Catholic University of Louvain

Identification of the reliability test (R^2) in case of an unknown common cause.

Submitted in partial fulfillment of the requirements for the degree of Master in Economics, Econometrics, Research focus (ETRI2MA)

Promoter: Sebastien Van Bellegem

Co-promoter: Ernesto San Martin Gutiérrez

Reader: Vincenzo Verardi

Cyril Ghislain

June, 2023

ACKNOWLEDGEMENTS

I could not thank enough Pr. Sebastien Van Bellegem for his patience, availability and every interesting discussion this work lead to. Not only did he teach me how to structure a mathematical proof, but he also gave necessary directions for the demonstrations.

I would also like to acknowledge the participation of Pr. Ernesto San Martin, who's been a great help understanding Spearman's intentions thanks to his previous work on the matter and was also of great help for the demonstrations.

Finally, thank you, Pr. Vincenzo Verardi for taking the time to read this work, which I hope will arouse your interest.

TABLE OF CONTENTS

LIST OF SYMBOLS

α, β, λ Estimator of X, Y and Z (in that order)

$\mathbb{E}(.)$ Expectation (weighted mean)

$\langle ., . \rangle$ Inner product

$\mathbb{R}$ Real numbers

ρ_{uv} Correlation between variables u and v

σ_θ^2 Variance of θ

$\theta \sim F(\mathbb{E}(\theta), \sigma_\theta^2)$ Distribution of θ whereas $F(.,.)$ gives two informations: the weighted mean $\mathbb{E}(\theta)$ and the variance σ_θ^2

L^2 Lebesgue square integrable space

Abstract

This work aims to investigate the identification of the reliability test in situations where there is a non-measured common cause. A common cause is one or more variables, say θ, that is responsible for the correlation of other variables, say X, Y and Z. In this work, we shall consider an univariate common cause, that fully explains the correlation between X, Y and Z, in other words, whose partial correlation is equal to zero, once the influence of θ is removed. The study considers three different scenarios, each resting on the base assumption that the variables have finite second moments: the first one assumes a linear model for the conditional expectations, the second one assumes a normal distribution of the random variables, and the third one demonstrates that the partial correlation formula can not work on non linear cases and leaves a trail to a new coefficient. The Projection Theorem in a Hilbert space (Rudin [1987]) and Spearman's partial correlation coefficient from his work "General Intelligence" (Spearman [1904]) are utilized as primary tools.

Chapter 1

INTRODUCTION

This thesis is structured into five chapters. The first chapter provides an introduction to the theoretical concepts being used, along with some history of Spearman's partial correlation and a confrontation of his ideas with a multifactoral take on intelligence (Thurstone and Thurstone [1914]), but also a few examples in order to better place the problem addressed in this work in a more concrete way. The following chapters focus on the three scenarios presented earlier, and the last one is focused on a multivariate case and the questions left to be answered.

The results of this study are new and intended to provide researchers with a more versatile tool to carefully measure the reliability of their models while avoiding misleading conclusions. Overall, this contributes to the broader understanding of causal inference and statistical analysis.

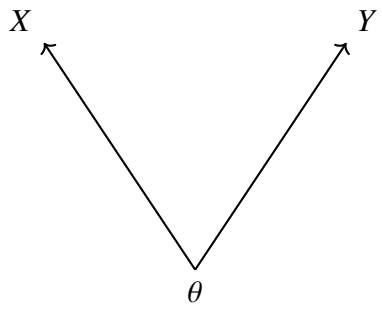
1.1 Common cause

In statistics, a common cause is one, or more, variable(s) that can influence or affect two or more other variables that are being studied. When two variables are found to be associated with each other, this work is helpful to find out whether there is a direct causal relationship between them or whether a common cause is responsible for the observed association.

For example, suppose we observe a positive correlation between the outcomes of a perfectly balanced roulette. It would be wrong to conclude that the outcomes affect each others. Instead, a common cause would be responsible for the outcomes, Walter A. Shewhart calls it the "Chance Cause" (Shewhart [1931]). One could argue, that chance is nothing but contingency, every single variable having so small of an impact, that we could not, would not, measure it.

In his work, Reichenbach [1956] writes as an example of common cause: "Suppose both lamps in a room go out suddenly. We regard it as improbable that, by chance, both bulbs burned out at the same time, and look for a burned-out fuse or some other interruption of the common power supply. The improbable coincidence is thus explained as the product of a common cause."

Reichenbach also introduces the following fork diagram, representing the case of three correlated variables, from which two of them are correlated by a common cause:



Representation of a common cause adapted from Reichenbach [1956]

In summary, a common cause is a third variable that is associated with both other variables, leading to a spurious association between the two. It is important to consider and control for common causes when analyzing data to draw accurate conclusions about causal relationships.

In a more formal language, the relationship between the variables can be denoted as $(X \perp Y \perp Z)|\theta$, meaning that X,Y and Z are independent conditionally to θ, defining θ as the common space between the random variables, we call this the Weak Axiom of Local Independence.

In his work on intelligence, Spearman [1904] defines it as an unifactorial common cause of partially independent tests, which lead him to construct the "partial correlation coefficient".

1.1.1 Correlation and causality

The term "common cause" is not to be confused with "common factor". To clarify the difference between the two, let us define causality and correlation:

Definition 1. *Consider two vectors $A, B \in \mathbb{R}^2$, their correlation is defined as such:*

$$\mathbb{C}orr(A, B) = \cos(A - \mathbb{E}(A), B - \mathbb{E}(B)). \tag{1.1}$$

$\mathbb{C}orr(A, B) \in [0, 1]$.

Therefore, correlation is a bilateral relationship between two variables, which is positive (resp. negative) if both evolve in a similar (resp. opposite) direction. The correlation is equal to zero if $A \perp B$, we then say that the two variables are independent. A common factor, would be a variable correlated with two other variables.

Causality is a more literal concept, it is a cause-and-effect relationship between variables that cannot be measured mathematically without a bit of intuition. The main difference with correlation is that causality is an unilateral relationship between variables. Here, we shall refer as "common cause" to the phenomena where the correlation between two or more variables can only be explained by another variable, which is therefore the "cause" of their correlation, whereas none of them is a cause of the other.

In other words, a common factor refers to a shared variable or characteristic that contributes to the correlation or similarity between observations, while a common cause refers to a source or event that simultaneously impacts several variables or observations without necessarily creating a direct dependency link between them.

1.2 History

Charles Spearman was a British psychologist who was born in London in 1863 and died in 1945. He is best known for his work on intelligence, particularly his

development of the theory of general intelligence.

In 1904, Spearman published a paper titled "General Intelligence, Objectively Determined and Measured." In this paper, he proposes the concept of a general factor of intelligence. According to Spearman, this factor is responsible for the overall performance on cognitive tasks including reasoning, problem-solving, and abstract thinking.

Spearman arrives at this theory through his work on factor analysis, a statistical technique that allows him to identify the underlying structure of intelligence. He analyzes data from a wide range of cognitive tests and finds that although people perform better on some tests than others, there is a common cause that accounts for their overall performance.

In order to measure the influence of that factor, he recognises four types of intelligence from which he decides to measure one of them that seems most relevant: "Common sense". To assess this type of intelligence, Spearman asks the older children separately about how they perceive their classmates outside of school. Finally, the rector's wife is also interviewed, but her answers are not usable as she does not know some of the children. The lists obtained, however, seem consistent.

This is representative of Spearman's method in this work, he mixes intuition and statistics to get to his goal, this is also by intuition that he creates the Partial Correlation coefficient, for which he does not provide any rigorous proof.

