Analyze important factors affecting total protein levels

Tingjia Lu 1005915306

2024-06-16

Contents

Abstract

The aim of this study was to identify the important factors affecting total protein levels through regression analysis of clinical data. Initial exploratory data analysis (EDA) and regression diagnostics identified key influencing factors including age, albumin, direct bilirubin, alanine aminotransferase and aspartate aminotransferase. The regression model constructed using these variables had strong predictive power with an adjusted R^2 of approximately 0.83. The findings highlight the critical role of these factors in determining total protein levels and emphasize the validity of regression analysis in the assessment of health data.

1. Introduction

Protein plays a variety of important roles in the body, including building and repairing tissues, acting as an enzyme to facilitate metabolic reactions, enhancing immune function, transporting and storing key molecules, regulating hormones, providing energy, maintaining acid-base and balancing fluids, and providing structural support. It is an essential component of muscle, skin, enzymes, antibodies, and hormones, and is critical for growth, repair, and proper functioning of daily body functions.

In this study, we used a dataset containing 416 records of patients with liver disease and 167 records of patients without liver disease collected from North East Andhra Pradesh, India. The "dataset" column is the category label used to categorize the group into patients with liver disease (hepatic) or non-hepatic (no disease). The dataset contains 441 male patient records and 142 female patient records.

The increased risk of IBD, in turn, results in the HFD diet probably due to increased intestinal permeability and the alteration of the intestinal microbiota (Mentella et al., 2020). It has been demonstrated that high-protein diets (HPDs) with low carbohydrates are a common strategy for weight loss (Blachier et al., 2019).

There are over 50,651+ reports in the University of Toronto's database of articles on the importance of high protein for the human body. Antibodies are also made up of proteins that play a key role in the immune system. Antibodies protect the body from infection by recognizing and neutralizing foreign invaders such as bacteria and viruses. So in many cases like cancer, inflammation, etc., there is an extremely strong correlation with PROTEIN LEVEL.

The aim of this study is to apply logistic regression analysis to explore the influence of various potential factors on protein levels, to identify which variables are statistically significantly correlated with high or low protein levels by combining a specific set of data, and to explore how optimization and evaluation of the model can improve the understanding of the strength of the influence of these variables, and to provide data support for the development of effective health promotion strategies.The main idea of this report is to go through the records of liver disease patients in India to find out the factors that can affect the protein levels, so that people can be more aware of their physical health.

2. method

The present study aimed to analyze the important factors affecting the total protein level and used a multi-step systematic approach. During the initial data exploration, we found missing values in the data and therefore used multiple interpolation (mice package) to process these missing values to ensure data integrity and reliability of the analysis.

In the model construction phase, we first constructed a full model with all potential predictor variables. Model assumptions and multicollinearity were checked through diagnostic tools such as residual plots and variance inflation factors (VIF) to ensure model plausibility and independence between predictor variables. Subsequently, we used stepwise regression to simplify the model by gradually removing non-significant predictor variables through AIC and BIC metrics for the purpose of balancing model fitting effectiveness and complexity.

To assess the performance of the models, we compared the adjusted R^2, AIC, BIC, and mean square error (MSE) on the test set for the full and simplified models. Adjusted R^2 measures the model's ability to explain data variability; AIC and BIC are used for model selection, balancing fitting effectiveness and model complexity; and MSE assesses the model's prediction error on the test set. The results show that the simplified model outperforms the full model on all these metrics, indicating better explanatory and predictive performance.

