respect to G_j is $C(S_i)$.

The product $n_g(S_i)C(S_i)$ is the total number of concordant pairs of G_j and its companion good statuses with the same score S_i in the same cell. Since the subscript $\{i = 1, \ldots, K|n_g(S_i) \neq 0\}$ represents all the K scores with at least one good status, (2.4) holds.

Lemma 2.6. *The number of tied pairs is*

$$n_T = \sum_{i=1}^{K} T\left(S_i\right) = \sum_{i=1}^{K} n_g\left(S_i\right) n_b\left(S_i\right). \tag{2.6}$$

Proof. Since $n_g(S_i)$ and $n_b(S_i)$ are the number of goods and bads for the same score S_i, respectively, product $n_g(S_i)n_b(S_i)$ is the number of tied pairs with score S_i, that is, □

$T(S_i) = n_g(S_i)n_b(S_i)$. Summing over all the K distinct scores, we obtain (2.6). Q.E.D.

As the total number of pairs is $n_G n_B$, the percent of concordant pairs is $\frac{n_C}{n_G n_B}$ and the percent of tied pairs is $\frac{n_T}{n_G n_B}$. To take the tied pairs into consideration for AUC, we tie the percent of concordant pairs $\frac{n_C}{n_G n_B}$ and the percent of tied pairs $\frac{n_{Ties}}{n_G n_B}$ together as follows.

Definition 2.7. Define the adjusted percent of concordant pairs to be

$$\frac{n_C}{n_G n_B} + 0.5\frac{n_{Ties}}{n_G n_B}.$$

As we will see later, the adjusted percent of concordant pairs equals AUC. The factor 0.5 is from the area of a triangle or a trapezoid.

2.3. Wilcoxon rank-sum statistic

The Wilcoxon Rank-Sum test (also called Mann-Whitney U test) is a nonparametric test based on the order in which the observations from the two samples fall (Wilcoxon 1945). To find the Wilcoxon Rank-Sum statistic or U statistic in credit scoring, we adopt the approach in Thomas (2009). We first order the sample in increasing order of score. The rank of each observation is its position in this ordered list, starting with rank 1 for the smallest observation. Next, we sum the ranks of the bads in the sequence. Let this be R_B. Similarly, we let R_G be the sum of the ranks of the goods in the sequence. Note that if there are $n = n_G + n_B$ in the sample, then

$$R_G + R_B = \frac{1}{2}n(n+1). \tag{2.7}$$

The Wilcoxon Rank-Sum statistic is defined as:

$$U = R_G - \frac{1}{2}n_G(n_G+1). \tag{2.8}$$

However, as in Hanley and McNeil (1982), tied samples are not considered in Thomas (2009). In case of tied samples, we assign each observation its average rank as in Narayanan and Watts (1996), and Wild and Seber (1999). Then, (2.6) still holds, and the Wilcoxon Rank-Sum statistics can be still defined as in (2.7) as can be seen.

We Note that the Wilcoxon Rank-Sum statistic was defined as a percent $\frac{U}{n_G n_B}$ in Hanley and McNeil (1982).

Definition 2.8. Define the adjusted Wilcoxon Rank-Sum statistic to be

$$\frac{U}{n_G n_B} = \frac{R_G - \frac{1}{2}n_G\left(n_G+1\right)}{n_G n_B}.$$

3. Analytic proofs of the three-way equivalence

3.1. Equivalence between AUC and the adjusted concordant

Theorem 3.1. *AUC equals adjusted percent of the concordant pairs, that is,*

$$AUC = \frac{n_C}{n_G n_B} + 0.5\frac{n_{Ties}}{n_G n_B}. \tag{3.1}$$

Proof. Our idea is to move along the curve from (0, 0) to find non-zero areas one by one. We will make use of the expressions of coordinates (x_i, y_i) for $i = 1, 2, \ldots, K$,

$$x_i = \frac{\sum_{k=1}^{i} n_g\left(S_k\right)}{n_G},$$
$$y_i = \frac{\sum_{k=1}^{i} n_b\left(S_k\right)}{n_B}.$$

□

Step 1. Beginning from $(0, 0)$, we calculate the first non-zero area of AUC. We consider 3 cases regarding the first score S_1: (1) S_1 has both bads and goods; (2) S_1 has only goods; and (3) S_1 has only bads. The 3 cases are equivalent to the 3 positions of (x_1, y_1) : (1) Inside the first quadrant; (2) on the x-axis; and (3) on the y-axis.

Table 2. Status-Score table for Step 1: (2)(i).

