KEPES
http://scholopress.com/kepes-journal

2023 Volume 21 Issue 3 pp. 727-740

Application of Regression Analysis in Advance Research: A Literature Review

***Ho, T. T. H**[1]

[1]Centre of Postgraduate Studies, Lincoln University College, Selangor, Malaysia

Abstract

This comprehensive study delves into the multifaceted applications and challenges of regression analysis across a wide array of academic disciplines, including economics, finance, healthcare, medicine, social sciences, environmental science, and engineering. Employing a rigorous systematic literature review methodology, the research scrutinizes peer-reviewed articles from the last 20 years to assess how different types of regression models—such as linear, logistic, ridge, lasso, and elastic net—are utilized to address complex research questions. The literature review further explores the adaptability and robustness of regression analysis in facilitating interdisciplinary investigations. It also critically examines the inherent limitations and challenges associated with the use of regression models, including issues like overfitting, multicollinearity, and ethical concerns. Special attention is given to the trade-offs between predictive power and interpretability, as well as the importance of domain-specific expertise for effective model application. The study concludes by highlighting the indispensable role of regression analysis in advancing human understanding through data-driven research, while also cautioning researchers about the challenges and limitations that must be taken into account for future rigorous and responsible scholarship.

Keywords: Regression Analysis, Systematic Literature Review, Multidisciplinary Applications, Ethical Concerns, Advanced Research, Challenges

1. Introduction

1.1 Background

In the current research landscape, the emergence of big data and the increasing complexity of research questions have presented significant challenges for researchers across various disciplines. Given this backdrop, the intricate nature of data as well as the nuanced relationships between variables mandates the utilization of analytical tools that are both robust and versatile. Emerging as a response to this need, regression analysis stands out as not merely a powerful methodological framework but also as an indispensable tool capable of tackling an expansive range of research queries. This technique transcends the bounds of basic statistical methods; it serves as a comprehensive analytical lens that empowers researchers to interpret multifaceted datasets, make accurate predictions, and subsequently draw insightful conclusions. Its applicability extends from simple linear relationships to intricate interactions involving multiple variables, making it a preferred method for contemporary, data-driven research (Wu et al., 2020).

In the broader context of big data research, numerous ongoing challenges necessitate concerted efforts from researchers. Younas (2019) explicitly highlights the pressing need for further exploration into the

*Corresponding author

DOI https://doi.org/10.6084/m9.figshare.24182547#126

challenges posed by big data. In line with this perspective, the author emphasizes that despite the significant advancements that have been made in big data systems, a myriad of open challenges persistently require attention. Complementing this viewpoint, Huang and associates (2015) argue that the concept of big data has evolved far beyond mere considerations of data volume. They advocate for a paradigm shift towards "No-Boundary Thinking" as a novel approach to tackle the multifaceted challenges intrinsic to big data research. Building on these insights, Mathias and colleagues (2018) pivot the discussion toward the specialized field of clinical research. They underscore the urgency for a comprehensive approach designed to fully leverage the untapped potential of big data in healthcare settings.

In summation, regression analysis has become a powerful analytical framework for addressing complex research questions in the era of big data. Its versatility and ability to handle intricate relationships between variables make it a valuable tool for data-driven research. However, challenges related to data availability, quality, and the broader context of big data research need to be addressed to fully harness the potential of regression analysis and big data in scientific inquiry.

1.2 Rationale

This paper aspires to serve as more than just a technical guide for using regression models; it aims to offer a panoramic view of how regression analysis is integrated into the very fabric of advanced research methodologies. Given its widespread usage, it is crucial to understand not just the 'how' but also the 'why' behind the choice of regression analysis in various research settings. By offering a comprehensive overview, this paper intends to be a pivotal resource for researchers, both novice and experienced, looking to navigate the complex terrains of advanced research through the lens of regression analysis.

1.3 Research Objectives

The objectives of this paper are multi-faceted:

1. **Classification of Regression Models**: To categorize the types of regression models that are most commonly employed in advanced research, including but not limited to linear regression, logistic regression, and ridge regression.
2. **Interdisciplinary Applications**: To explore how regression analysis serves as a universal language in the research community, finding applications in diverse disciplines such as psychology, economics, healthcare, and environmental science. Case studies from each of these disciplines will be examined to highlight the versatility of regression analysis.
3. **Challenges and Limitations**: To delve into the potential pitfalls, ethical considerations, and limitations of using regression analysis. This will include a discussion on issues such as overfitting, multicollinearity, and the interpretability of models.

1.4 Scope and Limitations

The scope of this paper is broad in terms of the disciplines covered but is specifically focused on the use of regression analysis within the context of advanced research. To maintain the rigor and credibility of the review, the paper is confined to studies that have been published in peer-reviewed academic journals. While this ensures a high standard of scholarship, it also introduces a limitation by excluding potentially valuable insights that may be found in gray literature, conference papers, or unpublished works.

2. Literature Review

2.1 Evolution of Regression Models

The research landscape has undergone significant changes over the past few decades, becoming increasingly complex and data-driven. In response to this evolving landscape, regression models have also advanced

considerably, adapting to accommodate a broader array of research questions and data structures. This section is about to offer a comprehensive overview of how regression models have evolved, transforming from simple linear models to more sophisticated forms to meet the rising demands of contemporary research.

