University of the Western Cape

Identifying Critical Blood Test Indicators for Liver Disorders through Multiple Linear Regression

A Report submitted in fulfilment of the requirements for the STA221 module

[GROUP 6]

By

Khanya Khosana

4353124

Minenhle Khuzwayo

4260321

Yoliswa Kokwana

4123238

Alyssa Jordan Krishna

4308998

Lunga Ntokozo Kubheka

4253243

Palesa Adjeley Le Tape

4395764

Supervisor:

Mr: Matthew Wayne Valentine

[12/10/2024]

Table of Contents

Main research Question

How can blood test results be used to predict excessive alcohol consumption in patients with liver disorders?

Proposed Research question

Does an increase in Gamma-glutamyl transferase (Gammagt) levels lead to an increase in the number of drinks, an indication in alcohol consumption?

Abstract

Excessive alcohol consumption is a leading factor in liver diseases. Biomarkers were established to predict diseases in patients. However, every biomarker was used in testing for multiple diseases. Understanding the relationship between alcohol consumption and blood test results is crucial, as liver disorders are one of the leading causes of morbidity and mortality worldwide. In this study, male participants were carefully studied, and it found that exceeding 21 drinks per week individuals are at risk for liver toxicity. While over 14 drinks per week alone are already considered risky. The reliability and accuracy of each blood test are still debated, necessitating further exploration.

This study seeks to answer the following key research questions: How can blood test results be used to predict excessive alcohol consumption in patients with liver disorders? Does an increase in Gamma-glutamyl transferase (Gammagt) levels lead to an increase in the number of drinks, an indication in alcohol consumption? Within the UCI BUPA liver disorder dataset, which blood test variables are most strongly associated with excessive alcohol consumption?

This study contributes to existing research by examining the relationship between alcohol consumption and liver disorders and also providing blood tests that are used in determining liver functioning. By using the UCI BUPA dataset, this research evaluates how diagnostic tests are used to predict excessive alcohol consumption.

Although diagnostic tests are unique, this research employs multiple linear regression, correlation, descriptive analysis and hypothesis testing to investigate the relationship between blood test results and alcohol consumption through male participants that recorded their results.

The findings of this study are expected to understand the value of diagnostic tests and how each test could be used to predict alcohol consumption. These relationships will assist physicians to better assess liver health and provide treatment plans.

Literature Review

The purpose of this research is to determine which of five blood tests namely mean corpuscular volume (Mcv), alkaline phosphatase (Alkphos), alanine aminotransferase (Sgpt), aspartate aminotransferase (Sgot) and gamma-glutamyl transferase (Gammagt) that are known to be sensitive to liver disorders to determine which blood tests are better predictors of liver disorders. Excessive alcohol consumption is one of the leading causes of liver damage and liver diseases alongside other factors such as lifestyle and diet. The literature review examines research articles that focus on the various causes of liver damage as well as similar studies that have researched the relationship between some of the aforementioned blood tests and their role in certain liver disease diagnosis and alcohol levels. Many assumptions have been made to justify the diagnosis and causes of liver disorders. Although there is a broad range of assumptions, this review will focus on blood tests that are examined on patients that consume alcohol in large volumes. Testing for alcohol hepatitis is linking a direct syndrome of liver failure to blood tests. These blood tests are biomarkers that serve as indicators for diagnosing and monitoring liver disease present in one’s blood alongside the influence of metabolic and genetic factors which is also carefully explored to make an accurate diagnosis. Throughout many studies, findings were that Mcv measures the average size of red blood cells and is elevated in alcohol consumption, Alkphos is an enzyme which indicates when there are elevated levels it signals liver dysfunction, Sgpt and Sgot are enzymes involved in metabolism and increased levels also indicate dysfunction, Gammagt is more specific to alcohol intake and is a key function in liver function experiments. Although many literature examine these factors, this literature will primarily focus on the role of Gammagt in diagnosing liver damage related to excessive alcohol consumption, examining its correlation with liver injury and discussing solely on this enzyme as an indicator of liver health. Quantifying variables such as drinks and selectors will be overviewed to validate and relate health risks to predictive models. By addressing these limitations, the review aims to provide a balanced view of blood tests' role in diagnosing liver disorders, helping clinicians and researchers understand its utility and constraints in the context of alcohol-related liver disease.

The literature on liver enzyme biomarkers highlights the role of various blood tests in diagnosing liver disorders. Pan et al. (2022) provides a comprehensive overview of liver enzymes, noting that alkaline phosphatase (Alkphos), alanine aminotransferase (Sgpt), and aspartate aminotransferase (Sgot) are significant markers of liver dysfunction. Alkaline phosphatase and alanine aminotransferase are particularly noted for their sensitivity to liver damage, being released into the bloodstream when liver cells are injured (Pan et al., 2022). Gamma-glutamyl transpeptidase (Gammagt) is also identified as an effective marker, especially for assessing alcohol-induced liver damage, though it is not specific to liver disease alone (Pan et al., 2022). These findings suggest that while individual tests are useful, their diagnostic power increases when considered in combination.

Chronologically, early research established the utility of specific liver enzymes such as ALT and AST. Pratt (2019) underscores that alanine aminotransferase (Sgpt) is the most specific marker for liver injury due to its high concentration in liver cells, whereas aspartate aminotransferase (Sgpt) is less specific because it is present in various tissues (Pratt, 2019, 1243). This specificity allows ALT to be a strong indicator of hepatocellular damage, though it does not necessarily reflect the severity of liver injury. Elevated levels of ALT and AST can signal liver cell damage, but distinguishing the extent of liver injury often requires additional tests (Pratt, 2019, 1243).

The literature also indicates that combining various blood tests enhances diagnostic accuracy. Pan et al. (2022) highlights the value of combining markers such as the Sgot/Sgpt ratio, which provides better detection of liver disorders compared to individual indicators. Additionally, the National Library of Medicine notes that the AST/ALT ratio greater than 2, alongside elevated Gammagt levels, is particularly indicative of alcoholic liver disease (National Library of Medicine, n.d.). This combination of tests captures different aspects of liver function and damage, offering a more comprehensive assessment.

Despite the advancements in liver enzyme testing, inconsistencies still remain. While ALT is recognized for its specificity, Pratt (2019) (Pratt, 2019, 1243) points out that other tests, like ALP and Gammagt, are less specific and can be influenced by non-liver-related conditions. For instance, elevated ALP might indicate cholestasis but can also result from bone disorders (Pratt, 2019, 1244). This necessitates careful interpretation of test results within the clinical context, as no single test is wholly sufficient for diagnosing liver disorders (Pratt, 2019, 1244). Furthermore, the role of mean corpuscular volume (Mcv) in liver function tests is less documented, suggesting a need for further research into its potential diagnostic value

In conclusion, determining liver function, relating to alcohol consumption, relies on blood tests that provide key insights on the malfunctioning of a liver. This review has highlighted the role of Gamma-Glutamyl Transferase (Gammagt) in detecting alcohol-induced liver damage, alongside mean corpuscular volume (Mcv), alkaline phosphatase (Alkphos), alanine aminotransferase (Sgpt), and aspartate aminotransferase (Sgot). While each test offers an identifiable factor in excessive alcohol consumption, Gammagt is known for its association with alcohol use.

However, the literature presents some inconsistencies in the interpretation of these markers, specifically regarding the Gammagt blood test. In some cases, elevated Gammagt levels have also been related to other diseases. This raises concerns about the reliance on Gammagt in diagnosing alcohol-related liver damage. Similarly, the Sgot/Sgpt, also does not always correlate consistently with different stages of patients liver disease.

Future research will focus on clarifying the accuracy of these biomarkers by exploring other factors. More comprehensive studies are needed to find the threshold levels of Gammagt and other markers that can reliably predict the difference between alcohol-related and non-alcohol-related liver conditions.

Finally, combining the Sgot/ALT ratio, elevated Gammagt, Mcv, Sgpt levels will create a strong indicator of alcohol-induced liver disease, but also emerge inconsistencies in further exploration. By addressing these gaps, further studies can help to develop a more clear diagnosis for liver disease.

Methodology

Introduction

Predicting liver disorders accurately is critical for improving early detection and effective treatment strategies, especially those related to excessive alcohol consumption. By identifying key blood test indicators, this study aims to contribute to more precise diagnostic tools, allowing physicians to intervene early and alter treatment plans based on reliable predictors.

Although existing research has explored blood tests for liver disease diagnosis, inconsistencies in methods and misinterpretation of data have led to mixed conclusions. Previously, studies using the UCI BUPA liver disorder dataset have incorrectly used a data-splitting variable as a diagnostic outcome, leading to flawed findings. This study seeks to correct these errors by focusing on particular variables to address the research question.

Overview of Methodologies and Rationale

In liver disorder research, methodologies such as logistic regression and non-parametric analysis will be used on SAS software to apply these analyses. However, multiple linear regression has been used for this study for its suitability for the continuous variable, drinks, analysis and its ability to model the relationship between alcohol consumption and blood test results. This offers a clear interpretation of how individual blood tests influence predicted variables (Mcv, Sgpt, Sgot and Alkphos) to detect alcohol consumption, making it an ideal approach for this research.

Key Methods and Approach

This study makes use a comprehensive set of statistical methods to address the research question:

- Descriptive analysis: to summarise the data distribution and central tendencies of each blood test variable.
- Correlation Analysis: to assess the relationship between blood test variables and alcohol consumption.
- Multiple Linear Regression: to model how the five blood test variables (Mcv, Alkphos, Sgpt, Sgot, Gammagt) predict alcohol consumption.
- Analysis of Variance (ANOVA): to test the significance of the regression models.
- Hypothesis testing: to evaluate the strength and significance of the relationships between the variables.

- Model Selection (Backward, Forward and Stepwise Regression): to refine the multiple regression model by selecting the most appropriate variables.
- R-squared and Coefficient of variation: to measure the model's accuracy.

Thesis statement

This study hypothesizes that gamma-glutamyl transferase (Gammagt) will be the strongest predictor of alcohol-related liver disorders, while multiple linear regression will provide the most reliable framework for modelling these relationships. By using regression techniques and model selection models, it aims to identify significant predictors for liver disorders for more accurate diagnosis.

Research Questions

1. Does an increase in Gamma-glutamyl transferase (Gammagt) levels lead to an increase in the number of drinks, an indication in alcohol consumption ?
2. Which blood test variables are the most significant predictors of liver disorders, particularly those associated with alcohol consumption?
3. Can multiple linear regression provide a reliable model to predict liver disorders based on these variables?

Methodology Overview

This methodology section will be covered as follows:

1. Descriptive Analysis: for the initial exploration and summary of the dataset
2. Correlation Analysis: examination of the relationships between variables.
3. Regression analysis and ANOVA: evaluation of the regression model.
4. Model Selection Techniques: application of backward, forward and stepwise methods to refine the model.
5. Model Interpretation: discussion of the results, R-squared and hypothesis testing outcomes.

1. Descriptive Analysis

This research seeks to understand the association of five blood tests with liver disorders with the aim of finding the best indicator for liver disorder. The dataset being investigated comprises various indicators which have a diverse range of statistics. The method of descriptive analysis gives a fundamental role of a basic understanding of the dataset. It gives an understanding of measures of central tendency i.e. mean, median, number of observations and the spread of the data. It additionally assesses the skewness of the data to evaluate any outliers as well as the distribution of the data.

This approach further enables the formation of hypotheses based on the chosen research question which wishes to investigate the most effective indicator for liver disorders. It will additionally highlight the importance of the 5-blood test being investigated in alignment with the primary focus of the research.

In conclusion, the above method will be used as a primary method of understanding the fundamentals of the dataset. It will also restore the effectiveness of assessing the significance of variables correctly with the aim to answer the research question at hand. By utilizing this method, this research aims to provide valuable insights which will further contribute to a deep understanding of which blood types should be used in identifying and diagnosing liver disorders.

2. Correlation Analysis

Correlation analysis is a statistical technique that evaluates the strength and direction of the linear relationship between two or more variables. This method is essential for understanding the extent of relationships between variables, predicting results, recognizing trends, and making informed decisions in research. Correlation analysis is frequently the first step in more advanced statistical analyses, like regression, and assists in determining the need for additional analysis.

Depending on the nature of the data, different correlation coefficients can be used. Pearson's Correlation Coefficient (r) is applied in cases when there is a linear relationship between variables, and both variables are continuous and normally distributed. This coefficient evaluates the magnitude and orientation of a linear correlation between two factors. Pearson's correlation can vary between –1 to +1, where +1 is a perfect positive linear relationship, 0 indicates that there is no relationship, and –1 is a perfect negative linear relationship. Spearman's Rank Correlation (ρ or rs) is utilized for non-parametric testing in cases where the data is either ordinal or not distributed normally. This coefficient evaluates the monotonic relationship between two variables, which indicates that one variable generally increases or decreases as the other one does, though it is not necessarily proportional. Pearson’s Correlation Coefficient is calculated using this formula: r=∑ (xi −x¯)2∑ (yi −y¯)2 ∑ (xi −x¯) (yi −y¯) , where xi and yi (Shaun Turney, 2024) represent individual data points,

x¯\bar{x}x¯ and y¯\bar{y}y¯ are the mean values of the respective variables. The steps taken is to first collect and organize the data, then confirm that the data satisfies the assumptions of linearity and normality, then determine the averages and calculate the deviations and finally calculate the correlation by multiplying the deviations of each pair, summing the products and then dividing the square root of the sum of squared deviations of each variable.

The Spearman Rank Correlation Coefficient is a measure of strength and direction of a relationship between two variables. First give rankings to the data points in each variable, then calculate the discrepancy in rankings by determining the variance between the ranks of relevant variables. The equation for calculating Spearman’s Correlation Coefficient is: ρ=1−n(n2−1)6∑di2 (Aryam Gupta, 2024), where di is the difference between the ranks of each pair of data points and n is the number of observations.

Though correlation analysis is strong, it does have restrictions. Casually, the existence of a relationship between two variables does not imply that one variable is the cause of changes in the other. Linearity, Pearson’s correlation only assesses linear relationships.

In conclusion, correlation analysis is a crucial method in research to determine the strength and direction of connections between continuous variables. Researchers use correlation coefficients to understand the relationship between variables informing future analysis and modelling endeavours.

3. Regression analysis and Anova

Multiple Regression to Identify Key Predictors of Alcohol Consumption

Multiple regression is a statistical tool that tries to find the relationship between one dependent variable and more than one independent variable. In such a way, it helps to predict the value of the dependent variable by using values of the independent variables. It basically fits a straight line equation through the data points, enabling analysts to assess changes in independent variables against the dependent variable. The model can be summarized as:

The Drinks= β0+β1∗Mcv + β2∗alkphos + β3∗Sgpt +β4* Sgot +β4* Gammagt +β5* Selector + ϵ. (Moone, McCabe, & Craig, 2017).