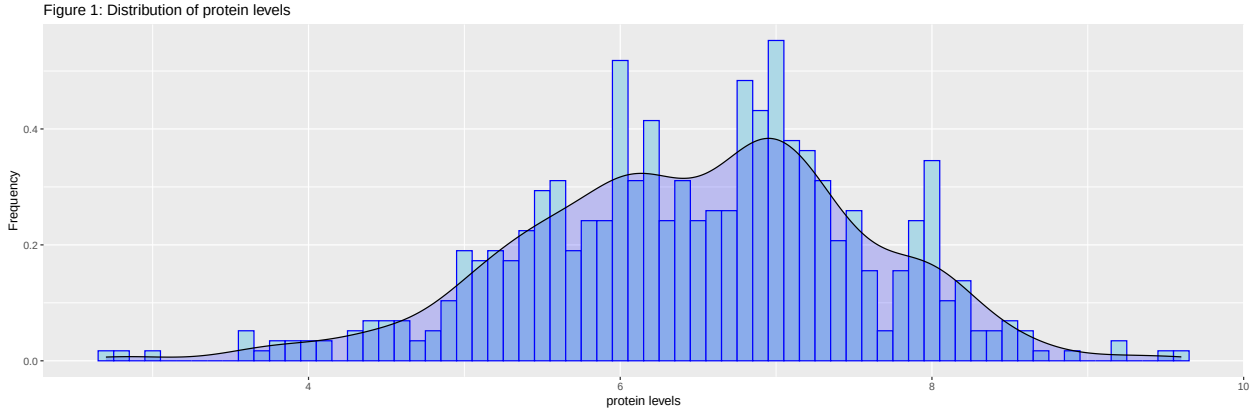
Finally, we use diagnostic plots (e.g., Dfbeta plots) to identify and deal with potentially influential observations to ensure the robustness and accuracy of the model. Through these steps, this study systematically identified the key factors affecting total protein levels and constructed a robust predictive model, which provides data support and theoretical basis for clinical diagnosis and health management.

3. Result

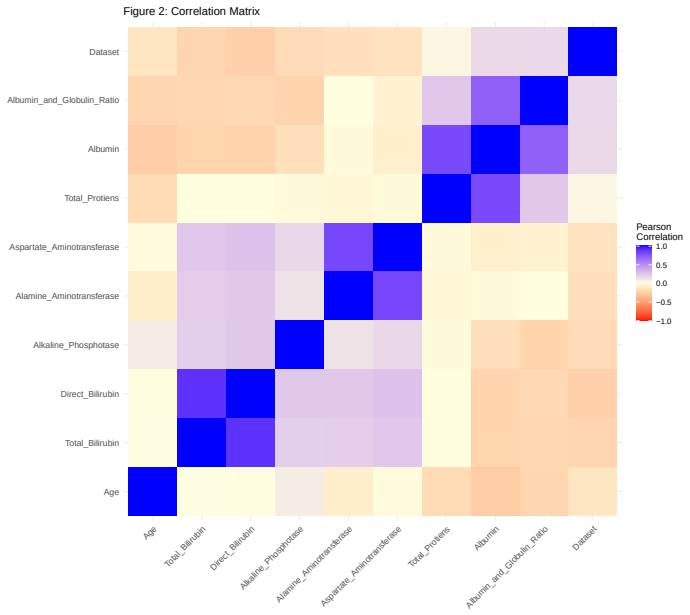
The dataset contains 583 observations and the variables include Age sex, gender, total bilirubin, direct bilirubin, alkaline phosphatase Total bilirubin Direct bilirubin Alkaline phosphatase alanine aminotransferase Aspartate aminotransferase Albumin Albumin-to-globulin ratio Initial exploration of the data revealed a number of missing values, and since the number of missing values was only four, we simply removed them and performed multiple estimation using the mice package..

In the raw data, we can observe that the PROTEIN LEVEL is close to conforming to the NORMAL DISTRIBUTION trend, so we don't have to make changes to the data. This figure demonstrates the distribution of protein levels in the dataset, showing a right-skewed distribution characteristic of a concentration in the range of 6 to 8, with two peaks near 6.8 and 7.5, indicating that protein values at these levels occur more frequently. The density estimation curve further shows the concentration trend and distribution pattern of the data.