2.1.1 Simple Linear Regression: The Genesis

Regression models have evolved significantly over the past few decades to meet the demands of contemporary research. The foundation of regression models lies in simple linear regression, which originated in the early 19th century. Simple linear regression was initially used to predict a single outcome variable based on one predictor variable (Nguyen et al., 2021). However, as research became more complex and data-driven, regression models needed to adapt. Researchers started developing data-driven models that could handle a broader array of research questions and data structures. These models were designed to improve the accuracy and flexibility of predictions. As a demonstration, in the field of hydrology, data-driven modeling techniques were investigated to improve predictive capabilities. Researchers conducted an extensive data-driven modeling experiment to explore the potential of these techniques (Elshorbagy et al., 2009). Similarly, in the field of concrete engineering, researchers compared the accuracy and flexibility of simple linear regression models with more advanced data mining techniques for predicting concrete properties. Interestingly, the use of advanced techniques was found to provide better results. To accommodate the increasing complexity of research questions, regression models have also evolved to include more sophisticated forms. Researchers have explored the use of ensemble methods, such as additive regression and bagging, to improve prediction accuracy (Omran et al., 2016). Furthermore, data-driven differential equation discovery techniques have been developed to allow for the discovery of partial derivative equations when a priori information is insufficient (Maslyaev et al., 2019). On the whole, regression models have undergone significant advancements to meet the rising demands of contemporary research. From simple linear regression to more sophisticated forms and data-driven techniques, these models have evolved to improve prediction accuracy and flexibility in various fields of study.

2.1.2 Multiple Linear Regression: A Natural Extension

Multiple linear regression is a statistical model that allows for the inclusion of multiple predictor variables, providing a more comprehensive understanding of complex relationships (Porter et al., 1994). This extension of simple linear regression is particularly useful in research scenarios where multiple factors contribute to an outcome, capturing the interconnectedness of variables. The use of multiple linear regression has been applied in various fields, including environmental science, computer science, and structural engineering. To illustrate, in the field of environmental science, Stelson (1990) used multiple linear regression to predict urban aerosol refractive index. In computer science, Kang and colleagues (2017) compared the performance of multiple linear regression with other models for concrete dam health monitoring. They found that multiple linear regression provided reliable results and was computationally efficient. Additionally, in the field of mathematics, Marill (2004) discussed the concepts and applications of multiple linear regression in medical research. Multiple linear regression is commonly used in medical research to model observational data and study the relationship between multiple independent predictor variables and a single dependent outcome variable.

One advantage of multiple linear regression is its ability to handle complex relationships between predictor variables and the outcome variable. Deterministic models based on physical laws often require solving complex differential equations, while multiple linear regression offers a simpler formulation and faster execution. Furthermore, multiple linear regression allows for the consideration of any type of correlation between predictor variables and the response variable, which can be established using various regression methods (Kang et al., 2017).

In brief, multiple linear regression is a valuable statistical tool that extends the capabilities of simple linear regression by allowing for the inclusion of multiple predictor variables. It has been widely used in various fields to model complex relationships and capture the interconnectedness of variables. Multiple linear regression offers advantages such as a simple formulation, faster execution, and the ability to consider correlations between variables. Its applications range from environmental science to computer science and medical research.

2.1.3 Logistic Regression: Bridging Dichotomous Outcomes

Logistic regression is a statistical method that allows for the prediction of categorical outcome variables, particularly binary outcomes (Howel & Kleinbaum, 1995). This makes it highly applicable in fields such as healthcare, where outcomes are often dichotomous, such as the presence or absence of a disease. The use of logistic regression in healthcare research has been well-established. As a case in point, in a study on education-based gaps in eHealth, logistic regression was used to predict various eHealth behaviors (Amo, 2016). Logistic regression has also been used in the analysis of subarachnoid hemorrhage outcomes, where factors associated with outcome were assessed using multivariable logistic regression analysis (Achrén et al., 2021). Logistic regression has also been applied in other fields, such as genetics and psychopathology. In a study on genotype '×' environment interaction in psychopathology, logistic regression was used to analyze dichotomous outcomes and detect significant interactions (Eaves, 2006). Similarly, logistic regression was used in simulations to predict membership of upper or lower groups based on the main effects and interactions of genes and environment. In addition to its application in various fields, logistic regression has been compared to other regression methods. For instance, logistic regression has been compared to XGBoost in predicting motor insurance claims using telematics data (Pesantez-Narvaez et al., 2019). The findings showed that logistic regression had good predictive capacity and interpretability, making it a suitable model for the task. It is worth noting that logistic regression assumes certain conditions, such as the linearity of the relationship between predictors and the log odds of the outcome. Violations of these assumptions can affect the validity of the results (Howel & Kleinbaum, 1995). Nevertheless, logistic regression remains a widely recognized and used method for predicting binary outcomes (Pesantez-Narvaez et al., 2019).

In summary, logistic regression is a valuable statistical method for predicting categorical outcomes, particularly binary outcomes. It has been widely applied in various fields, including healthcare, genetics, and psychopathology. Logistic regression offers interpretability and good predictive capacity, making it a suitable choice for many research tasks. However, researchers should be mindful of the assumptions and limitations of logistic regression when applying it to their data.

2.1.4 Ridge and Lasso Regression: Tackling Multicollinearity

Multicollinearity is a common issue in regression analysis when there are significant correlations among the predictor variables. It can lead to unstable and unreliable estimates of the regression coefficients, making it difficult to interpret the relationships between the predictors and the response variable accurately (Shrestha, 2020).