One of Spearman's main critics is L. L. Thurstone, in "Factorial studies of intelligence" Thurstone and Thurstone [1914], he argues that intelligence is made up of multiple independent factors, rather than a single general factor. Thurstone's theory proposes seven primary mental abilities:

1. "Verbal comprehension: This is the ability to understand and use language effectively, including vocabulary, grammar, and sentence structure.

2. Perceptual speed: This is the ability to quickly and accurately process visual

information.

3. Word fluency: This is the ability to produce a large number of words quickly and accurately.

4. Numerical ability: This is the ability to perform mathematical operations, including arithmetic, algebra, and geometry.

5. Associative memory: This is the ability to store and retrieve information.

6. Spatial ability: This is the ability to mentally manipulate and visualize spatial objects and their relationships.

7. Reasoning: This is the ability to apply logical thinking to solve problems and make decisions."

which he believed were unrelated to each other.

1.3 Hilbert Space

This section and the next one are mostly a mix between Rudin [1987] and Van Bellegem [2020], whereas the following properties are well-known and useful to understand the next chapters.

Definition 2 (Vector space)**.** *A vector space $\mathcal{V}$ is a set of vectors, on which we can apply two operations: addition and multiplication by a scalar satisfying the following properties:*

1. *For all x and y in $\mathcal{V}$, $x+y=y+x$ and $x+(y+z)=(x+y)+z$.*

2. *$\mathcal{V}$ contains a unique null vector that is such that $x+0=x$ for every $x \in \mathcal{V}$.*

3. *For every $x \in \mathbb{V}$, there is a unique $-x$ such that $x+(-x)=0$.*

4. *For every couple* (k,x) *and* (l,y) *where* x *and* y *are vectors and* $k,l \in \mathbb{R}$ *are scalars,* $kx \in \mathcal{V}$ *and* $ly \in \mathcal{V}$ *such that* $k(lx) = (kl)x$, $k(x+y) = kx+ky$ *and* $(k+l)x = kx+lx$.

Definition 3 (Pre-Hilbert space)**.** *A Pre-Hilbert space is a vector space* $\mathcal{H}$. *which can also be called an inner product space (or scalar product space), if, for each ordered pair of vector* $x, y \in \mathcal{H}$, *there exists an associated complex number* $\langle x, y\rangle$ *that satisfies the following properties for every* $x, y, z \in \mathcal{H}$ *and* $k \in \mathbb{R}$*:*

1. $\langle x, y\rangle = \langle y, x\rangle$
2. $\langle x+z, y\rangle = \langle x, y\rangle + \langle z, y\rangle$
3. $\langle kx, y\rangle = k\langle x, y\rangle$.
4. $\langle x, y\rangle \geq 0$
5. $\langle x, y\rangle = 0 \longrightarrow x = 0$
6. $\langle x, x\rangle = \|x\|^2$

Proposition 1 (Cauchy-Schwarz inequality)**.** *Let* $\mathcal{H}$ *be a pre-Hilbert space, then* $|\langle x, y\rangle| \leq \|x\|\,\|y\|$ *for all* $x, y \in \mathcal{H}$.

Proposition 2 (Triangular inequality)**.** *Let* $\mathcal{H}$ *be a pre-Hilbert space, then* $\|x+y\| \leq \|x\| + \|y\|$ *for all* $x, y \in \mathcal{H}$.

Definition 4 (Hilbert space)**.** *Hilbert space, which will now be associated with* $\mathcal{H}$*, is a complete pre-hilbert space, in which every Cauchy sequence converges.*

Theorem 1 (Projection Theorem)**.** *Let* θ *be a closed, convex subspace of the Hilbert space* $\mathcal{H}$. *There exists a couple of applications* P *and* Q *such that* $Px \in \theta$, $Qx \in S$ *whereas* $S \perp \theta$ *and*

$$x = Px + Qx \tag{1.2}$$

for all $x \in \mathcal{H}$. P *and* Q *satisfy the following properties:*

1. *If $x \in \theta$, $Px = x$ and $Qx = 0$. If $x \in S$, $Px = 0$ and $Qx = x$*

2. $\forall x \in \mathcal{H}, \|x - Px\| = \inf \|x - y\| : y \in \theta$

3. $\|x\|^2 = \|Px\|^2 + \|Qx\|^2$

4. *P and Q are linear*

1.4 Hilbert space $L^2(\Omega, \mathcal{A}, \mathbb{P})$

In statistics, a moment is a quantitative measure of the shape, center, and variability of a probability distribution. Specifically, moments are used to describe the distribution of a random variable and are calculated by raising the values of the variable to a certain power and multiplying them by a weighting factor. In a Hilbert space, we mean by L^2 that the second moment (the variance) of the random variables are finite, or more formally: $\mathbb{E}(X^2) < \infty$.

Definition 5. *The vector space $L^2(\Omega, \mathcal{A}, \mathbb{P})$ contains the equivalence classes of almost surely equal random variables on Ω with finite second-order moments. For any random variable $X, Y \in L^2(\Omega, \mathcal{A}, \mathbb{P})$ representing equivalent classes, the product $\langle X, Y \rangle = \mathbb{E}(X^T Y)$ is a scalar product in the sense of Definition 3.*

Proposition 3. *The space $L^2(\Omega, \mathcal{A}, \mathbb{P})$ is complete.*

Definition 6 (Covariance in Hilbert space)**.** *Let $X, Y \in \mathcal{H}$, be two vectors in a Hilbert space. The covariance of X and Y is the inner product of the two centered random variables such that:*

$$\mathbb{C}ov(X, Y) = \langle X - \mathbb{E}(X), Y - \mathbb{E}(Y) \rangle \tag{1.3}$$

Property 1 (Correlation in a Hilbert space)**.** *Let $X, Y \in \mathcal{H}$ and ρ_{XY} be the correlation between the two random variables, the following properties hold true:*

1. $\langle X, Y \rangle = \|X\| \, \|Y\| \cos(X, Y)$

2. $\rho_{XY} = \frac{\langle X - \mathbb{E}(X), Y - \mathbb{E}(Y) \rangle}{\|X - \mathbb{E}(X)\| \|Y - \mathbb{E}(Y)\|} = \cos(X - \mathbb{E}(X), Y - \mathbb{E}(Y)$

1.5 Spearman coefficient

Consider the regressions of X and Y,

$$X = \mathbb{E}(X|\theta) + (X - \mathbb{E}(X|\theta))$$
$$Y = \mathbb{E}(Y|\theta) + (Y - \mathbb{E}(Y|\theta)) \tag{1.4}$$

In psychometric, we call $\mathbb{E}(X|\theta)$ the true score and $(X - \mathbb{E}(X|\theta))$ the measurement error. In econometrics and because we are in a Hilbert space, we call $\mathbb{E}(X|\theta)$ the projection of X onto θ and $(X - \mathbb{E}(X|\theta))$ the residuals, which is the part of interest for this work, it being the random variable free of the influence of the common cause.

The idea behind Partial Correlation is that, once the correlation between variables and the common cause is removed, the true correlation between variables is observable. From this intuition we define partial correlation as such:

$$\rho_{XY \cdot \theta} \stackrel{\text{def}}{=} \mathbb{C}\text{orr}(X - \mathbb{E}(X|\theta), Y - \mathbb{E}(Y|\theta)). \tag{1.5}$$

From the Projection Theorem (1.1), because the residuals are always independent from θ, they are all in S, a hyperplane independent from θ, such that the residuals of X and Y projecting to θ are a function of S. For them to be independent, the angle formed by the residuals ("$Q_\theta X$" and "$Q_\theta Y$") must be of 90°, because the cosine between the centered variables in a vector space is the correlation between those variables, whereas the residuals are, by definition, centered. The space S is a $n - K$ dimensional linear subspace that is orthogonal to the columns of θ (where n is the number of observations and K is the number of variables of θ) (Van Bellegem [2020]).

