CONCORDANCE-BASED PREDICTIVE MEASURES IN REGRESSION MODELS FOR DISCRETE RESPONSES

Michel Denuit;
Mhamed Mesoui and Julien Trufin

DISCUSSION PAPER | 2019 / 05

CONCORDANCE-BASED PREDICTIVE MEASURES IN REGRESSION MODELS FOR DISCRETE RESPONSES

Michel Denuit

Institute of Statistics, Biostatistics and Actuarial Science
Louvain Institute of Data Analysis and Modeling, UC Louvain
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

November 22, 2018

Abstract

Dependence measures are often used in practice in order to assess the quality of a regression model. This is for instance the case with Kendall's tau and other association coefficients based on concordance probabilities. However, in case the response variable is discrete, correlation indices are often bounded and restricted to a sub-interval of $[-1, 1]$. Hence, in this context, small positive values of Kendall's tau may actually support goodness of prediction when getting close to its highest attainable value. In this paper, we derive the best-possible upper bounds for Kendall's tau when the response variable is discrete. Two cases are considered, depending on whether the score is continuous or discrete. Also, we illustrate the obtained upper bounds on a motor third-party liability insurance portfolio.

Keywords: concordance and discordance, Kendall's tau, upper bounds, discrete response, regression models.

1 Introduction and motivation

Consider a discrete response Y taking values in $\mathbb{N} = \{0, 1, 2, \dots\}$ and a set of explanatory variables $X_1, \dots, X_p$ gathered in the vector $\mathbf{X}$. The probability mass function of the response is denoted as $p_k = P[Y = k]$, $k \in \mathbb{N}$. Often in practice, the actuarial analyst aims to predict Y by means of $\mathbf{X}$. Typically, Y represents the number of claims in insurance applications. We refer the reader to Denuit et al. (2018) for binary responses $Y \in \{0, 1\}$.

The explanatory variables $X_1, \dots, X_p$ are generally combined in a linear (with a generalized linear model, or GLM) or additive way (with a generalized additive model, or GAM) to form a score S , or the feature space is partitioned to produce a piecewise constant score (with classification and regression tree, or CART). This score brings some information about Y and is mapped to the response scale to get the predictor $\pi(S)$. Throughout this paper, we assume that the predictor function $s \mapsto \pi(s)$ is strictly increasing. This assumption is fulfilled in the vast majority of regression and classification models (π being the inverse of the monotonic link function).

The score S can be discrete or continuous. In case all the explanatory variables are categorical, S will be discrete, while in case at least one explanatory variable is continuous, S will be either continuous or discrete, depending on the regression model considered. Indeed, in the latter case, a GAM will for instance produce a continuous S while a CART of limited size will provide a discrete S .

The quality of a regression model can be assessed in various ways. We can quote the overall model evaluation, tests about specific regression parameters or validation based on predicted values. In the latter case, measures of association for the pairs $(Y, \pi(S))$ or (Y, S) are very useful to evaluate the performances of a given model. Besides Gini coefficient, many criteria use concordance probabilities, such as Kendall's tau.

The probability of concordance between the predicted $\pi(S)$ and observed outcomes Y is frequently used to assess the predictive power of a regression model. It also often serves as a basis for ranking predictive strength between different models under consideration. In that respect, Kendall's tau is a classical measure of association defined as the probability of concordance minus the probability of discordance. It is an efficient tool for measuring the strength of dependence between continuous outcomes. However, when Kendall's tau is applied to discrete variables, it is no more distribution-free and its range is restricted to a sub-interval of $[-1, 1]$. This makes its interpretation more difficult as a relatively small value for Kendall's tau may in fact strongly support the fitted model if the maximal possible value for Kendall's tau is small as well.

In this paper, we aim to derive the best-possible upper bounds for Kendall's tau when the response variable is discrete. This enables the actuarial analyst to compare the observed Kendall's tau with its highest possible value in the situation under consideration and come up with more reliable conclusions. In particular, when $p_k = 0$ for $k \geq 2$, our results will coincide with those obtained in Denuit et al. (2018) for binary responses.

