The advent of ChatGPT: Job Made Easy or Job Loss to Data Analysts

ABIOLA TIMOTHY OWOLABI*, OLUWASEYI OLUWADAMILARE OKUNLOLA,
EMMANUEL TAIWO ADEWUYI, JANET IYABO IDOWU,
OLASUNKANMI JAMES OLADAPO
Department of Statistics, Ladoke Akintola,
University of Technology, Ogbomoso,
OYO STATE,
NIGERIA

Corresponding Author

Abstract: - Artificial Intelligence (AI) has proven valuable in almost every field of endeavour, including education, sciences, engineering, technology, medical sciences, and numerous other areas of application. Despite its widespread usefulness, concerns have arisen about AI potentially displacing jobs due to its highly advanced capabilities, commonly called the "god-man effect." One remarkable AI product is ChatGPT, a Chatbot developed by an open AI company in the USA that is capable of engaging in conversations that resemble human interactions. This study explores the strengths and limitations of ChatGPT for data analysis, with the primary objective of assessing whether ChatGPT poses a threat to the job of data analysts. An econometric dataset with a sample size of thirty (30), which consists of one dependent variable and three independent variables, was simulated. The dataset was intentionally generated with issues like multicollinearity, outliers, and heteroscedasticity. Subsequently, multiple tests were conducted on the datasets to confirm the presence of these problems. The ChatGPT 3.5 and 4.0 versions were then used to analyse the data to examine this chatbot's prowess in performing data analysis. ChatGPT 3.5 and 4.0 accurately predicted the suitable statistical tool for analyzing the simulated datasets. Both versions of ChatGPT emphasized that the expertise of a professional data analyst would be necessary. While they could offer guidance on data analysis, they cannot perform the analysis themselves as they are solely AI models. ChatGPT can help with what to do next when a data analyst gets stuck. However, they should not be recognized as an authority in making statistical decisions. Therefore, ChatGPT may not replace data analysts but could make their job easier by serving as a helpful resource to turn to when they encounter challenges.

Key-Words: - Artificial Intelligence, ChatGPT, Data Analysis, Simulation, Data Analysts.

Received: August 14, 2023. Revised: December 11, 2023. Accepted: February 16, 2024. Published: April 4, 2024.

1 Introduction

There is no doubt that the world is currently agog with the introduction of Artificial Intelligence (AI). Artificial intelligence has been found useful in almost every aspect of human life. During the pandemic, AI technology was used to deliver food items to people indoors. It has also been found helpful in many other areas of life like health, sports, and academics. Though helpful, some people are worried that their jobs might be at stake. It has been projected that automation will lead to 400 and 800 million jobs by 2030, [1]. This is because AI is faster than humans and even more accurate in most cases. They do exactly what they are instructed to do without getting tired or frustrated, even answering programming questions, [2]. This ability to process a lot of information with little or no mistakes has given AI an edge in the job recruitment world. An example of AI technology currently making waves all over the world is the ChatGPT, which was released to the public in late 2022, [3].

ChatGPT is a chatbot developed by the Open AI Company in the USA. The bot is designed to be able to hold human-like conversations. That is, it can converse with humans just like a human being. The ChatGPT was trained based on datasets from various sources, including books, websites, research articles, theses, newspapers, and several other sources. This dataset contains about 175 billion parameters. However, no information is

available to the public on the quality of data obtained from those sources, [4].

When asked questions, the ChatGPT responds by analysing the dataset earlier discussed using the concept called Natural Language Processing (NLP), [4], [5]. NLP is a technique of analysis in which information relevant to the questions being asked is extracted from a dataset. Utilising this technique, the ChatGPT may seem to be a question god who has mastered everything. That may not be true; it is merely working on the principle with which it was trained.

Concerning education, ChatGPT could aid students' learning and overall productivity by helping them in a variety of ways. It can provide useful information and resources, which can improve language skills, effectiveness, and motivation. However, the use of ChatGPT should be complementary and not a substitute for human interaction and students' diligence in learning and achieving their academic goals. Students should be encouraged to exercise caution in using ChatGPT to verify the credibility of the information sources it generates. While ChatGPT can engage in human-to-human conversations, it remains a machine without genuine human feelings despite its capacity for such interactions, [6]. The ChatGPT is also limited in the educational space because it can only produce information based on the existing dataset on which it was trained. Until it has been trained with new information, ChatGPT does not have access to it. Thus, teachers should focus on building competence in learning technology, enabling them to be more effective in their profession, [7].

One of the impediments of ChatGPT to the academic world is being unable to take responsibility for the contents it generates. Thus, when used for research, it can void the integrity of such research work. The academia is now confused about whether ChatGPT should be cited as a co-author. With this limitation, the ChatGPT can also be helpful but must not be wholly relied upon, [8]. In Statistics, ChatGPT lacks a profound understanding of statistical methodologies, including regression analysis, hypothesis testing, multivariate analysis, and machine learning algorithms, and it is unable to identify potential biases, confounding factors, or outliers that might affect the validity of the results, [9]. Recent concerns have been about whether ChatGPT would help or harm nurses and Doctors, [10], [11]. Recently, a ghostwriter and copywriter for a creative solutions agency and Adobe employees raised concerns about how the introduction of AI has led to job loss and shake-ups, [12].

In the face of millions losing their jobs to AI tools like ChatGPT, this study investigates the strengths and limitations of ChatGPT for Data Analysts. This study also examines whether ChatGPT poses a threat to Data Analysts by potentially encroaching on their roles and job opportunities.

2 Methodology

An econometric dataset is simulated for analysis by ChatGPT using R-programming 4.0. The objective is to evaluate the AI tool's strengths and limitations in data analysis in comparison to human analysts.

2.1 Description of the Dataset

The dataset that was simulated comprises several linear regression datasets deliberately designed to demonstrate various issues or violations of classical linear regression. The regression model is generated as follows:

$$y = x\beta + u_i \tag{1}$$

where $u_i \sim N(0, \Omega)$ and $\Omega = \sigma_i^2 \cdot \mathrm{M}$

M is a definite matrix, which is an identity matrix. In the simulated datasets, multicollinearity, outliers, and heteroscedasticity problems were intentionally introduced.

2.2 Introducing Multicollinearity, Outlier, and Heteroscedasticity

In introducing multicollinearity, the explanatory variables were generated following the works of [13], [14], [15], [16].

$$x_{ij} = z_{ij}\sqrt{1 - R^2} + Rz_{ij+1} \ \ where\ i = 1,2,\dots,30;\ j = 1,2,\dots,30 \tag{2}$$

Outliers can be introduced in two ways: in the x direction or the y direction. In this study, an independent variable is polluted with outliers using the expression given in [17] as:

$$x_{2j} = f_0 \times x_{2,\max} + x_{2j} \tag{3}$$

Where $j \subset [1,2,\dots,30]$ is the sampled index whose size is determined by the percentage of the outlier of interest.

Since the Heteroscedasticity has to do with the error terms, it was generated to change with one of

the explanatory variables, x_3. This is a major violation of Homoscedasticity. The error term is generated with the following steps:

Generate standard normal, z. Set $e_i = z_i \times \sigma^2$. Since the mean of the error term is 0, set $\sigma_i^2 = x_3^2$. To invoke heteroskedasticity so that the variance of the error term is not constant over x_3.

The heteroscedastic error becomes $e_j = z_j \times x_{3i}^2$ (4)

The Error e_i from (4) is modified to follow $AR(1)$ as with an autocorrelation value ρ:
given

$$u_i; i = 1, 2, \cdots, n.\ u_1 = \sigma_1^2 / \left(1 - \rho^2\right) u_i = \rho u_{i-1} + e_i \quad (5)$$

The u_i is added to the $X\beta$ vector to generate the y, the dependent variable.

