IN CORRELATION DO WE TRUST? SIMPSON'S PARADOX FOR KENDALL'S RANK COEFFICIENT

Michel Denuit, Julien Trufin, Pierre Zuyderhoff

LIDAM Discussion Paper ISBA
2024 / 26

ISBA | Louvain Institute of Data Analysis and Modeling in economics and statistics

ISBA
Voie du Roman Pays 20 - L1.04.01
B-1348 Louvain-la-Neuve
Email : lidam-library@uclouvain.be
https://uclouvain.be/en/research-institutes/lidam/isba/publication

In correlation do we trust? Simpson's paradox for Kendall's rank coefficient

Pierre Zuyderhoff
University of Nottingham
Ningbo, China

Michel Denuit
Institute of Statistics, Biostatistics and Actuarial Science
UCLouvain
Louvain-la-Neuve, Belgium

Julien Trufin [1]
Department of Mathematics
Université Libre de Bruxelles (ULB)
Brussels, Belgium

July 11, 2024

[1]Corresponding author, julien.trufin@ulb.be

Abstract

This note revisits Simpson's paradox in insurance and discusses confounding effects of hidden covariates on Kendall's tau. As a result, observed correlation may vanish or even revert. This phenomenon is shown to have important consequences in insurance risk assessment and ratemaking.

Keywords: Kendall's tau; covariates; Simpson's paradox.

1 Introduction and motivation

There are many applications in insurance where actuaries aim to assess the degree of dependency between two variables of interest. Examples include claim amounts (or claim frequencies) related to different guarantees or products owned by the same policyholder since actuaries started to move from marginal, policy-based analyses to joint, customer-based multiline risk assessment. See e.g Bermúdez and Karlis (2011, 2012) and Shi and Valdez (2014). The correlation structure among claim amounts related to policyholders belonging to the same household is another instance, since exploiting the correlation existing between the different products owned by the household appears to be useful for cash-flow projections, loyalty assessment, and the identification of cross-selling opportunities. See e.g. Pechon et al. (2018, 2019, 2021). In such circumstances, the dependence may only be apparent, produced by some hidden information impacting the variables under consideration. Dependencies can thus be partially explained by shared or correlated features. Features acting in opposite directions can also obscure the dependence existing between the variables of interest.

Another situation where dependency is of prime interest is certainly insurance ratemaking. Actuaries exploit the correlation structure between claims-related responses (frequency or severity) and features recorded in the database serving as potential risk factors. In a preliminary analysis, actuaries often consider the marginal impact of each candidate risk factor on the response of interest. The possible effect of the other features comprised in the database is thus disregarded. The aim at this stage is to make a first selection of potential risk factors, and to remove all the information seemingly unrelated with the response. This part of the analysis is often referred to as a one-way analysis. See e.g. Pechon et al. (2020). Multivariate methods (such as GLM/GAM regression approach) that adjust for correlations between features are then applied to a subset of the initial variables contained in the database. See e.g. Denuit et al. (2019) for an overview of risk classification based on GLM and GAM techniques. The actuary's task is much simplified when the features that do not play any significant role in explaining the insurance losses can be eliminated at an early stage, before starting the multivariate analysis. Again, one-way analyses are subject to confounding effects from omitted features and this may result in obscuring or even reverting the association. Actuaries must be aware of possible deficiencies accompanying this kind of analysis.

To detect an association between two variables, actuaries often use correlation measures. Usually in practice, the assessment of correlation between risk factors and loss responses is performed with the help of Cramer's V . This measure of association is based on contingency tables and assumes its values in the unit interval $[0, 1]$, with the extremities corresponding to independence and perfect dependence. Being based on data displayed in tabular form, Cramer's V is widely applicable, even to continuous variables after a preliminary banding. See e.g. Pechon et al. (2020). Kendall's tau is a classical measure of association between distributions also often used by actuaries. Recall that Kendall's tau for a random pair (Y, Z) is defined as

$$\tau[Y, Z] = \mathrm{P}[(Y_1 - Y_2)(Z_1 - Z_2) > 0] - \mathrm{P}[(Y_1 - Y_2)(Z_1 - Z_2) < 0], \tag{1.1}$$

where (Y_1, Z_1) and (Y_2, Z_2) are independent copies of the random pair (Y, Z). It turns out that Kendall's tau (as well as several other measures of association, including Goodman and Kruskal's gamma and Somers' D for instance) is based on concordance probabilities for pairs of observations. We refer the interested reader e.g. to Agresti (1996) for a general presentation of these association measures. Recall that in a concordant pair, both elements of one pair are either greater than, equal to, or less than the corresponding elements of the other pair. Concordance probabilities are also widely used to assess the quality of candidate premiums. See e.g. Ponnet et al. (2021) for a discussion.

This note aims to warn actuaries about the confusing effect of some features not included in the calculation of Kendall's tau. The phenomenon is closely related to the celebrated Simpson's paradox. The latter refers to situations where an association between two variables in a population emerges, disappears or reverses when the population is divided into sub-populations. For instance, two variables may be positively associated in a population, but be independent or even negatively associated in all sub-populations. A trend that is present when data are classified into groups may thus reverse or disappear when data are combined.