To address the problem of multicollinearity, researchers have developed regularization techniques such as ridge regression and lasso regression. Ridge regression adds a penalty term to the least squares' objective function, which shrinks the regression coefficients towards zero and reduces their variance. This helps to stabilize the estimates and mitigate the effects of multicollinearity (Friedman et al., 2010). Lasso regression, on the other hand, not only shrinks the coefficients but also performs variable selection by setting some coefficients to exactly zero. This feature makes lasso regression particularly useful in high-dimensional data settings, where there are many predictors but only a few are truly relevant. Given that, both ridge and lasso regression have been shown to be effective in handling multicollinearity and improving the predictive

performance of regression models. They provide a trade-off between bias and variance, allowing for more flexibility in model fitting while avoiding overfitting (Enwere et al., 2023). In terms of implementation, ridge and lasso regression can be solved using various algorithms. One popular approach is cyclical coordinate descent, which iteratively updates the coefficients by optimizing one coordinate at a time while keeping the others fixed. This algorithm has been shown to be computationally efficient and can handle large-scale problems with sparse features (Friedman et al., 2010).

To detect multicollinearity in regression analysis, researchers have proposed several techniques. These include examining the correlation matrix of the predictor variables, calculating the variance inflation factor (VIF), and conducting hypothesis tests for the presence of multicollinearity. More advanced regression procedures, such as principal components regression and weighted regression, can also be used (Shrestha, 2020). Overall, ridge and lasso regression are valuable tools for tackling multicollinearity in regression analysis. They provide a way to manage correlated predictors while minimizing the risk of overfitting. These techniques are particularly useful in high-dimensional data settings and have been shown to improve the predictive performance of regression models. Researchers can use various algorithms to solve ridge and lasso regression problems, and there are several techniques available for detecting multicollinearity in regression analysis.

2.1.5 Generalized Linear Models and Beyond

Generalized linear models (GLMs) have become increasingly popular in recent years due to their flexibility in accommodating different types of distributional assumptions and nested data structures. GLMs are particularly useful in the insurance industry, where they are used to support critical decisions (Jong & Heller, 2008). A comprehensive book by Jong & Heller (2008) provides a practical and rigorous treatment of GLMs, covering all standard exponential family distributions and extending the methodology to correlated data structures. The book also addresses specific issues related to insurance data, such as model selection in the presence of large data sets and the handling of varying exposure times. Exercises and data-based practicals are included to help readers consolidate their skills. While the book is package-independent, it includes SAS code and output examples in an appendix and on the companion website. Besides, R code and output for all examples are provided on the website. In addition to GLMs, generalized linear mixed models (GLMMs) and hierarchical linear models (HLMs) have also gained prominence. HLMs are particularly useful for analyzing data with a clustered structure, which is common in psychological research. HLMs allow for the simultaneous investigation of relationships within a given hierarchical level and across levels (Woltman et al., 2012). Woltman and associates (2012) provide an introduction to hierarchical linear modeling (HLM). They explain that HLM allows for the simultaneous investigation of relationships within a given hierarchical level and across levels. The notation employed by Raudenbush and Bryk is used in this journal article. Later, Mcneish et al. (2017) discuss the use of HLMs in psychology and the behavioral sciences. They highlight that HLMs and their extensions for discrete outcomes are popular methods for modeling clustered data. However, they also note that other methods exist for modeling clustered data and researchers should consider the specific requirements of their research design (McNeish et al., 2017).

All things considered, GLMs, GLMMs, and HLMs offer researchers a high degree of flexibility in analyzing complex research designs with different types of distributional assumptions and nested data structures. These models have found applications in various fields, including insurance, psychology, and the behavioral sciences. Researchers should carefully consider the specific requirements of their data and research design when choosing the appropriate model.

2.2 Interdisciplinary Applications

The utility of regression analysis extends far beyond the confines of any single academic discipline. In fact, its adaptability and methodological robustness make it a common analytical thread that weaves through a

diverse array of fields. This section aims to shed light on the interdisciplinary applications of regression analysis, illustrating how it has become an indispensable tool for researchers working on complex questions in various domains.

2.2.1 Economics and Finance

Regression analysis is a versatile and robust statistical tool that finds applications in various fields, including economics and finance. In these domains, regression models are commonly used to analyze market trends, understand consumer behavior, and evaluate the impact of fiscal policies (Lu, 2010). Particularly, multiple linear regression can be employed to examine how different economic indicators affect stock market performance.

In the field of economics, regression analysis is used to study the relationship between various economic variables and their impact on economic outcomes. Researchers often employ regression models to analyze the effectiveness of fiscal policies in promoting economic stability and growth. For instance, studies like the one by Lee & Sung (2008) focus on general fiscal policies, while others, like the meta-regression analysis conducted by Heinemann et al. (2016) examine the impact of fiscal rules on fiscal policy. The study of Heinemann and associates has found that fiscal rules have a constraining impact on fiscal policy, although this impact is weakened when refined identification strategies are employed. Another investigation led by Samsuddin (2021) has delved into the efficacy of monetary and fiscal policies under the conditions of market instability. This study infers that the execution of fiscal policy is largely influenced and governed by political dynamics.