Status	G	G	$\dots$	G	$B\cdots BG\cdots G$
Score	S_1	S_2	$\dots$	S_{i-1}	S_i

(1) If S_1 has both bads and goods, then $x_1 = \frac{n_g(S_1)}{n_G} > 0$ and $y_1 = \frac{n_b(S_1)}{n_B} > 0$. Thus, the 3 points $(0,0)$, (x_1, y_1) and $(x_1, 0)$ form a triangle and has the first non-zero area to AUC:

$$\frac{1}{2}x_1y_1 = \frac{1}{2}\frac{n_g(S_1)\,n_b(S_1)}{n_Gn_B} = \frac{1}{2}\frac{T(S_1)}{n_Gn_B} \tag{3.2}$$

Since $C(S_1) = 0$, the area in (3.2) equals $\frac{n_g(S_1)C(S_1)}{n_Gn_B} + \frac{1}{2}\frac{T(S_1)}{n_Gn_B}$.

(2) If S_1 has only goods, then $x_1 > 0$, $y_1 = 0$ and point (x_1, y_1) is on the x-axis. Assume S_i is the first score among the $(K-1)$ scores $S_2, \dots, S_K$ that has bads, then $y_1 = y_2 = \cdots = y_{i-1} = 0, x_1 < x_2 < \cdots < x_{i-1}, y_i > 0$, points $(x_1, y_1), \dots, (x_{i-1}, y_{i-1})$ are all on the x-axis and point (x_i, y_i) leaves the x-axis. Moreover, $(0,0), (x_1, y_1), \dots, (x_{i-1}, y_{i-1})$ will not contribute to AUC. We consider 2 subcases: (i) S_i has at least one good; (ii) S_i has no goods.

(i) If score S_i has at least one good, then

$$x_{i-1} = \sum_{j=1}^{i-1}\frac{n_g(S_j)}{n_G} < x_i = \sum_{j=1}^{i}\frac{n_g(S_j)}{n_G} \tag{3.3}$$

and the 3 points $(x_{i-1}, y_{i-1}) = (x_{i-1}, 0)$, (x_i, y_i), $(x_i, 0)$ forms a triangle and has the first non-zero area to AUC

$$(x_i - x_{i-1})y_i = \frac{1}{2}\frac{n_g(S_i)}{n_G}\sum_{j=1}^{i}\frac{n_b(S_j)}{n_B} = \frac{1}{2}\frac{n_g(S_i)}{n_G}\frac{n_b(S_i)}{n_B} = \frac{1}{2}\frac{T(S_i)}{n_Gn_B}. \tag{3.4}$$

It follows from Table 2 that $C(S_i) = 0$. Thus, the area in (3.4) equals $\frac{n_g(S_i)C(S_i)}{n_Gn_B} + \frac{1}{2}\frac{T(S_i)}{n_Gn_B}$.

(ii) If score S_i has no goods, then $x_{i-1} = x_i$ and so the line L connecting points $(x_{i-1}, y_{i-1}) = (x_{i-1}, 0)$ and (x_i, y_i) is perpendicular to the x-axis. Assume S_j is the first score among scores $S_{i+1}, \dots, S_K$ that has goods. Then $x_i = \cdots = x_{j-1}$ and point (x_j, y_j) leaves line L. If score S_j has at least one bad, then $y_{j-1} < y_j$, so points (x_{i-1}, y_{i-1}), (x_i, y_i), (x_{j-1}, y_{j-1}), (x_j, y_j), $(x_j, 0)$ form a trapezoid and has the first non-zero area to AUC:

$$\begin{aligned}
&(x_j - x_{j-1})y_{j-1} + \frac{1}{2}(x_j - x_{j-1})(y_j - y_{j-1}) \\
&= \frac{n_g(S_j)}{n_G}\sum_{k=1}^{j-1}\frac{n_b(S_k)}{n_B} + \frac{1}{2}\frac{n_g(S_j)}{n_G}\frac{n_b(S_j)}{n_B} \\
&= \frac{n_g(S_j)}{n_G}\sum_{k=i}^{j-1}\frac{n_b(S_k)}{n_B} + \frac{1}{2}\frac{n_g(S_j)}{n_G}\frac{n_b(S_j)}{n_B}.
\end{aligned} \tag{3.5}$$

Table 3. Status-Score table for Step 1: (2)(ii).

Status	G	G	$\dots$	G	$BB\cdots B$	$B\cdots BG\cdots G$
Score	S_1	S_2	$\dots$	S_{i-1}	$S_i, S_{i+1}, \dots, S_{j-1}$	S_j

Table 3 is the Status-Score table for the first j scores. We see that

$$C(S_j) = \sum_{k=i}^{j-1} n_b\,(S_k) \tag{3.6}$$

Multiplying both sides of (3.6) by $n_g(S_j)/n_G n_B$, we obtain

$$\frac{n_g(S_j)C(S_j)}{n_G n_B} = \frac{n_g(S_j)}{n_G} \sum_{k=i}^{j-1} \frac{n_b\,(S_k)}{n_B} \tag{3.7}$$

