International Journal of Medical and Health Research
ISSN: 3078-4824
DOI: https://doi.org/10.61424/ijmhr.v4i2.759
Journal Homepage: https://bluemarkpublishers.com/index.php/IJMHR

IJMHR
BLUEMARK PUBLISHERS

| RESEARCH ARTICLE

The Importance and Application of Regression Analysis in Advanced Health Sciences Research

Muhammad Ilyas[1] ✉ and Inam Ullah[2]
[1]PhD Nursing (Scholar), Lincoln University College, Petaling Jaya, Malaysia
[2]MPhil Biotechnology (Scholar), Kohat University of Science & Technology, Kohat, Pakistan
Corresponding Author: Muhammad Ilyas, **E-mail**: ilyas.phdscholar@lincoln.edu.my

| ABSTRACT
Regression Analysis is one of a well-known statistical technique used to examine and determine the relationship between two or more variables and to describe the variables in a mathematical equation that improves a crucial dimension needed to make investment decisions and predictions. Researchers may forecast or explain changes (variation) in one variable based on another using regression. Researchers across various disciplines, including health sciences, frequently use regression analysis as a statistical approach to describe the nature of the relationship between variables, which can be linear or non-linear, positive or negative. The study aimed to evaluate and assess the importance and application of regression analysis in advanced health sciences research. This review article systematically examines and synthesizes the existing literature on the importance and application of regression analysis in advanced health sciences research. This statistical technique is globally used in engineering, business, finance, health sciences, and other areas with the goal of determining the relationship between one dependent variable and a number of other independent variables, and is also widely used in the literature for scientific purposes, while its common methods include linear regression, multiple regression, logistic regression, and cox regression. Regression analysis is essential to health sciences research because it allows researchers to look at correlations between variables, pinpoint risk factors, and forecast health outcomes using data. In complicated biological and social systems, it makes it possible to control confounding variables, which enhances the validity and accuracy of results. It improves the capacity to convert data into useful insights, facilitating the formulation of healthcare policies, treatment plans, and disease preventive initiatives.

| KEYWORDS

Health sciences, research, linear regression, multiple regression, logistic regression, cox regression

| ARTICLE INFORMATION

ACCEPTED: 05 February 2026 **PUBLISHED:** 08 April 2026 **DOI:** https://doi.org/10.61424/ijmhr.v4i2.759

1. Introduction
Regression analysis is a method used in statistical modeling to determine the relationship between several variables, which examines the effects of changing one independent variable while holding the other independent variables constant on a dependent variable (Gupta et al., 2017). Regression analysis with a single independent variable is referred to as univariate regression analysis, and the regression analysis with two or more independent variables is referred to as multivariate regression analysis (Uyanık & Güler, 2013).

Regression analysis has three main advantages: it can generate predictions, show whether independent factors have a significant relationship with a dependent variable, and show the relative intensity of the effects of several independent variables on a dependent variable (Taylor, 2011). A researcher may obtain an equation for a graph using regression analysis to forecast research data. Since this is the only way to assess the regression model in scientific papers and hence appropriately interpret their conclusions (Goswami, 2018).

Regression analysis aims to find a relationship between a dependent variable and a set of independent variables. There are several regression techniques used for research purposes, including linear regression, multiple regression, logistic regression, cox regression, and power regression and each technique has its own advantages and disadvantages (Iqbal, 2021).

Regression analysis is employed in preventive healthcare research to identify risk factors, predict health outcomes, and inform targeted interventions. This approach examines the significance of various regression techniques, such as linear, logistic, Cox proportional hazards, quantile, linear mixed-effects, multilevel, and Poisson regression (Abdul Raheem, 2025).

Furthermore, regression models are extensively utilized within the health sciences to evaluate the effectiveness of therapies and to identify factors influencing patient outcomes. These models are also applied in the social sciences to elucidate the relationships between diverse social attributes, including education and income, and associated attitudes and behaviors (Akomodi, 2025). Both mathematicians and data scientists employ regression analysis for forecasting and prediction where it entails selecting the appropriate model to fit the provided data set, then utilizing that model to generate more predictions. The ideal model precisely depicts each relationship (Gupta et al., 2017).

Model utility test is a hypothesis testing procedure in regression to verify if there is a useful relationship between the dependent variable and the independent variable, and the effectiveness of the model utility test in testing the significance of regression model is evaluated using simple linear regression model with the significance level α = 0.01, 0.025 and 0.05 (Foong et al., 2018). Many assumptions regarding the model are made during regression analysis, including multicollinearity, nonconstant variance (non-homogeneity), linearity, and autocorrelation (Kang et al., 2017).

Likewise, health sciences and social sciences, regression models are globally used for their key benefits in all the fields of environmental sciences, economics and business, engineering and technology, education, clinical psychology, developmental psychology, and cognitive psychology (Akomodi, 2025).

2. Purposes of Regression Analysis

The four main purposes of regression analysis are description, estimation, prediction, and control (Ali & Younas, 2021). Regression analysis is often used in a way that requires a clear description of its goals, usually in a single sentence or a short paragraph. The following sections will outline different ways it can be used, and then connect those uses to specific statistical methods (Werner, A., 2004).

Description: Finding "laws" and mechanisms to describe the facts that researchers see is one of the fundamental objectives of the sciences. Researchers want to identify "explanatory" variables that influence the result of a certain target variable as well as the functional shape of this connection (Werner, A., 2004). It can describe how dependent and independent variables relate to one another (Ali & Younas, 2021).

Estimation: It is the process of determining the value of the dependent variable using the observed values of the independent variables (Ali & Younas, 2021). Regression analysis is a crucial subject in statistics because it may forecast future data behavior based on known data in addition to estimating the relationship between the x-values and y-values of the data points on a graph (Luo, 2016).

Prediction: Based on the interactions between dependent and independent variables, regression analysis can be beneficial for forecasting outcomes and changes in dependent variables (Ali & Younas, 2021). Learning from data and forecasting the results of a random process based on a small number of observations constitute prediction; the name "predictor" might be deceptive if it is understood to mean the capacity to forecast even outside the bounds of the data (Vogt & Johnson, 2015). With or without pretending that the model represents an underlying mechanism, the model can be used to forecast the value of the target variable from known values of the explanatory variables (Werner, A., 2004).

Control: Finally, when exploring the link between one independent variable and the dependent variable, regression allows for the control of the effect of one or more independent variables (Ali & Younas, 2021). Regression analyses use control variables to determine how a treatment affects an outcome, however, because even valid controls may be endogenous and represent a combination of multiple causal mechanisms acting jointly on the outcome, which is difficult to interpret theoretically, the estimated effect sizes of controls themselves are unlikely to have a causal interpretation (Hünermund & Louw, 2025).

3. Common Types of Regression Analysis

3.1 Simple Linear Regression

The concept and the term “regression” was first introduced by Sir Francis Galton in the 1880s in his work on hereditary stature and pea experiments, where he visualized and interpreted an approximately straight regression line, today we known as simple linear regression (Stanton, J., 2001, & Qu, K., 2024). Simple linear regression is an advanced and basic statistical technique used to examine and model the relationship between one independent i.e. predictor variable and one dependent i.e. outcome variable, that fits a straight line to observed data to describe how changes in the independent variable are associated with changes in the dependent variable (Schober & Vetter, 2021).

The technique is referred to as a simple linear regression analysis when there is only one continuous dependent variable and one independent variable, as these two variables are assumed to have a linear relationship in the study (Vogt & Johnson, 2015).

Simple linear regression is widely used in health sciences research to describe and predict how one continuous variable, for example, age relates to another such as blood pressure, it is intuitive and powerful, but is often misapplied or poorly reported, which can mislead clinical decisions, and models a linear relationship between one predictor X and one outcome Y, estimating how much Y changes on average for each 1-unit increase in X (Schober, P., & Vetter, T., 2021, Roustaei, N., 2024, Jones, L., et al., 2024).

