



MULTIPLE LINEAR REGRESSION ANALYSIS OF FACTORS INFLUENCING THE ABILITY TO USE DATA ANALYSIS SOFTWARE

Naufal Raihan Luthfi Hutagalung^{1*}

^{1*}Universitas Sumatera Utara, Medan, Indonesia, naufalhutagalung17@gmail.com

ABSTRACT

The ability to operate data analysis software is a fundamental competency for statistics students in the digital era; however, varying levels of proficiency are assumed to be influenced by internal factors (knowledge and attitudes) as well as external factors (usage frequency and barriers). This study aims to determine the influence of Data Literacy Knowledge (X1), Attitudes and Perceptions (X2), Usage Frequency and Intensity (X3), and Challenges and Barriers (X4) on the Ability of Statistics Students at Universitas Sumatera Utara to Use Data Analysis Software (Y). Using a quantitative method with an associative approach, the research involved 155 students selected through Stratified Random Sampling, with data collected via an online questionnaire and analyzed using multiple linear regression assisted by SPSS and validated with SmartPLS. The findings show that all four factors simultaneously have a significant effect on students' ability, contributing 66.2% ($R^2 = 0.662$; $F\text{-statistic} = 75.502$), where Knowledge, Attitudes, and Usage Frequency have a positive and significant effect, while Challenges and Barriers have a negative and significant effect, and Usage Frequency and Intensity emerges as the most dominant factor. These results indicate that knowledge, attitudes, and frequency of use positively enhance students' technical competence, whereas technical constraints diminish it, suggesting that improvement strategies should prioritize increasing practice intensity and minimizing technical obstacles.

Keywords: data literacy, statistical software, student ability, usage frequency, regression analysis.

Received: 2 October 2025

Revised: 4 November 2025

Published: 30 December 2025

INTRODUCTION

The ability to use data analysis software is an essential component of statistical learning, as nearly all modern analytical procedures from descriptive analysis to inferential modeling are carried out with the assistance of software tools. In the context of higher education, software proficiency is not only related to the technical skill of executing commands, but also to the ability to understand data structures, select

appropriate analytical techniques, and accurately interpret outputs. Thus, the ability to use data analysis software can be viewed as a concrete representation of students' statistical competence, which strongly determines their readiness to meet both academic and professional demands.

Empirically, this ability does not emerge spontaneously but is shaped by various interacting factors. Data literacy knowledge, attitudes and perceptions toward software, frequency and intensity of use, as well as the challenges and barriers encountered, all have the potential to produce variations in students' abilities. Identifying these factors is essential to explain why some students are able to operate software effectively, while others experience difficulties despite receiving relatively similar instructional exposure. However, without an appropriate quantitative approach, the influence of each factor on the ability to use data analysis software cannot be measured objectively and accurately.

Multiple linear regression analysis provides a relevant statistical framework for examining the simultaneous and partial effects of several independent variables on a single dependent variable within an integrated model. Through this technique, it is possible to assess the extent to which data literacy knowledge, attitudes and perceptions, usage frequency and intensity, as well as challenges and barriers contribute to the ability to use data analysis software, both collectively and individually. The use of multiple linear regression in this study enables the formulation of an empirical model that not only explains the strength of these relationships but also identifies the most dominant factor influencing this ability. Thus, the results of the analysis are expected to offer an evidential basis for formulating more targeted policies and instructional interventions aimed at improving students' software proficiency in data analysis.

METHODS

This study employs a quantitative approach with an associative correlational design, as the primary focus is to analyze the simultaneous and partial effects of several independent variables on a single dependent variable through multiple linear regression modeling. This approach was selected to enable the relationships between data literacy knowledge, attitudes and perceptions, usage frequency and intensity, as well as challenges and barriers, and the ability to use data analysis software to be measured objectively and accurately within a comprehensive statistical model. The research was conducted in the Statistics Study Program, Faculty of Vocational Studies, Universitas Sumatera Utara, targeting students who had used data analysis software in lectures, practicums, and project assignments. All stages of data collection were carried out during an active academic period, ensuring accessible respondents and maintaining the relevance of software usage to ongoing academic activities. Data were collected online using an electronic questionnaire distributed through official communication channels and student groups.

The population of this study consisted of all students of the Statistics Study Program at the Faculty of Vocational Studies, Universitas Sumatera Utara. A total sample of 155 students was determined using the Slovin formula with a 5% margin of error and selected using a stratified random sampling technique based on cohort year. This technique was employed to ensure that each cohort received a proportionate sample, thereby increasing the representativeness of the findings for the overall population. The independent variables included Data Literacy Knowledge (X1), Attitudes and Perceptions

toward Analytical Software (X2), Software Usage Frequency and Intensity (X3), and Challenges and Barriers in Software Use (X4). Data Literacy Knowledge is defined as students' understanding of statistical concepts, their ability to read and comprehend data, and their ability to interpret analytical results. Attitudes and Perceptions describe interest, perceived ease of use, confidence, and perceived usefulness of the software. Usage Frequency and Intensity reflect how often and how long the software is used for assignments, practicums, and independent practice. Challenges and Barriers include difficulties in understanding features, device limitations, lack of guidance, and technical issues. The dependent variable is the Ability to Use Data Analysis Software (Y), reflected through accuracy in performing analyses, ability to interpret output, correctness of result interpretation, and mastery of software features.