Hence, the area in (3.5) equals $\frac{n_g(S_j)C(S_j)}{n_G n_B} + \frac{1}{2}\frac{T(S_j)}{n_G n_B}$.
If otherwise score S_j has no bads, then $y_{j-1} = y_j$ and points (x_{i-1}, y_{i-1}), (x_i, y_i), (x_{j-1}, y_{j-1}), (x_j, y_j) form a rectangle and has the first non-zero area to AUC

$$\left(x_j - x_{j-1}\right) y_j = \frac{n_g(S_j)}{n_G} y_j = \frac{n_g(S_j)}{n_G} \sum_{k=1}^{j} \frac{n_b\,(S_k)}{n_B} = \frac{n_g(S_j)}{n_G} \sum_{k=i}^{j-1} \frac{n_b\,(S_k)}{n_B}. \tag{3.8}$$

Table 4 is the Status-Score table for the first j scores. We see that

$$C(S_j) = \sum_{k=i}^{j-1} n_b\,(S_k)\,. \tag{3.9}$$

Multiplying both sides of (3.9) by $\frac{n_g(S_j)}{n_G n_B}$, we obtain

$$\frac{n_g(S_j)C(S_j)}{n_G n_B} = \frac{n_g(S_j)}{n_G} \sum_{k=i}^{j-1} \frac{n_b\,(S_k)}{n_B}. \tag{3.10}$$

Since S_j has no bads, $T(S_j) = 0$. Therefore, the area in (3.8) equals

$$\frac{n_g(S_j)C(S_j)}{n_G n_B} + \frac{1}{2}\frac{T(S_j)}{n_G n_B}.$$

(3) If score S_1 has only bads, then $x_1 = 0$ and $y_1 > 0$ and point (x_1, y_1) is on the y-axis. Assume S_i is the first score among the K scores that has goods, then $x_1 = x_2 = \cdots = x_{i-1} = 0$, $x_i > 0$, points $(x_1, y_1), \dots, (x_{i-1}, y_{i-1})$ are all on the y-axis and point (x_i, y_i) leaves the y-axis. Moreover, $(0, 0)$, $(x_1, y_1), \dots, (x_{i-1}, y_{i-1})$ will not contribute to AUC. We consider 2 subcases: (i) S_i has at least one bad; (ii) S_i has no bads.
  (i) If score S_i has at least one bad, then $y_{i-1} = \sum_{j=1}^{i-1} \frac{n_b(S_j)}{n_B} < y_i = \sum_{j=1}^{j} \frac{n_b(S_j)}{n_B}$. Hence, points $(0, 0)$, $(x_{i-1}, y_{i-1}) = (0, y_{i-1})$, (x_i, y_i), $(x_i, 0)$ forms a trapezoid and has the

Table 4 . Status-Score table for Step 1: (2)(ii).

Status	G	G	$\dots$	G	$BB\cdots B$	$GG\cdots G$
Score	S_1	S_2	$\dots$	S_{i-1}	$S_i, S_{i+1}, \dots, S_{j-1}$	S_j

Table 5. Status-Score table for Step 1: (3)(i).

Status	$BB\cdots B$	$B\cdots BG\cdots G$
Score	$S_1, S_2, \ldots, S_{i-1}$	S_i

first non-zero area to AUC:

$$x_i y_{i-1} + \frac{1}{2} x_i \left(y_i - y_{i-1}\right) = \left(\sum_{j=1}^{i} \frac{n_g(S_j)}{n_G}\right)\left(\sum_{j=1}^{i-1} \frac{n_b(S_j)}{n_B}\right) + \frac{1}{2}\left(\sum_{j=1}^{i} \frac{n_g(S_j)}{n_G}\right)\frac{n_b\left(S_i\right)}{n_B}$$
$$= \frac{n_g\left(S_i\right)}{n_G}\sum_{j=1}^{i-1}\frac{n_b(S_j)}{n_B} + \frac{1}{2}\frac{n_g\left(S_i\right)}{n_G}\frac{n_b\left(S_i\right)}{n_B}. \tag{3.11}$$

Table 5 is the Status-Score table for the first i scores. We obtain as in (3.10) that

$$\frac{n_g\left(S_i\right)C\left(S_i\right)}{n_G n_B} = \frac{n_g\left(S_i\right)}{n_G}\sum_{j=1}^{i-1}\frac{n_b(S_j)}{n_B}. \tag{3.12}$$

Then, the area in (3.11) equals to $\frac{n_g(S_i)C(S_i)}{n_G n_B} + \frac{1}{2}\frac{T(S_i)}{n_G n_B}$.