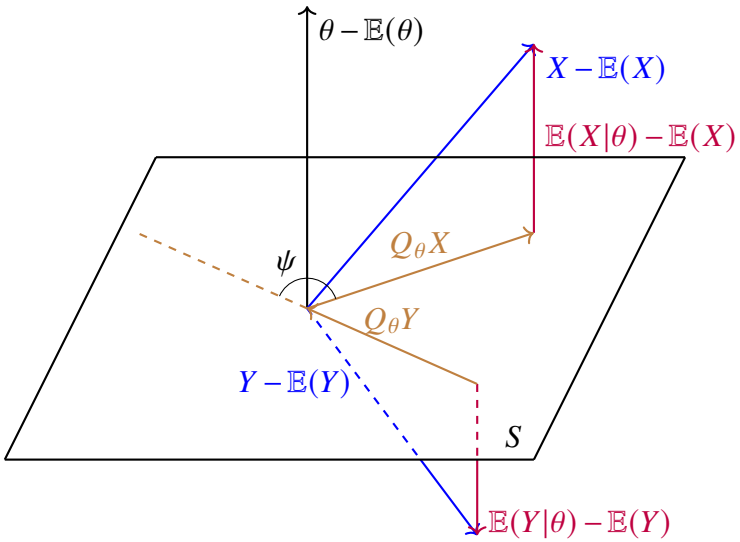


Figure 1.1: Projection Theorem: Linear space of residuals

Spearman had the intuition that the correlation between residuals could be represented as such:

$$\rho_{XY\cdot\theta} = \frac{\rho_{XY} - \rho_{Y\theta}\rho_{X\theta}}{\sqrt{1-\rho_{X\theta}^2}\sqrt{1-\rho_{Y\theta}^2}} \tag{1.6}$$

where ρ_{XY} is the correlation between X and Y which goes accordingly for $\rho_{X\theta}$ and $\rho_{Y\theta}$.

If $|\rho_{XY\cdot\theta}| < |\rho_{XY}|$ (resp. $|\rho_{XY\cdot\theta}| > |\rho_{XY}|$), θ is (resp. is not) correlating X and Y and, once removed, the correlation between the two is proportionally lesser (resp. greater). If $\rho_{XY\cdot\theta} = \rho_{XY}$, then the two variables are not correlated to θ. If $\rho_{XY\cdot\theta} = 0$ and $|\rho_{XY}| > 0$, the two variables are only correlated through θ.

In this paper, we shall demonstrate Equation 1.6 in three different ways then take the last case into account in order to identify reliability in case of a common cause correlating dependant variables.

1.6 Examples

1.6.1 Example of a common factor.

The tables presented below show a relationship between air pollution, cancer rates, and steel production in three countries (Austria, Netherlands, and USA) in the year 2007. PM2.5, is known to be a contributing factor to cancer death and is also associated with steel production. The data are taken from OECD [2023], OECD [2016] and OECD [2009].

Country	PM2.5 exposure (θ)	Death from cancer (X)	Steel prod. on mil. T (Y)
Austria	16.31	224.80	0.92
Netherlands	16.93	261.30	0.45
USA	11.33	224.60	0.33

Table 1.1: Air Pollution, Cancer Rates, and Steel Production by Country

ρ	θ	X	Y
θ	1.00		
X	0.58	1.00	
Y	0.59	-0.31	1.00

Table 1.2: Correlation Table

Table 1.1 provides the mean population exposure to PM2.5, expressed in micrograms per cubic meter, the death rate from cancer per 100,000 persons, and steel production in million tonnes per million persons. The second one (Table 1.2) shows the correlation coefficients between the variables, where θ represents PM2.5 exposure, X represents the death rate from cancer, and Y represents steel production.

According to the Health Effect Institute [2018], PM2.5 is a type of particulate matter that is found in the air and can be harmful to human health. It refers to particles that are smaller than two and a half micrometers in diameter, which is about three percent the diameter of a human hair. These particles can come from a variety of sources, and are mainly emitted from incomplete combustion linked

to industrial or domestic activities, as well as to agriculture and transport. When people breathe in PM2.5, it can penetrate deep into the lungs and even enter the bloodstream, leading to health problems such as asthma, heart disease, and lung cancer.

The correlation coefficients in Table 1.2 show that there is a positive correlation between PM2.5 exposure and the death rate from cancer ($\rho_{X\theta} = 0.58$), indicating that higher PM2.5 exposure could be associated with a higher death rate from cancer. There is also a moderate positive correlation between PM2.5 exposure and steel production ($\rho_{Y\theta} = 0.59$), suggesting that countries with higher levels of steel production tend to have higher levels of PM2.5 exposure.

Finally, there is a weak negative correlation between steel production and the death rate from cancer ($\rho_{XY} = -0.31$), indicating that countries with higher levels of steel production tend to have slightly lower death rates from cancer. It is important to note, however, that correlation does not necessarily imply causality, and further research would be needed to establish causal relationships between these variables, the point here is to mainly illustrate the concept of common factor and partial correlation. By using a standard t-test, we can measure the significance of the correlation coefficient between X and Y. Using the formula $t = \rho_{XY}\sqrt{(n-2)/1-\rho_{XY}^2}$, where $n = 3$ is the number of observations, we get $t = -0.33$ for a degree of freedom equal to one which leads to a p-value of 0.78. Therefore we cannot reject the hypothesis that $\rho_{XY} = 0$.

Using 1.6, the partial correlation between X (cancer death) and Y (steel production) without the influence of θ (PM2.5 exposure) is equal to -1, therefore, the residuals $X - \mathbb{E}(X|\theta)$ and $Y - \mathbb{E}(Y|\theta)$ are in opposite direction and there is a perfect negative correlation. Again, using the formula $t = \rho_{XY}\sqrt{(n-2)/(1-\rho_{XY}^2)}$, where $n = 3$ is the number of observations, we get $t = -\infty$ for a degree of freedom equal to one which leads to a p-value of 0 and therefore we can reject the hypothesis for which $\rho_{XY} = 0$.

An explanation for this negative partial correlation is that cancer deaths have an impact on steel production through workforce reduction. If a significant proportion of the population is dying from cancer, then there will be fewer individuals available to work in steel production, if everyone dies, there is no production at all. This seems intuitively correct in a world where producing steel does not, in any way, leads to cancer. Therefore, the common factor explaining the smaller correlation between X and Y is θ.

1.6.2 Example of a common cause.

The article "Storks deliver babies (p = 0.008)" by Matthews [2000] is a reference to a common myth that storks deliver babies, and the author uses the example to illustrate the concepts of correlation and causation.

Country	Area (km^2)	Storks (pairs)	Humans (10^6)	Birth rate (10^3/yr)
Albania	28750	100	3,2	83
Austria	83860	300	7,6	87
Belgium	30520	1	9,9	118
Bulgaria	111000	5000	9	117
Denmark	43100	9	5,1	59
France	544000	140	56	774
Germany	357000	3300	78	901
Greece	132000	2500	10	106
Holland	41900	4	15	188
Hungary	93000	5000	11	124
Italy	301280	5	57	551
Poland	312680	30000	38	610
Portugal	92390	1500	10	120
Romania	237500	5000	23	367
Spain	504750	8000	39	439
Switzerland	41290	150	6,7	82
Turkey	779450	25000	56	1576

Table 1.3: Table from Matthews [2000]

From these data, R. Matthews finds a strong correlation between birth rate and storks at a significant p-value and questions the meaning of both. A very low p-value

only means that the correlation is significantly different from zero.

