Original article

Interaction between numerical variables in regression model, and its graphical interpretation

Handan Ankaralı[1], Özge Pasin[2], Senem Gönenç[3], Abu Kholdun Al-Mahmood[4]

Abstract

Background and aim: One of the most important steps in the models to be established in order to define the relationships between the measured characteristics in health field research is to define the effects in the model correctly. One of these effects is the interaction effect, which is rarely used in practice, but on the contrary, should be required frequently used. The aim of this study, the concept of interaction between numerical independent variables, which is rarely used because it is little known in regression-like models, is to be presented with an easily interpretable graphical result. ***Materials and methods*****:** The data used to emphasize the interpretation and importance of the interaction in the regression model were produced by simulation based on the descriptive statistics and distribution patterns of the data in a real study. The data set includes the systolic blood pressures (SBP) of 167 people aged between 40 and 76 and a body mass index between 21 and 52. Age and body mass index were defined as independent variables and SBP as dependent variables. ***Results:*** In the model without interaction, it was observed that an increase in body mass index increased SBP when age was kept constant, and an increase in age increased SBP when body mass index was kept constant. Although this result is sufficient, appropriate, and meaningful for the practitioner, it will not make sense without knowing the importance and meaning of the interaction between body mass index and age. When the interaction term was added to the model, it was seen that the above described result could lead to an invalid and erroneous interpretation. It was seen that the effect of a 1-unit change observed in body mass index on SBP differs at various ages and the effect of a 1-year increase in age on SBP differs according to body mass index values. In this case, it has emerged that a physician who will make a clinical decision should also consider age when deciding on SBP according to body mass index (or vice versa). In addition, the contour graphic method, which will facilitate the work of the practitioners in the interpretation of the interaction, will make a significant contribution to the evaluation of this term in models. ***Conclusion:*** Using an incorrect or incomplete model in data analysis results in, erroneous or incomplete results. The modeling process in healthcare research involving complex relationships requires substantial knowledge, domain knowledge, modeling knowledge, and accurate interpretation of results. Examining the interaction terms is of great importance in the modeling process. If this effect is significant, the actual effects of the interacting effects are meaningless and their interpretation will yield erroneous results.

Keywords: Regression model; interaction; contour chart; independent variable

Bangladesh Journal of Medical Science Vol. 22 No. 01 January’23 Page : 189-194
DOI: https://doi.org/10.3329/bjms.v22i1.63078

Introduction

The mathematical equivalent of the hypotheses established in analytical research is the model. Models are a kind of mathematical expression of the subject to be investigated. Therefore, the correct model is synonymous with the correct hypothesis. When statistical models are evaluated in general, model nomenclature and definitions of variables in the model can be summarized as follows. In the established model, the variable on the left side of the equation is called the dependent variable, and the ones on the right are referred to as independent variables, risk factors, and covariates. The aim is to evaluate the effects of the variables on the right side alone or together or to estimate the value of the variable on the left by controlling the effects of

1. Handan Ankaralı, Istanbul Medeniyet University, Faculty of Medicine, Biostatistics Department, Istanbul.
2. Özge Pasin, Bezmialem University, Faculty of Medicine, Biostatistics Department, Istanbul.
3. Senem Gönenç, Ataturk University, Faculty of Science, Statistics Department, Erzurum.
4. Abu Kholdun Al-Mahmood, Department of Biochemistry, Ibn Sina Medical College. Bangladesh.

Correspondence: Handan Ankaralı; Istanbul Medeniyet University, Medical Faculty, Biostatistics Department, Istanbul-Türkiye; e-mail: handanankarali@gmail.com

the variables on the right. Considering the number of variables on the right, these models are classified as simple or multiple models[1]. When the number of variables on the left is one, a univariate model is defined, when there is more than one, a multivariate model is defined[1]. In addition, depending on the type of independent variables in the model, factor, block, or covariate naming is done. The factor and block represent the categorical predictor, while the covariate defines the continuous predictor[2].Besides, while the factor effect is a factor under investigation, the block effect refers to a categorical independent variable that is more or less known but whose effect should be taken into account. The covariate is a variable of numerical type, which is generally known to have an effect but needs to be controlled. The numeric type of independent variable whose effect is to be investigated is usually defined as a numeric independent variable. However, these definitions can be expressed with much more diverse concepts in various fields.

