

https://dspace.org/

Regression analysis basics making the right choice of regression analysis to model clinical data

Moses Kebalepile; Palesa Motshabi

https://hdl.handle.net/10539/39771
SCOPUS

Downloaded from a DSpace repository. https://dspace.org/

South Afr J Anaesth Analg. 2023;29(5 Suppl 1):S20-22
https://doi.org/10.36303/SAJAA.3055
Open Access article distributed under the terms of the
Creative Commons License [CC BY-NC 3.0]
http://creativecommons.org/licenses/by-nc/3.0

South Afr J Anaesth Analg
ISSN 2220-1181 EISSN 2220-1173
© 2023 The Author(s)
FCA REFRESHER COURSE

Regression analysis basics: making the right choice of type of regression analysis to model clinical data

MM Kebalepile, PM Chakane

Department of Anaesthesia, School of Clinical Medicine, Faculty of Health Sciences, Charlotte Maxeke Johannesburg Academic Hospital, University of the Witwatersrand, South Africa
Corresponding author, email: *moses.kebalepile@wits.ac.za*

Data-driven clinical practice and evidence-based medicine present themselves as critical tenets of medical practice the world over. The use of data to achieve these tenets often involves using known clinical patient presentations and biological data to predict disease outcomes and other post-care events (i.e. major adverse effects of treatment or mortality). These predictions are achieved using statistical techniques, such as hypothesis testing and forecasting or prediction analysis, also accepted as regression analysis.

In its simplicity, regression analysis represents the statistical technique of applying certain mathematical computational equations to determine the relationship between variables. Therefore, in the use of such mathematical equations, a change in a unit of one variable (the independent variable) can lead to a predictable corresponding change in the dependent variable. This corresponding change can be a unit change, in the independent variable, or it can be a quantifiable factor of the unit of change observed in the independent variable.

The current paper takes an introductory and non-technical approach to discuss processes involved in regression analysis. The focus is the understanding and application of linear and logistic regression to clinical data. Other types of regression methods such as lasso, ridge, and polynomial regression analysis are mentioned but not fully discussed.

Keywords: linear regression, logistic regression, prediction, relationship between variables

Introduction

In statistical analysis, regression analysis is a technique used to test the relationships between variables and make predictions based on those relationships.[1] Regression analysis allows researchers and analysts to understand how changes in one variable (often called an independent variable) are associated with changes in another (called a dependent variable).[2] The ability to use the relationship between a dependent variable and an independent variable allows for the prediction of future values of the dependent variable.

This relationship might be causal in nature or deemed to just predict an association. The relationships are explained through variances, simple and multiple correlations, and regression coefficients, in an iterative process of fitting the regression of one variable on others.[1] The current paper serves to introduce the reader to the two most common linear regression models: linear regression analysis and logistic regression analysis.

Linear regression

There are various types of regression analysis methods, which include but are not limited to linear regression, polynomial regression, logistic regression, lasso and ridge regressions.[3] Figure 1 is a representation of the common types of regression analysis methods.

It is important to note that Figure 1 represents linear models in general and not specifically linear regression analysis. Linear regression analysis is one type or form of a linear model, and so are the other regression types in Figure 1.

Linear regression can be termed simple or multiple linear regression. Simple linear regression tests the effect or influence of one variable (independent variable) on a single dependent variable.[2,4] Equation 1 is a mathematical representation of the simple linear regression analysis. This is a general equation for a straight line. The equation is a derivative of a pair of simultaneous equations called normal equations, and the method of using these normal equations to derive a straight-line equation is called the least squares method.[5]

The normal equations are useful for a simple straight-line equation, while linear relations with curving lines require different sets of derivatives.[5] As in the name, multiple linear regression has multiple explanatory or predictor variables, often called independent variables, used to estimate the future value of a dependent variable.[2,4] Equation 2 represents multiple linear regression analysis. Figure 2 describes the statistical assumptions that must be met for linear regression analysis to be performed.