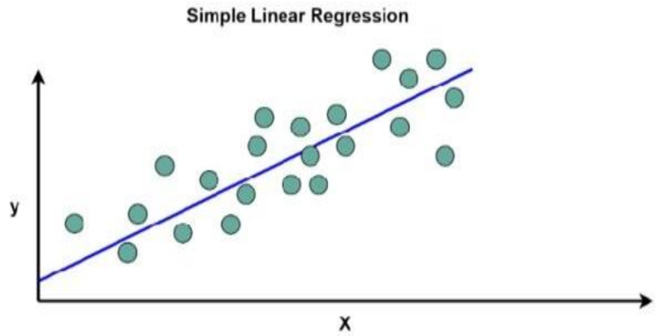
A single dependent variable (Y) and a single independent variable (X) that are linearly connected to one another constitute simple linear regression, the most basic type of regression which goals are to demonstrate the relationship between X and Y in order to forecast Y for a given value of X and to look for a broad underlying pattern linking two variables (Roustaei, 2024). According to Starbuck (2023), the mathematical formula for simple linear regression is usually written as;

$$\mathbf{Y = \beta 0 + \beta 1\, X + \varepsilon}$$

Where;

- Y = dependent (outcome) variable
- X = independent (predictor) variable
- $\beta 0$ = intercept (value of Y when X=0)
- $\beta 1$ = slope (change in Y for one-unit change in X)
- ε = random error (Starbuck, 2023).

According to Md Rasel Uddin (2023), the diagram for simple regression for variables on “Y” and “X” axis is as follows;



[Simple Linear Regression Structure, (Md Rasel Uddin, 2023)]

To find out a better set of variables, independent variables are added to the regression model using a variety of techniques for which some of the most popular techniques are enter, forward, reverse, and sequential selection. The "enter method" involves simultaneously entering each independent variable into the regression model. Finding the impact of an independent variable (X) on a dependent variable (Y) is the aim of this approach (Roustaei, 2024).

3.1.1 Advantages of Linear Regression Analysis

A highly manageable optimization algorithm that can produce a solid answer is the linear regression technique. As opposed to sophisticated algorithms, models produced by linear regression approaches can be implemented quickly and effectively on systems with little computational power (Iqbal, 2021).

Linear regression is widely used because it provides a straightforward method for modeling relationships between variables and making predictions. Their work emphasizes that regression models are easy to construct and interpret in applied research (Douglas C. Montgomery et al. 2012).

The anticipated value (Y') and the observed (Y) data values differ since regression is a model and only approximates values; this discrepancy is known as the residuals, or prediction errors. The regression line with the lowest sum of squared errors of prediction is the one that fits the data the very best (Bazdaric et al., 2021).

Although linear regression is easier to use, analyze, and workout, it can be over-becoming. However, this can be avoided by using cross-validation, regularization techniques, and a few dimensionality discount algorithms (Anandhi & Nathiya, 2023).

In diagnostic and therapeutic investigations where the result depends on multiple factors, as well as in medical research, linear regression is used to model observational data (Marill, 2004).

3.1.2 Dis-advantages of Linear Regression Analysis

Linear regression method assumes that the relevant variables are independent and because the majority of naturally occurring phenomena are non-linear, the linear regression technique cannot adequately fit complicated data sets because it assumes that the input and output variables have a linear relationship (Iqbal, 2021).

The concept of linearity between the established variable and the impartial variables is the primary issue with linear regression and the facts are rarely linearly separable in the real world which makes the false assumption that there may be a straight-line connection between the identified and unbiased factors (Anandhi & Nathiya, 2023).

Because there are usually several pertinent predictor variables, simple linear regression has limited applications in medical research. One predictor variable is used in univariate statistical procedures, like basic linear regression, which are frequently clinically deceptive despite being technically sound (Marill, 2004).

Multipurpose software like Microsoft Excel can be used to compute a basic linear regression. Regretfully, while the algorithms are capable of computing a regression, they are unable to perform the further actions required to assess the method's suitability (Bazdaric et al., 2021).

3.2 Multiple Linear Regression

Multiple linear regression was first used by Galton with the core idea of multiple causes in his heredity studies, in 1880s, later on, Karl Pearson extended Galton's ideas, developing multiple correlation and multiple regression, using determinantal matrix algebra and linking regression to the multivariate normal distribution, and Yule in 1899, used the first widely recognized complete multiple regression analysis (Stanton, J., 2001).

Multiple linear regression is an advanced method in statistics, used to describe the simultaneous association of several variables with one continuous outcome and to make inferences and predictions based on these relationships (Eberly, L.E.,2007). This statistical technique permits other factors to join the study independently, in order to evaluate the impact of each item, as it is useful for measuring how different concurrent influences affect a single dependent variable, even when the researcher is only interested in the effects of one of the independent variables, multiple regression is frequently necessary due to the bias caused by omitted factors in simple regression (Sykes & Sykes, 1993).

In health sciences research, multiple linear regression is a helpful method for modeling a variety of events as it provides estimates of the predictor variable coefficients and their Standard Error or uncertainty for data sets that satisfy the required assumptions and it is typically possible to solve this well-developed model exactly (Marill, 2004).

Simple linear regression is expanded further by multiple regression as it is employed when we wish to forecast a dependent variable's value (also known as a target or criterion variable) by using the values of two or more independent variables (also known as predictor or explanatory variables) as multiple regression estimates how changes in several predictors simultaneously influence a single continuous outcome variable (Vogt & Johnson, 2015). Multiple regression represents the relationship between a dependent variable and a weighted linear combination of independent variables, and it can be used to describe relationships, test hypotheses, and predict outcomes. (Aiken, L. S. & West, S. G., 2003).

According to Muda et al., 2020, general equation for multiple linear regression is:

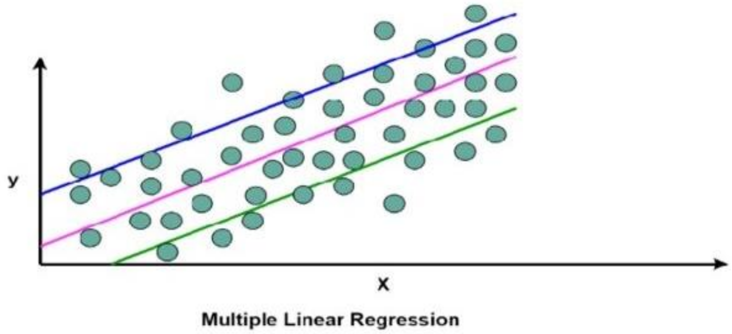
$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta k\, Xk + \varepsilon$$

Where:

- Y = Dependent variable
- $X1, X2, \ldots, Xk$ = Independent variables
- $\beta 0$ = Intercept
- $\beta 1, \beta 2, \ldots, \beta k$ = Regression coefficients
- ε = Error term

The above equation shows how several predictors combine linearly to estimate or predict the outcome variable (Muda et al., 2020).

According to Md Rasel Uddin (2023), the diagram for multiple regression for different variables on "Y" and "X" axis is as follows;



[Multiple Linear Regression Structure, (Md Rasel Uddin, 2023)]

3.2.1 Advantages of Multiple Regression Analysis

Multiple regression analysis is especially valuable in social science and health research, domains characterized by the influence of numerous factors, as it allows for the simultaneous examination of multiple independent variables' effects on a single dependent variable, thereby offering a more accurate representation of intricate occurrences (Fletcher J., 2009).

Furthermore, multiple regression facilitates the estimation of individual predictor effects as it furnishes distinct regression coefficients for each independent variable, enabling researchers to ascertain the unique contribution of each predictor while controlling for the influence of others (Kundu, P. et al., 2019).

Multiple regression generally improve prediction accuracy as it produces more accurate predictions than simple regression because it incorporates more relevant variables into the model (Bloniarz et al., 2016).

Researchers from many fields including health sciences, can use multiple regression to test hypotheses about relationships between variables, including statistical significance and direction of effects (Montgomery et al., 2012).

Multiple regression has the ability to control the confounding variables, which help researchers to isolate the true relationship between predictors and the outcome as this improves the internal validity of research findings (Fletcher J., 2009).