Primary data were collected using a Likert scale questionnaire (1–5) developed based on the indicators of variables X1–X4 and Y. The questionnaire consisted of respondent identity information and statements measuring each research construct. Prior to distribution, the instrument underwent content validation by lecturers or experts to ensure clarity, relevance, and appropriateness of the items. Secondary data were obtained from books, scientific articles, and relevant research reports to support the theoretical foundation and result interpretation. Data analysis was conducted through several stages. First, descriptive analysis was used to describe respondent characteristics and the general profile of the research variables. Second, validity testing was performed using Pearson Product Moment correlation and reliability testing using Cronbach's Alpha to ensure that the instrument measured each variable consistently and accurately. Third, classical assumption tests particularly normality tests were conducted to ensure that the data met the statistical assumptions required for multiple linear regression. Fourth, multiple linear regression analysis was employed to examine the simultaneous and partial effects of Data Literacy Knowledge, Attitudes and Perceptions, Usage Frequency and Intensity, and Challenges and Barriers on the Ability to Use Data Analysis Software. The parameters analyzed included regression coefficients, t-values (partial test), F-value (simultaneous test), and the coefficient of determination (R^2). The results were then interpreted to identify the significant factors and the most dominant factor in explaining variations in students' software proficiency.

RESULT AND DISCUSSION)

The multiple linear regression model indicates that the four independent variables (X1, X2, X3, and X4) simultaneously predict the dependent variable Y with varying levels of contribution. Three of the variables have a positive effect, while one variable has a negative effect on Y. Form of the regression equation, From the *Unstandardized Coefficients (B)* column, the resulting multiple linear regression equation is: $Y = 1.667 + 0.453X_1 + 0.174X_2 + 0.463X_3 - 0.169X_4$. This equation implies that when all independent variables (X1, X2, X3, and X4) are equal to zero, the predicted mean value of Y is 1.667. However, from a substantive or practical perspective, this constant often has limited interpretative value if a zero value for all independent variables is unlikely to occur in real-world conditions. Interpretation of coefficients x1-x4, The coefficient (B) for X1 is 0.453 with a significance value (Sig.) of 0.000, indicating that for every one-unit increase in X1, the value of Y increases by 0.453 units, assuming the other variables remain constant. This effect is statistically significant at the 5% significance level. Similarly, the

coefficient for X2 is 0.174 with a significance value of 0.043, which also indicates a statistically significant effect. This means that a one-unit increase in X2 is associated with an increase of 0.174 units in Y, although the magnitude of this effect is smaller compared to those of X1 and X3.

The coefficient for X3 is 0.463 with a significance value (Sig.) of 0.000, representing the largest positive effect in the model. This indicates that for every one-unit increase in X3, the value of Y increases by 0.463 units, and this effect is highly statistically significant. Therefore, X3 can be regarded as the strongest predictor in the model when evaluated based on the unstandardized coefficient (B). In contrast, the coefficient for X4 is -0.169 with a significance value (Sig.) of 0.006, meaning that each one-unit increase in X4 leads to a decrease in Y of 0.169 units. This negative effect is also statistically significant, indicating that X4 functions as a factor that reduces the value of Y. Comparison of effect strength (Beta coefficients), The *Standardized Coefficients (Beta)* column shows that X3 has the largest Beta value (0.444), followed by X1 (0.369) and X2 (0.149), while X4 has a negative Beta (-0.140). This comparison suggests that X3 exerts the strongest influence on Y, whereas X4 has a weaker but significant negative effect.

This indicates that after all variables are standardized to a common scale, X3 is the most dominant variable influencing Y, followed by X1 and X2, while X4 contributes in the opposite direction to Y. Partial significance test (t-test), The calculated t-values for X1 (4.740), X2 (2.043), X3 (7.250), and X4 (-2.786) are each accompanied by significance values (Sig.) below 0.05. Therefore, all four independent variables have a statistically significant partial effect on Y. Meanwhile, the constant has a significance value (Sig.) of 0.244 (> 0.05), indicating that statistically the intercept is not significantly different from zero. However, this does not undermine the main interpretation of the model, as the primary focus of the study lies on the direction and strength of the effects of variables X1–X4 on Y.

