DISCUSSION PAPER

2016/46

Bounds on Concordance-Based Validation Statistics in
Regression Models for Binary Responses

Denuit, M., Mefioui, M. and J. Trufin

BOUNDS ON CONCORDANCE-BASED VALIDATION STATISTICS IN REGRESSION MODELS FOR BINARY RESPONSES

Michel Denuit

Institute of Statistics, Biostatistics and Actuarial Science (ISBA)
Université Catholique de Louvain (UCL)
Louvain-la-Neuve, Belgium

Mhamed Mesfioui

Département de mathématiques et d'informatique
Université du Québec à Trois-Rivières
Trois-Rivières (Québec)
Canada G9A 5H7

Julien Trufin

Department of Mathematics
Université Libre de Bruxelles (ULB)
Bruxelles, Belgium

December 15, 2016

Abstract

Association measures based on concordance, such as Kendall's tau, Somers' delta or Goodman and Kruskal's gamma are often used to measure explained variations in regression models for binary outcomes. As responses only assume values in $\{0, 1\}$, these association measures are constrained, which makes their interpretation more difficult as a relatively small value may in fact strongly support the fitted model. In this paper, we derive the set of attainable values for concordance-based association measures in this setting so that the closeness to the best-possible fit can be properly assessed.

Keywords: concordance and discordance, correlation, conditional expectation, logistic regression, GLM.

1 Introduction and motivation

Consider a binary response $Y \in \{0, 1\}$, with 1 coding “success” and 0 coding “failure”, say. There are numerous situations in which the analyst wants to predict Y by means of a set of covariates $X_1, \dots, X_p$. See e.g. Lombrozo (2007) or Martignon et al. (2008) and the references therein for applications in psychology. The covariates are generally combined in a linear or additive way to form a score S . This score brings some information about Y . Throughout this paper, we assume that the regression function $s \mapsto E[Y|S = s]$ is strictly increasing, i.e. that the larger S , the larger Y on average. This assumption is fulfilled in the vast majority of regression and classification models.

There are various ways to assess the quality of a regression model. We refer the interested reader e.g. to Forster (2000). For a dichotomous response Y , this includes the overall model evaluation, tests about specific regression parameters as well as validation based on predicted probabilities. In this paper, we are interested in the latter aspect. Once the model has been fitted, we have a set of observed response values together with the corresponding predicted success probabilities. It is then natural to check whether high probabilities are indeed associated with observed successes and low probabilities with observed failures. The degree to which the predicted probabilities agree with the actual outcomes can be expressed using a measure of association such as Kendall’s tau, Goodman and Kruskal’s gamma or Somers’ delta. See e.g. Mittlbok and Schemper (1996) or Peng et al. (2002). These measures of association rely on the concept of concordance. Recall that a pair of observations is said to be concordant if the observation with the larger value of the first component has also the larger value for the second component. The pair is said to be discordant if the observation with the larger value of the first component has the smaller value of the second component. In our case, this means that larger predicted success probabilities, or higher scores are associated with more responses equal to 1, and vice versa.

Kendall’s tau is a classical measure of association based on this construction, defined as the probability of concordance minus the probability of discordance. Whereas Kendall’s tau is an efficient tool for measuring the strength of dependence between continuous outcomes, it loses many of its good properties when it is applied to discrete variables. In particular, it is no more distribution-free and its range is restricted to a sub-interval of $[-1, 1]$. As we will see further in this work, Kendall’s tau cannot attain a very large value because a large proportion of the pairs are tied. Therefore, several dependence measures based on concordance and discordance probabilities have been introduced to deal with discrete random variables. Their respective differences lie in the treatment of ties. Such concordance-based dependence measures include Kendall’s tau b, Stuart’s tau c, Goodman and Kruskal’s gamma and Somers’ delta. For a general presentation of these association measures, we refer the interested reader e.g. to Agresti (1996).

It has been documented that some of these association measures cannot attain a value of 1 even when the outcome is completely determined by the predictor, i.e. when the fit is perfect. See e.g. Table 3 in Mittlbok and Schemper (1996). In this paper, we derive best-possible bounds on these measures of association. Comparing the actual values to these bounds helps data analysts in deciding whether the agreement between predicted probabilities and observed outcomes is high enough to support the candidate model. The remainder of this paper is organized as follows. Section 2 describes the regression models for dichotomous

responses considered in this paper. In Section 3, we recall the definition of the association measures considered in this paper and we establish several useful representations. In Section 4, we derive the best-possible bounds on these association measures. The final Section 5 discusses the results based on numerical illustrations and concludes the paper.

2 Regression model

2.1 Regression function

Let us consider a binary (or dichotomous) response $Y \in \{0, 1\}$ predicted by means of a score S , with

$$P[Y = 1] = E[P[Y = 1|S]] = p \in (0, 1).$$

Throughout the paper, we assume that the regression function

$$s \mapsto E[Y|S = s] = P[Y = 1|S = s]$$

is strictly increasing, with

$$\lim_{s \rightarrow -\infty} P[Y = 1|S = s] = 0 \text{ and } \lim_{s \rightarrow +\infty} P[Y = 1|S = s] = 1.$$

The monotonicity assumption on the conditional expectation ensures that Y and S are positively related. In particular, the following inequalities for the covariances

$$C[Y, S] = C[E[Y|S], S] \geq 0$$

and

$$C[Y, E[Y|S]] = V[E[Y|S]] \geq 0$$

both hold when the regression function is increasing.