(ii) If score S_i has no bads, then $y_{i-1} = y_i$, then points $(0, 0)$, $(x_{i-1}, y_{i-1}) = (0, y_{i-1})$, (x_i, y_i), $(x_i, 0)$ form a rectangular and has first non-zero area:

$$x_i y_{i-1} = \left(\sum_{j=1}^{i} \frac{n_g(S_j)}{n_G}\right)\left(\sum_{j=1}^{i-1} \frac{n_b(S_j)}{n_B}\right) = \frac{n_g\left(S_i\right)}{n_G}\sum_{j=1}^{i-1}\frac{n_b(S_j)}{n_B}. \tag{3.13}$$

Table 6 is the Status-Score table for the first i scores. We obtain as in (3.10) that

$$\frac{n_g\left(S_i\right)C\left(S_i\right)}{n_G n_B} = \frac{n_g\left(S_i\right)}{n_G}\sum_{j=1}^{i-1}\frac{n_b(S_j)}{n_B} \tag{3.14}$$

Since S_i has no bads, $T(S_i) = 0$. Therefore, the area in (3.13) equals

$$\frac{n_g\left(S_i\right)C\left(S_i\right)}{n_G n_B} + \frac{1}{2}\frac{T\left(S_i\right)}{n_G n_B}.$$

Step 2. We start with the last position (x_{i_1}, y_{i_1}) at Step 1, where $i_1 = 1$ for case (1), $i_1 = j$ for case (2)(ii) and $i_1 = i$ otherwise. Let S_l be the first score among $S_{i_1+1}, \ldots, S_K$ that has goods. Then $x_{i_1} = \cdots = x_{l-1}$, $y_{i_1} < y_{i_1+1} < \cdots < y_{l-1}$. If S_l has at least one bad, then $(x_{i_1}, 0)$, (x_{i_1}, y_{i_1}), (x_{i_1+1}, y_{i_1+1}), $\ldots$, (x_{l-1}, y_{l-1}), (x_l, y_l), $(x_l, 0)$ form a trapezoid with area

$$\left(x_l - x_{l-1}\right)y_{l-1} + \frac{1}{2}\left(x_l - x_{l-1}\right)\left(y_l - y_{l-1}\right) = \frac{n_g\left(S_l\right)}{n_G}\sum_{k=1}^{l-1}\frac{n_b\left(S_k\right)}{n_B} + \frac{1}{2}\frac{n_g\left(S_l\right)}{n_G}\frac{n_b\left(S_l\right)}{n_B}. \tag{3.15}$$

Table 6. Status-Score table for Step 1: (3)(ii).

Status	$BB\cdots B$	$G\cdots G$
Score	$S_1, S_2, \ldots, S_{i-1}$	S_i

Table 7. Status-Score table for Step 2.

Status	$\cdots$	$BB\cdots$	$B\cdots BG\cdots G$
Score	$\cdots$	$S_{i_1}, S_{i_1+1}, \ldots, S_{l-1}$	S_l

Table 8. Status-Score table for Step 2.

Status	$\cdots$	$BB\cdots B$	$G\cdots G$
Score	$\cdots$	$S_{i_1}, S_{i_1+1}, \ldots, S_{l-1}$	S_l

Table 7 is the Status-Score table for the first l scores. We obtain as in (3.10) that

$$\frac{n_g(S_l)\,C(S_l)}{n_G n_B} = \frac{n_g(S_l)}{n_G}\sum_{k=1}^{l-1}\frac{n_b(S_k)}{n_B} \tag{3.16}$$

Therefore, the area in (3.15) equals $\frac{n_g(S_l)C(S_l)}{n_G n_B} + \frac{1}{2}\frac{T(S_l)}{n_G n_B}$.

If S_l has no bads, then $y_{l-1} = y_l$ and $(x_{i_1}, 0)$, (x_{i_1}, y_{i_1}), $(x_{i_1+1}, y_{i_1+1}), \ldots, (x_{l-1}, y_{l-1})$, (x_l, y_l), $(x_l, 0)$ forms a rectangle with area:

$$(x_l - x_{l-1})\,y_l = \frac{n_g(S_l)}{n_G}\sum_{k=1}^{l}\frac{n_b(S_k)}{n_B} = \frac{n_g(S_l)}{n_G}\sum_{k=1}^{l-1}\frac{n_b(S_k)}{n_B}. \tag{3.17}$$

Table 8 is the Status-Score table for the first l scores. We obtain as in (3.10) that

$$\frac{n_g(S_l)\,C(S_l)}{n_G n_B} = \frac{n_g(S_l)}{n_G}\sum_{k=1}^{l-1}\frac{n_b(S_k)}{n_B}. \tag{3.18}$$

Since S_l has no bads, $T(S_l) = 0$. Therefore, the area in (3.17) equals $\frac{n_g(S_l)C(S_l)}{n_G n_B} + \frac{1}{2}\frac{T(S_l)}{n_G n_B}$.

Step 3. Continue Step 2 to find the next non-zero area of AUC until we reach (x_K, y_K). We see that the area in each part equals the percent of concord so far plus one half of the percent of ties.

Step 4. In conclusion,

$$\mathrm{AUC} = \frac{n_C}{n_G n_B} + \frac{1}{2}\times\frac{n_T}{n_G n_B}. \tag{3.19}$$

$\square$

If we switch goods and bads and let n_D denote the number of discordant pairs, we may find AUC based on bads.

$$\mathrm{AUC} = \frac{n_D}{n_G n_B} + \frac{1}{2} \times \frac{n_T}{n_G n_B}. \tag{3.20}$$

Since $n_C + n_D + n_{Ties} = n_G n_B$, we immediately have the following results.