The paper is structured as follows. In Section 2, we derive useful expressions for concordance probabilities and we establish the best-possible upper bounds both for S (and hence $\pi(S)$) discrete and continuous. Section 3 is devoted to Kendall's tau. We apply the results obtained in Section 2 to derive the general expression for Kendall's tau as well as the best-possible upper bounds. Section 4 illustrates the obtained upper bounds for Kendall's tau on

a motor third-party liability insurance portfolio.

2 Concordance probability

2.1 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.

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.

2.2 Concordance and discordance

Consider independent copies (Y_1, Z_1) and (Y_2, Z_2) of (Y, Z) . Here, Y is the discrete response and Z can be either the score S or the predicted outcome $\pi(S)$ for some increasing function $\pi(\cdot)$. 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 $(Y_1 - Y_2)(Z_1 - Z_2) < 0$.

Tied pairs (that is, pairs of observations that have equal values of Y or Z) may 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 2.1. *If the function $\pi(\cdot)$ is continuous and strictly increasing,*

$$P[(Y_1 - Y_2)(S_1 - S_2) > 0] = P[(Y_1 - Y_2)(\pi(S_1) - \pi(S_2)) > 0].$$

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 $Z_i = S_i$ ($i = 1, 2$) and for g the function $\pi(\cdot)$ yields the announced result. □

Based on Property 2.1, the assumption that the function $\pi(\cdot)$ is continuous and strictly increasing implies that the concordance probability for the pair $(Y, \pi(S))$ coincides with the concordance probability for the pair (Y, S) .

We first establish expressions for concordance probabilities that will be useful in the following.

Property 2.2. *Let H denote the joint distribution function of the pair (Y, Z) . We then have*

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

Proof. The first equality can be found in Denuit et al. (2018, Property 3.2). Now, since

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

it suffices to notice that

$$P[Y_1 < Y_2, Z_1 < Z_2] = \sum_{k=0}^{\infty} \sum_{l=1}^{\infty} P[Y_1 = k, Y_2 = k + l, Z_1 < Z_2],$$

which gives the second announced equality and ends the proof. $\square$

Property 2.2 shows that concordance probabilities get higher when we replace the joint distribution function H with H_+ whose graph lies everywhere above H , provided the marginals are kept unchanged. A natural candidate for H_+ is the Fréchet-Höfding upper bound H^u defined as

$$H^u(y, z) = \min\{F_Y(y), F_Z(z)\}.$$

We then get

$$P[(Y_1 - Y_2)(Z_1 - Z_2) > 0] \leq P[(Y_1^u - Y_2^u)(Z_1^u - Z_2^u) > 0] \quad (2.1)$$

where (Y_1^u, Z_1^u) and (Y_2^u, Z_2^u) are independent copies of the random pair (Y^u, Z^u) obeying the Fréchet-Höfding upper bound H^u , i.e.

$$Z^u = F_Z^{-1}(U) \text{ and } Y^u = \sum_{k=0}^{\infty} k \mathbf{I}[F_Y(k-1) \leq U < F_Y(k)] \quad (2.2)$$

with U being uniformly distributed over the unit interval $[0, 1]$. We refer the reader to Denuit et al. (2018, Proposition 3.3) for a formal proof. This argument will be used repeatedly in the remainder of this paper.

We are now in a position to establish the two following results which provide upper bounds for concordance probabilities.