Multiple regression has the also the ability to examine the complex relationship between the variables as it allows researchers to study interactions and combined effects among variables, providing deeper insights into complex systems (Kundu, P., et al., 2019).

3.2.2 Dis-advantages of Multiple Regression Analysis

Linearity, proper model specification, independent errors, homoscedasticity, no or limited multicollinearity, and a sufficient sample size are all necessary for multiple regression; failure to meet these requirements results in inaccurate or biased estimates and conclusions (Duncan, G. M., 1986, Grant, S. et al., 2018, Venter, A., & Maxwell, S., 2000).

Small samples reduce power and generalizability, raising Type-II errors, and nonlinearity, heteroscedasticity, and autocorrelation can significantly skew results if they are not identified and addressed in multiple regression analysis (Venter, A., & Maxwell, S., 2000, Seabrook, J., 2025).

Even at low correlations for example $r \approx 0.3$, multicollinearity (correlated predictors) results in unstable coefficients, large confidence ranges, reduced power, and deceptive sign/magnitude of effects in multiple regression analysis and predictor connections have the potential to "confound" parameters, making meaningful interpretation such as determining which variable is truly important is questionable while using multiple regression analysis (Venter, A., & Maxwell, S., 2000, Graham, M., 2003, Ellsworth, S. et al., 2023, Seabrook, J., 2025).

Inappropriate causal assertions from correlational models result from the frequent use of regression because there are several independent variables, without making a clear distinction between prediction and explanation/causality. When theory of variable relations is poor, high-order partial coefficients can be extremely deceptive (also known as the "partialling fallacy"), because they can incorrectly rank predictors when they are correlated, standardized betas are sometimes abused as importance indices (Yoshida, T., & Murai, J., 2021, Mizumoto, A., 2022, Groenwold, R., & Dekkers, O., 2023).

Stepwise techniques frequently result in contradicting models across samples by inflating Type I error, producing biased parameters, and creating unstable "best models" and the outcomes of many meta-regression studies are often deceptive due to ecological fallacy, overfitting, and improper regressing of effects on baseline risk (Whittingham, M. et al., 2006, Ray-Mukherjee, J. et al., 2014, Geissbühler, M. et al., 2021).

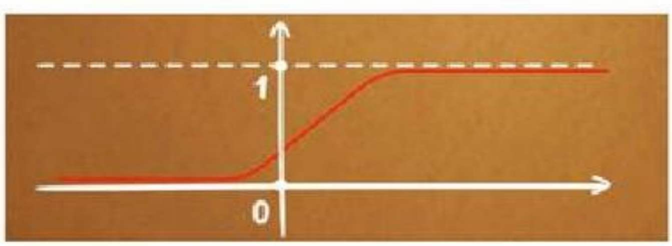
3.3 Logistic Regression

Logistic regression curve was first introduced by Pierre-François Verhulst in 1838 to the model population growth, and has grown in popularity as a statistical tool in health sciences research, particularly in the last 20 years as it is commonly considered the preferred statistic when one or more independent (predicting) variables are used to predict the occurrence of a binary (dichotomous) outcome variable (Boateng & Abaye, 2019).

Logistic regression is a type of regression analysis, used for categorical outcomes, most commonly used when the dependent variable is binary for example, male vs. female, diseased vs. healthy, yes vs. no, dead vs. alive (Sperandei, S., 2014, Boateng & Abaye, 2019, Schober, P., & Vetter, T., 2021). This method has become a standard multivariable tool in social, educational, and health sciences research for modeling the relationship between one or more predictors and the probability that an event occurs (Peng, C., et al., 2002, Bewick, V., et al., 2005, Boateng, E., & Abaye, D., 2019).

Unlike linear regression, which assumes a continuous outcome and a straight-line relationship, logistic regression assumes a binary outcome and models the logit i.e. natural log of the odds, of the outcome as a linear function of the predictors, if (P) is the probability that an event occurs for example, disease present (Hosmer, D., et al., 2005, Stoltzfus, J. 2011, Schober, P., & Vetter, T., 2021). Logistic regression formula and figure according to Shah & Patel (2022), is given in a diagram below;

$$\frac{p(X)}{1-p(X)} = e^{(\beta_0+\beta_1 X_1+\ldots+\beta_k X_k)}$$



[Logistic Regression equation and diagram, (Shah & Patel, 2022)]

The above formulation guarantees that predicted probabilities lie between 0 and 1 and yields an S-shaped (sigmoid) relationship between predictors and probability (Bewick, V., et al., 2005, Stoltzfus, J. 2011, Schober, P., & Vetter, T., 2021). Exponentiating coefficients gives odds ratios (ORs), which express how the odds of the event change for a one-unit increase in a predictor, holding other variables constant (Bewick, V., et al., 2005, Ranganathan, P. et al., 2017, Boateng, E., & Abaye, D., 2019).

3.3.1 Advantages of Logistic Regression Analysis

Logistic regression is well-suited for binary and categorical outcomes because it is specifically built for binary results like disease presence or absence, or success versus failure. In situations where, linear regression assumptions are not met, logistic regression and its variations, such as multinomial and ordinal logistic regression, can effectively model outcomes with multiple categories (LaValley, M., 2008, Boateng, E., & Abaye, D., 2019, Yay, M., 2023).

Logistic regression simultaneously allows inclusion of many explanatory variables i.e. continuous or categorical, estimating each variable's independent effect on the outcome (Stoltzfus, J., 2011, Sperandei, S., 2014, Yay, M., 2023).

Logistic regression offers a significant benefit by accounting for all variables simultaneously, thereby mitigating confounding effects, in contrast to examining variables individually (Sperandei, S., 2014, Pal, A., 2021, Ezeonu, T. et al., 2025).

Logistic regression accommodates continuous, categorical, or a combination of predictor variables because they are not required to follow a normal distribution or exhibit homoscedasticity. This method models the logit of the outcome, ensuring that predicted probabilities remain within the 0 to 1 range and circumventing issues encountered when applying linear regression to binary data (Stoltzfus, J., 2011, Lever, J. et al., 2016, Yay, M., 2023).

Logistic regression is extensively employed within the fields of epidemiology, clinical medicine, public health, and meta-analysis for purposes of prediction, diagnosis, and evaluation of treatment efficacy, and is incorporated in all principal statistical software packages, utilizing well-established regression diagnostic techniques and model development methodologies (Boateng, E., & Abaye, D., 2019, Josephine, K. et al., 2024, Hua, Y. et al., 2025).

3.3.2 Dis-advantages of Logistic Regression Analysis

A key dis-advantage logistic regression is that it necessitates the independence of observations, a linear relationship in the log-odds for continuous variables, the absence of multicollinearity, and the exclusion of influential outliers; any violations of these assumptions may result in biased estimates and predictions (Stoltzfus, J., 2011, Harris, J., 2021).

Logistic regression needs an adequate number of events per variable i.e. 10 – 20, to avoid overfitting and unstable odds ratios and the separation (predictors perfectly classify outcomes) leads to infinite or unstable coefficients, especially with rare events or sparse data (Park, H., 2013, Mansournia, M. et al., 2018, Boateng, E., & Abaye, D., 2019).

In logistic regression, odds ratios are on a log-odds scale, which is unintuitive; they are frequently misinterpreted as risk or prevalence ratios, exaggerating the effects when outcomes are common (Niu, L., 2018 & Howell-Moroney, M., 2023).

Another dis-advantage of logistic regression lies in its propensity to significantly overstate relative risks or prevalence ratios for outcomes that occur frequently, specifically those exceeding 10%, which can result in erroneous interpretations; thus, methodologies such as log-binomial or robust Poisson regression are frequently considered more advantageous (Pinheiro-Guedes, L. et al., 2024).

A notable limitation of logistic regression is its characteristic of non-collapsibility; specifically, the odds ratios exhibit variability when additional covariates are incorporated, even in the absence of confounding variables, thereby rendering model comparisons across disparate samples or specifications invalid (Mood, C., 2010 & Howell-Moroney, M., 2023).