For binary responses, the monotonicity condition imposed to the regression function is equivalent to regression dependence as defined by Lehmann (1966), which requires that Y increases with S in first-degree stochastic dominance. In general, however, this is not true and we refer the reader to Shea (1979) for a detailed analysis.

2.2 Examples

Typical examples include the logistic link function such that

$$s = \ln \frac{P[Y = 1|S = s]}{P[Y = 0|S = s]} \Leftrightarrow P[Y = 1|S = s] = \frac{\exp(s)}{1 + \exp(s)},$$

the probit link function such that

$$s = \Phi^{-1}(P[Y = 1|S = s]) \Leftrightarrow P[Y = 1|S = s] = \Phi(s)$$

where Φ denotes the distribution function of the standard Normal distribution, and complementary log-log link such that

$$s = \ln(-\ln P[Y = 0|S = s]) \Leftrightarrow P[Y = 1|S = s] = 1 - \exp(-\exp(s)).$$

Notice that the validation procedures discussed in the present paper do not question the appropriateness of the chosen link function as long as it remains increasing, but they only consider the association of observed successes with larger predicted success probabilities. As it will become clear in the remainder of this paper, only the dependence between the score S and the response Y matters. Our approach thus also applies to the case where the link function is estimated in a nonparametric way, as with projection pursuit estimates for instance, as well as when machine learning tools are used to predict Y .

2.3 Two cases of interest

In the remainder of the paper, we consider two cases:

Case 1 all the covariates $X_1, \dots, X_p$ are discrete or categorical so that the score S assumes its values in a finite or countable subset of the real line. Henceforth, we denote the support of S as $\{s_1, s_2, \dots, s_m\}$ with $s_1 < s_2 < \dots < s_m$. In this case, the distribution function F_S of S is a step function, equal to j/m when its argument falls between s_j and s_{j+1} . We further assume that $s \mapsto E[Y|S = s]$ is continuous in this case.

Case 2 the score is continuous, with (an interval of) the real line as support. This is for instance the case when at least one of the covariates is continuous so that the score also is. Further, we assume that the distribution function F_S of the score S is continuous and strictly increasing.

Notice that, even in Case 2, we are back to Case 1 when dealing with observed data as soon as we use the empirical distribution function of the fitted scores for inference purposes.

3 Concordance-based association measures

3.1 Concordance and discordance

Consider independent copies (Y_1, Z_1) and (Y_2, Z_2) of (Y, Z) . Here, Y is the binary response and Z can be either the score S or the predicted success probability $P[Y = 1|S]$. Then, (Y_1, Z_1) and (Y_2, Z_2) are said to be concordant if $(Y_1 - Y_2)(Z_1 - Z_2) > 0$ holds true whereas they are said to be discordant when $(Y_1 - Y_2)(Z_1 - Z_2) < 0$.

Many tied pairs (that is, pairs of observations that have equal values of Y or Z) occur in practice. If all the covariates are categorical (Case 1) so that the score is discrete then pairs of observations that have equal values of Z may be encountered. In Case 2, scores are continuous so that no ties occur for the second component. Specifically, the probability that a tie occurs is given by

$$P[(Y_1 - Y_2)(Z_1 - Z_2) = 0] = \begin{cases} P[Y_1 = Y_2 \text{ or } Z_1 = Z_2] & \text{in Case 1} \\ P[Y_1 = Y_2] & \text{in Case 2.} \end{cases}$$

The following property will be useful in the remainder of this paper.

Property 3.1. *If the regression function is continuous and strictly increasing,*

$$P[(Y_1 - Y_2)(S_1 - S_2) > 0] = P[(Y_1 - Y_2)(P[Y_1 = 1|S_1] - P[Y_2 = 1|S_2]) > 0].$$

Proof. The probability of concordance is not modified when Z is transformed using a continuous strictly increasing function g , i.e.

$$P[(Y_1 - Y_2)(Z_1 - Z_2) > 0] = P[(Y_1 - Y_2)(g(Z_1) - g(Z_2)) > 0].$$

Considering $Z_i = S_i$ and for g the regression function, i.e. $g(s) = E[Y|S = s]$, yields the announced result. $\square$

Based on Property 3.1, the assumption that the regression function $s \mapsto E[Y|S = s]$ is continuous and strictly increasing implies that the concordance probability for the pair $(Y, E[Y|S])$ coincides with the concordance probability for the pair (Y, S) .

Let us now establish some useful expressions for concordance probabilities.

Property 3.2. *Let H denote the joint distribution function of the pair (Y, Z) . We then have*

$$\begin{aligned} P[(Y_1 - Y_2)(Z_1 - Z_2) > 0] &= 2E[H(Y, Z)] - P[Y_1 = Y_2] - P[Z_1 = Z_2] \\ &= 2P[Y_1 = 0, Y_2 = 1, Z_1 < Z_2]. \end{aligned}$$

Proof. As Z_1 and Z_2 are independent and identically distributed, we have

$$\begin{aligned} P[Z_1 \leq Z_2] &= P[Z_1 < Z_2] + P[Z_1 = Z_2] \\ &= P[Z_1 > Z_2] + P[Z_1 = Z_2] \\ &= \frac{1 - P[Z_1 = Z_2]}{2} + P[Z_1 = Z_2] \\ &= \frac{1 + P[Z_1 = Z_2]}{2}. \end{aligned}$$