There are many examples of situations where this happens in insurance. We refer the interested reader to Stenmark and Wu (2004) for more examples. For instance, mortality may decrease for both smokers and non-smokers but increase overall, when both profiles are considered together. This is the case when smokers experience higher mortality and the proportion of smokers increases over time. Thus, mortality seems to increases overall whereas it exhibits decreasing trends for both smokers and non-smokers. As another example, consider underwriting rules. Assume that the company aims to offer more inclusive coverage and therefore accepts more applicants in both the standard and impaired life segments. Acceptance rate may however decrease if more applicants are classified as impaired over time (under the reasonable assumption that rejection rate is higher in the impaired life segment).

The remainder of the text is organized as follows. In Section 2, we establish a decomposition of Kendall's tau in presence of subroups. Section 3 proposes a formal definition of a Simpson's paradox for Kendall's tau inspired by Good and Mittal (1987) and provides necessary and sufficient conditions for a Simpson's paradox to occur. In Section 4, we give some simple examples of paradoxical situations in insurance. The final Section 5 concludes the paper.

2 Kendall's tau

Consider a variable $X \in \{0, 1, \dots, k\}$ with $k \in \mathbb{N}^*$ defining $k + 1$ groups among the observations. We denote $p_i = \mathrm{P}[X = i]$ for $i = 0, 1, \dots, k$ and (Y_1, Z_1, X_1) and (Y_2, Z_2, X_2) are independent copies of the random vector (Y, Z, X).

The concordance probability $\mathrm{P}[(Y_1 - Y_2)(Z_1 - Z_2) > 0]$ can then be expressed as

$$\begin{aligned} \mathrm{P}[(Y_1 - Y_2)(Z_1 - Z_2) > 0] &= \sum_{i,j=0}^{k} \mathrm{P}[(Y_1 - Y_2)(Z_1 - Z_2) > 0 | X_1 = i, X_2 = j]\, \mathrm{P}[X_1 = i, X_2 = j] \\ &= \sum_{i,j=0}^{k} \mathrm{P}[(Y_1 - Y_2)(Z_1 - Z_2) > 0 | X_1 = i, X_2 = j]\, p_i p_j. \end{aligned}$$

Similarly, the discordance probability $\mathrm{P}[(Y_1-Y_2)(Z_1-Z_2)<0]$ can be rewritten as

$$\mathrm{P}[(Y_1-Y_2)(Z_1-Z_2)<0]=\sum_{i,j=0}^{k}\mathrm{P}[(Y_1-Y_2)(Z_1-Z_2)<0|X_1=i,X_2=j]\,p_ip_j,$$

so that Kendall's tau in (1.1) can be expressed as

$$\tau[Y,Z]=\sum_{i,j=0}^{k}\tau_{ij}[Y,Z]\,p_ip_j, \tag{2.1}$$

where $\tau_{ij}[Y,Z]=\mathrm{P}[(Y_1-Y_2)(Z_1-Z_2)>0|X_1=i,X_2=j]-\mathrm{P}[(Y_1-Y_2)(Z_1-Z_2)<0|X_1=i,X_2=j]$, $i,j=0,1,\ldots,k$. Since for all $i,j=0,1,\ldots,k$, we have $\mathrm{P}[(Y_1-Y_2)(Z_1-Z_2)>0|X_1=i,X_2=j]=\mathrm{P}[(Y_1-Y_2)(Z_1-Z_2)>0|X_1=j,X_2=i]$ by symmetry, expression (2.1) can be rewritten as

$$\tau[Y,Z]=\sum_{i=0}^{k}\tau_{ii}[Y,Z]\,p_i^2+2\sum_{i<j}\tau_{ij}[Y,Z]\,p_ip_j. \tag{2.2}$$

This latter expression shows that Kendall's tau $\tau[Y,Z]$ can be decomposed as the weighted sum of Kendall's tau $\tau_{ii}[Y,Z]$ within each group, namely $\sum_{i=0}^{k}\tau_{ii}[Y,Z]\,p_i^2$, plus the weighted sum of Kendall's tau $\tau_{ij}[Y,Z]$ between groups ($i\neq j$), namely $2\sum_{i<j}\tau_{ij}[Y,Z]\,p_ip_j$.

