Journal of Technological Sciences (JTS)
Chinhoyi University of Technology
Journal of Technological Sciences
http://journals.cut.ac.zw/index.php/jts

Linear Regression in the Spotlight: From Statistical Staple to Misused Tool

Chimwanda, Peter[1], Rupi, Edwin[2]

[1]Department of Mathematics, Chinhoyi University of Technology. Chinhoyi
[2]Department of Mathematics, Masvingo Teachers College Masvingo
pchimwanda@cut.ac.zw
edwinrupi@gmail.com

Abstract

This article explores the application, assumptions, and frequent misuses of linear regression analysis in research, particularly within the business, social sciences and medical fields. While linear regression remains one of the most widely used and accessible statistical tools due to its simplicity and interpretability, it is often misapplied, especially when researchers overlook the foundational assumptions required for valid inferences. The paper reviews the key assumptions of linearity, normality, homoscedasticity, and independence of errors, and discusses the appropriate use of linear regression in descriptive, predictive, and causal research. Through a critical review of published studies and an empirical analysis of customer satisfaction data from Kaggle, the article identifies common violations, including the inappropriate modelling of discrete and ordinal dependent variables, unjustified covariate adjustments, and misinterpretation of regression coefficients. Residual plots and diagnostic tests further reveal that linear regression is frequently applied where it is not suitable, leading to misleading conclusions. The study concludes with practical recommendations to improve statistical literacy and rigor among researchers, emphasizing the importance of involving statisticians and aligning statistical instruction with domain-specific research contexts.

Key words: Residual, Autocorrelation, Homoscedasticity, Linearity, Regression.

Introduction

Linear Regression is a commonly employed statistical technique across different fields to model the relationship between a dependent variable and one or more independent variables (Montgomery et al., 2012). In Economics, it is notably used to examine the link between consumer spending and personal income, while in Finance, it helps predict financial risks and stock prices. In Healthcare, linear regression is utilized to analyse relationships between variables such as age and blood pressure. Additionally, in business, effective marketing strategies are crucial for success, and linear regression aids in optimizing these strategies and forecasting sales based on advertising expenditures. Linear regression is extensively utilized in Environmental Science to examine how factors such as rainfall, temperature, and fertilizer affect crop yield. In Hydrology, it helps identify relationships between variables like rainfall and water table depth. Various sports employ linear regression to analyse performance metrics. In Real Estate, it is used to forecast house prices based on factors like location, square footage, and number of rooms. Additionally, the Social Sciences rely on linear regression for preliminary data analysis and predicting future trends.

Regression analysis helps determine how changes in one variable such as price can influence another, like sales (Leffondré et al., 2014). The technique comes in many forms that include linear, logistic, ordinal, multinomial, ridge, Lasso, hedonic and Gompertz models.

The method is perhaps the most commonly used form of statistical analysis and is invaluable when making a large number of business and economic decisions (Nwachukwu et al 2000). In areas such as the social and behavioural sciences, medicine and public health, linear regression, in particular, stands out as one of the most widely used analytical tools (Darlington & Hayes, 2017). One of the reasons for linear regression's wide spread use is that there are several natural phenomenal laws that can be captured by it.

The use of linear regression models is generally justified, provided particular assumptions are satisfied. While some statistical techniques are complex and require specialized knowledge, others are more accessible (Mehta, 2023). Linear regression stands out as one of the simplest, hence most commonly used method, which explains why it features prominently in a vast number of research studies. Allen, M. P. (2004) encourages researchers to employ linear regression analysis because its linear functional form is simpler than most mathematical equations.

Regardless of it being the most widely used technique, linear regression has moved from being a statistical staple to a misused tool. This article examines a variety of situations in which the technique is misused. The remaining part of the paper looks at Linear Regression Analysis assumptions, uses of linear regression, misuses of linear regression, methodology, data analysis, findings, conclusion and recommendations, in that order.