This allows us to write

$$\begin{aligned} P[Y_1 \leq Y_2, Z_1 \leq Z_2] &= 1 - P[Y_1 > Y_2] - P[Z_1 > Z_2] + P[Y_1 < Y_2, Z_1 < Z_2] \\ &= P[Y_1 < Y_2, Z_1 < Z_2] + \frac{P[Y_1 = Y_2]}{2} + \frac{P[Z_1 = Z_2]}{2}. \end{aligned}$$

The concordance probability can finally be expressed as

$$\begin{aligned} P[(Y_1 - Y_2)(Z_1 - Z_2) > 0] &= 2P[Y_1 < Y_2, Z_1 < Z_2] \\ &= 2P[Y_1 \leq Y_2, Z_1 \leq Z_2] - P[Y_1 = Y_2] - P[Z_1 = Z_2], \quad (3.1) \end{aligned}$$

as announced. Considering the second expression, starting from

$$P[Y_1 < Y_2, Z_1 < Z_2] = P[Y_1 > Y_2, Z_1 > Z_2] = \frac{P[(Y_1 - Y_2)(Z_1 - Z_2) > 0]}{2}.$$

Then,

$$P[(Y_1 - Y_2)(Z_1 - Z_2) > 0] = 2P[Y_1 = 0, Y_2 = 1, Z_1 < Z_2],$$

which ends the proof. $\square$

Notice that in (3.1), only the first term involves both Y and Z , the two other ones depending only on marginal distributions of Y and Z , respectively.

The next result shows that the maximum value for the concordance probability is attained when Y and Z are perfectly dependent. It extends previous results by Denuit and Lambert (2005) and Mesfioui and Tajar (2005) who considered pairs of counting random variables whereas here, we deal with a binary response Y together with a possibly continuous score S .

Proposition 3.3. *Let us consider the random pair (Y^u, Z^u) obeying the Fréchet-Höfding upper bound, i.e.*

$$Z^u = F_Z^{-1}(U) \text{ and } Y^u = \begin{cases} 0 & \text{if } U \leq 1 - p \\ 1 & \text{if } U > 1 - p \end{cases}$$

where U is uniformly distributed over the unit interval $[0, 1]$. Then, the inequality

$$P[(Y_1 - Y_2)(Z_1 - Z_2) > 0] \leq P[(Y_1^u - Y_2^u)(Z_1^u - Z_2^u) > 0]$$

holds for every (Y, Z) with the same marginals as (Y^u, Z^u) .

Proof. The joint distribution function of the random pair (Y, Z) satisfies

$$H(y, z) \leq \min\{F_Y(y), F_Z(z)\} \text{ for all } y \text{ and } z.$$

This ensures that

$$E[H(Y, Z)] \leq E[\min\{F_Y(Y), F_Z(Z)\}]$$

holds true. Now, the inequality $E[g(Y, Z)] \leq E[g(Y^u, Z^u)]$ is known to be valid for every supermodular function g (see e.g. Denuit et al., 2005, Section 6.2.4). As every joint distribution function is supermodular, we also have

$$E[\min\{F_Y(Y), F_Z(Z)\}] \leq E[\min\{F_Y(Y^u), F_Z(Z^u)\}],$$

so that

$$E[H(Y, Z)] \leq E[\min\{F_Y(Y^u), F_Z(Z^u)\}]$$

is true. Hence, as

$$P[Y^u \leq y, Z^u \leq z] = \min\{F_Y(y), F_Z(z)\},$$

we have the announced result by Property 3.2. □

3.2 Kendall's tau

Kendall's tau (also known as Kendall's tau a, to distinguish it from its variants discussed in Remark 3.5 below) is a widely used measure of dependence between Y and Z , defined as

$$\tau[Y, Z] = P[(Y_1 - Y_2)(Z_1 - Z_2) > 0] - P[(Y_1 - Y_2)(Z_1 - Z_2) < 0].$$

With continuous random variables, one can show that Kendall's tau is completely determined by the copula and unrelated to the marginal distributions. This is no more true in general as shown for instance in Neslehova (2007) who studies Kendall's tau for random variables that are not necessarily continuous. When the involved random variables are valued in the non-negative integers, we refer the reader to Denuit and Lambert (2005) for a detailed study. The general discrete case is covered by Mesfioui and Tajar (2005).

The next result establishes the general expression for Kendall's tau in our context.

Property 3.4. *Let H denote the joint distribution function of the pair (Y, Z) . We then have*

$$\tau[Y, Z] = 4E[H(Y, Z)] - P[Y_1 = Y_2] - P[Z_1 = Z_2] - P[Y_1 = Y_2, Z_1 = Z_2] - 1.$$

If Z is continuous (Case 2), the latter expression simplifies into

$$\tau[Y, Z] = 4E[H(Y, Z)] - 2(1 - p(1 - p)).$$

Proof. As

$$P[(Y_1 - Y_2)(Z_1 - Z_2) > 0] + P[(Y_1 - Y_2)(Z_1 - Z_2) < 0] + P[(Y_1 - Y_2)(Z_1 - Z_2) = 0] = 1$$

we have

$$\tau[Y, Z] = 2P[(Y_1 - Y_2)(Z_1 - Z_2) > 0] - 1 + P[(Y_1 - Y_2)(Z_1 - Z_2) = 0].$$