The article discusses the white stork and the number of breeding pairs available in 17 European countries, as well as demographic data. There seems to be a possible correlation between the number of stork pairs and the number of births in each country. A linear regression of the annual number of births against the number of breeding pairs of white storks is performed, giving a correlation coefficient of $\rho = 0.62$. The statistical significance of this correlation is assessed using the standard t-test, which gives a p-value of 0.008 with 15 degrees of freedom. The most plausible reason for the observed correlation is the existence of a confounding variable: a common cause to both birth rates and the number of stork pairs that can lead to a statistical correlation between two variables that are not directly related to each other. A potential confounding variable is the land area, for which we shall investigate.

ρ	Area (km^2)	Storks (pairs)	Birth rate (10^3/yr)
Area (km^2)	1		
Storks (pairs)	0,58	1	
Birth rate (10^3/yr)	0,92	0,62	1

Table 1.4: Correlation table.

Using this table of correlation computed with the data from Table 1.3, we can, if we trust Spearman, compute its partial correlation coefficient between the pair of storks and birth rate by removing the influence of the area. This gives us the following value $\rho = 0.27$, for $t = 0.098$ and a p-value of 0.29, meaning we now cannot reject the hypothesis for which $\rho = 0$. The most probable case would be that the common cause is multivariate.

1.6.3 Small economic example.

In a most recent crisis event, we saw the oil price increasing, having an impact on the prices of different goods, correlated to each others by the prices of transports,

energy, some are also correlated through climate change or seasonality affecting the crops. We also use partial correlation in time series analysis, the goal of this work is to verify the efficiency of the tools that we use in econometrics, but also other areas using statistics.

1.7 Reliability

Consider random variables $(X, \theta) \in L^2(\Omega, \mathcal{A}, \mathbb{P})$, reliability is computed as such:

$$\eta_{X|\theta} = \frac{\mathbb{V}\text{ar}(\mathbb{E}(X|\theta))}{\mathbb{V}\text{ar}(X)} \tag{1.7}$$

where $\eta_{X|\theta} \in [0, 1]$.

It is a measure of how much of X is explained by θ, which can be explained graphically thanks to the projection Theorem. The closer to 1, the higher θ explains the variation of X. If $\eta_{X|\theta} = 1$, $X = \mathbb{E}(X|\theta)$ and is a function of θ as in Figure 0.2. If $\eta_{X|\theta} = 0$, X is independent of θ as in Figure 0.1. $\eta_{X|\theta}$ cannot exceed 1 whereas the length of the projection is at most the length of X, and it cannot be negative by definition of variance.

Figure 1.2: $X \perp \theta$

Figure 1.3: X as a function of θ

Note that , in econometrics, we call the reliability test: R^2.

Chapter 2

IDENTIFICATION OF RELIABILITY UNDER LINEARITY

Assumption 1. *$\exists \alpha, \beta$ and $\lambda \in \mathbb{R}$ such that:*

The marginal-conditional decomposition of (θ, Y, X, Z) is:

$$\begin{cases} \theta \sim F(0, \sigma_\theta^2) \\ Y|\theta \sim F(\beta\theta, \gamma_Y^2(\theta)) \\ X|Y, \theta \sim X|\theta \sim F(\alpha\theta, \tau_X^2(\theta)) \\ Z|Y, X, \theta \sim Z|\theta \sim F(\lambda\theta, \zeta_Z^2(\theta)) \end{cases} \tag{2.1}$$

Theorem 2. *Let the random vector (X, Y, Z, θ) in $L^2(\Omega, \mathcal{A}, \mathbb{P})$ such that X, Y, Z are observed but θ is not. Under Assumption 1, the reliability is identified such that:*

$$\eta_{X|\theta} = \frac{\mathbb{V}ar(\mathbb{E}(X|\theta))}{\mathbb{V}ar(X)} = \frac{\rho_{XY}\rho_{XZ}}{\rho_{YZ}} \tag{2.2}$$

where ρ_{UV} denotes the correlation between two random variables U and V.

Because we are in a Hilbert space, the correlation is equivalent to the cosine of the angle formed by the centered variables and $X - \mathbb{E}(X)$ is projected onto $\theta - \mathbb{E}(\theta)$, the assumption of θ being zero mean does not affect the results whereas it does not change anything in terms of correlation or conditional expectations.

Lemma 1. *Let the random vector (X, Y, Z, θ) in $L^2(\Omega, \mathcal{A}, \mathbb{P})$. Under Assumption 1, the partial correlation $\rho_{XY \cdot \theta}$ defined in 1.5 is such that:*

$$\rho_{XY \cdot \theta} = \frac{\rho_{XY} - \rho_{Y\theta}\rho_{X\theta}}{\sqrt{1 - \rho_{X\theta}^2}\sqrt{1 - \rho_{Y\theta}^2}}. \tag{2.3}$$

If $\rho_{X\theta} = \pm 1$ (resp.$\rho_{Y\theta} = \pm 1$), then X (resp. Y) is a function of θ and the norm of the residuals is equal to zero. Therefore, by definition of correlation (Property 1), the correlation between residuals does not exist.

Proof. Let us first expand the correlation between residuals, that is:

$$\mathbb{C}\text{orr}(X - \mathbb{E}(X|\theta), Y - \mathbb{E}(Y|\theta)) = \frac{\mathbb{C}\text{ov}(X - \mathbb{E}(X|\theta), Y - \mathbb{E}(Y|\theta))}{\sqrt{\mathbb{V}\text{ar}(X - \mathbb{E}(X|\theta))}\sqrt{\mathbb{V}\text{ar}(Y - \mathbb{E}(Y|\theta))}}. \tag{2.4}$$

Because the conditional expectations $\mathbb{E}(X|\theta)$ and $\mathbb{E}(Y|\theta)$ are orthogonal to the residuals, the numerator is such that:

$$\mathbb{C}\text{ov}(X - \mathbb{E}(X|\theta), Y - \mathbb{E}(Y|\theta)) = \mathbb{C}\text{ov}(X, Y) - \mathbb{C}\text{ov}(\mathbb{E}(X|\theta), \mathbb{E}(Y|\theta)). \tag{2.5}$$

From Assumption 1, by law of iterated expectations and because θ is distributed around zero, the following equalities hold true:

$$\mathbb{C}\text{ov}(\theta, Y) = \mathbb{E}(\theta Y) = \mathbb{E}\mathbb{E}(\theta Y|\theta) = \mathbb{E}(\theta\mathbb{E}(Y|\theta)) = \mathbb{E}(\theta\beta\theta) = \beta\sigma_\theta^2.$$

Therefore:

$$\beta = \frac{\mathbb{C}\text{ov}(\theta, Y)}{\mathbb{V}\text{ar}(\theta)}. \tag{2.6}$$

We get similar results for α and λ.