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad \textbf{Equation 1}^2$$

Where **y** is the dependent variable that is made of continuous data points; β_1 is the slope of the line; **x** is the independent

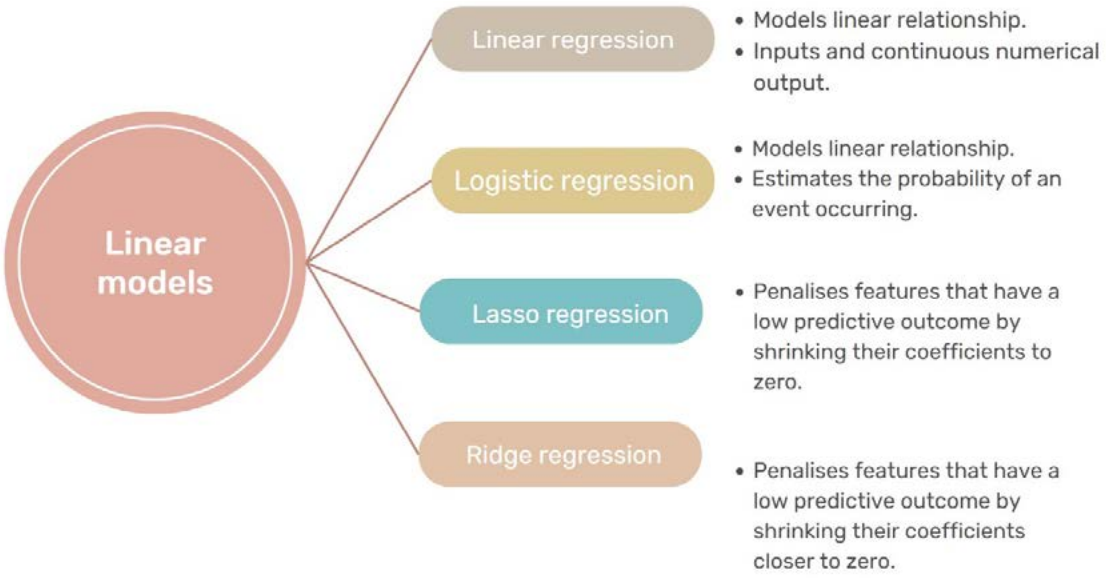


Figure 1: Types of regression models or algorithms (in this case all models assume linearity between inputs and output)

variable; and β_0 represents the intercept or the point where the straight line cuts the y-axis. Ɛ is the error term.

$$y = \beta_0 + \beta_1 x + \beta_2 x + \ldots\ldots \beta_n x_n + Ɛ \qquad \textbf{Equation 2}^2$$

Where x_1, x_2 **and** x_n represent the multiple independent variables. The rest of the equation has the same parameters as the simple linear regression equation (Equation 1).

When plotting a regression model, and fitting the best line, some points will fall on the line and others above or below the line of best fit. Those points vertically above or below the line of best fit are called residuals. The error terms in both equations 1 and 2 relate to these data points that do not perfectly fall on the line of best fit.

When applied in linear regression analysis, the least squares method aims to produce a fit where the sum of all the residuals approximates zero, or the sum is the least or smallest.[3] Therefore, this computation will produce two outcomes, where a) the squared sum of the residuals (SSRes) represents deviation that the model failed to explain or estimate, hence the relationship to the error term, and b) the regression sum of squares (RegSS), which represent the variation in the dependent variable (y) that is explained by the line of best fit.[4]

In the model, the RegSS would be derived as a sum of the squares of vertical distances from the line of best fit to the horizontal line where y is equal to the mean or average. Both the RegSS and the residual sum of squares (RSS) represent deviations from the line of best fit. The former deviation is called explained deviation, while the latter is called unexplained. Figure 3 demonstrates the relationship of the measures of variation in linear regression analysis.

The mathematical relationship of the measures of variation represented in Figure 3 can be described by Equation 3.