The announced result then directly follows from Property 3.2. The second part of the result comes from

$$P[Z_1 = Z_2] = P[Y_1 = Y_2, Z_1 = Z_2] = 0$$

when Z is continuous, and

$$P[Y_1 = Y_2] = p^2 + (1 - p)^2.$$

Hence,

$$\tau[Y, Z] = 4E[H(Y, Z)] - P[Y_1 = Y_2] - 1 = 4E[H(Y, Z)] - 2(1 - p(1 - p)),$$

which ends the proof □

Remark 3.5. Variants of Kendall's tau have been proposed in the literature, to address the issue of ties. For instance, Kendall's tau b has been proposed by Kendall (1945) and is defined as

$$\tau_b[Y, Z] = \frac{\tau[Y, Z]}{\sqrt{P[Y_1 \neq Y_2]} \sqrt{P[Z_1 \neq Z_2]}}.$$

The denominator does not involve the joint distribution of (Y, Z) so that the upper bound on $\tau_b[Y, Z]$ is easily derived from the upper bound on $\tau[Y, Z]$ established in the next section. Also, we do not discuss Kendall's tau c proposed by Stuart (1953) as it reduces to $\tau_c = 2\tau$ when binary variables are involved.

3.3 Goodman and Kruskal's gamma

The measure gamma proposed by Goodman and Kruskal (1954) is a conditional version of Kendall's tau, given that no tie occurs. Specifically,

$$\begin{aligned} \gamma[Y, Z] &= \frac{P[(Y_1 - Y_2)(Z_1 - Z_2) > 0] - P[(Y_1 - Y_2)(Z_1 - Z_2) < 0]}{P[(Y_1 - Y_2)(Z_1 - Z_2) > 0] + P[(Y_1 - Y_2)(Z_1 - Z_2) < 0]} \\ &= \frac{\tau[Y, Z]}{1 - P[(Y_1 - Y_2)(Z_1 - Z_2) = 0]}. \end{aligned}$$

For continuous random variables, Goodman and Kruskal's gamma obviously coincides with Kendall's tau. Goodman and Kruskal's gamma is based on the numbers of concordant and discordant pairs of observations and ignores tied pairs of observations.

For a study of the properties of Goodman and Kruskal's gamma, we refer the interested reader e.g. to Agresti (1990).

3.4 Somers' delta

Somers (1962) proposed a measure similar to Goodman and Kruskal's gamma, but for which pairs of untied on Y serve as the base rather than only those untied on both Y and Z . Population version of Somers' delta is

$$\begin{aligned}\delta[Y, Z] &= \frac{P[(Y_1 - Y_2)(Z_1 - Z_2) > 0] - P[(Y_1 - Y_2)(Z_1 - Z_2) < 0]}{P[Y_1 \neq Y_2]} \\ &= \frac{\tau[Y, Z]}{1 - P[Y_1 = Y_2]}.\end{aligned}$$

It is worth mentioning that Somers' delta is not symmetric, a fact often regarded as undesirable except in regression problems where the response and the score do not play the same role. As ties can only occur in one component in the binary regression case under Case 2, Somers' delta and Goodman and Kruskal's gamma coincide.

The denominator in $\delta[Y, Z]$ does not depend on the joint distribution of (Y, Z) so that bounds on Somers' delta are easily derived from the bounds on Kendall's tau obtained in the next section.

4 Bounds on concordance-based association measures

In order to measure the goodness-of-fit, we aim to measure the strength of the association between the response $Y \in \{0, 1\}$ and the corresponding predicted success probability $P[Y = 1|S] \in [0, 1]$. Hereafter, we derive the upper bound on such goodness-of-fit measures.

4.1 Bounds on concordance probabilities

4.1.1 Case 1

In this case, Z is valued in $\{z_1, z_2, \dots, z_m\}$ with

$$z_j = \begin{cases} s_j & \text{when } Z = S \\ E[Y|S = s_j] & \text{when } Z = E[Y|S]. \end{cases}$$

Notice that $z_1 < z_2 < \dots < z_m$. Define

$$z_{j^*} = F_Z^{-1}(1 - p) = \inf\{z \in \mathbb{R} | F_Z(z) \geq 1 - p\}$$

and set $z_0 = 0$. As F_Z is a step function, we have

$$z_{j^*} = z_j \text{ when } 1 - \frac{j+1}{m} < p \leq 1 - \frac{j}{m}.$$

Proposition 4.1. *In Case 1,*

$$P[(Y_1 - Y_2)(Z_1 - Z_2) > 0] \leq 2p(1 - p) - 2(F_Z(z_{j^*}) - 1 + p)(1 - p - F_Z(z_{j^*-1}))$$

and the upper bound can be attained.

