Advances in Economics, Business and Management Research, volume 203
Proceedings of the 2021 3rd International Conference on Economic Management and Cultural Industry (ICEMCI 2021)

Survey Data Analysis Using Information Theory – A New Method for Business Research

Pak Hou Che[1,*] Caleb Huanyong Chen[1] Chunsheng Li[1]

[1]*School of Business, Macau University of Science and Technology, Macao SAR, China*
[*]*Corresponding author. Email: pahche@must.edu.mo*

ABSTRACT
As a traditional method to test relationships between variables, regression has been used widely but there are also limitations. Based on information theory, this study proposed an entropy method for researchers to use as an alternative option. Firstly, the mathematical foundation basing on entropy theory was built. Then a computer-simulated model and data were generated and tested. Finally, real survey data were used to test the method. Both computer-simulated and survey models proved the equivalence between the entropy method and traditional regression, while the entropy method told more details behind the data. This study extends methodologies and research in business.

Keywords: *Survey analysis, information theory, entropy, regression*

1. INTRODUCTION

In business research, regression has been a traditional method to test relationships between variables. However, regression fails to verify the relationships between two variables in some simple cases. Noticing the methodological gap, this study attempts to propose an alternative method based on information theory to strengthen the test of relationships.

Information theory is a mathematical research field that studies the transmission of information through communication systems, e.g. the amount of information that we could transmit in a 5G network. In 1948, Shannon [1] formally define the mathematical representation of information, and studied and answered the questions about the maximum amount of information that can be transmitted through some noisy channels.

Information can be thought as uncertainty, the information theoretic function – entropy in information theory measures the amount of uncertainty that could produce of random variables. Consider a scenario that Alice talking to Bob in a very noisy place. It is natural that sometimes Alice may need to repeat her words multiple times, so that Bob could understand what Alice is talking about with low probability of error. Shannon made this concept concrete, and shows that, the fundamental limit of the transmission rate given the noise level.

Rather than in communication, information theory has been widely adopted in different research area including economy, finance, and psychology. The psychology society was excited after the birth of information theory as the researchers believed that the measures in information theory can be adopted in psychological experiments and behaviour analysis. A work [2] showed that the response sequences could be analyzed using information theory. Miller applied information measures to show the "magical seven" [3]. Since then, ideas from information theory have then been used in communication, group decision-making [4, 5], and sequences of talk and silence [6, 7, 8] etc.

In this work, based on entropy function in information theory, the mathematical foundation with definitions and formulas are built. Then computer-simulated data are employed to illustrate the entropy method. And the results are compared with those using traditional regression. To confirm the application of the entropy method, a dataset from a real survey is used to test the relationship between two variables. Both the computer-simulated and survey models prove the equivalence between the entropy method and traditional regression, while the entropy method can tell more details behind the data. Therefore, this study contributes to

methodology and business research by providing an alternative method for researchers.

2. BACKGROUND

2.1. Information Theory

In this section, information theoretic functions will be used in analyzing the survey data.

Definition 1. (Discrete Entropy) Given a discrete random variable X with alphabet $\mathcal{X}$ and the probability mass function $p(x) = P(X = x), x \in \mathcal{X}$. The entropy $H(X)$ of X is defined by

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\ \log p(x), \qquad (1)$$

where the logarithm is of base 2.

From equation (1), we can see that the maximum value of Shannon's entropy is $\log|\mathcal{X}|$ when $p(x) = 1/|\mathcal{X}|$ (we consider only $|\mathcal{X}|$ is finite) for all $x \in \mathcal{X}$. Also, we assume that $0 \log 0 = 0$.

We can see that from the above definition, the entropy function $H(X)$ measures the amount of uncertainty (randomness) of the random variable X, that is, when the probability $p(x)$ is uniformly distributed (completely random), the entropy $H(X)$ is maximized, and if X is deterministic, which means, there is no randomness, then the entropy $H(X) = 0$.

Similar to **Definition 1.**, the discrete entropy can be further extended to the multiple random variables. In this work, we use two random variables in measuring the "relationship" between an independent variable and a dependent variable, and the following definition gives us the two random variables entropy.

