African Journal of Social and Behavioural Sciences (AJSBS)
Volume 15, Special Issue (2025) ISSN: 2141-209X

REGRESSION ANALYSIS IN PSYCHOLOGICAL RESEARCH: METHODOLOGICAL CHALLENGES AND CONTEMPORARY ISSUES

Francisca Nkechi Enenta[1]* & Caleb Chukwuebuka Iwuala[2]

[1]Department of Psychology, Nasarawa State University, Keffi, Nigeria

[2]Department of Psychology, Kingsley Ozumba Mbadiwe University, Ideato, Imo State, Nigeria

*nkeonyebuchi@gmail.com

ABSTRACT: Regression analysis, an essential statistical technique in psychological research, is used to examine relationships among variables and to predict outcomes. However, its application is often hindered by various methodological challenges that undermine the reliability of results. These include issues with model specification, assumption violations, multicollinearity, overfitting, and inadequate interpretation. This paper highlights critical gaps in the current use of regression analysis, particularly in psychological research. The research underscores the importance of carefully considering model assumptions and diagnostic tests to avoid flawed conclusions. It also discusses the implications of these issues for psychological research, especially regarding the interpretation of causal relationships and the generalizability of findings. The paper proposes several innovations and recommendations to mitigate these challenges, including improving model specification through theoretical knowledge and exploratory data analysis, conducting rigorous assumption checks, employing advanced techniques such as bootstrapping for model validation, and ensuring clear and accurate reporting of results. These steps are essential for enhancing the validity and reliability of regression-based findings in psychology.

INTRODUCTION

Regression analysis is a fundamental part of statistical methods, providing valuable insights into how variables relate to each other. In its most basic form, it models how one variable, often called the dependent or response variable, depends on one or more independent variables. (Cohen et al., 2020; Ruan et al., 2024). This technique is indispensable in fields as diverse as economics, engineering, and particularly psychology, where it is employed to explore the complex dynamics of human behaviour, predict outcomes, and infer causal relationships (Ruan et al., 2024; Singh et al., 2024). Despite its widespread utility, regression analysis is not without its challenges, especially within psychological research. The method's application in this domain has often been complicated by issues such as model mis-specification, assumption violations, and misinterpretation of results, which can undermine its effectiveness and reliability.

In psychological research, regression techniques are commonly used for two distinct purposes: prediction and causal inference. While the former aims to forecast future outcomes based on historical data, the latter seeks to establish cause-and-effect relationships between variables. However, as has been noted by scholars (e.g., Hastie et al., 2017; Wang et al., 2022), while regression can identify relationships between variables, it does not, by itself, confirm causality. To infer causal effects, researchers must justify the assumptions of causality, ensuring that the relationships hold beyond the observed data. This critical issue highlights a gap in current regression practice, particularly in observational studies where confounding variables may distort the purported causal links.

Despite the essential role of regression in psychological modelling, its application often faces significant hurdles. Key methodological challenges include improper model specification, where the choice of variables or the form of the relationship between them is misjudged; the violation of fundamental assumptions such as linearity and homoscedasticity; and the dangers of overfitting, where models are excessively complex, fitting not only the signal but also the noise in the data (Field, 2018; Tabachnick & Fidell, 2019). These pitfalls can severely compromise the validity of conclusions drawn from statistical models, making it crucial for researchers to adopt more robust and methodologically sound approaches.

The aim of this paper is to address these concerns by critically examining the use of regression analysis in psychological research. Specifically, it seeks to highlight the main methodological gaps that compromise the effectiveness of regression techniques in this field, and to offer practical recommendations for improving the accuracy and reliability of regression models. This discussion is particularly timely, as the increasing complexity of psychological data, often involving large datasets with multiple interacting variables, necessitates more nuanced and rigorous applications of regression techniques.