Figure 1: Distribution of protein levels



Based on the heat map, we can see a strong positive correlation between albumin and total protein, albumin and albumin-globulin ratio, and direct bilirubin and total bilirubin. The correlations between the other variables are weak or no significant correlation. Subsequently, we will continue to use other ways to see if we should remove the variables with less correlation.



A fully linear regression model with all potential predictors was initially constructed. We performed preliminary diagnostics on the initial model. All VIF values in the initial model were <5 and there were no violations of linear regression and multicollinearity in the residual plots. We next chose to use stepwise regression to simplify the model by removing unimportant predictors. By using stepwise regression, the final model included direct bilirubin, alanine aminotransferase, aspartate aminotransferase, albumin, and albumin-to-globulin ratio as significant predictors. Together, these variables explained a large portion of the variability in total protein levels, with an adjusted R^2 of 0.83.

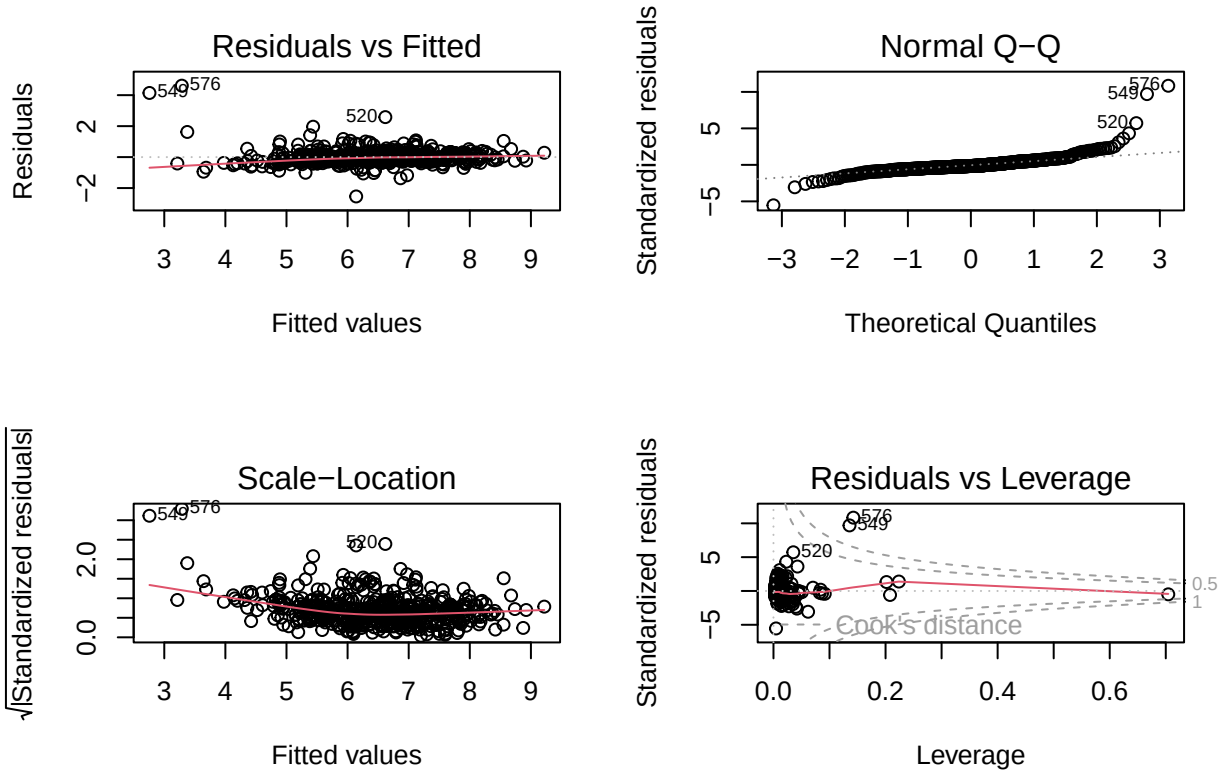
Through stepwise regression screening, we retained five variables, and the following table shows the data for the variables we screened: Table 1