Proof. By Proposition 3.3, it suffices to show that

$$P[(Y_1 - Y_2)(Z_1 - Z_2) > 0] = 2p(1 - p) - 2(F_Z(z_{j^*}) - 1 + p)(1 - p - F_Z(z_{j^*-1}))$$

when the random pair (Y, Z) obeys the upper Fréchet-Höfding bound, i.e. when

$$Z_i = F_Z^{-1}(U_i) \text{ and } Y_i = \begin{cases} 0 & \text{if } U_i \leq 1 - p \\ 1 & \text{if } U_i > 1 - p \end{cases} \quad (4.1)$$

for $i = 1, 2$, where U_1 and U_2 are independent random variables, uniformly distributed over the unit interval $[0, 1]$. In that case, we have

$$\begin{aligned} P[(Y_1 - Y_2)(Z_1 - Z_2) > 0] &= 2P[Y_1 = 0, Y_2 = 1, Z_1 < Z_2] \\ &= 2P[U_1 \leq 1 - p, U_2 > 1 - p, F_Z^{-1}(U_1) < F_Z^{-1}(U_2)] \\ &= 2P[F_Z^{-1}(U_1) < F_Z^{-1}(U_2) | U_1 \leq 1 - p, U_2 > 1 - p] \\ &\quad P[U_1 \leq 1 - p, U_2 > 1 - p] \\ &= 2(1 - P[F_Z^{-1}(U_1) = F_Z^{-1}(U_2) | U_1 \leq 1 - p, U_2 > 1 - p])(1 - p)p. \end{aligned}$$

Now, since

$$\begin{aligned} &P[F_Z^{-1}(U_1) = F_Z^{-1}(U_2) | U_1 \leq 1 - p, U_2 > 1 - p] \\ &= P[U_1 > F_Z(z_{j^*-1}), U_2 \leq F_Z(z_{j^*}) | U_1 \leq 1 - p, U_2 > 1 - p] \\ &= P[U_1 > F_Z(z_{j^*-1}) | U_1 \leq 1 - p] P[U_2 \leq F_Z(z_{j^*}) | U_2 > 1 - p] \\ &= \frac{P[F_Z(z_{j^*-1}) < U_1 \leq 1 - p]}{1 - p} \frac{P[1 - p < U_2 \leq F_Z(z_{j^*})]}{p} \\ &= \frac{(1 - p - F_Z(z_{j^*-1}))}{1 - p} \frac{(F_Z(z_{j^*}) - 1 + p)}{p} \end{aligned} \quad (4.2)$$

we get

$$P[(Y_1 - Y_2)(Z_1 - Z_2) > 0] = 2p(1 - p) - 2(F_Z(z_{j^*}) - 1 + p)(1 - p - F_Z(z_{j^*-1})).$$

This ends the proof. $\square$

4.1.2 Case 2

Let us now turn to the case where Z is continuous.

Proposition 4.2. *In Case 2,*

$$P[(Y_1 - Y_2)(Z_1 - Z_2) > 0] \leq 2p(1 - p)$$

and the upper bound can be attained.

Proof. By Proposition 3.3, it suffices to show that the concordance probability is equal to $2p(1-p)$ when the random pair (Y, Z) obeys the upper Fréchet-Höfding bound, i.e. when (4.1) holds. This is indeed the case as

$$\begin{aligned} P[(Y_1 - Y_2)(Z_1 - Z_2) > 0] &= 2P[Y_1 = 0, Y_2 = 1, Z_1 < Z_2] \\ &= 2P[U_1 \leq 1-p, U_2 > 1-p, U_1 < U_2] \\ &= 2P[U_1 \leq 1-p]P[U_2 > 1-p] \\ &= 2p(1-p). \end{aligned}$$

This ends the proof. $\square$

Notice that the upper bound established in Proposition 4.2 is related to the one of Proposition 4.1 as $F_Z(z_{j*})$ becomes $1-p$ in the continuous case. The second term appearing in the upper bound of Proposition 4.1 is an improvement compared to Proposition 4.2 when the range of the score is constrained to be discrete.

4.2 Kendall's tau

In her study of Kendall's tau for random variables not necessarily continuous, Neslehova (2007) derives the following upper bound on Kendall's tau:

$$\tau[Y, Z] \leq \sqrt{(1 - E[F_Y(Y) - F_Y(Y-)])(1 - E[F_Z(Z) - F_Z(Z-)])}, \quad (4.3)$$

where $F_Y(y-) = P[Y < y]$ and $F_Z(z-) = P[Z < z]$. For binary responses Y ,

$$E[F_Y(Y) - F_Y(Y-)] = (1-p)^2 + p^2$$

and

$$E[F_Z(Z) - F_Z(Z-)] = \begin{cases} \sum_{j=1}^m (P[Z = z_j])^2 & \text{in Case 1} \\ 0 & \text{in Case 2.} \end{cases}$$

Thus, (4.3) becomes

$$\tau[Y, Z] \leq \begin{cases} \sqrt{2p(1-p) \left(1 - \sum_{j=1}^m (P[Z = z_j])^2\right)} & \text{in Case 1} \\ \sqrt{2p(1-p)} & \text{in Case 2.} \end{cases} \quad (4.4)$$

Mesfioui and Quessy (2010) provide the following upper bound on Kendall's tau

$$\tau[Y, Z] \leq 2 \min \{E[F_Y(Y-)], E[F_Z(Z-)]\}, \quad (4.5)$$

which improves the one obtained in Neslehova (2007) since

$$2E[F_Y(Y-)] = 1 - E[F_Y(Y) - F_Y(Y-)]$$

and

$$2E[F_Z(Z-)] = 1 - E[F_Z(Z) - F_Z(Z-)].$$

For a binary response Y , (4.5) can be rewritten as

$$\tau[Y, Z] \leq \begin{cases} \min \left\{ 2p(1-p), 1 - \sum_{j=1}^m (\mathbb{P}[Z = z_j])^2 \right\} & \text{in Case 1} \\ 2p(1-p) & \text{in Case 2.} \end{cases} \quad (4.6)$$

In the remainder of this section, we derive upper bounds sharper than (4.6) in Case 1 while we find back the same upper bound in Case 2.

