# A Perspective on Analyzing Ordinal Data using Statistics Analysis and Data Mining Techniques

Aisyah Larasati
Department of Industrial Technology
Universitas Negeri Malang
Malang, Indonesia
aisyah.larasati.ft@um.ac.id

Apif Miftahul Hajji
Department of Civil Engineering
Universitas Negeri Malang
Malang, Indonesia
apif.miptahul.ft@um.ac.id

***Abstract*— Ordinal data is different from interval data since it has unknown absolute distances although the rank order of the level is clearly defined. Numerous model algorithms may be used to analyze ordinal data, such as regression and structural equation model, which both of them represents statistics analyses, and data mining techniques. Each of these algorithms requires different assumptions and has some advantages as well as disadvantages. This study aims to present a perspective on analyzing ordinal data by delivering a comparison between structural equation model and regression (as a representation of statistics analyses) and artificial neural network (as a representation of data mining techniques) when these approaches are used to analyze ordinal data. One aspect to be considered to decide which model used to analyze a problem is characteristics and quality of data set to be analyze, which includes data distribution, data pattern and type of data. The chosen model parameter and the evaluation criteria used are also important.**

***Keywords—ordinal data; statistics analysis; data mining techniques***

## I. INTRODUCTION

Interval data and ordinal data have different characteristics. Interval data has a clearly determined distance between scales. On the other hand, ordinal data has unknown absolute distances although the rank order of the level is clearly defined. Therefore, applying analyses that maintain rank order of data promises a higher capability to detect meaningful trend of explanatory variables on the response variable than other analyses that treats ordinal data as interval data [1, 2]. In line with those studies, Stevens [3] outlines the statistical procedures that are permissible for each type of data. All the permissible statistics for ordinal data is also permissible to interval and ratio data, but not all permissible statistics for interval data or ratio data is permissible for ordinal data. This statement implies that selecting data analysis accordingly to the type of measurement scale assures a study delivering meaningful results. Thus, performing data analysis without considering the type of measurement scale can lead to meaningless results.

Many studies in the field of education, health, behavioral and social sciences have been using ordinal data for decades. In the social and behavioral sciences, an ordinal data is often collected as a result of attitudes and opinions measurements. For example, a question administered to an employee to rate his/her overall job satisfaction using ordered categories such as "strongly dissatisfied," "dissatisfied," "neutral," "satisfied," and "strongly satisfied." The overall job satisfaction measure is an ordinal because an employee who chooses "dissatisfied" feels more negative feeling toward his/her job than if he/she chooses "neutral." Although the distance between "satisfied" and "neutral" cannot be numerically measured, thus certainly cannot be assumed to be equal, the rank-order in this measure is clearly defined.

Statistics analysis, such as Ordinal Logistic Regression (OLR) and Structural Equation Model (SEM), as well data mining technique such as Artificial Neural Network (ANN), are some available models to analyze ordinal data without assuming data as interval scale [4, 5]. Ordinal regression model is an extension of logistic regression that is capable of handling data on an ordinal scale, including investigating the relationship between independent and dependent variables [6]. SEM is also capable of analyzing ordinal data. Although the dependent variables are in ordinal scale, some researchers tend to treat ordinal data as continuous variables and to analyze them using multivariate methods. Thus, when a dependent variable is on ordinal scale, the use of ordinal regression is more appropriate than multiple regressions [7]. Artificial neural networks model is built through an iterative process, in which the model learns the pattern of complex relationships between independent and dependent variables. The interconnection weight in artificial neural networks model is harder to explain than the one in a logistic regression or SEM. If a logistic regression or SEM model only includes a main effect, then each parameter refers to the weight of each predictor variable. Furthermore, this parameter can be statistically tested to examine the significance of each parameter to the model [8]. A perspective on the comparison between OLR, SEM and ANN model provide more perspectives on robustness of each model and its accuracy when the model is used to analyze ordinal data. Thus, this study aims to present a perspective on analyzing ordinal data using statistical analysis, which is represented by ordinal logistic regression and structural equation model, and data

mining techniques, which is represented by artificial neural network.

## II. Statistical Analysis

### A. Ordinal Regression

Regression model is useful method to investigate the relationship between multiple independent variables and dependent variable. This method is useful to determine explanatory variable related to the dependent variable. Moreover, regression also supports the effort to examine the effect of explanatory variables on the dependent variable [9]. The decision to choose multiple regression or logistic regression depends on the measurement scale of dependent variable. When the dependent variable is on continuous scale, linear regression is more appropriate to be used. On the other hand, logistic regression performs better with binary variable.