2.3 Test for Multiple Linear Regression Assumptions

Various tests were conducted to confirm the presence of multicollinearity, outliers, and heteroscedasticity problems introduced earlier in the datasets and to verify the violation of assumptions of the multiple linear regression (MLR) model.

The test for multicollinearity using Variance Inflation Factor (VIF) shows x1=2094.253609, x2 = 1.040304, and x3 = 2092.433043. A VIF value greater than 10 indicates the presence of severe multicollinearity. The test for Autocorrelation using **the** Durbin-Watson test shows the Durbin-Watson value (DW) = 0.42608 and p-value = 8.226e-09. This shows the presence of autocorrelation. Studentized Breusch-Pagan test for Heteroscedasticity gives the value 2.9299 and p-value = 0.4026. This confirms the presence of Heteroscedasticity. Cook's distance, [18] and dffits diagnostic tools, [19], for detecting outliers show some observations as outliers in the datasets. The detected observations are bolded in Table 1 (Appendix).

3 Data Analysis

For analysis purposes, two versions of ChatGPT (3.5 and 4.0) were utilized to determine whether there would be any differences in the responses based on the version. Note that all the Figures are in the Appendix.

3.1 When ChatGPT 3.5 is used to Analyse the Data

Analyst: (After copying the data set in text format into the chat environment) check this data out. How would it be displayed in R studio?
ChatGPT's Response:Fig. (Appendix)

Comment: The ChatGPT got this instruction right. It was even impressive that it had to put the data inside a data frame when asked how it would be displayed in R studio.

Analyst: What would the output look like?

ChatGPT's Response: Fig. (Appendix)

Comment: When the researchers scrolled down, they discovered some parts of the output had been cut off to save space. Additionally, ChatGPT could only generate a few parts of the output, even though there were 30 observations; it only produced 24 of them. This is the first limitation noticed. Therefore, the ChatGPT may produce an incomplete output when not accurately supervised during use.

Analyst: Alright. Having done that, could you suggest a statistical tool that could be considered appropriate for analysing the data?

ChatGPT's Response: Fig. **1**a and 3b (Appendix)

Comment: A lot of unnecessary comments by ChatGPT have been eliminated. The ChatGPT seems to have a good glimpse of the nature of the data. It even gave a hands-on code useful for performing multiple linear regression using the R studio. It however never suggested that assumptions of multiple linear regression be tested first before carrying out regression analysis. This is another limitation of ChatGPT; being able to decipher what tool to use and not being able to take precautions on using such a tool.

Analyst: (It's now being asked to perform multiple linear regression on the data.) Since the multiple linear regression could be appropriate, after taking all necessary measures into consideration, proceed with performing multiple regression on the data.

ChatGPT's Response: Fig. (Appendix)

Comment: There seems to be a "going back-and-forth" in the output generated by ChatGPT. It failed to carry out multiple regression analysis on the data. Also, it encountered difficulties in completing the code when attempting to place the data inside a "data frame" in R Studio. This is a failure on the part of the chatbot.

Analyst: Were you not trained to always test assumptions before performing any analysis, especially when using parametric test statistics?

ChatGPT's Response: Fig. (Appendix)

Comment: The bot acknowledged the importance of testing assumptions before data analysis. It compiled a list of assumptions that needed to be tested before performing multiple linear regression analysis. However, it failed to carry out these assumption tests and proceeded directly to conduct the regression analysis with the datasets. This is another limitation of ChatGPT; it could be misleading when the person using it does not have adequate knowledge of what they intend to use it for. The Ordinary Least Square Estimator (OLSE) is the Best Linear Unbiased Estimator (BLUE). Nevertheless, in the presence of multicollinearity, the OLSE can lead to regression coefficients with incorrect signs, considerably large absolute values, and extensive standard errors, [20], [21]. It means using the OLSE like the ChatGPT in the presence of multicollinearity will give a misleading result. Further chats with ChatGPT about its inefficiency resulted in some arguments. Here is the final submission of the bot: Fig. (Appendix)

Comment: ChatGPT explicitly acknowledges that it lacks personal qualities or capabilities. Consulting with a qualified human data analyst or Statistician is recommended for anyone seeking to perform data analysis.

Analyst: Okay. Could you recall the data you put inside the data frame?

ChatGPT's Response: Fig. (Appendix)

Comment: It's pretty impressive that the bot can still recall the data inside a "data frame" after some conversations.

Analyst: Alright. Now, test the assumptions of multiple linear regression on the data. You have to generate the code in R studio and the output of the code as well. Also, write a brief note on the outcome of each test.

ChatGPT's Response: Fig. (Appendix)

Comment: The ChatGPT, instead of testing the assumptions, listed and explained the underlying assumptions of MLR. After further prompting, it later generates the codes as seen in the next response of the ChatGPT: Fig. (Appendix)

Looking at the response in Fig. (Appendix), it can be realized that ChatGPT failed to write out codes for testing the assumptions of the Multiple Linear Regression (MLR) model. It still repeated the cycle of putting the data inside a data frame.

Analyst: Now perform those tests on the data using the R studio.

ChatGPT's response: Fig. (Appendix)

Comment: In certain instances, ChatGPT has been requested to place a dataset inside a data frame, run a code, and successfully generate an output. However, in a new development, ChatGPT raises eyebrows about its inability to execute codes. Upon closer examination of the output it did produce, it becomes evident that complete codes have not been written to test all the assumptions of the multiple linear regression (MLR) model.

Analyst: Complete the rest from where you stopped.

ChatGPT's Response: Fig. (Appendix)

Comment: In this response, ChatGPT has provided the complete codes for testing assumptions for the MLR model. However, it's worth noting that ChatGPT may not have been able to generate these codes without prompting from the data analyst. Furthermore, none of the codes have been executed yet. This raises a crucial point: if researchers lack the necessary knowledge for data analysis, the bot could execute an incomplete code, resulting in erroneous results.

Analyst: Run those codes in R studio, generate an output, and write a brief note indicating whether or not an assumption has been violated.

ChatGPT's Response: Fig. (Appendix)
Comment: Up there, ChatGPT is lamenting its limitation in running and executing codes. Another

limitation of the ChatGPT is being noticed here. It is worth noting that all the outputs generated thus far by ChatGPT have been obtained through the collaborative efforts of the data analyst, who assists ChatGPT by providing contextual knowledge for each chat.
After several attempts to get ChatGPT to execute codes for testing of MLR model assumptions, the final response by ChatGPT is found in Figure 13 (Appendix).

Note: Some parts were repeated in the response and have hence been cut off. It could also be observed from ChatGPT's response that the bot lacks the capability to run codes and generate textual outputs.

3.2 When ChatGPT 4.0 is used to Analyse the Data

Researchers undertook an investigation to assess the capability of ChatGPT 4.0 for independent data analysis without the involvement of human analysts. This was prompted by the belief that ChatGPT 4.0 possesses greater robustness and intelligence compared to ChatGPT-3.5. The conversations between the researchers and ChatGPT 4.0 are outlined
as follows:

Analyst: Check this data out: Table 2. Econometric Data Put Inside Table by ChatGPT 4 (Appendix)

Do these on the data
1) Put the data inside a table
2) What statistical tool is appropriate for analysing the data?
3) Use that tool to analyse the data.
4) Are there any assumptions that must be met before the statistical tool can be used?
5) Test those assumptions.
6) Why did you employ the statistical tool you used in analysing the data?
7) Generate output for the test of assumptions and the analysis you conducted.
8) Write a brief note on steps 1 to 7.