Proposition 2.3. *If Z is continuous, we get*

$$P[(Y_1 - Y_2)(Z_1 - Z_2) > 0] \leq 2E[F_Y(Y-)]. \quad (2.3)$$

Proof. By (2.1), since $P[(Y_1^u - Y_2^u)(Z_1^u - Z_2^u) > 0] = 2P[Y_1^u < Y_2^u, Z_1^u < Z_2^u]$, it suffices to show that $P[Y_1^u < Y_2^u, Z_1^u < Z_2^u] = E[F_Y(Y-)]$ with

$$Z_i^u = F_Z^{-1}(U_i) \text{ and } Y_i^u = \sum_{k=0}^{\infty} kI[F_Y(k-1) \leq U_i < F_Y(k)]$$

for $i = 1, 2$, where U_1 and U_2 are independent random variables, uniformly distributed over the unit interval $[0, 1]$. We get

$$\begin{aligned} P[Y_1^u < Y_2^u, Z_1^u < Z_2^u] &= \sum_{k=0}^{\infty} P[Y_1^u = k, Y_2^u > k, Z_1^u < Z_2^u] \\ &= \sum_{k=0}^{\infty} P[F_Y(k-1) \leq U_1 < F_Y(k), F_Y(k) \leq U_2, U_1 < U_2] \\ &= \sum_{k=0}^{\infty} P[F_Y(k-1) \leq U_1 < F_Y(k), F_Y(k) \leq U_2] \\ &= \sum_{k=0}^{\infty} (F_Y(k) - F_Y(k-1)) \bar{F}_Y(k) \\ &= \sum_{k=0}^{\infty} p_k \bar{F}_Y(k) \\ &= E[\bar{F}_Y(Y)], \end{aligned}$$

which ends the proof since $E[F_Y(Y)] + E[F_Y(Y-)] = 1$. $\square$

Let us now consider the case where Z is discrete. We assume that Z is valued in $\{z_1, z_2, \dots, z_m\}$ with $z_1 < z_2 < \dots < z_m$, and we define

$$j_k = \min\{j \in \{1, 2, \dots, m\} : F_Z(z_j) \geq F_Y(k)\}, \quad k \in \mathbb{N}.$$

Proposition 2.4. *If Z is discrete, we have*

$$P[(Y_1 - Y_2)(Z_1 - Z_2) > 0] \leq 2E[F_Y(Y-)] - 2 \sum_{k=0}^{\infty} (F_Y(k) - \max\{F_Y(k-1), F_Z(z_{j_k-1})\}) (F_Z(z_{j_k}) - F_Y(k)).$$

Proof. From (2.1), it amounts to show that

$$P[Y_1^u < Y_2^u, Z_1^u < Z_2^u] = E[F_Y(Y-)] - \sum_{k=0}^{\infty} (F_Y(k) - \max\{F_Y(k-1), F_Z(z_{j_k-1})\}) (F_Z(z_{j_k}) - F_Y(k))$$

with

$$Z_i^u = \sum_{j=1}^m z_j I[F_Z(z_{j-1}) \leq U_i < F_Z(z_j)] \text{ and } Y_i^u = \sum_{k=0}^{\infty} kI[F_Y(k-1) \leq U_i < F_Y(k)]$$

for $i = 1, 2$, where U_1 and U_2 are independent random variables, uniformly distributed over the unit interval $[0, 1]$ and $z_0 = 0$. We then have

$$\begin{aligned}
P[Y_1^u < Y_2^u, Z_1^u < Z_2^u] &= \sum_{k=0}^{\infty} P[Y_1^u = k, Y_2^u > k, Z_1^u < Z_2^u] \\
&= \sum_{k=0}^{\infty} P[Z_1^u < Z_2^u | Y_1^u = k, Y_2^u > k] P[Y_1^u = k, Y_2^u > k] \\
&= \sum_{k=0}^{\infty} P[Z_1^u < Z_2^u | (U_1, U_2) \in \mathcal{A}_k] p_k \bar{F}_Y(k) \\
&= \sum_{k=0}^{\infty} \left(1 - P[Z_1^u = Z_2^u | (U_1, U_2) \in \mathcal{A}_k]\right) p_k \bar{F}_Y(k)
\end{aligned}$$

where

$$\mathcal{A}_k = [F_Y(k-1), F_Y(k)] \times [F_Y(k), 1], \quad k \in \mathbb{N}.$$