The importance of regression analysis in psychology cannot be overstated. It serves as an essential tool for understanding the nuanced relationships between psychological variables, from predicting mental health outcomes to analysing the impact of cognitive and social factors on behaviour (Rosenberg, 2013). Regression methods, ranging from simple linear regression to more sophisticated multivariate and hierarchical models, allow researchers to explore complex interrelationships between variables and test hypotheses about causal mechanisms. However, these techniques are not without limitations, and their application must be approached with caution. The use of regression analysis in psychology is complicated by the inherent variability of human behaviour, the challenges of accurately measuring psychological constructs, and the need to account for confounding factors that can obscure true relationships (David, 2009).

The proliferation of machine learning techniques and other advanced statistical methods has further complicated the landscape of regression analysis. While these methods offer exciting possibilities, they also introduce new complexities, particularly in terms of model interpretation and validation. As the field continues to evolve, it is imperative that researchers remain vigilant to the limitations of the tools at their disposal, while also embracing innovations that can enhance the precision and utility of regression models in psychological research (Bollen, 2023).

This paper explores these issues in greater detail, outlining the current state of regression analysis in psychological research, the common methodological challenges it faces, and the potential avenues for improving its application. Through this analysis, it aims to provide a framework for researchers to more effectively employ regression methods, ultimately leading to more accurate and generalisable findings in the study of psychological phenomena.

How Regression Analysis Works

Regression analysis serves as a vital tool for researchers across disciplines, enabling them to explore relationships between variables and derive predictive insights. By quantifying these relationships, regression analysis allows for inferences about how one variable influences another and aids in forecasting future outcomes based on observed data (David, 2009). However, the accurate application of regression depends on a deep understanding of the different types of variables involved and the assumptions that underpin this statistical method. Below, the categories of variables critical to regression analysis are examined, offering a clearer insight into their roles and implications for empirical research.

Dependent Variable

The dependent variable, often regarded as the "outcome" or "response" variable, is the primary focus of any regression analysis. It represents the phenomenon that the researcher aims to understand or predict, whether in terms of sales performance, behavioural changes, or psychological outcomes. In psychological studies, for example, the dependent variable may involve complex constructs such as mental health scores, cognitive performance, or emotional responses. Several key criteria guide the identification of this variable: it must be the variable whose variation is explained by the independent variables, and it is measured after changes in the independent variables have occurred. Thus, the dependent variable is inherently dependent on other factors, both directly and indirectly. Careful consideration of the dependent variable is crucial, as it shapes the entire analytical approach and helps delineate the focus of the study (Hosmer et al., 2013).

Independent Variables

Independent variables, also referred to as predictors or explanatory variables, are those factors assumed to cause or influence changes in the dependent variable. These variables are not influenced by the dependent variable but instead serve as inputs that potentially explain variations in the outcome. In experimental settings, independent variables can be manipulated to observe their effect on the dependent variable, as seen in controlled experiments where the researcher changes one factor while holding others constant. For instance, in an economic study, the price of a product (independent variable) might be adjusted to investigate its effect on consumer demand (dependent variable) (Chicco et al., 2021). A crucial task for researchers is to ensure that independent variables are appropriately selected, based on both theoretical foundations and empirical evidence, to avoid spurious associations.

Explanatory Variables

Explanatory variables are employed in regression models to offer insights into the underlying causes of observed outcomes. These variables help clarify why certain effects or patterns emerge in the dependent variable. For example, in a psychological study examining the impact of stress on performance, explanatory variables such as age, gender, and prior experiences could help elucidate why some individuals are more affected by stress than others. By including explanatory variables in regression models, researchers can control for potential confounding factors and provide a more comprehensive understanding of the relationships between variables (Rosenberg, 2013). It is worth noting that the distinction between independent and explanatory variables can sometimes blur, especially in complex models, necessitating careful conceptual framing and precise statistical modelling.