Another important step in the modeling process is the correct definition of the effects in the model in case the number of independent variables to be included in the model is more than one. Also, it is necessary to decide whether to take the interaction effects together with the main effects of the predictors whose effects are being investigated. However, it should be known that this decision is directly related to knowing or understanding very well what the researcher wants in her hypothesis. Generally, interaction terms are not included in regression models because the meaning of interaction is often not well known or it is difficult to interpret the interaction. In addition often, hypotheses can be incomplete or incorrectly constructed because the researcher does not reflect their wishes well or clearly does not know what to examine.

As a general definition, the interaction between any two independent variables, either a categorical or continuous variable, is called first-order interaction and is interpreted as the change of the relationship between one of these variables and the outcome variable according to the values of the other independent variable[3]. Interpreting the interaction is relatively well known and easier when the independent variables are categorical. However, the interpretation of the interaction does not differ according to the type of variables. The main problem is that researchers have problems with what the interactions between continuous variables, what role they play in the model, and how to interpret the results, and accordingly they do not include these terms in their models. When the literature including health field research is examined, it is seen that the models containing the interactions between the numerical variables in the continuous structure are used almost non-existent and there are few methodological studies on the subject[4,5]. On the other hand, in cases where it is desired to examine the relations between variables in the field of health, it is almost always seen that more than one factor and result occur. This situation requires maximum attention in the modeling process of the relations between these variables.

The aim of this study is to describe the interactions between continuous independent variables which are little known and rarely used, in regression-like models, and present the results with an easily interpretable graphic.

Material and methods

Regression models with interaction terms

A regression model with two independent variables is set up as follows[6]:

$$\Delta Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} X_{2i}) + u_i \text{ X}$$

The interaction term is expressed as $(X_{1i} X_{2i})$.

When X_2 is given, the effect of the change in X_1 on the Y is calculated with the help of the following equation and a positive β_3 indicates that the effect of an additional unit increase in X_1 on Y increases linearly with X_2.

$$\frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2$$

Or when X_1 is given, the effect of the change in X_2 on Y is calculated with the help of the following equation.

$$\frac{\Delta Y}{\Delta X_2} = \beta_2 + \beta_3 X_1$$

These statements show that when examining the effects of variables on the outcome in regression models containing more than one variable, it is not enough to obtain the adjusted effects by controlling the effects of other variables, and it may even be misleading. In this case, it is also necessary to test whether the effects of the variables on the resulting change according to the various values of other variables.

Hypothetical data set

The data used for application purposes in the study were produced by simulation based on the descriptive statistics and distribution shape of the data in real research. The dataset includes systolic blood pressure (***Y, SBP***), body mass index (***X_1, BMI***), and age (***X_2***) measurements of 167 people aged between 40 and 76 years who applied to the internal medicine outpatient clinic. Using the available information, the effect of bmı and age on the systolic blood pressure will be investigated.

Data for application were simulated with Minitab macro (ver. 18.0) and Stata (ver. 14.0) was used for data analysis.

- **Ethical clearence:** No material requiring ethical permission was used in the study. The data were generated by simulation.

Results

The effects of age and BMI, which will be included as independent variables in the linear regression model, on the SBP will be investigated. In order to understand the subject, the results of the model without interaction are given in the first stage. Descriptive statistics of the variables included in the model are presented in Table 1, and the model results without interaction terms are presented in Table 2.

Table 1. Descriptive statistics of variables

	N	Mean	Sd	Minimum	Maximum
SBP	167	130,03	19,547	90	210
Age	167	54,92	8,988	40	76
BMI	167	32,93	5,833	21	52

Suppose there is a group of people of the same age but with different BMI values. Logically, as people's BMI increases, we would expect their SBP to increase. In other words, "SBP as BMI increases when age is kept constant".

In addition, in individuals with the same body mass index, SBP is expected to increase as age increases. This interpretation means that when the body mass index is kept constant, SBP increases with age.

