Data Analysis on Liver Disorders

Stats 101A

Author: Pinyi Li

Professor: Shirong Xu

Introduction

This project aims to investigate the relationship between blood test measurements and the amount of alcohol consumed per day. The dataset consists of seven variables, with the first five variables representing blood test results that are sensitive indicators of potential liver disorders associated with excessive alcohol consumption. The sixth variable, "drinks," serves as the dependent variable in the analysis.

Table 1 Variables Table

Variable name	Role	Type	Description
mcv	Feature	Continuous	mean corpuscular volume
alkphos	Feature	Continuous	alkaline phosphotase
sgpt	Feature	Continuous	alanine aminotransferase
sgot	Feature	Continuous	aspartate aminotransferase
gammagt	Feature	Continuous	gamma-glutamyl transpeptidase
drinks	Target	Continuous	number of half-pint equivalents of alcoholic beverages drunk per day
selector	Other	Categorical	field created by the BUPA researchers to split the data into train/test sets

The objective of this study is to examine the association between the response variable, drinking, and five blood test measures. The selector variable, which lacks a mathematical relationship with the “drinks” variable, is excluded from the analysis. Multiple linear regression is employed to establish the model. Diagnostic tools are utilized to assess the model's reliability. The mean, variance, and standard deviation of the variables are presented in Table 2.

Table 2 Mean, Variance, and Standard Deviation of the Response Variable and 5 Blood Test Measures

Variable Name	Mean	Variance	Standard Deviation
mcv	90.16	19.79	4.45
alkphos	69.87	336.64	18.35
sgpt	30.41	380.73	19.51
sgot	24.64	101.29	10.06
gammagt	38.28	1540.92	39.25
drinks	3.46	11.14	3.34

Linear Regression Model

To assess the linear relationship between the predictors and the response variable, multiple linear regression is employed to create a full model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$$

Where:

- Y represents the response variable (drinks)
- $x_1, x_2, x_3, x_4,$ and x_5 denote the predictors (mcv, alkphos, sgpt, sgot, and gammagt)
- $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4,$ and β_5 are the coefficients
- ϵ is the error term

To evaluate the linear relationship between the predictors and the response variable, scatter plots are utilized to visualize the pairwise relationships between all variables. This allows for an initial assessment of the potential linear associations between the predictors and the response variable.

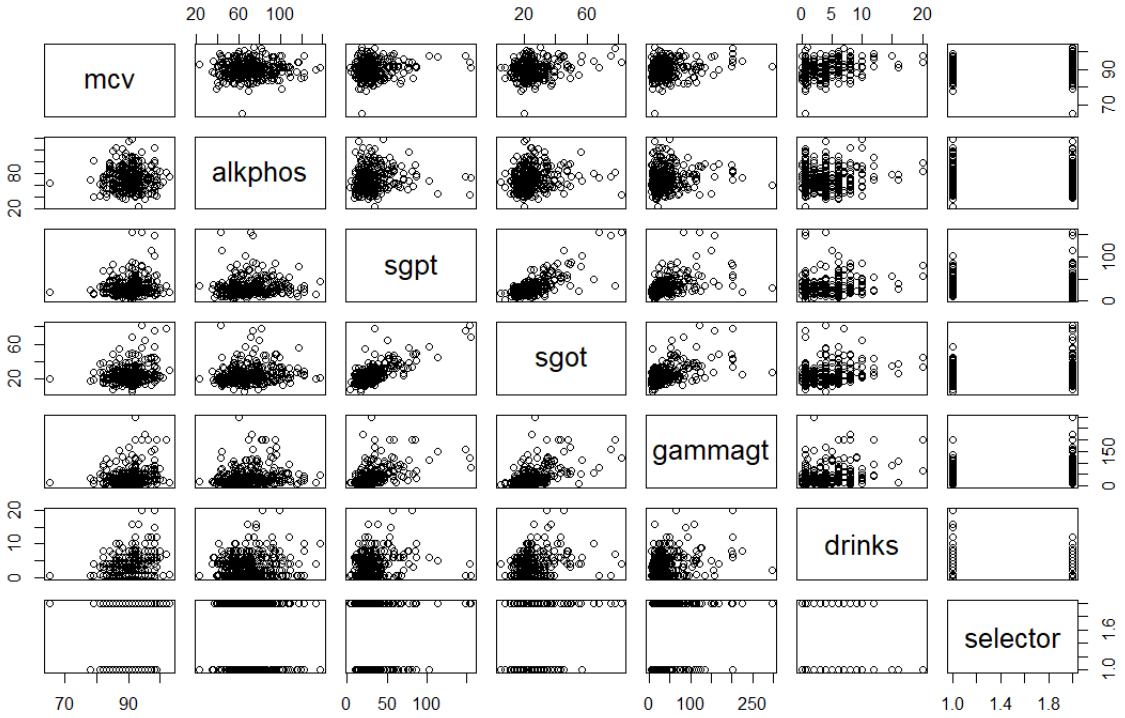

Figure 1 Pairwise Scatterplots Showing the Relationship Between Blood Test Predictors and Alcohol Consumption