Linear Regression Analysis assumptions

i. **Linearity:**
The core assumption of linear regression is that the relationship between the independent variable(s) and the dependent variable is a straight line. This implies that a change in the dependent or response variable due to one unit change in independent or predictor variable is constant, regardless of the value of the predictor variable. In multiple regression, this linear relationship should exist between the response variable and each of the predictor variables. This linearity assumption is stressed by Hadi and Chatterjee (2006).

ii. **Residual Expectation**
$\boldsymbol{E\{e_i\} = 0}$ **for all** ***i.*** The expected value of the errors is zero. Suppose there is a number of observations with the same value of the independent variable. If the relationship between the dependent variable and this independent variable is exact then all the observations mentioned above have the same value of the dependent variable. Each residual, which is the deviation of the estimated value from the observed value, is zero hence the expectation is zero as stated. Alternatively, if the relationship is statistical, this assumption implies that the deviations are distributed with both positive and negative values, balancing out so that their sum, and therefore their expected value, is zero.

iii. **Homoscedasticity**
$\boldsymbol{Var\{e_i\} = \sigma^2}$ **for all** ***i.*** The variance of the errors is constant. This assumption suggests that because there are multiple residuals for each value of the independent variable, there is an associated variance, and this variance remains constant across all values of the independent variable.

iv. **Autocorrelation**
$\boldsymbol{\mathrm{Cov}(e_i, e_j) = 0}$ **for** $\boldsymbol{i \neq j}$. The errors are uncorrelated to each other. When we talk about uncorrelated errors in the context of regression, we mean that the covariance of the residuals of different observations is zero.

v. **Normality**
$\boldsymbol{\varepsilon \sim N(0, \sigma^2)}$. The errors are normally distributed. This implies that the errors are symmetrically distributed around zero, with no skewness or kurtosis.

In order for regression analysis to generate valid results the above assumptions must be satisfied.

Uses of Linear Regression

Linear regression analysis is a powerful statistical tool that can be used for descriptive, predictive, and causal research. Below is a detailed explanation of how it is applied in each of these types of research, along with concrete examples:

Descriptive Purpose

A linear regression model is said to answer a descriptive question if it seeks to provide a broad characterization of populations or subpopulations (in the latter case, perhaps with the aim of describing the difference between subpopulations (Carlin and Moreno-Betancur 2024). The model describes the strength and direction of relationships between variables, without implying causation. The focus of such research is to provide summary statistics such as means and standard deviations of continuous variables along with percentage breakdowns into key categories of interest (Carlin and Moreno-Betancur 2024). As an example, a researcher may want to explore the relationship between hours studied and examination scores among university students. The corresponding linear regression model is:

$$\text{Examination Score} = \beta_0 + \beta_1(\text{Hours Studied}) + \varepsilon$$

Where β_1 is the change in examination score per unit change in hours studied. This gives a clear picture of how strongly the two variables are related. Descriptive regression is often exploratory, it tells us what is happening, not why or what will happen next.

Predictive Research

In predictive research, linear regression helps build a model that can be used to predict the value of a dependent variable based on one or more independent variables. Prediction problems invariably involve multiple predictors and seek to develop an algorithm (i.e. in our usage, a procedure) for reliably forecasting the value of Y for individuals for whom only the values of the *X*'s are available (Carlin and Moreno-Betancur 2024). An insurance company, for example, wants to predict a client's car insurance premium based on age, driving and car type. The linear regression model is:

$$\text{Premium} = \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{experience}) + \beta_3(\text{Car TYPE}) + \varepsilon$$

The model is developed on historical data. Once developed, it can be used to predict the premium for a new client. The model tells us how each factor contributes to the predicted premium. Predictive regression focuses on what will happen or estimating unknowns, using patterns in existing data.

Causal Research

Causal research aims to establish cause-and-effect relationships, often using techniques like controlled experiments, instrumental variables, or difference-in-differences within a regression framework to address confounding factors. These studies seek to answer a "What if…" question. Such questions are answered, ideally, by experimentation. However, regression modelling through observational data is often used. A policymaker who wants to know if increasing the minimum

wage causes a change in employment levels may build the following model: The simple linear regression model is:

$$\text{Employment Rate} = \beta_0 + \beta_1(\text{Minimum Wage})) + \varepsilon$$

Other variables (e.g., economic gross domestic product growth, industry type) could influence employment. To control for their effect, we could use a more controlled regression (e.g., include control variables, fixed effects, or natural experiments) to isolate the causal impact of minimum wage on employment. An improved model is as follows:

$$\text{Employment Rate} = \beta_0 + \beta_1(\text{Minimum Wage})) + \beta_1(\text{GDP Growth}) + \alpha_i + \delta_t + \varepsilon_{it}$$

Causal regression is about answering "what happens if we change X?" it requires careful design to avoid misleading conclusions.