Because the model of regression analyzes more than one variable at a time, the model could identify which of the individual blood test results are the best predictors of heavy alcohol consumption. It may indicate, for instance, that liver enzymes have a close relation with increased intake of alcohol, while the rest of the biomarkers have a lesser impact. Clinicians will find such information useful in the identification of biomarkers on which they should place greater emphasis when assessing patients.

Study on Alcohol Consumption and Liver Enzymes: A comparison was drawn between alcohol consumption and liver enzymes, amongst others, in respect to the Gammagt and alcohol consumption. The authors identified that increased Gammagt levels were strongly related to increased alcohol intake, thereby proving that through regression analysis, predictive biomarkers can be identified clinically.

In the end, the multiple regression model effectively carries out the analysis of a dependent variable and several independent variables, hence making a prediction about alcohol consumption based on various blood test results. This model will highlight significant indicators with the aim of helping the clinicians focus on biomarkers to prioritize during their evaluation. It has also been established that higher Gammagt levels are linked to higher consumptions of alcohol, further reflecting the utility of this model in clinical settings to identify predictive biomarkers.

Analysis of Variance

The goal of this analysis is to statistically find the differences in the mean number of drinks and how the blood tests relate to the patients being diagnosed with liver disorders. The ANOVA test is to grasp the concept of the relationship between the continuous response variable also known as the dependent variable, number of drinks and the independent variables, Gammagt, Mcv, Alkphos, Sgpt, Sgot. This method makes use of classing the variable of interest and modelling them by the respective independent variables. Although, in this study one will firstly, use the one-way ANOVA test to consider one factor, drinks with two or more independent variables. Secondly, perform a two-way ANOVA test to determine the interactions between the number of drinks in association with what the blood tests indicate. The output of the F-statistics corresponding to the p-values will help one determine the significant effects. Tukey's test will be used to understand which blood test differs significantly in terms of the number of drinks. The plots will help analyze the interactions and relationships between blood tests and the number of drinks variations.

4. Understanding the Role of R-Squared in Predicting Alcohol Consumption through Blood Tests

R-squared is a statistical measure showing the proportion of variance in the dependent variable of drinks explained by the independent variables of mean corpuscular volume, alkaline phosphatase, alanine aminotransferase, aspartate aminotransferase, and gamma-glutamyl transpeptidase when fitted in a regression model. R-squared is primarily used to assess the strength of fitness of a regression model. It helps researchers to know exactly how their model is serving its purpose in predicting outcomes given certain inputs. A higher R-squared normally indicates a better fit; that is, a larger part of the dependent variable's variance is explained by the model.

R-squared can help in the quantification of how well the independent variables, the blood test results, predict the dependent variable that is excessive alcohol consumption. Suppose you consider different biomarkers within your regression model; a high value of R-squared will tell you that those biomarkers effectively explain the variation among patients suffering from liver disorders due to excessive alcohol consumption. This will guide clinicians in using some of the blood tests as predictive tools to assess the consumption of alcohol. R-squared, in this regard, will help us in determining the goodness of fit of the relationship between alcohol consumption, being the independent variable, and Gammagt level, which is the dependent variable.

A high value of R-squared will indicate that alteration in alcohol intake is highly associated with changing levels of Gammagt and will, therefore, support our hypothesis that increased intake of alcohol is associated with a corresponding rise in the level of Gammagt.

5. Hypothesis Testing

Hypothesis testing involves making inferences or arriving at conclusions about a population using data from a sample. This includes comparing the null hypothesis to an alternative hypothesis. Hypothesis testing plays a crucial role in research by helping researchers evaluate the statistical importances of their findings and differentiate between results that may the random or reflect actual effects in the population.

The procedure for hypothesis testing consists of the following steps:

Create Hypotheses: Define the null hypothesis(H0) and alternative hypothesis (Ha), where the null hypothesis suggests that there is no impact or difference between variables, H0: μ1 =μ2, whereas the alternative hypothesis suggests that there is a difference, H1: μ1 偃=μ2. Then select a level of significance (α) , which is usually at 0.05, is the likelihood of incorrectly rejecting the null hypothesis when it is true. Then choose the suitable test, then determine the Test statistic and make a choice on the results.

The most used hypothesis tests include the t-Test which is used to compare the means of two groups to a known population mean. And the F-Test, which is used to compare variances, testing whether the variances of the two populations are equal or not.

In conclusion, hypothesis testing offers a systematic approach for making decisions based on data and assessing the statistical importance of research results. By choosing the appropriate tests, researchers can determine if their sample data supports conclusions about relationships or differences in the broader population.

Conclusion

This study employed a rigorous methodology to explore the relationship between blood test indicators and alcohol consumption in diagnosing liver disorders, particularly focusing on Gamma-Glutamyl Transferase (Gammagt). By utilizing the UCI BUPA liver disorder dataset, it aimed to clarify the predictive value of various blood tests, correcting past misunderstandings in the literature.

The analysis began with descriptive statistics to understand the dataset, followed by correlation analysis to assess relationships between blood test variables and alcohol consumption. The primary technique, multiple linear regression, revealed Gammagt as a key predictor of alcohol-related liver disorders, supported by ANOVA testing to validate the significance of our findings.These results emphasize the importance of specific blood tests in early diagnosis and treatment planning for liver disorders associated with excessive alcohol intake.

Initial Analysis

Introduction

This study searches into the UCI BUPA liver disorder dataset to identify which blood tests can effectively predict liver disorders, particularly those related to alcohol consumption. The dataset includes 276 male subjects, detailing their blood test results and self-reported drinking habits. Accurate prediction of liver disorders is essential for early diagnosis, allowing physicians to act quick enough. Previous research on this dataset has misinterpreted the variables which led to inaccurate conclusions. This study aims to correct these misunderstandings by properly analyzing the dataset, focusing on the research question: *Which blood tests are reliable predictors for liver disorders, particularly in the context of alcohol consumption?*

As discussed in the methodology the research aims to find the best indicator for patients with liver disorder due to excessive alcohol consumption. The following blood test will be evaluated in the aim to find the most suitable indicator for the patients: Mean Corpuscular Volume (Mcv), Alkaline Phosphatase (alkphos), Alanine Aminotransferase (Sgpt), Aspartate Aminotransferase (Sgot), and Gamma-Glutamyl Transpeptidase (Gammagt). Statistical Methods will be used to further examine the relationship between the above variables/blood tests and their association with liver disorders using a UCI BUPA liver disorder dataset consisting of a sample of 276 men.

Aims and Objectives

The aim of the research is to assist physicians to make accurate diagnosis using the dataset provided by UCI BUPA to order to make inferences from the dataset through the use of statistical methods, namely through descriptive analysis, coloration analysis , multiple linear regression model, ANOVA, multiple linear regression (reduced model) using selection techniques (backward , forward and stepwise).This methods will be evaluated on SAS Studio with aim to provide a better understanding of relationships between the blood test and their association with liver disorders.

Results and Discussion

Descriptive analysis

The statistical measure involves conducting a descriptive analysis. This analysis aims to evaluate the distribution of the data and acquire a foundational understanding of how the data of each variable is spread through measures of central tendencies.

Mean Corpuscular Volume(Mcv):

The mean of the variable is 90.1558 which is slightly greater than the median suggesting a right skew in the data. The standard deviation 4.5712 suggests a moderate variability in the data. The data is negatively skewed with a kurtosis of 2.7209 which suggests a left-skewed distribution with slightly heavier tails.

Alkaline Phosphatase(Alkphos):

The mean of the variable is 69.6239 which is slightly greater than the median suggesting a right skew in the data. The standard deviation 18.4804 suggests a moderate variability in the data. The data is positively skewed with a kurtosis of 0.7345 which suggests a right-skewed distribution with tails similar to a normal distribution.

Alanine Aminotransferase(Sgpt):

The mean of the variable is 30.3478 which is slightly greater than the median suggesting a right skew in the data. The standard deviation 19.8963 indicates a high variability in the data. The data is positively skew with a kurtosis of 2.7209 which strongly suggests a right-skewed distribution with heavy tails.

Aspartate Aminotransferase(Sgot):

The mean of the variable is 24.5688 which is slightly greater than the median suggesting a slight right skew in the data. The standard deviation 10.5644 suggests a moderate variability in the data. The data is positively skewed with a kurtosis of 14.9740 suggesting a right-skewed distribution with heavier tails.

Gamma-Glutamyl Transpeptidase(Gammagt):

The mean of the variable is of 36.9493 which is slightly greater than the median suggesting a strong right skew in the data. The standard deviation 38.8124 suggests a high variability in the data. The data is positively skew with a kurtosis of 12.4122 suggesting a right-skewed distribution with heavy tails.

Drinks:

The mean of the variable is 3.4094 which is slightly greater than the median suggesting a slight right skew in the data. The standard deviation 3.2834 suggests a moderate variability in the data. The data is positive skew with a kurtosis of 4.3178 suggesting a right-skewed distribution with slightly heavier tails.

This further suggests that the), Alkaline Phosphatase (alk phos), Alanine Aminotransferase (Sgpt),Aspartate Aminotransferase (Sgot), and Gamma-Glutamyl Transpeptidase (Gammagt) are seen to be the most skewed with heavier tails suggesting that they might be good indicators of liver disorders.

Correlation Analysis with Pearson correlation coefficient

The Pearson analysis coefficient was utilized to measure the linear relationship between the variable drinks and the blood tests that were given. The value of r can vary from –1 to 1, where r= –1 indicates a negative perfect linear relationship, r = 1 indicates a positive perfect linear relationship.One will choose between the independent variable such as Mcv, Alkphos, Sgot, Sgpt and Gammagt, and drinks as the dependent variable.The aim of this analysis is to know which blood tests are strongly linked to the number of drinks and serve as important indicators of alcohol-related liver disorders.

Analysis between Mcv and Drinks

A correlation coefficient of r= 0.3577 signifies a positive relationship between drinks and Mcv. The p-value is below 0.0001 falling under the 0.05 significance level, therefore rejecting the null hypothesis.This indicates that there is a relationship between Mcv and drinks.

Analysis between Alkphos and Drinks

The correlation coefficient of r=0.1152 indicates a weak positive relationship between Drinks and Alkphos. However, the p-value of 0.0560 is significantly higher than the usual threshold of 0.05. Thus the null hypothesis is not rejected and suggests that the finding is insignificant.

Analysis between Sgpt and Drinks

A correlation coefficient of r= 0.1886 indicating a weak positive relationship between Drinks and Sgpt. The p-value is of 0.0016 which is lower than 0.05 significance level. Thus rejecting the null hypothesis, indicating a relationship between Sgpt and Drinks.

Analysis between Sgot and Drinks

The statistical correlation of r = 0.2582 which suggests a significant positive relationship between Drinks and. The p-value is below 0.0001, much smaller than the 0.05 significance level. Thus rejecting the null hypothesis, declaring a relationship between Drinks and Sgot.

Analysis between Gammagt and Drinks

A strong positive correlation of r = 0.3652 and a p-value below 0.0001. This suggests a strong relationship between Drinks and GAMMA GT. With a p-value below 0.0001, lower than the typical significant levels of 0.05, thus rejecting the null hypothesis, confirming a relationship between Drinks and Gammagt.

In conclusion, the correlation uncovers the important relationships between alcohol intake and the different liver function-related blood tests.Mcv, Sgot and Gammagt demonstrate significant positive correlations with alcohol consumption. Sgpt and Alkphos have weak correlations with the variable 'Drinks' and do not present significant findings.However analysis between Drinks and Gammagt showed the strongest correlation of r=0.3652.

Hypothesis testing Using the T-Test.

The purpose of the T-test is to compare the means of two groups to determine if there is a significant difference between them. The Pooled T-test assumes that the variances between the two groups are equal, while the Satterthwaite T-test assumes that the variances differ. Each variable is compared across two groups using both the pooled and Satterthwaite methods.

H0: There is no significant difference between alcohol consumption and all blood tests(Mcv, Alkphos, Sgpt, Sgot, Gammagt)

H1: There is a significant difference between alcohol consumption and different liver enzymes.

Mean Corpuscular Volume (Mcv) : The Pooled t-test is 0.1350 and the Satterthwaite t-test is 0.1249. Since both p-values are greater than 0.05, meaning that the null hypothesis is not rejected, there is no significant difference between Mcv levels and drinks.

Alkaline Phosphatase (Alkphos): The Pooled t-test p-value is 0.2936 and the Satterthwaite t-test p-value is 0.2977. Since both p-values are greater than 0.05, the null hypothesis is not rejected, so there is no significant difference between Alkphos levels and Drinks.

Alanine Aminotransferase (Sgpt): The Pooled t-test p-value is 0.5661 and the Satterthwaite p-value is 0. 5486.Due to both p-values being greater than 0.05, the null hypothesis is rejected, indicating that there is no difference between Sgpt levels and Drinks.

Aspartate Aminotransferase (Sgot): The Pooled t-test p-value is 0.0104 and the Satterthwaite p-value is 0.0070. Since both p-values are less than 0.05, the null hypothesis is rejected, indicating that there is a difference in Sgot levels and Drinks.

Gamma-Glutamyl Transferase (Gammagt): The Pooled t-test p-value is 0.0390 and the Satterthwaite p-value is 0.0340. Since both p-values are less than 0.05, the null hypothesis is rejected, meaning that there is a difference in the Gammagt levels and Drinks.

In conclusion, out of the five blood tests analyzed, it is only Sgot and Gammagt that shows a significant difference based on alcohol consumption, and this suggests that the enzymes could be associated with drinking disorder. Mcv, Alkphos and Sgpt do not show a significant difference with alcohol consumption, implying that they may not be related to alcohol intake.

R-squared (R²) and Coefficient of determination:

That is, the proportion of variability in the dependent variable explained by a regression model is measured by the coefficient of determination, normally denoted as R². The ranges for R² view from 0 to 1, where 0 signifies no explanation of the variance by the model and 1 signifies that the model explains all of the variance. The closer the R² value is to 1, the better the model is at predicting the outcome (Tunery, 2022). That is, it helps estimate the general fit of the model in predicting the outcomes based on the independent variables.

With an R² value of 0.2229, that is 22.29%, this model explains only 22.29% of variation in alcohol consumption ("drinks"), leaving 77.71% not accounted for by the model. Thus, a weak relation exists between the blood test results of mean corpuscular volume, alkaline phosphatase, alanine aminotransferase, aspartate aminotransferase, and gamma-glutamyl transpeptidase and alcohol intake. That is to say, independent variables do not stand that strong in predicting an outcome, thus making the model less powerful for correct predictions.

The low R² value of 22.29% suggests that only a small portion of variation in alcohol consumption is explained by this model.

This would also imply that blood tests from the model, such as gamma-glutamyl transpeptidase, are pretty weak indicators of alcohol consumption. While Gammagt does have a correlation with alcohol intake up to a point, this variable is a lousy predictor when considered with the other variables. In other words, Gammagtitself may not serve very well as a good index of the use of alcohol but may require other predictors for a better model.