4.2.1 Case 1

Let us start with the case where Z is valued in $\{z_1, \dots, z_m\}$.

Property 4.3. *In Case 1,*

$$\tau[Y, Z] \leq 2p(1-p) - 2(F_Z(z_{j^*}) - 1 + p)(1 - p - F_Z(z_{j^*-1}))$$

and the upper bound can be attained.

Proof. By Propositions 3.3 and 4.1, it suffices to notice that when the random pair (Y, Z) obeys the upper Fréchet-Höfding bound, i.e. under (4.1), we get

$$\begin{aligned} \mathbb{P}[(Y_1 - Y_2)(Z_1 - Z_2) < 0] &= 2\mathbb{P}[Y_1 = 1, Y_2 = 0, Z_1 < Z_2] \\ &= 2\mathbb{P}[U_1 > 1 - p, U_2 \leq 1 - p, F_Z^{-1}(U_1) < F_Z^{-1}(U_2)] = 0. \end{aligned}$$

□

4.2.2 Case 2

Let us now turn to the case where Z is continuous.

Property 4.4. *In Case 2,*

$$\tau[Y, Z] \leq 2p(1-p)$$

and the upper bound can be attained.

Proof. We know from Proposition 4.2 that there is a joint distribution for (Y, Z) maximizing the concordance probability and setting the discordance probability to zero. The upper bound is obtained in this case and is equal to

$$1 - \mathbb{P}[Y_1 = Y_2] = 1 - (p^2 + (1-p)^2) = 2p(1-p).$$

This ends the proof. □

The upper bound in Property 4.4 is the one obtained by Mesfioui and Quessy (2010), which appears to be the best-possible one. It is not affected by the distribution of the score S . It only depends on p and cannot exceed 0.5, the value obtained when $p = 0.5$. This constraint must be taken into account when interpreting the values obtained for Kendall's tau in a regression analysis as a relatively small value may in reality be so close to the upper bound that it strongly supports the model fit.

Comparing Properties 4.3 and 4.4, we see that the second term in the upper bound of Property 4.3 improves the upper bound of Property 4.4 when the range of the score is constrained to be discrete. In Case 2, $F_Z(z_{j^*})$ becomes $1 - p$ and this term disappears.

Remark 4.5. Let us further investigate the particular case when (Y, S) obeys the upper Fréchet-Hoeffding bound. This means that there exists a unit uniform random variable U such that

$$Y = F_Y^{-1}(U) = \mathbf{I}[U > 1 - p] \text{ and } S = F_S^{-1}(U) \quad (4.7)$$

where $\mathbf{I}[A]$ is the indicator of the event A , equal to 1 when A is realized and to 0 otherwise. In this case,

$$\begin{aligned} \mathbb{E}[Y|S = s] &= \mathbb{P}[Y = 1 | F_S^{-1}(U) = s] \\ &= \mathbb{P}[U > 1 - p | U = F_S(s)] \\ &= \mathbf{I}[s > F_S^{-1}(1 - p)] \end{aligned}$$

which implies that $\mathbb{E}[Y|S]$ is a Bernoulli random variable with probability of success p . Thus, Y and $\mathbb{E}[Y|S]$ are perfectly dependent Bernoulli random variables with the same probability of success p .

Recall that Kendall's tau associated to two Bernoulli random variables B_1 and B_2 , with respective means p_1 and p_2 , is of the form

$$\tau[B_1, B_2] = 2p_{00} - 2(1 - p_1)(1 - p_2) \text{ with } p_{00} = \mathbb{P}[B_1 = B_2 = 0].$$

This implies that the upper bound in Property 4.4 is $2p_{00} - 2(1 - p)^2$ with $p_{00} = 1 - p$ under the dependence structure (4.7). This provides an alternative derivation of the result established in Property 4.4.

4.3 Goodman and Kruskal's gamma

4.3.1 Case 1

Let us first consider the case where Z is discrete.

Property 4.6. *In Case 1, $\gamma[Y, Z] \leq 1$ and the upper bound can be attained.*

Proof. From the monotonicity property of Goodman and Kruskal's gamma with respect to the concordance order established in Mesfioui and Tajar (2005, Proposition 2.7), we have that

$$\gamma[Y, Z] \leq \gamma[Y^u, Z^u].$$

It is easily seen that Goodman and Kruskal's gamma can be rewritten as

$$\gamma[Y, Z] = \frac{\tau[Y, Z]}{\tau[Y, Z] - 2\mathbb{P}[(Y_1 - Y_2)(Z_1 - Z_2) < 0]}.$$

As $\mathbb{P}[(Y_1^u - Y_2^u)(Z_1^u - Z_2^u) < 0] = 0$ when (Y^u, Z^u) obeys the upper Fréchet-Höfdding upper bound, as shown in the proof of Property 4.3, we have

$$\gamma[Y^u, Z^u] = \frac{\tau[Y^u, Z^u]}{\tau[Y^u, Z^u]} = 1.$$

This ends the proof. □

4.3.2 Case 2

Let us now turn to the continuous case.