Misuses of Linear Regression

Many applications of regression analysis in the medical and health research literature lack clarity of purpose and exhibit misunderstanding of key concepts. Linear regression analysis can be a very effective way to model data as long as the assumptions being made are true, but if they are violated least squares can potentially lead to misleading results (Chatterjee and Simonoff, 2013). The technique is suited for a dependent variable which is quantitative and continuous.

Pearl (2000) asserts that while linear regression can reveal relationships between variables, misinterpretation may result in incorrect causal conclusions, highlighting the necessity of distinguishing correlation from causation. Hastie et al. (2009) address the problem of over fitting, which arises when too many predictors are used, and emphasize the need to balance model complexity with the ability to generalize to new data. Belsley et al. (1980) stress the critical nature of verifying the assumptions of linear regression which are linearity, independence, and homoscedasticity since neglecting these can lead to invalid conclusions and misleading outcomes. Harlow and Mulaik (2014) criticize the excessive dependence on R-squared as the only indicator of model fit, noting that it can lead to misguided conclusions about predictive accuracy and overall model quality. They argue that R-squared fails to give a comprehensive view of model fit, which can be deceptive. Wright (1934) highlights the risks associated with using linear regression predictions outside the data's range, stressing that extrapolation can result in considerable errors and inaccurate forecasts.

Belsley (1991) points out that multi-collinearity can increase standard errors and produce misleading coefficient estimates, complicating the evaluation of individual predictors' effects. Steyerberg et al. (2010) stress the importance of model validation to ensure reliability, noting that misuse frequently happens when models are not evaluated on independent datasets.

Carlin and Moreno-Betancur (2024) examined 57 papers published in three leading journals of clinical research: *Pediatrics*, *Neurology* and *BMJ Open* in June 2022. (The journals were selected from top 20 most influential medical journals. 36 of the papers used linear regression. Among

these papers, 25 (69%, or 44% of all papers) exhibited a type of misuse of regression along the lines that we have identified in the table below. 10 papers applied multiple regression to ill-posed questions. Frequent misuse of regression, such as inadequately justified "adjustment for covariates" and erroneous interpretation of estimated coefficients was observed. Table 1 was generated from that data by Carlin and Moreno-Betancur (2024).

Table 1: Research classification and challenges faced by Researchers

Type of research question	N	**Problems found**	Types of Problems
Descriptive	5	3	no justification for adjustment.
Predictive	7	4	inappropriate model developed, also interpretation of coefficients.
Causal	14	8	Univariable, ignoring confounding, unclear about confounding adjustment, problems in method used to identify adjustment set.
Vague/unclear	10	10	all risk factor identification and version of type 2 fallacy.
Total	36	25	

Papers using regression analysis (36 of a total of 57 reviewed) in the 3 journals (June 2022), classified according to the purpose (type of research question) underlying the analysis.

In business and social sciences, linear regression is frequently used to model qualitative data. Researchers in these fields often mistakenly believe that assigning numerical codes to qualitative variables effectively transforms them into quantitative data. A common example of this practice is the modelling of customer satisfaction, a qualitative construct, using linear regression. Customer satisfaction reflects how pleased customers are with their overall experience and is typically assessed through questionnaires. These questionnaires contain opinion-based statements, with responses ranging from "strongly disagree" to "strongly agree," which are then coded numerically from 1 to 5.

In such studies, customer satisfaction is treated as the dependent variable, while factors like price fairness, service quality and product quality, also measured on 5-point Likert scales, are used as independent variables. This approach was observed in seven articles from journals published in the years 2010, 2011, 2018, 2022 (two articles), and 2023 (two articles). Notably, the results of these studies were often reported without meaningful interpretation. Researchers seem to overlook the fact that regression is not merely a model-fitting tool but a method for answering specific research questions. Concerns about the validity of results are largely ignored, and the fundamental assumptions of regression, namely, linearity, independence, homoscedasticity and normality are often disregarded altogether.

Some published articles from these fields have objectives whose achievement has nothing in line with regression analysis. However, because regression analysis is the only tool that is readily available for most of these researchers, regression models are run and results are adopted.

Methodology

Articles on uses and misuses of linear regression were studied. Seven articles that utilized regression analysis to model customer satisfaction were examined. This was meant for establishing the scale on which customer satisfaction was measured. The study aimed at revealing misuses of linear regression analysis.