Literature show the widespread poor practice of logistic regression such as, inadequate sample size, missing diagnostics, no validation, and failure to account for complex survey design, dependence, outliers, or missing data, all of which can bias results (Boateng, E., & Abaye, D., 2019 & Dey, D. et al., 2025).

Logistic regression, in complex surveys, ignoring weights, clustering, and stratification is common and leads to biased estimates and standard errors (Dey, D. et al., 2025).

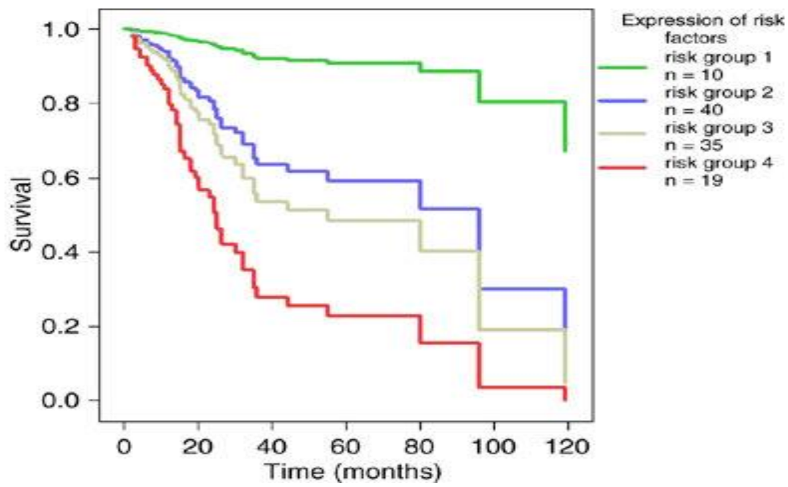
3.4 Cox Regression (Multivariate Cox's Proportional Hazards Regression)

Cox regression or proportional hazards regression, was first discovered and introduced by Sir David Cox in 1972, in his paper "Regression Models and Life Tables" (Kalbfleisch, J., & Schaubel, D. (2022). This is a semiparametric model for time-to-event (survival) data, widely used in clinical and epidemiologic research to estimate how covariates affect the hazard i.e. instantaneous risk of an event over time (Christensen, E., 1987 & ElHafeez, S. et al., 2021).

Cox regression expresses the hazard for subject with covariates (z) as, $\lambda(t/z) = \lambda 0(t)\exp(\beta z)$, where $\exp(\beta)$ are hazard ratios (HRs), and the continuous covariates for example age, biomarkers, and categorical covariates for example sex, disease status, can be modeled simultaneously, allowing adjustment for confounding (Tsiatis, A., 1981, Christensen, E., 1987, Zapf, A. et al., 2024).

In cox or Proportional hazards, the covariate effects are constant over time, i.e. the hazards for two individuals are proportional at all times, as this model is often checked via Schoenfeld residuals, log-minus-log plots, time-by-covariate interactions, or extended models (Dessai, S., & Patil, V., 2019; Kuitunen, I. et al., 2021; Hua, K. et al., 2025).

According to Taubert et al., (2007), multivariate Cox's proportional hazards regression model for overall survival of cancer patients for different age groups is given below;



[Multivariate Cox's proportional hazards regression model for overall survival of cancer patients for different age groups, (Taubert et al., 2007)]

3.4.1 Advantages of Multivariate Cox's Proportional Hazards Regression Analysis

Multivariate Cox proportional hazards regression is widely used in health sciences research because it has several important advantages than other regression methods.

A big advantages of cox regression is the simultaneous adjustment for multiple covariates i.e. control of confounding, as it handles multiple predictors at once, estimating adjusted hazard ratios (HRs) while controlling for confounders such as age, sex, and health conditions (diseases) and it allows more valid causal interpretation than other regression analyses (Cioci et al., 2021; ElHafeez, S. et al., 2021 & Lee, S., 2023).

Cox regression is semi-parametric in nature as it does not assume a specific distribution for survival times, only proportional hazards, making it robust across many settings and the baseline hazard is left unspecified, while covariate effects are modelled via regression coefficients (McGregor, D. et al., 2019; Lee, S., 2023 & Hua, K. et al., 2025).

Another advantage of cox regression is that it appropriately accommodates right-censored data and uses full time-to-event information, unlike binary outcome models that ignore timing and time-varying covariates can be incorporated, improving dynamic risk prediction and triage decisions (ElHafeez, S. et al., 2021; Lee, S., 2023 & Blythe, R. et al., 2024).

Cox regression provides clinically interpretable hazard ratios for continuous and categorical variables, analogous to relative risks but time-to-event based, and it facilitates prognostic modelling and individualized risk stratification in clinical decision-making (Cioci et al., 2021; Lee, S., 2023 & Monikapreethi S.K. et al., 2024).

The Cox proportional hazards model possesses the capability of accommodating enhanced frameworks, such as shared frailty models, marginal models, copula-based approaches, and multivariate joint models, which facilitate the examination of recurrent events, clustered datasets, and various event types, all while preserving the fundamental structure of the Cox model (Li, G. et al., 2020; Knafl, G., 2023 & Ben-Assuli, O. et al., 2023).

Cox regression has partial likelihood estimation which is computationally efficient, enabling large-scale and high-dimensional applications, including omics-based web tools and Genome-wide Association Studies (GWAS) time-to-event analyses (Lánczky, A., & Győrffy, B., 2021; Li, Y. et al., 2025).

3.4.2 Dis-advantages of Multivariate Cox's Proportional Hazards Regression Analysis

Multivariate Cox regression is widely used in health science, but its reliance on proportional hazards, sensitivity to time-dependent covariates, small-sample bias, non-collapsibility, and frequent neglect of assumption checking make it vulnerable to biased or misleading estimates if not applied and tested carefully.

The core proportional hazards (PHs) assumption i.e. constant hazard ratio over time, is often unrealistic in clinical and epidemiologic data where violations can yield false models and misleading time-independent risk factors (Babińska, M. et al., 2015; Kuitunen, I. et al., 2021 & ElHafeez, S. et al., 2021).

Even with thorough adjustments, strong time dependent risk variables like smoking can skew hazard ratios for other correlated covariates, and proportional hazards violations are frequent in high dimensional settings (Moolgavkar, S. et al., 2018 & Zeng, Z. et al., 2022).

With small samples or many variables, multivariate Cox regression yields low power, exaggerated standard errors, a nd skewed interaction estimates; penalization or firth correction is more effective, especially for treatment-biomarker interactions (Jóźwiak, K. et al., 2024).

Penalized proportional hazards models may have high false positive rates and biased selection when proportional hazards are violated, and cox regression cannot naturally accommodate many confounders in some applied settings, where propensity score methods can include more covariates more stably (Martens, E. et al., 2008 & Sheng, A., & Ghosh, S., 2020).

The Cox regression model is generally non-collapsible, meaning that even in the absence of confounding, the marginal and adjusted hazard ratios differ, making the interpretation of causality more difficult (Samuelsen, S., 2022).

Erroneous minor connections can arise in the Cox Regression model for correlated factors due to residual confounding and inaccurate time varying covariate specification (Moolgavkar, S. et al., 2018 & Jiang, N. et al., 2024).

Naïve Cox implementations may violate assumptions or misread cause specific hazards in complicated compositional or competing risk structures unless they are reformulated for example, log-ratio coordinates and augmented data (Lunn, M., & McNeil, D., 1995 & McGregor, D. et al., 2019).

4. Application of Regression Analysis in Health Sciences

Linear, multiple and cox regression is widely used in epidemiology to quantify relationships between exposures i.e. risk factors and health outcomes, especially when the outcome variable is continuous e.g., blood pressure, cholesterol level, body mass index (BMI), (Kenneth J. Rothman, 1998). Researchers in epidemiology, use both linear and multiple regression to control for confounding variables such as age, sex, or socioeconomic status which allows estimation of the independent effect of a risk factor (Bhattacharyya et al., 1979).