Total sum of squares (TSS) = RegSS + ResSS **Equation 3**

Where RegSS is the regression sum of squares (derived from the sum of the square vertical distances between the line of best fit and y is equal to the mean, and ResSS is the residual sum of squares (representing the sum of squared distances of all the vertical points above and below the line of best fit).[4]

Then, to understand how well the regression model is performing, a ratio of the RegSS over the TSS would indicate model performance. As with all ratios, this ratio falls between 0 and 1, with 1 indicating a good and ideal model performance and 0 indicating a poor model. The outcome of this ratio represents a regression correlation coefficient, called the coefficient of

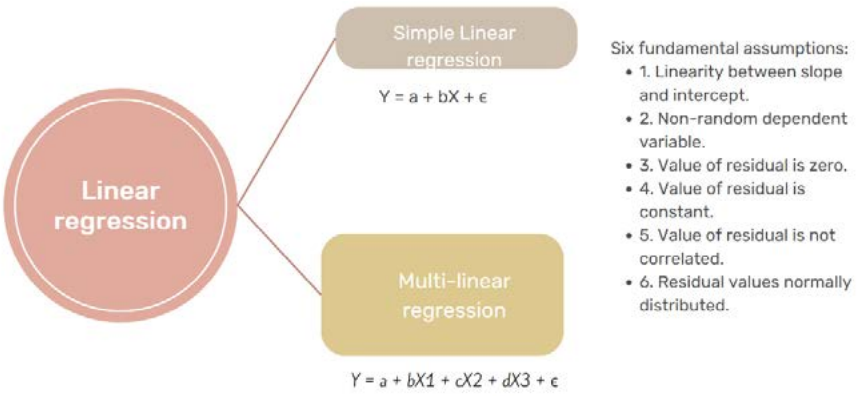


Figure 2: Types of linear regression analysis: simple versus multiple

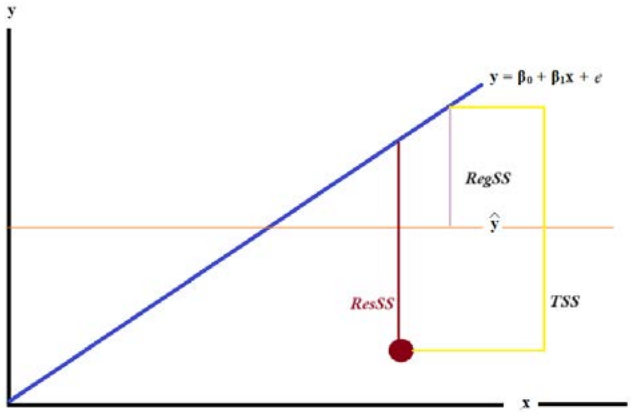


Figure 3: Measures of variation in linear regression analysis

determination (CD), denoted by R^2. It is a measure of correlation because, in classical linear regression analysis, the square root of the CD gives an output mathematically equal to the Pearson correlation coefficient.

To illustrate the simple linear regression, computer-generated dummy data with patient ages and STAT 2020 mortality scores were used. The mortality score was developed in 2009 and updated in 2020. It is a risk stratification tool used to stratify congenital heart disease surgical intervention into groups of increasing risk of mortality.[6] These variables were generated at random and represent no actual patients. To illustrate the concept of simple linear regression, a mortality score is regressed on age, which means age as the independent variable is used to predict the value of a mortality score.

In Figure 4, the *p*-value of the model is significant, indicating that the model is significant as a whole and can be used to describe the linear relationship between age and risk of mortality. Age is statistically significant and has a negative coefficient. This negative statistically significant relationship indicates that as age (in months) increases for paediatric patients with congenital heart defects, which require surgical intervention, the risk score decreases. Older paediatric patients are at a reduced risk of mortality compared to younger paediatric patients, who would in this example have a characteristically high risk score.

In this model, Model SS represents RegSS, which are the deviations explained by the model. The Residual SS represents the error term or the sum of squares of the residuals. Although the model is significant, the R^2 of the model is 6%. As discussed earlier, this comes from the ratio ResSS / TSS. The smaller the ratio, the weaker the model and the closer the ratio is to 1 or 100%, the better the model is at representing the variation in y, which is explained by the model.

The equation of the line is therefore:

STAT 2020 mortality score = 0.45 + (-0.02) * age (in months)

Therefore, when age increases by a unit (one month increase in age), the risk score decreases by 0.02. As such, this model allows us to predict what the risk score can be at any known value of age. At 12 months of age, the risk score can be calculated to be 0.21 = 0.45 + (-0.02 * 12).