ChatGPT-4 Response:
ChatGPT responded by putting the data in Table format (Table 2. Econometric Data Put Inside Table by ChatGPT 4, Appendix). It continued its discussions as follows:

There are several statistical tools that can be used to analyze this data, depending on the research question and the nature of the data. One possible tool is multiple linear regression, which can be used to model the relationship between a dependent variable (y) and multiple independent variables (x_1,x_2,x_3). This tool can help determine if there is a significant relationship between the independent variables and the dependent variable. It can also provide information about the strength and direction of those relationships.

Before using multiple linear regression, several assumptions must be met:
- Linearity: A linear relationship between the independent and dependent variables should exist.
- Independence: The observations should be independent of each other.
- Normality: The residuals (errors) should be normally distributed.
- Homoscedasticity: The variance of the residuals should be constant across the range of the independent variables.

These assumptions can be tested using various statistical tests and visual methods such as scatter plots, histograms, and Q-Q plots.

I employed multiple linear regression because it is a widely used statistical tool for analyzing the relationship between dependent and multiple independent variables. It can provide valuable information about the nature of those relationships and can help answer research questions about the data.

I'm sorry, but I am not able to generate an output for the test of assumptions and the analysis as I am an AI language model and do not have the capability to run statistical tests or generate visual outputs. However, you can use statistical software such as R or SPSS to conduct these tests and generate outputs.

In summary, I have put the data into a table format, suggested multiple linear regression as a possible statistical tool for analyzing the data, explained the assumptions that must be met before using this tool, and provided a rationale for why this tool was employed. However, I am not able to generate outputs for the test of assumptions or the analysis. Is there anything else you would like to know?

Comment: Based on the response of ChatGPT-4.0, it can be concluded that its answers align with those of ChatGPT-3.5. Therefore, both versions of the Chabot exhibit the same limitations.

4 Discussions

The ChatGPT is a fantastic tool that anyone who understands its use will find helpful. However, it is not without its limitations. Its use can only be productive when its user has full knowledge of what they intend to achieve with it.

In this study, ChatGPT's prowess in data analysis was examined. Using ChatGPT 3.5, the analyst must make several promptings before it can understand instructions, and at times, it generates wrong outputs and misleading results. However, ChatGPT 4.0 can generate the correct response. When asked to analyse data, a lot of discrepancies were noticed. The outputs it generated were mostly erroneous; it had to be queried and prompted severally before it could generate the correct outputs. The ChatGPT itself submitted that the service of a human data analyst would be required to validate any result it produces. This is because at some points in time, after several promptings and queries, it could still not respond correctly. When asked to generate codes to test assumptions of the MLR model, it took a lot of effort to get it to generate the needed codes, the codes it could not even execute. How about writing a conclusion? It's a no for ChatGPT.

The peak of our findings in this study is the submission of ChatGPT with respect to the need for human Statistical analysts. In its words, after a query due to its failure to perform a task as expected, it responded, 'I can offer guidance on various topics, including data analysis, but it is important to note that the final responsibility for data analysis and interpretation rests with the user. It's always recommended to consult a qualified human data analyst or Statistician to ensure accurate and reliable results for complex data analysis tasks." Therefore, we may conclude that ChatGPT will not be able to replace human data analysts. The tool may come handy to data analysts when stuck on what to do but should never be considered an authority in statistical decisions.

Although ChatGPT may not be efficient for data analysis, it can still be helpful by offering valuable suggestions and writing codes. However, it is essential that the analyst thoroughly checks these suggestions and codes before adopting them. Human data analysts need not fear losing their jobs anytime soon; instead, they should focus on enhancing their data analysis knowledge. This will enable them to discern the correct approaches from the incorrect ones, ensuring their continued relevance and success in their field. This is consistent with the submission of Ausat et al. (2023), who submitted that ChatGPT can be used as a tool in the learning process and not a substitute for the role of the teacher in the classroom.

The ChatGPT may appear to be a valuable tool for data analysis. However, it is referred to as a tool for a reason, indicating that only individuals with mastery of the tool can utilize it for optimal results. It is undeniable that GPT is faster and can operate at incredible speed when analyzing data compared to humans.

Nevertheless, it remains reliant on humans as it can only generate codes and offer suggestions for data analysis, unable to perform any data analysis independently. This signifies that human data analysts are still irreplaceable in the field of data analysis. Sometimes, ChatGPT can even produce erroneous codes and provide incorrect suggestions when asked for help. Data analysis with the tool will always require the service of an expert in most cases, [22].

5 Conclusion

This study has established that the role of ChatGPT in data analysis is incredibly crucial, serving as a valuable tool for knowledgeable data analysts. It provides a natural language interface, making it easier for analysts to ask questions, aiding in data exploration, cleaning, and pre-processing. Additionally, ChatGPT can explain complex concepts in data analysis and provide information in domain-specific knowledge.

However, to maximize its potential in data analysis and interpretation for decision-making in various domains of life, the expertise of a knowledgeable and experienced Statistical Analyst is necessary. This study has also revealed the weaknesses and limitations of ChatGPT, particularly when used by novices in the field of Statistics.

The requirement for statistical knowledge, critical thinking, ethical awareness, continuous learning, and effective communication skills to convey complex findings in a clear and understandable manner to non-technical stakeholders in data analysis can only be fulfilled by human analysts. Therefore, we can infer that the introduction of ChatGPT has enhanced the efficiency and speed of a Data Analyst's job, especially for those with a solid grasp of theoretical statistics. This tool simplifies the analysis process for them, providing a valuable resource when they encounter challenges in statistical analysis.

References:

[1] C. McClelland, "The Impact of Artificial Intelligence - Widespread Job Losses," 30 January 2023. [Online]. https://www.iotforall.com/impact-of-artificial-intelligence-job-losses (Accessed Date: July 22, 2023).

[2] Savelka, J.; Agarwal, A.; Bogart, C.; Song, Y. and Sakr, M., "Can generative pre-trained transformers (GPT) pass assessments in higher education programming courses?" *arXiv preprint,* p. arXiv:2303.09325, 2023, https://doi.org/10.1145/3587102.3588792

[3] Wamba, S. F., Bawack, R. E., Guthrie, C., Queiroz, M. M., & Carillo, K. D., "Are we preparing for a good AI society? A bibliometric review and research agenda," *Technological Forecasting and Social Change,* p. 120482, 2020, https://doi.org/10.2139/ssrn.3735322.

[4] OpenAI, "ChatGPT," Open AI, 2022. [Online], https://chat.openai.com (Accessed Date: March 31, 2024).

[5] Adamopoulou, E., & Moussiades, L., *An overview of chatbot technology*, Switzerland: Springer Nature, 2020, pp. 373 - 383.

[6] Fauzi, F.; Tuhuteru, L.; Sampe, F.; Ausat, A. M. A. & Hatta, H. R., "Analysing the Role of ChatGPT in Improving Student Productivity in Higher Education," *J. Educ.,* vol. 5, no. 4, p. 14886–14891, 2023, https://doi.org/10.31004/joe.v5i4.2563.

[7] Ausat,A. M. A.; Massang, B.; Efendi, M.; Nofirman; Riady, Y., "Can Chat GPT Replace the Role of the Teacher in the Classroom: A Fundamental Analysis," *Journal on Education,* vol. 5, no. 4, pp. 16100-16106, 2023, https://doi.org/10.31004/joe.v5i4.2745.

