Some Malpractices on Regression Analysis

Epimaco A. Cabanlit, Jr.*

* Mathematics Department, Mindanao State University, General Santos City, Philippines

DOI: 10.29322/IJSRP.11.05.2021.p11342
http://dx.doi.org/10.29322/IJSRP.11.05.2021.p11342

Abstract- Regression analysis is the most frequently used statistical tool or treatment in Education and Social Sciences, especially when the conceptual framework is on causes and effects. As seen in many theses and dissertations results, regression analysis has been abused in the sense that the basic assumptions are not justified. Moreover, results show unrealistic outcomes and even to a point where no sense of reality has been arrived, where the authors and advisers have not realized. Some of these illustrations with anonymity are presented in this paper. We have made some recommendations to the Commission on Higher Education, Philippines, with the hope that such practices be terminated. Lastly, we believe that these are not practiced alone in the Philippines and we hope that our recommendations will be adapted in other countries.

Index Terms- Regression analysis, Regression coefficients, Normality, Homoscedasticity, Error distribution and Commission on Higher Education, Philippines.

I. Introduction

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. https://www.statisticssolutions.com/assumptions-of-multiple-linear-regression/.

In regression analysis there are four basic assumptions that must be justified for purposes of prediction. They are the following: (http://www.duke.edu/~rnau/testing.htm) Linearity of the relationship between dependent and independent variable, Independence of the errors (no serial correlation), Homoscedasticity (constant variance) of the errors, and lastly, Normality of the error distribution.

In practice, particularly in some theses and dissertations, the authors and advisers do not present the justification of the four basic requirements. Sad to say in an investigation being conducted, models which are supposed to be outcomes of these analyses show an unrealistic results and interpretations.

In some instances, even if the problem is on finding out if there is a significant relationship or effect between and among the independent variables and the dependent variable, regression analysis is carried out. These do not require a mathematical model to be framed at. A simple correlation analysis can solve this problem.

The paper shows some of the evaluations of these outcomes.

II. Some Malpractices and the Unrealistic results

We present three cases from three universities.

The first case is all about the study of one dependent variable and four dependent variables. The problem seeks if there is a significant effect of the independent variables to the dependent variable. The independent variables are measured by five-point scale with 1 as the lowest and 5 as the highest, while the dependent is measured by a five-point scale with 1 as the lowest and 5 as the highest. For anonymity, we do not include the values of Multiple R, R-square and the F-value.

This publication is licensed under Creative Commons Attribution CC BY.

The paper does not show the results of the basic four requirements/assumptions in regression analysis.

Table 1 shows the regression result of the study. The author's interpretation from the table implies that x_2, x_3,and x_4 significantly affect the dependent variable.

The author does not even mind to interpret the meaning of the positive and negative signs of the regression coefficients. If one has to examine the contents of Table 1, the model that can be derived is $y = a + cx_2 - dx_3 + ex_4$, where y is the dependent variable, x_2, x_3 and x_4 are the significant predictors. If $x_2 = x_3 = x_4 = 1$, the dependent variable y<1, which is out of the range for the dependent variable. The dependent variable has the minimum value of 1. If one has to consider predicting the dependent variable with this regression model, we would expect of a value outside of the range.

Table 1
Regression Result between Dependent and Independent Variables (University A)

Predictors	Coefficients	t-value	p-value	Remarks
Constant	a	f	<0.05	Significant
x_1	-b	-g	>0.05	Not significant
x_2	c	h	<0.05	Significant
x_3	-d	-i	<0.05	Significant
x_4	e	j	<0.05	Significant

This is a practice wherein there is no proper investigation and proper scrutiny of the model.

The main problem of the paper is to look for the significant relationship between the dependent and independent variables. A good approach for this case is to simply employ the simple correlation analysis, instead of using the regression analysis. The presented results in the table can even mislead the reader.

The second case is about one dependent variable and one independent variable. Table 2 shows the results. For anonymity we do not include Multiple R, R-square, and the F-value.

The author claims that the model is given by $y = k + lx$, where y is the dependent variable and x is the independent variable. He stressed that for every increase in the independent variable the dependent variable will increase by l starting with k.