Predictor Variables

Predictor variables serve a specific purpose within regression analysis: they are used to predict the future values of the dependent variable. These variables are not necessarily causal but are indicative of trends or associations that can inform predictions. For example, in predicting academic achievement, predictor variables might include hours spent studying, class attendance, or socio-economic status. Although these variables can provide valuable forecasts, their predictive power must be assessed with caution. Predictors are inherently more reliable when drawn from a solid theoretical framework and validated through empirical evidence, ensuring that the predictions are both meaningful and generalisable (David, 2009).

Experimental Variables

Experimental variables are those that researchers actively manipulate during an experiment to observe their effects on the dependent variable. These variables are key components of controlled experimental designs, where researchers test hypotheses by introducing systematic changes. For instance, an experiment testing the effects of different teaching methods on student performance would manipulate the teaching method while measuring the resulting performance (Chicco et al., 2021). The manipulation of experimental variables provides a means of establishing causality, allowing researchers to determine whether changes in the dependent variable are attributable to the specific manipulation of independent variables.

Subject Variables (Fixed Effects)

Subject variables, also referred to as fixed effects, are characteristics that vary across subjects but are not manipulated in the experimental design. These include demographic variables such as age, gender, and socioeconomic status, which may influence the outcome but cannot be directly controlled. While researchers cannot alter these subject variables, their inclusion in the model is crucial for understanding how different groups may respond to the independent variables. For example, in a study examining the impact of therapy on mental health, subject variables such as pre-treatment severity of symptoms could significantly affect the treatment's effectiveness, thus necessitating their inclusion as fixed effects in the analysis (Chicco et al., 2021).

Basic Assumptions of Regression Analysis

Regression analysis is built on a set of fundamental assumptions, each of which is critical for ensuring the validity and reliability of the results. These assumptions must be met to avoid biased or misleading conclusions. The following are the key assumptions underlying regression analysis, as outlined by Tabachnick and Fidell (2013):

1. **Linearity**: There is an assumption that the relationship between the independent and dependent variables is linear. This implies that changes in the independent variables are expected to produce proportional changes in the dependent variable. If this assumption is violated, the regression model may fail to accurately capture the nature of the relationship between variables accurately, leading to model mis-specification (Hosmer et al., 2013).
2. **Independence**: Observations must be independent of one another. In longitudinal or time-series data, this assumption is particularly challenging to meet, as data points collected over time may exhibit autocorrelation. In such cases, advanced techniques such as time-series analysis or mixed-effects models may be required to account for dependence structures (Hastie et al., 2017).
3. **Homoscedasticity**: The variance of the residuals (the differences between observed and predicted values) should be constant across all levels of the independent variable. When this assumption is violated, the model may suffer from heteroscedasticity, which can lead to inefficient estimates and incorrect statistical inferences (David, 2009).
4. **Normality**: It is assumed that the residuals follow a normal distribution. While this assumption is less critical in large samples due to the Central Limit Theorem, significant deviations from normality, especially in small samples, can distort the results and affect the statistical tests (Tabachnick & Fidell, 2019).
5. **No Multicollinearity**: The independent variables should not be highly correlated with one another. When multicollinearity exists, it becomes difficult to discern the individual effect of each independent variable, leading to inflated standard errors and unstable coefficient estimates. In practice, diagnostic tools such as Variance Inflation Factors (VIFs) can be used to assess and mitigate multicollinearity (Field, 2018).

Each of these assumptions is fundamental to the robustness of regression analysis. Failure to meet these assumptions can result in misleading conclusions, particularly when used for causal inference or predictive modelling. As such, it is incumbent upon researchers to test these assumptions thoroughly and consider appropriate remedial actions where violations are identified.

Types of Regression

Regression analysis, as a versatile statistical tool used in various fields to understand relationships between variables and make predictions, has many types but for the purpose of this paper, the major ones commonly used by researchers will be discussed below

i. Simple Linear Regression
ii. Multiple Linear Regression
iii. Hierarchical Regression
iv. Logistic Regression
v. Polynomial Regression
vi. Non-linear Regression
vii. Multivariate Linear Regression

viii. Stepwise Regression

Simple Linear Regression

Simple linear regression is used to model the relationship between two variables, with one variable serving as the independent variable (predictor) and the other as the dependent variable (outcome). It is frequently used to determine how a change in one variable affects another. For example, simple linear regression might be used to investigate how the number of prior arrests (independent variable) predicts recidivism rates (dependent variable). Researchers could explore whether individuals with more prior arrests are more likely to re-offend, using regression to quantify this relationship (David, 2009).