[8] A. Iskender, "Holy or Unholy? Interview with Open AI's ChatGPT," *European Journal of Tourism Research,* vol. 34, no. 3414, 2023, https://doi.org/10.54055/ejtr.v34i.3169.

[9] U. Bhosale, "9 Things Researchers Shouldn't Expect From ChatGPT: Setting realistic boundaries.," Enago Academy, 8 June 2023, [Online]. https://www.enago.com/academy/chatgpt-cannot-do-for-researchers (Accessed Date: July 22, 2023).

[10] S. O'Connor, "ChatGPT. Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse?" *Nurse Educ Pract. ,* vol. 66:103537, p. Medline:36549229, 2023, https://doi.org/10.1016/j.nepr.2022.103537.

[11] S. Huh, "Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study," *J Educ Eval Health Prof.,* vol. 20, no. 1, 2023, https://doi.org/10.3352/jeehp.2023.20.01.

[12] B. Kato, "AI stole my job — pay is down 90%, and my family is suffering," 2 8 2023, [Online]. https://nypost.com/2023/08/02/ai-stole-my-job-pay-is-down-90-and-my-family-is-suffering/ (Accessed Date: August 4, 2023).

[13] Idowu, J.I., Oladapo, O. J., Owolabi, A. T., Ayinde, K. and Akinmoju, O., "Combating Multicollinearity: A new two-parameter Approach," *Journal of Quantitative Sciences,* vol. 5, no. 1, 2023, https://doi.org/10.51541/nicel.1084768.

[14] Owolabi, A. T., Ayinde, K. and Alabi, O. O., "A Modified Two Parameter Estimator with Different Forms of Biasing Parameters in the Linear Regression Model," *African Scientific Reports,* vol. 1, p. 212–228, 2022, https://doi.org/10.46481/asr.2022.1.3.62.

[15] Oladapo, O.J., Owolabi, A.T., Idowu, J.I. and Ayinde, K. , "A New Modified Liu Ridge-Type Estimator for the Linear Regression Model: Simulation and Application," *Int J Clin Biostat Biom,* vol. 8, no. 048, 2022, https://doi.org/10.23937/2469-5831/1510048

[16] Ayinde, K.; Lukman, A.F.; Samuel, O.O.; & Ajiboye, S.A., "Some new adjusted ridge estimators of linear regression model," *Int. J. Civ. Eng. Technol.,* p. 2838-2852, 2018.

[17] Lukman, A. F., Arashi, M. and Prokaj, V., "Robust biased estimators for Poisson regression model: Simulation and applications," *Concurrency and Computation: Practice and Experience,* vol. 35, p. e7594, 2023, https://doi.org/10.1002/cpe.7594.

[18] R. Cook, "Detection of Influential Observations in Linear Regression," *Technometrics,* 1977.

[19] Paul, S.R. and Fung, K. Y., "A generalized extreme studentized residual multiple outlier

detection procedure in Linear Regression," *Technometrics,* vol. 33, pp. 339-348, 1991, https://doi.org/10.1080/00401706.1991.10484839.

[20] Lukman, A.F. and Ayinde, K., "Review and classifications of the ridge parameter estimation techniques," *Hacettepe Math Stat.,* vol. 46, no. 5, p. 953-967, 2017,https://doi.org/10.15672/hjms.201815671.

[21] Owolabi, A. T., Ayinde, K., Idowu, J. I., Oladapo, O. J., & Lukman, A. F., "A New Two-Parameter Estimator in the Linear Regression Model with Correlated Regressors," *Journal of Statistics Applications & Probability,* vol. 11, pp. 499-512, 2022, https://doi.org/10.18576/jsap/110211.

[22] A. Sakamoto, "ChatGPT Data Analysis vs. Human Analysis: Which One Should You Choose?," Kanaries, 31 07 2023, [Online]. https://docs.kanaries.net/articles/chatgpt-data-analysis-human (Accessed Date: August 4, 2023).

APPENDIX

Table 1. Identification of outliers in the datasets

dfb.1_	dfb.x1	dfb.x2	dfb.x3	dffit	cov.r	cook.d	hat
-0.056	-0.045	0.026	0.046	-0.067	1.300	0.001	0.106
-0.460	-0.183	0.210	0.185	-0.471	0.709	0.050	0.056
0.060	0.243	-1.171	**-0.238**	-1.308	0.457	0.335	0.173
-0.286	0.137	0.148	-0.146	-0.584	0.739	0.077	0.085
-0.311	0.084	0.152	-0.088	-0.419	0.757	0.040	0.051
-0.206	0.042	0.085	-0.037	-0.334	1.235	0.028	0.146
0.088	-0.084	-0.033	0.080	0.196	1.441	0.010	0.207
0.060	0.022	-0.030	-0.021	0.082	1.235	0.002	0.067
-0.034	0.058	0.015	-0.058	-0.078	1.279	0.002	0.094
0.046	0.041	-0.024	-0.040	0.077	1.316	0.002	0.118
0.055	0.032	-0.025	-0.032	0.066	1.292	0.001	0.100
0.060	0.075	-0.030	-0.074	0.088	1.382	0.002	0.159
-0.029	-0.044	-0.147	0.047	-0.208	1.485	0.011	**0.230**
-0.012	-0.101	-0.234	0.100	-0.306	1.387	0.024	**0.204**
0.007	0.008	-0.003	-0.008	0.010	1.445	0.000	0.191
-0.127	0.069	0.061	-0.071	-0.189	1.125	0.009	0.056
-0.069	-0.040	0.034	0.039	-0.084	1.231	0.002	0.065
0.276	0.089	-0.114	-0.096	0.427	1.271	0.046	0.187
0.062	0.453	0.610	**-0.448**	0.934	1.091	0.206	**0.262**
0.242	-0.141	-0.125	0.149	0.508	0.846	0.060	0.086
0.009	-0.244	0.572	**0.232**	0.794	1.221	0.153	**0.267**
-0.109	-0.279	0.427	**0.282**	0.558	1.430	0.078	**0.279**
0.207	-0.140	-0.088	0.137	0.306	1.049	0.023	0.073
0.068	-0.165	-0.027	0.164	0.204	1.323	0.011	0.149
0.090	-0.121	-0.045	0.124	0.220	1.196	0.012	0.092
0.134	-0.009	-0.062	0.009	0.142	1.136	0.005	0.042
0.029	-0.026	-0.015	0.027	0.064	1.272	0.001	0.087
-0.031	0.130	0.015	-0.131	-0.157	1.395	0.006	0.177
0.169	0.157	-0.088	-0.153	0.286	1.220	0.021	0.125
0.083	0.014	-0.042	-0.013	0.122	1.212	0.004	0.066

NOTE: The bolded observations are outliers.