In the field of finance, regression analysis serves as a pivotal analytical tool for scrutinizing market trends, projecting stock price movements, and assessing investment strategies. Numerous studies, such as the one conducted by Lu (2010), have leveraged regression models to explore the intricate relationships between diverse financial indicators and stock market performance. In a similar case, Liu & Zhan (2019) employ regression analysis to evaluate the financing efficiency of agricultural companies listed in China. Their research findings indicate that the incorporation of machine learning techniques holds potential for enhancing the financing efficiency of such entities.

Collectively, regression analysis is a valuable tool in economics and finance, allowing researchers to analyze complex relationships between variables and make informed decisions based on empirical evidence. By employing regression models, researchers can gain insights into market trends, consumer behavior, and the effectiveness of fiscal policies, among other areas of interest.

2.2.2 Healthcare and Medicine

Logistic regression is a commonly used statistical method in the healthcare sector for predicting binary outcomes, such as the presence or absence of a disease, based on a set of predictors (Rao et al., 1998). It is a valuable tool for assessing associations between various factors and health outcomes. For example, in a study conducted by Risher et al. (2013), logistic regression was used to assess the relationship between fear of seeking healthcare and disclosure of same-sex practices among men who have sex with men (MSM) in Swaziland. The study found that fear of seeking healthcare was significantly associated with factors such as experiencing legal discrimination, feeling suicidal, and having been raped. Meanwhile, disclosure of same-sex practices to a healthcare provider was significantly associated with factors such as education level, condom use, and suicidal ideation (Risher et al., 2013).

Survival analysis, a specialized form of regression, is another commonly employed method in healthcare research. It is used to study time-to-event data, such as patient survival rates. During the investigation by Rao and associates (1998), survival analysis techniques were used to analyze censored and

truncated data. This type of analysis is particularly useful in studying diseases with long-term outcomes, such as cancer.

Beyond the conventional applications of logistic regression and survival analysis in healthcare research, other statistical approaches also play pivotal roles. To elaborate further, Karapanagiotis et al. (2021) delved into the significance of risk prediction models that center on binary outcomes, emphasizing the imperative to factor in unequal misclassification costs in healthcare contexts. In response to this identified need, they innovatively proposed a risk modeling framework designed to account for these variances in misclassification costs. Complementing this, logistic regression remains a cornerstone in the field, especially for exploring relationships between independent predictor variables and binary dependent outcome variables. Case in point, Moeti and associates (2023) employed logistic regression to scrutinize the factors that influence access to public healthcare facilities in the City of Tshwane, South Africa. Their findings revealed a noteworthy correlation between income and insurance, thereby reinforcing the argument that socioeconomic determinants are integral to understanding healthcare access.

Overall, logistic regression and survival analysis are valuable statistical methods in healthcare research for predicting binary outcomes and studying time-to-event data, respectively. These methods allow researchers to assess associations between various factors and health outcomes, providing valuable insights for improving healthcare services and interventions.

2.2.3 Social Sciences

Regression analysis is a widely used statistical method in social sciences, including fields such as psychology, sociology, and political science. It allows researchers to explore and quantify relationships between variables, whether they are behavioral traits, societal factors, or political inclinations (Sovey & Green, 2010).

In the specialized domain of political science, instrumental variables regression has notably transitioned from a relatively obscure technique to a cornerstone of methodological inquiry (Sovey & Green, 2010). Despite this rise in prominence, the quality of its application varies considerably. Specifically, numerous studies employing this method fall short in providing the necessary logical or empirical substantiation to validate their statistical assertions. To address this gap, Sovey and Green (2010) have proactively formulated reporting standards, accompanied by a checklist that serves as a guide for readers in the critical evaluation of the method's applications.

Transitioning to sociology, regression analysis is undeniably pervasive, yet it often appears in contexts that are tangential to the discipline's central concerns, thereby diminishing its pedagogical impact. To rectify this, Maio (2013) advocates for framing regression analysis within discussions of pivotal sociological issues, such as income inequality and mortality rates. Beyond the confines of conventional regression analysis, there is a burgeoning recognition among social scientists of the potential embedded in social network analysis. This innovative approach accounts for social structure as an integral variable, particularly when elucidating the mechanisms underlying group behavior. As emphasized by Wölfer et al. (2015), social network analysis provides invaluable tools for investigating intra- and intergroup dynamics, and multilevel analysis is frequently employed to manage the nested structures inherent to social networks.

Complementing these traditional and emerging methods, meta-analysis offers yet another analytical pathway in sociological research. It consolidates the findings from multiple studies to achieve a more comprehensive understanding of specific research questions. Significantly, meta-analysis can be leveraged to scrutinize regression coefficients reported across various social science studies, thereby offering insights into heterogeneity among research outcomes (Tong & Guo, 2019).

In summary, regression analysis is a widely used and versatile method in social sciences. It allows researchers to explore and quantify relationships between variables, and it can be applied in various contexts,

such as political science, sociology, and the study of social networks. However, it is important to ensure the quality of its implementation and provide the necessary evidence for readers to evaluate the statistical claims.

2.2.4 Environmental Science

Regression models are widely used in environmental science research to study the impact of various factors on ecological systems. These models allow researchers to understand how changes in temperature and precipitation patterns affect crop yields or animal migration (Marriott et al., 1985; Poudel & Shaw, 2016).