Expression (2.1) can be rewritten as

$$\tau[Y,Z]=\mathbf{p}^t\mathcal{T}[Y,Z]\mathbf{p},$$

where

$$\mathbf{p}^t=\begin{pmatrix} p_0 & p_1 & \cdots & p_k\end{pmatrix}$$

is a vector containing the probabilities p_i, $i=0,1,\ldots,k$, and

$$\mathcal{T}[Y,Z]=\begin{pmatrix}\tau_{00}[Y,Z] & \tau_{01}[Y,Z] & \cdots & \tau_{0k}[Y,Z]\\ \tau_{10}[Y,Z] & \tau_{11}[Y,Z] & \cdots & \tau_{1k}[Y,Z]\\ \vdots & \vdots & \ddots & \vdots\\ \tau_{k0}[Y,Z] & \tau_{k1}[Y,Z] & \cdots & \tau_{kk}[Y,Z]\end{pmatrix}$$

is a matrix containing the Kendall's tau $\tau_{ij}[Y,Z]$, $i,j=0,1,\ldots,k$. The decomposition of Kendall's tau $\tau[Y,Z]$ in (2.2) can then be reexpressed as

$$\tau[Y,Z]=\mathbf{p}^t\mathcal{T}_1[Y,Z]\mathbf{p}+\mathbf{p}^t\mathcal{T}_2[Y,Z]\mathbf{p},$$

where

$$\mathcal{T}_1[Y,Z]=\begin{pmatrix}\tau_{00}[Y,Z] & \frac{\tau_{00}[Y,Z]+\tau_{11}[Y,Z]}{2} & \cdots & \frac{\tau_{00}[Y,Z]+\tau_{kk}[Y,Z]}{2}\\ \frac{\tau_{11}[Y,Z]+\tau_{00}[Y,Z]}{2} & \tau_{11}[Y,Z] & \cdots & \frac{\tau_{11}[Y,Z]+\tau_{kk}[Y,Z]}{2}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\tau_{kk}[Y,Z]+\tau_{00}[Y,Z]}{2} & \frac{\tau_{kk}[Y,Z]+\tau_{11}[Y,Z]}{2} & \cdots & \tau_{kk}[Y,Z]\end{pmatrix}$$

and

$$\mathcal{T}_2[Y,Z]=\begin{pmatrix}0 & \tau_{01}[Y,Z]-\frac{\tau_{00}[Y,Z]+\tau_{11}[Y,Z]}{2} & \cdots & \tau_{0k}[Y,Z]-\frac{\tau_{00}[Y,Z]+\tau_{kk}[Y,Z]}{2}\\ \tau_{10}[Y,Z]-\frac{\tau_{11}[Y,Z]+\tau_{00}[Y,Z]}{2} & 0 & \cdots & \tau_{1k}[Y,Z]-\frac{\tau_{11}[Y,Z]+\tau_{kk}[Y,Z]}{2}\\ \vdots & \vdots & \ddots & \vdots\\ \tau_{k0}[Y,Z]-\frac{\tau_{kk}[Y,Z]+\tau_{00}[Y,Z]}{2} & \tau_{k1}[Y,Z]-\frac{\tau_{kk}[Y,Z]+\tau_{11}[Y,Z]}{2} & \cdots & 0\end{pmatrix}.$$

3 Simpson's paradox for Kendall's tau

Simpson's Paradox is the phenomenon that appears in some data sets, where groups with a common trend show the reverse trend when they are aggregated. We refer the interested reader to Selvitella (2017) and the references therein for examples of Simpson's paradox in different research fields and to Stenmark and Wu (2004) for examples in insurance.

3.1 Definition

In their Definition 1.1, Good and Mittal (1987) provide an interesting manner to define a Simpson's paradox (also called amalgamation paradox) for two-by-two contingency tables. This note is concerned with a paradox for Kendall's tau defined in the following manner, largely inspired by Good and Mittal (1987).

Definition 3.1. We say that a Simpson's paradox for Kendall's tau occurs if

$$\max_{0\le i\le k} \tau_{ii}[Y,Z] < \tau[Y,Z] \qquad \text{or} \qquad \tau[Y,Z] < \min_{0\le i\le k} \tau_{ii}[Y,Z].$$

In particular, this definition indeed implies that a Simpson's paradox occurs if $\tau_{ii}[Y,Z] < 0$ for all i and $\tau[Y,Z] > 0$ or if $\tau_{ii}[Y,Z] > 0$ for all i and $\tau[Y,Z] < 0$, which correspond to situations where groups with a common trend (negative (resp. positive) Kendall's tau $\tau_{ii}[Y,Z]$ for all i) show the reverse trend when they are aggregated (positive (resp. negative) Kendall's tau $\tau[Y,Z]$).

3.2 Necessary condition

The next proposition provides a necessary condition for a Simpson's paradox to occur.

Proposition 3.2. *If a Simpson's paradox occurs, then*

$$\max_{0\le i\le k} \tau_{ii}[Y,Z] < \frac{\sum_{i<j} p_i p_j \tau_{ij}[Y,Z]}{\sum_{i<j} p_i p_j} \qquad \textit{or} \qquad \frac{\sum_{i<j} p_i p_j \tau_{ij}[Y,Z]}{\sum_{i<j} p_i p_j} < \min_{0\le i\le k} \tau_{ii}[Y,Z]. \tag{3.1}$$

Proof. It suffices to notice that expression (2.2) can be rewritten as

$$\tau[Y,Z] = \sum_{i=0}^{k} p_i^2 \tau_{ii}[Y,Z] + \left(1 - \sum_{i=0}^{k} p_i^2\right) \frac{\sum_{i<j} p_i p_j \tau_{ij}[Y,Z]}{\sum_{i<j} p_i p_j}.$$

□

Notice that condition (3.1) is also a sufficient condition for a Simpson's paradox to occur when $k = 1$, as shown below in Proposition 3.3.