Once a model such as in Figure 4 has been developed, the quality of the model must be tested. Several post-model tests, based on the assumptions that were made before the modelling process, are used to assess whether the developed model is reliable. The most used tests for the quality of the model include:

a. The y-estimates must fall within the range of the sample y-values. If this condition is not satisfied, it may be possible that at extreme ends, the relationship of y to x is non-linear.
b. The second condition to test is the distribution of the residuals. In a normal distribution, the data are symmetrical around the mean, and when testing the distribution of the residuals, it would be expected for the polygon to peak around zero.[7]
c. Test the variation of the residuals. A test of homoscedasticity or heteroscedasticity is often performed, where the variance of the residuals across a range of x-values is tested. An even variance, called homoscedasticity, is a desired outcome. There are statistical tests such as the Breusch-Pagan test or the White test for this equality of variance.
d. The independence of residuals otherwise represented as autocorrelation should be tested. It is desired that the residuals must be independent and not autocorrelated. A Breusch-

```
. regress STAT2020_score Age
```

Source	SS	df	MS		
				Number of obs	= 243
				F(1, 241)	= 14.82
Model	1.55362034	1	1.55362034	Prob > F	= 0.0002
Residual	25.2654743	241	.104835993	R-squared	= 0.0579
				Adj R-squared	= 0.0540
Total	26.8190947	242	.110822705	Root MSE	= .32378

STAT2020_s~e	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
Age	-.0194887	.0050625	-3.85	0.000	-.0294611	-.0095163
_cons	.4452594	.0267026	16.67	0.000	.392659	.4978598

Figure 4: STAT output for linear regression analysis where age is used to predict the risk of mortality score

Godfrey LM test may be computed to test the null hypothesis on the independence of the residuals.

e. Finally, a test for missing variables may be performed to assess the quality of a simple linear regression model.[2] A Ramsey RESET test can be performed and it will indicate if the model performs better if additional variables are added.

Figures 5, 6, and 7 represent the STATA outputs for the tests in b to e.

```
. predict residuals, resid
(6 missing values generated)

. swilk resid

                    Shapiro-Wilk W test for normal data

    Variable |   Obs        W         V         z       Prob>z
-------------+-------------------------------------------------
   residuals |   243    0.79745   35.829     8.315    0.00000
```

Figure 5: STATA output showing the generation of residuals as a new variable, then testing distribution

In Figure 5, the residuals are first generated as a variable in the data set. Then a Shapiro-Wilk test is applied to ascertain the nature of the probability of distribution of the residuals. The high significant *p*-value indicates that the residuals are not normally distributed. Figure 6 presents the two possible tests for testing the equality of variance of the residuals.

```
. hettest, rhs

Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: Normal error terms
Variable: All independent variables

H0: Constant variance

    chi2(1) =  10.87
Prob > chi2 = 0.0010
```

```
. imtest, white
White's test
H0: Homoskedasticity
Ha: Unrestricted heteroskedasticity

    chi2(2) =  11.12
Prob > chi2 = 0.0038

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df       p
---------------------+-----------------------------
  Heteroskedasticity |      11.12      2   0.0038
            Skewness |      12.39      1   0.0004
            Kurtosis |       1.83      1   0.1758
---------------------+-----------------------------
               Total |      25.35      4   0.0000
---------------------------------------------------
```

Figure 6: Test for equality of variance of the residuals

A Breusch-Godfrey test for autocorrelation tests the null hypothesis that postulates no autocorrelation and if significant, the null would be rejected. Finally, Figure 7 represents the result of the Ramsey RESET test.

After developing the model and testing the quality of the model, it can be accepted that although the model has a significant F statistic, and the relationship of age to the mortality risk score was found to be significant, a better model may be possible when other variables are added to the model to enhance the understanding of factors related to mortality in this case.