The observational research using linear regression analysis were the ones that made the connection between smoking and mortality and illnesses as for instance, we have a linear regression model where the dependent variable is a person's lifespan expressed in years and the explanatory variable is cigarette smoking (Iqbal, 2021).

Linear regression is often used to examine dose–response relationships, which are central in epidemiology for example; cigarettes smoked per day vs lung function, alcohol intake vs liver enzyme levels, pollution exposure vs respiratory symptoms where dose–response relationships strengthen the causal inference (Szklo, M., & Nieto, F. J. 2019).

Linear regression is used to examine trends in disease rates over time for example; trends in malaria incidence over years, changes in mortality rates, increase or decrease in vaccination coverage. Regression can estimate whether rates are increasing or decreasing and by how much annually (Gordis L., 2014). Linear and multiple regression is used to predict future or unknown values of health-related variables based on known predictors, for example; predicting hospital stay duration based on disease severity, estimating birth weight from maternal characteristics, predicting patient recovery time after treatment and, predictive models assist healthcare professionals in planning treatments and allocating resources efficiently (Kleinbaum, D et al., 2013).

Linear regression is applied to evaluate healthcare quality, efficiency, and outcomes for example; relationship between nurse staffing and patient outcomes, hospital costs and length of stay, patient satisfaction and waiting timea, as these analyses help improve healthcare delivery and policy decisions (Kutner et al. 2005). Linear regression is commonly used to determine how one health-related variable changes with another. Researchers use it to quantify relationships between biological and behavioral variables for example, linear regression can estimate how much systolic blood pressure increases with each year of age (Bland, M. 2015).

In epidemiology, linear regression is used to analyze trends and relationships in population health data for example; disease trends over time, nutritional factors and health outcomes, environmental exposures and disease rates. Regression analysis helps to estimate relationships while controlling for the effects of other variables (WHO, 2020). In clinical trials and intervention studies, multiple regression helps assess treatment effectiveness while controlling for baseline characteristics for example; evaluating the effect of a new antihypertensive drug on blood pressure while controlling for age, sex, and baseline blood pressure, measuring improvement in quality of life after therapy, adjusting for socioeconomic status and comorbidities (Vittinghoff, E., et al. 2012).

Multiple regression helps assess relationships between psychological variables and health outcomes for example; predicting depression scores from stress levels, social support, and income and, studying factors influencing treatment adherence in psychiatric patients (Tabachnick, B. G., & Fidell, L. S., 2019). Logistic regression is commonly used to identify factors associated with the presence or absence of disease while controlling for confounding variables, for example; determining risk factors for diabetes i.e. obesity, age, and physical inactivity are all key factors, and it's important to identify predictors of hypertension, such as smoking, diet, and stress, while also evaluating factors related to infectious diseases. This approach also helps in calculating adjusted odds ratios, which show how strongly risk factors are linked to disease outcomes (David W. Hosmer Jr. et al., 2013).

Cox regression is commonly used in clinical trials to evaluate the effect of treatments on patient survival while adjusting for covariates for example; comparing survival time between patients receiving two different cancer treatments, assessing the effect of a new drug on mortality while controlling for age and disease severity and cox regression also estimates the hazard ratios (HRs), which indicates the relative risk of an event occurring at any time point (Cox, 1972). Cox regression is used to study disease progression and prognosis by examining how patient characteristics affect survival time for example; time to cancer recurrence after treatment, progression of HIV infection to AIDS, and time to kidney failure in chronic kidney disease patients (Frank E. & Harrell, Jr., 2015).

5. Factors Affecting Regression Analysis

Regression analysis is susceptible to breaches of its underlying assumptions, multicollinearity, confounding variables, the selection of variables, and the integrity or potential biases present within the data. Consequently, the generation of valid and interpretable outcomes necessitates meticulous verification, judicious model construction, and the appropriate management of collinearity, missing data, and bias.

Employing linear models when the actual relationship deviates from linearity will yield erroneous specifications, thereby producing biased predictions and coefficients (Greenland, S., 1989 & Tamhane, A.C., 2020).

Violations of normality, homoscedasticity, and independence can distort standard errors, statistical tests, and the R^2 statistic, for example, in models of gross domestic product (GDP), correcting for these violations caused the R^2 value to drop from 97% to 26%, which indicated a significant overestimation of the model's fit (Tamhane, A.C., 2020 & Alanazi, B., 2025).

Exceptions and significant points i.e. isolated or high leverage observations can distort coefficients and residual variance therefore the diagnostics are essential (Tamhane, A.C., 2020 & Lapach, S., 2025).

Heterogeneity and huge dispersion i.e. change which estimation method performs best and can degrade are the most model quality indicators (Lapach, S., 2025).

Sample size, missing data, and sample characteristics are all factors that have an impact on both linear and multiple regression studies (Ali & Younas, 2021).

Correlation among predictors i.e. multicollinearity, inflates standard errors, reduces power, and can make important variables appear nonsignificant, while also increasing parameter bias and Type-II error (Daoud, J., 2017 & Kim, J., 2019).

Multiple regression results are strongly shaped by data properties, assumption compliance, and modeling choices, when there are poor coefficients, p-values, and R^2 can be badly misleading (Tamhane, A.C., 2020).

In multivariate models, confounding and shared variance between predictors change individual coefficients after adjustment, so variables significant in unadjusted analyses may lose significance (Johnston, R., 2017; Kim, J., 2019 & Andrade, C., 2024).

6. Conclusion

Regression is an essential statistical tool in health sciences, used to examine relationships between variables, identify risk factors, predict outcomes, evaluate treatments, and support public health decision-making. Its simplicity and interpretability make it one of the most widely used methods in health sciences research. In respect to any researched phenomenon, regression analysis helps researchers to characterize, forecast, estimate, and draw plausible conclusions about the connected variables. When researchers want to look at the relationship between particular factors, regression also enables controlling one or more variables. In regression analysis researchers may find some of the important factors offered to be helpful, therefore they should consider the kind and number of dependent and independent variables, as well as the nature and size of the sample, while planning and carrying out regression analysis. Different regression models are employed in various contexts and their usability varies from situation to situation and is typically based on the type of data and the relationships between the data.

6.1 Limitations of the Study

Regression analysis in health sciences research has a number of limitations that need to be recognized despite its extensive use and significance. The use of underlying assumptions, such as linearity, independence, homoscedasticity, and normalcy, which, if broken, can produce biased or deceptive results, is one significant drawback. Additionally, multicollinearity, missing data, and outliers can all skew parameter estimate and make regression models less interpretable. Furthermore, it is challenging to prove actual cause-and-effect without complementing study designs because regression analysis mostly finds connections rather than causal links. Overfitting also restricts generalizability to different populations or environments, particularly in intricate models with several predictors.

6.2 Areas for Further Study

Future research should concentrate on creating more resilient and adaptable modeling techniques that can manage the high-dimensional data and intricate, non-linear interactions that are frequently present in contemporary health sciences. Additionally, better techniques are required to deal with confounding variables, measurement problems, and missing data. Regression techniques can be combined with more sophisticated approaches like machine learning and longitudinal data analysis to improve prediction accuracy and gain a deeper understanding of disease trends. Additionally, encouraging reproducibility, transparency, and appropriate model validation procedures will increase the validity of regression-based results and broaden their use in clinical and public health decision-making.