Linear regression models serve as a fundamental tool in environmental science research. For instance, Marriott and colleagues (1985) delineate the utility of these models in forecasting snowfall amounts. Further extending the scope, Zhao's research team (2020) incorporate nonparametric Bayesian methods to analyze count value data. In addition to linear models, spatial regression techniques offer another layer of analytical depth in environmental science. Specifically, these models enable researchers to investigate the intricate relationships between response variables and spatially complex covariates (Zhao et al., 2020). Within this subset, Geographically Weighted Regression (GWR) stands out as a specialized form, finding applications in diverse fields including, but not limited to, environmental science (Ma et al., 2020).

In the context of climate change and crop yields, Poudel & Shaw (2016) employed regression models to examine the interplay between climatic variables and crop yields in Lamjung District, Nepal. Their study revealed a set of nuanced relationships: a negative correlation between maize yield and summer precipitation, and another negative correlation between wheat yield and winter minimum temperatures. Conversely, they identified a positive relationship between millet yield and summer maximum temperatures. Transitioning from natural to human environments, regression analysis has also been instrumental in exploring the behavioral aspects of environmental science. Case in point, Yayla et al. (2020) utilized regression techniques to assess the influence of environmental commitment on responsible environmental behavior among hotel employees. Their findings indicate that higher levels of environmental commitment exert a positive, albeit moderate, impact on such behavior.

As a whole, regression models are widely used in environmental science research to study the impact of various factors on ecological systems. Linear regression models are commonly used for prediction and analysis of count value data, while spatial regression models are used to explore relationships with complex spatial patterns. Regression analysis has been applied to assess the impact of climate variables on crop yields and to study the relationship between environmental commitment and environmental responsibility behavior.

2.2.5 Engineering

Regression analysis is a widely used technique in engineering disciplines for various purposes such as optimizing processes, predicting system failures, and improving performance metrics (Dieuleveut & Bach, 2016). In datasets with a large number of variables, multicollinearity can be a challenge. To manage multicollinearity, techniques like ridge and lasso regression are commonly applied (Kibria & Banik, 2016).

Ridge regression was first proposed as a method to handle the multicollinearity problem in engineering data (Kibria & Banik, 2016). It was found that by introducing a nonzero value for the ridge parameter, the mean squared error (MSE) for the ridge regression estimator can be smaller than the variance of the ordinary least squares (OLS) estimator. Thus, this makes ridge regression a useful tool for improving the stability and interpretability of regression models in practical engineering applications (Li et al., 2010). Lasso regression, on the other hand, is a sparse modeling method that can be used to select important variables and eliminate irrelevant ones (Lv & Xu, 2020). It penalizes the 'l1' norm of the linear coefficients, which helps reduce the number and strength of correlated or unhelpful predictors (Jain et al., 2016). To date, lasso regression has been successfully applied in many fields of engineering and science (Lv & Xu, 2020).

Beyond the well-established ridge and lasso regression techniques, elastic net regression emerges as another frequently employed method in both engineering and science (Mai, 2019). Uniquely, this approach amalgamates the l1 and l2 norms, thereby striking a nuanced balance between variable selection and coefficient shrinkage (Jain et al., 2016). Also in the study of Jain's research team, elastic net regression has found applications in a variety of complex problems within materials informatics, including but not limited to band gap prediction and the assessment of mechanical properties in alloys.

Overall, regression analysis, including techniques like ridge, lasso, and elastic net regression, plays a crucial role in engineering disciplines for optimizing processes, predicting system failures, and improving performance metrics. These techniques help manage multicollinearity, select important variables, and improve the stability and interpretability of regression models in practical engineering applications.

2.3 Conclusion

In conclusion, the remarkable versatility of regression analysis transcends disciplinary boundaries, manifesting its significance across a diverse spectrum of academic domains. As evidenced through its applications in economics, finance, healthcare, medicine, social sciences, environmental science, and engineering, regression analysis emerges as a unifying thread that weaves together intricate relationships among variables. This methodological framework empowers researchers to unravel complex phenomena, make informed decisions, and contribute to the collective understanding of multifaceted questions. Its adaptability and robustness underline its indispensable role in facilitating interdisciplinary investigations, enabling a holistic comprehension of multifarious issues that extend beyond the confines of individual academic spheres. In the ever-expanding realm of knowledge, regression analysis stands as a resolute pillar, providing researchers with a reliable means to navigate the intricacies of interdisciplinary research and contribute to the advancement of human understanding.

3. Methodology

3.1 Systematic Literature Review

The core of this study is built upon a systematic literature review; a methodological approach designed to collect and analyze relevant scholarly articles in a rigorous and replicable manner. This approach was chosen to ensure the comprehensiveness and quality of the review.

3.2 Data Sources

The literature search targeted peer-reviewed articles across a wide spectrum of academic disciplines, including but not limited to psychology, economics, healthcare, and environmental sciences. The search mainly surveyed publications from the last 20 years to ensure that the findings are both contemporary and relevant.

3.3 Search Criteria and Databases

The primary criteria for article selection were the application and discussion of regression analysis within the framework of advanced research methodologies. This emphasis was intended to narrow down the field of potential articles to those most relevant for this review. A variety of academic search engines and databases were utilized for this review, including PubMed, Scopus, Google Scholar, and the Web of Science. These platforms were chosen for their wide coverage of academic disciplines and their focus on peer-reviewed articles.

3.4 Article Selection Process and Analytical Framework

Articles were initially screened based on their titles and abstracts to assess relevance. Following this, full-text articles were obtained and further scrutinized to determine their suitability for inclusion in this review. Factors

considered during this stage included the quality of the research design, the rigor of the analytical methods, and the relevance of the findings to the objectives of this study. After finalizing the selection, the articles underwent a detailed analysis focusing on how regression analysis was employed, the types of regression models used, and the insights gained through this analytical method.