When Table 2 is examined, it is observed that the above described results have been achieved. That is, when body mass index is held constant, SBP increases by 0.912 mmhg for each year increase in age. Similarly, when age is kept constant, it is observed that an increase of 1 kg/m2 in body mass index increases the SBP by 1.070 mmHg. These results are defined as adjusted effects in models where multiple factors are investigated.

Table2. Model results without interaction terms

			T	P
	B	**SE**		
Constant	44,738	10,082	4,437	**<0,001**
BMI	1,070	0,219	4,883	**<0,001**
Age	0,912	0,142	6,411	**<0,001**

B: Regression coefficient; **SE:** Standart Error of B

The model whose results are given in Table 2 requires that the effect of increasing body mass index on SBP be constant at all ages. However, this may not be the case. The change in the SBP per body mass index for 40-year-olds may be different for 60-year-olds. In other words, the amount of the relationship between body mass index and SBP may differ in different ages or age groups. It is possible to say this in terms of the relationship between the age and SBP. In order to give the correct answer to this question, the interaction term must be added to the model.

When the interaction term between age and body mass index is added to the model results that do not contain the interaction term in Table 2, the results given in Table 3 are obtained.

When Table 3 is examined, it is seen that the interaction between age and BMI is significant. In this model, the interaction(s) are tested first, and if the result is not found statistically significant, only the main effects of the variables involved in the interaction should be examined. In this direction, when the interaction term is examined, it is seen that it is statistically significant and according to this result, it is inferred that the main effects do not make any sense for the results. In addition, it is seen that the main effects are not significant unlike the ones in Table 2. This result once again emphasizes the importance of accurate modelling. When the term interaction is interpreted, it highlights that the effect of 1-unit change observed in BMI on SBP differs at different ages (P=0.022). Similarly, the effect of 1-year increase in age on SBP appears to differ according to bmı levels. In this case, it would not be correct to explain how a 1-unit change in BMI changes SBP without considering age (main effect of BMI) or to explain how a 1-year change in age changes sbpwithout considering BMI values (main effect of age).

Table 3. Model results with the interaction term

	B	SE	T	P
Constant	149,46	46,48	3,22	**0,002**
BMI	-2,15	1,412	4,88	0,130
Age	1,05	,862	-1,52	0,225
Age x BMI interaction	0,061	,0261	2,31	**0,022**

B: Regression coefficient; **SE:** Standart Error of B

When the regression coefficients of BMI for different ages are estimated, the results given in Table 4 are obtained. When Table 4 is examined, it is seen that a 1-unit increase in BMI at age 40 does not significantly affect SBP, but a 1-unit increase in BMI at other ages significantly increases SBP, and this effect increases with increasing age.

Table 4. Regression coefficients of BMI for different ages

	Coefficients for BMI (B)	SE	T	P
Yaş=40	0,25	0,41	0,61	0,542
Yaş=50	0,85	0,24	3,63	**<0,001**
Yaş=60	1,46	0,27	5,32	**<0,001**
Yaş=70	2,06	0,48	4,29	**<0,001**
Yaş=80	2,66	0,72	3,68	**<0,001**

B: Regression coefficient; **SE:** Standart Error of B

The graphical representation of the results obtained in Table 4 is presented in Figure 1.The interaction term can be interpreted much more simply than the odds ratio or regression coefficient with the this contour plot.

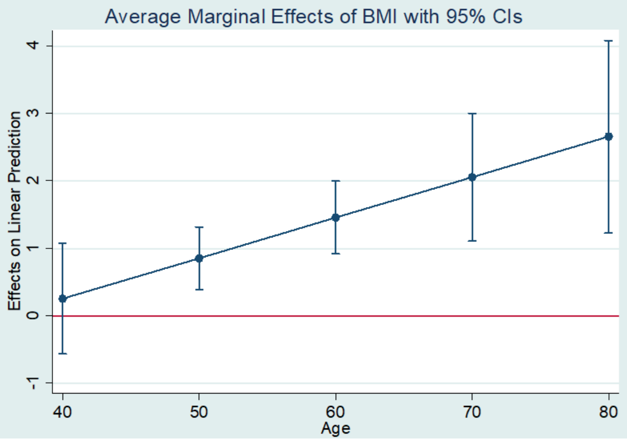

Figure 1. Regression coefficients of BMI for different ages

The graph showed that if the BMI level of 40-year-olds is between 20-50, the SBP will vary between 120 and 140.