Tables

Accordingly, the results presented in the table above can be summarized as follows:

Koefisien Determinasi (R^2):

Multiple Linear Regression | Hutagalung

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.814 ^a	0,662	0,653	2,44154
a. Predictors: (Constant), X4, X3, X2, X1				

Uji Signifikansi Simultan (Uji F):

ANOVA ^a						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1753,702	4	438,425	73,547	.000 ^b
	Residual	894,169	150	5,961		
	Total	2647,871	154			
a. Dependent Variable: Y						
b. Predictors: (Constant), X4, X3, X2, X1						

Uji Signifikansi Parsial dan Koefisien Regresi (Uji t):

Coefficients ^a						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1,667	1,425		1,170	0,244
	X1	0,453	0,096	0,369	4,740	0,000
	X2	0,174	0,085	0,149	2,043	0,043
	X3	0,463	0,064	0,444	7,250	0,000
	X4	-0,169	0,061	-0,140	-2,786	0,006
a. Dependent Variable: Y						

CONCLUSION

The regression results indicate that X1, X2, X3, and X4 have a significant effect on Y. Variables X1, X2, and X3 exert a positive influence, whereas X4 has a negative effect. Among these variables, X3 is the most dominant factor influencing Y, followed by X1 and X2. Recommendations, It is recommended to prioritize improvements in X3 while continuing to give attention to X1 and X2, as they have a positive influence on Y. Meanwhile, the negative effect of X4 should be properly managed or controlled. Future studies may incorporate additional variables in order to produce more comprehensive and robust analytical results.

BIBLIOGRAPHY

- Brezavšček, A. Š. (2017). Factors Influencing the Behavioural Intention to Use Statistical Software: The Perspective of the Slovenian Students of Social Sciences. *EURASIA Journal of Mathematics, Science and Technology Education*, 953–986.
- Cipta, E. S. (2025). Optimizing statistics education: Analyzing students' perceptions of SPSS usage. *Journal of Honai Math*, 297-308.
- Ghozali, I. (2018). *Aplikasi Analisis Multivariate dengan Program IBM SPSS 25*. Diponegoro: Badan Penerbit Universitas Diponegoro.
- Handayani, L. (2023). Pemanfaatan Python dalam Pembelajaran Statistik Mahasiswa. *Jurnal Informatika dan Statistik*, 36-38.
- Hasabo, E. A. (2022). Statistics for Undergraduate Medical Students in Sudan: Usage and Associated Factors for Statistical Analysis Software. *BMC Medical Education*.
- Khairunnisa, & Y. (2025). Factors Affecting Students' Statistical Literacy Ability in Indonesia (Meta-Analysis). *MATHEMA: Jurnal Pendidikan Matematika*, 7(2).
- Lukma A, P. S. (2025). Pelatihan SPSS sebagai upaya optimalisasi literasi statistik mahasiswa. *Jurnal JPM Widina*, 56-63.
- Pasaribu, R. (2022). Pemahaman mahasiswa terhadap literasi data dan kebutuhan penggunaan SPSS/Excel dalam penelitian. *Jurnal Literasi Sains dan Pendidikan*, 290-295.
- Prihastari, E. B., Hidayah, I., Masrukan, M., & Susilo, B. E. (2023). Analisis literasi statistik pada mahasiswa PGSD dalam mata kuliah statistik pendidikan. *Jurnal Basicedu*, 671-680.
- Silalahi, E., & Limbong, B. (2021). Analisis pemahaman literasi data mahasiswa PGSD dalam merepresentasikan data software Excel dan SPSS pada mata kuliah statistik. *Mutiara Pendidikan*, 60-65.
- Silalahi, T. M., & Limbong, W. S. (2021). Analisis pemahaman literasi data mahasiswa PGSD dalam merepresentasikan data software Excel dan SPSS pada mata kuliah statistik. *Jurnal Mutiara Pendidikan Indonesia*, 57-65.
- Sudrajat A, H. L. (2024). Pelatihan penggunaan aplikasi SPSS dalam pengolahan data penelitian. *Cendekia: Jurnal Pengabdian Masyarakat*, 6014-6018.
- Suryati E, K. D. (2023). Efektivitas penerapan pembelajaran statistika berbantuan aplikasi SPSS terhadap hasil belajar mahasiswa. *Jurnal Paedagoria*, 447-455.
- Syahrudin, A. Z., Ilham, H. N., V, M., & Mappanyompa. (2025). Peningkatan Kemampuan Literasi Akademik melalui Pelatihan Penulisan Artikel Ilmiah Menggunakan

ChatGPT dan VOSviewer. Jurnal Kegiatan Pengabdian Mahasiswa , 43-52.

Wulandari, T. (2021). Analisis Kesiapan Mahasiswa dalam Menggunakan Software Statistik. Jurnal Ilmu Pendidikan, 60-62.