Table 2. Econometric Data Put Inside Table by ChatGPT 4

y	x1	x2	x3
0.355772049	-0.366547941	-1.182745938	-0.91417413
-0.416631134	-0.25264793	-0.206310012	-0.371176738
0.768840797	-0.535158825	13.49243216	-0.299338951
1.870846452	0.778642334	1.120420843	1.188066458
1.538532252	0.663850856	-0.066041848	1.000284307
-0.949063094	-1.107773201	-1.903406801	-0.747583634
-1.216966558	-1.972597717	-1.288179357	-1.125329398
1.960095769	0.645110609	0.03686517	0.593879336
0.884327286	-0.575699024	-0.008114294	0.378645915
2.25594158	0.992529211	0.820315829	0.561067384
0.455959839	-0.205369354	-0.212007131	-0.519583762
1.294878123	0.706875387	0.045052258	-0.246070919
0.407722362	-1.039792628	11.71049892	-1.392955829
3.437947651	1.070833737	14.01180889	0.54517308
-0.094945274	-0.246727249	-0.991681436	-1.219339862
2.172972119	0.545802769	0.494296884	1.018028015
1.152682348	0.178637641	0.696243381	-0.050153627
-0.534216222	-1.577309852	-1.623539592	-1.679847471
4.625167052	1.328776544	13.63763941	0.466952523
5.36849768	1.225683711	1.300765358	1.678838253
2.455415364	-1.412064924	11.68088058	-0.893731385
4.887618937	0.063436956	13.87780516	0.918925503
1.992924213	-0.711888317	0.17855623	-0.106977664
1.463271696	-1.207861548	0.647646888	0.028174873
3.525824643	0.2803768	1.052345338	1.023760812
2.373526151	-0.093445044	0.41218895	0.138662914
3.446087155	0.636232781	0.542666176	1.223515765
3.276991747	0.287532367	0.899982	1.635599667
3.225472496	1.085633778	0.848451595	0.613920679
3.0287778	0.745206701	0.982665575	0.823590469

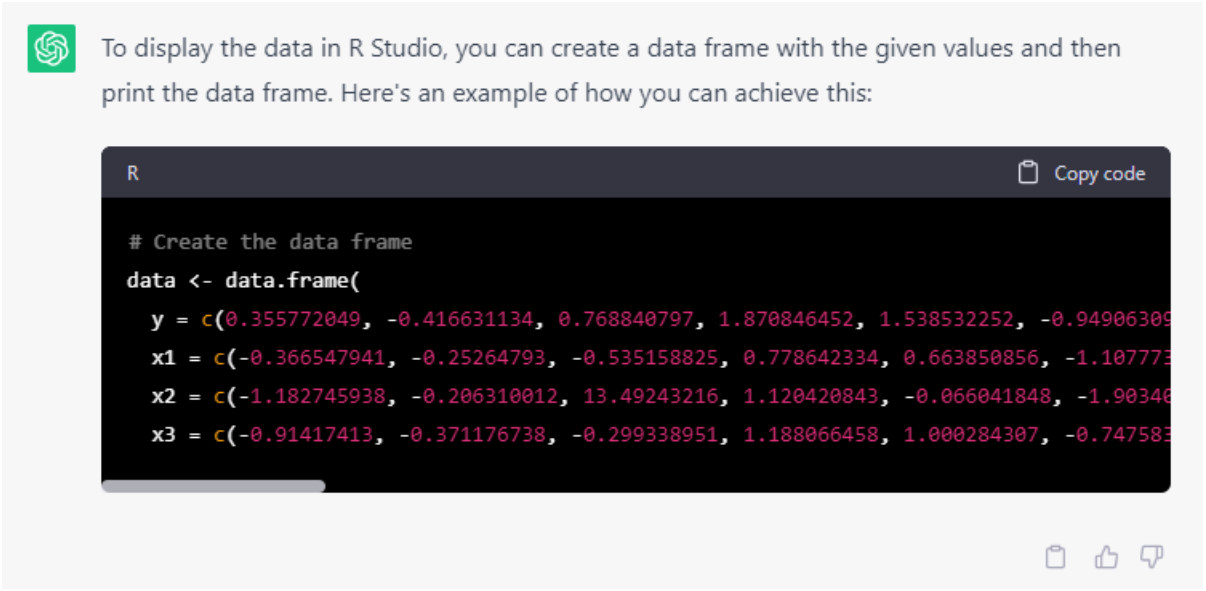
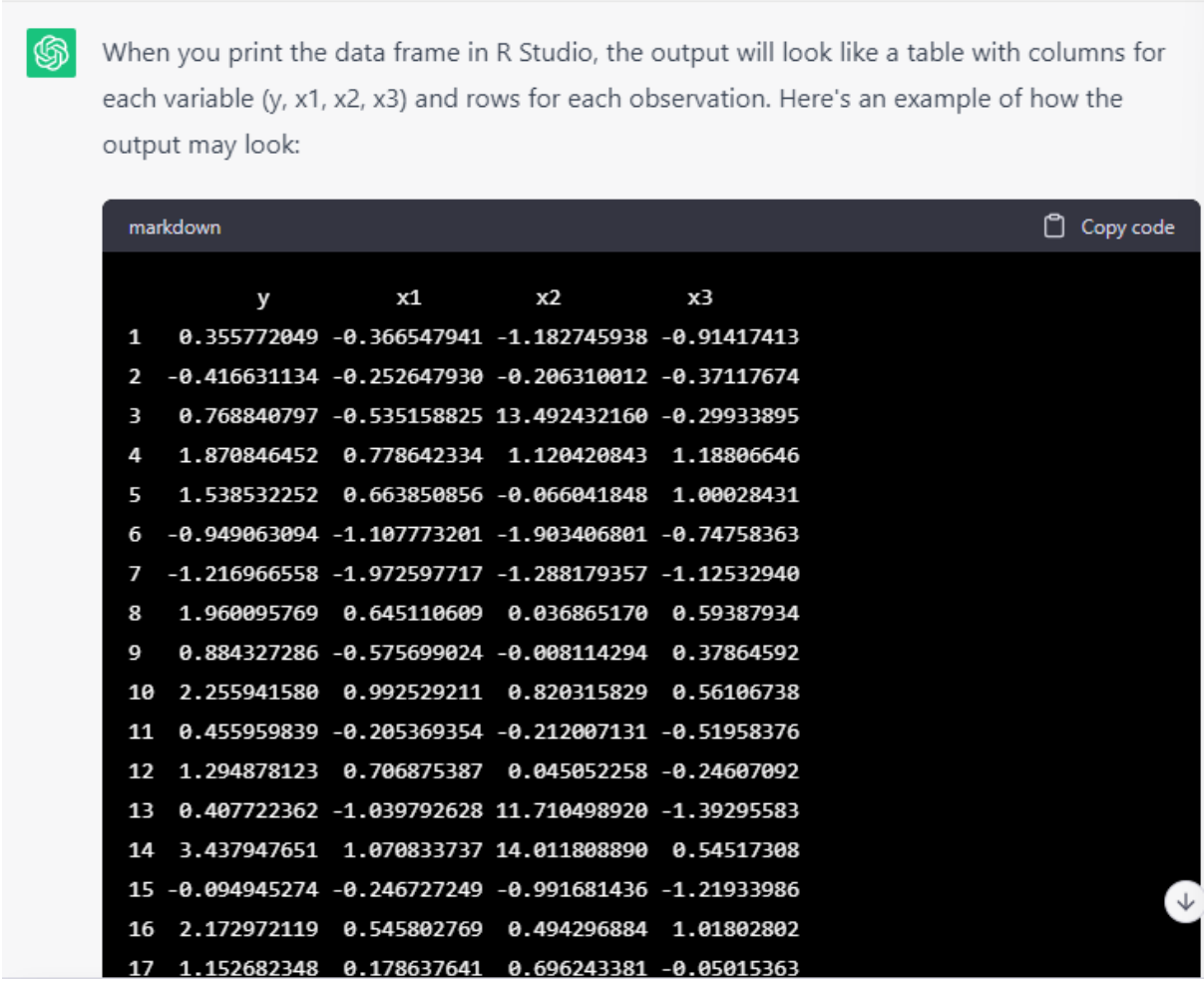