Funding: This research received no external funding
Conflicts of Interest: The authors declare no conflict of interest
Muhammad Ilyas: ORCID ID: https://orcid.org/0009-0006-1795-884X

References

[1] AbdulRaheem, Y. (2025). The Role of Regression Analysis in Preventive Research Modalities: A Medical-Focused Comprehensive Review. *Journal of Public Health Issues and Practices*, *9*(1), 1–5. https://doi.org/10.33790/jphip1100238
[2] Akomodi, J. O. (2025). The Efficacy of Statistics in All Major Fields of Research: A Focus on Regression Analysis. *Open Journal of Statistics*, *15*(01), 53–72. https://doi.org/10.4236/ojs.2025.151004
[3] Alanazi, B. (2025). The Effects of Assumption Violations on Coefficient of Determination and Regression Model Accuracy: A Study of Construction GDP Data. *Fractals*. https://doi.org/10.1142/s0218348x25400961.
[4] Ali P, Younas A. Understanding and interpreting regression analysis. Evid Based Nurs. 2021 Oct;24(4):116-118. doi: 10.1136/ebnurs-2021-103425. Epub 2021 Sep 8. PMID: 34497132.
[5] Anandhi, P., & Nathiya, D. E. (2023). Application of linear regression with their advantages, disadvantages, assumption and limitations. *International Journal of Statistics and Applied Mathematics*, *8*(6), 133–137. https://doi.org/10.22271/maths.2023.v8.i6b.1463
[6] Andrade, C. (2024). Regression: Understanding What Covariates and Confounds Do in Adjusted Analyses. *The Journal of clinical psychiatry*, 85 4. https://doi.org/10.4088/jcp.24f15573.
[7] Babińska, M., Chudek, J., Chełmecka, E., Janik, M., Klimek, K., & Owczarek, A. (2015). Limitations of Cox Proportional Hazards Analysis in Mortality Prediction of Patients with Acute Coronary Syndrome. *Studies in Logic, Grammar and Rhetoric*, 43, 33 - 48. https://doi.org/10.1515/slgr-2015-0040.
[8] Bazdaric, K., Sverko, D., Salaric, I., Martinović, A., & Lucijanic, M. (2021). The abc of linear regression analysis: What every author and editor should know. *European Science Editing*, *47*, 1–9. https://doi.org/10.3897/ese.2021.e63780
[9] Ben-Assuli, O., Ramon-Gonen, R., Heart, T., Jacobi, A., & Klempfner, R. (2023). Utilizing shared frailty with the Cox proportional hazards regression: Post discharge survival analysis of CHF patients. *Journal of biomedical informatics*, 104340. https://doi.org/10.1016/j.jbi.2023.104340.
[10] Bewick, V., Cheek, L., & Ball, J. (2005). Statistics review 14: Logistic regression. *Critical Care*, 9, 112 - 118. https://doi.org/10.1186/cc3045.
[11] Bhattacharyya, H. T., Kleinbaum, D. G., & Kupper, L. L. (1979). Applied Regression Analysis and Other Multivariable Methods. *Journal of the American Statistical Association*, *74*(367), 732. https://doi.org/10.2307/2287012
[12] Bloniarz, A., Liu, H., Zhang, C. H., Sekhon, J. S., & Yu, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(27), 7383–7390. https://doi.org/10.1073/pnas.1510506113
[13] Blythe, R., Parsons, R., Barnett, A., Cook, D., McPhail, S., & White, N. (2024). Prioritising deteriorating patients using time-to-event analysis: prediction model development and internal–external validation. *Critical Care*, 28. https://doi.org/10.1186/s13054-024-05021-y.
[14] Boateng, E. Y., & Abaye, D. A. (2019). A Review of the Logistic Regression Model with Emphasis on Medical Research. *Journal of Data Analysis and Information Processing*, *07*(04), 190–207. https://doi.org/10.4236/jdaip.2019.74012
[15] Christensen, E. (1987). Multivariate survival analysis using Cox's regression model. *Hepatology*, 7. https://doi.org/10.1002/hep.1840070628.
[16] Cioci, A. C., Cioci, A. L., Mantero, A. M. A., Parreco, J. P., Yeh, D. D., & Rattan, R. (2021). Advanced Statistics: Multiple Logistic Regression, Cox Proportional Hazards, and Propensity Scores. *Surgical Infections*, *22*(6), 604–610. https://doi.org/10.1089/sur.2020.425
[17] Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *34*(2), 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x
[18] Daoud, J. (2017). Multicollinearity and Regression Analysis. *Journal of Physics: Conference Series*, 949. https://doi.org/10.1088/1742-6596/949/1/012009.
[19] Dessai, S., & Patil, V. (2019). Testing and interpreting assumptions of COX regression analysis. *Cancer Research, Statistics, and Treatment*, 2, 108 - 111. https://doi.org/10.4103/crst.crst_40_19.
[20] Dey, D., Haque, M., Islam, M., Aishi, U., Shammy, S., Mayen, M., Noor, S., & Uddin, M. (2025). The proper application of logistic regression model in complex survey data: a systematic review. *BMC Medical Research Methodology*, 25. https://doi.org/10.1186/s12874-024-02454-5.
[21] Duncan, G. M. (1986), Review of *Multiple Regression in Practice*, by W. D. Berry & S. Feldman. *Journal of Marketing*

Research, *23*(3), 309–310. https://doi.org/10.2307/3151494
[22] Ellsworth, S., Van Rossum, P., Mohan, R., Lin, S., Grassberger, C., & Hobbs, B. (2023). Declarations of independence: How embedded multicollinearity errors affect dosimetric and other complex analyses in radiation oncology. *International journal of radiation oncology, biology, physics*. https://doi.org/10.1 016/j.ijrobp.2023.06.015
[23] ElHafeez, S., D'Arrigo, G., Leonardis, D., Fusaro, M., Tripepi, G., & Roumeliotis, S. (2021). Methods to Analyze Time-to-Event Data: The Cox Regression Analysis. *Oxidative Medicine and Cellular Longevity*, 2021. https://doi.org/10.1155/2021/1302811.
[24] Ezeonu, T., Narayanan, R., Huang, R., & Sherman, M. (2025). Constructing and Interpreting Logistic Regression Analyses in Orthopedic Clinical Research. *Clinical spine surgery*. https://doi.org/10.1097/bsd.0000000000001816.
[25] Fletcher, J. (2009). Multiple linear regression. *BMJ*, 338:b167, https://doi.org/10.1136/bmj.b167
[26] Foong, N. S., Ming, C. Y., Eng, C. P., & Shien, N. K. (2018). *An Insight of Linear Regression Analysis*.
[27] Geissbühler, M., Hincapié, C., Aghlmandi, S., Zwahlen, M., Jüni, P., & Da Costa, B. (2021). Most published meta-regression analyses based on aggregate data suffer from methodological pitfalls: a meta-epidemiological study. *BMC Medical Research Methodology*, 21. https://doi.org/10.1186/s12874-021-01310-0
[28] Goswami, A. (2018). Utilization of regression analysis in clinical research. *International Journal of Unani and Integrative Medicine*, *2*(1), 01–05. https://doi.org/10.33545/2616454x.2018.v2.i1a.15
[29] Graham, M. (2003). CONFRONTING MULTICOLLINEARITY IN ECOLOGICAL MULTIPLE REGRESSION. *Ecology*, 84, 2809-2815. https://doi.org/10.1890/02-3114
[30] Grant, S., Hickey, G., & Head, S. (2018). Statistical primer: multivariable regression considerations and pitfalls†. *European Journal of Cardio-Thoracic Surgery*, 55, 179–185. https://doi.org/10.1093/ejcts/ezy403
[31] Greenland, S. (1989). Modeling and variable selection in epidemiologic analysis. *American journal of public health*, 79 3, 340-9 . https://doi.org/10.2105/ajph.79.3.340.
[32] Groenwold, R., & Dekkers, O. (2023). Is it a risk factor, a predictor, or even both? The multiple faces of multivariable regression analysis. *European journal of endocrinology*, 188 1. https://doi.org/10.1093/ejendo/lvac012
[33] Gupta, A., Sharma, A., & Goel, A. (2017). *Review of Regression Analysis Models*. *6*(08), 58–61.
[34] Harris, J. (2021). Primer on binary logistic regression. *Family Medicine and Community Health*, 9. https://doi.org/10.1136/fmch-2021-001290.
[35] Hua, Y., Stead, T., George, A., & Ganti, L. (2025). Clinical Risk Prediction with Logistic Regression: Best Practices, Validation Techniques, and Applications in Medical Research. *Academic Medicine & Surgery*. https://doi.org/10.62186/001c.131964.
[36] Hua, K., Wojdyla, D., Carnicelli, A., Granger, C., Wang, X., & Hong, H. (2025). Network Meta-Analysis With Individual Participant-Level Data of Time-to-Event Outcomes Using Cox Regression. *Statistics in Medicine*, 44. https://doi.org/10.1002/sim.70027.
[37] Hünermund, P., & Louw, B. (2025). On the Nuisance of Control Variables in Causal Regression Analysis. *Organizational Research Methods*, *28*(1), 138–151. https://doi.org/10.1177/10944281231219274
[38] Hosmer, D., Lemeshow, S., & Sturdivant, R. (2005). Introduction to the Logistic Regression Model., 1-30. https://doi.org/10.1002/9781118548387.ch1.
[39] Howell-Moroney, M. (2023). Inconvenient truths about logistic regression and the remedy of marginal effects. *Public Administration Review*. https://doi.org/10.1111/puar.13786.
[40] Iqbal, M. A. (2021). *Article 4 Application of Regression Techniques with their Advantages and Disadvantages*. *September*.
[41] Jiang, N., Wu, Y., & Li, C. (2024). Limitations of using COX proportional hazards model in cardiovascular research. *Cardiovascular Diabetology*, 23. https://doi.org/10.1186/s12933-024-02302-2.
[42] Johnston, R., Jones, K., & Manley, D. (2017). Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour. *Quality & Quantity*, 52, 1957 - 1976. https://doi.org/10.1007/s11135-017-0584-6.
[43] Josephine, K., O., Olowe, K., Edoh, N., Jean, S., Zouo, C., & Olamijuwon, J. (2024). Comprehensive review of logistic regression techniques in predicting health outcomes and trends. *World Journal of Advanced Pharmaceutical and Life Sciences*. https://doi.org/10.53346/wjapls.2024.7.2.0039.
[44] Jones, L., Barnett, A., & Vagenas, D. (2024). Common misconceptions held by health researchers when interpreting linear regression assumptions, a cross-sectional study. *PLOS One*, 20. https://doi.org/10.1101/2024.02.15.24302870.
[45] Jóźwiak, K., Nguyen, V., Sollfrank, L., Linn, S., & Hauptmann, M. (2024). Cox proportional hazards regression in small studies of predictive biomarkers. *Scientific Reports*, 14. https://doi.org/10.1038/s41598-024-64573-9.
[46] Kalbfleisch, J., & Schaubel, D. (2022). Fifty Years of the Cox Model. *Annual Review of Statistics and Its Application*. https://doi.org/10.1146/annurev-statistics-033021-014043.
[47] Kang, N., Elsner, J. B., Ren, Q., Nishioka, S., Shirato, H., Vega, J., & Murari, A. (2017). *Multicollinearity and Regression Analysis Multicollinearity and Regression Analysis*.
[48] Kim, J. (2019). Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology*, 72, 558 - 569. https://doi.org/10.4097/kja.19087.
[49] Knafl, G. (2023). Adaptive Conditional Hazard Regression Modeling of Multiple Event Times. *Open Journal of*