3.3 Sufficient conditions

In order to find sufficient conditions for a Simpson's paradox to occur or for its absence, let us consider $\tau[Y,Z]$ as a function ϕ of $\mathbf{p}$, namely

$$\phi\colon \Delta_{k+1} \to [-1,1], \mathbf{p} \mapsto \phi(\mathbf{p}) = \mathbf{p}^t \mathcal{T}[Y,Z]\mathbf{p},$$

where $\Delta_{k+1} = \{\mathbf{p} \in (\mathbb{R}_0^+)^{k+1} | \sum_{i=0}^k p_i = 1\}$ is the probability simplex. We have

$$\phi(\mathbf{e}_i) = \tau_{ii}[Y,Z] \quad \text{for all } i \in \{0,1,...,k\},$$

where $\mathbf{e}_i^t$ is the $(i+1)$th unit vector $(0,...,0,1,0,...,0)$. The function $\phi(\mathbf{p})$ attains both its maximum and minimum on its domain Δ_{k+1}, which is compact. Therefore,

$$\max_{\mathbf{p}\in\Delta_{k+1}} \phi(\mathbf{p}) \geq \max_i \tau_{ii}[Y,Z]$$

and

$$\min_{\mathbf{p}\in\Delta_{k+1}} \phi(\mathbf{p}) \leq \min_i \tau_{ii}[Y,Z].$$

If the inequality is strict in one of the two latter inequalities, then there exists $\mathbf{p} \in \Delta_{k+1}$ such that a Simpson's paradox occurs for $\tau[Y,Z] = \phi(\mathbf{p})$. Notice that the problem of finding an optimum value for $\phi(\mathbf{p})$ on Δ_{k+1} is a quadratic programming problem with standard simplex constraints. There are efficient algorithms to compute such an optimum (see e.g. Lawrence and Tits (2001)).

The case $k=1$ is straightforward and leads to a necessary and sufficient condition for a Simpson's paradox to occur, as shown in the next proposition.

Proposition 3.3. *The three following cases are mutually exclusive for $k=1$:*

1) $\tau_{01}[Y,Z] > \max_i \tau_{ii}[Y,Z] \Leftrightarrow \max_{\mathbf{p}\in\Delta_2} \phi(\mathbf{p}) > \max_i \tau_{ii}[Y,Z]$. *Moreover,*

$$\max_{\mathbf{p}\in\Delta_2} \phi(\mathbf{p}) = \phi(\mathbf{p}^*) = \frac{\tau_{01}^2[Y,Z] - \tau_{00}[Y,Z]\tau_{11}[Y,Z]}{2\tau_{01}[Y,Z] - \tau_{00}[Y,Z] - \tau_{11}[Y,Z]},$$

where

$$\mathbf{p}^* = \left(\frac{\tau_{11}[Y,Z] - \tau_{01}[Y,Z]}{\tau_{00}[Y,Z] + \tau_{11}[Y,Z] - 2\tau_{01}[Y,Z]}, \frac{\tau_{00}[Y,Z] - \tau_{01}[Y,Z]}{\tau_{00}[Y,Z] + \tau_{11}[Y,Z] - 2\tau_{01}[Y,Z]}\right).$$

2) $\tau_{01}[Y,Z] < \min_i \tau_{ii}[Y,Z] \Leftrightarrow \min_{\mathbf{p}\in\Delta_2} \phi(\mathbf{p}) < \min_i \tau_{ii}[Y,Z]$. *Moreover,*

$$\min_{\mathbf{p}\in\Delta_2} \phi(\mathbf{p}) = \phi(\mathbf{p}^*).$$

3) $\min_i \tau_{ii}[Y,Z] \leq \tau_{01}[Y,Z] \leq \max_i \tau_{ii}[Y,Z] \Leftrightarrow \min_i \tau_{ii}[Y,Z] \leq \phi(\mathbf{p}) \leq \max_i \tau_{ii}[Y,Z]$, *for all* $\mathbf{p} \in \Delta_2$, *and there is no* $\mathbf{p} \in \Delta_2$ *such that a Simpson's paradox occurs for* $\tau[Y,Z] = \phi(\mathbf{p})$.