Definition 2. (Discrete Entropy for Two Random Variables) Given two discrete random variables (X, Y) with the joint probability mass function $p(x, y)$. The joint entropy $H(X, Y)$ is defined by

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\ \log p(x,y). \qquad (2)$$

Next, we define the condition entropy.

Definition 3. (Discrete Conditional Entropy) Given two discrete random variables (X, Y) with the joint probability mass function $p(x, y)$. The conditional entropy $H(Y|X)$ is defined by

$$H(Y|X) = \sum_{x\in\mathcal{X}} p(x)H(Y|X = x). \qquad (3)$$

So, we have,

$$H(Y|X) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x). \qquad (4)$$

To understand the discrete conditional entropy, one may interpret as a measure of uncertainty given prior knowledge. That is, the conditional entropy $H(Y|X)$ measures the uncertainty of Y given that X is known.

From the above definitions about the entropy, we have,

$$H(X,Y) = H(X) + H(Y|X) \qquad (5)$$
$$= H(Y) + H(X|Y) \qquad (6)$$

Next, a two variables measure is introduced, which is called mutual information, is a measure the amount between two random variables in common.

Definition 4. (Mutual Information) Given two discrete random variables (X, Y) with the joint probability mass function $p(x, y)$, marginal probability mass function $p(x)$ and $p(y)$. The mutual information $I(X; Y)$ is defined by

$$H(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}. \qquad (7)$$

From the definition of mutual information, the relationships to the Shannon's entropy functions is shown as follows,

$$I(X;Y) = H(Y) - H(Y|X) \qquad (7)$$
$$= H(X) - H(X|Y) \qquad (8)$$
$$= H(X) + H(Y) - H(X,Y). \qquad (9)$$

2.2. Relationship Between Information Theory and Survey Analysis

In communication theory, the scenario about "talking in a noisy place" we described earlier in the introduction can be modelled by a communication model as shown in **Figure 1**. That is, the receiver (Bob) is trying to recover the original message M from the sender (Alice). The message M first encodes into a codeword X^n (one may imagine that X^n is a 0-1 sequence of length n, as the fundamental transmission unit is in bits, which is 0 and 1), and the output of the channel depends probabilistic on the input, it is characterized by a transition probability $p(y|x)$. The receiver decoded the message $\hat{M}$ by observing Y^n.

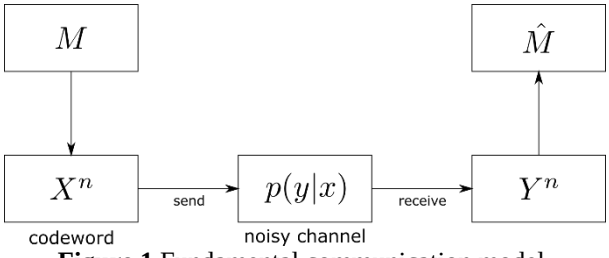

Figure 1 Fundamental communication model.

On the other hand, in survey analysis, consider a construct A and construct B, then hypotheses the relationship between construct A and construct B, such as positively influence or negatively influence. Regression analysis can be conducted from the dataset, and used in answering such questions by checking the statistical parameters of

$$y = \beta_0 + \beta_1 x. \tag{10}$$

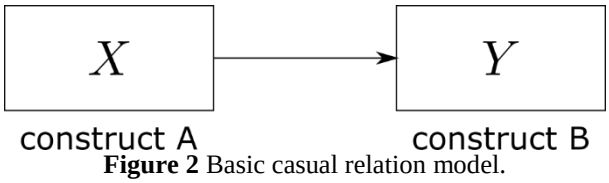

Figure 2 Basic casual relation model.

We see that this basic communication model is like the casual relationship model in survey analysis, see **Figure 2**. The survey analysis focuses in understanding the relationship between construct A and construct B. In terms of information theory, the survey analysis can be viewed as the problem of finding the channel parameters given the empirical distributions to be X and Y.