Using Lemma 2.1, properties of variance and covariance and definition 2.4, the following equalities hold true:

$$\begin{aligned}
&Cov(\mathbb{E}(X|\theta), \mathbb{E}(Y|\theta)) = Cov(\alpha\theta, \beta\theta) \\
&= Cov(\frac{Cov(\theta, X)}{Var(\theta)}\theta, \frac{Cov(\theta, Y)}{Var(\theta)}\theta) \\
&= Cov(\frac{\sigma_{\theta X}}{\sigma_\theta^2}\theta, \frac{\sigma_{\theta Y}\theta}{\sigma_\theta^2}) \\
&= \frac{\sigma_{\theta X}}{\sigma_\theta^2}\frac{\sigma_{\theta Y}}{\sigma_\theta^2}Cov(\theta, \theta) \\
&= \frac{\sigma_{\theta X}}{\sigma_\theta^2}\frac{\sigma_{\theta Y}}{\sigma_\theta^2}\sigma_\theta^2 = \frac{\sigma_{\theta X}}{\sigma_\theta}\frac{\sigma_{\theta Y}}{\sigma_\theta} \\
&= \rho_{\theta X}\rho_{\theta Y}\sigma_X\sigma_Y
\end{aligned} \tag{2.7}$$

There is yet to solve the denominator which we shall do for X, using properties of variance and covariance, as well as Equation 2.7:

$$\begin{aligned}\mathbb{V}\text{ar}(X - \mathbb{E}(X|\theta)) &= \mathbb{V}\text{ar}(X) - \mathbb{V}\text{ar}(\mathbb{E}(X|\theta)) \\ &= \sigma_X^2 - \mathbb{C}\text{ov}(\mathbb{E}(X|\theta), \mathbb{E}(X|\theta)) \\ &= \sigma_X^2(1 - \rho_{X\theta}^2)\end{aligned}$$

We can now recompose the correlation of residuals:

$$\begin{aligned}\rho_{XY\cdot\theta} &= \frac{\sigma_{XY} - \rho_{Y\theta}\rho_{X\theta}\sigma_X\sigma_Y}{\sigma_X\sigma_Y\sqrt{1-\rho_{X\theta}^2}\sqrt{1-\rho_{Y\theta}^2}} \\ &= \frac{\rho_{XY} - \rho_{Y\theta}\rho_{X\theta}}{\sqrt{1-\rho_{X\theta}^2}\sqrt{1-\rho_{Y\theta}^2}}\end{aligned} \tag{2.8}$$

which concludes Lemma 1. □

Proof of Theorem 2. The common cause assumption implies $\rho_{XY\cdot\theta} = \rho_{XZ\cdot\theta}$ $= \rho_{ZY\cdot\theta} = 0$. Therefore, from Lemma 1 we can write:

$$\begin{cases}\rho_{XY} = \rho_{Y\theta}\rho_{X\theta} \\ \rho_{YZ} = \rho_{Y\theta}\rho_{Z\theta} \\ \rho_{XZ} = \rho_{X\theta}\rho_{Z\theta}.\end{cases}$$

From the precedent equations, we compute the following:

$$\mathbb{C}\text{orr}^2(X, \theta) = \rho_{X\theta}^2 = \frac{\rho_{XY}\rho_{XZ}}{\rho_{Y\theta}\rho_{Z\theta}} = \frac{\rho_{XY}\rho_{XZ}}{\rho_{YZ}}. \tag{2.9}$$

2.9 implies that the reliability is identified as the following tetrachoric relation:

$$\begin{aligned}\eta_{X|\theta} &= \frac{\mathbb{V}\text{ar}(\mathbb{E}(X|\theta))}{\mathbb{V}\text{ar}(X)} = \frac{\mathbb{V}\text{ar}(\alpha\theta)}{\mathbb{V}\text{ar}(X)} \\ &= \frac{\mathbb{V}\text{ar}(\frac{\mathbb{C}\text{ov}(\theta,X)}{\mathbb{V}\text{ar}(\theta)}\theta)}{\mathbb{V}\text{ar}(X)} = \frac{\sigma_{\theta X}^2}{\sigma_X^2\sigma_\theta^2} \\ &= \rho_{X\theta}^2 = \frac{\rho_{XY}\rho_{XZ}}{\rho_{YZ}}\end{aligned} \tag{2.10}$$

Whereas $\rho_{X\theta}^2 = \rho_{\theta X}^2$, the following equality holds true: $\eta_{X|\theta} = \eta_{\theta|X}$. □

Chapter 3

IDENTIFICATION OF RELIABILITY UNDER NORMALITY

Assumption 2. *The marginal-conditional decomposition of* (θ, Y, X, Z) *is:*

$$\begin{cases} \theta \sim \mathcal{N}(0, \sigma_\theta^2) \\ Y|\theta \sim \mathcal{N}(\mathbb{E}(Y|\theta), \gamma_Y^2(\theta)) \\ X|Y, \theta \sim X|\theta \sim \mathcal{N}(\mathbb{E}(X|\theta), \tau_X^2(\theta)) \\ Z|Y, X, \theta \sim Z|\theta \sim \mathcal{N}(\mathbb{E}(Z|\theta), \zeta_X^2(\theta)) \end{cases} \tag{3.1}$$

First, a reminder of the properties that are relevant to this demonstration (Van Bellegem [2020]).

Definition 7. *Multivariate Normal distribution*

We assumed a normal distribution, our three uni-variate random variables can be seen as one multivariate random variable H *such that* $H \in \mathbb{R}^3$ *which follows a normal distribution with parameters* (μ, Σ) *where* $\mu \in \mathbb{R}^3$ *and* Σ *is a 3 x 3 positive definite, finite matrix. We can write its formal distribution as such:*

$$\begin{bmatrix} \theta \\ Y \\ X \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_0 \\ \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{bmatrix} \right) \tag{3.2}$$

Where $\mu_0 = 0$ *and such that:*

$$\left(\begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{bmatrix} = \begin{bmatrix} \sigma_\theta^2 & \sigma_{\theta Y} & \sigma_{\theta X} \\ \sigma_{\theta Y} & \sigma_Y^2 & \sigma_{XY} \\ \sigma_{\theta X} & \sigma_{XY} & \sigma_X^2 \end{bmatrix} \right)$$

Property 2. *Conditional distribution*

In order to find the distribution of $X|\theta$ and $Y|\theta$, we need the property of conditional distribution:

$$\begin{aligned} Y|\theta &\sim \mathcal{N}(\mu_1 - \Sigma_{21}\Sigma_{11}^{-1}(\mu_0 - \theta), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}) \\ X|Y,\theta \sim X|\theta &\sim \mathcal{N}(\mu_2 - \Sigma_{31}\Sigma_{11}^{-1}(\mu_0 - \theta), \Sigma_{33} - \Sigma_{31}\Sigma_{11}^{-1}\Sigma_{13}) \end{aligned} \tag{3.3}$$

Theorem 3. *Let the random vector (X,Y,Z,θ) in $L^2(\Omega,\mathcal{A},\mathbb{P})$ such that X,Y,Z are observed but θ is not. Under Assumption 1, the reliability is identified such that:*

$$\eta_{X|\theta} = \frac{\mathbb{V}ar(\mathbb{E}(X|\theta))}{\mathbb{V}ar(X)} = \frac{\rho_{XY}\rho_{XZ}}{\rho_{YZ}} \tag{3.4}$$

Lemma 2. *Let the random vector (X,Y,Z,θ) in $L^2(\Omega,\mathcal{A},\mathbb{P})$. Under 2, the partial correlation $\rho_{XY\cdot\theta}$ defined in 1.5 is such that:*

$$\rho_{XY\cdot\theta} = \frac{\rho_{XY} - \rho_{Y\theta}\rho_{X\theta}}{\sqrt{1-\rho_{X\theta}^2}\sqrt{1-\rho_{Y\theta}^2}}. \tag{3.5}$$

Proof. From definition 7 and property 2 come the next equalities:

$$\mathbb{E}(Y|\theta) = \mu_1 - \Sigma_{21}\Sigma_{11}^{-1}(\mu_0 - \theta) = \mu_1 + \frac{\sigma_{\theta Y}\theta}{\sigma_\theta^2}. \tag{3.6}$$