4. Challenges and Limitations

Regression analysis is a cornerstone in various scientific disciplines, offering researchers a robust statistical tool to model relationships between variables. While its versatility and applicability have made it indispensable, it is not without its set of challenges and limitations. Understanding these challenges is imperative for both the accurate interpretation of results and the integrity of scientific research.

4.1 Overfitting: The Double-Edged Sword of Complexity

Overfitting emerges as one of the most pervasive issues in regression analysis. It occurs when a model, in an attempt to minimize error during the training phase, learns the noise in the data rather than capturing the underlying relationship between variables. While such a model may exhibit impressive performance on the training data, it often fails to generalize well to new, unseen data. Overfitting is especially prevalent in complex models with many predictors or higher-degree polynomials. Researchers must be vigilant about this issue and should consider techniques like regularization or cross-validation to mitigate its impact.

4.2 Multicollinearity: A Threat to Interpretability

Multicollinearity occurs when predictor variables are highly correlated, leading to unstable and unreliable estimates of the regression coefficients. This poses a significant threat to the interpretability of the model, making it difficult to discern the individual impact of each predictor. Solutions like variable elimination or ridge regression can be employed, but these come with their own sets of trade-offs, such as reduced model interpretability or biased estimates.

4.3 Interpretability vs. Predictive Power

The use of advanced regression techniques like ensemble methods or neural networks often enhances predictive accuracy. However, this comes at the cost of interpretability. Such complex models, though effective in prediction, make it challenging to understand the specific contributions of individual predictors. This issue is especially problematic in fields like healthcare or public policy, where understanding the 'why' behind a prediction is as crucial as the prediction itself.

4.4 Assumption Violations: Undermining Validity

Regression models come bundled with a set of underlying assumptions like linearity, normality, and homoscedasticity. The violation of these assumptions not only affects the validity of the model but also risks drawing incorrect or misleading conclusions. Researchers must rigorously check and, if necessary, transform their data to meet these assumptions to maintain the integrity of their findings.

4.5 Data Quality: The Foundation of Reliable Models

The quality of data used for regression analysis is pivotal. Outliers or missing data can significantly skew results, leading to erroneous conclusions. It is essential to employ robust data cleaning methods and outlier detection techniques before building regression models to mitigate these risks.

4.6 Ethical Concerns: The Responsibility of Accurate Modeling

In sensitive fields like healthcare and social sciences, incorrect conclusions drawn from flawed regression models can have severe implications, such as inappropriate treatments or biased policy decisions. Researchers

bear the ethical responsibility to ensure their models are as accurate and reliable as possible to prevent such outcomes.

4.7 Domain Knowledge: The Understated Necessity

Effective application of regression models often requires domain-specific expertise to select the most relevant predictors and to interpret the results in the correct context. Lack of domain knowledge can lead to spurious results and misleading interpretations, thus degrading the quality of the research.

4.8 Causality vs. Correlation: The Eternal Conundrum

Regression analysis is fundamentally correlational, limiting its ability to infer causal relationships between variables. Hence, researchers must exercise caution while interpreting results, explicitly stating the limitations regarding causal inference to avoid misleading conclusions.

4.9 Computational Complexity: The Resource Dilemma

Advanced regression techniques, especially those designed to handle large datasets or high-dimensional spaces, can be computationally intensive. This requires significant computing resources and can be a limiting factor in some research settings.

In summation, regression analysis, while powerful and versatile, comes with its own set of challenges that researchers must be aware of. From overfitting and multicollinearity to ethical concerns and computational complexities, each issue presents a unique obstacle requiring careful consideration and, specialized techniques to overcome. As the role of data-driven research continues to expand, understanding those challenges and limitations becomes increasingly essential for conducting rigorous, ethical, and effective research.

6. Conclusion

This comprehensive study set out with the ambitious goal of exploring the wide-ranging applications and implications of regression analysis across multiple academic disciplines. From economics and finance to healthcare and beyond, our research has illuminated the role of regression analysis as a unifying tool that enables the modeling of intricate relationships among variables.

One of the standout revelations of this study is the remarkable adaptability of regression analysis. Whether it's decoding market trends in economics, predicting patient outcomes in healthcare, or understanding societal behaviors in social sciences, regression models have proven to be invaluable. Their utility extends to environmental science, where they help gauge ecological impacts, and to engineering, where they optimize processes and predict system failures.

However, it's not all smooth sailing. The study also highlights some of the inherent challenges in using regression models, such as the risks of overfitting and multicollinearity. To navigate these issues, researchers often resort to specialized techniques like regularization and variable selection, aiming to enhance both the reliability and interpretability of their models. Additionally, the ethical considerations of applying regression models cannot be overstated, especially in sensitive fields like healthcare and social sciences. Our study underscores the imperative for rigorous data validation and ethical scrutiny to ensure responsible interpretation and application of the findings.

Looking to the future, as our world becomes increasingly data-driven, the importance of regression analysis is set to grow exponentially. There's promising potential for integrating advanced computational methods and machine learning algorithms to further bolster the predictive power of these models, all while maintaining ethical integrity.

