JUST AGRICULTURE
multidisciplinary e-Newsletter

Role of Regression Modeling in Agricultural Research in General

Karukumalli Sindhura*, Meena Yadav[1] and Pusala Srujana[2]
*PhD scholar, Division of Agricultural Extension & Communication, SKUAST-Jammu
[1]PhD scholar, Division of Soil Science &Agricultural Chemistry, SKUAST-Jammu
[2]MSc scholar, Division of Agricultural Extension & Communication, SKUAST-Jammu

ARTICLE ID: 042

Introduction

Statistics is a science that has its own methodology, which is used in research and analysis of phenomena around us. Phenomena are in fact subjects of statistical research. The subjects of statistical researches are mass phenomena that are inherently variable, and should be viewed in a number of cases and based on these observations there are made conclusions. Therefore, the statistics are often defined as a scientific method of quantitative research of mass phenomena (Zizic *et al*., 1996).

Agriculture has always been one of the vital occupations that serve mankind, both in terms of livelihood and employment. Due to the substantial increase in the population, the nutritional status of the poor is growing bad, which must be improved. The major effect of population increase has been prominently shown on the environment, the damage of which is increasing rapidly, which ultimately hinders agricultural production. Studies show that the modern techniques used in agriculture have not been environment-friendly, though they are technologically advanced than the primitive techniques. The past achievements in the field of agriculture clearly depict the power and ability of man being able to meet the agricultural demand in spite of the population growth. Hence, a balanced relationship between the nature's major creations i.e., human beings and their environment has to be maintained in order to lead a sustainable life.

Some of the modern methods of statistical analysis were grounded for more than a century. Statistical methodology consists of four main phases: data collection, presentation, analysis, and interpretation of numerical data/method are a manner to decipher the facts and their interpretation. As a scientific method, the statistics origins simultaneously from Germany and England in the seventeenth century. In Germany, the task of statistics was based primarily on the description, whereas in England began introducing the mathematical

JUST AGRICULTURE
multidisciplinary e-Newsletter

processing of data in order to detect regularities in the behaviour of the observed phenomena. The application of statistical methodology allows, not only to detect the general characteristics of variable phenomena, but also to detect regularities in the tendency of these phenomena. The task of statistics is to make a selection of those statistical methods appropriate to the nature of economic processes and phenomena. Statistical survey method phenomena can be divided into two main groups. First method involves the collection, sorting, presentation and describing the basic characteristics of statistical data series, and it belongs to the domain of descriptive statistical analysis. The second group consists of methods of statistical analysis dealing with relationship issues, links (correlation), and conclusions on the basis of sample. These methods are included in the area of analytical statistics, although there is no strict underlined line between these two groups of methods.

Regression analysis

Regression analysis and correlation examine the connection of dependent and independent variables, except that regression analysis in addition to establishing the existence of links between two or more variables also achieves prediction of change in dependent variable regards to changes in the independent variables. In other words, by use of the regression analysis the type and intensity of the connections are defined, as well as the quantity of change in one variable in relation to a unit change of the second variable. The regression is analysis of causality. How we can get to the cause of what is now emerging as a consequence.

In regression analysis, there are at least two variables: the criterion (dependent) variable and the predictor (independent) variable. The dependent variable is a variable of interest, or variable that we want to examine, explain, and predict. Independent variable attempts to explain the dependent variable. When there is only one independent variable, then we are talking about a simple regression, and when there are several independent variables, then we are talking about multiple regression. The goal of regression analysis is to estimate the regression model that minimizes the total distances of the dependent variable from the regression line (Horvat and Mijoc, 2012).

Multiple Regression

When the dependence of a phenomenon with two or more independent variables is examined, then we talk about multiple regressions.

JUST AGRICULTURE
multidisciplinary e-Newsletter

Name of multiple linear regression means:

1. multiple –there are more independent variables X;
2. linear –the regression function is linear by the coefficients B1, B2, ... Bn;
3. regression –the regression function as the best prediction of Y based on Xi, i = 1, .., n is used. The goal of multiple regressions is to determine the intensity of the relationships as more independent variables and the dependent variable.