Five datasets in which customer satisfaction served as the dependent variable were downloaded from Kaggle. Of these, one had age, gender, country, income, product quality, service quality, purchase frequency, feedback score, and loyalty level as independent variables. A scrutiny of the types of the variables, especially the dependent variables, in the study was carried out.

Since Linear regression models are designed for continuous dependent variables, modelling other types of dependent variables is a violation of this important condition. Linear regression assumptions were checked using residual plots to ensure model validity. Key assumptions include linearity, normality of errors, independence of errors, and homoscedasticity.

Data Analysis

The analysis looked mainly at determining when linear regression is misused. Jamovi software was used to generate the tables and plots that follow. The Kaggle dataset that had customer satisfaction as the dependent variable and age, gender, country, income, product quality, service quality, purchase frequency, feedback score, and loyalty level as independent variables was used in the generation of theses plots and tables.

Table 2: Correlation Matrix

		Product-Quality	Service-Quality	Feedback-Score	Loyalty-Level	Age	Income	Satisfaction
Product-Quality	r_s	1.00						
	p-value							
Service-Quality	r_s	0.005	1.00					
	p-value	0.307						
Feedback-Score	r_s	-0,010	0.000	1.00				
	p-value	0.050	0.969					
Loyalty-Level	r_s	0.000	-0.006	-0.001	1.00			
	p-value	0.999	0.233	0.841				
Age	r_s	-0.009	-0.005	0.003	-0.006	1.00		
	p-value	0.071	0.342	0.608	0.243			
Income	r_s	-0.002	0.005	-0.004	-0.006	0.000	1.00	
	p-value	0.753	0.314	0.396	0.224	0.985		
Satisfaction	r_s	0.542	0.547	-0.009	-0.006	0.156	0.242	1.00
	p-value	<0.001	<0.001	0.063	0.287	<0.001	<0.001	

The matrix in table 2 was meant for checking multicollinearity. It is evident from the matrix that there was no multicollinearity among the independent variables. Product-quality, service-quality, age and income were significantly correlated to satisfaction.

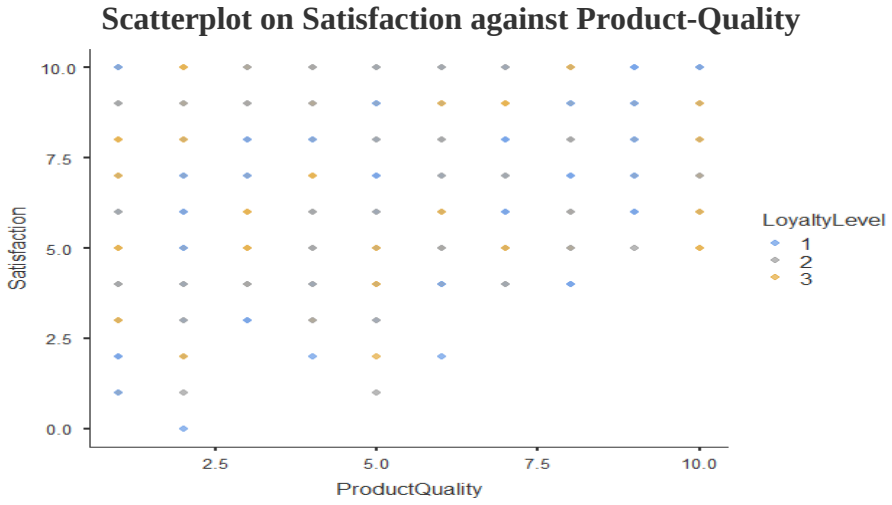

Figure 1

Figure 1 is a scatterplot on Satisfaction against product-quality, grouped by loyalty-level. Although the data satisfied the multicollinearity assumption, the linearity assumption is violated here. It is clear that there is no linear relationship between satisfaction and product-quality. This, in addition to the dependent variable being non-continuous, makes linear regression unsuitable.

