*Wits Journal of Clinical Medicine,* 2025, 7(2) 99–104 http://dx.doi.org/10.18772/26180197.2025.v7n2a12

Statistics @ WJCM

# Making sense of regression models in clinical research: a guide to interpreting beta coefficients and odds ratios

Okechinyere Achilonu*, Nneamaka O. Echendu, Glory Chidumwa

Division of Epidemiology and Biostatistics, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand

***Corresponding Author:*** *okechinyere.achilonu@wits.ac.za*

## ABSTRACT

Regression models play a central role in clinical research by quantifying the relationship between outcomes and explanatory variables. Accurate interpretation of model outputs, such as beta coefficients in linear regression and odds ratios in logistic regression, is critical for drawing valid conclusions. This article focuses on the key principles for interpreting linear and logistic regression results in clinical research. A publicly available heart disease dataset was used to demonstrate this. Linear regression was applied to a continuous outcome (cholesterol), while logistic regression modelled a binary outcome (presence of heart disease). Model building involved purposeful variable selection to account for confounding. Model adequacy and multicollinearity were assessed using goodness-of-fit statistics and variance inflation factors, respectively. Interpretation of results focused on beta coefficients and odds ratios. Regression models are robust methods for analysing clinical data and identifying key predictors. The concepts presented in this manuscript provide a foundational framework for applying and understanding linear and logistic regression in clinical research.

## INTRODUCTION

Regression models play a fundamental role in clinical research, providing a statistical framework to understand the relationships between a clinical outcome of interest (response or dependent variable) and one or more explanatory variables (also called predictors, exposure, or risk factors) that the researchers choose to evaluate.(1–4) For instance, whether estimating how age affects blood pressure or determining which patient characteristics predict clinical complication status, regression analysis helps researchers derive meaningful and quantifiable insights from data.

Understanding the nature of the outcome variable is essential when selecting the appropriate regression model.(2,5,6) Linear regression is suitable for outcome variables measured continuously, such as cholesterol level or blood pressure. In contrast, Logistic regression is appropriate for binary outcomes such as disease status (yes/no) and provides estimates in Log odds or odds ratios. In prospective (cohort) studies, risk ratios may offer more intuitive interpretations of effect size. They can be estimated using alternative regression models, such as log-binomial or Poisson regression with robust standard errors.(7–9) All regression models are governed by assumptions about the data and relationships being modelled; these assumptions are essential to justify the validity of model estimates and interpretations.(7,10) This article introduces key principles for interpreting linear and logistic regression results in clinical research. We focus on how linear regression applies to a continuous outcome and logistic regression to a binary outcome, emphasising the interpretation of beta coefficients and odds ratios.

To illustrate the regression concepts, we used a widely cited heart disease prediction dataset from the Kaggle Repository.(11) The dataset includes demographic, clinical, and diagnostic data as described in Table 1.

## LINEAR REGRESSION

### Simple linear regression

Simple or univariable linear regression is used to estimate the effect of one explanatory variable on the continuous outcome variable. In our example, the impact of each explanatory variable on cholesterol level (outcome variable) was examined. Simple linear regression models generate a beta or regression coefficient ($\beta_1$) which represents the expected change in $Y$ (outcome variable) for a one-unit increase in $X$ (explanatory variable). A positive beta coefficient indicates a direct relationship between the explanatory