Corollary 3.2. *The sum of the AUC based on goods and the AUC based on bads is 1.*

3.2. Equivalence between the adjusted percent of concordant pairs and the adjusted wilcoxon rank-sum statistic

Theorem 3.3. *The adjusted percent of the concordant pairs equals the adjusted Wilcoxon Rank-Sum Statistic, that is*

$$\frac{n_C}{n_G n_B} + \frac{1}{2} \times \frac{n_T}{n_G n_B} = \frac{R_G - \frac{1}{2} n_G (n_G + 1)}{n_G n_B}. \tag{3.21}$$

Proof. Sort all n records by their scores in ascending order and label them 1, 2, 3, ⋯, by their positions. Consider the ith good in this increasing score sequence. Let's denote the rank of that good by $R(i)$. If the ith good has no tied score, then the number of bads with lower scores is $(R(i) - i)$. If otherwise the ith good has a tied score S, let's assume this score S has j goods and k bads and there are a total of m accounts including l goods before the tied scores:

$$G_{l+1}, G_{l+2} \ldots, G_{l+j}, B, B, \ldots, B$$

with positions $m+1, m+2, \ldots, m+j, m+j+1, m+j+2, \ldots, m+j+k$, respectively. Thenthere are $(m-l)$ bads with lower scores than the score S of $G_{l+1}, G_{l+2} \ldots, G_{l+j}$. Hence, $C(S) = j(m-l)$, and so the cordant percent regarding score S is

$$\Pr\{S_G > S_B\} = \frac{(m-l)\, j}{n_G n_B}. \tag{3.22}$$

□

On the other hand, the rank of the j goods is the average position of the $(j+k)$ tied scores:

$$\frac{(m+1) + (m+2) + \cdots (m+j+k)}{j+k}.$$

Thus, we have

$$\begin{aligned}
\sum_{i=l+1}^{l+j} (R(G_i) - i) &= j\frac{(m+1) + (m+2) + \cdots (m+j+k)}{j+k} - (l+1+l+2+\cdots+l+j) \\
&= j\frac{(j+k)\, m + \frac{(j+k)(j+k+1)}{2}}{j+k} - \left(jl + \frac{j(j+1)}{2}\right) \\
&= mj + \frac{j(j+k+1)}{2} - jl - \frac{j(j+1)}{2} \\
&= (m-l)\, j + \frac{jk}{2} = C(S) + 0.5 n_g(S)\, n_b(S).
\end{aligned} \tag{3.23}$$

Summing $(R(G_i) - i)$ over all the goods, we have

$$\sum_i (R(G_i) - i) = n_C + 0.5 n_{Ties}. \tag{3.24}$$

Dividing $n_G n_B$, we obtain

$$\frac{\sum_i (R(G_i) - i)}{n_G n_B} = \frac{n_C + 0.5 n_{Ties}}{n_G n_B}. \tag{3.25}$$

Since

$$\frac{\sum_i (R(G_i) - i)}{n_G n_B} = \frac{\sum_i R(G_i) - \frac{n_G(n_G+1)}{2}}{n_G n_B}, \tag{3.26}$$

We obtain

$$\frac{\sum_i R(G_i) - \frac{n_G(n_G+1)}{2}}{n_G n_B} = \frac{n_C + 0.5 n_{Ties}}{n_G n_B} = \frac{n_C}{n_G n_B} + \frac{1}{2} \times \frac{n_T}{n_G n_B}. \tag{3.27}$$

Since $\sum_i R(G_i) = R_G$, we can rewrite (3.27) as

$$\frac{R_G - \frac{n_G(n_G+1)}{2}}{n_G n_B} = \frac{n_C}{n_G n_B} + \frac{1}{2} \times \frac{n_T}{n_G n_B}, \tag{3.28}$$

which is identical to (3.21).

4. A Numerical example

Let's revisit example 2.1 one more time. We first calculate AUC in 3 ways and then illustrates the proof in Section 3.1.

4.1. Calculation of AUC in 3 ways

(A). AUC by Geometry

To calculate AUC directly and geometrically, we notice that the area under curve consists of one triangle, 3 rectangles and 2 trapezoids

$$\begin{aligned} \text{AUC} = {} & \frac{0.125 \times 0.2}{2} + (0.25 - 0.125) \times 0.2 + \frac{(0.4 + 0.6) \times (0.375 - 0.25)}{2} \\ & + (0.625 - 0.375) \times 0.6 + \frac{(0.8 + 1) \times (0.75 - 0.625)}{2} \\ & + (1 - 0.75) \times 1 = 0.6125. \end{aligned}$$

(B). AUC by Concordant Pairs and Tied Pairs

We calculate AUC by Concordant Pairs and Tied Pairs. From the Status-Score table, we see that $n_C = 1 + 2 + 2 \times 3 + 4 + 2 \times 5 = 23$, $n_g(150) = 1$, $n_b(150) = 1$; $n_g(200) = 1$, $n_b(200) = 1$; $n_g(260) = 1$, $n_b(260) = 1$, and $n_T = 1 \times 1 + 1 \times 1 + 1 \times 1 = 3$.