Most attitude and behavioral studies employs ordinal scale to measure dependent and independent variables. Although the variables are in ordinal scale, some researchers tend to treat them as continuous variables and to analyze them using multivariate methods. However, this approach may lead to bias and misleading result [10]. Thus, when dependent variable is on ordinal scale, the use of ordinal regression is more appropriate than multiple regressions.

Ordinal regression model is an extension of logistic regression that capable to handle data on an ordinal scale. Basically, logistic regression is used to investigate relationship between independent and dependent variable, in which the dependent variable is a binary/dichotomous variable. However, logistic regression can be modified to analyze nominal or ordinal scale data by changing the link function such as change the link function of simple logistic to cumulative logits [11].

Several cumulative link functions that are available to build an OLR model include the cumulative logits, probit, cauchit, complementary log-log, and the related log-log link [1]. The outcome (dependent) variable distribution affects the analyst decision on choosing the type of link function used in the OLR model. The most commonly used link function in the OLR model is the cumulative logit model [12, 13]. The cumulative logit link function is used when an OLR model is applied to the *k* levels of a dependent variable, the model incorporates *k-1* logits into a single model. The effects of the ordinal logistic regression coefficient ($\beta$) are the same for each cumulative logit.

If the distribution of the ordinal data is mainly located on the higher response levels, such as 'very satisfied' on a satisfaction rating, or "extremely agree" on an agreement statement, the complementary log-log link function is the most appropriate link function to be used to build the ordinal logistics regression model [9].

To interpret OLR results, a researcher should consider the signs and coefficients used in the model. The signs represent the existence of negative or positive effects of the independent variables on the ordinal outcome. The coefficient, β, indicates that a one unit change in independent variables results in a change of the odds of the event occurring by a factor of $e^{\beta}$, holding other independent variables as constant [13].

### B. Structural Equation Model (SEM)

Structural equation modeling (SEM) is a type of statistical analysis that is capable of testing a conceptual or theoretical model. Thus, Structural equation model (SEM) is able to examine measurement indicator issues and structural relationship among variables. SEM is also capable to examine relationship among variables that are measured using a single indicator [14]. Several approaches that are commonly used in SEM include confirmatory factor analysis (CFA), path analysis, and latent growth modeling.

The term structural equation model implies that the model contains a combination of two variables: a measurement model and a structural regression model, a model of simultaneous regression equations. A measurement model describes the latent variables in the model using one or more observed variables. On the other hand, the structural regression model presents the causal dependencies between exogenous and endogenous variables, in which each part of the structural equation model is linked simultaneous by regression equations.

SEM has ability to minimized observational error from measurement of latent variables, thus this model is widely applied in the behavioral and social sciences. Observational errors commonly occurred due to the measurement cannot be conduct directly, such as the concept of human intelligence. In the case of human intelligence measurement model, the test items become the observed variables, while the latent variable is the human intelligent itself.

Three main issues on SEM application are: 1) the hypothesized theoretical model as the initial specification model; 2) data screening to use in the model estimation and testing; and 3) the estimation and testing of hypothesized model on empirical data [14].

The more desirable model in SEM is a model with simple structure and no correlated measurement error. Simple structure leads to hold parsimonious principle, while no correlated measurement error represents one dimensional construct measurement, in which each observed variable is related to a single latent variable. Parsimony principle is a principle that explains a simpler a model the better it is. Other aspects to be considered in SEM is number of observable variables should be measured for each latent variable and how it should be correlated each other. Having more indicators for a factor may lead to a lower improper solution and non-convergence estimation problem. Non-convergence means the model is not able to find a specific (simple solutions). However, it may also lead to over-fitted data and the more difficult to obtain a parsimoniously model [14]. Over-fitted data leads to inflexible model to be applied to other group of similar data.

Issues related to data screening prior to model estimation and testing is one of the most crucial steps in SEM [14]. Data screening includes checking coding error, the presence of outliers, and fulfillment of multivariate normality assumption. Coding error should be minimized to ensure validity of data used as observed variables. Eliminating the outliers helps a distortion that happens in the causal relationships between endogenous and exogenous variables. Moreover, the violation on normality may affect goodness of fit indices and standard errors.