2.3. A Counter Example That Traditional Statistical Measurements are Failed

Consider a simple survey analysis with two constructs A and B. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be 3-level Likert scale variable with respective to construct A and B. Also, the level of survey namely "agree", "neutral", "disagree", and values $-\mathbf{1}, \mathbf{0}, \mathbf{1}$ respectively. Also, we let $\boldsymbol{Y}$ to be deterministic, that is, $\boldsymbol{Y} = \mathbf{1}$, if $\boldsymbol{X} = -\mathbf{1}$; $\boldsymbol{Y} = \mathbf{0}$, if $\boldsymbol{X} = \mathbf{1}$; and $\boldsymbol{Y} = -\mathbf{1}$, if $\boldsymbol{X} = \mathbf{0}$. Given a certain input distribution of $\boldsymbol{X}$, $\boldsymbol{P}(\boldsymbol{X} = -\mathbf{1}) = \boldsymbol{P}(\boldsymbol{X} = \mathbf{0}) = \mathbf{4}/\mathbf{9}$ and $\boldsymbol{P}(\boldsymbol{X} = \mathbf{1}) = \mathbf{1}/\mathbf{9}$, assuming $\boldsymbol{n}$ survey results that we have collected, and regardless of the size of $\boldsymbol{n}$, from regression analysis, $\boldsymbol{\beta_1} = \mathbf{0}$ and $\boldsymbol{R^2} = \mathbf{0}$ in expectation.

From this simple counter example, we see that the measurement of construct B should be completely determined by construct A, which means the "relationship" is strong. But, instead, from the traditional statistical measurements, it does not tell a certain result. Therefore, in the next section, we introduce a new method adopting information theory that tries to measure the "relationship" between two constructs.

2.4. Defining Information Theoretic Measure of Significance

We define an information theoretic measure for two random variables (which is in fact the empirical distributions of the collected surveys in this case) as follows:

Definition 5. (Information Theoretic Measure of Significance) Given two discrete random variables (X, Y) with the joint probability mass function $p(x, y)$, marginal probability mass function $p(x)$ and $p(y)$. The information theoretic measure of significance / determination $s(X, Y)$ is defined by

$$\boldsymbol{s(X,Y)} = \frac{\boldsymbol{I(X;Y)}}{\boldsymbol{H(X,Y)}}. \tag{11}$$

If $\boldsymbol{s(X,Y)} > \mathbf{0.05}$, we say $\boldsymbol{X}$ and $\boldsymbol{Y}$ are significant information theoretically.

From the above definition, we see that $\mathbf{0} \leq \boldsymbol{s(X,Y)} \leq \mathbf{1}$, and $\boldsymbol{s(X,Y)} = \mathbf{0}$ if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, $\boldsymbol{s(X,Y)} = \mathbf{1}$ if $\boldsymbol{X} = \boldsymbol{Y}$, which shares similar property of the coefficient of determination. The following Figure 3 is one of our experiments that shows the relationship between the measure $\boldsymbol{s(X,Y)}$ and the t-statistics of the regression analysis.

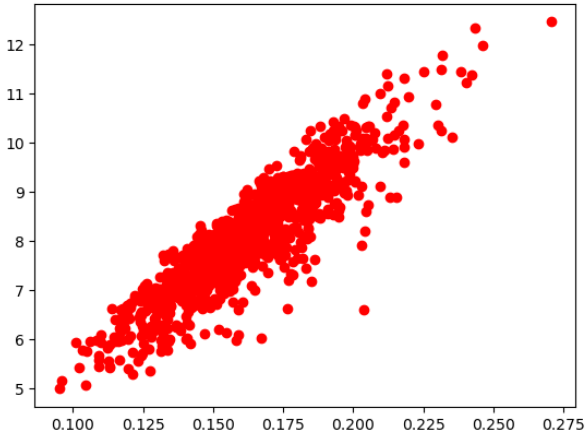

Figure 3 A 7-point Likert scale (values from -3 to 3) survey with 2 questions, which corresponds to construct A and B respectively. The question corresponding to the construct A is generated according to a distribution $\boldsymbol{X} \sim (\mathbf{0}, \mathbf{0.2}, \mathbf{0.6}, \mathbf{0.2}, \mathbf{0}, \mathbf{0}, \mathbf{0})$, where the question of construct B follows $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{N}$, where $\boldsymbol{N}$ is the noise and values -1, 0, and 1 with the distribution $\boldsymbol{N} \sim (\mathbf{0.3}, \mathbf{0.4}, \mathbf{0.3})$. For each round, 100 samples are generated according to the above setting. 1000 tests were conducted, where x-axis and y-axis corresponds to $\boldsymbol{s(X,Y)}$ and the t-statistics (significance) of the regression analysis.