From this model, the predictors used in this scenario, including Gammagt, are not good indicators of alcohol consumption. There is some improved correlation in Gammagt with higher consumption, but this is poor as a predictor combined with other factors. From these results, it would appear that Gammagt as an independent predictor is not very good, and further exploration is required in coming up with suitable predictors for the more realistic evaluation of alcohol use.

Analysis Of Variance (ANOVA) analysis

Introduction

To address the research question, the analysis of variance (ANOVA) is used to assess the predictive power of the five blood test variables - Gammagt, Mcv, alkphos, sgpt, and sgot. This approach will help identify which test strongly correlates with liver disorders. For each variable the p-value will be analyzed, if it is below 0.05 it suggests the variable has a significant relationship and could be a good predictor. The F-statistic will be considered to indicate the ratio between-group variance to within-group variance, a large F-value will indicate that the independent variable has a significant effect, which means a good predictor. R-squared will measure how the independent variable explains the continuous variable, higher r-squared explains the proportion of variability in the dependent variables. Tukey's test will compare the means between different levels of the predicator, which will help identify which groups are significantly different.

Analysis of Variance Code and Graph

ANOVA Discussion and Relation to research questions

1. Gamma-Glutamyl Transferase (ammagt):

The p-value (<0.05) indicates a statistically significant relationship, with an F-statistic is 3.15 and the r-squared value of 0.568 which is also relatively high, suggesting that Gammagt is a strong predictor of alcohol-related liver disorders. This supports the hypothesis that Gammagt is highly relevant in this study.

2. Mean Corpuscular Volume (Mcv):

The p-value is below 0.05 but the F-statistic is 2.54 and the r-squared value is only 0.19. This relatively low R-squared value suggests that Mcv does not have significant effects of predicting liver disorders to explain the variability in alcohol consumption, making it a weaker predictor.

3. Alkaline Phosphatase (Alkphos):

Although the p-value is below 0.05, the F-statistic is 1.56 and the r-squared value is 0.356 this suggests that Alkphos does not provide strong predictive power for liver disorders related to alcohol consumption. Its weak explanatory value is consistent with its limited role in alcohol-related liver damage.

4. Alanine Aminotransferase (Sgpt):

The p-value is below 0.05, with an F-statistic of 2.99 and the r-squared value of 0.443, making Sgpt another strong predictor. This result aligns with existing research showing Sgpt's relevance in liver function, particularly when combined with Gammagt.

5. Aspartate Aminotransferase (Sgot):

Despite a p-value below 0.05, the F-statistic is 2.25 and the r-squared value is 0.299, suggesting that Sgot is a weaker predictor compared to Gammagt and Sgpt. This may be because Sgot is less specific to liver damage compared to other enzymes.

In summary, the ANOVA analysis highlights Gammagt and sgpt as the most significant predictors, directly addressing the main research question. The findings support the proposed research question that elevated Gamma-Glutamyl Transferase (Gammagt) and Alanine Aminotransferase (Sgpt) levels are significantly associated with an increase in the number of drinks consumed, making Gammagt and Sgpt strong indicators of alcohol consumption.

<u>Multiple linear regression model analysis</u>

Introduction

Multiple linear regression(MLR) is used to predict the value of a of a dependent variable(in this research the dependent variable is 'drinks') based on two or more dependent variables also known as predictors(blood tests namely mean corpuscular volume (Mcv), alkaline phosphatase (ALP), alanine aminotransferase (Sgpt), aspartate aminotransferase (Sgot) and gamma-glutamyl transferase (Gammagt). This helps to assess how each independent variable influences the dependent variable, to estimate future values of the dependent variable and helps to capture the more complex relationships between dependent and independent variables. This leads to improved prediction accuracy and a better understanding of which variables have significant effect and its magnitude.To ensure this model's reliability the given data used for the model was split into training data set (selector 2) and testing data set (selector 1). The training data set is used to develop the model while the testing data is used to evaluate the model’s performance on unseen data to ensure that that model is not overfitted to the training model and assess how consistent the predictions are when given different data sets.

Comparison and data analysis of Training data(selector 2) Multilinear Regression Model and Testing data(selector 2) Multilinear Regression Model - Stepwise selection for training data with 0.05 Significance level

Using stepwise selection, the model concluded that mean corpuscular value (Mcv) and gamma-glutamyl transferase (Gammagt) are the best predictors of the drinks variable. Therefore Mcv and Gammagt were the two variables used in the Multilinear Regression model. When deciding which selector to use for the training data set, the larger selector 2 data set with a frequency of 156 as a sufficient amount of data is needed in order to allow the model to learn the patterns and relationships between the variables.

Analysis of R-squared value of the Training model and Testing model

The R-squared value is used to show the strength of the relationship between independent variables and dependent variables. It indicates what proportion of the variance in the dependent variable can be explained by the model providing insight into how well fitted the model is to the data.

In the Training data model R-squared model is equal to 0.1202 meaning approximately only 12.02% of the variability in drinks can be explained by both Mcv and Gammagt which indicates only a small portion of the variance is explained by the model.The Testing data set indicates that approximately 39.94% of the variability in the dependent variable (drinks) can be explained by the model with both Mcv and Gammagt. This is a significant improvement compared to the training model’s R-Square of 0.1202, suggesting that the testing model captures more of the variability in drinks. Smaller datasets are often subject to higher variance, meaning that the performance metrics (like R-Square) can fluctuate more. This could explain why the testing model shows a better fit. While the results explain a moderate portion of the variance in the testing data, the model is underfitted to the training data. Both R-squared values are still quite low and could suggest that the model is under fitted. The performance of the model could be improved by the addition of other predictor variables as stepwise may not always capture the best overall fit for the model and taking into account the complexities of the relationship between the predictors and the drinks variable.

Analysis of C(p) value of the Training model and Testing model

Mallows’ C(p) is used to assess how well the regression model fits the data without the model being too complex. A good model will have a lower C(p) value closer to the number of predictors used in the model creating a more parsimonious model, meaning it uses the fewest number of variables while still having enough predictors to show the most significant relationships between dependent and independent variables. A higher C(p) value may suggest overfitting of the model or that it is using too many predictor variables.

The C(p) value in the training model is 4.1063 which is considerably in relation to the number of predictors indicating the model is a good fit for the training data set. The C(p) value is not provided. Both models use Mcv and Gammagt which contradicts existing research that does not typically include Mcv as a predictor of alcohol. However the inclusion of Gammagt as a good predictor of alcohol consumption is consistent with other research and supports the research hypothesis that an increase in Gammagt levels leads to higher alcohol consumption predictions.

Analysis of p-values of the Training model and Testing model

The p-value is a statistic that shows the probability of observing the data and helps to determine whether the research findings are significant. In this report, a p-value less than 0.05 is considered significant and would indicate that the results are not due to random variation. Conversely, a p-value greater than 0.05 would indicate the results are simply the result of chance or random variation.

According to the ANOVA table for the Training data model, the p-value(labelled Pr>F) is equal to 0.0001 which is less than 0.05 indicating that the model is significant meaning that at least one of the predictors is significantly related to the drinks variable. Similarly, the p-value of the Testing model also had a p-value of 0.001 indicating that the model is significant meaning that at least one of the predictors is significantly related to the drinks variable. These predictors being the Mcv and Gammagt. Overall, both models are significant predictors.

When observing specific p-values of the individual predictors Mcv and Gammagt in each model; the predictors Mcv and Gammagt in the Training model were found to be significant predictors of the drinks variable evidenced by their p-values being 0.0011 and 0.0220 respectively (both less than 0.05). This shows that changes in the Mcv and Gammagt variables play a crucial role in influencing the drinks variable or observed amount of alcohol consumption.

The Testing model only provides p-values associated with the t-test (labelled Pr>|T|) for individual predictors which can also be used to evaluate their significance which can be found in the parameter estimates table. Both Mcv and Gammagt are less than 0.0001 which is less than 0.05 thus reinforcing that Mcv and Gammagt play a crucial role in predicting alcohol consumption.

In summary, the p-values mentioned indicate the strong predictive power of the model and chosen predictors Mcv (mean corpuscular volume) and Gammagt (gamma-glutamyl transferase).

Conclusion for testing model based on training model

The multiple linear regression models for the Training data set and the Testing data set both produced p-values that are less than 0.05 indicating that the predictor variables have a significant influence on the drinks variable and demonstrates the combined predictive power of the mean corpuscular volume(Mcv) and gamma-glutamyl transferase (Gammagt). The higher R-squared value of the Testing data shows that the model is better fitted to that specific data set but ideally both models need to have a higher R-squared value to ensure that the model is a good fit for the data and future unseen data sets. This would indicate that a greater portion of the drinks variable can be explained by the model. Furthermore, graphical representations demonstrated clustering of values, suggesting that the model's fit may be compromised by extreme outliers, which can skew performance. Although there is an improvement in the variance of the multiple linear regression model of the testing data set, it is still relatively low and therefore more investigation into the data set, consideration of other predictor variables and addressing extreme outliers in the data set could improve the models accuracy. Ultimately, the findings support the conclusion that combining blood tests, particularly gamma-glutamyl transferase (Gammagt), offers a more accurate assessment of alcohol consumption.

Multiple Regression Model (Reduced Model)

A multiple regression model is used to predict the value of a dependent variable based on two or more independent variables (predictors). The full model includes all potential predictors, while the reduced model focuses on a fixed number of significant predictors after removing irrelevant or less important ones based on their p-values which determines whether they are good predictors or not. A reduced model is often more detailed and can avoid overfitting. There are selection techniques involved in reducing the multiple regression model; namely the forward selection method, backward selection method and the stepwise selection method that combines both forward and backward approaches; and all these techniques help to eliminate every irrelevant predictor according to its significance, while also improving the reliability of our results and helps to reject or not reject our null hypothesis and to support our research question. I

Selection Techniques:

- The first variable of interest selected is Gammagt with an R^2 = 0.13333, indicating that it explains 13.33% of the variance in the dependent variable, "Drinks”. The model is statistically significant with F=42.16and p<0.0001.

- The second variable, Mcv, is added, increasing the R^2 to 0.2114 (21.14% of the variance explained). The model remains highly significant (p<0.0001).

The full model starts with all the five predictors; namely Gammagt, Mcv, Alkphos, Sgpt, sgot, yielding R^2 = 0.2189, explaining 21.89% of the variance in the dependent variable "drinks." Using the forward selection method, the variable Sgpt is removed, slightly reducing R^2 to 0.2164. After removing Sgot with the same technique, R^2 decreases further to 0.2152.

Full model R^2: 0.2189

Reduced Model R^2: 0.2114 (after applying forward selection).

Removing Alkphos levels Gammagt and Mcv as the final predictors, resulting in R^2 = 0.2114, which is the same as the forward selection reduced model.

Evaluation of R^2:

- The full model has an R^2 = 0.2189, explaining 21.89% of the variance in "drinks."
- The reduced model (both Forward Selection and Backward Elimination) results in an R^2 = 0.2114, explaining 21.14% of the variance, which is very close to the full model.

Interpretation:

- Forward Selection: The reduced model retains Gammagt and Mcv, which explain 21.14% of the variance in the dependent variable "drinks." This selection keeps only the most significant predictors while maintaining a high percentage of variance explained.
- Backward Elimination: By removing non-significant predictors (Alkphos, Sgpt, Sgot), the reduced model becomes simpler with minimal loss in the variance explained (from 21.89% to 21.14%). This brings us close to precision at concluding which ones of the predictors can be used.

Both approaches yield a model with similar self-supporting results, suggesting that Gammagt and Mcv are the most important predictors in this research.

Conclusion

The blood tests Gammagt (Gammagt) and Sgpt (ALP have been carefully studied and have shown that both are definitive biomarkers for diagnosing patients with excessive alcohol consumption which could lead to liver disorders. The high levels of Gammagt is sensitive to alcohol which makes it a good predictor for excessive alcohol consumption. While Sgpt is more focused on liver cell damage and has also shown high levels in this study.

Through statistical analysis, both the blood tests have significant predictive value and high levels correlating strongly with liver disorders. Therefore, using Gammagt and Sgpt together will assist in accurately diagnosing patients of a high risk of a liver disorder and excessive alcohol consumption as the contributing factor.

Summary and Conclusion

The study was aimed to investigate the effectiveness of various blood tests in excessive alcohol consumption in patients with liver disorders. Findings and analysis of the dataset provided by the UCI BUPA suggest that gamma-glutamyl transferase (Gammagt) is the strongest indicator among the five tests which were examined namely, mean corpuscular volume (Mcv), alkaline phosphatase (ALP), alanine aminotransferase (Sgpt), aspartate aminotransferase (S) and gamma-glutamyl transferase (Gammagt).

The Pearson correlation analysis results highlight this, with Gammagt a strong positive correlation (r=0.3652) with the number of drinks consumed. This was statistically significant with a p-value below 0.0001, which validates the relationship between higher Gammagt levels and increased alcohol intake. Similarly, Sgot and Mcv also show a positive correlation with alcohol consumption.However, the correlation for Sgpt is shown to be weak, and no significant relationship was found between Alkphos levels and alcohol consumption.

Furthermore, the analysis using the Student T-test further validates these findings. as a significant difference in Mcv and Gammagt levels between groups with varying alcohol consumption suggesting that ALP and Sgpt may not be reliable indicators of excessive alcohol consumption.

In summary, the research indicates that Gamma-gt, along with Sgot and Mcv, serve as a reliable marker for assessing excessive alcohol consumption in patients with liver disorders.Although individual tests provide valuable insight, the diagnostic accuracy improves when used in conjunction. Further research is recommended to refine the predictive models in this study and further enhance their application in clinical practice.