Simple linear regression analysis assumes several conditions to be met in order to produce reliable and valid results

i. There must be a relationship between the independent variable (x) and the dependent variable (y) that is linear.
ii. Each observation should be independent of the others.
iii. The variance of the residuals should be constant across all levels of the independent variable.
iv. The residuals should be normally distributed.
v. The independent variable should not be highly correlated with other independent variables.

Type of Hypotheses

Simple Linear Regression can test the following types of hypotheses:

Directional hypothesis: e.g., There is a positive relationship between the amount of fertilizer used and crop yield.

Non-directional hypothesis: e.g., There is a relationship between the amount of fertilizer used and crop yield.

Type of Data to be used

Simple Linear Regression requires the following types of data:

i. **Continuous data:** The dependent variable (y) should be continuous, such as crop yield or height.
ii. **Interval or ratio data**: The independent variable (x) should be interval or ratio data, such as the amount of fertilizer used or temperature.

Multiple Linear Regression

Multiple Linear Regression (MLR) is a statistical technique used to model the relationship between one dependent variable and two or more independent variables. It helps in understanding how various predictors simultaneously influence the outcome. In psychology, MLR can be used to predict academic performance (dependent variable) from multiple behaviors (independent variables), such as study time, sleep habits, and class attendance. This model helps identify how each behavior contributes to academic success, controlling for the influence of the others.

Assumptions of Multiple Linear Regression:

i. The relationship between each independent variable and the dependent variable should be linear.
ii. The residuals (errors) are independent of each other (no autocorrelation).
iii. The variance of the residuals should be constant across all levels of the independent variables.
iv. The residuals should be normally distributed.
v. The independent variables should not be highly correlated with each other.
vi. The model includes all relevant variables and excludes irrelevant ones.

Type of Hypotheses

Multiple Linear Regression can test the following types of hypotheses:

i. **Directional hypothesis:** e.g., There is a positive relationship between the amount of fertilizer used, temperature, and crop yield.
ii. **Non-directional hypothesis:** e.g., There is a relationship between the amount of fertilizer used, temperature, and crop yield.
iii. **Comparative hypothesis:** e.g., The relationship between the amount of fertilizer used and crop yield is stronger than the relationship between temperature and crop yield.

Type of Data

Multiple Linear Regression requires the following types of data:

i. **Continuous data**: The dependent variable (y) should be continuous, such as crop yield or height.
ii. **Interval or ratio data:** The independent variables (x) should be interval or ratio data, such as the amount of fertilizer used, temperature, or pH level.
iii. **Categorical data**: Multiple Linear Regression can also accommodate categorical independent variables, such as the type of fertilizer used or soil type.

Hierarchical Regression

Hierarchical Regression is a statistical technique used to examine the incremental value of adding new predictors to a model. Researchers add predictors step by step to determine how much each new variable improves the model's explanatory power. In psychology, hierarchical regression can be used to examine how social support (Step 1) and coping strategies (Step 2) jointly predict mental health outcomes. Initially, social support would be entered, followed by coping strategies, to see how much coping strategies explain additional variance in mental health after accounting for social support.

Assumptions of Hierarchical Regression:

i. The relationship between each independent variable and the dependent variable should be linear.
ii. Each observation should be independent of the others.
iii. The variance of the residuals should be constant across all levels of the independent variables.
iv. The residuals should be normally distributed.
v. The independent variables should not be highly correlated with each other.
vi. The independent variables should be entered into the model in a hierarchical or sequential manner, with the most important variables entered first.