The findings from Figure 1 suggest that there is a poor linear relationship between the dependent variables (mcv, alkphos, sgpt, sgot, and gammagt) and the response variable (drinks). This implies that the blood test measures may not be strongly correlated with the amount of alcohol consumed per day. Consequently, variable selection becomes necessary to identify and remove poorly correlated variables from the model. Since the dependent variables do not exhibit any pattern or relationship with the response variable, data transformation is considered unnecessary.

Table 3 presents the p-value of t-test for the coefficients in the full model. The model summary indicates that certain coefficients are not statistically significant, prompting the need for variable selection. By choosing a subset of predictors, the model's complexity can be reduced while retaining predictive power. The adjusted R^2 is 0.1763 suggests that the model explains approximately 17.63% of the variance in the response variable, indicating moderate predictive ability. Further refinement through variable selection may improve the model's performance and interpretability.

Table 3 p-value of t-Test for the Coefficients of Predictors for Full Model

Variable	Pr(> | t |)
mcv	2.91e-06
alkphos	0.400164
sgpt	0.420052
sgot	0.050208
gammagt	0.000107

Multicollinearity can lead to unstable and sensitive coefficients in regression models, hindering their interpretability and predictive reliability. To avoid multicollinearity, variance inflation factor (VIF) is calculated. Table 4 displays the VIF values for all predictors, with none exceeding 5, indicating low correlation among predictors. This implies that coefficients can be estimated with higher accuracy, enhancing the reliability of predictions and facilitating interpretation of the model.

Table 4 VIF of Predictors

Predictor	VIF
mcv	1.059993
alkphos	1.03042
sgpt	2.307708
sgot	2.419507
gammagt	1.483186

The Akaike Information Criterion (AIC) serves as a measure of goodness of fit, where lower values indicate better model fit. Using backward stepwise regression (AIC), the optimal subset model can be identified. The analysis reveals that the five-variable model is prone to overfitting, whereas the three-variable model (including mcv, sgot, and gammagt predictors) offers a better fit. Thus, the selected model can be represented as $Y = \beta_0 + \beta_1 x_1 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$. Subsequently, the coefficients are estimated through linear regression, resulting in the multiple linear regression model $Y = -14.5 + 0.18x_1 + 0.038x_4 + 0.019x_5$. Table 5 presents the p-values of the t-tests for the coefficients in the reduced model, indicating statistical significance for both the overall model and individual coefficients (all p-values < 0.05). With a p-value for F statistics of 4.753e-15, the reduced model is also deemed statistically significant. Additionally, the adjusted R-squared value of 0.1776 for the reduced model surpasses that of the full model, indicating improved fitness.

Table 5 p-value of t-Test for the Coefficients of Predictors of Reduced Model

Variable	Pr(> | t |)
mcv	2.59e-06
sgot	0.048357
gammagt	0.000113

Model Diagnosis

Multiple linear regression relies on several key assumptions, including linearity, normality, homoscedasticity, and absence of multicollinearity. Multicollinearity was assessed previously using VIF, indicating low correlation between predictors.

To evaluate the normality assumption, a Normal Q-Q plot is employed. As shown in Figure 2, the plot exhibits a trend that deviates from a straight line, appearing right-skewed. This departure from the linear trend line suggests a violation of the normality assumption.