By using definitions and property of covariance and expectation (more specifi-

cally the fact that the expectation of a constant is that constant):

$$\begin{aligned}
\mathbb{C}\text{ov}(\mathbb{E}(Y|\theta),\mathbb{E}(X|\theta)) &= \mathbb{C}\text{ov}(\mu_1+\frac{\sigma_{\theta Y}\theta}{\sigma_\theta^2},\mu_2+\frac{\sigma_{\theta X}\theta}{\sigma_\theta^2}) \\
&= \mathbb{E}([\mu_1+\frac{\sigma_{\theta Y}\theta}{\sigma_\theta^2}-\mathbb{E}(\mu_1+\frac{\sigma_{\theta Y}\theta}{\sigma_\theta^2})] \\
&\qquad [\mu_2+\frac{\sigma_{\theta X}\theta}{\sigma_\theta^2}-\mathbb{E}(\mu_2+\frac{\sigma_{\theta X}\theta}{\sigma_\theta^2})]) \\
&= \mathbb{E}([\frac{\sigma_{\theta Y}\theta}{\sigma_\theta^2}-\mathbb{E}(\frac{\sigma_{\theta Y}\theta}{\sigma_\theta^2})][\frac{\sigma_{\theta X}\theta}{\sigma_\theta^2}-\mathbb{E}(\frac{\sigma_{\theta X}\theta}{\sigma_\theta^2})]) \\
&= \mathbb{C}\text{ov}(\frac{\sigma_{\theta Y}\theta}{\sigma_\theta^2},\frac{\sigma_{\theta X}\theta}{\sigma_\theta^2}) = \frac{\sigma_{\theta X}}{\sigma_\theta^2}\frac{\sigma_{\theta Y}}{\sigma_\theta^2}\mathbb{C}\text{ov}(\theta,\theta) \\
&= \frac{\sigma_{\theta Y}}{\sigma_\theta}\frac{\sigma_{\theta X}}{\sigma_\theta} = \rho_{\theta X}\rho_{\theta Y}\sigma_X\sigma_Y. \qquad (3.7)
\end{aligned}$$

Then, by definition of variance:

$$\mathbb{V}\text{ar}(\mathbb{E}(X|\theta)) = \rho_{X\theta}^2\sigma_X^2 \qquad (3.8)$$

Therefore, we can recompose the correlation of residuals as such:

$$\rho_{XY\cdot\theta} = \frac{\sigma_{XY}-\rho_{Y\theta}\rho_{X\theta}\sigma_X\sigma_Y}{\sigma_X\sigma_Y\sqrt{1-\rho_{X\theta}^2}\sqrt{1-\rho_{Y\theta}^2}} \qquad (3.9)$$

$$= \frac{\rho_{XY}-\rho_{Y\theta}\rho_{X\theta}}{\sqrt{1-\rho_{X\theta}^2}\sqrt{1-\rho_{Y\theta}^2}} \qquad (3.10)$$

which concludes Lemma 2. □

Proof of Theorem 3. The common cause assumption implies $\rho_{XY\cdot\theta} = \rho_{XZ\cdot\theta}$ $= \rho_{ZY\cdot\theta} = 0$. Therefore, from Lemma 2 we can write:

$$\begin{cases}
\rho_{XY} = \rho_{Y\theta}\rho_{X\theta} \\
\rho_{YZ} = \rho_{Y\theta}\rho_{Z\theta} \\
\rho_{XZ} = \rho_{X\theta}\rho_{Z\theta}.
\end{cases}$$

From the precedent equations, we compute the following:

$$\mathbb{C}\text{orr}^2(X,\theta) = \rho_{X\theta}^2 = \frac{\rho_{XY}\rho_{XZ}}{\rho_{Y\theta}\rho_{Z\theta}} = \frac{\rho_{XY}\rho_{XZ}}{\rho_{YZ}}. \qquad (3.11)$$

3.8 implies that the reliability is identified as the following tetrachoric relation:

$$\eta_{X|\theta} = \frac{\mathbb{V}\text{ar}(\mathbb{E}(X|\theta))}{\mathbb{V}\text{ar}(X)} = \frac{\rho^2_{X\theta}\sigma^2_X}{\sigma^2_X} = \rho^2_{X\theta} = \frac{\rho_{XY}\rho_{XZ}}{\rho_{YZ}}. \tag{3.12}$$

Whereas $\rho^2_{X\theta} = \rho^2_{\theta X}$, the following equality holds true: $\eta_{X|\theta} = \eta_{\theta|X}$. □

Chapter 4

GENERAL THEOREM

In this chapter, we shall demonstrate that Spearman's partial correlation is generally not true in case of a non linear regression. First, let us demonstrate by a counterexample the relationship between Spearman's formula and the correlation between residuals. Then, we shall give intuition on why the formula cannot be used in the presented case and present a new formula.

4.1 Counterexample

Let three random variables $X, Y, \Theta \in \mathbb{R}^3$ such that $X = (3,2,4,1)$, $Y = (2,4,6,8)$ and $\Theta = (1,2,-2,2)$. We know nothing about the relationship between those variables but we want to compute two regressions:

1. $X = \mathbb{E}(X|\Theta) + X - \mathbb{E}(X|\Theta)$
2. $Y = \mathbb{E}(Y|\Theta) + Y - \mathbb{E}(Y|\Theta)$

In order to do so, we can either compute a linear regression, take the risk of assuming the linearity of $\mathbb{E}(X|\Theta)$ and $\mathbb{E}(Y|\Theta)$, or we can compute manually $\mathbb{E}(X|\Theta)$ and $\mathbb{E}(Y|\Theta)$. We shall do both.

	X	Y	Θ
X	1		
Y	0,79	1	
Θ	-0,89	-0,87	1

Table 4.1: Correlation table

From the correlation table, we can see that the three variables are correlated.

4.1.1 Linear regression

The linear regression is a specific restriction that assumes that the regression function is a linear function of θ components, that is

$$g(\Theta) = \mathbb{E}(X|\Theta) = \beta\theta^T \tag{4.1}$$

We compute the linear regression for X and Y and we get the following values:

$\mathbb{E}(X\|\Theta)$	$\mathbb{E}(Y\|\Theta)$
2,35	3,4
1,74	1,98
4,16	7,65
1,74	1,98

Table 4.2: Conditional expectations of a linear regression

We follow by computing the correlation between residuals:

X-E(X|O)	Y-E(Y|O)
0,65	-1,4
0,26	2,02
-0,16	0,35
-0,74	-0,98

Table 4.3: Residuals of a linear regression

We compute the correlation between residuals $\mathbb{C}\text{orr}(X - \mathbb{E}(X|\Theta), Y - \mathbb{E}(Y|\Theta)) = 0,1$. From Correlation Table 4.1 and formula 1.6, $\rho_{XY\cdot\theta} = 0,1$, which is equal to the correlation between residuals as we proved theoretically in Chapter two.

4.1.2 Non linear regression

For this model, no assumptions are needed, we construct $\mathbb{E}(X|\Theta)$ by definition of conditional expectation: $\mathbb{E}(X|\Theta = \theta) = \sum x\mathbb{P}(X = x|\Theta = \theta)$ (Van Bellegem [2020]).

$\mathbb{E}(X\|\Theta)$	$\mathbb{E}(Y\|\Theta)$
3	2
1,5	2,5
4	8
1,5	2,5

Table 4.4: Conditional expectations of a non linear regression

We get different values for each observation than from the precedent model, which is therefore biased. We can now compute the correlation between residuals and compare with Spearman's formula and precedent results. The residuals have the following values:

$X-\mathbb{E}(X\|\Theta)$	$Y-\mathbb{E}(Y\|\Theta)$
0	0
0,5	1,5
0	0
-0,5	-1,5

Table 4.5: Residuals of a non linear regression

We compute the correlation between residuals $\mathbb{C}\text{orr}(X-\mathbb{E}(X|\Theta), Y-\mathbb{E}(Y|\Theta)) = 1$. From Correlation Table 4.1 and formula 1.6, $\rho_{XY\cdot\theta} = 0,1$, which is far from equal to the correlation between residuals. We can therefore conclude that the formula does not work for non linear regressions.