Type of Hypotheses

Hierarchical Regression can test the following types of hypotheses.

i. **Sequential hypothesis**: e.g., The addition of variable X to the model significantly improves the prediction of the dependent variable, beyond what is predicted by variable Y.
ii. **Hierarchical hypothesis**: e.g., Variable Y moderates the relationship between the dependent variable and variable X.
iii. **Mediation hypothesis:** e.g., Variable X affects the dependent variable through its effect on variable Y.

Type of Data

Hierarchical Regression requires the following types of data.

i. **Continuous data:** The dependent variable (y) should be continuous, such as test scores or employee satisfaction.
ii. **Interval or ratio data**: The independent variables (x) should be interval or ratio data, such as age, income, or years of experience.
iii. **Categorical data**: Hierarchical Regression can also accommodate categorical independent variables, such as gender, ethnicity, or occupation.

Logistic Regression

Logistic Regression is a statistical method used to model binary or categorical outcomes based on one or more predictor variables. It estimates the probability that a given input will result in a particular outcome (Efron & Tibshirani, 1993). In psychology, logistic regression could be used to predict whether an individual engages in risky behavior (e.g., substance use) based on predictors like peer influence, stress levels, and family history. The outcome is binary: either the person engages in risky behavior (1) or does not (0).

Assumptions of Logistic Regression

i. The dependent variable should be binary, meaning it has only two possible outcomes (e.g., 0/1, yes/no, etc.).
ii. Each observation should be independent of the others.
iii. The relationship between the independent variables and the logit (i.e., the log-odds of the dependent variable) should be linear.
iv. The independent variables should not be highly correlated with each other.
v. The variance of the residuals should be constant across all levels of the independent variables.

Type of Hypotheses

Logistic Regression can test the following types of hypotheses:

i. **Directional hypothesis:** e.g., There is a positive relationship between the amount of education and the likelihood of being employed.
ii. **Non-directional hypothesis**: e.g., There is a relationship between the amount of education and the likelihood of being employed.
iii. Comparative hypothesis: e.g., The likelihood of being employed is higher for individuals with a university degree compared to those with only a secondary education.

Type of Data

Logistic Regression requires the following types of data:

i. **Binary dependent variable**: The dependent variable should be binary, meaning it has only two possible outcomes (e.g., 0/1, yes/no, etc.).
ii. **Continuous or categorical independent variables**: The independent variables can be continuous (e.g., age, income) or categorical (e.g., gender, education level).

Polynomial Regression

Polynomial Regression is a type of regression analysis that models the relationship between the independent and dependent variables as an nth-degree polynomial. It is used when the relationship between variables is nonlinear, allowing for curves rather than straight lines. In psychology, polynomial regression could be used to examine how levels of daily exercise (independent variable) impact mood (dependent variable), where the effect of exercise may not be linear (Fotheringham & Wong, 2022).. For example, moderate exercise could improve mood, but excessive exercise might negatively affect it. Polynomial regression would capture this curve.

Assumptions of Polynomial Regression:

i. The relationship between the independent variable(s) and the dependent variable should be linear in the parameters.

ii. Each observation should be independent of the others.
iii. The variance of the residuals should be constant across all levels of the independent variable(s).
iv. The residuals should be normally distributed.
v. The independent variables should not be highly correlated with each other.

Type of Hypotheses

Polynomial Regression can test the following types of hypotheses:

i. **Directional hypothesis**: e.g., There is a positive curvilinear relationship between the amount of fertilizer used and crop yield.
ii. Non-directional hypothesis: e.g., There is a curvilinear relationship between the amount of fertilizer used and crop yield.
iii. Comparative hypothesis: e.g., The relationship between the amount of fertilizer used and crop yield is more curvilinear than linear.