Thus, AUC, in view of concordant pairs and tied pair, is

$$\text{AUC} = \frac{n_C}{n_G n_B} + \frac{1}{2} \times \frac{n_T}{n_G n_B} = \frac{23}{5 \times 8} + \frac{1}{2} \times \frac{3}{5 \times 8} = 0.6125.$$

(C). AUC by Wilcoxon Rank-sum Statistic

The sum of the ranks of the goods in the sequence is

$$R_G = 1.5 + 2 + 5.5 + 7 + 8 + 10.5 + 12 + 13 = 60.5.$$

Therefore, AUC in view of Wilcoxon Rank-sum Statistic is

$$\text{AUC} = \frac{R_G - \frac{n_G(n_G+1)}{2}}{n_G n_B} = \frac{60.5 - 0.5 \times 8 \times 9}{5 \times 8} = 0.6125.$$

4.2. Illustration of the proof in Section 3.1

Let us use this example to illustrate the proof in Section 3. First of all, $n_G = 8$, $n_B = 5$. Next, using the Status-Score table (Table 1), we find the following values, which will be used in the subsequent calculations:

- $S_1 = 150 : C(S_1) = 0,\ T(S_1) = n_g(S_1)n_b(S_1) = 1.$
- $S_2 = 180 : n_g(S_2) = 1, C(S_2) = 1,\ T(S_2) = n_g(S_2)n_b(S_2) = 0.$
- $S_4 = 200 : n_g(S_4) = 1, C(S_4) = 2,\ T(S_4) = n_g(S_4)n_b(S_4) = 1.$
- $S_5 = 205 : n_g(S_5) = 1, C(S_5) = 3,\ T(S_5) = n_g(S_5)n_b(S_5) = 0.$
- $S_6 = 230 : n_g(S_6) = 1, C(S_6) = 3,\ T(S_6) = n_g(S_6)n_b(S_6) = 0.$
- $S_8 = 260 : n_g(S_8) = 1, C(S_8) = 4,\ T(S_8) = n_g(S_8)n_b(S_8) = 1.$
- $S_9 = 280 : n_g(S_9) = 1, C(S_9) = 5,\ T(S_9) = n_g(S_9)n_b(S_9) = 0.$
- $S_{10} = 300 : n_{10}(S_9) = 1, C(S_{10}) = 5,\ T(S_{10}) = n_g(S_{10})n_b(S_{10}) = 0.$

Step 1: Since the first score $S_1 = 150$ has a good and a bad, the first non-zero area to AUC is from a triangle (formed by points O, A and A1 in Fig. 1.) by Step 1(1). The area of this triangle can be found geometrically to be $\frac{1}{2} \times 0.2 \times 0.125 = 0.0125$. Since $\frac{n_g(S_1)C(S_1)}{n_G n_B} + \frac{1}{2}\frac{T(S_1)}{n_G n_B} = \frac{1}{2}\frac{T(S_1)}{n_G n_B} = \frac{1}{2} \times \frac{1}{5\times 8} = 0.0125$ too, (3.2) in the proof of Theorem 3.1 is verified.

Step 2. Next, let us move to Step 2 beginning from point A = (0.125, 0.2). The first score with good is $S_2 = 180$. Since S_2 has no bad, the next non-zero area to AUC is from a rectangle (formed by points A1, A, B, B1 in Figure 1). The area of this rectangle can be found geometrically to be $(0.25 - 0.125) \times 0.2 = 0.025$. Since $\frac{n_g(S_2)C(S_2)}{n_G n_B} + \frac{1}{2}\frac{T(S_2)}{n_G n_B} = \frac{1\times 1}{5\times 8} = 0.025$ too, (3.17) in the proof of Theorem 3.1 is verified.

Step 3. (Continuing Step 2). Repeating Step 2 starting from point B = (0.25, 0.2). The first score with good is $S_4 = 200$. Since S_4 has a bad, the next non-zero area is from a trapezoid (formed by points B1, C, D, D1 in Fig. 1.). The area of this trapezoid can be found geometrically to be $(0.4 + 0.6) \times (0.375 - 0.25)/2 = 0.0625$. Since $\frac{n_g(S_4)C(S_4)}{n_G n_B} + \frac{1}{2}\frac{T(S_4)}{n_G n_B} = \frac{1\times 2}{5\times 8} + \frac{1}{2}\frac{1}{5\times 8} = 0.06256$ too, (3.15) in the proof of Theorem 3.1 is verified.