3. RESULTS

In this section, survey data is used to test the entropy method. The data include two variables: powerlessness and workplace deviance (in this study, we focus on organizational deviance rather than interpersonal deviance). Powerlessness, defined as a lack of autonomy and participation, is a component of job insecurity [9]. Being disempowered, employees are more likely to resort to violence as a way of capturing some influence over their environment [10]. Powerless employees tend

to adopt outcome-based moral thinking [11]. To restore a control on a "fair" outcome or maintain gains, they may undertake deviant behaviour such as sabotage or theft toward the organization. The data were collected in two waves, with an interval of three months, from a telecom company in China. The scales of powerlessness (time 1) and workplace deviance (time 2) were adopted from Ashford, Lee and Bobko [9] and Bennett and Robinson [12], respectively.

Results show that the correlation between powerlessness and workplace deviance is just 0.105, and the significance is 0.112. Also, the regression coefficient $\beta_1 = 0.0497$ with p-value 0.113. It means that from the traditional statistical measurements do not give any concrete evidence about the relationship between powerlessness and workplace deviance.

Therefore, the information theoretic measure that defined in Section 2.4. results $s = 0.132 > 0.05$. Therefore, the result is significant using the method in this work.

4. CONCLUSION

Although it has been widely used, regression still has limitations in testing relationships. This study introduces the logics of information theory and employs it to develop a new method. Information can be thought as uncertainty, and entropy measures the amount of uncertainty inherent in random variables. Borrowing the theory, a new method to test the relationship between an independent variable and a dependent variable is established. In a computer-simulated model and a survey-based model, the results prove the equivalence between the entropy method and the traditional regression. However, the entropy method can provide more detail behind the data.

This study extends the methodology by providing an alternative method, while the value and contribution of the entropy method is clearly demonstrated. Future research in business or social science may include the method as a confirmation to the regression results, enhancing examinations on hypothesized relationships. Nevertheless, this study also has limitations as more experiments can be conducted.

AUTHORS' CONTRIBUTION

This work is equally contributed by the three authors.

ACKNOWLEDGMENT

This work was partially supported by Macau University of Science and Technology.

REFERENCES

[1] C. E. Shannon, A mathematical theory of communication. The Bell system technical journal, 27(3), 1948, pp. 379-423

[2] Miller, G. A., & Frick, F. C. Statistical behavioristics and sequences of responses. Psychological Review, 56(6), 1949, pp. 311.

[3] Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological review, 63(2), 1956, pp. 81.

[4] Fisher, B. A. Decision emergence: Phases in group decision-making. Communications Monographs, 37(1), 1970, pp. 53-66.

[5] Ellis, D. G., & Fisher, B. A. Phases of conflict in small group development: A Markov analysis. Human Communication Research, 1(3), 1975, pp. 195-212.

[6] Cappella, J. N. Talk-silence sequences in informal conversations I. Human Communication Research, 6(1), 1979, pp. 3-17.

[7] Cappella, J. N. Talk and silence sequences in informal conversations II. Human Communication Research, 6(2), 1980, pp. 130-145.

[8] Cappella, J. N., & Planalp, S. Talk and silence sequences in informal conversations III: Interspeaker influence. Human Communication Research, 7(2), 1981, pp. 117-132.

[9] Ashford, S. J., Lee, C., & Bobko, P. Content, causes, and consequences of job insecurity: A theory-based measure and substantive test. Academy of Management Journal, 32, 1989, pp. 803-829.

[10] Spreitzer, G. Giving peace a chance: Organizational leadership, empowerment, and peace. Journal of Organizational Behavior, 28, 2007, pp. 1077-1095.

[11] Lammers, J., & Stapel, D. A. How power influences moral thinking. Journal of Personality and Social Psychology, 97(2), pp. 279-289, 2009.

[12] Bennett, R. J., & Robinson, S. L. Development of a measure of workplace deviance. Journal of Applied Psychology, 85, 2000, pp. 349–360.