Fig. 1: ChatGPT's Response to the first question of the Analyst



	y	x1	x2	x3
1	0.355772049	-0.366547941	-1.182745938	-0.91417413
2	-0.416631134	-0.252647930	-0.206310012	-0.37117674
3	0.768840797	-0.535158825	13.492432160	-0.29933895
4	1.870846452	0.778642334	1.120420843	1.18806646
5	1.538532252	0.663850856	-0.066041848	1.00028431
6	-0.949063094	-1.107773201	-1.903406801	-0.74758363
7	-1.216966558	-1.972597717	-1.288179357	-1.12532940
8	1.960095769	0.645110609	0.036865170	0.59387934
9	0.884327286	-0.575699024	-0.008114294	0.37864592
10	2.255941580	0.992529211	0.820315829	0.56106738
11	0.455959839	-0.205369354	-0.212007131	-0.51958376
12	1.294878123	0.706875387	0.045052258	-0.24607092
13	0.407722362	-1.039792628	11.710498920	-1.39295583
14	3.437947651	1.070833737	14.011808890	0.54517308
15	-0.094945274	-0.246727249	-0.991681436	-1.21933986
16	2.172972119	0.545802769	0.494296884	1.01802802
17	1.152682348	0.178637641	0.696243381	-0.05015363

Fig. 2: ChatGPT's Response to the second question of the Analyst

Based on the given data, which consists of multiple variables (y, x1, x2, x3), and assuming the goal is to explore relationships and patterns within the data, a suitable statistical tool for analysis could be multiple linear regression.

Multiple linear regression allows for the examination of the linear relationship between a dependent variable (y) and multiple independent variables (x1, x2, x3) simultaneously. It helps in understanding how the independent variables collectively contribute to the variation in the dependent variable and can provide insights into the strength and significance of those relationships.

Fig. 1a: ChatGPT's Response to the third question of the Analyst

Here's an example of how you can fit a multiple linear regression model in R using the given data:

R Copy code

```
# Fit a multiple linear regression model
model <- lm(y ~ x1 + x2 + x3, data = data)

# Print the summary of the model
summary(model)
```

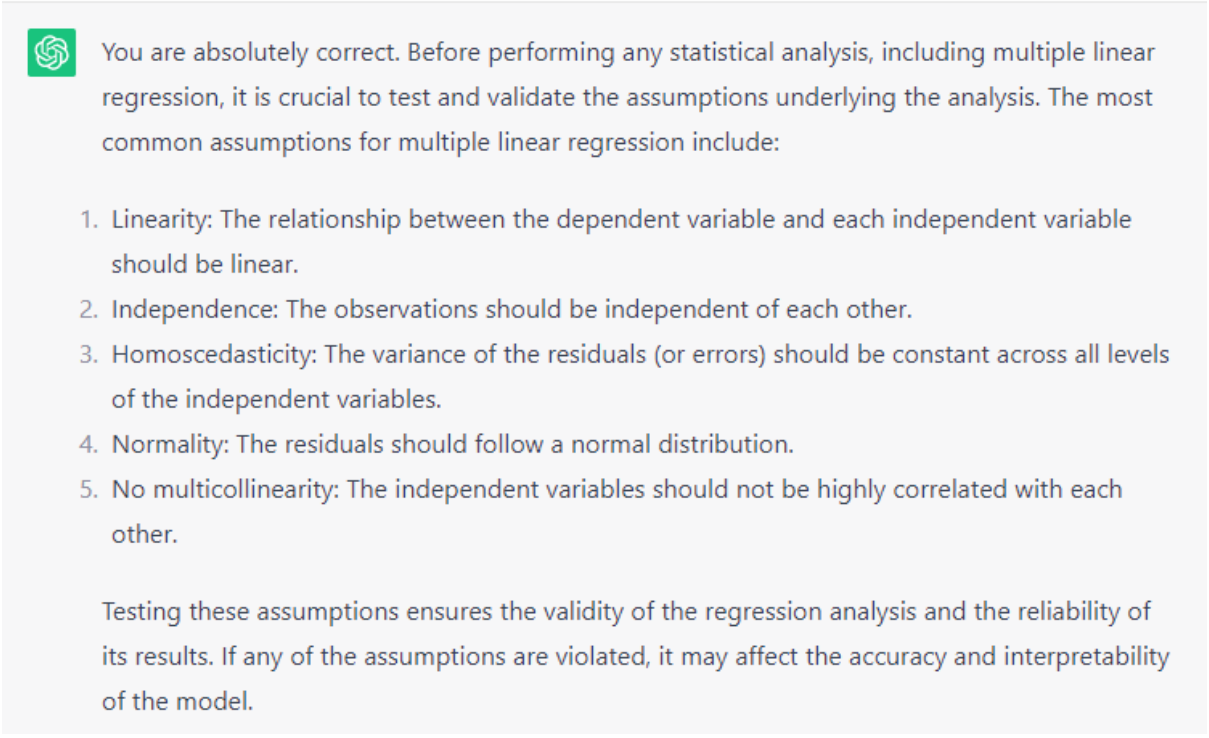
Fig. 3b: Part of ChatGPT's Response to the third question of the Analyst

To perform multiple linear regression on the given data, you can use the `lm()` function in R. Here's an example of how you can fit a multiple linear regression model and examine the results:

R Copy code

```
# Create the data frame
data <- data.frame(
  y = c(0.355772049, -0.416631134, 0.768840797, 1.870846452, 1.538532252, -0.94906309
  x1 = c(-0.366547941, -0.25264793, -0.535158825, 0.778642334, 0.663850856, -1.107773
  x2 = c(-1.182745938, -0.206310012, 13.49243216, 1.120420843, -0.066041848, -1.90340
  x3 = c(-0.91417413, -0.371176738, -0.299338951, 1.188066458, 1.000284307, -0.747583
```

Fig. 4: ChatGPT's Response to the fourth question of the Analyst

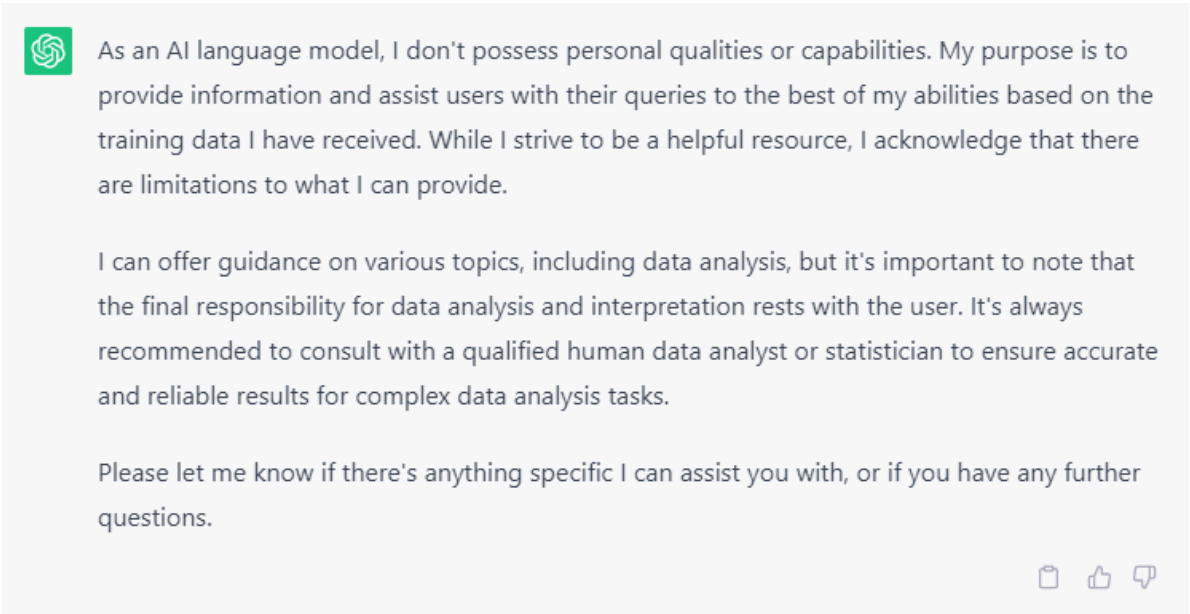


You are absolutely correct. Before performing any statistical analysis, including multiple linear regression, it is crucial to test and validate the assumptions underlying the analysis. The most common assumptions for multiple linear regression include:

1. Linearity: The relationship between the dependent variable and each independent variable should be linear.
2. Independence: The observations should be independent of each other.
3. Homoscedasticity: The variance of the residuals (or errors) should be constant across all levels of the independent variables.
4. Normality: The residuals should follow a normal distribution.
5. No multicollinearity: The independent variables should not be highly correlated with each other.