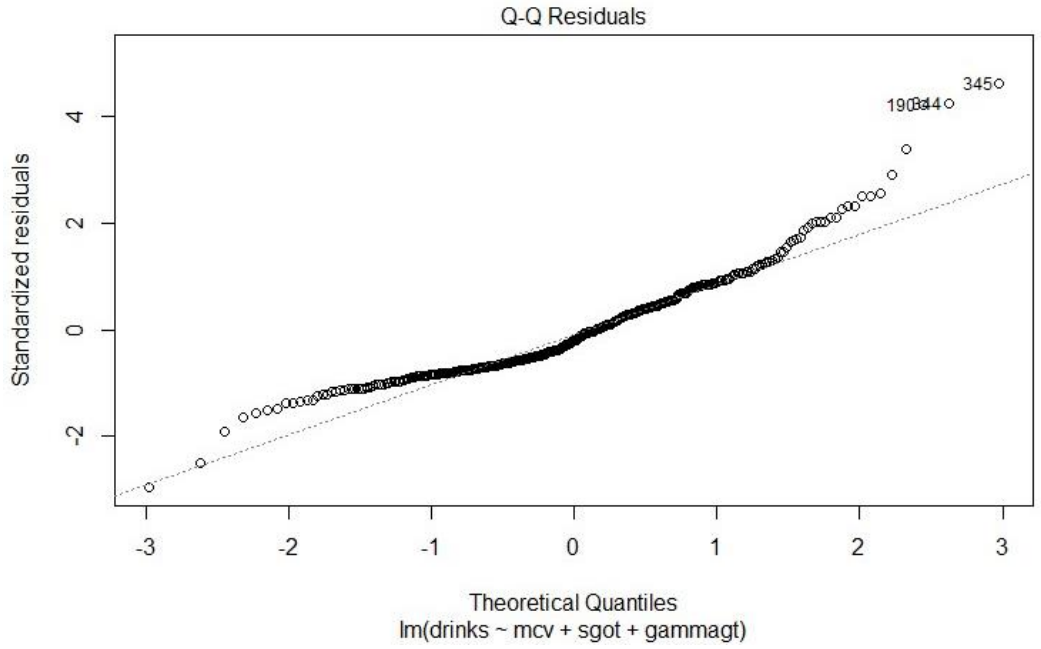

Figure 2 Normal Q-Q Plot for Multiple Linear Model

To assess the constant variance assumption, standardized residuals are plotted against four predictors. The average of residuals should form a horizontal line, indicating consistent variance across the range of predictor values. In Figure 3, no discernible pattern is observed in the distribution of residuals, suggesting that the assumption of constant variance is not violated. This implies that the variability of the residuals remains stable across different levels of the predictors, reinforcing the reliability of the regression model.

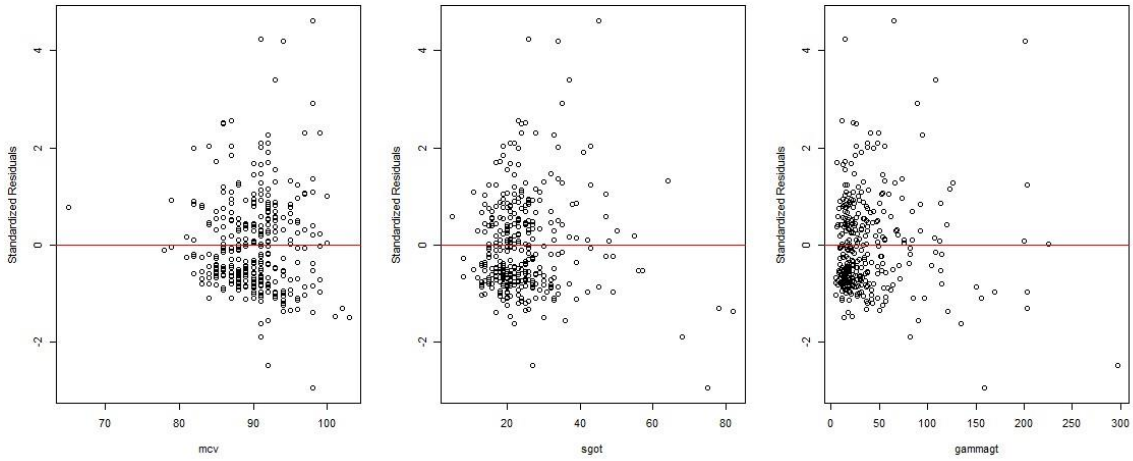

Figure 3 Standardized Residuals for Predictors

Outliers are data points that significantly deviate from the rest of the observations and can potentially affect the results of statistical analysis. In this case, a criterion of standardized residuals falling outside the range (-2, 2) is used to detect outliers. In the liver dataset with 345 observations, 18 outliers were identified based on this criterion. These outliers may represent extreme or unusual observations that merit further investigation. Further examination of these outliers may be warranted to determine their impact on the analysis and whether they should be retained or excluded from the dataset.

Summary

This project aimed to examine the relationship between alcohol consumption and blood test results using multiple linear regression. Initially, all five blood test predictors were assumed to have a linear relationship with the response variable. However, upon examining the full model summary, some predictor coefficients were found to be statistically insignificant.

To address this, backward stepwise regression was used for variable selection, resulting in a reduced model with three significant predictors. The model suggests that $Y = -14.5 + 0.18x_1 + 0.038x_4 + 0.019x_5$. The best-fitted model indicated that increases in mean corpuscular volume, aspartate aminotransferase, and gamma-glutamyl transpeptidase were associated with positive changes in alcohol consumption.

Diagnostic tools revealed violations of the normality assumptions, suggesting that the linear regression model may not be suitable for the dataset. Despite statistically significant regression output, the adjusted R^2 value remained low after variable selection, indicating poor model fit. Further investigation into potential non-linear relationships or additional influencing factors can be considered.