4.2 The General Theorem

Assumption 3. *The marginal-conditional decomposition of* (θ, Y, X, Z) *is:*

$$\begin{cases} \theta \sim F(\mathbb{E}(\theta), \sigma_\theta^2) \\ Y|\theta \sim F(\mathbb{E}(Y|\theta), \gamma_Y^2(\theta)) \\ X|Y,\theta \sim X|\theta \sim F(\mathbb{E}(X|\theta), \tau_X^2(\theta)) \\ Z|Y,X,\theta \sim Z|\theta \sim F(\mathbb{E}(Z|\theta), \zeta_Z^2(\theta)) \end{cases} \tag{4.2}$$

Theorem 4. *Let the random vector* (X, Y, Z, θ) *in* $L^2(\Omega, \mathcal{A}, \mathbb{P})$ *such that* X, Y, Z *are observed but* θ *is not. Under Assumption 3, the reliability is identified such that:*

$$\eta_{X|\theta} = \frac{\mathbb{V}ar(\mathbb{E}(X|\theta))}{\mathbb{V}ar(X)} = \frac{\rho_{XY}\rho_{XZ}}{\rho_{YZ}} \frac{\rho_{\mathbb{E}(Y|\theta)\mathbb{E}(Z|\theta)}}{\rho_{\mathbb{E}(X|\theta)\mathbb{E}(Y|\theta)}\rho_{\mathbb{E}(X|\theta)\mathbb{E}(Z|\theta)}}. \tag{4.3}$$

Where ρ_{UV} *denotes the correlation between two random variables* U *and* V.

This Theorem is based on the sole assumption that the random variables are univariate, belong to the space $L^2(\Omega, \mathcal{A}, \mathbb{P})$. Therefore, they are square-integrable, meaning that the expected value of the square of the variables are finite. In other words, those are said to be square-integrable if their variances are finite and their inner products are well-defined in the L^2 space. This result is new and the following proof is based on the projection Theorem.

Lemma 3. *Let the random vector* (X, Y, Z, θ) *in* $L^2(\Omega, \mathcal{A}, \mathbb{P})$. *Then the partial correlation* $\rho_{XY\cdot\theta}$ *defined in 1.5 is such that:*

$$\rho_{XY\cdot\theta} = \frac{\rho_{XY} - \rho_{X\mathbb{E}(X|\theta)}\rho_{Y\mathbb{E}(Y|\theta)}\rho_{\mathbb{E}(X|\theta)\mathbb{E}(Y|\theta)}}{\sqrt{1-\rho^2_{X\mathbb{E}(X|\theta)}}\sqrt{1-\rho^2_{Y\mathbb{E}(Y|\theta)}}}. \tag{4.4}$$

If $\rho_{X\mathbb{E}(X|\theta)} = \pm 1$, then X is a function of θ and the norm of the residuals is equal to zero. Therefore, by definition of correlation (Property 1), the correlation between residuals does not exist, this goes accordingly if $\rho_{Y\mathbb{E}(Y|\theta)} = \pm 1$. It is because $\rho_{X\mathbb{E}(X|\theta)} \neq \rho_{X\theta}$ when $\mathbb{E}(X|\theta)$ is multivariate and non linear that the partial correlation formula cannot be applied in this case. Indeed, when θ is univariate and $\mathbb{E}(X|\theta)$ is linear, because $\mathbb{E}(X|\theta)$ is a projection of X onto the space generated by θ of dimension one, the following relations hold true:

1. $\rho_{X\mathbb{E}(X|\theta)} = \rho_{X\theta}\rho_{\mathbb{E}(X|\theta)\theta}$.

2. $\rho_{\mathbb{E}(X|\theta)\mathbb{E}(Y|\theta)} = \rho_{\mathbb{E}(X|\theta)\theta}\rho_{\mathbb{E}(Y|\theta)\theta}$.

3. $\rho_{X\mathbb{E}(X|\theta)}\rho_{Y\mathbb{E}(Y|\theta)}\rho_{\mathbb{E}(X|\theta)\mathbb{E}(Y|\theta)} = \rho_{X\theta}\rho_{Y\theta}$.

Which is trivial once showed graphically, but not true in case of non linear regressions.

Proof. First, we expand the correlation between residuals, that is:

$$\mathbb{C}\text{orr}(X-\mathbb{E}(X|\theta),Y-\mathbb{E}(Y|\theta)) = \frac{\mathbb{C}\text{ov}(X-\mathbb{E}(X|\theta),Y-\mathbb{E}(Y|\theta))}{\sqrt{\mathbb{V}\text{ar}(X-\mathbb{E}(X|\theta))}\sqrt{\mathbb{V}\text{ar}(Y-\mathbb{E}(Y|\theta))}}. \tag{4.5}$$

Because the conditional expectations $\mathbb{E}(X|\theta)$ and $\mathbb{E}(Y|\theta)$ are orthogonal to the residuals, the numerator is such that:

$$\mathbb{C}\text{ov}(X-\mathbb{E}(X|\theta),Y-\mathbb{E}(Y|\theta)) = \mathbb{C}\text{ov}(X,Y) - \mathbb{C}\text{ov}(\mathbb{E}(X|\theta),\mathbb{E}(Y|\theta)). \tag{4.6}$$

Remind that:

$$\mathbb{V}\text{ar}(\mathbb{E}(X|\theta)) = \mathbb{C}\text{ov}(X,\mathbb{E}(X|\theta)). \tag{4.7}$$

We use that identity to write:

$$\begin{aligned}\mathbb{C}\text{ov}(X,\mathbb{E}(X|\theta))\mathbb{C}\text{ov}(Y,\mathbb{E}(Y|\theta)) =& \mathbb{V}\text{ar}(\mathbb{E}(X|\theta))\mathbb{V}\text{ar}(\mathbb{E}(Y|\theta)) \\ =& \sigma_X\sqrt{\mathbb{V}\text{ar}(\mathbb{E}(X|\theta))}\rho_{X\mathbb{E}(X|\theta)} \\ & \sigma_Y\sqrt{\mathbb{V}\text{ar}(\mathbb{E}(Y|\theta))}\rho_{Y\mathbb{E}(Y|\theta)}\end{aligned} \tag{4.8}$$

By definition of correlation and 4.7, we get the following equality:

$$\mathbb{C}\text{ov}(\mathbb{E}(X|\theta),\mathbb{E}(Y|\theta)) = \rho_{X\mathbb{E}(X|\theta)}\rho_{Y\mathbb{E}(Y|\theta)}\rho_{\mathbb{E}(X|\theta)\mathbb{E}(Y|\theta)}\sigma_X\sigma_Y. \tag{4.9}$$

From 4.9 comes the next equality:

$$\mathbb{V}\text{ar}(\mathbb{E}(X|\theta)) = \rho^2_{X\mathbb{E}(X|\theta)}\sigma^2_X. \tag{4.10}$$

Therefore, the variance of the residuals is such that:

$$\mathbb{V}\text{ar}(X-\mathbb{E}(X|\theta)) = \sigma^2_X(1-\rho^2_{X\mathbb{E}(X|\theta)}). \tag{4.11}$$

We can now recompose the correlation of residuals:

$$\begin{aligned}\rho_{XY\cdot\theta} &= \frac{\sigma_{XY} - \rho_{X\mathbb{E}(X|\theta)}\rho_{Y\mathbb{E}(Y|\theta)}\rho_{\mathbb{E}(X|\theta)\mathbb{E}(Y|\theta)}\sigma_X\sigma_Y}{\sigma_X\sigma_Y\sqrt{1-\rho^2_{X\mathbb{E}(X|\theta)}}\sqrt{1-\rho^2_{Y\mathbb{E}(Y|\theta)}}} \\ &= \frac{\rho_{XY} - \rho_{X\mathbb{E}(X|\theta)}\rho_{Y\mathbb{E}(Y|\theta)}\rho_{\mathbb{E}(X|\theta)\mathbb{E}(Y|\theta)}}{\sqrt{1-\rho^2_{X\mathbb{E}(X|\theta)}}\sqrt{1-\rho^2_{Y\mathbb{E}(Y|\theta)}}}.\end{aligned} \tag{4.12}$$

which concludes Lemma 3.

