Prediction and Analysis of Liver Patient Data Using Linear Regression Technique

Deepankar Garg and Akhilesh Kumar Sharma

Abstract In the current scenario, it is very difficult for the doctors to diagnose liver patient and there should be some kind of automated support based on machine intelligence that can help to diagnose in advance so that doctors start the treatment faster and save time. The machine intelligence is a way to predict the liver-related problems; in this study, the linear regression is used to predict the same, more accurately. The albumin levels are highly related in diagnosing these kinds of liver problems. The proposed model worked efficiently on 583 observations provided as well as on new datasets. The total average accuracy achieved in this proposed model was 89.34% which is much more than the previously identified research work of Wold et al. (SIAM J Sci Stat Comput, 5(3), 735–743, 1984, [1]) of 84.22%.

Keywords Liver patient · Albumin · Regression data mining model R-Miner

1 Introduction

India is a country with a population of 1.2 billion. There are just around 1 million certified doctors to attend to these people. Hence, we tried implementing a multiple linear regression algorithm, to predict the 'Albumin' level of a patient based on some significant variables. The original dataset provided consisted of 583 observations with 11 variables. Prediction of albumin level will help the doctors save time, and they need not run test, and the patients would save money as well. It would also enable you to diagnose the patient faster and with much more appropriate measures.

D. Garg (✉)
Department of Computer Science & Engineering, Manipal University, Jaipur, India
e-mail: deepankar.garg5@gmail.com

A. K. Sharma
Department of Information Technology, Manipal University, Jaipur, India
e-mail: akhileshshm@gmail.com

D. Reddy Edla et al. (eds.), *Advances in Machine Learning and Data Science*,
Advances in Intelligent Systems and Computing 705,
https://doi.org/10.1007/978-981-10-8569-7_8

In the model designed, 'Albumin' is considered as the dependent/outcome variable and 'Age', 'Gender', 'Total Bilirubin', and 'A/G Ratio' as the independent/observational variables. The independent variables are highly significant and are independent of each other. They are not correlated among themselves. Finally, a test model has been designed using multiple linear regression which predicts the albumin level with 90% accuracy, based on test dataset as well.

2 Related Work

As proposed by Tomohiro et al., the hypoalbuminemia separately predicts cardiac morbidity, and also mortality, for various kinds of chronic kidney disease-related patients by CRT. These kinds of observations are helpful in identifying the albumin levels, by the information of long-term prognosis in chronic kidney disease-related patients who willingly undergo CR [2]. Multiple linear regression was used [3] to implement our algorithm. According to Mark Tranmer et al., a multiple linear regression analysis is used to estimate the values of outcome variable, Y, provided with a set of p explanatory variables (x1, x2, …, xp). Mu-Jung Huang et al. adopts data cleaning and analysis methods [4] to explore meaningful guidelines from health-related data and employs case-based reasoning that always favors the highly severe diseases diagnosis and their treatments and working on these processes for more efficient working system. Swati Gupta et al. provided with introduction to linear regression algorithm [5] and explained its implementation. The researcher concentrated on the formulation and test data for linear regression. According to Dimitris Bertsimas et al., linear regression model is considered with response vector, model matrix, regression coefficients, and errors [6]. The linear regression models the relationship between a dependent variable and explanatory variables. Linear discriminant analysis and logistic regression are the most widely used statistical methods [7] for analyzing the numerical (or categorical) outcome variables. Logistic regression is useful when the dependent variable is of binary outcome.

According to David Broadhurst et al., variables in a linear regression can be selected [8] using backward or forward selection methods. These methods of elimination will help in building a linear model. Various logics were used to differentiate diffuse liver disorders, to categorize liver disorders under healthy and unhealthy liver patients, to diagnose hepatitis, and to perform the necessary operation required [9–15].

3 Methodology

The multiple linear regression algorithm is used, on the basis of which this model has been developed. In this study, the dependent/outcome variable used is 'Albumin'. And independent/observation variables are 'Age', 'Gender', 'Total Bilirubin', 'Direct Bilirubin', 'Alkphos', 'SGOT', 'SGPT', 'Total Proteins', and 'A/G Ratio'.