**Table 1:** Demographic and clinical characteristics of study participants

| SN | Variable | Description (Mean ± SD or Frequency) |
|---|---|---|
| 1 | Age | Age in years (mean = 54.3, SD = 9.07) |
| 2 | Sex | Male = 1 (713); Female = 0 (299) |
| 3 | Cp | Chest pain type [0 = typical angina (490), 1 = atypical angina (167), 2 = non-anginal pain (278), 3 = asymptomatic (77)] |
| 4 | Restbps | Resting blood pressure in mm Hg on admission to the hospital (131.6 ±17.6) |
| 5 | Chol | Serum cholesterol in mg/dL (243.4 ± 46.1) |
| 6 | Fbs | Fasting blood sugar > 120 mg/dL [1 = true (150); 0 = false (862)] |
| 7 | Restecg | Resting electrocardiographic results [0 = normal (484), 1 = abnormal (528)] |
| 8 | Thalach | Maximum heart rate achieved (149.0 ± 23.1) |
| 9 | Exang | Exercise induced angina [1 = yes (342); 0 = no (670)] |
| 10 | Oldpeak | ST depression induced by exercise relative to rest (1.1 ± 1.2) |
| 11 | Slope | Slope of the peak exercise ST segment [(0 = upsloping (74), 1 = flat (472), 2 = down sloping (466)] |
| 12 | Ca | Number of major vessels coloured by fluoroscopy [0 = (575), 1 = (223),2 = (131), 3 (83)] |
| 13 | Thal | Thalassemia status 1 = normal (71) ; 2 = fixed defect (541); 3 = reversible defect (400) |
| 14 | Heart disease | 0 = no disease (492) and 1 = disease (520) |

variable and the outcome, a negative coefficient indicates an inverse relationship, and a coefficient of zero suggests no linear relationship. The constant value ($\beta_0$) is also commonly provided by statistical packages and is interpreted as the expected value of $Y$ when $X = 0$. It is important to note that the intercept is often not clinically meaningful, especially when "zero" is not a realistic or interpretable value of the explanatory variable.

In the univariable model presented here, age was significantly associated with cholesterol level (Table 2). Each additional year of age was associated with an estimated 0.98 mg/dL increase in cholesterol level (95% CI: 0.67,1.29; $p < 0.001$). The confidence interval (CI) and p-value are key outputs for interpreting regression coefficients. The 95% confidence interval reflects the precision of the estimated effect. The narrower the interval, the more precise the estimate. The CI also indicates statistical significance if it excludes zero. P-values below 0.05 suggest the relationship is unlikely due to chance (statistically significant). The regression model also estimates the coefficient of determination ($R^2$)and the F-statistic. The $R^2$ value measures the proportion of variance in the outcome variable explained by the explanatory variable. In the univariable model (Table 2), the relationship between cholesterol levels and age yielded an $R^2$ of 0.037, indicating that approximately 3.7% of the variability in cholesterol levels is explained by age. Low values are common in clinical research due to the multifactorial nature of health outcomes.(12) Despite the low $R^2$, the model is statistically significant (F-statistic (1, 1010) = 38.96, $p < 0.001$), suggesting that age is a significant explanatory variable of cholesterol levels in this dataset. Linear regression assumes that the outcome variable is linearly related to the predictors and that residuals are normally distributed, independent, and homoscedastic.(7) These assumptions were valid for this dataset.

## Multiple linear regression

Multiple (Multivariable) linear regression allows the inclusion of additional explanatory variables to adjust for confounding. In multiple linear regression, the goal is not only to model the outcome variable $Y$, but also to identify which explanatory variables contribute most significantly to explaining the variability in $Y$. In the multivariable linear regression model presented here, we estimated the independent effect of each variable on cholesterol levels while adjusting for the influence of other covariates.

Multivariable linear regression identified age, sex, chest pain type (cp), resting blood pressure (trestbps), and resting ECG results (restecg) as significant factors associated with cholesterol levels (Table 2). The interpretation of the effects in the above model follows that of simple linear regression; however, the researcher needs to acknowledge the adjustment of other variables in the model. For instance, one would state that a one-year increase in age significantly increases the cholesterol level by 0.72 (95% CI: 0.40, 1.04; $p < 0.001$) on average, adjusting for other explanatory variables in the model. Similarly, the model shows that after adjusting for other variables, higher resting blood pressure was associated with higher cholesterol, while being male and having abnormal ECG or certain chest pain types were associated with lower cholesterol levels.