Bibliography

1. Dufour, D. R., Lott, J. A., Nolte, F. S., Gretch, D. R., Koff, R. S., & Seeff, L. B. (n.d.). Diagnosis and monitoring of hepatic injury. II. Recommendations for use of laboratory tests in screening, diagnosis, and monitoring. *Clin Chem*, *46*(12), 2050-2068. https://doi.org/10.1093/clinchem/46.12.2050
2. Jain, P., Batta, A. K., & Singh, P. (2023, August 25). Comparative Study of Serum Levels of Gamma-glutamyl Transferase, Aspartate Aminotransferase (AST), Alanine Transaminase (ALT), AST:ALT, and Bilirubin in Patients with Chronic Hepatitis. *Indian Journal of Medical Biochemistry*, *26*(3), 74-76. https://www.ijmb.in/doi/IJMB/pdf/10.5005/jp-journals-10054-0208
3. Nyblom, H., Berggren, U., Balldin, J., & Olsson, R. (2004, July). HIGH AST/ALT RATIO MAY INDICATE ADVANCED ALCOHOLIC LIVER DISEASE RATHER THAN HEAVY DRINKING. *Alcohol and Alcoholism*, *39*(4), 336-339. https://doi.org/10.1093/alcalc/agh074
4. Patel, R., & Mueller, M. (2023, July 13). National Library of Medicine. *Alcoholic Liver Disease*, 150. https://www.ncbi.nlm.nih.gov/books/NBK546632/
5. Pratt, D. S. (2019). Liver Chemistry and Function Tests. In *Sleisenger and Fordtran's Gastrointestinal and Liver Disease* (11th ed., pp. 1243-1252). Clinical Key. http://dx.doi.org/10.1016/B978-1-4160-6189-2.00073-1
6. Torruellas, C., French, S. W., & Medici, V. (2014, September 7). Diagnosis of alcoholic liver disease. *World Journal of Gastroenterology*, *20*(33), 11684-11699. https://doi.org/10.3748/wjg.v20.i33.11684
7. Gupta, A. (2024, June 4). *Spearman's Rank correlation: The Definitive Guide to understand*. Simplilearn.com. https://www.simplilearn.com/tutorials/statistics-tutorial/spearmans-rank-correla

tion#:~:text=Spearman's%20rank%20correlation%20measures%20the,represented%20using%20a%20monotonic%20function.

8. Turney, S. (2024, February 10). Pearson Correlation Coefficient (r) | Guide & Examples. Scribbr. Retrieved October 10, 2024, from https://www.scribbr.com/statistics/pearson-correlation-coefficient/

Appendix

Timeline

DATE	TASK	COMMENT
27 July – 1st Team meeting	- Discussing the project - Distributing tasks to each team member for the first submission - Setting a deadline for each task assigned by group members	
2 Aug – 2nd Team meeting	- Finalizing first part of abstract	
5 Aug – 1st Submission	- Abstract - Research questions - Timeline Plan	- Receive feedback on the 8th
6 Aug–3rd Team meeting	- Discussing literature review	-
9 Aug-4th Team meeting	- Work on the feedback for the 1st submission	
13th Aug – 5th Team meeting	- Checking progress on literature review	

16th Aug – 2nd Submission	- Literature review	- Receive feedback on the 21st
20th Aug-6th Team meeting	- Discussing methodology	
23rd Aug-7th Team meeting	- Work on the feedback for the 2nd submission	
27th Aug – 8th Team meeting	- Checking progress on methodology	
30th Aug – 3rd Submission	- Methodology	- Receive feedback on the 4th of Sep
3rd Sep-9th Team meeting	- Discussing analysis	
6th Sep-10th Team meeting	- Work on the feedback	
10th Sep – 11th Team meeting	- Finalizing analysis	
20th Sep – 4th Submission	- Analysis	- Receive feedback on the 25th
27th Sep-12th Team meeting	- Work on the feedback	
4 Oct-13th Team meeting	- Draft report	
7th Oct – Final submission	- Final report	

Descriptive Analysis

Code:

```
ODS HTML;

PROC UNIVARIATE DATA = "/home/u63783556/SAS/group_6_train.sas7bdat";

VAR Mcv alkphos sgpt sgot Gammagt;

TITLE "Descriptive analysis of  UCI BUPA liver disorder dataset";
```

```
HISTOGRAM/NORMAL (COLOR=BLUE W=5) NROWS=1;

RUN;

ODS HTML CLOSE;
```

Graphs and Tables:

Descriptive analysis of UCI BUPA liver disorder dataset

The UNIVARIATE Procedure

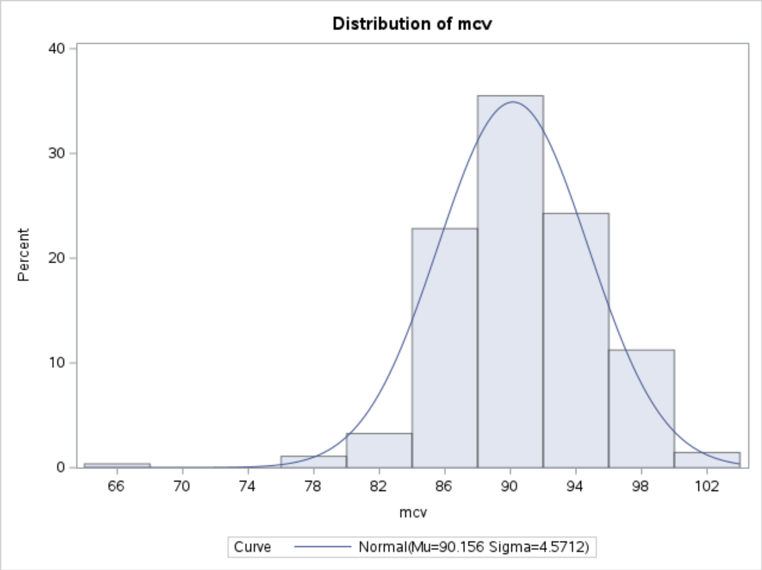
Variable: Mcv

Moments			
N	276	Sum Weights	276
Mean	90.1557971	Sum Observations	24883
Std Deviation	4.57117479	Variance	20.895639
Skewness	-0.4667643	Kurtosis	2.72094595
Uncorrected SS	2249093	Corrected SS	5746.30072
Coeff Variation	5.070306	Std Error Mean	0.27515239

Basic Statistical Measures			
Location		Variability	
Mean	90.15580	Std Deviation	4.57117
Median	90.00000	Variance	20.89564
Mode	91.00000	Range	37.00000
		Interquartile Range	6.00000

Descriptive analysis of UCI BUPA liver disorder dataset

The UNIVARIATE Procedure



Descriptive analysis of UCI BUPA liver disorder dataset

The UNIVARIATE Procedure

Variable: alkphos

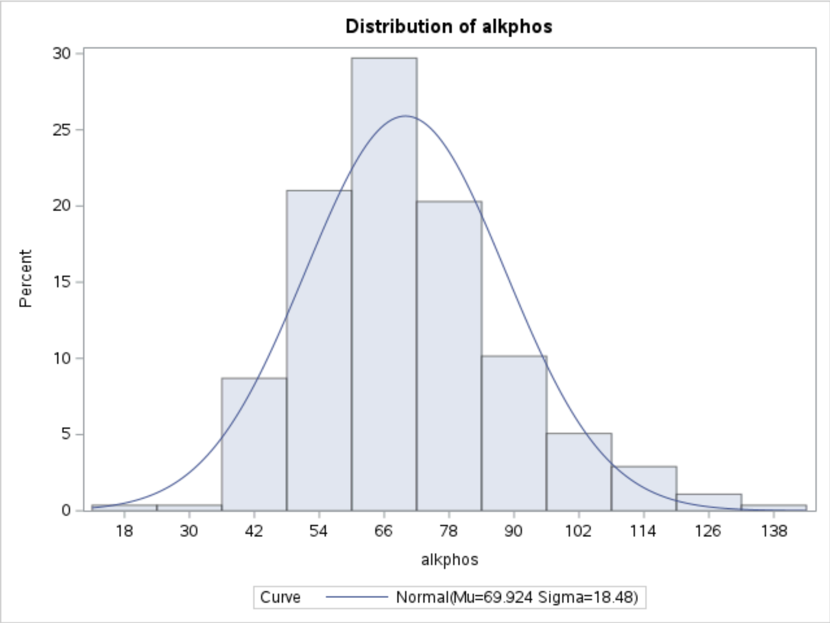
Moments			
N	276	Sum Weights	276
Mean	69.923913	Sum Observations	19299
Std Deviation	18.4803977	Variance	341.525099
Skewness	0.73451927	Kurtosis	0.69084625
Uncorrected SS	1443381	Corrected SS	93919.4022
Coeff Variation	26.4292956	Std Error Mean	1.11238922

Basic Statistical Measures			
Location		Variability	
Mean	69.92391	Std Deviation	18.48040
Median	67.00000	Variance	341.52510

Basic Statistical Measures			
Location		Variability	
Mode	62.00000	Range	115.00000
		Interquartile Range	22.00000

Descriptive analysis of UCI BUPA liver disorder dataset

The UNIVARIATE Procedure



Descriptive analysis of UCI BUPA liver disorder dataset

The UNIVARIATE Procedure

Variable: sgpt

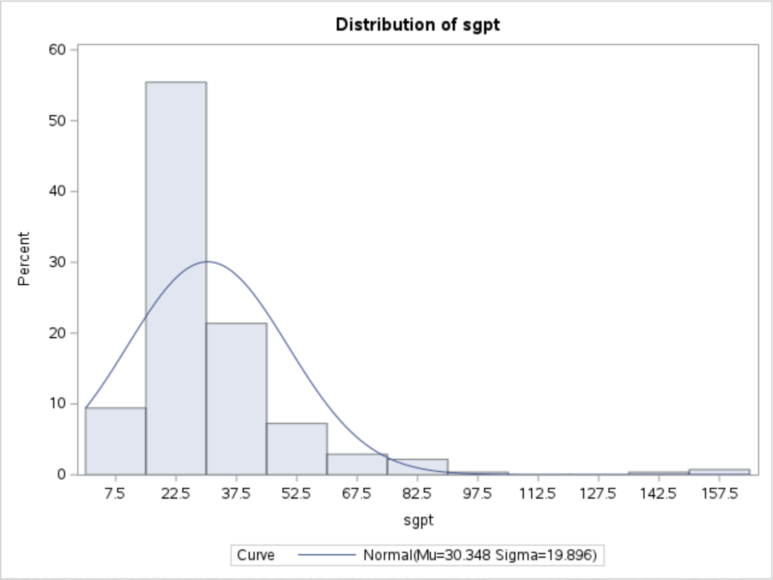
Moments			
N	276	Sum Weights	276
Mean	30.3478261	Sum Observations	8376
Std Deviation	19.8963321	Variance	395.864032
Skewness	3.21712749	Kurtosis	14.9740438

Moments			
Uncorrected SS	363056	Corrected SS	108862.609
Coeff Variation	65.5609797	Std Error Mean	1.19761846

Basic Statistical Measures			
Location		Variability	
Mean	30.34783	Std Deviation	19.89633
Median	25.50000	Variance	395.86403
Mode	17.00000	Range	151.00000
		Interquartile Range	15.00000

Descriptive analysis of UCI BUPA liver disorder dataset

The UNIVARIATE Procedure



Descriptive analysis of UCI BUPA liver disorder dataset

The UNIVARIATE Procedure

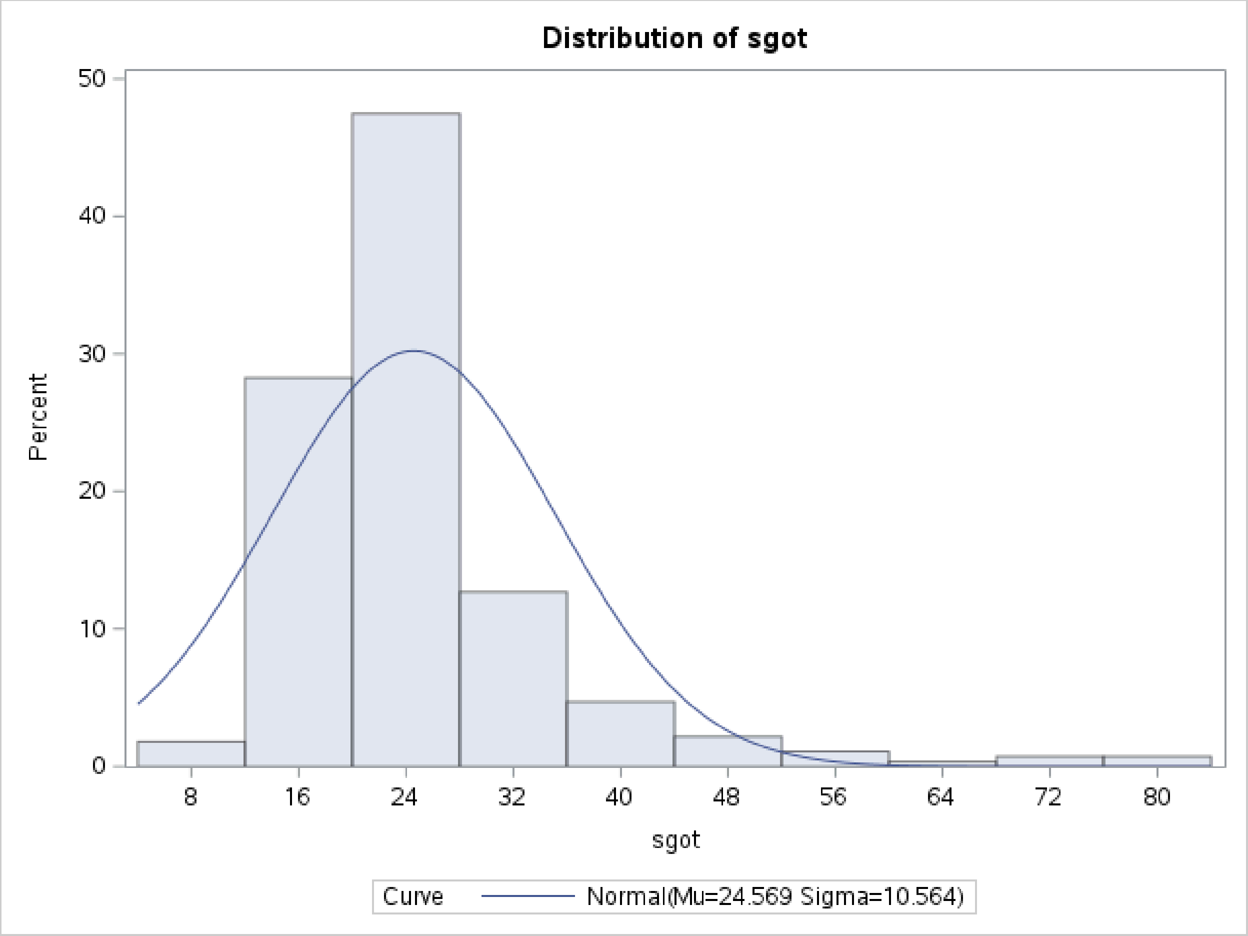
Variable: sgot

Moments			
N	276	Sum Weights	276
Mean	24.5688406	Sum Observations	6781
Std Deviation	10.5643813	Variance	111.606153
Skewness	2.39091651	Kurtosis	8.28326132
Uncorrected SS	197293	Corrected SS	30691.692
Coeff Variation	42.9991041	Std Error Mean	0.63590103

Basic Statistical Measures			
Location		Variability	
Mean	24.56884	Std Deviation	10.56438
Median	22.00000	Variance	111.60615
Mode	20.00000	Range	77.00000
		Interquartile Range	8.00000