Proof. We have

$$\begin{aligned}
\phi(\mathbf{p}) &= \phi(1-p_1, p_1) \\
&= (1-p_1)^2\tau_{00}[Y,Z] + p_1^2\tau_{11}[Y,Z] + 2p_1(1-p_1)\tau_{01}[Y,Z] \\
&= p_1^2(\tau_{00}[Y,Z] + \tau_{11}[Y,Z] - 2\tau_{01}[Y,Z]) + p_1(-2\tau_{00}[Y,Z] + 2\tau_{01}[Y,Z]) + \tau_{00}[Y,Z].
\end{aligned}$$

So, ϕ attains a maximum on the interior of the set Δ_2 if and only if

$$0 < \frac{\tau_{00}[Y,Z] - \tau_{01}[Y,Z]}{\tau_{00}[Y,Z] + \tau_{11}[Y,Z] - 2\tau_{01}[Y,Z]} < 1$$

and

$$\tau_{00}[Y,Z] + \tau_{11}[Y,Z] - 2\tau_{01}[Y,Z] < 0.$$

It is easy to show that the two last conditions are equivalent to

$$\tau_{01}[Y,Z] > \max_i \tau_{ii}[Y,Z],$$

which is Case 1. Case 2 is obtained by symmetry and Case 3 follows from Cases 1 and 2. □

For the cases where $k > 1$, we get the following result.

Proposition 3.4. *If*

$$\min_i \tau_{ii}[Y,Z] \leq \tau_{jj'}[Y,Z] \leq \max_i \tau_{ii}[Y,Z] \quad \textit{for all } 0 \leq j, j' \leq k,$$

then

$$\min_i \tau_{ii}[Y,Z] \leq \phi(\mathbf{p}) \leq \max_i \tau_{ii}[Y,Z] \quad \textit{for all } \mathbf{p} \in \Delta_{k+1},$$

and there is no $\mathbf{p} \in \Delta_{k+1}$ *such that a Simpson's paradox occurs for* $\tau[Y,Z] = \phi(\mathbf{p})$.

Proof. The statement is a weaker version of the contrapositive of Proposition 3.2. □

The converse of Proposition 3.4 does not hold. However, we have the following weaker result.

Corollary 3.5. *Let* $j \in \{0,1,...,k\}$ *be such that* $\tau_{jj}[Y,Z] = \max_i \tau_{ii}[Y,Z]$. *If there exists* $i \in \{0,1,...,k\}$ *such that the inequality* $\tau_{jj}[Y,Z] < \tau_{ij}[Y,Z]$ *(resp.* $\tau_{jj}[Y,Z] > \tau_{ij}[Y,Z]$*) holds true, then*

$$\max_{\mathbf{p}\in\Delta_{k+1}} \phi(\mathbf{p}) > \max_i \tau_{ii}[Y,Z], \qquad (\textit{resp. } \min_{\mathbf{p}\in\Delta_{k+1}} \phi(\mathbf{p}) < \min_i \tau_{ii}[Y,Z],)$$

and there exists $\mathbf{p} \in \Delta_{k+1}$ *such that a Simpson's paradox occurs for* $\tau[Y,Z] = \phi(\mathbf{p})$.

Proof. The result follows directly from Proposition 3.3. □

3.4 Particular cases

3.4.1 Case 1

Let the matrix $\mathcal{T}[Y,Z]$ be such that the off-diagonal elements are constant, namely

$$\tau_{ij}[Y,Z] = a \quad \text{for } 0 \le i < j \le k,$$

with $-1 \le a \le 1$. Moreover, let us assume that $\tau_{ii}[Y,Z] \neq a$ for $i \in \{0,1,...,k\}$. Then, we have

$$\begin{aligned}\phi(\mathbf{p}) &= \sum_{i=0}^{k} \tau_{ii}[Y,Z]p_i^2 + a\left(1 - \sum_{i=0}^{k} p_i^2\right)\\ &= a + \sum_{i=0}^{k} p_i^2(\tau_{ii}[Y,Z] - a).\end{aligned}$$

To find the optimums for ϕ on Δ_{k+1}, we consider the Lagrangian

$$\mathcal{L}(\mathbf{p},\lambda) = a + \sum_{i=0}^{k} p_i^2(\tau_{ii}[Y,Z] - a) - \lambda\left(\sum_{i=0}^{k} p_i - 1\right).$$

The unique stationary point $(\mathbf{p}^*,\lambda^*)$ of $\mathcal{L}$ on $\mathbb{R}^{k+1}\times\mathbb{R}$ verifies $\sum_{i=0}^{k} p_i = 1$ and

$$p_i^*(\tau_{ii}[Y,Z] - a) = p_j^*(\tau_{jj}[Y,Z] - a) \quad \text{for } i,j \in \{0,1,...,k\},$$

which implies

$$\mathbf{p}^* = \left(\sum_{i=0}^{k} \frac{1}{\tau_{ii}[Y,Z] - a}\right)^{-1}\left(\frac{1}{\tau_{00}[Y,Z] - a}, \frac{1}{\tau_{11}[Y,Z] - a}, ..., \frac{1}{\tau_{kk}[Y,Z] - a}\right).$$

Therefore, ϕ has a maximum (resp. a minimum) point $\mathbf{p}^*$ on the interior of its domain Δ_{k+1} if and only if $\max_i \tau_{ii}[Y,Z] < a$ (resp. $\min_i \tau_{ii}[Y,Z] > a$) for $i \in \{0,1,...,k\}$. In that case, there exists $\mathbf{p} \in \Delta_{k+1}$ such that a Simpson's paradox occurs for $\tau[Y,Z] = \phi(\mathbf{p})$, which is in line with Corollary 3.5.