Define

$$\mathcal{B}_j = [F_Z(z_{j-1}), F_Z(z_j)] \times [F_Z(z_{j-1}), F_Z(z_j)], \quad j \in \{1, \dots, m\}.$$

Then,

$$\begin{aligned}
P[Z_1^u = Z_2^u | (U_1, U_2) \in \mathcal{A}_k] &= \sum_{j=1}^m P[(U_1, U_2) \in \mathcal{B}_j | (U_1, U_2) \in \mathcal{A}_k] \\
&= \frac{1}{p_k \bar{F}_Y(k)} \sum_{j=1}^m P[(U_1, U_2) \in \mathcal{A}_k \cap \mathcal{B}_j] \\
&= \frac{1}{p_k \bar{F}_Y(k)} \sum_{j=1}^m \alpha_{k,j} \beta_{k,j}
\end{aligned}$$

with

$$\alpha_{k,j} = (\min\{F_Y(k), F_Z(z_j)\} - \max\{F_Y(k-1), F_Z(z_{j-1})\})_+$$

and

$$\beta_{k,j} = (F_Z(z_j) - \max\{F_Y(k), F_Z(z_{j-1})\})_+,$$

where, for any real number r , we let r_+ denote the positive part of r ; that is, $r_+ = r$ if $r \geq 0$ and $r_+ = 0$ if $r < 0$.

For $j < j_k$, we get $\beta_{k,j} = 0$ since $F_Y(k) \geq F_Z(z_j) \geq F_Z(z_{j-1})$. Also, for $j > j_k$, we have $\alpha_{k,j} = 0$ since $F_Z(z_j) \geq F_Z(z_{j-1}) \geq F_Y(k) \geq F_Y(k-1)$. Now, in the remaining case $j = j_k$, it comes $F_Z(z_{j_k}) \geq F_Y(k) \geq F_Z(z_{j_k-1})$ and hence

$$\alpha_{k,j_k} = F_Y(k) - \max\{F_Y(k-1), F_Z(z_{j_k-1})\} \text{ and } \beta_{k,j_k} = F_Z(z_{j_k}) - F_Y(k).$$

Finally, we then get

$$P[Y_1^u < Y_2^u, Z_1^u < Z_2^u] = E[\bar{F}_Y(Y)] - \sum_{k=0}^{\infty} \left(F_Y(k) - \max\{F_Y(k-1), F_Z(z_{j_k-1})\} \right) (F_Z(z_{j_k}) - F_Y(k)),$$

which completes the proof. $\square$

Let us notice that in the particular case where Y is a Bernoulli random variable, namely when $p_k = 0$ for $k \geq 2$, the upper bounds derived in Propositions 2.3 and 2.4 coincide with those obtained in Denuit al. (2018).

3 Kendall's tau

Kendall's tau 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).

More specifically, Neslehova (2007) derives the following upper bound on Kendall's tau:

$$\tau[Y, Z] \leq \sqrt{(1 - E[F_Y(Y) - F_Y(Y-)])(1 - E[F_Z(Z) - F_Z(Z-)]). \quad (3.1)$$

The latter upper bound has been improved by Mesfioui and Quessy (2010) who got the following result:

$$\tau[Y, Z] \leq 2 \min \{E[F_Y(Y-)], E[F_Z(Z-)]\}. \quad (3.2)$$

This is well an improvement compared to (3.1) since

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

and

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

Notice that (3.2) simplifies to $2E[F_Y(Y-)]$ when Z is continuous.

In the following, we derive the best possible upper bounds on Kendall's tau for discrete responses Y . We obtain an upper bound sharper than (3.2) when Z is discrete while we obtain the upper bound (3.2) for Z continuous.

Let us start with the case where Z is continuous.