Property 4.7. *In Case 2, $\gamma[Y, Z] \leq 1$ and the upper bound can be attained.*

Proof. Since

$$P[(Y_1 - Y_2)(Z_1 - Z_2) = 0] = P[Y_1 = Y_2] = p^2 + (1 - p)^2,$$

we get

$$\begin{aligned} \gamma[Y, Z] &= \frac{\tau[Y, Z]}{1 - P[(Y_1 - Y_2)(Z_1 - Z_2) = 0]} \\ &= \frac{\tau[Y, Z]}{2p(1 - p)} \end{aligned}$$

so that we deduce from Property 4.4 that the upper bound is 1 and can be attained. This ends the proof. $\square$

5 Discussion

In the preceding sections, we have established upper bounds on concordance probabilities and then on concordance-based association measures, such as Kendall's tau, Somers' delta or Goodman and Kruskal's gamma, holding when the response variable Y is valued in $\{0, 1\}$. The commonly-used Kendall's tau is constrained in such a case and cannot attain the value 1, which makes its interpretation more difficult. Let us now investigate the behavior of Kendall's tau with the help of numerical illustrations.

Let us first assume that the joint distribution function H of the random pair (Y, S) is a member of the Fréchet family. This means that, for all $(y, s) \in \{0, 1\} \times \mathbb{R}$,

$$H(y, s) = \theta \max \{F_Y(y) + F_S(s) - 1, 0\} + (1 - \theta) \min \{F_Y(y), F_S(s)\}, \quad \theta \in [0, 1].$$

Under the Fréchet-Höfding lower bound $\max \{F_Y + F_S - 1, 0\}$, the concordance probability is zero and the discordance probability is $2p(1 - p)$ in Case 2. This directly follows from Property 4.4 by noting that (Y, S) obeys the Fréchet-Höfding lower bound if, and only if $(1 - Y, S)$ obeys the Fréchet-Höfding upper bound with modified success probability $1 - p$. Thus, we get in Case 2 that

$$\begin{aligned} \tau[Y, E[Y|S]] &= \tau[Y, S] \\ &= \theta 2p(1 - p) + (1 - \theta)(-2p(1 - p)) \\ &= 2(2\theta - 1)p(1 - p) \end{aligned}$$

when H belongs to the Fréchet family. In this case, we thus see that $\tau[Y, E[Y|S]]$ linearly increases with the dependence parameter θ while it varies quadratically with the marginal parameter p .

Copulas can also be used to define bivariate distributions with discrete margins. Recall that a copula C is a joint distribution with unit uniform marginals. In opposition to the

situation found in the continuous case, there is in general no unique way to express the joint distribution as a function of their marginal distributions. Sklar's representation in terms of copulas can nevertheless be used in a constructive way to define the joint distribution function of (Y, S) as

$$H(y, s) = P[Y \leq y, S \leq s] = C(F_Y(y), F_S(s)). \quad (5.1)$$

Let us now take for C in (5.1) a member of Ali-Mikhail-Haq family, defined for all $(u, v) \in [0, 1]^2$ by

$$C_\theta(u, v) = \frac{uv}{1 - \theta(1-u)(1-v)}, \quad \theta \in [-1, 1].$$

The next result is useful to compute Kendall's tau when H is obtained from (5.1) for some copula C (in particular, when C is a Ali-Mikhail-Haq copula).

Property 5.1. *In Case 2, when H is obtained from (5.1), we have*

$$\begin{aligned} \tau[Y, E[Y|S]] = \tau[Y, S] &= 4E[H(0, S)] - 2(1-p) \\ &= 4E[C(1-p, U)] - 2(1-p) \end{aligned}$$

where U is a random variable uniformly distributed over $[0, 1]$.

Proof. Define

$$F_{S|Y=1}(s) = P[S \leq s | Y = 1] \text{ and } F_{S|Y=0}(s) = P[S \leq s | Y = 0].$$

Clearly,

$$F_S(s) = pF_{S|Y=1}(s) + (1-p)F_{S|Y=0}(s)$$

and

$$E[F_{S|Y=1}(S) | Y = 1] = E[F_{S|Y=0}(S) | Y = 0] = \frac{1}{2}.$$

We then have

$$\begin{aligned} E[H(Y, S)] &= E[H(0, S) | Y = 0](1-p) + E[F_S(S) | Y = 1]p \\ &= E[F_{S|Y=0}(s) | Y = 0](1-p)^2 + E[F_S(S) | Y = 1]p \\ &= \frac{(1-p)^2}{2} + E[F_S(S) | Y = 1]p. \end{aligned} \quad (5.2)$$

Now,

$$\begin{aligned} E[F_S(S) | Y = 1] &= pE[F_{S|Y=1}(S) | Y = 1] + (1-p)E[F_{S|Y=0}(S) | Y = 1] \\ &= \frac{p}{2} + (1-p) \left(\frac{1}{p}E[F_{S|Y=0}(S)] - \frac{1-p}{p}E[F_{S|Y=0}(S) | Y = 0] \right) \\ &= \frac{p}{2} + \frac{1}{p}E[H(0, S)] - \frac{(1-p)^2}{2p}. \end{aligned} \quad (5.3)$$