Type of Data

Polynomial Regression requires the following types of data.

i. **Continuous dependent variable**: The dependent variable should be continuous, such as crop yield or height.
ii. **Continuous independent variable(s):** The independent variable(s) should be continuous, such as the amount of fertilizer used or temperature.
iii. **Polynomial terms:** The model can include polynomial terms, such as squared or cubed terms, to capture non-linear relationships

Non-linear Regression

Non-linear regression is a statistical technique used when a straight line cannot adequately describe the relationship between the independent and dependent variables. Instead, the relationship is represented by a curve or a more complex model. For example, non-linear regression could be used to study how stress levels (independent variable) affect performance (dependent variable). Research might find that performance improves up to a certain point as stress increases, but then declines once stress surpasses an optimal level, indicating a curvilinear or non-linear relationship.

Assumptions of Non-linear Regression

i. The relationship between the independent variable(s) and the dependent variable should be non-linear.
ii. Each observation should be independent of the others.
iii. The variance of the residuals should be constant across all levels of the independent variable(s).
iv. The residuals should be normally distributed.

v. The independent variables should not be highly correlated with each other.
vi. The non-linear relationship should be smooth and continuous.

Type of Hypotheses

Non-linear Regression can test the following types of hypotheses:

i. **Directional hypothesis**: e.g., There is a positive non-linear relationship between the amount of fertilizer used and crop yield.
ii. **Non-directional hypothesis**: e.g., There is a non-linear relationship between the amount of fertilizer used and crop yield.
iii. Comparative hypothesis: e.g., "The relationship between the amount of fertilizer used and crop yield is more non-linear than linear."

Type of Data

Non-linear Regression requires the following types of data:

i. **Continuous dependent variable**: The dependent variable should be continuous, such as crop yield or height.
ii. **Continuous independent variable(s):** The independent variable(s) should be continuous, such as the amount of fertilizer used or temperature.
iii. **Non-linear terms**: The model can include non-linear terms, such as exponential, logarithmic, or polynomial terms, to capture non-linear relationships.

Multivariate Linear Regression

Multivariate linear regression extends multiple linear regression by including more than one dependent variable alongside multiple independent variables. This method is particularly valuable for addressing complex real-world situations in which multiple factors influence multiple outcomes. For example, a business might use multivariate linear regression to assess the impact of multiple marketing strategies across different regions, with sales figures, customer engagement, and brand awareness as dependent variables. This approach provides a more holistic and realistic analysis, but its complexity often requires advanced statistical software to interpret the data accurately.

Assumptions of Multivariate Linear Regression:

i. The relationship between each independent variable and the dependent variable should be linear.
ii. Each observation should be independent of the others.
iii. The variance of the residuals should be constant across all levels of the independent variables.
iv. The residuals should be normally distributed.
v. The independent variables should not be highly correlated with each other.

vi. The residuals should not be autocorrelated.

Type of Hypotheses

Multivariate Linear Regression can test the following types of hypotheses.

i. Directional hypotheses: e.g., There is a positive relationship between the amount of fertilizer used, temperature, and crop yield.
ii. Non-directional hypotheses: e.g., There is a relationship between the amount of fertilizer used, temperature, and crop yield.
iii. Comparative hypotheses: e.g., The relationship between the amount of fertilizer used and crop yield is stronger than the relationship between temperature and crop yield.

Type of Data

Multivariate Linear Regression requires the following types of data:

i. Continuous dependent variable: The dependent variable should be continuous, such as crop yield or height.
ii. Continuous independent variables: The independent variables should be continuous, such as the amount of fertilizer used, temperature, or pH level.
iii. Multiple independent variables: The model can include multiple independent variables to capture the relationships between each variable and the dependent variable.

Stepwise Regression

Stepwise Regression is a method for selecting the most important predictors for a dependent variable by adding or removing variables based on statistical criteria (e.g., AIC or p-values). It proceeds in steps, either forward (by adding variables) or backward (by removing variables), to build the best model (Cohen et al., 2013). In psychology, stepwise regression could be used to predict aggressive behavior in adolescents. Independent variables might include family conflict, peer influence, and school performance. The model would determine which of these factors are the most significant predictors of aggression by systematically adding or removing them.