Multicollinearity among predictor variables can be checked using the variance inflation factor (VIF), as high multicollinearity can inflate the standard errors of the

**Table 2:** Univariable and multivariable linear regression results

| Variable | Category | Univariable analysis | | Multivariable analysis | |
|---|---|---|---|---|---|
| | | **Beta Coeff (95% CI)** | ***p*-value** | **Beta Coeff (95% CI)** | ***p*-value** |
| Age | | 0.98 (0.67,1.29) | <0.001 | 0.72 (0.40,1.04) | <0.001 |
| Cp | 0 | ref | | ref | |
| | 1 | −3.04 (11.10,5.03) | 0.460 | 1.83 (−6.16,9.82) | 0.724 |
| | 2 | −10.90 (−17.66,−4.14) | 0.002 | −8.59 (−15.25,−1.93) | 0.012 |
| | 3 | −10.67 (−21.71,0.36) | 0.058 | −11.99 (−22.77,−1.21) | 0.029 |
| Ca | | ref | | | |
| | 1 | 5.06 (−2.06,12.18) | 0.163 | | |
| | 2 | 10.16 (1.42,18.89) | 0.023 | NS | |
| | 3 | 6.56 (−4.03,17.15) | 0.224 | | |
| Sex | 0 | ref | | Ref | |
| | 1 | −14.19 (−20.36,−8.03) | <0.001 | −9.89 (−15.56,−4.23) | <0.001 |
| Thal | 0 | ref | | | |
| | 1 | 17.46 (6.09,28.82) | 0.003 | NS | |
| | 2 | 17.99 (6.40,29.59) | 0.002 | | |
| Oldpeak | | 0.97 (−1.47,3.41) | 0.435 | | |
| Exang | 0 | ref | | | |
| | 1 | 9.15 (3.17,15.14) | 0.003 | NS | |
| Thalach | | −0.08 (−0.21,0.04) | 0.188 | NS | |
| Trestbps | | 0.37 (0.21,0.53) | <0.001 | 0.23 (0.06,0.39) | 0.006 |
| Restecg | 0 | ref | | ref | |
| | 1 | −12.93 (−17.74,−7.26) | 0.000 | −9.89 (−15.56,−4.23) | 0.001 |
| Fbs | 0 | ref | | NS | |
| | 1 | 2.97 (−5.03,10.97) | 0.466 | | |
| Slope | 0 | ref | | | |
| | 1 | 7.14 (−4.16,18.44) | 0.215 | NS | |
| | 2 | 4.68 (−6.63,15.99) | 0.417 | | |

regression coefficients, leading to unstable estimates and reducing the model's interpretability. Although there are no universally accepted rules for determining when a VIF is excessively high, commonly used guidelines suggest that VIF values of 5 or 10 and above may indicate sufficiently strong collinearity to warrant corrective action.(15)

## LOGISTIC REGRESSION

### Simple logistic regression

Simple logistic regression is used to model the log odds of an outcome with one explanatory variable. The regression/Beta coefficient reflects the change in the log odds of the outcome for a one-unit increase in the explanatory variable, while the intercept or constant represents the log odds of the outcome when $X = 0$.

In the logistic regression example presented here, we investigated factors associated with the presence of heart disease using both univariable and multivariable logistic regression models, as carried out in linear regression analyses. When looking at the association of age with heart disease, the coefficient for age is −0.055, indicating that for each additional year of age, the log odds of having heart disease decrease by 0.055.

Although this is a valid interpretation, log odds are not easily interpretable for many clinical researchers and are not conventionally presented. For this reason, it is common practice to express the regression coefficient in its