## III. Data Mining Technique

Data mining techniques have been developed from statistics, artificial intelligence, machine learning, and database research. This development goes along with the development of related tools and software. Data mining techniques is a part of knowledge discovery from database (KDD), which is used to discover a useful pattern from data. Data mining techniques that are adopted from statistics, such as Bayesian decision models. Some others data mining techniques are developed from artificial intelligence, which include decision tree, support vector machines, and artificial neural network. The classification of data mining technique can be different, such as some experts classify artificial neural network as computational intelligence [15].

Artificial neural network model is one of data mining technique. Garver [16] indicated that data-mining techniques has capability to perform better than traditional statistical techniques since data mining techniques are able to mitigate several assumption of statistical techniques, such linearity, multi-collinearity, and normal distributed data. The key of neural network analysis is the way to model complexity and uncertainty.

The simplest form of artificial neural network consists three layers. The first layer comprises one or more neurons that represent independent (predictor) variables, while the output layer contains one or more neurons that are dependent (outcome) variables. The output layer consists of output nodes that represent the models' classification decision and has one node for each output. The hidden neurons in the model connect input and output layer indirectly, and reflect an actual level of uncertainty and complexity in real company operation. Thus, these hidden nodes are called as internal representations. In general, there can be one or more hidden layers between input and output layer [16, 17].

Within an artificial neural network model, a specific activation function is used to connect two layers in the model. The type of the activation function used in the model depends on the outcome range in the output layer. The most common activation function used in an artificial neural network model is the sigmoid activation function, which is similar to the logit function used in the logistic regression model. In addition, other aspects to be considered during the building process are the network architecture and topology, learning algorithm. This study discussing a supervised artificial neural network model, which means this type of model works by training the model using a dependent variable (output layer). In other word, the training process in the model is supervised by the existing of the output layer. This process is similar to that happened in the regression. The most popular learning algorithm within the supervised artificial neural network is the back propagation algorithm [18].

The parameters used in back propagation involve momentum, learning rate and weight decay parameters [19]. The challenges to build ANN model are to make decision in determining all of those aspect that effect on the model accuracy. To overcome ANN model, a sensitivity analysis should be performed. Two others important things to be considered when analyzing ordinal data using artificial neural network are over training and random starting positions. Preventing over training is important to avoid memorization of data and getting result that are not generalized to the population. On the other hand, random starting position has a significant role to influence the confidence of the results [16]. A major limitation of neural network is its incapability to clarify the reasoning resembling a black box. However, it can be minimize by conducting sensitivity analysis and provide number to measure the importance of each neuron in the input layer [16].

## IV. Comparison Between Analyses

A model to study the relationship between independent and dependent variable can be classified as a prediction model. This model depends on the quality set of data employed, model parameter chosen, and the evaluation criteria used [8]. Some of potential methods to study the link between ordinal dependent variable and ordinal independent variables are ordinal logistic regression, artificial neural network, and structural equation model for ordinal data.

In the context of analyzing ordinal data, artificial neural network does not require any presumption on a parametric relationship between independent (attributes) and dependent variable, such as linearity or additive relationship [20]. Ordinal logistic regression cannot deal with multi-collinearity in the predictor variables, since the colinearity can make the model unstable and lead to bias. To deal with colinearity, logistic regression requires the colinier variables to be trimmed or combined into single score [19].

Ordinal logistic regression also has capability to handle non-linear relationship as artificial neural network model, since it incorporate an exponential term on its function. However, logistic regression requires the form of the nonlinearity relationship have been known a priori. Therefore, artificial neural network provides more flexibility nature of model and has higher robustness to model misspecification than ordinal logistic regression [20].

Artificial neural network model is built through an iterative process, in which the model learns the pattern of complex relationship between input and output. The interconnection weight in neural network model is hard to explain as the one in the logistic regression. More hidden

layers used in neural network model create more complex interconnection weight and interdependencies [20]. This drawback can be reduced by conducting sensitivity analysis. On the other hand, the model parameter in logistic regression is easy to be interpreted. If the model only includes main effect, then each parameter refers to the weight of each predictor variable. Furthermore, this parameter can be statistically tested to examine the significance of each parameter to the model [8].

Another potential drawback of artificial neural network model is this model can lead to local minimum error rate since the iteration process depends on the sample used to learn the pattern. Local minimum error rate means the optimal condition (the best goodness of fit) that applies in a certain iteration is not always provides an optimum results in other iteration. Thus, validation set data is needed to decrease these potential weaknesses [20].