One of the assumptions for the use of multiple regression analysis is the existence of a linear dependence between the variables. Multiple regression has the form Y = A + B1X1 + B3X3 + b2x2 + ... + BnXn, where Y is the dependent variable, and the X1, X2, X3, Xn independent variables, and it is represented by Venn diagram (Figure 1). Understanding the correlation can be easily explained by Venn diagrams. The overlaps that can be observed (interference field) are indicating a correlationbetween the variables and indicate which part of one variable explains another variable.What the interfering field is greater, the greater is the level of correlation, or if it is lower, the lower the correlation is.

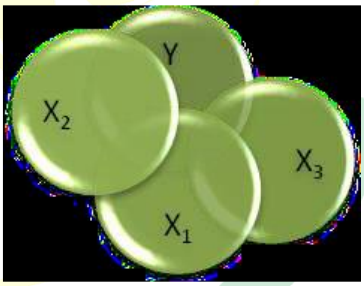

Figure 1: Venn diagram

Multiple regressions provide answers to the following questions:

- How well the independent variables can explain the dependent variable (R2)?
- What is the relative importance of each independent variable in explaining the change in the dependent variables (beta coefficients), under the condition that there is no significant multi co linearity?
- What change of the dependent variable is expected for each unit change of each independent variable (which shows the simple correlation coefficients)? If X1 increases by one unit, what is the expected change in Y (response gives B1) assuming that does not change the impact of other explanatory variables X2, X3, ..., Xn?

JUST AGRICULTURE
multidisciplinary e-Newsletter

An important goal while creating a regression model is to explain the largest possible percentage of the dependent variable. In simple regression, also multiple regression, the percentage of the explained dependent variable is labelled as R2 (R –squared) and it tells what percentage of the dependent variable is explained by the included independent variables. Which variables will be taken as predictor variables or independent variables depends on the aim of survey, as well as on the data availability. When creating the model, it is important to examine the problem of multi–co linearity between independent variables. Namely, if there is a problem of multi–co linearity, then small changes on independent variables can cause large changes in the dependent variable, which does not provide a realistic picture of the model.

There are several statistical indicators that examine multi–co linearity: correlation coefficients between the independent variables, Eigen values and condition index, tolerance, VIF, etc. Eigen values shows how many different dimensions among the independent variables, while condition index calculated as the square root relationship largest Eigen values and Eigen values for a given independent variable. If there are multiple independent variables in which the Eigen value close to 0, there is probably a problem of multi–co linearity. It is similar if the condition index greater than 15, if it is greater than 30, then multi–co linearity is a serious problem in the regression model.

Tolerance and VGF (variance growth factor)10 are associated indicators, because the one is calculated on the basis of another. Tolerance shows a part of the variance of the independent variable which is not included through other independent variables. The high level of tolerance, over 0.8, means that the variable is relatively uncorrelated with other variables. The low level of tolerance, up to 0.2, is indicating high multi–co linearity and that the variable contributes little to explaining the dependent variable in the model. Tolerance level: 1 - R2i VGF is a reciprocal of the tolerance; high values indicate that there is very little contribution to the model.

One of the solutions for the problem of multi–co linearity in the model is the use of stepwise regression. Multiple regressions provide a model that includes all variables with which the analysis was initiated, regardless of their different importance, and also in the case when there is large multi-co linearity. Stepwise regression allows solving the problem of multi–co linearity, with independent variables which are of little importance. The Stepwise

JUST AGRICULTURE
multidisciplinary e-Newsletter

regression allows eliminating the variables that overlap with each other, and therefore they have little or no contribution to the prediction accuracy of the model.11 The rating of model is achieved by using analysis of residuals. The residuals represent the difference between the prediction and the original variables, i.e. the part of the dependent variable which failed to be explained by the independent variables. A good model exists when the difference between the sums of squares model and the residual sum of squares (F test) is higher possible (Gegaj, 2011).

References

Gegaj, P. (2011) Methodology of Social and Political Sciences, Department of Political Science, Podgorica

Horvat, J., Mijoc, J. (2012): Fundamentals of Statistics; Ljevak doo, Zagreb

Vukotic, V. (2002): Statistical Analysis of Labour Productivity, University word, Titograd.

Zizic, M., Lovric, M., Pavlicic, D. (1996): "Methods of Statistical Analysis", Faculty of Economics, Belgrade

Zvizdojević.J&Vukotic,M (2015) Application of statistical methods in analysis of agriculture - correlation and regression analysis, *Ag*