Property 3.1. *If Z is continuous, we get*

$$\tau[Y, Z] \leq 2E[F_Y(Y-)]. \quad (3.3)$$

Proof. If the random pair (Y, Z) obeys the upper Fréchet-Höfding bound, i.e. under (2.2), we get

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

The maximal value of Kendall's tau corresponds to a random pair distributed according to the Fréchet-Höfding upper bound because this distribution simultaneously maximizes the concordance probability and leads to zero discordance probability. This ends the proof. $\square$

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

Property 3.2. *If Z is discrete, we have*

$$\tau[Y, Z] \leq 2E[F_Y(Y-)] - 2 \sum_{k=0}^{\infty} (F_Y(k) - \max\{F_Y(k-1), F_Z(z_{j_k-1})\}) (F_Z(z_{j_k}) - F_Y(k)). \quad (3.4)$$

Proof. Similarly to the continuous case for Z , when the random pair (Y, Z) obeys the upper Fréchet-Höfding bound, we obviously have $P[(Y_1 - Y_2)(Z_1 - Z_2) < 0] = 0$, so that we directly get the desired result. $\square$

4 Numerical example

Let us illustrate the computation of the upper bounds (3.3) and (3.4) in a situation of practical relevance. In that goal, we consider the motor third-party liability insurance portfolio of an insurance company operating in the EU. The portfolio comprises 124 524 insurance policies that have all been observed during a whole year.

For each policy, we have the numbers of claims filed by the policyholder, that is the response variable Y , and eight explanatory variables $X_1, \dots, X_8$, namely

- AgePh: policyholder's age;
- AgeCar: age of the car;
- Fuel: fuel of the car, with two categories (gas or diesel);
- Split: splitting of the premium, with four categories (annually, semi-annually, quarterly or monthly);
- Cover: extent of the coverage, with three categories (from compulsory third-party liability cover to comprehensive);
- Gender: policyholder's gender, with two categories (female or male);
- Use: use of the car, with two categories (private or professional);
- PowerCat: the engine's power, with five categories.

Figure 4.1 displays for each explanatory variable the number of policies by category/value and Table 4.1 shows the observed numbers of claims.

If we denote by n the number of policies and by Y_i the number of claims of policy i ($i = 1, \dots, n$), then the probabilities p_k can be directly estimated using the empirical proportions

$$\hat{p}_k = \frac{\sum_{i=1}^n \mathbf{I}[Y_i = k]}{n}, \quad k \in \mathbb{N},$$

Figure 4.1: The categories/values of the explanatory variables and their corresponding numbers of policies.

Number of claims	Number of policies
0	110 231
1	13 024
2	1149
3	115
4	5
≥ 5	0

Table 4.1: Descriptive statistics for the number of claims.

where $I[\cdot]$ stands for the indicator function. Hence, based on the observations for the number of claims summarized in Table 4.1, we get the empirical proportions depicted in Table 4.2.

The expectation $E[F_Y(Y-)]$ that we find in (3.3) and (3.4) can be estimated using the

k	0	1	2	3	4	≥ 5
$\hat{p}_k$	88.52189	10.45903	0.92271	0.09235	0.00402	0

Table 4.2: Empirical proportions $\hat{p}_k$ displayed as a percentage.

empirical proportions $\hat{p}_k$ since we have

$$E[F_Y(Y-)] = \sum_{k=1}^{\infty} p_k F_Y(k-1) = \sum_{k=1}^{\infty} \sum_{l=0}^{k-1} p_k p_l.$$

Here, we get $\hat{E}[F_Y(Y-)] = 0.1026812$ with the $\hat{p}_k$ displayed in Table 4.2. Of course, other estimators can be envisaged, exploiting the regression model structure (whereas the proposed one is purely nonparametric).

Therefore, when Z is continuous, which is the case when AgePh and AgeCar are treated as continuous variables using GAM for instance, the upper bound (3.3) for Kendall's tau can be estimated by $2\hat{E}[F_Y(Y-)] = 0.2053624$. The actuarial analyst wishing to assess the predictive performances of the regression model for claim counts using Kendall's tau must therefore compare the obtained value to the upper bound 20.5% and not to 1 (which cannot be attained with the data under consideration).