Remark 3.6. In particular, if $\tau_{ii}[Y,Z] = b$ for all $0 \le i \le k$ with $-1 \le b \le 1$ and $a \neq b$, then

$$\mathbf{p}^* = \left(\frac{1}{k+1}, \frac{1}{k+1}, ..., \frac{1}{k+1}\right),$$

and a Simpson's paradox occurs for all $\mathbf{p}$ in the interior of Δ_{k+1}. Note that the optimum value for $\tau[Y,Z]$ is then given by

$$\phi(\mathbf{p}^*) = \frac{ka+b}{k+1}.$$

3.4.2 Case 2

Let us consider the case where the marginal distributions of the random vectors $[(Y,Z)|X = i]$, $i = 0,\ldots,k$, do not depend on i. In this setting, Proposition 3.8 below provides conditions for the absence of a Simpson's paradox. The particular case where X is independent of (Y,Z) is not interesting since we obviously have $\tau[Y,Z] = \tau_{ij}[Y,Z]$ for all $i,j = 0,1,\ldots,k$.

Recall from Denuit et al. (2005) that two random variables are said to be concordant if they tend to be both large together or small together. The concordance order expresses the idea that large and small values tend to be more often associated under the distribution that dominates the other one. Precisely, let us consider two random couples (Y_1,Z_1) and (Y_2,Z_2) with the same marginal distributions, i.e. $\mathrm{P}[Y_1 \le y] = \mathrm{P}[Y_2 \le y]$ for all y and $\mathrm{P}[Z_1 \le z] = \mathrm{P}[Z_2 \le z]$ for all z. If

$$\mathrm{P}[Y_1 \le y, Z_1 \le z] \le \mathrm{P}[Y_2 \le y, Z_2 \le z] \text{ for all } y \text{ and } z,$$

then (Y_1,Z_1) is said to be less concordant than (Y_2,Z_2). This is henceforth denoted as $(Y_1,Z_1) \preceq_{\text{corr}} (Y_2,Z_2)$.

The next result is useful to derive Proposition 3.8.

Lemma 3.7. *Let $i,j \in \{0,\cdots,k\}$ such that $i \neq j$. If $[(Y,Z)|X = i] \preceq_{corr} [(Y,Z)|X = j]$, then*

$$\tau_{ii}[Y,Z] \le \tau_{ij}[Y,Z] \le \tau_{jj}[Y,Z].$$

Proof. Assume that

$$F_{(Y,Z)|X=i}(y,z) \le F_{(Y,Z)|X=j}(y,z) \quad \text{for all } y,z \in \mathbb{R},$$

where $F_{(Y,Z)|X=i}(y,z) = P[Y \le y, Z \le z | X = i]$. Therefore, for all $(p_i,p_j) \in \Delta_2$,

$$F_{(Y,Z)|X=i}(y,z) \le p_i F_{(Y,Z)|X=i}(y,z) + p_j F_{(Y,Z)|X=j}(y,z) \le F_{(Y,Z)|X=j}(y,z) \quad \text{for all } y,z \in \mathbb{R}.$$

Hence, by the consistency of Kendall's tau with the concordance order (see e.g. Denuit et al., 2005), we have

$$\tau_{ii}[Y,Z] \le \phi(p_i\mathbf{e_i} + p_j\mathbf{e_j}) \le \tau_{jj}[Y,Z] \quad \text{for all } (p_i,p_j) \in \Delta_2,$$

and the result follows from Proposition 3.3. □

We are now in position to prove the next result.

Proposition 3.8. *If*

$$[(Y,Z)|X = 0] \preceq_{corr} [(Y,Z)|X = i] \preceq_{corr} [(Y,Z)|X = k] \quad \textit{for all } i \in \{1,...,k-1\},$$

then

$$\tau_{00}[Y,Z] = \min_i \tau_{ii}[Y,Z] \le \phi(\mathbf{p}) \le \max_i \tau_{ii}[Y,Z] = \tau_{kk}[Y,Z] \quad \textit{for all } \mathbf{p} \in \Delta_{k+1},$$

and there is no $\mathbf{p} \in \Delta_{k+1}$ such that a Simpson's paradox occurs for $\tau[Y,Z] = \phi(\mathbf{p})$.