In closing, regression analysis emerges as an indispensable cornerstone in interdisciplinary academic research. Its versatility and adaptability are its strengths, but a nuanced understanding of its limitations is essential for its effective and ethical application. As we forge ahead in this data-centric era, this study serves as a timely guidepost, contributing to our collective understanding of this fundamental analytical tool.

Funding Information

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Conflict

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Achrén, A., Raj, R., Siironen, J., Laakso, A., & Marjamaa, J. (2022). Spontaneous angiogram-negative subarachnoid hemorrhage: A retrospective Single Center cohort study. *Acta Neurochirurgica*, *164*(1), 129–140. doi:10.1007/s00701-021-05069-7
2. Amo, L. C. (2016). Education-based gaps in ehealth: A weighted logistic regression approach. *Journal of Medical Internet Research*, *18*(10), e267. doi:10.2196/jmir.5188
3. Dieuleveut, A., & Bach, F. (2016). Nonparametric stochastic approximation with large step-sizes. *Annals of Statistics*, 44(4). doi:10.1214/15-AOS1391
4. Eaves, L. J. (2006). Genotype × environment interaction in psychopathology: Fact or artifact? *Twin Research and Human Genetics*, *9*(1), 1–8. doi:10.1375/183242706776403073
5. Elshorbagy, A., Corzo, G., Srinivasulu, S., & Solomatine, D. P. (2009). Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology – Part 1: Concepts and methodology. *Hydrology and Earth System Sciences*, *14*(10), 1931–1941. doi:10.5194/hess-14-1931-2010
6. Enwere, K., Nduka, E., & Ogoke, U. (2023). Comparative analysis of ridge, bridge and lasso regression models in the presence of multicollinearity. *IPS Intelligentsia Multidisciplinary Journal*, *3*(1), 1–8. doi:10.54117/iimj.v3i1.5
7. Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. doi:10.18637/jss.v033.i01
8. Heinemann, F., Moessinger, M.-D., & Yeter, M. (2016). Do fiscal rules constrain fiscal policy? A meta-regression-analysis. *SSRN Electronic Journal*. doi:10.2139/ssrn.2762314
9. Howel, D., & Kleinbaum, D. G. (1995). Logistic regression: A self learning text. *Statistician*, *44*(3). doi:10.2307/2348716
10. Huang, X., Jennings, S. F., Bruce, B. D., Buchan, A. M., Cai, L., Chen, P., . . . Moore, J. H. (2015). Big data – a 21st century science Maginot Line? No-boundary thinking: Shifting from the big data paradigm. *BioData Mining*, *8*, 7. doi:10.1186/s13040-015-0037-5
11. Jain, A., Hautier, G., Ong, S. P., & Persson, K. A. (2016). New opportunities for materials informatics: Resources and data mining techniques for uncovering hidden relationships. *Journal of Materials Research*, *31*(8), 977–994. doi:10.1557/jmr.2016.80
12. Jong, P. d, and Heller, G. Z. (2008). *Generalized linear models for insurance Data*. Cambridge: Cambridge University Press. doi:10.1017/cbo9780511755408
13. Kang, F., Liu, J., Li, J., & Li, S. (2017). Concrete dam deformation prediction model for health monitoring based on extreme learning machine. *Structural Control and Health Monitoring*, *24*(10). doi:10.1002/stc.1997
14. Karapanagiotis, S., Benedetto, U., Mukherjee, S., Kirk, P. D. W., & Newcombe, P. J. (2022). Tailored Bayes: A risk modeling framework under unequal misclassification costs. *Biostatistics*, *24*(1), 85–107. doi:10.1093/biostatistics/kxab023
15. Kibria, B. M. G., & Banik, S. (2016). Some ridge regression estimators and their performances. *Journal of Modern Applied Statistical Methods*, *15*(1), 206–238. doi:10.22237/jmasm/1462075860
16. Lee, Y. H., & Sung, T. (2008). Fiscal policy, business cycles and economic stabilisation: Evidence from industrialised and developing countries. *Fiscal Studies*, *28*(4), 437–462. doi:10.1111/j.1475-5890.2007.00063.x
17. Li, Y.-F., Xie, M., & Goh, T. N. (2010). Adaptive ridge regression system for software cost estimating on multi-collinear datasets. *Journal of Systems and Software*, *83*(11), 2332–2343. doi:10.1016/j.jss.2010.07.032