Descriptive analysis of UCI BUPA liver disorder dataset

The UNIVARIATE Procedure



Descriptive analysis of UCI BUPA liver disorder dataset

The UNIVARIATE Procedure

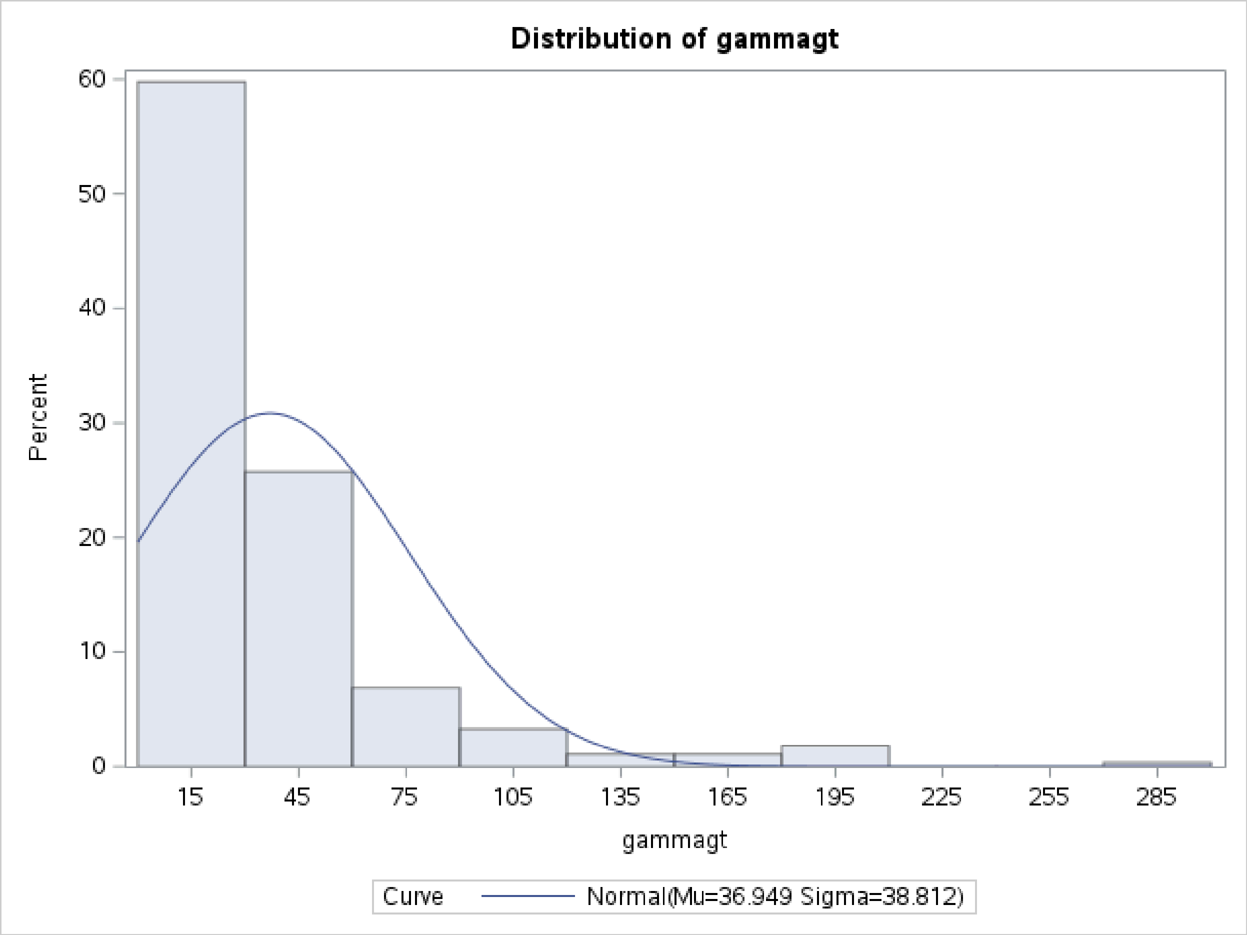
Variable: Gammagt

Moments			
N	276	Sum Weights	276
Mean	36.9492754	Sum Observations	10198
Std Deviation	38.8124296	Variance	1506.40469
Skewness	3.11632812	Kurtosis	12.4122306
Uncorrected SS	791070	Corrected SS	414261.29
Coeff Variation	105.042465	Std Error Mean	2.33623372

Basic Statistical Measures			
Location		Variability	
Mean	36.94928	Std Deviation	38.81243
Median	24.00000	Variance	1506
Mode	11.00000	Range	292.00000
		Interquartile Range	27.00000

Descriptive analysis of UCI BUPA liver disorder dataset

The UNIVARIATE Procedure



Code for Correlation Analysis

```
ods html;

proc corr data = "/home/u63783268/group_6_train (1).sas7bdat";

var Mcv alkphos sgpt sgot Gammagt; with drinks;

title "blood test correlation using with statement";

run;

ods html close;
```

Graphs and Tables

Correlations Analysis

blood test correlation using with statement

The CORR Procedure

1 With Variables:	drinks
5 Variables:	Mcv alkphos sgpt sgot Gammagt

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
drinks	276	3.40942	3.28342	941.00000	0	20.00000
Mcv	276	90.15580	4.57117	24883	65.00000	102.00000
alkphos	276	69.92391	18.48040	19299	23.00000	138.00000
sgpt	276	30.34783	19.89633	8376	4.00000	155.00000
sgot	276	24.56884	10.56438	6781	5.00000	82.00000
Gammagt	276	36.94928	38.81243	10198	5.00000	297.00000

Pearson Correlation Coefficients, N = 276 Prob > |r| under H0: Rho=0					
	Mcv	alkphos	sgpt	sgot	Gammagt
drinks	0.35770 <.0001	0.11519 0.0560	0.18860 0.0016	0.25817 <.0001	0.36516 <.0001

Code for Hypothesis Testing

```
ods html;
proc ttest data = "/home/u63783268/group_6_train (1).sas7bdat";
        class selector;
        var Mcv alkphos sgpt sgot Gammagt;
title "Blood test ttest";
run;
ods html close;
```

Blood test ttest

The TTEST Procedure

Variable: Mcv

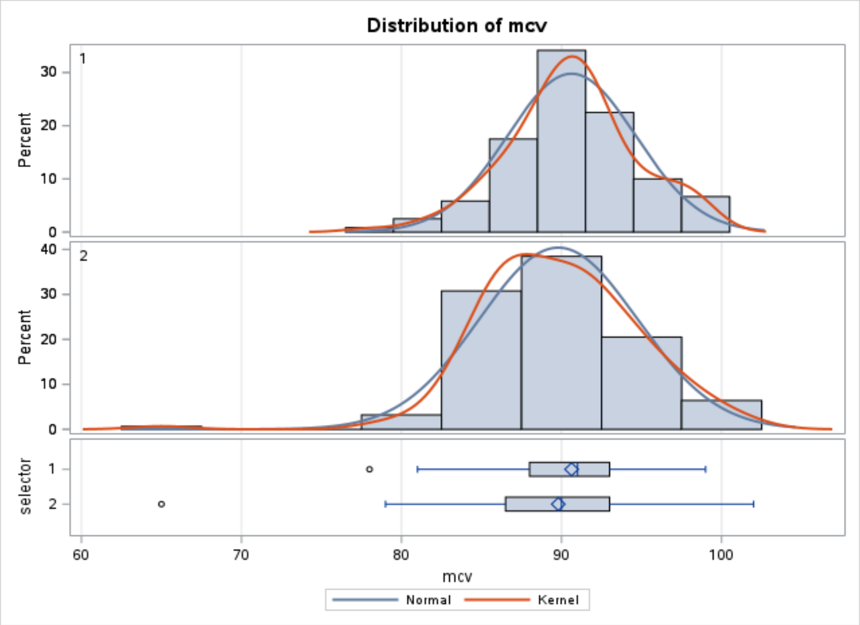
selector	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
1		120	90.6250	4.0190	0.3669	78.0000	99.0000
2		156	89.7949	4.9367	0.3952	65.0000	102.0
Diff (1-2)	Pooled		0.8301	4.5608	0.5538		
Diff (1-2)	Satterthwaite		0.8301		0.5393		

selector	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		90.6250	89.8985	91.3515	4.0190	3.5668	4.6035

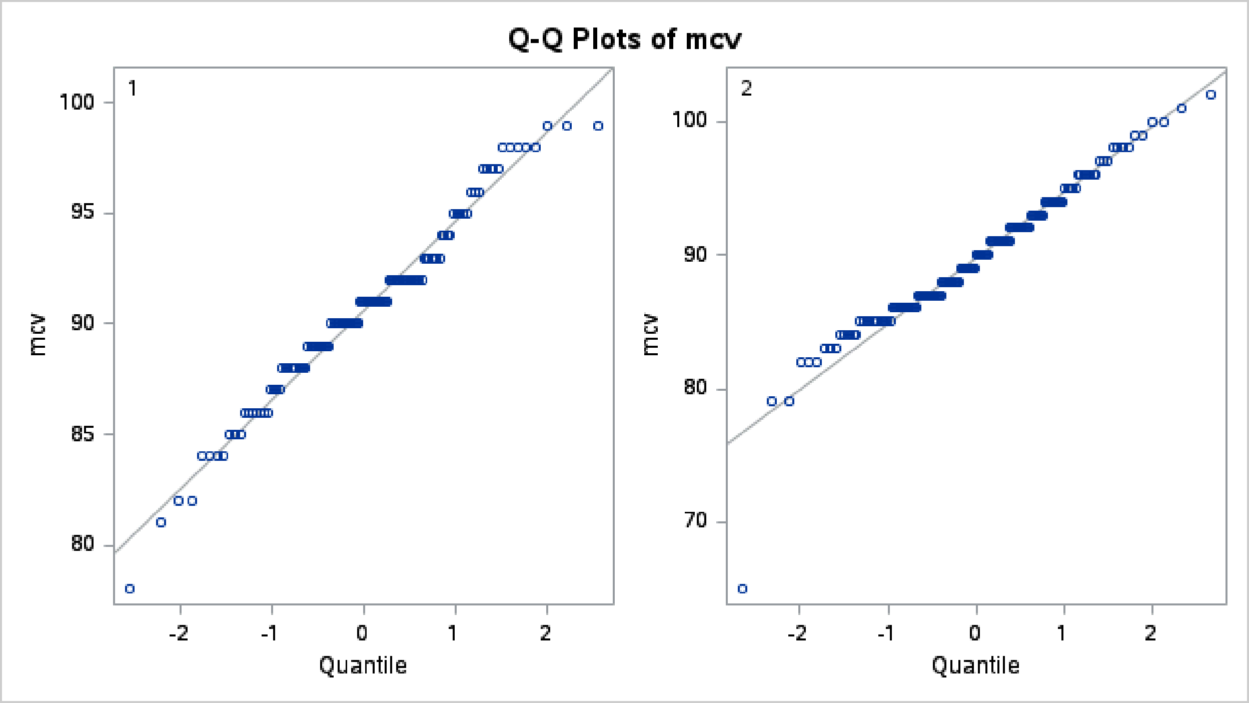
selector	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
2		89.7949	89.0141	90.5756	4.9367	4.4429	5.5548
Diff (1-2)	Pooled	0.8301	-0.2601	1.9204	4.5608	4.2089	4.9776
Diff (1-2)	Satterthwaite	0.8301	-0.2315	1.8918			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	274	1.50	0.1350
Satterthwaite	Unequal	273.1	1.54	0.1249

Equality of Variances



Method	Num DF	Den DF	F Value	Pr > F
Folded F	155	119	1.51	0.0190



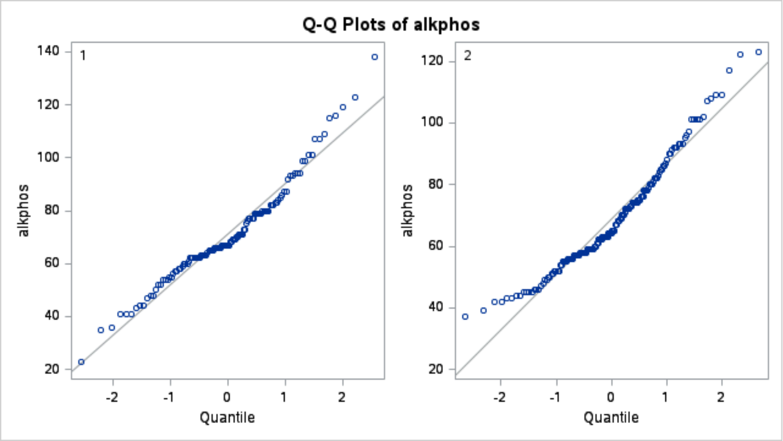
Variable: alkphos

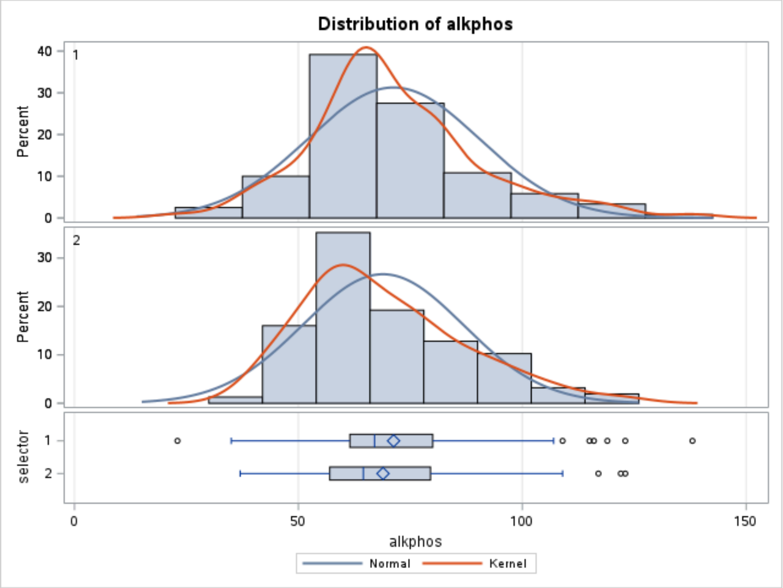
selector	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
1		120	71.2583	19.1327	1.7466	23.0000	138.0
2		156	68.8974	17.9570	1.4377	37.0000	123.0
Diff (1-2)	Pooled		2.3609	18.4768	2.2435		
Diff (1-2)	Satterthwaite		2.3609		2.2622		

selector	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		71.2583	67.8000	74.7167	19.1327	16.9801	21.9152
2		68.8974	66.0574	71.7375	17.9570	16.1611	20.2055
Diff (1-2)	Pooled	2.3609	-2.0558	6.7776	18.4768	17.0508	20.1650
Diff (1-2)	Satterthwaite	2.3609	-2.0947	6.8165			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	274	1.05	0.2936
Satterthwaite	Unequal	247.62	1.04	0.2977