Statistics. https://doi.org/10.4236/ojs.2023.134025.
[50] Kuitunen, I., Ponkilainen, V., Uimonen, M., Eskelinen, A., & Reito, A. (2021). Testing the proportional hazards assumption in cox regression and dealing with possible non-proportionality in total joint arthroplasty research: methodological perspectives and review. *BMC Musculoskeletal Disorders*, 22. https://doi.org/10.1186/s12891-021-04379-2.
[51] Kundu P, Tang R, Chatterjee N. Generalized meta-analysis for multiple regression models across studies with disparate covariate information. Biometrika. 2019 Sep;106(3):567-585. doi: 10.1093/biomet/asz030. Epub 2019 Jul 13. PMID: 31427822; PMCID: PMC6690173.
[52] Lánczky, A., & Győrffy, B. (2021). Web-Based Survival Analysis Tool Tailored for Medical Research (KMplot): Development and Implementation. *Journal of Medical Internet Research*, 23. https://doi.org/10.2196/27633.
[53] Lapach, S. (2025). Conflict of user accuracy requirements and regression model indicators. *Mathematical machines and systems*. https://doi.org/10.34121/1028-9763-2025-1-91-102.
[54] LaValley, M. (2008). Logistic Regression. *Circulation*, 117, 2395-2399. https://doi.org/10.1161/circulationaha.106.682658.
[55] Lee, S. (2023). Kaplan-Meier and Cox proportional hazards regression in survival analysis: statistical standard and guideline of Life Cycle Committee. *Life Cycle*. https://doi.org/10.54724/lc.2023.e8.
[56] Lever, J., Krzywinski, M., & Altman, N. (2016). Points of Significance: Logistic regression. *Nature Methods*, 13, 541-542. https://doi.org/10.1038/nmeth.3904.
[57] Li, G., Lesperance, M., & Wu, Z. (2020). Joint Modeling of Multivariate Survival Data With an Application to Retirement. *Sociological Methods & Research*, 51, 1920 - 1946. https://doi.org/10.1177/0049124120914928.
[58] Li, Y., Y., Xu, H., Sun, Y., Zhu, M., Yue, W., Zhou, W., & Bi, W. (2025). Applying weighted Cox regression to genome-wide association studies of time-to-event phenotypes. *Nature Computational Science*, 5, 1064 - 1079. https://doi.org/10.1038/s43588-025-00864-z.
[59] Lunn, M., & McNeil, D. (1995). Applying Cox regression to competing risks. *Biometrics*, 51 2, 524-32 . https://doi.org/10.2307/2532940.
[60] Luo, X. (2016). *A Comparison of Three Estimation Methods In Linear Regression Analysis*. *71*(Icmmita 2016), 0–4. https://doi.org/10.2991/icmmita-16.2016.92
[61] Mansournia, M., Geroldinger, A., Greenland, S., & Heinze, G. (2018). Separation in Logistic Regression: Causes, Consequences, and Control. *American journal of epidemiology*, 187 4, 864-870. https://doi.org/10.1093/aje/kwx299.
[62] Marill, K. A. (2004). Advanced Statistics: Linear Regression, Part II: Multiple Linear Regression. *Academic Emergency Medicine*, *11*(1), 94–102. https://doi.org/10.1197/j.aem.2003.09.006
[63] Martens, E., De Boer, A., Pestman, W., Belitser, S., Stricker, B., & Klungel, O. (2008). Comparing treatment effects after adjustment with multivariable Cox proportional hazards regression and propensity score methods. *Pharmacoepidemiology and Drug Safety*, 17. https://doi.org/10.1002/pds.1520.
[64] McGregor, D., Palarea-Albaladejo, J., Dall, P., Hron, K., Chastin, S., & Chastin, S. (2019). Cox regression survival analysis with compositional covariates: Application to modelling mortality risk from 24-h physical activity patterns. *Statistical Methods in Medical Research*, 29, 1447 - 1465. https://doi.org/10.1177/0962280219864125.
[65] Mizumoto, A. (2022). Calculating the Relative Importance of Multiple Regression Predictor Variables Using Dominance Analysis and Random Forests. *Language Learning*. https://doi.org/10.1111/lang.12518
[66] Rothman, Kenneth J, Sander Greenland, and Timothy L Lash. *Modern Epidemiology*. 3rd ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins, 2008. Print.
[67] Monikapreethi S.K., Preetha, J., Reddy, K., Ramya, S., S, Y., & Murugan, S. (2024). Survival Analysis with Cox Proportional Hazards Model in Predicting Patient Outcomes. *2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 1155-1161. https://doi.org/10.1109/icesc60852.2024.10689732.
[68] Moolgavkar, S., Chang, E., Watson, H., & Lau, E. (2018). An Assessment of the Cox Proportional Hazards Regression Model for Epidemiologic Studies. *Risk Analysis*, 38. https://doi.org/10.1111/risa.12865.
[69] Mood, C. (2010). Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It. *European Sociological Review*, 26, 67-82. https://doi.org/10.1093/esr/jcp006.
[70] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. Wiley. https://books.google.com.pk/books?id=0yR4KUL4VDkC
[71] Muda, M. A. D., Affandi, A., & Suprapto, Y. K. (2020). *Forecasting Medicine Purchase Budget using Multiple Linear Regression Method: Case Study - For Ende Regency Health Office*. *Icasess 2019*, 186–192. https://doi.org/10.5220/0009880501860192
[72] Niu, L. (2018). A review of the application of logistic regression in educational research: common issues, implications, and suggestions. *Educational Review*, 72, 41 - 67. https://doi.org/10.1080/00131911.2018.1483892.
[73] Pal, A. (2021). Logistic regression: A simple primer. *Cancer Research, Statistics, and Treatment*, 4, 551 - 554. https://doi.org/10.4103/crst.crst_164_21.
[74] Park, H. (2013). An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. *Journal of Korean Academy of Nursing*, 43 2, 154-64. https://doi.org/10.4040/jkan.2013.43.2.154.
[75] Peng, C., Lee, K., & Ingersoll, G. (2002). An Introduction to Logistic Regression Analysis and Reporting. *The Journal of*