Step 3. (Continuing Step 2). Repeating Step 2 starting from point D = (0.375, 0.6). The first score with good is $S_5 = 205$. Since S_5 has no bad, the next non-zero area is from a rectangle (formed by points D1, D, E, E1in Fig. 1.). The area of this rectangle can be found geometrically to be $(0.5 - 0.375) \times 0.6 = 0.075$. Since $\frac{n_g(S_5)C(S_5)}{n_G n_B} + \frac{1}{2}\frac{T(S_5)}{n_G n_B} = \frac{1\times 3}{5\times 8} = 0.075$ too, (3.17) in the proof of Theorem 3.1 is verified.

Step 3. (Continuing Step 2). Repeating Step 2 starting from point E = (0.5, 0.6). The first score with good is $S_6 = 230$. Since S_6 has no bad, the next non-zero area is from a rectangle (formed by points E1, E, F, F1in Fig. 1.). The area of this rectangle can be found geometrically to be $(0.5 - 0.375) \times 0.6 = 0.075$. Since $\frac{n_g(S_6)C(S_6)}{n_G n_B} + \frac{1}{2}\frac{T(S_6)}{n_G n_B} = \frac{1\times 3}{5\times 8} = 0.075$ too, (3.17) in the proof of Theorem 3.1 is verified.

Step 3. (Continuing Step 2). Repeating Step 2 starting from point F = (0.625, 0.6). The first score with good is $S_8 = 260$. Since S_8 has a bad, the next non-zero area is from a trapezoid (formed by points F1, G, H, H1in Fig. 1.). The area of this trapezoid can be found geometrically to be $(0.4 + 0.8) \times (0.375 - 0.25)/2 = 0.1125$. Since $\frac{n_g(S_8)C(S_8)}{n_G n_B} + \frac{1}{2}\frac{T(S_8)}{n_G n_B} = \frac{1\times 4}{5\times 8} + \frac{1}{2}\frac{1}{5\times 8} = 0.1125$ too, (3.15) in the proof of Theorem 3.1 is verified.

Step 3. (Continuing Step 2). Repeating Step 2 starting from point H = (0.75, 1). The first score with good is $S_9 = 280$. Since S_9 has no bad, the next non-zero area is from a rectangle (formed by points H1, H, I, I1 in Fig. 1.). The area of this trapezoid can be found geometrically to be $(0.875 - 0.75) \times 1 = 0.125$. Since $\frac{n_g(S_9)C(S_9)}{n_G n_B} + \frac{1}{2}\frac{T(S_9)}{n_G n_B} = \frac{1\times 5}{5\times 8} = 0.125$ too, (3.17) in the proof of Theorem 3.1 is verified.

Step 3. (Continuing Step 2). Repeating Step 2 starting from point I = (0.875, 1). The last score is $S_{10} = 300$ with a good status. Since S_{10} has no bad, the next non-zero area is from a rectangle (formed by points I1, I, J, J1 in Fig. 1.). The area of this trapezoid can be found geometrically to be $(1 - 0.875) \times 1 = 0.125$. Since $\frac{n_g(S_{10})C(S_{10})}{n_G n_B} + \frac{1}{2}\frac{T(S_{10})}{n_G n_B} = \frac{1\times 5}{5\times 8} = 0.125$ too, (3.17) in the proof of Theorem 3.1 is verified.

Step 4. Adding up all the areas above, we obtain AUC = 0.0125 + 0.025 + 0.0625 + 0.075 + 0.075 + 0.1125 + 0.125 +0.125 = 0.6125.

6. Conclusions

In this paper, we have extended the three-way equivalence among AUC, the probability of a correct ranking of a good and bad pair, and the Wilcoxon Rank Sum Test statistic in credit scoring to the case of tied scores. We have analytically proved the three-way equivalence.

References

Bradley, A. P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30 (7):1145–59. doi:10.1016/S0031-3203(96)00142-2.

Clemencon, S., N. Vayatis, and M. Depecker. 2009. AUC optimization and the two-sample problem. *In Advances in Neural Information Processing Systems* 22:360–68.

Cortes, C., and M. Mohri. 2003. AUC optimization vs. error rate minimization. *Neural information processing systems 15 (NIPS)*. MIT Press.

Denis, J., et al. 2010. A program for computing the predication probability and the related receiver operating characteristic graph. *Anesthesia & Analgesia* 111:1416–21. doi:10.1213/ANE.0b013e3181fb919e.

Fawcett, T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27:861–74. doi:10.1016/j.patrec.2005.10.010.

Finlay, S. 2012. Credit Scoring, Response Modeling, and Insurance Rating: A Practical Guide to Forecasting Consumer Behavior 2nd ed.

Gönen, M. 2006. Receiver Operating Characteristic (ROC) Curves. Paper 210–31, in: *Proceedings of the SUGI 31*, San Francisco, USA.

Hand, D. J., and J. T. Robert. 2001. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning* 45:171–86. doi:10.1023/A:1010920819831.