Proof. We have

$$F_{(Y,Z)|X=0}(y,z) \leq F_{(Y,Z)|X=i}(y,z) \leq F_{(Y,Z)|X=k}(y,z) \quad \text{for all } y,z \in \mathbb{R} \text{ and } i \in \{1,...,k-1\}.$$

Then, for all $\mathbf{p} \in \Delta_{k+1}$, we have

$$F_{(Y,Z)|X=0}(y,z) \leq F_{(Y,Z)}(y,z) = \sum_{i=0}^{k} p_i F_{(Y,Z)|X=i}(y,z) \leq F_{(Y,Z)|X=k}(y,z) \quad \text{for all } y,z \in \mathbb{R}.$$

Hence

$$[(Y,Z)|X=0] \preceq_{\text{corr}} (Y,Z) = \sum_{i=0}^{k} p_i[(Y,Z)|X=i] \preceq_{\text{corr}} [(Y,Z)|X=k] \quad \text{for all } \mathbf{p} \in \Delta_{k+1},$$

and the absence of a Simpson's paradox for $\tau[Y,Z]$ follows from the consistency of Kendall's tau with the concordance order. □

4 Examples of Simpson's paradox in insurance

4.1 Example 1

This first example illustrates the confusing effect of a feature not included in the calculation of Kendall's tau. Consider that Y corresponds to claim severity (assumed to be absolutely continuous, restricting the analysis to claimants), Z corresponds to the power of the car and X corresponds to the insured's gender ($X = 0$ for males and $X = 1$ for females). It is well documented that Y and Z are positively correlated, so that we consider $\tau[Y,Z] > 0$, which is equivalent to considering

$$\tau_{01}[Y,Z] > -\frac{1}{2p_0p_1}\left(\tau_{00}[Y,Z]p_0^2 + \tau_{11}[Y,Z]p_1^2\right) \tag{4.1}$$

from equation (2.2). In this setting, we know from Proposition 3.3 that a Simpson's paradox occurs if and only if

$$\tau_{01}[Y,Z] > \max(\tau_{00}[Y,Z], \tau_{11}[Y,Z]) \qquad \text{or} \qquad \tau_{01}[Y,Z] < \min(\tau_{00}[Y,Z], \tau_{11}[Y,Z]). \tag{4.2}$$

Quite unexpectedly, inequalities (4.1) and (4.2) are both possible even if Y and Z are negatively associated for both male and female drivers. Indeed, even with $\tau_{00}[Y,Z] < 0$ and $\tau_{11}[Y,Z] < 0$, there are situations where both inequalities (4.1) and (4.2) can be fulfilled since $\tau_{01}[Y,Z]$ can be made arbitrarily large (up to 1, in case for instance males always file more expensive claims and drive more powerful cars than females, which would lead to $\mathrm{P}[Z_1 - Z_2 > 0|X_1 = 0, X_2 = 1] = 1$ and $\mathrm{P}[Y_1 - Y_2 > 0|X_1 = 0, X_2 = 1] = 1$ and hence to $\tau_{01}[Y,Z] = 1$).

4.2 Example 2

As already mentioned in the introduction of the paper, multivariate mixture models are used in actuarial science to account for dependencies between claim amounts of different guarantees owned by the same policyholder. This second example illustrates the use of Kendall's tau in a simple situation where the mixture variable X is bivariate (for example an hidden binary feature) and where the variables of interest Y and Z are occurrences of claims for two guarantees owned by the same policyholder. We assume that the random variables Y and Z are both Bernoulli distributed, such that

$$\mathrm{P}[Y=1] = q(X) = 1 - \mathrm{P}[Y=0] \text{ and } \mathrm{P}[Z=1] = h(X) = 1 - \mathrm{P}[Z=0]$$

with

$$q(X) = \begin{cases} q_0 & \text{if } X=0 \\ q_1 & \text{if } X=1 \end{cases},$$

and

$$h(X) = \begin{cases} h_0 & \text{if } X=0 \\ h_1 & \text{if } X=1 \end{cases},$$

so that conditionally to X, the random variables Y and Z are independent. We thus have

$$\tau_{00}[Y,Z] = \tau_{11}[Y,Z] = 0.$$

Hence, from expression (2.2), we get

$$\tau[Y,Z] = 2\tau_{01}[Y,Z]\, p_0 p_1.$$

In this setting, Proposition 3.3 tells us that a Simpson's paradox occurs if and only if $\tau_{01}[Y,Z] \neq 0$, which basically means here that $\tau[Y,Z] \neq 0$. Now, since

$$\begin{aligned}
&\mathrm{P}[(Y_1-Y_2)(Z_1-Z_2)>0|X_1=0,X_2=1] \\
&= \mathrm{P}[Y_1>Y_2, Z_1>Z_2|X_1=0,X_2=1] + \mathrm{P}[Y_1<Y_2, Z_1<Z_2|X_1=0,X_2=1] \\
&= \mathrm{P}[Y_1=1, Y_2=0, Z_1=1, Z_2=0|X_1=0,X_2=1] \\
&\quad + \mathrm{P}[Y_1=0, Y_2=1, Z_1=0, Z_2=1|X_1=0,X_2=1] \\
&= \mathrm{P}[Y_1=1, Z_1=1|X_1=0]\,\mathrm{P}[Y_2=0, Z_2=0|X_2=1] \\
&\quad + \mathrm{P}[Y_1=0, Z_1=0|X_1=0]\,\mathrm{P}[Y_2=1, Z_2=1|X_2=1] \\
&= \mathrm{P}[Y_1=1|X_1=0]\,\mathrm{P}[Z_1=1|X_1=0]\,\mathrm{P}[Y_2=0|X_2=1]\,\mathrm{P}[Z_2=0|X_2=1] \\
&\quad + \mathrm{P}[Y_1=0|X_1=0]\,\mathrm{P}[Z_1=0|X_1=0]\,\mathrm{P}[Y_2=1|X_2=1]\,\mathrm{P}[Z_2=1|X_2=1] \\
&= q_0 h_0 (1-q_1)(1-h_1) + (1-q_0)(1-h_0) q_1 h_1
\end{aligned}$$