Educational Research, 96, 14 - 3. https://doi.org/10.1080/00220670209598786.

[76] Pinheiro-Guedes, L., Martinho, C., & Martins, M. (2024). Logistic Regression: Limitations in the Estimation of Measures of Association with Binary Health Outcomes. *Acta medica portuguesa*, 37 10, 697-705. https://doi.org/10.20344/amp.21435.

[77] Qu, K. (2024). Research on linear regression algorithm. *MATEC Web of Conferences*. https://doi.org/10.1051/matecconf/202439501046.

[78] Ranganathan, P., Pramesh, C., & Aggarwal, R. (2017). Common pitfalls in statistical analysis: Logistic regression. *Perspectives in Clinical Research*, 8, 148 - 151. https://doi.org/10.4103/picr.picr_87_17.

[79] Ray-Mukherjee, J., Nimon, K., Mukherjee, S., Morris, D., Slotow, R., & Hamer, M. (2014). Using commonality analysis in multiple regressions: a tool to decompose regression effects in the face of multicollinearity. *Methods in Ecology and Evolution*, 5. https://doi.org/10.1111/2041-210x.12166

[80] Roustaei, N. (2024). Application and interpretation of linear-regression analysis. *Medical Hypothesis, Discovery, and Innovation in Ophthalmology*, *13*(3), 151–159. https://doi.org/10.51329/mehdiophthal1506

[81] Samuelsen, S. (2022). Cox regression can be collapsible and Aalen regression can be non-collapsible. *Lifetime Data Analysis*, 29, 403 - 419. https://doi.org/10.1007/s10985-022-09578-0.

[82] Schober, P., & Vetter, T. R. (2021). Linear Regression in Medical Research. *Anesthesia and Analgesia*, *132*(1), 108–109. https://doi.org/10.1213/ANE.0000000000005206

[83] Seabrook, J. (2025). Powering Nutrition Research: Practical Strategies for Sample Size in Multiple Regression. *Nutrients*, 17. https://doi.org/10.3390/nu17162668

[84] Shah, A., & Patel, R. (2022). Heart Disease Prediction Based on Machine Learning. *International Journal for Research in Applied Science and Engineering Technology*, *10*(8), 1027–1036. https://doi.org/10.22214/ijraset.2022.46341

[85] Sheng, A., & Ghosh, S. (2020). Effects of Proportional Hazard Assumption on Variable Selection Methods for Censored Data. *Statistics in Biopharmaceutical Research*, 12, 199 - 209. https://doi.org/10.1080/19466315.2019.1694578.

[86] Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica*, 24, 12 - 18. https://doi.org/10.11613/bm.2014.003.

[87] Stanton, J. (2001). Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors. *Journal of Statistics Education*, 9. https://doi.org/10.1080/10691898.2001.11910537.

[88] Starbuck, C. (2023). The Fundamentals of People Analytics. In *The Fundamentals of People Analytics*. https://doi.org/10.1007/978-3-031-28674-2

[89] Stoltzfus, J. (2011). Logistic regression: a brief primer. *Academic emergency medicine: official journal of the Society for Academic Emergency Medicine*, 18 10, 1099-104. https://doi.org/10.1111/j.1553-2712.2011.01185.x.

[90] Sykes, A. O., & Sykes, A. O. (1993). *An Introduction to Regression Analysis An Introduction to Regression Analysis*. *20*.

[91] Tamhane, A.C. (2020). Multiple linear regression: model diagnostics. In Predictive Analytics, A.C. Tamhane (Ed.). https://doi.org/10.1002/9781119464761.ch4

[92] Taubert, H., Würl, P., Greither, T., Kappler, M., Bache, M., Bartel, F., Kehlen, A., Lautenschläger, C., Harris, L. C., Kaushal, D., Füssel, S., Meye, A., Böhnke, A., Schmidt, H., Holzhausen, H. J., & Hauptmann, S. (2007). Stem cell-associated genes are extremely poor prognostic factors for soft-tissue sarcoma patients. *Oncogene*, *26*(50), 7170–7174. https://doi.org/10.1038/sj.onc.1210530

[93] Taylor, M. Z. (2011). Regression analysis. In *21st Century Political Science: A Reference Handbook* (Issue March 2014). https://doi.org/10.4135/9781412979351.n57

[94] Tsiatis, A. (1981). A Large Sample Study of Cox's Regression Model. *Annals of Statistics*, 9, 93-108. https://doi.org/10.1214/aos/1176345335.

[95] Uddin, R. (2023). *COVID-19 Pandemic Data Analysis and Prediction Using Machine-Learning Algorithms*. 1–11.

[96] Uyanık, G. K., & Güler, N. (2013). A Study on Multiple Linear Regression Analysis. *Procedia - Social and Behavioral Sciences*, *106*, 234–240. https://doi.org/10.1016/j.sbspro.2013.12.027

[97] Venter, A., & Maxwell, S. (2000). Issues in the Use and Application of Multiple Regression Analysis., 151-182. https://doi.org/10.1016/b978-012691360-6/50007-0

[98] Vogt, W., & Johnson, R. (2015). Correlation and Regression Analysis. *Correlation and Regression Analysis*. https://doi.org/10.4135/9781446286104

[99] Werner, A., (2004), Purposes and strategies in regression analysis, Journal of Statistical Planning and Inference 122(1-2):175-186, DOI:10.1016/j.jspi.2003.06.018

[100] Whittingham, M., Stephens, P., Bradbury, R., & Freckleton, R. (2006). Why do we still use stepwise modelling in ecology and behaviour? *The Journal of animal ecology*, 75 5, 1182-9. https://doi.org/10.1111/j.1365-2656.2006.01141.x

[101] Yay, M. (2023). Assessment of the factors that affect fast-track or early extubation following pediatric cardiac surgery: Logistic regression in clinical studies. *Turkish Journal of Thoracic and Cardiovascular Surgery*, 31, 8 - 10. https://doi.org/10.5606/tgkdc.dergisi.2023.98550.

[102] Yoshida, T., & Murai, J. (2021). Various issues surrounding the use of multiple regression analysis in psychological research. *Japanese Journal of Psychology*. https://doi.org/10.4992/jjpsy.92.19226

[103] Zapf, A., Wiessner, C., & König, I. (2024). Regression Analyses and Their Particularities in Observational Studies—Part 32 of a Series on Evaluation of Scientific Publications. *Deutsches Arzteblatt international*, Forthcoming. https://doi.org/10.3238/arztebl.m2023.0278.

[104] Zeng, Z., Gao, Y., Li, J., Zhang, G., Sun, S., Wu, Q., Gong, Y., & Xie, C. (2022). Violations of proportional hazard assumption in Cox regression model of transcriptomic data in TCGA pan-cancer cohorts. *Computational and Structural Biotechnology Journal*, 20, 496 - 507. https://doi.org/10.1016/j.csbj.2022.01.004.