Assumptions of Stepwise Regression:

i. The relationship between each independent variable and the dependent variable should be linear.
ii. Each observation should be independent of the others.
iii. The variance of the residuals should be constant across all levels of the independent variables.
iv. The residuals should be normally distributed.
v. The independent variables should not be highly correlated with each other.
vi. The residuals should not be autocorrelated.
vii. The F-statistic should be significant, indicating that the model is a good fit to the data.

Type of Hypotheses

Stepwise Regression can test the following types of hypotheses:

i. **Directional hypothesis:** e.g., There is a positive relationship between the amount of fertilizer used and crop yield.
ii. Non-directional hypothesis: e.g., There is a relationship between the amount of fertilizer used and crop yield.
iii. Comparative hypothesis: e.g., The relationship between the amount of fertilizer used and crop yield is stronger than the relationship between temperature and crop yield.

Type of Data

Stepwise Regression requires the following types of data:

i. **Continuous dependent variable**: The dependent variable should be continuous, such as crop yield or height.
ii. Continuous independent variables: The independent variables should be continuous, such as the amount of fertilizer used, temperature, or pH level.
iii. Multiple independent variables: The model can include multiple independent variables to capture the relationships between each variable and the dependent variable.

How to Perform Regression Analysis

i. Data collection and preparation: Gather and clean data, ensuring it meets assumptions like linearity and independence.
ii. Selecting the appropriate regression model: Choose the correct type of regression (linear, polynomial, etc.) based on the data and research objectives.
iii. Data analysis and interpretation: Analyze results, assess model accuracy, and interpret coefficients to draw meaningful conclusions.
iv. Model evaluation and validation: Test the model's performance using metrics like R-squared, mean-squared error, or cross-validation.
v. Using software tools: Use Excel, Python, or R to perform regression analysis efficiently.

Application of regression analysis

Regression analysis has numerous applications in various fields, including: Business and Economics, Healthcare and Medicine, Social Sciences, Environmental Science, Engineering and Technology. These are just a few examples of the many applications of regression analysis. The technique can be applied in any field where relationships between variables need to be understood and predicted.

Uses of regression analysis

Below are the uses of regression analysis in psychological research

i. **Predicting outcomes:** Regression analysis can be used to predict continuous outcomes, such as student grades, stock prices, or blood pressure.
ii. **Identifying relationships:** Regression analysis can help identify the relationships between variables, such as the relationship between exercise and weight loss.
iii. **Controlling for confounding variables**: Regression analysis can help control for confounding variables that may affect the relationship between variables.
iv. **Analyzing the impact of variables**: Regression analysis can help analyze the impact of individual variables on an outcome, such as the impact of education on income.
v. **Modeling complex relationships**: Regression analysis can be used to model complex relationships between variables, such as non-linear relationships or interactions between variables.
vi. **Validating models**: Regression analysis can be used to validate models by comparing predicted outcomes with actual outcomes.
vii. **Identifying outliers**: Regression analysis can help identify outliers or unusual patterns in data.
viii. **Segmenting data**: Regression analysis can be used to segment data into different groups based on their characteristics.

Critical Issues of Regression Analysis

Regression analysis has, over the years, proven to be a powerful statistical method used to examine the relationships between variables. In psychological research, regression analysis is widely used to examine the relationships between predictor variables and outcome variables. However, there are critical issues and methodological challenges associated with the use of regression analysis in psychological research.

Model Specification

Model specification refers to the process of selecting the predictor variables and the functional form of the relationships between the predictor variables and the outcome variable. Poor model specification can result in biased estimates and incorrect conclusions (Cohen et al., 2013).

Assumption Checking

Regression analysis assumes that the data meet certain assumptions, including linearity, homoscedasticity, normality, and independence. Failure to check these assumptions can result in incorrect conclusions (Tabachnick & Fidell, 2013).