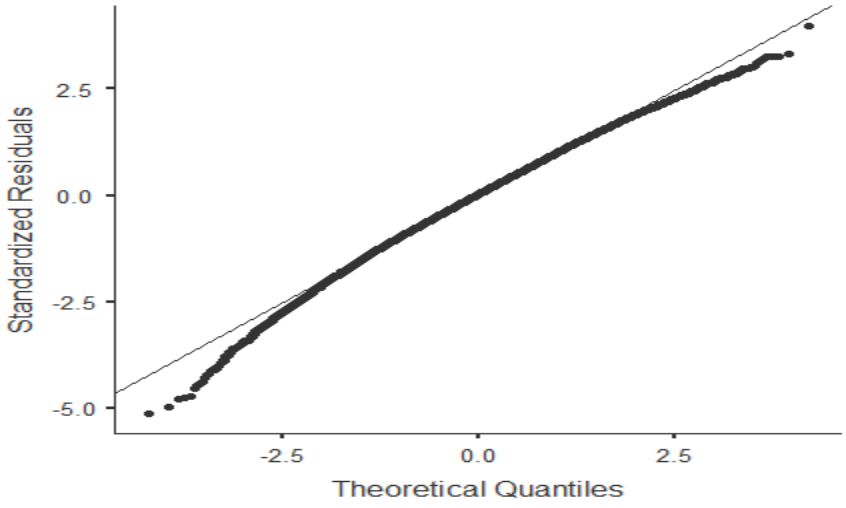

Figure 2

The Q-Q plot in figure 2 shows that the residuals are approximately normally distributed generally, especially around the centre where the points closely follow the line, supporting the normality assumption in that region. There are, however, noticeable deviations in the tails where the points fall away from the line, suggesting potential skewness. The slight deviations from the line indicate mild outliers or non-normal errors.

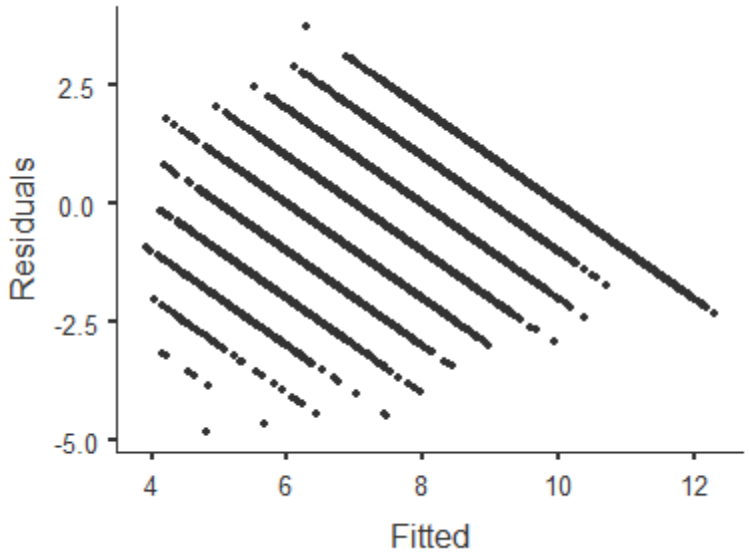

Figure 3

Figure 3 is a residuals-fitted plot which is meant for testing for a number of assumptions that include the homogeneity of variance assumption. The residuals are arranged in a very regular, diagonal banded pattern. The spread of residuals decreases as fitted values increase. This is not typical of good residual plots as these residuals are expected to be randomly scattered around zero. The structured residuals suggest that the model violates the linearity, independence and homoscedasticity assumptions. It may also indicate the presence of ordinal or discrete outcome values, like count data, where residuals naturally cluster in patterns. This plot suggests that a linear regression model is not appropriate for the dataset.

Jamovi software was also used to run linear regression models with each of four other Kaggle datasets on customer satisfaction. For all the datasets, the scatter-plots

Findings

The reviewed articles revealed that the requirement for the dependent variable to be continuous is often overlooked, with researchers mistakenly applying linear regression, not only to discrete but also to ordinal data. Additionally, it was observed that many researchers appear unaware of the assumptions underlying least-squares regression and therefore fail to verify their validity. A lack of familiarity with alternative methods to least-squares regression when assumptions are violated was also noted. Misinterpretation of regression coefficients emerged as another common challenge faced by researchers.

Conclusion

Linear regression remains one of the most widely used and accessible statistical tools across various disciplines due to its simplicity and interpretability. When properly applied, it offers powerful insights for descriptive, predictive, and causal analysis. However, this article highlights

a critical and growing concern: the frequent misuse of linear regression, particularly in fields such as health research, business, and the social sciences. Common errors include applying the model to inappropriate types of data, especially non-continuous dependent variables, failing to test key assumptions, and interpreting results without contextual understanding. The data analysis presented further demonstrates that when foundational assumptions such as linearity, homoscedasticity, normality, and independence are violated, the validity of the results is compromised. The findings underscore the urgent need for researchers to treat regression not just as a model-fitting procedure, but as a question-driven analytical tool that demands careful model selection, assumption checking, and critical interpretation. Without such rigor, the risk of drawing invalid conclusions increases, ultimately undermining the reliability of scientific and applied research.