18. Liu, L., & Zhan, X. (2019). Analysis of financing efficiency of Chinese agricultural listed companies based on machine learning. *Complexity*, *2019*, 1–11. doi:10.1155/2019/9190273
19. John Lu, Z. Q. J. (2010). The elements of statistical learning: Data mining, inference, and prediction. *Journal of the Royal Statistical Society Series A*, *173*(3), 693–694. doi:10.1111/j.1467-985X.2010.00646_6.x
20. Lv, J., & Xu, X. (2020). *Prediction of daily maximum ozone levels using lasso sparse modeling method*. doi:10.48550/arxiv.2010.08909
21. Ma, Z., Xue, Y., & Hu, G. (2021). Geographically weighted regression analysis for spatial economics Data: A bayesian recourse. *International Regional Science Review*, *44*(5), 582–604. doi:10.1177/0160017620959823
22. Mai, V. V. (2019). *Curvature-exploiting acceleration of elastic net computations*. doi:10.48550/arxiv.1901.08523
23. De Maio, F. (2014). Regression analysis and the sociological imagination. *Teaching Statistics*, *36*(2), 52–57. doi:10.1111/test.12019
24. Marill, K. A. (2004). Advanced statistics: Linear regression, Part II: Multiple linear regression. *Academic Emergency Medicine*, *11*(1), 94–102. doi:10.1197/j.aem.2003.09.006
25. Marriott, F. H. C., Neter, J., Wasserman, W., & Kutner, M. H. (1985). Applied linear regression models. *Biometrics*, *41*(2). doi:10.2307/2530893
26. Maslyaev, M., Hvatov, A., & Kalyuzhnaya, A. V. (2019). *Data-driven partial derivative equations discovery with evolutionary approach*. doi:10.1007/978-3-030-22750-0_61
27. Mathias, D. E., Goveas, D. R., & Rajak, M., & Innvocept Solutions Mh India. (2018). Clinical research – A big data science approach. *International Journal of Trend in Scientific Research and Development*, *Volume–2*, 1075–1078. doi:10.31142/ijtsrd9547
28. McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, *22*(1), 114–140. doi:10.1037/met0000078
29. Moeti, T., Mokhele, T., Weir-Smith, G., Dlamini, S., & Tesfamicheal, S. (2023). Factors affecting access to public healthcare facilities in the City of Tshwane, South Africa. *International Journal of Environmental Research and Public Health*, *20*(4). doi:10.3390/ijerph20043651
30. Nguyen, D. H., Kim, J. H., & Bae, D.-H. (2021). Improving radar-based rainfall forecasts by long short-term memory network in urban basins. *Water*, *13*(6). doi:10.3390/w13060776
31. Omran, B. A., Chen, Q., & Jin, R. (2016). Comparison of data mining techniques for predicting compressive strength of environmentally friendly concrete. *Journal of Computing in Civil Engineering*, *30*(6). doi:10.1061/(ASCE)CP.1943-5487.0000596
32. Pesantez-Narvaez, J., Guillén, M., & Alcañiz, M. (2019). *Predicting Motor Insurance Claims Using Telematics Data—XGBoost vs. logistic Regression*. doi:10.20944/preprints201905.0122.v1
33. Porter, M. A., Aike, L. S., & West, S. G. (1994). Multiple regression: Testing and interpreting interactions. *Statistician*, *43*(3). doi:10.2307/2348581
34. Poudel, S., & Shaw, R. (2016). The relationships between climate variability and crop yield in a mountainous environment: *A Case* study in Lamjung District, Nepal. *Climate*, *4*(1). doi:10.3390/cli4010013
35. Rao, M. B., Klein, J. P., & Moeschberger, M. L. (1998). Survival analysis techniques for censored and truncated Data. *Technometrics*, *40*(2). doi:10.2307/1270658
36. Risher, K., Adams, D., Sithole, B., Ketende, S., Kennedy, C. E., Mnisi, Z., . . . Baral, S. D. (2013). Sexual stigma and discrimination as barriers to seeking appropriate healthcare among men who have sex with men in Swaziland. *Journal of the International AIDS Society*, *16*(3-Suppl. 2), 18715. doi:10.7448/IAS.16.3.18715
37. Samsuddin, M. A. (2021). Monetary vs fiscal policy, which is more effective? Case studies of Asean-5 countries. *Economica*, *9*(1), 172–181. doi:10.22202/economica.2020.v9.i2.4562
38. Shrestha, N. (2020). Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics*, *8*(2), 39–42. doi:10.12691/ajams-8-2-1
39. Sovey, A. J., & Green, D. P. (2011). Instrumental variables estimation in political science: A readers' guide. *American Journal of Political Science*, *55*(1), 188–200. doi:10.1111/j.1540-5907.2010.00477.x
40. Stelson, A. W. (1990). Urban aerosol refractive index prediction by partial molar refraction approach. *Environmental Science and Technology*, *24*(11), 1676–1679. doi:10.1021/es00081a008
41. Tong, G., & Guo, G. (2022). Meta-analysis in sociological research: Power and heterogeneity. *Sociological Methods and Research*, *51*(2), 566–604. doi:10.1177/0049124119882479
42. Wölfer, R., Faber, N. S., & Hewstone, M. (2015). Social network analysis in the science of groups: Cross-sectional and longitudinal applications for studying intra- and intergroup behavior. *Group Dynamics: Theory, Research, and Practice*, *19*(1), 45–61. doi:10.1037/gdn0000021
43. Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 52–69. doi:10.20982/tqmp.08.1.p052

44. Wu, X., Nethery, R. C., Sabath, M. B., Braun, D., & Dominici, F. (2020). Air pollution and COVID-19 mortality in the United States: Strengths and limitations of an ecological regression analysis. *Science Advances*, *6*(45). doi:10.1126/sciadv.abd4049
45. Yayla, Ö., Kendİr, H., & Arslan, E. (2020). Moderator role of gender in the effect of environmental commitment on environmental responsibility behaviour in hotel employees. *Business and Management Studies: An International Journal*, *8*(5), 3971–3990. doi:10.15295/bmij.v8i5.1626
46. Younas, M. (2019). Research challenges of big Data. *Service Oriented Computing and Applications*, *13*(2), 105–107. doi:10.1007/s11761-019-00265-x
47. Zhao, P., Yang, H.-C., Dey, D. K., & Hu, G. (2020). *Bayesian spatial homogeneity pursuit regression for count value Data*. doi:10.48550/arxiv.2002.06678