Although linear regression analysis has numerous applications in medical research, it is often limited by the nature of the outcome variable being studied.[2] It is limited to studying outcomes that are continuous, such as changes in blood pressure. However, the presence or absence of a specific outcome of medical intervention and factors associated with such outcomes are often the desired knowledge. These intervention outcomes can be adverse effects of a new treatment regime compared to a standard of care, or in-hospital mortality, and have binary dependent variables that cannot be studied using linear regression analysis. For binary outcomes, logistic regression analysis is the correct type of regression to test the linear relationship between inputs and output.[2,3]

```
. estat ovtest

Ramsey RESET test for omitted variables
Omitted: Powers of fitted values of STAT2020_score

H0: Model has no omitted variables

F(3, 238) =   7.13
 Prob > F = 0.0001
```

Figure 7: Ramsey RESET test with a significant *p*-value, indicating that the model misses some additional predictors

Logistic regression

Unlike in linear regression analysis, the estimating equation or method in logistic regression is the maximum likelihood (ML) method. The ML method uses a likelihood function, which is a probability function, where the probability of a certain binary outcome (Y) is the highest. This method allows for determining how the probability of success [P(Y) = 1)] can be affected by the presence of predictor variables (predictor variables that also have a probability of occurring in the population being studied).

In the previous linear regression model, the effect of age on the mortality risk score was tested. In a logistic regression, where the dependent variable can be mortality (where 1 means mortality present and 0 indicates the absence, with either state having an associated probability), it may be important to establish how the probability of mortality can be affected by the presence of comorbid disease in a patient. In this case, it may be possible that the presence of the comorbid disease may increase the probability of mortality. This conjecture is based on the premise that for non-mutually exclusive events (the presence of comorbid disease can occur together with mortality, and each event has a probability), probabilities are multiplicative and not additive.[8]

The logistic regression model also applies a mathematical formula in testing the relationship of the binary outcome, with independent variables that can also be binary or categorical. When the coefficients are exponentiated, the relationship of how one status affects the probability of the other status can then be presented as the odds of the latter occurring.[9] It is similar to linear regression but differs in the outcome variable (binary outcome) and has similar assumptions. Since it models the probability of an outcome that has a chance of being in one state or the other (outcome present versus outcome absent), this ratio is modelled as a logarithm of the chance of that outcome, hence the need to exponentiate the coefficient to attain interpretable odds ratios.[10] Pandis describes the mathematical equation to model this outcome and is called a logit.[9] Equation 4 represents the mathematical equation of the logistic regression. The deferential steps to derive the equation can be found in Pandis, 2017.[9]

$$\text{Log}(p/1 - p) = \beta_0 + \beta_1 x_1 + \beta_2 X_2 + \ldots \beta_n X_n$$

When developing logistic regression analysis, the analyst can either start from a full model and progress to remove very insignificant variables, one variable at a time (a process called stepwise backward regression analysis); or the starting point can be a null model, with only the intercept, and progress forward to add predictors.[10] To illustrate a logistic regression and interpret the results, the same data that were used for the linear regression analysis were used here.

Additionally, the computer-generated data included data on morbidity and mortality. As in linear regression analysis, once the logistic regression model is built, the next step is to assess the quality of the model, and similarly, there are a set of tests that can be performed.

An important consideration when setting off to compute a logistic regression analysis is the sample size. Smaller sample sizes with multiple variables produce bad-quality logistic regression models (models that have a high chance of overestimating the effect measured).[11] An events per variable (EPV) approach is a popular method used to avoid the overestimating problem due to insufficient sample size.[11] EPV is derived using the number of observations in the smaller of the two outcome groups, in relation to the number of predictor variables identified to be used in the model.[11,12] These variables are called candidate predictors because it may compromise the model to use all of them in a saturated model. A rule of thumb often cited when using the EPV is that an EPV of 10 is acceptable.[12]

```
. logit  Mortality  Morbidity STAT2020_score Age

Iteration 0:   log likelihood = -47.745809
Iteration 1:   log likelihood = -46.081753
Iteration 2:   log likelihood =  -44.20935
Iteration 3:   log likelihood = -44.196562
Iteration 4:   log likelihood = -44.196531
Iteration 5:   log likelihood = -44.196531

Logistic regression                               Number of obs =     242
                                                  LR chi2(3)    =    7.10
                                                  Prob > chi2   =  0.0688
Log likelihood = -44.196531                       Pseudo R2     =  0.0743
```