**Table 3:** Univariable and multivariable logistic regression results

| Variables | Categories | Univariable analysis | | | Multivariable analysis | |
|---|---|---|---|---|---|---|
| | | Log-Odds Coef (95% CI) | OR (95%CI) | P-value | OR (95% CI) | P-value |
| Age | | −0.05 (−0.07,−0.04) | 0.95 (0.93,0.96) | <0.001 | NS | |
| Cp | 0 | ref | ref | | ref | |
| | 1 | 2.51 (2.07,2.94) | 12.25 (7.95,18.8) | <0.001 | 3.11 (1.71,5.65) | <0.001 |
| | 2 | 2.29 (1.95,2.64) | 9.88 (7.00,13.96) | <0.001 | 7.24 (4.20,12.50) | <0.001 |
| | 3 | 1.78 (1.26,2.29) | 5.92 (3.54,9.90) | <0.001 | 10.66 (5.07,22.40) | <0.001 |
| Ca | 0 | ref | ref | | ref | |
| | 1 | −1.86 (−2.20,−1.52) | 0.16 (0.11,0.22) | <0.001 | 0.12 (0.07,0.20) | <0.001 |
| | 2 | −2.58 (−3.08,−2.08) | 0.08 (0.05,0.12) | <0.001 | 0.05 (0.02,0.10) | <0.001 |
| | 3 | −1.83 (−2.33,−1.32 | 0.16 (0.10,0.27) | <0.001 | 0.29 (0.13,0.66) | 0.003 |
| Sex | 0 | ref | ref | | ref | |
| | 1 | −1.34 (−1.64,−1.05) | 0.26 (0.19,0.35) | <0.001 | 0.15 (0.09,0.28) | <0.001 |
| Thal | 1 | ref | ref | | ref | |
| | 2 | 1.80 (1.27,2.33) | 6.07 (3.57,10.30) | <0.001 | 0.87 (0.38,1.99) | 0.744 |
| | 3 | −0.61 (−1.15,−0.06) | 0.54 (0.32,0.94) | 0.029 | 0.23 (0.10,0.51) | <0.001 |
| Oldpeak | | −0.98 (−1.13,−0.82) | 0.38 (0.32,0.44) | <0.001 | 0.62 (0.49,0.79) | <0.001 |
| Exang | 0 | ref | ref | | ref | |
| | 1 | −2.05 (−2.36,−1.74) | 0.13 (0.09,0.18) | <0.001 | 0.50 (0.31,0.80) | 0.004 |
| Thalach | | 0.04 (0.04,0.05) | 1.05 (1.04,1.05) | <0.001 | 1.02 (1.01,1.03) | 0.003 |
| Chol | | −0.01 (−0.01,0.00) | 0.99 (0.99,1.00) | <0.001 | 0.99 (0.99,1.00) | 0.002 |
| Trestbps | | −0.02 (−0.02,−0.01) | 0.98 (0.98,0.99) | <0.001 | 0.98 (0.97,0.99) | 0.001 |
| Restecg | 0 | ref | ref | | NS | |
| | 1 | 0.65 (0.40,0.90) | 1.92 (1.49,2.46) | <0.001 | | |
| Fbs | | −0.28 (−0.63,0.06) | 0.75 (0.53,1.07) | 0.109 | NS | |
| Slope | 1 | ref | ref | | ref | |
| | 2 | −0.22 (−0.73,0.29) | 0.80 (0.48,1.33) | 0.398 | 0.56 (0.23,1.37) | 0.203 |
| | 3 | 1.46 (0.94,1.97) | 4.29 (2.57,7.16) | <0.001 | 1.85 (0.71,4.85) | 0.211 |

exponential form, yielding the odds ratio (OR). The OR compares the odds (likelihood) of the outcome per unit change in the explanatory variable, which can be either continuous or categorical. An OR of 1 indicates no association between the predictor and the outcome. An OR greater than 1 suggests increased odds of the outcome, while an OR less than 1 indicates decreased odds.(7) For example, the OR corresponding to the age coefficient of −0.055 is calculated as $e^{-0.055} = 0.946$, indicating that each additional year of age is associated with a 5.4% decrease in the odds of heart disease.

For categorical variables such as sex (Table 3), where male is coded as "1" and female as "0", the odds ratio comparing males to females is $e^{-1.344} = 0.266$. This can also be interpreted as a per cent change using the formula: $\%\ change = 100 \times (\mathrm{OR} - 1)\%$. Thus, the odds of having heart disease are approximately 73.4% lower in males than in females.

As with linear regression, a p-value below 0.05 suggests the relationship is unlikely to be due to chance; furthermore, the 95% confidence interval indicates significance if it excludes one.