In case Z is discrete, we need the distribution function of Z in order to compute the upper bound (3.4). Let us assume that the number of claims Y is predicted using CART (see Breiman et al. (1984) for more details). For the illustration, we fit several regression trees, controlling the size of the trees by the minimum number of observations required in the final nodes. The results are presented in Table 4.3. The corresponding regression trees and distribution functions of Z are displayed in Figures 4.2 to 4.8.

Min. nb of obs.	m	Kendall's tau	Upper bound	$\frac{\text{Kendall's tau}}{\text{Upper bound}}$
35 000	2	0.0167340	0.1645063	0.1017228
30 000/25 000	3	0.0227923	0.1695944	0.1343933
20 000	4	0.0257349	0.1893385	0.1359203
15 000	5	0.0281253	0.1941500	0.1448638
10 000	7	0.0287892	0.1941500	0.1482834
5000	14	0.0313762	0.2000066	0.1568756
1000	66	0.0355663	0.2050653	0.1734387

Table 4.3: Upper bound (3.4), Kendall's tau and its normalized version $\left(\frac{\text{Kendall's tau}}{\text{Upper bound}}\right)$ as well as the corresponding m (i.e. number of values taken by Z) for trees whose sizes are controlled by the minimum number of observations (Min. nb of obs.) required in the final nodes.

Considering the values displayed in Table 4.3, we notice that the upper bound (3.4) increases with m to ultimately tend to the upper bound (3.3) derived in the continuous case. Also, the values for Kendall's tau are rather small (around 2%), as it is generally the case with claim frequency data. However, they do not compare to 1 (that cannot be attained) but with the upper bounds ranging from 16% to 20% (depending on m). This may lead the

analyst to a different conclusion than the one deduced from its normalized version obtained by dividing Kendall's tau by its corresponding upper bound. Let us notice that both the upper bound and Kendall's tau become larger when m increases, which produces normalized Kendall's tau that penalizes in a way the model complexity. This is illustrated by the fact that Kendall's tau more than doubles from the simplest model ($m = 2$) to the most complex one ($m = 66$) while its normalized version only increases by 70.5%.

Figure 4.2: Regression tree (on the left) and distribution function of Z (on the right) when the minimum number of observations per final node is set to 35 000.

Figure 4.3: Regression tree (on the left) and distribution function of Z (on the right) when the minimum number of observations per final node is set to 30 000 (same results when minimum number of observations per node set to 25 000).

Figure 4.4: Regression tree (on the left) and distribution function of Z (on the right) when the minimum number of observations per final node is set to 20 000.

Acknowledgements

The authors would like to thank Frédéric Melle, Robin Van Oirbeek and Kelly Verheyen all from Belfius Insurance, as well as Gireg Willame from Detralytics, for interesting discussions about the use of concordance-based association measures in ratemaking studies. Mhamed Mesfioui acknowledges the financial support of the Natural Sciences and Engineering Research Council of Canada No 06536-2018.

References

- Breiman, L., Friedman, J., Olshen, R., Stone, C. (1984). Classification and Regression Trees, Wadsworth, New York.
- Denuit, M., Lambert, P. (2005). Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis* 93, 40-57.
- Denuit, M., Mesfioui, M., Trufin, J. (2018). Bounds on concordance-based validation statistics in regression models for binary responses. *Methodology and Computing in Applied Probability*, in press.
- 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.
- Neslehova, J. (2007). On rank correlation measures for non-continuous random variables. *Journal of Multivariate Analysis* 98, 544-567.

Figure 4.5: Regression tree (on the left) and cdf of Z (on the right) when minimum number of observations = 15 000.

Figure 4.6: Regression tree (on the left) and cdf of Z (on the right) when minimum number of observations = 10 000.

Figure 4.7: Regression tree (on the left) and cdf of Z (on the right) when minimum number of observations = 5000.

Figure 4.8: Regression tree (on the left) and cdf of Z (on the right) when minimum number of observations = 1000.