Mortality	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
Morbidity	1.02557	.8520574	1.20	0.229	-.6444321	2.695571
STAT2020_score	1.210641	.7198585	1.68	0.093	-.2002554	2.621538
Age	-.114241	.1227597	-0.93	0.352	-.3548456	.1263636
_cons	-3.357176	.6080281	-5.52	0.000	-4.548889	-2.165463

Figure 8: A saturated logistic regression model, built to derive a final model through stepwise backward regression analysis

```
. logit  Mortality STAT2020_score Age

Iteration 0:   log likelihood = -47.745809
Iteration 1:   log likelihood = -45.903483
Iteration 2:   log likelihood = -44.816492
Iteration 3:   log likelihood = -44.809873
Iteration 4:   log likelihood = -44.809855
Iteration 5:   log likelihood = -44.809855

Logistic regression                               Number of obs =     242
                                                  LR chi2(2)    =    5.87
                                                  Prob > chi2   =  0.0531
Log likelihood = -44.809855                       Pseudo R2     =  0.0615
```

Mortality	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
STAT2020_score	1.261616	.697345	1.81	0.070	-.1051552	2.628387
Age	-.1237151	.123521	-1.00	0.317	-.3658119	.1183817
_cons	-3.257578	.5848614	-5.57	0.000	-4.403885	-2.111271

Figure 9: Stepwise backward univariate analysis regression excluding morbidity

To demonstrate the logistic regression analysis, a saturated model (model with all variables: morbidity, mortality risk score and age, predicted mortality) was developed. Using a flexible *p*-value of 0.25, the non-predictive variables were dropped out. This is often called a univariate logistic regression analysis.[10]

An important first observation of the model in Figure 8 is that it is not a statistically significant model. The univariate analysis principle, as discussed in Sperandei, would suggest that variable age be dropped out first.[10] Age has the highest *p*-value of 0.35, which is greater than the recommended 0.25 suggested by Sperandei.[10] The *p*-value of 0.25 is not a rigid selection, as the intention of the regression at this stage is not to predict relationships but rather to select candidate predictors.[10] Based on this notion, age was kept in the model at the first backward dropping of predictors. Morbidity was the initially excluded predictor as in Figure 9.

Figure 9 still reflects an insignificant model, but the *p*-value has improved to what some researchers might term marginally significant. In this paper, even if the *p*-value is above 0.05 by a small fraction of a number, it was considered insignificant. The Figure 9 model still has age as highly insignificant (*p*-value of 0.32). Therefore, in the second iteration of model building, in the direction that removes insignificant variables, age was omitted in the next model.

The iterations in Figure 10 (4 iterations) indicate how rapidly the modelling process converged. The model had a log-likelihood of -45.60, which is a measure of the goodness-of-fit (GOF). It can range from negative infinity to positive infinity. The bigger the log-likelihood of a model is, the better the model. However, it is often not a used indicator, unless it is used to help compare nested models.

The model without the two highly insignificant variables (age and morbidity), is a significant model. The likelihood ratio chi-square has a significant *p*-value. This indicates that adding the variable surgical risk of mortality score made the model predict mortality better than the null model (a model with just the intercept). However, it is possible that in practice a reduced model may exclude some variables that have clinical relevance. Therefore, the determination to identify a model as final should always weigh in the clinical relevance of the factors studied for the outcome of interest. For the current paper, the model in Figure 10 was accepted as the final logistic regression model. Figure 11 represents the final model with odds ratios. To have the confidence that the selected model is indeed the best, a LR test with a null hypothesis can be computed to establish if the models are statistically different (significantly different based on the null hypothesis that the full model is the same as the reduced model).

```
. logit  Mortality  STAT2020_score

Iteration 0:   log likelihood = -47.847097
Iteration 1:   log likelihood = -46.597586
Iteration 2:   log likelihood = -45.577526
Iteration 3:   log likelihood = -45.576949
Iteration 4:   log likelihood = -45.576949

Logistic regression                                     Number of obs =    244
                                                        LR chi2(1)    =   4.54
                                                        Prob > chi2   = 0.0331
Log likelihood = -45.576949                             Pseudo R2     = 0.0474

------------------------------------------------------------------------------
   Mortality | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
STAT2020_score |   1.514173   .6632878     2.28   0.022     .2141525    2.814193
       _cons |  -3.677442   .4858115    -7.57   0.000    -4.629615   -2.725269
------------------------------------------------------------------------------
```