The correlation between independent variables was found out. It has been observed that 'Direct Bilirubin' and 'Total Bilirubin' were highly correlated. Hence, 'Direct Bilirubin' was eliminated from the model (in preprocessing phase). In this study, the first regression model was formed including uncorrelated independent variables. This model helped to classify the variables as 'significant' and 'nonsignificant'. All the significant variables were included in the next regression model, and the nonsignificant variables were removed. The final regression model consisted of just the significant and uncorrelated independent variables (Fig. 1).

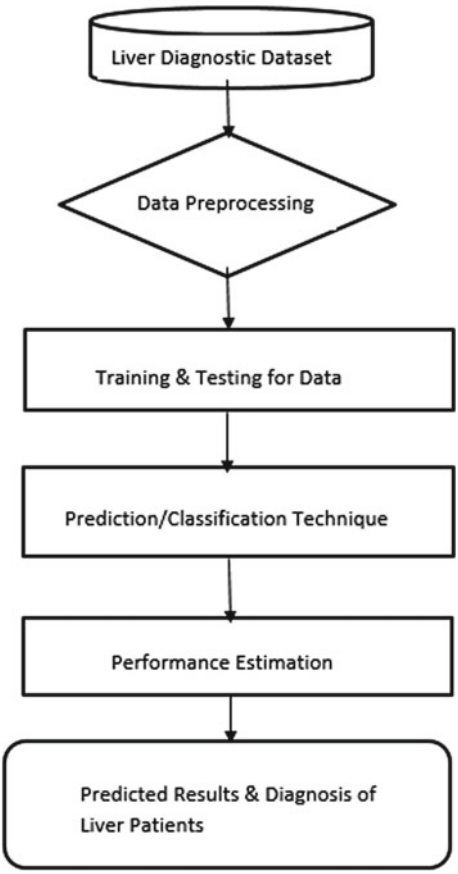

Fig. 1 Proposed model

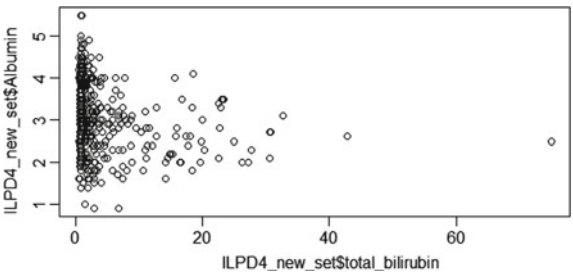

Fig. 2 Plot bet. albumin and total bilirubin

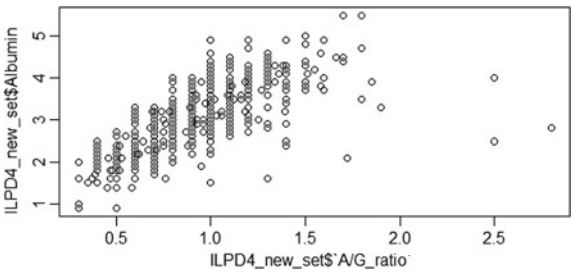

Fig. 3 Plot bet. albumin and A/G ratio

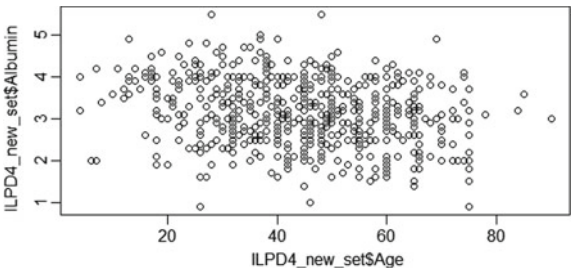

Fig. 4 Plot bet. albumin and age

In the next phase of this study, relation between dependent and independent variables was established (as shown in Figs. 2, 3, and 4). It was found that dependent variable was highly correlated with each of the independent variable.

To check for near normal residuals, a histogram (Fig. 5) and a qqplot (Fig. 6) were plotted. Histogram elaborates the normal residual values about 0 and the qqplot displays a linear graph, both of which satisfy the near normal residual condition.

In order to check for constant variability (i.e., whether the errors remain constant throughout), a graph was plotted between residuals and the fitted values as shown in Fig. 7. The graph seems to be constant in a particular region.