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	119	155	1.14	0.4570





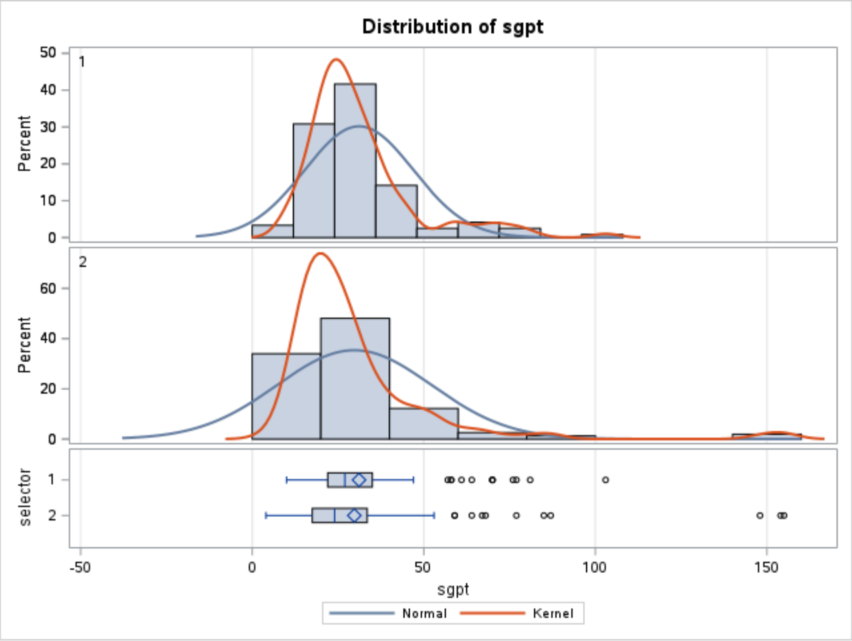
Variable: sgpt

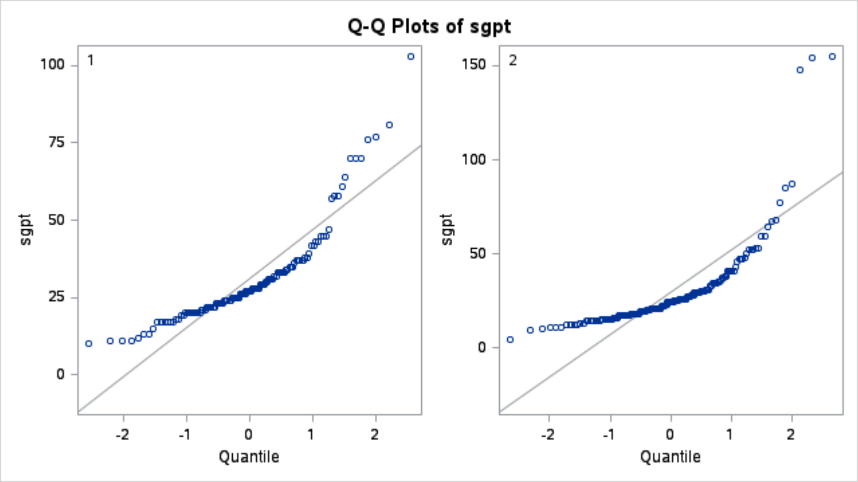
selector	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
1		120	31.1333	15.8581	1.4476	10.0000	103.0
2		156	29.7436	22.5483	1.8053	4.0000	155.0
Diff (1-2)	Pooled		1.3897	19.9206	2.4188		
Diff (1-2)	Satterthwaite		1.3897		2.3140		

selector	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		31.1333	28.2669	33.9998	15.8581	14.0739	18.1644
2		29.7436	26.1774	33.3098	22.5483	20.2932	25.3716
Diff (1-2)	Pooled	1.3897	-3.3721	6.1516	19.9206	18.3832	21.7408
Diff (1-2)	Satterthwaite	1.3897	-3.1660	5.9455			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	274	0.57	0.5661
Satterthwaite	Unequal	271.96	0.60	0.5486

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	155	119	2.02	<.0001





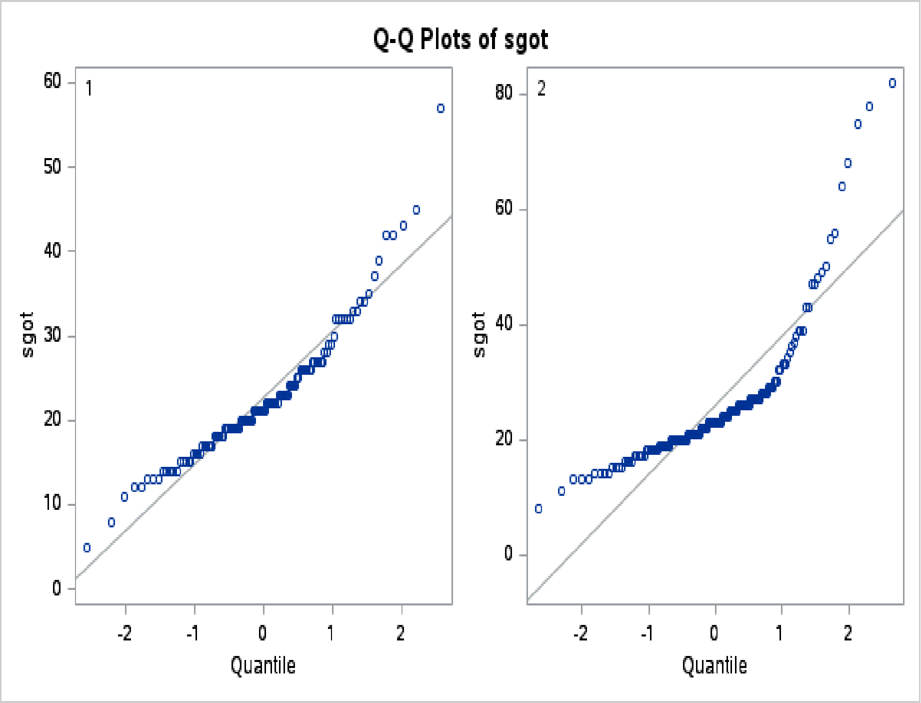
Variable: sgot

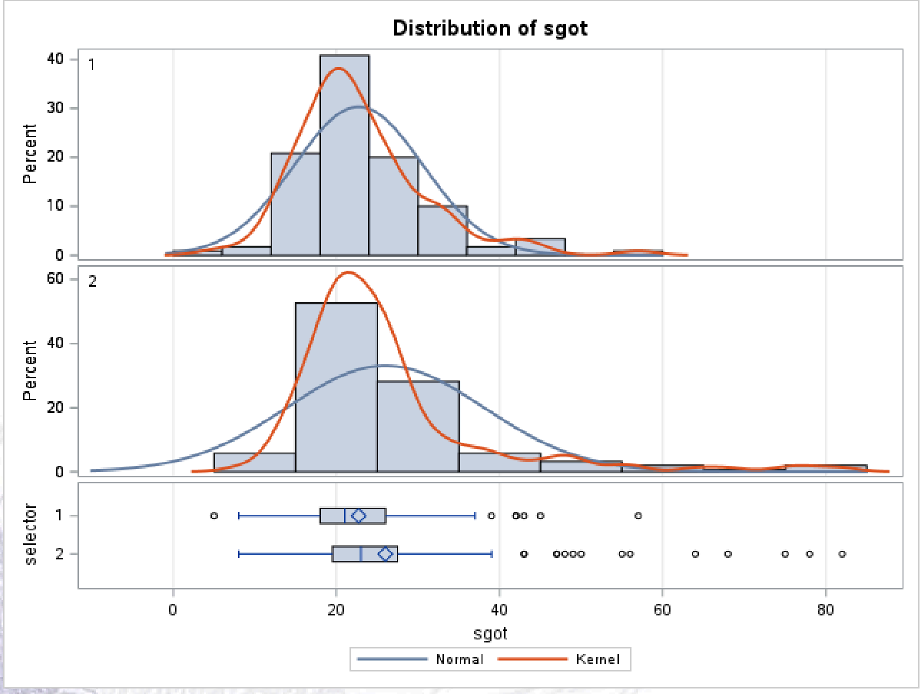
selector	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
1		120	22.7167	7.8998	0.7211	5.0000	57.0000
2		156	25.9936	12.0582	0.9654	8.0000	82.0000
Diff (1-2)	Pooled		-3.2769	10.4573	1.2698		
Diff (1-2)	Satterthwaite		-3.2769		1.2050		

selector	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		22.7167	21.2887	24.1446	7.8998	7.0110	9.0487
2		25.9936	24.0865	27.9007	12.0582	10.8523	13.5680
Diff (1-2)	Pooled	-3.2769	-5.7766	-0.7772	10.4573	9.6503	11.4128
Diff (1-2)	Satterthwaite	-3.2769	-5.6495	-0.9044			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	274	-2.58	0.0104
Satterthwaite	Unequal	267.68	-2.72	0.0070

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	155	119	2.33	<.0001





Variable: Gammagt

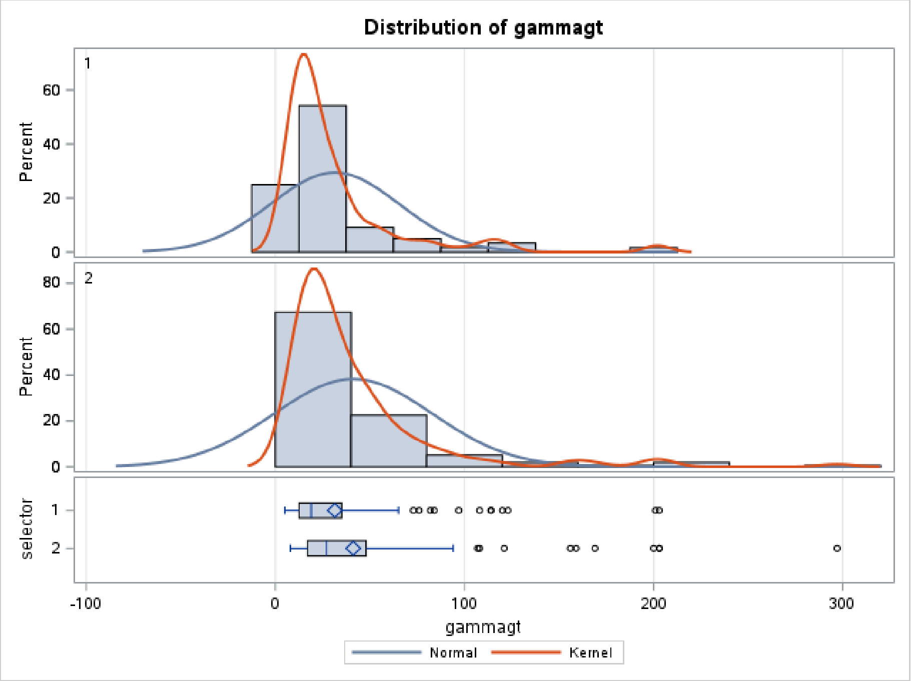
selector	Method	N	Mean	Std Dev	Std Err	Minimum	Maximu m
1		120	31.4583	33.9064	3.0952	5.0000	203.0

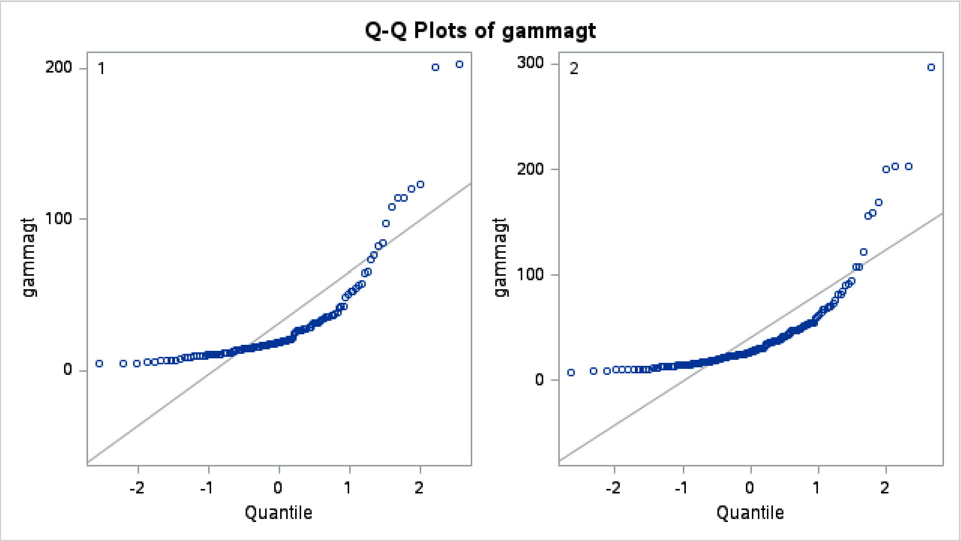
selector	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
2		156	41.1731	41.8178	3.3481	8.0000	297.0
Diff (1-2)	Pooled		-9.7147	38.5816	4.6847		
Diff (1-2)	Satterthwaite		-9.7147		4.5596		

selector	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		31.4583	25.3295	37.5872	33.9064	30.0916	38.8376
2		41.1731	34.5593	47.7869	41.8178	37.6356	47.0539
Diff (1-2)	Pooled	-9.7147	-18.9373	-0.4922	38.5816	35.6040	42.1069
Diff (1-2)	Satterthwaite	-9.7147	-18.6912	-0.7383			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	274	-2.07	0.0390
Satterthwaite	Unequal	273.22	-2.13	0.0340