We conclude from (5.2) and (5.3) that

$$E[H(Y, S)] = \frac{p^2}{2} + E[H(0, S)] = \frac{p^2}{2} + E[C(1-p, U)].$$

and the announced result is then deduced from Property 3.4. $\square$

In Case 2 with C in (5.1) a member of Ali-Mikhail-Haq family, one deduces from Property 5.1 that

$$\tau[Y, S] = 4E[H(0, S)] - 2(1 - p) = 4 \int_0^1 C_\theta(1 - p, u) du - 2(1 - p),$$

which, after standard calculations, leads to

$$\tau[Y, S] = \frac{4(1 - p)}{\theta^2 p^2} ((1 - \theta p) \ln(1 - \theta p) + p\theta) - 2(1 - p), \quad \theta \in [-1, 1] \setminus \{0\}$$

with $\tau[Y, S] = 0$ when $\theta = 0$. We notice that the copulas $\max\{u + v - 1, 0\}$ and $\min\{u, v\}$ corresponding to the Fréchet-Höfding lower and upper bounds, respectively, do not belong to the Ali-Mikhail-Haq family. Thus, the bounds for $\tau[Y, S]$ corresponding to this model are different from the ones established in Property 4.4. In fact, the upper bound on $\tau[Y, S]$ for the current model is

$$\frac{4(1 - p)}{p^2} ((1 - p) \ln(1 - p) + p) - 2(1 - p). \quad (5.4)$$

Figure 5.1 displays for $p \in \{0.05, 0.1, 0.3, 0.5, 0.9\}$ the values for $\tau[Y, S]$ in Case 2 when the joint distribution function H of (Y, S) belongs to the Fréchet family or is obtained from (5.1) with C in the Ali-Mikhail-Haq copula family. The horizontal line corresponding to the upper bound obtained in Property 4.4 is also visible. When p is small or large (i.e. close to 0 or 1), the upper bound is small. For instance, a Kendall's tau above 0.08 or 0.15 when $p = 0.05$ or $p = 0.1$, respectively, can be considered as large given the restricted range of admissible values. Such values thus support the model fit.

As mentioned previously, the copula $\min\{u, v\}$ corresponding to the Fréchet-Höfding upper bound does not belong to the Ali-Mikhail-Haq family so that the value of Kendall's tau in this case stays below the upper bound, being constrained by (5.4). Also, the upper bound from Property 4.4 and the value for Kendall's tau when H belongs to the Fréchet family remains unaffected when p is replaced with $1 - p$. However, this is no more true when H is obtained from (5.1) with C in the Ali-Mikhail-Haq copula family.

To end with, let us mention that squared values of association measure are sometimes used (see e.g. Mittlbock and Schempen, 1996). The bounds derived in the present paper are easily adapted to this setting. Notice that the increasingness of the regression function ensures that the response and the predicted success probability are positively correlated.

Acknowledgements

This work originates from discussions with members of the Addactis team, a consulting company offering software solutions to the insurance industry. Michel Denuit and Julien Trufin would like to thank Michael Casalnuovo and Stéphanie Dausque for interesting discussions about the use of association measures in binary regression models. Michel Denuit acknowledges the financial support from the contract “Projet d’Actions de Recherche Concertées” No 12/17-045 of the “Communauté française de Belgique”, granted by the “Académie universitaire Louvain”. Mhamed Mesfioui acknowledges the financial support of the Natural Sciences and Engineering Research Council of Canada No 261968-2013.

Figure 5.1: Values for $\tau[Y, S]$ as a function of $\theta \in [0.5, 1]$ in Case 2 when the joint distribution function H of (Y, S) belongs to the Fréchet family or is obtained from (5.1) with C in the Ali-Mikhail-Haq (AMH, in short) copula family together with the upper bound obtained in Property 4.4. From upper left to lower right: $p = 0.05, 0.1, 0.3, 0.5$ and 0.9 .

References

- Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.
- Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. Wiley, New York.
- Denuit, M., Dhaene, J., Goovaerts, M.J., Kaas, R. (2005). *Actuarial Theory for Dependent Risks: Measures, Orders and Models*. Wiley, New York.
- Denuit, M., Lambert, P. (2005). Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis* 93, 40-57.
- Forster, M. R. (2000). Key concepts in model selection: Performance and generalizability. *Journal of Mathematical Psychology* 44, 205-231.
- Goodman, L.A., Kruskal, W.H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association* 49, 732-764.
- Kendall, M.G. (1945). The treatment of ties in rank problems. *Biometrika* 33, 239-251.
- Lehmann, E.L. (1966). Some concepts of dependence. *Annals of Mathematical Statistics* 37, 1137-1153.
- Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology* 55, 232-257.
- Martignon, L., Katsikopoulos, K. V., Woike, J. K. (2008). Categorization with limited resources: A family of simple heuristics. *Journal of Mathematical Psychology* 52, 352-361.
- Mesfioui, M., Quessy, J. F. (2010). Concordance measures for multivariate non-continuous random vectors. *Journal of Multivariate Analysis* 101, 2398-2410.
- Mesfioui, M., Tajar, A. (2005). On the properties of some nonparametric concordance measures in the discrete case. *Nonparametric Statistics* 17, 541-554.
- Mittlbock, M., Schemper, M. (1996). Explained variation for logistic regression. *Statistics in Medicine* 15, 1987-1997.
- Neslehova, J. (2007). On rank correlation measures for non-continuous random variables. *Journal of Multivariate Analysis* 98, 544-567.
- Peng, C. Y. J., Lee, K. L., Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research* 96, 3-14.
- Shea, G. (1979). Monotone regression and covariance structure. *The Annals of Statistics* 7, 1121-1126.
- Somers, R.H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review* 27, 799-811.
- Stuart, A. (1953). The estimation and comparison of strengths of association in contingency tables. *Biometrika* 40, 105-110.