□

Proof of Theorem 4. The common cause assumption implies $\rho_{XY\cdot\theta} = \rho_{XZ\cdot\theta}$ $= \rho_{ZY\cdot\theta} = 0$. Therefore, from Lemma 3 we can write:

$$\begin{cases} \rho_{XY} = \rho_{X\mathbb{E}(X|\theta)}\rho_{Y\mathbb{E}(Y|\theta)}\rho_{\mathbb{E}(X|\theta)\mathbb{E}(Y|\theta)} \\ \rho_{YZ} = \rho_{Y\mathbb{E}(Y|\theta)}\rho_{Z\mathbb{E}(Z|\theta)}\rho_{\mathbb{E}(Y|\theta)\mathbb{E}(Z|\theta)} \\ \rho_{XZ} = \rho_{X\mathbb{E}(X|\theta)}\rho_{Z\mathbb{E}(Z|\theta)}\rho_{\mathbb{E}(X|\theta)\mathbb{E}(Z|\theta)}. \end{cases}$$

From the precedent equations, we compute the following:

$$\begin{aligned} \mathbb{C}\mathrm{orr}^2(X, \rho_{X\mathbb{E}(X|\theta)}) = \rho^2_{X\rho_{X\mathbb{E}(X|\theta)}} &= \frac{\rho_{XY}\rho_{XZ}}{\rho_{Y\theta}\rho_{Z\theta}} \frac{1}{\rho_{\mathbb{E}(X|\theta)\mathbb{E}(Y|\theta)}\rho_{\mathbb{E}(X|\theta)\mathbb{E}(Z|\theta)}} \\ &= \frac{\rho_{XY}\rho_{XZ}}{\rho_{YZ}} \frac{\rho_{\mathbb{E}(Y|\theta)\mathbb{E}(Z|\theta)}}{\rho_{\mathbb{E}(X|\theta)\mathbb{E}(Y|\theta)}\rho_{\mathbb{E}(X|\theta)\mathbb{E}(Z|\theta)}}. \end{aligned} \tag{4.13}$$

4.7 implies that the reliability is identified as the following tetrachoric relation:

$$\begin{aligned} \eta_{X|\theta} = \frac{\mathbb{V}\mathrm{ar}(\mathbb{E}(X|\theta))}{\mathbb{V}\mathrm{ar}(X)} = \frac{\rho^2_{X\mathbb{E}(X|\theta)}\sigma^2_X}{\sigma^2_X} &= \rho^2_{X\mathbb{E}(X|\theta)} \\ = \frac{\rho_{XY}\rho_{XZ}}{\rho_{YZ}} \frac{\rho_{\mathbb{E}(Y|\theta)\mathbb{E}(Z|\theta)}}{\rho_{\mathbb{E}(X|\theta)\mathbb{E}(Y|\theta)}\rho_{\mathbb{E}(X|\theta)\mathbb{E}(Z|\theta)}}. \end{aligned} \tag{4.14}$$

Reliability is, therefore, not identified in case of a non linear regression. □

Chapter 5

MULTIVARIATE CASE AND CONCLUSION

5.1 Partial correlation and reliability for a multivariate θ.

In case of a multivariate θ, if $\mathbb{E}(X|\theta)$ and $\mathbb{E}(Y|\theta)$ are linear and we observe θ, we compute the partial correlation as such. Take, for example, the case of a bivariate θ, which is well resumed in the article of Chidiebere [2015]:

$$\rho_{XY\cdot\theta_1,\theta_2} = \frac{\rho_{XY\cdot\theta_1} - \rho_{X\theta_2\cdot\theta_1}\rho_{Y\theta_2\cdot\theta_1}}{\sqrt{1-\rho^2_{X\theta_2\cdot\theta_1}}\sqrt{1-\rho^2_{Y\theta_2\cdot\theta_1}}}. \tag{5.1}$$

This way, we can remove the impact of every θ from the correlation between X and Y. If the common factor is composed of more than one variable, repeat this operation until every variables' influence is removed from the correlation between X and Y.

Therefore, the results are not limited to an univariate case, Spearman's formula can be applied as long as the second moments of the variables are finite, assuming a linear model of conditional expectations.

5.2 Conclusion.

The identification of reliability allows us to measure the impact of a non measured variable, if that said variable is univariate and we assume that the conditional expectations are linear. Therefore, what is left is a matter of inference, interpretation and decomposition of the variables, as it often seems to be in statistics.

This work leaves a few questions to be answered: Now that we demonstrated that Spearman's partial correlation does not work in case of a non linear model, we need a new generalized tool to measure partial correlation between random variables in order to avoid strong assumptions.

An interesting following to the work on correlations is "Generalized Correlation and Kernel Causality with Applications in Development Economics" from Vinod [2017], who treats kernel causality in multivariate non linear relations with asymmetric correlations.

BIBLIOGRAPHY

O. C. Chidiebere. Multivariate approach to partial correlation analysis. *Science Journal of Applied Mathematics and Statistics*, Vol. 3(No. 3):165–170, 2015. doi: 10.11648/j.sjams.20150303.20.

H. E. Institute. State of global air 2018. *Special Report. Boston*, 2018.

R. Matthews. Storks deliver babies (p=0.008). *Teaching Statistics*, 22:36–38, June 2000. doi: 10.1111/1467-9639.00013.

OECD. Steel production (accessed on 07 may 2023). 2009. doi: https://doi.org/10.1787/factbook-2009-table20-en. URL https://www.oecd-ilibrary.org/content/component/factbook-2009-table20-en.

OECD. Exposition aux particules fines pm2,5 - pays et régions (accessed on 07 may 2023). 2016. doi: https://doi.org/10.1787/ba4edead-fr. URL https://www.oecd-ilibrary.org/content/data/ba4edead-fr.

OECD. Deaths from cancer (accessed on 07 may 2023). 2023. doi: 10.1787/8ea65c4b-en.

H. Reichenbach. *The Direction of Time*. Mineola, N.Y.: Dover Publications, 1956.

W. Rudin. *Real and Complex Analysis*. 1987.

W. A. Shewhart. *Economic Control of Quality of Manufactured Product*. D. Van Nostrand Company, Inc., 1931.

C. Spearman. General intelligence, objectively determined and measured. *The American Journal of Psychology*, Vol.15(N°02):201–292, April 1904.

L. L. Thurstone and T. G. Thurstone. Factoral studies of intelligence. *Psychometric Monographs*, Vol.94(N°02), 1914.

S. Van Bellegem. Statistics for economists. 2020.

H. Vinod. Generalized correlation and kernel causality with applications in development economics. *Communications in Statistics - Simulation and Computation*, 0:1–22, January 2017. doi: 10.1080/03610918.2015.1122048.