Figure 10: Final significant model, presented with the coefficients that need to be exponentiated to get the odds ratios

```
. logit, or

Logistic regression                                     Number of obs =    244
                                                        LR chi2(1)    =   4.54
                                                        Prob > chi2   = 0.0331
Log likelihood = -45.576949                             Pseudo R2     = 0.0474

------------------------------------------------------------------------------
   Mortality | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
STAT2020_score |   4.545659   3.01508     2.28   0.022     1.238812    16.67971
       _cons |   .0252876   .012285    -7.57   0.000     .0097585    .0655285
------------------------------------------------------------------------------
Note:  cons estimates baseline odds.
```

Figure 11: Final model with odds ratios and their 95% confidence interval (CI)

For the model in Figure 11, for every one-unit change in the STAT 2020 mortality risk score, the log odds of mortality (versus survival) increased by 4.55. The final model equation is:

$$\log(p/1 - p) = 0.25 + 4.545659*\text{STAT2020score}$$

To test the quality of the model, there are a few techniques to establish the prediction error of the model. A useful technique is the training and testing approach, where the data can be subset with a bigger portion, often 70% used to develop the model in a process called training, and the rest of the data are used to validate the model in a process called testing.[13] Another method is the use of the area under the receiver operating characteristic curve (ROC curve). Additional methods include computation of a confusion matrix, using GOF tests (Pearson GOF and the Hosmer-Lemeshow GOF tests), information criteria statistics, which are also a form of GOF tests, and analysis of residuals following the model command.[13]

The vertical line in Figure 12 can be taken as a model that has no predictive value. A model with a high predictive value has an area under the ROC curve that approximates or goes closer to 1. The current model has an area under the ROC curve of 0.71. A ROC curve of 0.72 is often described as representative of a strong model, and the current model nearly achieved that level of predictive value.

Figure 13 represents the result of a Pearson GOF test. The null hypothesis is that there is no difference in the number of observed and predicted successes in each group of the outcome. The *p*-value is significant and therefore we can reject the null hypothesis. This test statistic corrodes the confidence that regression has achieved good agreement between the model predictions and the observed mortality outcomes.

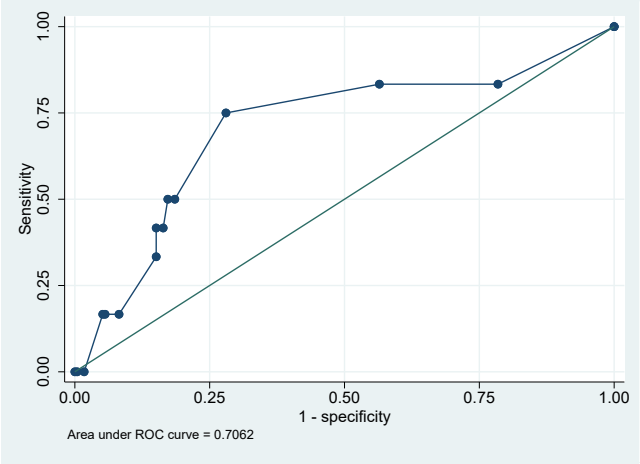


Figure 12: ROC curve for the final logistic regression model

```
. estat gof

Goodness-of-fit test after logistic model
Variable: Mortality

      Number of observations =    244
Number of covariate patterns =     14
           Pearson chi2(12) =  24.97
               Prob > chi2 = 0.0150
```

Figure 13: A GOF test statistic with a significant *p*-value

```
. estat class

Logistic model for Mortality

                 True
Classified |         D          ~D |      Total
-----------+----------------------+-----------
         + |         0           0 |          0
         - |        12         232 |        244
-----------+----------------------+-----------
     Total |        12         232 |        244

Classified + if predicted Pr(D) >= .5
True D defined as Mortality != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   0.00%
Specificity                     Pr( -|~D)  100.00%
Positive predictive value       Pr( D| +)      .%
Negative predictive value       Pr(~D| -)  95.08%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   0.00%
False - rate for true D         Pr( -| D)  100.00%
False + rate for classified +   Pr(~D| +)      .%
False - rate for classified -   Pr( D| -)   4.92%
--------------------------------------------------
Correctly classified                       95.08%
--------------------------------------------------
```