Lastly, the checking was conducted to find the existence of a pattern on x-axis (Fig. 8). There was no such pattern observed. Data was uniformly distributed. Finally, the testing of model was performed. It has been tried to predict the value of 'Albumin' on existing (training) as well as new (test) dataset. The model predicts

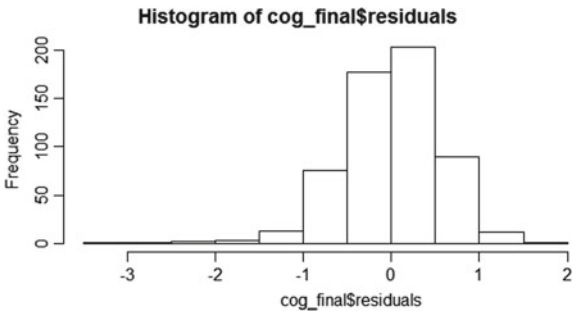

Fig. 5 Histogram plot of residuals

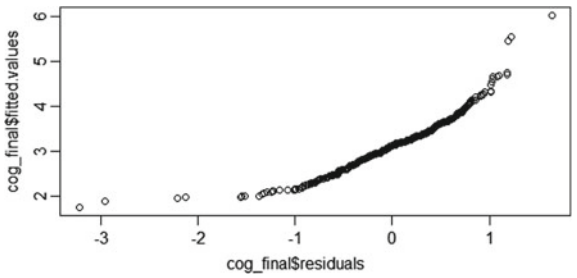

Fig. 6 qqplot between residuals and fitted values

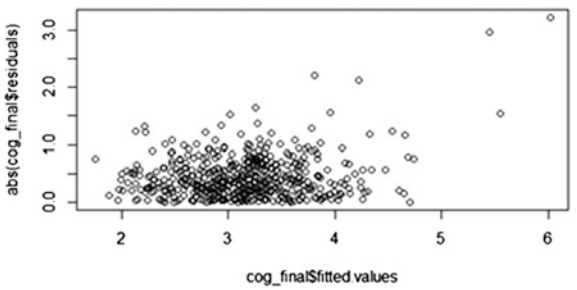

Fig. 7 Plot bet. absolute residuals and fitted values

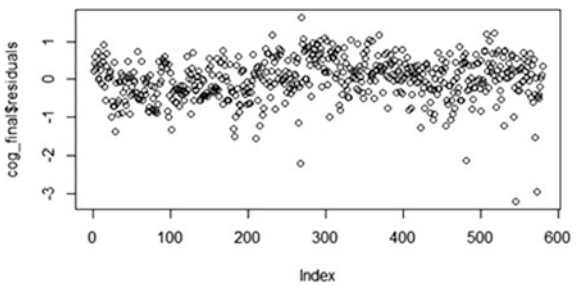

Fig. 8 Plot of residual values

the 'Albumin' level with an average accuracy of 89.353% which is a significant amount of accuracy this study observed and obtained.

3.1 Mathematical Formulation

The equation for linear regression is:

$$y = \alpha + \beta X$$

X and *y* are the two variables involved in the equation. The equation that describes how *y* is related to *x* is known as the **regression model** [13, 14]. α is the *y* intercept of the regression line, and *β* is the slope. The equation that describes how *y* is related to *x* is known as the **regression model** [16, 17].

There are three types of relationships in a regression line—first is positive linear relationship, second is a negative linear relationship, and third is no relationship [18–20].

Y is known as the outcome or dependent variable as a linear function [5] of another variable X also known as independent variable that is represented by the equation

$$Y = \alpha + \beta X \tag{1}$$

Here, the regression coefficients which are represented as α and β are given by

$$\beta = \Sigma(xi - x)(yi - y)/\Sigma(xi - x) \tag{2}$$

$$\alpha = y - \beta x \tag{3}$$

The mentioned values of regression coefficients of α and β can be computed in Eqs. (2) and (3) which are updated and putted in Eq. (1) to observe relationship among the independent variables and the dependent variable.