and similarly

$$\mathrm{P}[(Y_1-Y_2)(Z_1-Z_2)<0|X_1=0,X_2=1] = q_0(1-h_0)(1-q_1)h_1 + (1-q_0)h_0 q_1 (1-h_1),$$

we get

$$\tau_{01}[Y,Z] = (q_0-q_1)(h_0-h_1). \tag{4.3}$$

Hence, expression (4.3) shows that despite τ_{00} and τ_{11} are both zero, this is not necessarily the case for τ_{01} and hence for Kendall's tau τ. When $q_0 = q_1$ or when $h_0 = h_1$, we get $\tau[Y, Z] = 0$ (Y and Z are independent) and there is no Simpson's paradox. When $q_0 < q_1$ and $h_0 < h_1$ or when $q_0 > q_1$ and $h_0 > h_1$, we get $\tau[Y, Z] > 0$ (Y and Z are positively correlated) and a Simpson's paradox occurs. Finally, when $q_0 < q_1$ and $h_0 > h_1$ or when $q_0 > q_1$ and $h_0 < h_1$, we get $\tau[Y, Z] < 0$ (Y and Z are negatively correlated) and a Simpson's paradox occurs.

5 Discussion

This short note discusses Kendall's tau in the vein of the celebrated Simpson's paradox. Coming back to the title of this note, paraphrasing the official motto of the United States printed in every dollar currency, actuaries are warned to be very cautious when interpreting correlations since the vast majority of insurance studies are purely observational by nature and there are generally many hidden features correlated with those recorded in the database. This may obscure or even revert association between claim-related responses and candidate risk factors.

Competing Interest

The authors declare that they have no competing interests or other interests that might be perceived to influence the results and/or discussion reported in this paper.

Availability of Data and Materials

Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Agresti, A. (1996). An Introduction to Categorical Data Analysis. Wiley, New York.

Bermúdez, L., Karlis, D. (2011). Bayesian multivariate poisson models for insurance ratemaking. Insurance: Mathematics and Economics 48(2), 226-236.

Bermúdez, L., Karlis, D. (2012). A finite mixture of bivariate poisson regression models with an application to insurance ratemaking. Computational Statistics and Data Analysis 56(12), 3988-3999.

Denuit, M., Dhaene, J., Goovaerts, M.J., Kaas, R. (2005). Actuarial Theory for Dependent Risks: Measures, Orders and Models. Wiley, New York.

Denuit, M., Hainaut, D., Trufin, J. (2019). Effective Statistical Learning Methods for Actuaries I: GLMs and Extensions. Springer Actuarial Series.

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 21, 491-509.

Good, I. J., Mittal, Y. (1987). The amalgamation and geometry of two-by-two contingency tables. The Annals of Statistics, 694-711.

Lawrence, C.T., Tits, A.L. (2001). A computationally efficient feasible sequential quadratic programming algorithm. SIAM Journal on optimization, 11(4), 1092-1118.

Pechon, F., Denuit, M., Trufin, J. (2018). Multivariate modelling of household claim frequencies in motor third-party liability insurance. ASTIN Bulletin 48, 969-993.

Pechon, F., Denuit, M., Trufin, J. (2019). Multivariate modelling of multiple guarantees in motor insurance of a household. European Actuarial Journal 9, 575-602.

Pechon, F., Denuit, M., Trufin, J. (2021). Home and Motor insurance joined at a household level using multivariate credibility. Annals of Actuarial Science 15, 82-114.

Pechon, F., Trufin, J., Denuit, M. (2020). Preliminary selection of risk factors in P&C ratemaking. Variance 13, 124-140.

Ponnet, J., Van Oirbeek, R., Verdonck, T. (2021). Concordance Probability for Insurance Pricing Models. Risks 9, paper # 178.

Selvitella, A. (2017). The ubiquity of the Simpson's Paradox. Journal of Statistical Distributions and Applications 4, paper # 2.

Shi, P., Valdez, E.A. (2014). Multivariate negative binomial models for insurance claim counts. Insurance: Mathematics and Economics 55(1), 18-29.

Stenmark, J. A., Wu, C. S. P. (2004). Simpsons paradox, confounding variables and insurance ratemaking. Proceedings of the Casuality Actuarial Society Annual Meeting, Montreal, Canada.