Multicollinearity

Multicollinearity occurs when two or more predictor variables are highly correlated. This can result in unstable estimates and inflated variance (Cohen et al., 2013).

Overfitting

Overfitting occurs when a model is too complex and fits the noise in the data rather than the underlying patterns. This can result in poor generalizability of the findings (Hastie et al., 2009).

Interpretation

Regression analysis provides estimates of the relationships between predictor variables and outcome variables. However, these estimates must be interpreted appropriately, taking into account the research context and the study's limitations (Cohen et al., 2013).

Methodological Challenges of Regression Analysis

Measurement Error: Measurement error occurs when the measures used to assess the predictor and outcome variables are unreliable or invalid. This can result in biased estimates and incorrect conclusions (Bollen, 1989).

Sampling Bias: Sampling bias occurs when the sample is not representative of the population. This can result in biased estimates and incorrect conclusions (Cohen et al., 2013).

Longitudinal Designs: Longitudinal designs involve collecting data from the same participants over time. This can result in complex data structures and require specialized analytical techniques.

Missing Data: Missing data occurs when some participants do not provide complete data. This can result in biased estimates and incorrect conclusions (Schafer & Graham, 2002).

Conclusion

Conclusively, regression analysis is a powerful statistical method used in psychological research to examine relationships between variables. However, there are critical issues and methodological challenges associated with the use of regression analysis in psychological research. Regression aids forecasting, risk assessment, and identifying trends. Regression is widely employed in various fields, including psychology, to understand associations, make predictions, and test hypotheses about causal relationships. Yet regression is not free of issues and challenges that temper the predicted outcome. By being aware of these issues and challenges, researchers can use regression analysis effectively and ensure the validity and reliability of their findings.

Recommendations

The following recommendations are made.

i. It is recommended that a careful Model Specification should be done to use theoretical knowledge and exploratory data analysis to inform model specification
ii. It was also recommended that there should be an assumption check by researchers, diagnostic tests, and visual inspections to verify assumptions
iii. The study further recommended that there should be a model validation, and the use of higher advanced techniques like bootstrapping and jackknifing to evaluate model performance will be better
iv. Finally, the study recommended that there should be a clear interpretation and reporting: Provide clear and accurate interpretations of results, including effect sizes and confidence intervals.

REFERENCES

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

Chicco, D. O., Warrens, M., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE, and RMSE in regression analysis evaluation. *PeerJ Computer Science, 7*, e623. https://doi.org/10.7717/peerj-cs.623

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Routledge.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2020). *Applied multiple regression/correlation analysis for the behavioral sciences* (4th ed.). Routledge.

David, A. F. (2009). *Statistical models: Theory and practice*. Cambridge University Press.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman and Hall.

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage Publications.

Fotheringham, A., & Wong, D. (2022). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A, 23*(7), 1025–1044. https://doi.org/10.1177/0305731X20958568

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.

Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2005). *Applied linear regression models* (5th ed.). McGraw-Hill.

Ratkowsky, D. A. (2004). *Handbook of nonlinear regression models*. Marcel Dekker.

Rosenberg, J. (2013). *The Oxford handbook of quantitative methods*. Oxford University Press.

Ruan, Y. (2024). Exploring multiple regression models: Key concepts and applications. *Journal of Quantitative and Predictive Techniques, 1*(1), 1–4. https://doi.org/10.61173/yjpt3s59

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Singh, R., & Singh, R. (2024). Analyzing the relationship between product buying behavior and individual salary: A classification and regression analysis. *ShodhKosh: Journal of Visual and Performing Arts, 5*(6), 1110–1116. https://doi.org/10.29121/shodhkosh.v5.i6.2024.2110

Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson.

YangJing, L. (2009). Human age estimation by metric learning for regression problems. In *Proceedings of the International Conference on Computer Analysis of Images and Patterns* (Vol. 3, Issue 2, pp. 74–82). https://doi.org/10.1007/978-3-642-04463-8_10