However, if the BMI level of 60-year-olds is 35 and above, the SBP will vary between 140-160. In addition, if the BMI level is higher than 40 in 80-year-olds, the SBP is a minimum of 180. The interaction term in our model causes the curvature of the contour lines in the chart. Without interaction, the contour lines will be straight[7].

The variation of corrected SBP predictive values by age for various BMI levels and by BMI for various ages are given in Figures 3a and 3b.When Figure 3 is examined, it shows that the effect of age on sbp increases as BMI increases. Similarly, as age increases, BMI change appears to change sbp even more.

Estimates of SBP for different ages and bmı levels are given in Table 5.

Table 5. Predicted values of systolic blood pressures for different ages and BMI levels

Different subgroups	Predicted systolic blood pressure	Standard error of prediction (SE)
Age =40 ve BMI =21	112,78	5,20
Age =40 ve BMI=31	115,32	2,47
Age =40 ve BMI=41	117,86	4,42
Age =40 ve BMI=51	120,39	8,21
Age =60 ve BMI =21	117,02	3,69
Age =60 ve BMI=31	131,57	1,58
Age =60 ve BMI=41	146,12	2,52
Age =60 ve BMI=51	160,68	5,01
Age =80 ve BMI =21	121,25	9,47
Age =80 ve BMI=31	147,82	4,02
Age =80 ve BMI=41	174,39	6,83
Age =80 ve BMI=51	200,96	13,47

Discussion

Incorrect modeling means misleading results. This means producing false evidence unknowingly. The modeling process in healthcare research involving complex relationships requires substantial knowledge, domain knowledge, modeling knowledge, and accurate interpretation of results. As can be seen from the sample data set, the results of which are given in the findings section, the term interaction offers a different perspectiveondiagnosis and treatment in health research.

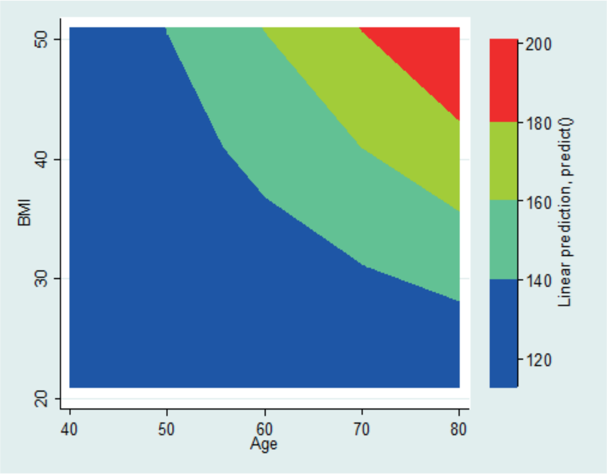

Figure 2. Graphical interpretation of BMI*x*Age interaction

The number of independent variables included in the statistical models and the relationship between these variables and the dependent variable should be determined based on the established hypothesis. Interaction terms are often omitted from the model due to difficulties in interpretation and the often unknown biological meaning. On the contrary, interaction terms are so important that if there is a significant interaction effect in the model, this effect is interpreted. The main effects of interacting variables are not interpreted. In addition, depending on the number of variables whose interaction will be examined, interaction levels are named first-order, second-order, third-order.[6]. However, first-order interaction, which is the simplest interaction, should be evaluated in the process of establishing the model, as it will be the solution to the researched problem in many cases and it is the most easily interpreted interaction term. Third-order or higher-order interactions should be interpreted with the support of experts.The contour graphic proposed to be used in the interpretation of the interaction in this study is suitable for the first-order interaction result[7]. In addition, it can be interpreted by drawing a 3-dimensional contour graphic in second-order interaction. However, a graphical representation is not yet available for higher-order interactions. If there are important interactions for researhers are found, these terms can be interpreted without using graphics.