Testing these assumptions ensures the validity of the regression analysis and the reliability of its results. If any of the assumptions are violated, it may affect the accuracy and interpretability of the model.

Fig. 5: ChatGPT's Response to the fifth question of the Analyst



As an AI language model, I don't possess personal qualities or capabilities. My purpose is to provide information and assist users with their queries to the best of my abilities based on the training data I have received. While I strive to be a helpful resource, I acknowledge that there are limitations to what I can provide.

I can offer guidance on various topics, including data analysis, but it's important to note that the final responsibility for data analysis and interpretation rests with the user. It's always recommended to consult with a qualified human data analyst or statistician to ensure accurate and reliable results for complex data analysis tasks.

Please let me know if there's anything specific I can assist you with, or if you have any further questions.

Fig. 6: ChatGPT's Response to the sixth question of the Analyst

Certainly! Here is the data that was used to create the data frame:

diff Copy code

y	x1	x2	x3
0.355772049	-0.366547941	-1.182745938	-0.91417413
-0.416631134	-0.25264793	-0.206310012	-0.371176738
0.768840797	-0.535158825	13.49243216	-0.299338951
1.870846452	0.778642334	1.120420843	1.188066458
1.538532252	0.663850856	-0.066041848	1.000284307
-0.949063094	-1.107773201	-1.903406801	-0.747583634
-1.216966558	-1.972597717	-1.288179357	-1.125329398
1.960095769	0.645110609	0.03686517	0.593879336
0.884327286	-0.575699024	-0.008114294	0.378645915
2.25594158	0.992529211	0.820315829	0.561067384
0.455959839	-0.205369354	-0.212007131	-0.519583762
1.294878123	0.706875387	0.045052258	-0.246070919

Fig. 7: ChatGPT's Response to the seventh question of the Analyst

To test the assumptions of multiple linear regression, we will perform the following tests:

1. Linearity: We can assess linearity by examining scatterplots of the dependent variable (y) against each independent variable (x1, x2, x3) and verifying if the relationship appears approximately linear.
2. Independence: Since the data is already collected, we assume independence of observations.
3. Homoscedasticity: We can assess homoscedasticity by plotting the residuals against the predicted values and checking for a consistent spread of residuals across different levels of the predicted values.
4. Normality: We can check the normality assumption by examining a histogram or a Q-Q plot of the residuals.
5. No multicollinearity: We can calculate the variance inflation factor (VIF) for each independent variable to assess multicollinearity.