Hanley, J. A., and B. J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143:29–36. doi:10.1148/radiology.143.1.7063747.

Kraus, A. 2014. Recent methods from statistics and machine learning for credit scoring. Ph.D. Dissertation. Ph.D. thesis, Ludwig-Maximilians-Universitat Munchen.

Mason, S. J., and N. E. Graham. 2002. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society* 128:2145–66. doi:10.1256/003590002320603584.

Narayanan, A., and D. Watts. 1996. Exact methods in the NPAR1WAY procedure. *Proceedings of the 21st Annual SUGI Conference*, Chicago, Illinois, USA.

Pandey, M., et al. 2016. *ROC Curve: Making way for correct diagnosis*. SP11 – PharmaSUG 2016.
Rasouliyan, L., and D. P. Miller. 2012. *The logic and logistics of logistic regression including new features in SAS® 9.2*. Budapest, Hungary: The Pharmaceutical Users Software Exchange (PhUSE).
Refaat, M. 2011. *Credit risk scorecards: Development and implementation using SAS*. Raleigh, North Carolina, USA: LULU.COM.
Siddiqi, N. 2006. *Credit risk scorecards: Developing and implementing intelligent credit scoring*. Hoboken, New Jersey: John Wiley & Sons, Inc.
Thomas, L. C. 2009. *Consumer credit models: Pricing, profit and portfolios*. Oxford University Press, 1st Edition.
Wilcoxon, F. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin* 1 (6):80–83. doi:10.2307/3001968.
Wild, C. J., and G. A. F. Seber. 1999. The wilcoxonon rank-sum test. *Chapter 10 in chance encounters: A first course in data analysis and inference*. New York: Wiley & Sons.
Zeng, G. 2013. Metric divergence measures and information value in credit scoring. *Journal of Mathematics* 2013, Article ID 848271. doi:10.1155/2013/848271.
Zeng, G. 2014a. A rule of thumb for reject inference in credit scoring. *Mathematical Finance Letters* vol. 2014, article 2, 1–13.
Zeng, G. 2014b. A necessary condition for a good binning algorithm in credit scoring. *Applied Mathematical Sciences* 8 (65):3229–42. doi:10.12988/ams.2014.44300.
Zeng, G. 2015. A unified definition of mutual information with applications in machine learning. *Mathematical Problems in Engineering* vol. 2015, Article ID 201874. doi:10.1155/2015/201874
Zeng, G. 2016a. A comparison study of computational methods of Kolmogorov–Smirnov statistic in credit scoring. *Communications in Statistics: Simulation and Computation*. doi:10.1080/03610918.2016.1249883
Zeng, G. 2016b. On the existence of maximum likelihood estimates for weighted logistic regression. *Communications in Statistics: Simulation and Computation*. doi:10.1080/03610926.2016.1260742
Zeng, G. 2017. Invariant properties of logistic regression model in credit scoring under monotonic transformations. *Communications in Statistics: Theory and Methods* 46 (17):8791–807. doi:10.1080/03610926.2016.1193200.

Appendix

A.1. SAS for AUC

SAS procedure npar1way can be used to find as follows:

```
proc npar1way Wilcoxonon data = auc_data;
  class target;
  var score;
run;
```

Here, auc_data is the dataset under calculation, target is the binary dependent variable (1 for bad and 0 for good), and score is the score variable. Let us use Example 2.1 again. The output for our Example is shown in Table 9. The AUC can be calculated as

$$\frac{60.5 - 56}{5 \times 8} + 0.5 = 0.6125.$$

Table 9. Wilcoxon output from SAS.

Wilcoxon Scores (Rank Sums) for Variable score Classified by Variable target			
target	N	Sum of Scores	Expected Under H0
1	5	30.5	35
0	8	60.5	56
Average scores were used for ties.			

In general, Sum of Scores for Class 1 (for bads), denoted by $SS_{1,}$ is R_B, Expected Under H0 for class 1, denoted by $ES_{1,}$ is $\frac{1}{2}n_B(n_B+1)+0.5n_G n_B$. Similarly, Sum of Scores for Class 0 (for goods), denoted by $SS_{0,}$ is R_G, Expected Under H0 for class 0, denoted by $ES_{0,}$ is $\frac{1}{2}n_G(n_G+1)+0.5n_G n_B$. Then,

$$\begin{aligned}(SS_1-ES_1)+(SS_0-ES_0) &= R_B+R_G-\frac{1}{2}n_B\left(n_B+1\right)-\frac{1}{2}n_G\left(n_G+1\right)-n_G n_B \\ &= R_B+R_G-\frac{\left(n_B+n_G\right)\left(n_B+n_G+1\right)}{2}=R_B+R_G-\frac{n\left(n+1\right)}{2} \\ &= 0.\end{aligned}$$

A.2. R for AUC

The AUC can be also easily found in R by function auc() from R package pROC as follows

auc(roc(auc_data$target, auc_data$score))

The output will be “*Area under the curve: 0.6125*”.