## Multiple logistic regression

The simple logistic regression model can be extended to include multiple explanatory variables, resulting in a multivariable logistic regression model where several explanatory variables are included in the model, and each *Beta coefficient* represents the change in the log odds of the outcome associated with a one-unit increase in the corresponding variable, holding all other variables constant.

In the example shown in Table 3, 10 variables were retained in the multivariable logistic regression model based on the purposeful variable selection method. Specifically, patients with asymptomatic chest pain (cp = 3) had approximately 10.7-fold higher odds of developing heart disease compared to those with typical angina (reference group) (OR = 10.66, 95% CI: 5.07–22.40). Similarly, atypical angina (cp = 2) also showed a strong positive association with heart disease (OR = 7.24), while being male was associated with significantly lower odds of heart disease (OR = 0.15, 95% CI: 0.09–0.28). Variables such as slope were not statistically significant ($p > 0.05$), indicating no strong independent effect in the presence of other explanatory variables.

Like linear regression, the multivariable model showed good fit to the data, with a highly significant likelihood ratio test ($\chi^2$ (16) = 773.88, $p < 0.001$) and a pseudo $R^2$ of 0.55.

## CONCLUSION

Regression models are powerful tools in clinical research that enable relationships between outcome and explanatory variables. This paper underscores the importance of selecting appropriate models based on the nature of the outcome variable, while ensuring that model assumptions hold, and correctly interpreting results such as beta coefficients and odds ratios. Using linear and logistic regression examples, including univariable and multivariable analyses, we demonstrated how these models can be used to model continuous and binary outcomes while accounting for confounding variables.

## REFERENCES


1. Del Águila MR, Benítez-Parejo N. Simple linear and multivariate regression models. Allergol Immunopathol (Madr). 2011; 39(3):159–173.
2. Nunez E, Steyerberg EW, Nunez J. Regression modeling strategies. Rev Esp Cardiol (Engl Ed). 2011; 64(6):501–507.
3. Sheather S. A modern approach to regression with R. Berlin: Springer Science & Business Media; 2009.
4. Worster A, Fan J, Ismaila A. Understanding linear and logistic regression analyses. Can J Emerg Med. 2007; 9(2):111–113.
5. Castro HM, Ferreira JC. Linear and logistic regression models: when to use and how to interpret them? SciELO Brasil; 2022.
6. Gail M, Krickeberg K, Samet J, Tsiatis A, Wong W. Statistics for biology and health. Berlin: Springer; 2007.
7. Vittinghoff E, Glidden DV, Shiboski SC, McCulloch CE. Regression methods in biostatistics: linear, logistic, survival, and repeated measures models. Berlin: Springer Science & Business Media; 2012.
8. Zhu C, Blizzard C, Stankovich J, Wills K, Hosmer DW. Be wary of using Poisson regression to estimate risk and relative risk. Biostat Biom Open Access J. 2018; 4(5):555649.
9. Andrade C. Understanding relative risk, odds ratio, and related terms: as simple as it can get. J Clin Psychiatry. 2015; 76(7):21865.
10. Achilonu OJ, Fabian J, Musenge E. Modeling long-term graft survival with time-varying covariate effects: an application to a single kidney transplant centre in Johannesburg, South Africa. Front Public Health. 2019; 7:201.
11. Ali MM, Paul BK, Ahmed K, et al. Heart disease prediction using supervised machine learning algorithms: performance analysis and comparison. Comput Biol Med. 2021; 136:104672.
12. Schober P, Boer C, Schwarte LA. Correlation coefficients: appropriate use and interpretation. Anesth Analg. 2018; 126(5):1763–1768.
13. Bursac Z, Gauss CH, Williams DK, Hosmer DW. Purposeful selection of variables in logistic regression. Source Code Biol Med. 2008; 3:1–8.
14. Hosmer DW, Lemeshow S, May S. Applied survival analysis: regression modeling of time-to-event data (Wiley series in probability and statistics). Hoboken, NJ: Wiley-Interscience; 2008: 60.
15. Craney TA, Surles JG. Model-dependent variance inflation factor cutoff values. Qual Eng. 2002; 14(3):391–403.