Fig. 8: ChatGPT's Response to the eighth question of the Analyst

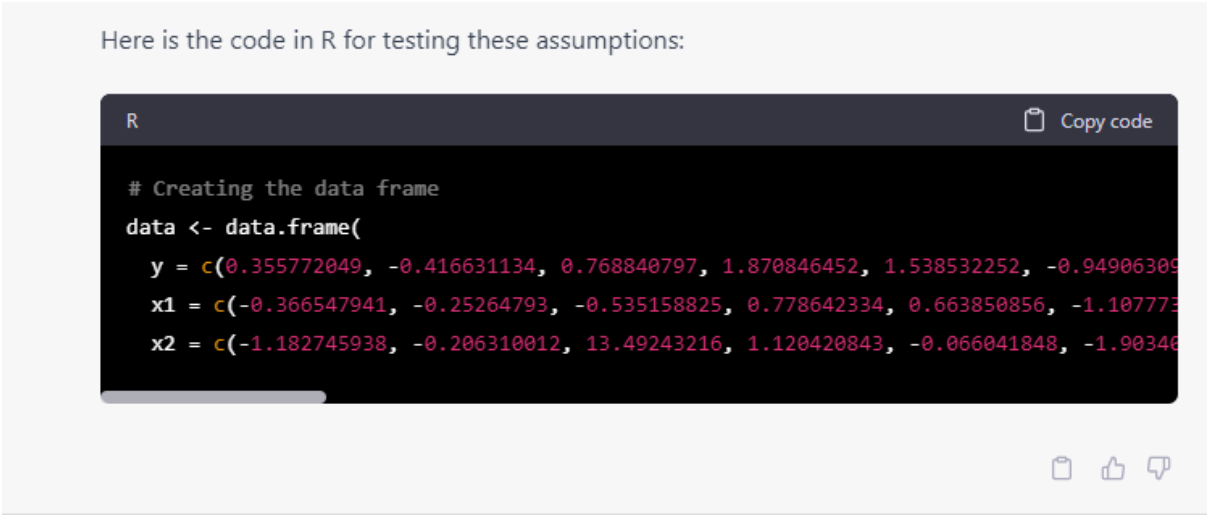


Fig. 9: ChatGPT's Response to the ninth question of the Analyst

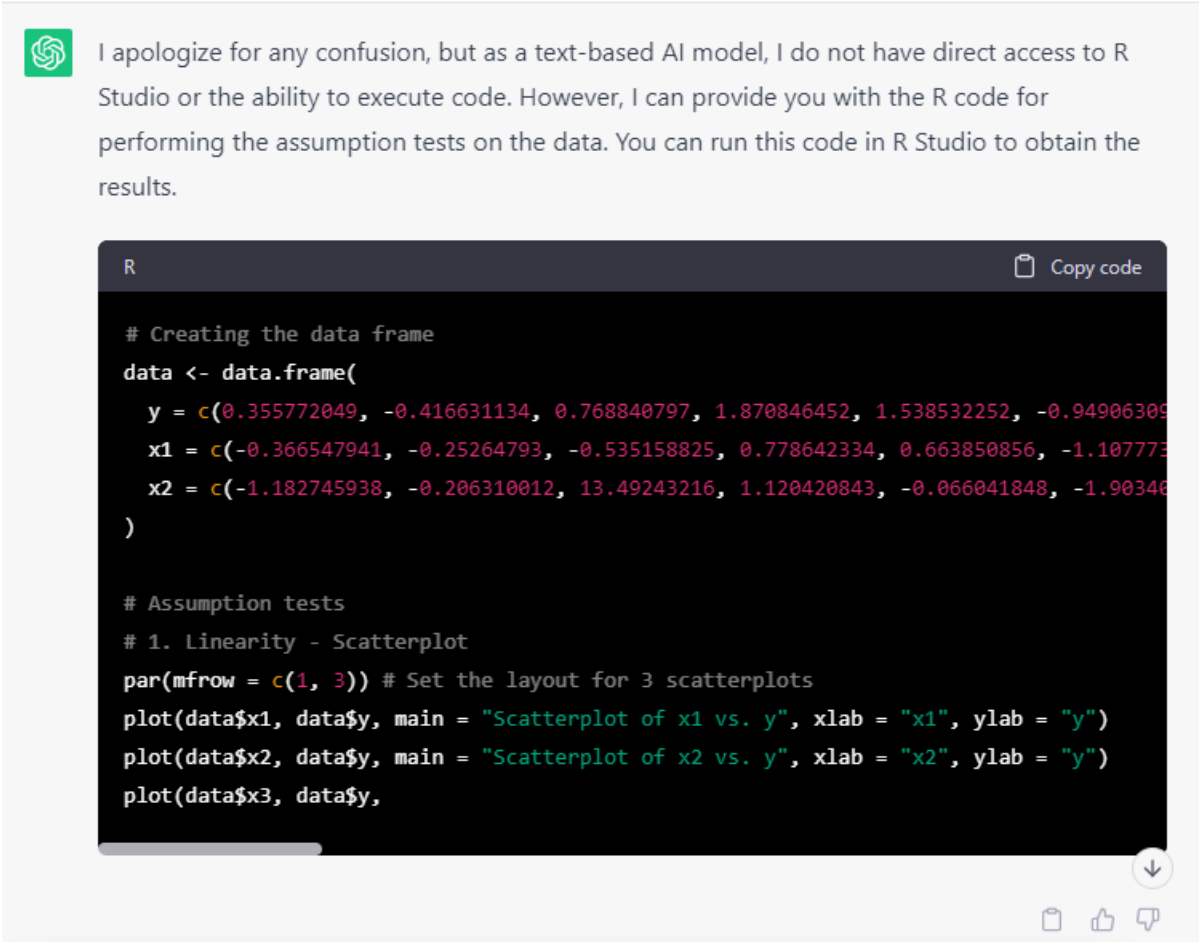


Fig. 10: ChatGPT's Response to the tenth question of the Analyst

Apologies for the incomplete response. Here's the continuation of the code:

R Copy code

```r
plot(data$x3, data$y, main = "Scatterplot of x3 vs. y", xlab = "x3", ylab = "y")

# 2. Independence: No code needed as we assume independence of observations.

# 3. Homoscedasticity: Residuals vs. Fitted plot
model <- lm(y ~ x1 + x2 + x3, data = data)  # Fit the multiple linear regression mode
residuals <- residuals(model)  # Get the residuals
fitted <- fitted(model)  # Get the predicted values
plot(fitted, residuals, main = "Residuals vs. Fitted", xlab = "Fitted values", ylab =

# 4. Normality: Histogram and Q-Q plot of residuals
par(mfrow = c(1, 2))  # Set the layout for 2 plots
hist(residuals, main = "Histogram of Residuals")
qqnorm(residuals, main = "Q-Q Plot of Residuals")
qqline(residuals)

# 5. No multicollinearity: Variance Inflation Factor (VIF)
library(car)  # Load the 'car' package for VIF calculation
vif <- vif(model)  # Calculate VIF for each independent variable
vif
```

Fig. 11: ChatGPT's Response to the eleventh question of the Analyst

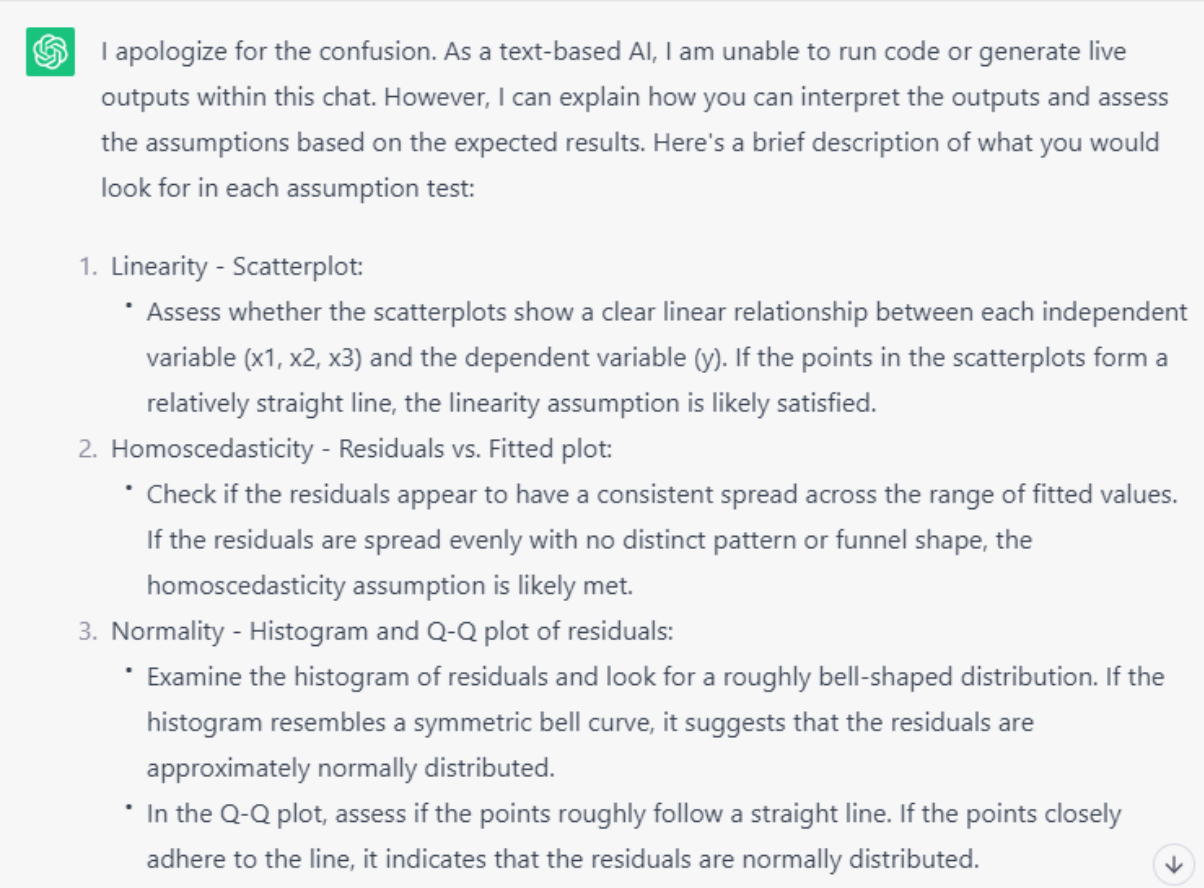


Fig. 12: ChatGPT's Response to the twelfth question of the Analyst

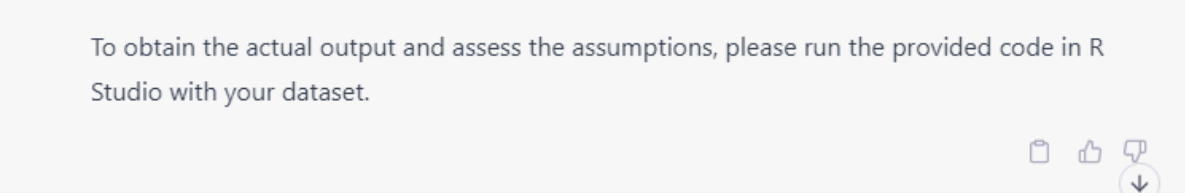


Fig. 13: ChatGPT's Response to the thirteenth question of the Analyst

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)
The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself
No funding was received for conducting this study.

Conflict of Interest
The authors have no conflicts of interest to declare.