- **Source of fund: (if any).**

There was no fund received for this work from anywhere

- **Conflict of ınterest:**

The authors declared no conflict of interest

Authors's contribution:

- **Data gathering and idea owner of this study:** HA, MAK
- **Study design:** HA, ÖP, SG
- **Data gathering:** HA, ÖP
- **Writing and submitting manuscript:** HA, ÖP, SG, MAK
- **Editing and approval of final draft:** HA, ÖP, SG, MAK

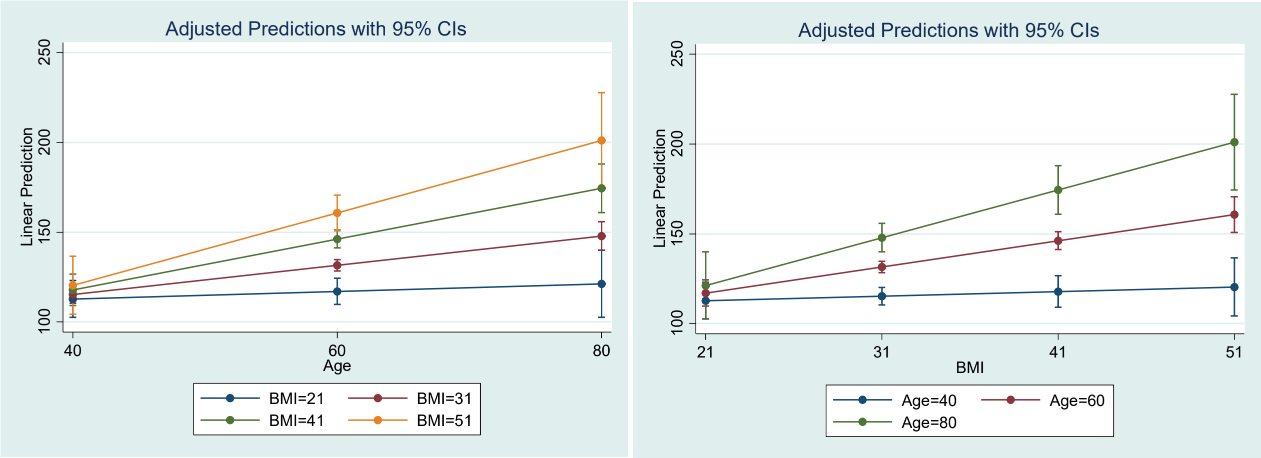

Figure 3. Changes in adjusted SBP predictors (a) by age for various BMI levels, (b) by BMI for various ages

References

1. Grace-Martin, K. Confusing Statistical Term #9: Multiple regression model and multivariate regression model. First Published 4.29.2009; Updated 2.23.2021, https://www.theanalysisfactor.com/multiple-regression-model-univariate-or-multivariate-glm/

2. Schwenke, JR. Comparing the use of block and covariate information in analysis of variance. 1997, Conference on Applied Statistics in Agriculture, 9th Annunal Conference Proceedings. https://newprairiepress.org/cgi/viewcontent.cgi?article=1299&context=agstatconference

3. Rekka, M. Dealing with interaction effects in regression. 2019.https://stattrek.com/multiple-regression/interaction.aspx?Tutorial=reg.

4. Liu, Y., West, SG., Levy, R., Aiken, LS.Tests of simple slopes in multiple regression models with an interaction: Comparison of four approaches. Multivariate Behav Res.2017;52(4):445-464.doi:10.1080/00273171.2017.1309261.

5. Xinhai Li, X., Li, B., Wang, G., Zhana, X. Holyoak, M. Deeply digging the interaction effect in multiple linear regressions using a fractional-power interaction term. 2020, Methods X, doi: 10.1016/j.mex.2020.101067.

6. Jaccard, J and Turrisi, R. Interaction effects in multiple regression (Quantitative applications in the social Sciences). 2003, 2nd Ed.,ISBN-13: 978-0761927426, ISBN-10: 0761927425, Sage Publications, Inc. International Educational and Professional Publisher, London.

7. Rising, B. StataCorp LP, 2011 Stata Conference, Chicago, IL, July 15, 2011.