The linear regression technique works as following algorithm:

Step 1: Take the values of variable Xi and Yi
Step 2: Calculate average of variable Xi such that average is x = (X1 + X2 + · · · + Xi)/Xi
Step 3: Calculate average for variable Yi such that average is y = (Y1 + Y2 + · · · + Yi)/Yi
Step 4: Obtain the value of regression coefficient β by substituting the values of Xi, Yi, average of Xi, and average of Yi in the Eq. 2
Step 5: Compute the value of another regression coefficients α by substituting the values of β (calculated in step 4), average of Xi, and average of Yi in the Eq. (3).
Step 6: At last, update the value of regression coefficients α and β in the equation Y = α + βX

1. **Relationship Between Dependent and Independent Variables: Linear Relation**

Figure 2 shows plot between 'Albumin' at y-axis and 'Total Bilirubin' at x-axis. It shows a linear relation between the two variables.

Figure 3 shows plot between 'Albumin' at y-axis and 'A/G Ratio' at x-axis. It shows a linear relation between the two variables.

2. **Nearly Normal Residuals: Condition Met**

The histogram of the residual as shown in Fig. 5 is taken to prove that variance is normally distributed. The histogram is evenly distributed about zero which indicates that it is normally distributed. If it were not normally distributed, it would mean that the model's assumptions may have been contradicted.

Relationship Between Dependent and Independent Variables: Linear Relation.

3.2 Database Terminology and Their Description

For dataset, the bilirubin test conducted to find any increased level in the blood. It can be used to determine the cause of jaundice or diagnose other liver diseases, hemolytic type anemia, the blockage of bile duct, etc. Direct bilirubin that moves much freely in the blood is known as conjugated bilirubin. Alkphos estimates the quantity of alkaline phosphate enzyme in anybody's bloodstream.

Increased level SGPT suggests medical problems such as viral hepatitis, diabetes, congestive heart failure, liver damage. This is a common way of screening for liver problems.

AST or aspartate aminotransferase is single or two liver enzymes. It is always found as serum glutamic oxaloacetic transaminase type or SGOT. The AST is a protein, i.e., made by liver cells. And when liver cells are damaged, AST leaks into the bloodstream and the level of the AST in the blood becomes more than basic/normal. An estimated serum protein test measures the total amount of protein in blood. It measures the amounts for two major groups of proteins in the blood: albumin and the globulin.

A/G ratio is the comparison between the amounts of albumin with those of globulin. This test is useful when your healthcare provider suspects you have liver disease. Albumin is a plasma (blood) protein. If it is low, it could be due to decreased synthesis by liver indicating liver problem.

Table 1 Table for accuracy and predicted albumin level

Levels	Total bilirubin	A/G ratio	Given albumin	Predicted albumin	Accuracy (%)
Level 1	0.8	0.8	3.7	3.03	81.89
Level 2	0.6	1.1	2.6	3.34	77.84
Level 3	0.8	1.1	3.7	3.46	93.51
Level 4	0.9	0.9	3.9	3.15	80.76
Level 5	9.4	0.8	2.8	2.79	99.64

4 Implementation and Result

Dataset collected in this work shows the different parameters that could affect a person's health, particularly the liver. All the variables are of numeric form, and a model is built with one independent/outcome variable and four dependent/observant variables. A multiple linear regression model is built which has an average accuracy of 89.34%. The model works well on new dataset as well.

Creating regression models # First regression model

model1 = lm(Albumin ~ Age + Gender + total_biliru-
bin + Alkphos + SGPT + SGOT + 'A/G_ratio', data = ILPD4_new_set summary (model1)

Removing insignificant variables from model1 # Final regression model created

model_final = lm(Albumin ~ Age + Gender + total_bilirubin + 'A/G_ratio', data = ILPD4_new_set)

summary(model_final) **# Testing the model**

Age = c(43), Gender = ("Male"), total_bilirubin = c(0.1), 'A/G_ratio' = c(1.2), new_patient_data <- data.frame(Age, Gender, total_bilirubin, 'A/G_ratio')

model_test <- lm(Albumin ~ Age + Gender + total_bilirubin + 'A/G_ratio', data = ILPD4_new_set) predict(model_test, newdata = new_patient_data)

Model Accuracy: Training and Test Datasets

The main parameters include 'Total Bilirubin' and 'A/G Ratio'. From these parameters, albumin level can be predicted, which will help in diagnosing the patient faster. The prediction has an average accuracy of 89.353%. Age and gender also play a significant role in determining the albumin level (Table 1).