—-	Estimate	Std. Error	t value	Pr(>
(Intercept)	3.0137049	0.1102299	27.340	0.0000000
direct_bilirubin	0.0644850	0.0084852	7.600	1.69e-13
Alamine_Aminotransferase	-0.0006888	0.0002089	-3.296	0.00106
Aspartate_Aminotransferase	0.0003170	0.0001160	2.734	0.00651
Albumin	1.6905475	0.0371037	45.563	<2e-16
Albumin and Globulin Ratio	-1.9168072	0.0936098	-20.477	<2e-16

We can see that the p value of each variable is less than 0.05, and we will next look at the other variables to see if they meet the criteria for linear regression. Regarding the value of vif, we can see that in the screened model, the value of vif is still <5 and the vif of each variable is significantly lower than that of the original model.

The graph demonstrates the relationship between protein levels and direct bilirubin, with the y-axis being the Dfbeta value, which indicates the effect of deleting a particular observation on the regression coefficients. The overall trend shows that the Dfbeta value fluctuates roughly around 0, indicating that the overall effect of direct bilirubin on the model is relatively stable. Points above or below the critical line in the graph may be observations that have a greater impact on the model.

Figure 3: The Residual Plots of the Simplified Model



The simplified model performs better than the full model in several indicators: the adjusted R^2 of the simplified model is 0.8291822, which is higher than the 0.8182361 of the full model, indicating better explanatory ability; its AIC value is 603.3118, which is lower than that of the full model's 609.8012, and the BIC value is 636.4481, which is lower than that of the full model's 659.5056, showing the more advantageous in model fitting effect and complexity balance; moreover, the simplified model has a mean square error of 0.2178628 on the test set, indicating better prediction performance. Therefore, the simplified model is a better choice.

4.Conclusion

The analysis identified direct bilirubin, alanine aminotransferase, aspartate aminotransferase, albumin, and the albumin-to-globulin ratio as key factors influencing total protein levels. The final regression model showed strong predictive power, making it a useful tool for assessing protein levels in the clinical setting. These findings highlight the importance of these biomarkers in managing liver health and protein metabolism.

5.Disscussion

This study highlights the complex interactions between various biochemical indicators and total protein levels. Although the predictive accuracy of the model was high, other potential influences were not included in the dataset, suggesting that further research is needed. Perhaps the inclusion of other variables such as dietary intake and genetic factors could provide a more comprehensive understanding of the determinants of protein levels.

6.Limitation

The conclusions of this study are limited by the cross-sectional nature of the data, which precludes causal interpretation. In addition, the dataset may not be fully representative of the wider population, and potential measurement error in the clinical data could also affect the results. Future studies using longitudinal data and more variables could improve the predictive power and generalizability of the model. In R^2, the difference between the simplified model and the original model is actually not that great, and the data from the original model is better in AIC. Maybe we can use more complex models later to analyze the factors that are more influential on protein levels.

7.Reference

Khorsand Zaker, B. S., Saghebjoo, M., & Islami, F. (2022). Effectiveness of high-intensity interval training and high-protein diet on TNF- protein level in colon tissue of obese male rats: The importance of diet modifying. Obesity Medicine, 31, 100403-. https://doi.org/10.1016/j.obmed.2022.100403

Nuttall FQ. Body Mass Index: Obesity, BMI, and Health: A Critical Review. Nutr Today. 2015 May;50(3):117-128. doi: 10.1097/NT.0000000000000092. Epub 2015 Apr 7. PMID: 27340299; PMCID: PMC4890841.

Hannah Ritchie and Max Roser (2017) - "Obesity" Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/obesity' [Online Resource]

7.2 Reference for Datasets

CC0 1.0 Deed | CC0 1.0 Universal | Creative Commons. (n.d.). https://creativecommons.org/publicdomain/zero/1.0/

8.Appendix

Figure 4

```
## 549 576
## 545 572
```