Recommendations

The following recommendations were made:

i. Advanced statistics courses focusing on statistical issues and data analysis should be incorporated into both taught and research-based Master's and Doctoral programmes.

ii. At least one of the research supervisors at master's and doctoral levels should involve a Statistician from the onset through to the completion of the study.

iii. Reviewers of journals should have an understanding of the basics of statistics so that articles that do not conform to appropriate statistical analysis techniques are not accepted.

References

Agbenyegah, A.T., Zogli, L-K.J., Dlamini, B., Mofokeng, N.E.M. & Kabange, M.M. (2022). Ambient Situation and Customer Satisfaction in Restaurant Businesses: A Management Perspective. African Journal of Hospitality, Tourism and Leisure, 11(2):394-408. DOI: https://doi.org/10.46222/ajhtl.19770720.232.

Allen, M. P. (2004). Understanding regression analysis. Springer Science & Business Media.

Douglas C. Montgomery, Elizabeth A. Peck and G. Geoffrey Vining (2012) INTRODUCTION TO LINEAR REGRESSION ANALYSIS, A JOHN WILEY & SONS, INC., PUBLICATION

Carlin John B. and Moreno-Betancur Margarita (2024) On the uses and abuses of regression models: a call for reform of statistical practice and teaching, Murdoch Children's Research Institute and The University of Melbourne.

Chatterjee, S., & Hadi, A. S. (2006). Regression analysis by example. John Wiley & Sons.

Chatterjee Samprit and Simonoff Jeffrey S. (2013) Handbook of Regression Analysis, John Wiley & Sons, Inc., Hoboken, New Jersey.

Darlington, R. B., & Hayes, A. F. (2017). Regression Analysis and Linear Models: Concepts, Applications, and Implementation. New York: The Guilford Press.

Hanif Muzammil, Hafeez Sehrish and Riaz Adnan (2010) Factors Affecting Customer Satisfaction, International Research Journal of Finance and Economics ISSN 1450-2887 Issue 60.

Leffondré, K., Jager, K. J., Boucquemont, J., Stel, V. S., & Heinze, G. (2014). Representation of exposures in regression analysis and interpretation of regression coefficients: basic concepts and pitfalls. *Nephrology Dialysis Transplantation*, *29*(10), 1806-1814.

Lei Z, Duan H, Zhang L, Ergu D and Liu F (2022) The main influencing factors of customer satisfaction and loyalty in city express delivery. Front. Psychol. 13:1044032. doi: 10.3389/fpsyg.2022.1044032.

Karen Leffondré1, Kitty J. Jager, Julie Boucquemont, Vianda S. Stel and Georg Heinze3 (2013) Representation of exposures in regression analysis and interpretation of regression coefficients: basic concepts and pitfalls, Oxford University Press on behalf of ERA-EDTA. All rights reserved.

Kubasu, K. (2018). Factors influencing customer satisfaction with services offered by Safaricom mobile cellular network (Thesis). Strathmore University. Retrieved from https://suplus.strathmore.edu/handle/11071/6127.

Kumbhar Vijay M. (2011) Factors affecting on customers' satisfaction: an empirical investigation of atm service, Munich Personal RePEc Archive, Online at https://mpra.ub.uni-muenchen.de/32713.

Mehta, D. (2023). What Is Regression Analysis? Types, Importance, and Benefits. Retrieved from https://www.g2.com/articles/regression-analysis.

Nwachukwu, Vitals Obioma and Egbulonu, K.G (2000). Elements of Statistical inference. Owerri Imo State, Nigeria: Peace Enterprises.pp108-109.

Putta Santhosh Samuel (2023) The Effect of Service Quality on Consumer Satisfaction in Restaurants, International Journal for Research Trends and Innovation (www.ijrti.org).

Rahman Md. Habibur, Huq Md. Mahmudul and Ullah Mohammad Enayet (2023) Factors Affecting Customer Satisfaction: An Empirical Study on Telecommunication Sector in Bangladesh, Academic Journal on Science, Technology, Engineering & Mathematics Education (AJSTEME) Volume 3, Issue 2.