Table 1 consists of total_bilirubin, a/g_ratio, and albumin, which were already given. The fifth column predicts the albumin level from the model built with an average accuracy of 86.72%.

5 Conclusion

A multiple linear regression model has been prepared, which has one dependent variable and four independent variables. Initial model was prepared using the backward elimination method. The variables were eliminated based on their significance level. Least significant variables were removed from the model. Highly correlated independent variables were excluded as well. A final linear model was prepared where the dependent variable was linearly related to each of the independent variable and the model consisted of just the significant, nonlinear independent variables.

The model prepared predicts with an average accuracy of 89.34%. It predicts the level of albumin in a patient, which is highly dependent on the four independent variables defined in the model. Albumin level will help the doctor think in the right direction and help diagnose the patient faster. Usually, it is assumed that significant changes in the albumin level indicate that the patient is probably having problems in his lever. The model made works well on new dataset as well.

References

1. Wold, S., et al.: The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM J. Sci. Stat. Comput. **5**(3), 735–743 (1984)
2. Uchikawa, T., et al.: Serum albumin levels predict clinical outcomes in chronic kidney disease (CKD) patients undergoing cardiac resynchronization therapy. Intern. Med. **53**(6), 555–561 (2014)
3. Aiken, L.S., Stephen G.W., Steven, C.P.: Multiple linear regression. Handbook of psychology (2003)
4. Huang, M.-J., Chen, M.-Y., Lee, S.-C.: Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. Expert Syst. Appl. **32**(3), 856–867 (2007)
5. Gupta, S.: Int. J. Comput. Appl. **116**(9), 0975–8887 (2015)
6. Bertsimas, D., King, A.: OR forum—an algorithmic approach to linear regression. Oper. Res. **64**(1), 2–16 (2016). https://doi.org/10.1287/opre.2015.1436
7. Metodološki zvezki, vol. 1, No. 1, pp. 143–161 (2004)
8. Broadhurst, D., et al.: Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry. Anal. Chim. Acta **348**(1–3), 71–86 (1997)
9. Badawi, A.M., Derbala, A.S., Youssef, A.B.M.: Fuzzy logic algorithm for quantitative tissue characterization of diffuse liver diseases from ultrasound images. Int. J. Med. Inform. **55**, 135–147 (1999)
10. Gadaras, I., Mikhailov, L.: An interpretable fuzzy rule-based classification methodology for medical diagnosis. Artif. Intell. Med. **47**, 25–41 (2009). Luukka, P.: Fuzzy beans in classification. Expert Syst. Appl. **38**, 4798–4801 (2011)
11. Ming, L.K., Kiong, L.C., Soong, L.W.: Autonomous and deterministic supervised fuzzy clustering with data imputation capabilities. Appl. Soft Comput. **11**, 1117–1125 (2011)
12. Neshat, M., Yaghobi, M., Naghibi, M.B., Esmaelzadeh, A.: Fuzzy expert system design for diagnosis of liver disorders. In: Proceedings of the 2008 International Symposium on Knowledge Acquisition and Modeling KAM 2008, pp. 252–256 (2008)

13. Breese, J.S., Heckerman, D., Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of UAI-1998: The Fourteenth Conference on Uncertainty in Artificial Intelligence
14. Burges, C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. **2**(2), 955–974 (1998)
15. CiteSeer: CiteSeer Scientific Digital Library (2002). http://www.citeseer.com
16. Singh A., et al.: Liver disorder diagnosis using linear, nonlinear and decision tree classification algorithms. IJET **8(**5) 2059–2069 (2016)
17. Qual Quant: Linear versus Logistic Regression vol. 43, pp. 59–74 (2009). https://doi.org/10.1007/s11135-007-9077-3
18. Fox, J.: Applied Regression Analysis, Linear Models, and Related Methods. Sage Publications, Thou-sand Oaks, CA (1997)
19. Robbins and Cotran's Pathologic Basis of Disease, 9th edn.
20. Hocking, Ronald R.: A biometrics invited paper. The analysis and selection of variables in linear regression. Biometrics **32**(1), 1–49 (1976)
21. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. Wiley (1973)