Table 2
Regression between Dependent and Independent Variables (University B)

Predictors	Coefficients	t-value	Significant
Constant	k	O	>0.05
X	l	P	<0.05

The author does not notice that the constant is not significant, meaning it is already zero. The model should just be $y = lx$. Basing from the original data and for the purpose of anonymity, we use characters, it is given that the minimum average value of the dependent variable is m. The minimum annual average value of the independent variable is $\overline{m}$. Thus, the expected minimum value of the dependent variable is $\widetilde{m}$, a very far value from the expected one. Here $|\widetilde{m} - m| > 500$,. It is also given that the maximum average value of the dependent variable is M. The maximum average value of the independent variable is $\overline{M}$. Thus, the expected maximum average of the dependent variable is $\widetilde{M}$. Here $\left|\widetilde{M} - M\right| > 500$. This makes the forecasted value to be erroneous.

The third case is seen in Table 3 The authors claim that the dependent variable is a function of all the five indicators, but as seen in the table only two indicators are significantly acceptable. This shows that the authors just performed the regression analysis and claim that all the independent variables are significant indicators. It is better if the authors employ a step-wise regression analysis.

This publication is licensed under Creative Commons Attribution CC BY.

Table 3
Regression Result between Dependent and Independent Variables (University C)

Indicators	Extent of Relationships				
	Beta	t-value	p-value	Remarks	Decision
x_1	q	t_1	<0.05	S	Reject H_o
x_2	-r	$-t_2$	>0.05	NS	Accept H_o
x_3	-s	$-t_3$	>0.05	NS	Accept H_o
x_4	t	t_4	>0.05	NS	Accept H_o
x_5	u	t_5	<0.05	S	Reject H_o
Model: $y = qx_1 - rx_2 - sx_3 + tx_4 + ux_5$					

These are some lapses on the authors and advisers by not examining the basic assumptions in regression analysis. If any of these assumptions is violated, then the forecasts, confidence intervals and economic insights yielded by a regression model may be (at best) inefficient or (at worst) seriously biased or misleading (http://www.duke.edu/~rnau/testing.htm) .

III. Concluding Remarks

The above scenarios pose for a further and thorough review and investigation. However, there are no existing preventive measures on how to control such malpractices. If such practices could not be prevented, a chain effect is highly expected. We will continue producing theses and dissertations with no proper directions, resulting to erroneous and unrealistic models.

The sad thing about this is that we will be giving wrong models for the policy makers in the government just in case these theses and dissertations are used. On the other hand, copies of these theses and dissertations are just dusted in the library shelves. This is not the priority of the government. So, we are very safe that this will not be used by these people in the government. This is not the main point. The point is that we are producing misleading theses and dissertations year after year.

To minimize this prevailing problem, it is suggested that there must be a thorough review on theses and dissertations. The Commission on Higher Education (CHED) will initiate and implement this by conducting a public review on theses and dissertations before they will be bind. There must be a clearance or permit from the CHED that said theses or dissertations are ready for binding. The panel shall be composed of professors of the graduate schools in the region. There must be at least one prominent practicing statistician.

Acknowledgment

The author is grateful to the valuable suggestions and comments of Dr. Aldwin Teves and to the participants attending to the presentation of this paper in various conferences.

References

DUDEWICZ, E.J. and MISHRA, S.N. (1988). Modern Mathematical Statistics. John Wiley & Sons, Inc., USA.

HOGG, R.V. and CRAIG, A.T. (1970) Introduction to Mathematical Statistics. Mcmillan, London.
http://www.duke.edu/~rnau/testing.htm

https://www.statisticssolutions.com/assumptions-of-multiple-linear-regression/.

WALPOLE, R.E. (1982). Introduction to Statistics, 3rd ed., Macmillan Publishing Company, New York.

Author

Author – Epimaco A. Cabanlit, Jr. is a Ph.D. in Mathematical Sciences, Professor in Applied Mathematics, Mindanao State University, General Santos City, Philippines. e-mail address: maco_727@yahoo.com.

Correspondence Author – Epimaco A. Cabanlit, Jr., e-mail address: maco_727@yahoo.com.

This publication is licensed under Creative Commons Attribution CC BY.