Figure 14: A confusion matrix for the logistic regression model

Although a confusion matrix is one of the most popular post-regression tests, the results can be affected by the proportion of the outcome in the sample or the positive cases.[14] This problem was a reality for the current model developed using data that was computer generated and might not represent real-life patient data. However, the data worked well to demonstrate the concepts. This limitation can be addressed by earlier steps of considering the representation of the outcome in the sampling techniques and calculations. Figure 14 reports the indicators in the confusion matrix.

Whether the data types favour linear regression analysis or logistic regression analysis, the process of data collection, data preprocessing, and finally deciding on the correct regression analysis to be performed require diligence and careful understanding of the statistical assumptions associated with the chosen analysis.

Conclusion

Regression analysis is a powerful tool that enables researchers and analysts to uncover relationships, identify trends, and make predictions. If adequate data is used correctly, studied relationships using regression analysis can provide insights about the phenomena in question. Simple linear and multiple linear regression can be used successfully to understand health data that is continuous, but when data of interest has dependent variables that are binary or categorical, a linear regression analysis will be incorrect. The logistic regression can adequately analyse data where the outcome is binary. It is important to perform post-model testing to ensure the quality of the regression model. This quality assurance will enhance the usefulness of the insights derived from regression analysis and modelling.

ORCID

MM Kebalepile https://orcid.org/0000-0002-5346-5798
PM Chakane https://orcid.org/0000-0001-9990-6336

References

1. Arnab R. Survey sampling theory and applications. Academic Press; 2017. p. 673-89. https://doi.org/10.1016/B978-0-12-811848-1.00020-0.
2. Olsson U. Generalized linear models: an applied approach. Professional Pub Service; 2002. p. 18.
3. Kumar S, Bhatnagar V. A review of regression models in machine learning. J Intell Inf Syst. 2022;3(1):40-7.
4. Marill KA. Advanced statistics: linear regression, part ii: multiple linear regression. Acad Emerg Med. 2004;11(1):94-102. https://doi.org/10.1197/j.aem.2003.09.006.
5. Gujarati DN. Basic econometrics. 4th ed. New Delhi: Tata McGraw-Hill; 2004.
6. Jacobs ML, Jacobs JP, Thibault D, et al. Updating an empirically based tool for analyzing congenital heart surgery mortality. World J Pediatr Congenit Heart Surg. 2021;12(2):246-81. https://doi.org/10.1177/2150135121991528.
7. Kebalepile MM, Chakane PM. Commonly used statistical tests and their application. South Afr J Anaesth Analg. 2022:S80-S4.
8. Howson C. The curious case of Frank Ramsey's proof of the multiplication rule of probability. Analysis. 2018;78(3):431-9. https://doi.org/10.1093/analys/anx161.
9. Pandis N. Logistic regression: part 1. Am J Orthod Dentofacial Orthop. 2017;151(4):824-5. https://doi.org/10.1016/j.ajodo.2017.01.017.
10. Sperandei S. Understanding logistic regression analysis. Biochem Med (Zagreb). 2014;24(1):12-8. https://doi.org/10.11613/BM.2014.003.
11. Bujang MA, Sa'at N, Sidik TMITAB, Joo LC. Sample size guidelines for logistic regression from observational studies with large population: emphasis on the accuracy between statistics and parameters based on real life clinical data. Malays J Med Sci. 2018;25(4):122-30. https://doi.org/10.21315/mjms2018.25.4.12.
12. Van Smeden M, Moons KG, de Groot JA, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. Stat Methods Med Res. 2019;28(8):2455-74. https://doi.org/10.1177/0962280218784726.
13. Dankers FJWM, Traverso A, Wee L, van Kuijk SMJ. Prediction modeling methodology. Springer International Publishing; 2019. p. 101-20. https://doi.org/10.1007/978-3-319-99713-1_8.
14. Fielding AH, Bell JF. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ Conserv. 1997;24(1):38-49. https://doi.org/10.1017/S0376892997000088.