## V. CONCLUSION

Ordinal data requires a different analysis technique since it has different characteristics compared to interval data. Some available technique that is capable of analyzing ordinal data are ordinal logistic regression, structural equation model, and artificial neural network. Each of these algorithms requires different assumptions and has some advantages as well as disadvantages.

The aspects to be considered to decide which model used to analyze ordinal data is characteristics and quality of data set to be analyze, which includes data distribution, data pattern and type of data. In addition, the chosen model parameter chosen and the evaluation criteria used are also important to be considered since these factors affect the model goodness of fit. The chosen model parameter include: 1) number of independent variable, variable selection criteria, and the link function chosen in the ordinal logistic regression; 2) number of observed item variable and latent variable as well as the causal path used in the structural equation model; 3) starting point of the iteration, decay rate, learning rate, momentum rate and number of layer in the artificial neural network. The evaluation criteria on choosing the best technique also affects the model goodness of fit. A certain model can be better in one evaluation criteria, but it worse in other evaluation criteria.

In general, the number of predictor variables in all three techniques (ordinal logistic regression, structural equation model, and artificial neural network) has a significant influence on the model accuracy. More predictor variables in the model may increase the prediction accuracy, however it may also cause over fitting.

## REFERENCES


[1] A.Agresti. *Analysis of ordinal categorical data* (Second ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.2010.

[2] S. Jamieson. Likert scales: how to (ab)use them. *Medical Education, 38*(12), 1217-1218. 2004.

[3] S. S. Stevens. On the Theory of Scales of Measurement. *Science, 103*(2684), 677-680. 1946.

[4] W.J. Deng., W.C. Chen, & W.Pei. Back-propagation neural network based importance-performance analysis for determining critical service attributes. *Expert Systems with Applications, 34*(2), 1115-1125. 2008.

[5] C. Jianlin,W. Zheng, W., & G. Pollastri. (2008, 1-8 June 2008). *A neural network approach to ordinal regression.* Paper presented at the Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). 2008.

[6] C.Lawson, & D. C. Montgomery. Logistic regression analysis of customer satisfaction data. *Quality and Reliability Engineering International, 22*(8), 971-984.2006.

[7] N. Lundahl, F. Vegholm, & L. Silver. Technical and functional determinants of customer satisfaction in the bank-SME relationship. *Managing Service Quality, 19*(5), 581 - 594.2009.

[8] S. Dreiseitl, & L. Ohno-Machado. Logistic regression and artificial neural network classification models. *Journal of Biomedical Informatics, 35*(5-6), 352-359. 2002.

[9] C.-K.Chen, , J. Hughes. Using ordinal regression model to analyze student satisfaction questionnaire. *IR Applications, 1*, 1-13.2004.

[10] N. Lundahl, F. Vegholm, & L. Silver. Technical and functional determina nts of customer satisfaction in the bank-SME relationship. *Managing Service Quality, 19*(5), 581 - 594.2009

[11] C. Lawson, & D.C. Montgomery. Logistic regression analysis of customer satisfaction data. *Quality and Reliability Engineering International, 22*(8), 971-984.2006.

[12] C. C.Clogg, & E. S. Shihadeh. *Statistical model for ordinal variables*. Thousand Oak, California: Sage Publication, Inc.1994.

[13] A. S. Fullerton. A Conceptual Framework for Ordered Logistic Regression Models. *Sociological Methods & Research, 38*(2), 306-347. 2009.

[14] H. Baumgartner, & C. Homburg. Applications of structural equation modeling in marketing and consumer research: A review. *International Journal of Research in Marketing, 13*(2), 139-161. 1996.

[15] R. Mikut, and M. Reischl. *Data mining tools*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, *1*(5), 431-443.2011.

[16] M. S.Garver. Using data mining for customer satisfaction research. *Marketing Research, 14*(1), 8-12.2002.

[17] R. S. Behara, W.W. Fisher, & J. G. A. M. Lemmink. Modeling and evaluating service quality measurement using neural networks. *International Journal of Operations & Production Management, 22*(9/10), 1162.2002.

[18] J.V. Tu. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology, 49*(11), 1225-1231.1996.

[19] K. B.Detienne, D.H. Detienne, & S.A. Joshi. Neural networks as statistical tools for business researchers. *Organizational Research Methods, 6*(2), 236-265. 2003.

[20] P.M. West, P.L. Brockett, & L.L. Golden. A comparative analysis of neural networks and statistical methods for predicting consumer choice. *Marketing Science, 16*(4), 370-391.1997.