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	155	119	1.52	0.0168

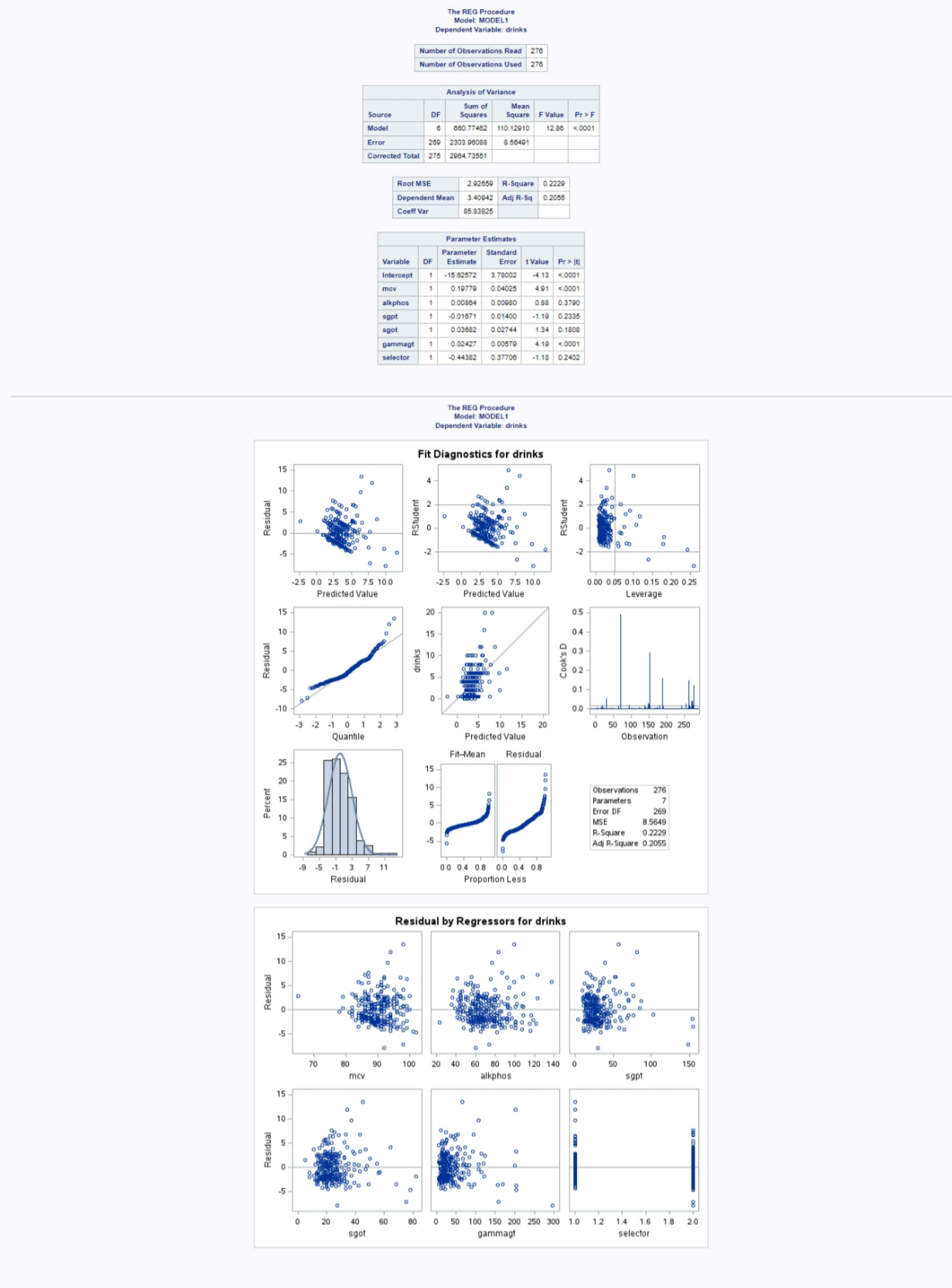




<u>R-squared</u>

<u>Code:</u>

```
ODS HTML;

PROC REG DATA ="/home/u63363624/group_6_train.sas7bdat";

MODEL Drinks = Mcv alkphos sgpt sgot Gammagt selector;

RUN;

ODS HTML CLOSE;
```



The REG Procedure
Model: MODEL1
Dependent Variable: drinks

Number of Observations Read	276
Number of Observations Used	276

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	660.77462	110.12910	12.86	<.0001
Error	269	2303.96088	8.56491		
Corrected Total	275	2964.73551			

Root MSE	2.92659	R-Square	0.2229
Dependent Mean	3.40942	Adj R-Sq	0.2055
Coeff Var	85.83825		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-15.62572	3.78002	-4.13	<.0001
mcv	1	0.19779	0.04025	4.91	<.0001
alkphos	1	0.00864	0.00980	0.88	0.3790
sgpt	1	-0.01671	0.01400	-1.19	0.2335
sgot	1	0.03682	0.02744	1.34	0.1808
gammagt	1	0.02427	0.00579	4.19	<.0001
selector	1	-0.44382	0.37706	-1.18	0.2402

The REG Procedure
Model: MODEL1
Dependent Variable: drinks

Fit Diagnostics for drinks

Observations	276
Parameters	7
Error DF	269
MSE	8.5649
R-Square	0.2229
Adj R-Square	0.2055

Residual by Regressors for drinks

Analysis of Variance Code and Graphs

Each variable changed for every procedure using the following code:

```
ODS HTML;
PROC ANOVA DATA= '/home/u63790646/my_shared_file_links/u63790646/GROUP6/group_6_train (6).sas7bdat';
 CLASS Gammagt;
 MODEL drinks=Gammagt;
 MODEL drinks=Mcv;
 MODEL drinks=alkphos;
 MODEL drinks=sgpt;
 MODEL drinks=sgot;
 MEANS Gammagt/TUKEY CLDIFF;
TITLE 'Analysis of Variance for the relationship between the number of drinks and Gammagt';
RUN;
ODS HTML CLOSE;
```

Class Level Information		
Class	Levels	Values

Gammagt	82	5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
		37 38 39 41 42 43 44 46 47 48 49 50 52 53
		59 60 62 64 65 67 68 69 70 71 73 76 81 82
		169 200 201 203 297

Mcv	25	65 78 79 81 82 83 84 85 86 87 88 89 90 91

alkphos	73	23 35 36 37 39 41 42 43 44 45 46 47 48 49
		67 68 69 70 71 72 73 74 75 76 77 78 79 80
		85 86 87 88 90 91 92 93 94 95 96 97 99 101

sgpt	59	4 9 10 11 12 13 14 15 16 17 18 19 20 21 22
		39 41 42 43 45 46 47 48 50 52 53 57 58 59
		70 76 77 81 85 87 103 148 154 155

sgot	45	5 8 11 12 13 14 15 16 17 18 19 20 21 22 23 43 45 47 48 49 50 55 56 57 64 68 75 78 82

Analysis of Variance for the relationship between the number of drinks and the blood tests

The ANOVA Procedure

Number of Observations Read	276
Number of Observations Used	276

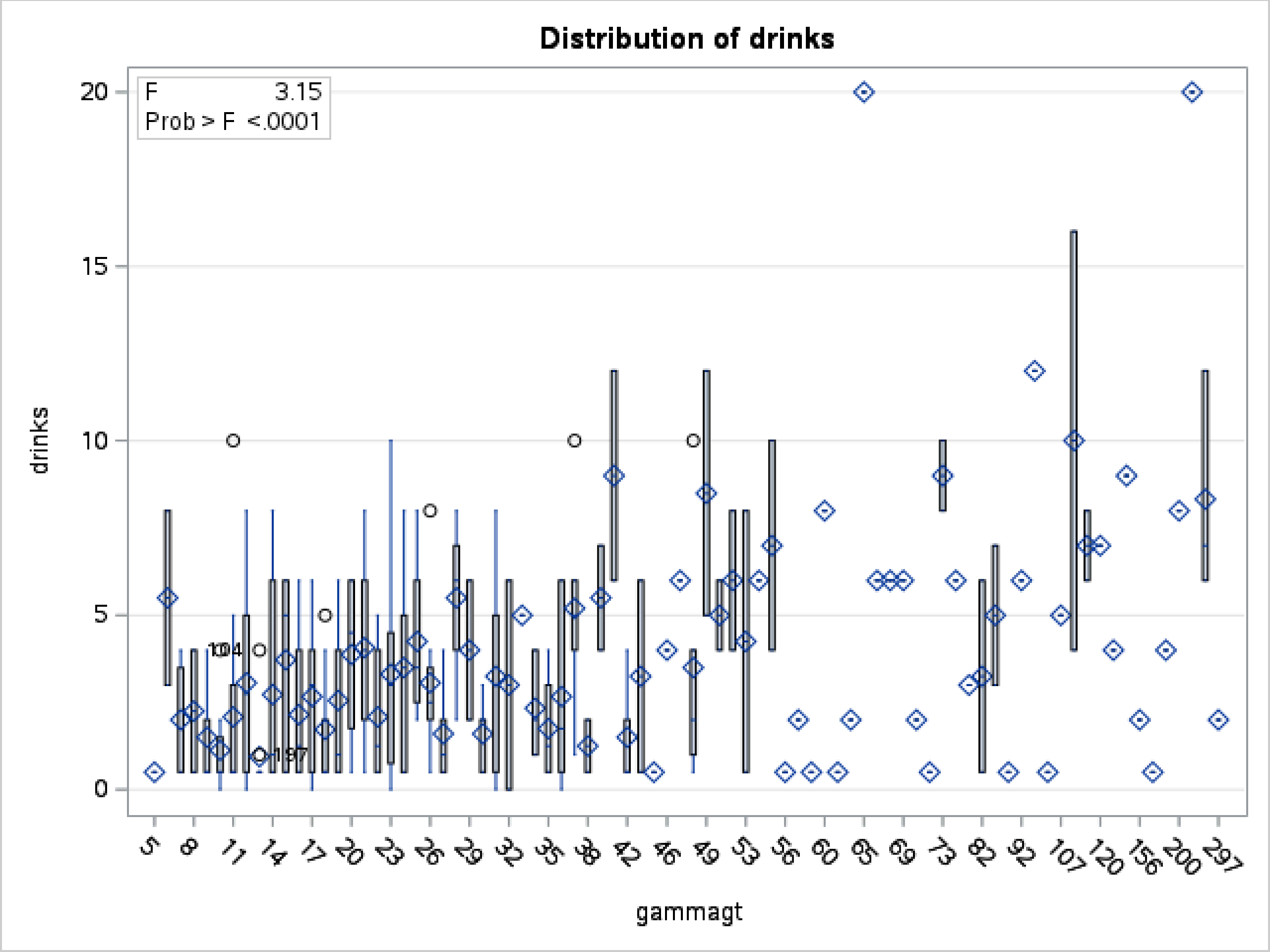
Gammagt

Analysis of Variance for the relationship between the number of drinks and Gammagt

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	81	1683.298431	20.781462	3.15	<.0001
Error	194	1281.437076	6.605346		
Corrected Total	275	2964.735507			

R-Square	Coeff Var	Root MSE	drinks Mean
0.567774	75.38193	2.570087	3.409420

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Gammagt	81	1683.298431	20.781462	3.15	<.0001


Distribution of drinks
F 3.15
Prob > F <.0001
drinks
20
15
10
5
0
104
197
5
8
11
14
17
20
23
26
29
32
35
38
42
46
49
53
56
60
65
69
73
82
92
107
120
156
200
297
gammagt

Tukey's Studentized Range (HSD) test for drinks

Note: This test controls the Type I experiment wise error rate.

Mcv

Alpha	0.05
Error Degrees of Freedom	194
Error Mean Square	6.605346
Critical Value of Studentized Range	6.07404

Analysis of Variance for the relationship between the number of drinks and Mcv

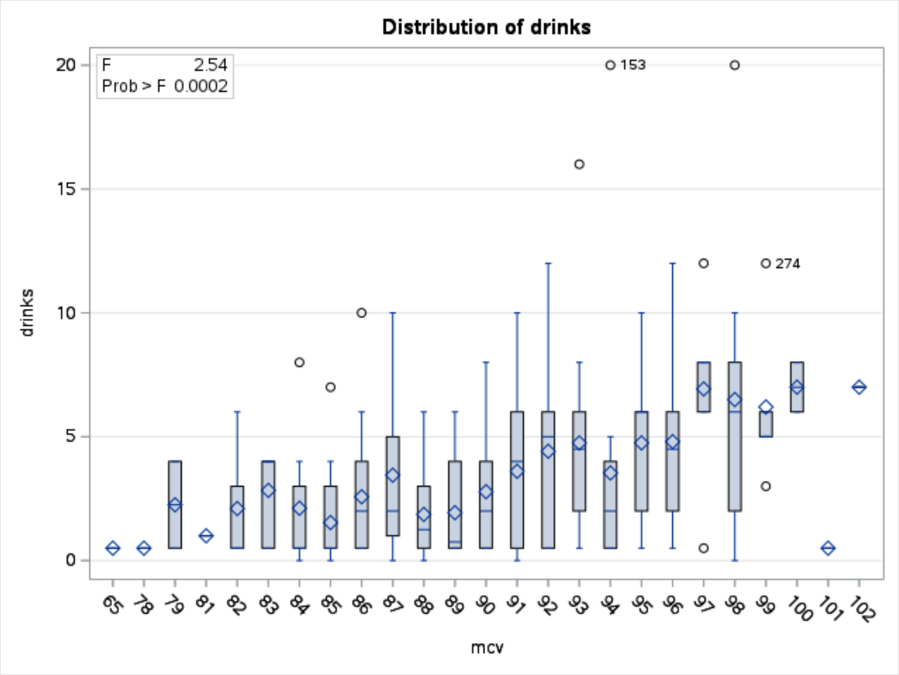
The ANOVA Procedure

Dependent Variable: drinks

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	24	579.187547	24.132814	2.54	0.0002
Error	251	2385.547960	9.504175		
Corrected Total	275	2964.735507			

R-Square	Coeff Var	Root MSE	drinks Mean
0.195359	90.42253	3.082884	3.409420

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Mcv	24	579.1875471	24.1328145	2.54	0.0002


Distribution of drinks
F 2.54
Prob > F 0.0002
153
274
drinks
20
15
10
5
0
65
78
79
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
mcv

Tukey's Studentized Range (HSD) Test for drinks Note: This test controls the Type I experiment wise error rate.

Alpha	0.05
Error Degrees of Freedom	251
Error Mean Square	9.504175
Critical Value of Studentized Range	5.23224

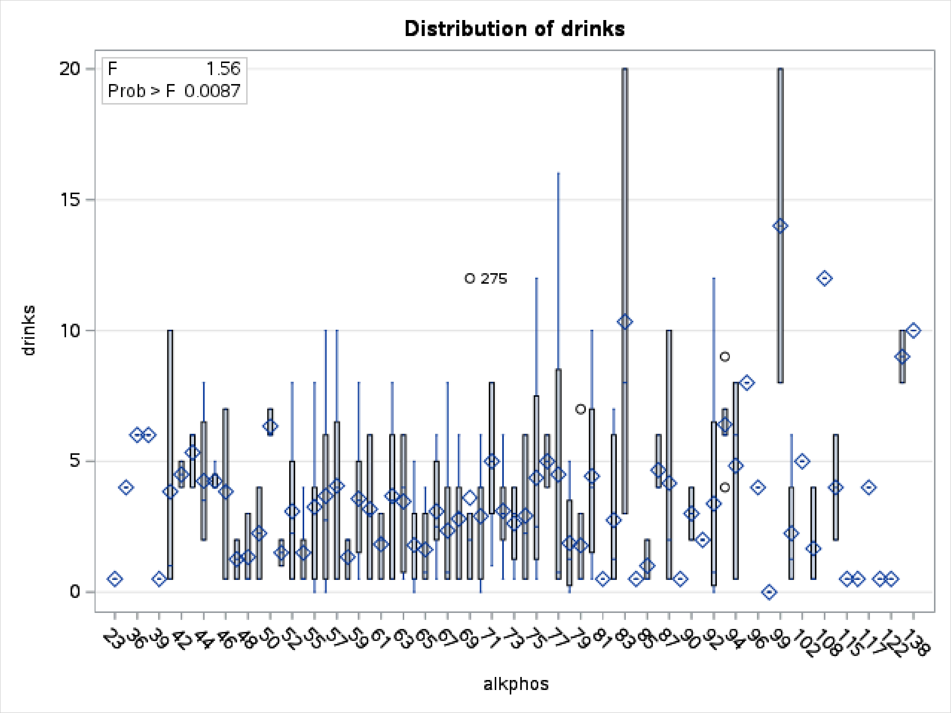
Alkphos

Analysis of Variance for the relationship between the number of drinks and alkphos

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	72	1054.204852	14.641734	1.56	0.0087
Error	203	1910.530655	9.411481		
Corrected Total	275	2964.735507			

R-Square	Coeff Var	Root MSE	drinks Mean
0.355581	89.98051	3.067814	3.409420

Source	DF	Anova SS	Mean Square	F Value	Pr > F
alkphos	72	1054.204852	14.641734	1.56	0.0087



Tukey's Studentized Range (HSD) Test for drinks Note: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	203
Error Mean Square	9.411481
Critical Value of Studentized Range	5.99275

Sgpt

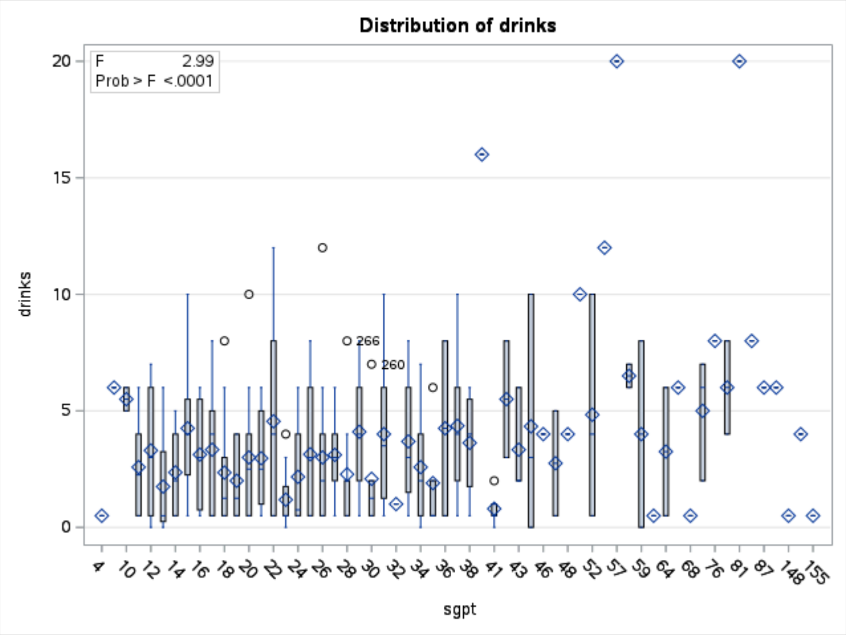
Analysis of Variance for the relationship between the number of drinks and sgpt

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	58	1316.161680	22.692443	2.99	<.0001
Error	21	1648.5738	7.597114		

	7	28			
Corrected Total	275	2964.735507			

R-Square	Coeff Var	Root MSE	drinks Mean
0.443939	80.84326	2.756286	3.409420

Source	DF	Anova SS	Mean Square	F Value	Pr > F
sgpt	58	1316.161680	22.692443	2.99	<.0001



Analysis of Variance for the relationship between the number of drinks and sgpt

The ANOVA Procedure

Tukey's Studentized Range (HSD) Test for drinks Note: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	217
Error Mean Square	7.597114
Critical Value of Studentized Range	5.84455

Sgot

Analysis of Variance for the relationship between the number of drinks and sgot

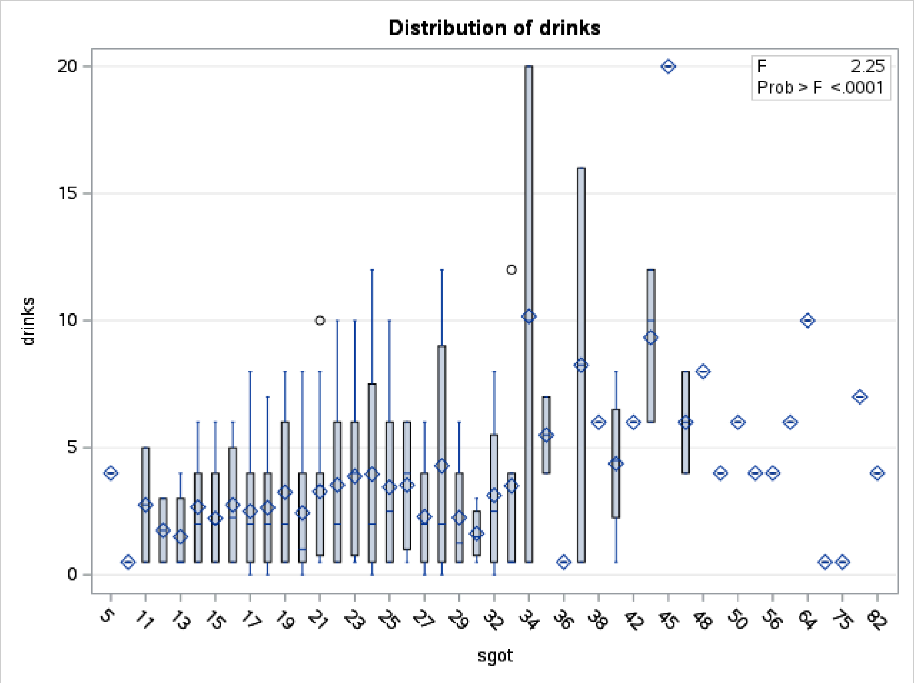
The ANOVA Procedue Dependent Variable: drinks

Source	D	Sum of	Mean	F	Pr >

	F	Squares	Square	Value	F
Model	44	888.471650	20.192538	2.25	<.0001
Error	231	2076.263857	8.988155		
Corrected Total	275	2964.735507			

R-Square	Coeff Var	Root MSE	drinks Mean
0.299680	87.93358	2.998025	3.409420

Source	DF	Anova SS	Mean Square	F Value	Pr > F
sgot	44	888.4716503	20.1925375	2.25	<.0001



Analysis of Variance for the relationship between the number of drinks and sgot

The ANOVA Procedure

Tukey's Studentized Range (HSD) Test for drinks
Note: This test controls the Type I experimentwise error rate.

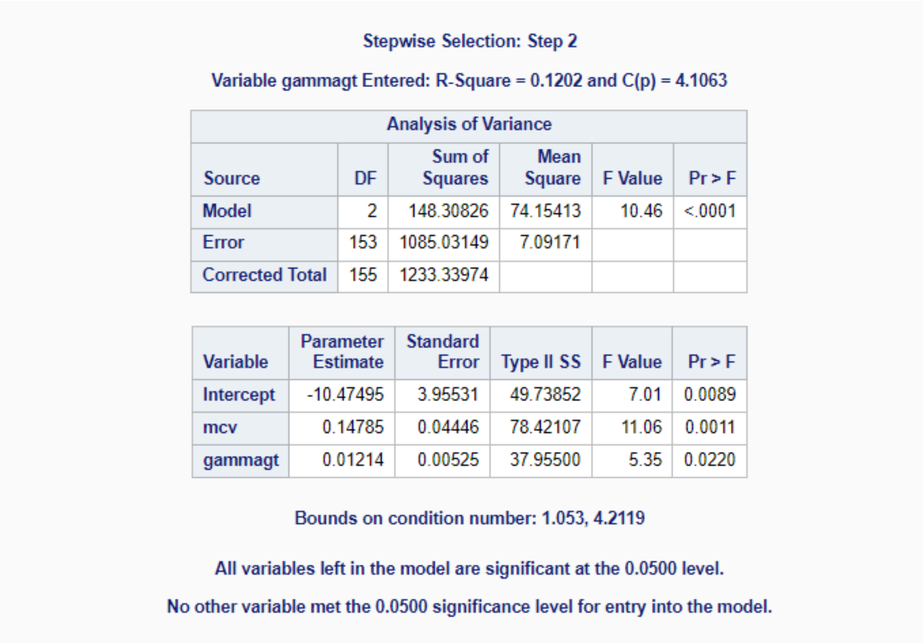
Alpha	0.05
Error Degrees of Freedom	231
Error Mean Square	8.988155
Critical Value of Studentized Range	5.65517

<u>SAS code for the Multiple linear regression model using the training data set(selector 2) and the testing data set(selector 1)</u>

```
LIBNAME mylib '/home/u63793201/My folder/SAS Project';

PROC PRINT DATA = mylib.group_6_train;

RUN;

PROC CONTENTS DATA = mylib.group_6_train;

RUN;

DATA mylib.test_data mylib.train_data;

   SET mylib.group_6_train;

   IF selector = 1 THEN OUTPUT mylib.test_data;

   ELSE IF selector = 2 THEN OUTPUT mylib.train_data;
```

```
RUN;
PROC FREQ DATA=mylib.group_6_train;
    TABLES selector;  /* To determine which selector to use to train the model, must use the one that is greater */
   TITLE "Frequency Count of Selector Values";
RUN;
ODS RTF FILE='/home/u63793201/My folder/SAS Project/results.rtf' STARTPAGE=NO STYLE= JOURNAL;
PROC REG DATA = mylib.train_data OUTEST=mylib.model_parameters;
        MODEL drinks = Mcv alkphos sgpt Gammagt/SELECTION=STEPWISE SLENTRY=0.05 SLSTAY=0.05;
        OUTPUT OUT=mylib.train_predictions;
        TITLE 'Training data(selector 2) multilinear regression model-Stepwise selection for training data with 0.05 Significance level';
RUN;
PROC REG DATA = mylib.test_data
                OUTEST=mylib.test_model_parameters;
                MODEL drinks = Mcv Gammagt;
                OUTPUT OUT=mylib.test_predictions;
        TITLE 'Testing Data(selector 1) Multilinear Regression Model';
RUN;
ODS RTF CLOSE;
```

Multiple Linear Regression Model of selector 1 and Model 1



Stepwise Selection: Step 2

Variable gammagt Entered: R-Square = 0.1202 and C(p) = 4.1063

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	148.30826	74.15413	10.46	<.0001
Error	153	1085.03149	7.09171		
Corrected Total	155	1233.33974			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-10.47495	3.95531	49.73852	7.01	0.0089
mcv	0.14785	0.04446	78.42107	11.06	0.0011
gammagt	0.01214	0.00525	37.95500	5.35	0.0220

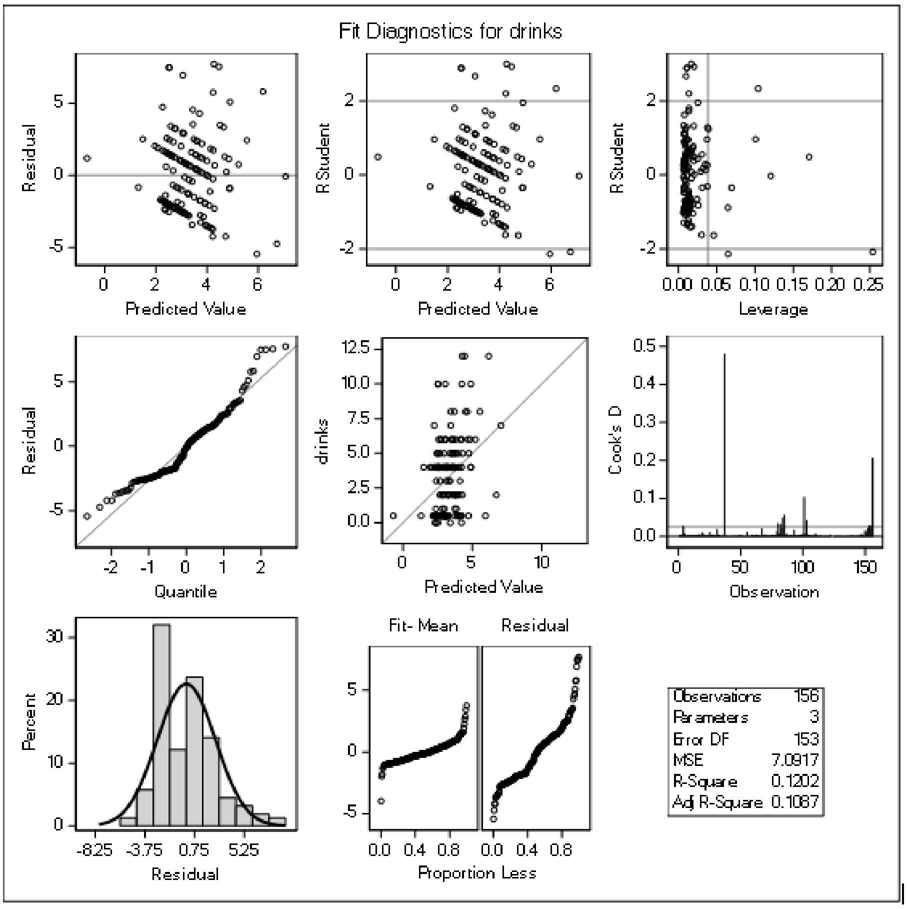
Bounds on condition number: 1.053, 4.2119

All variables left in the model are significant at the 0.0500 level.

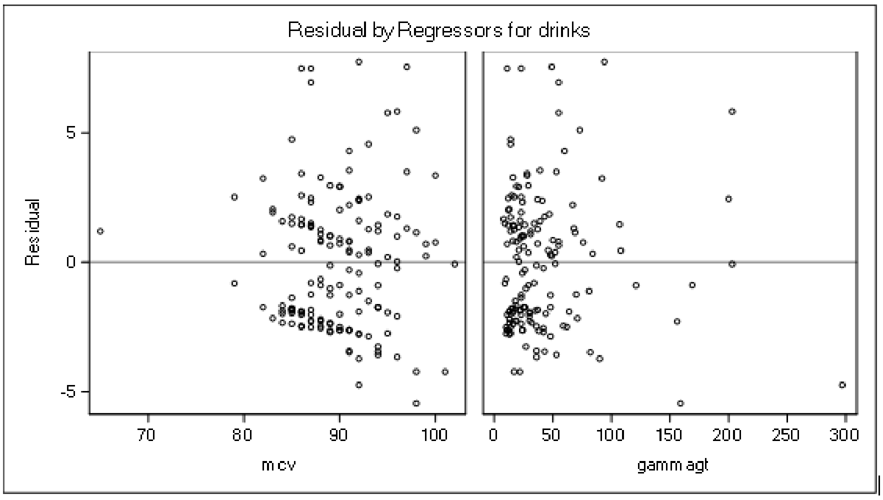
No other variable met the 0.0500 significance level for entry into the model.

Training data (selector 2) multilinear regression model-Stepwise selection for training data with 0.05 Significance level

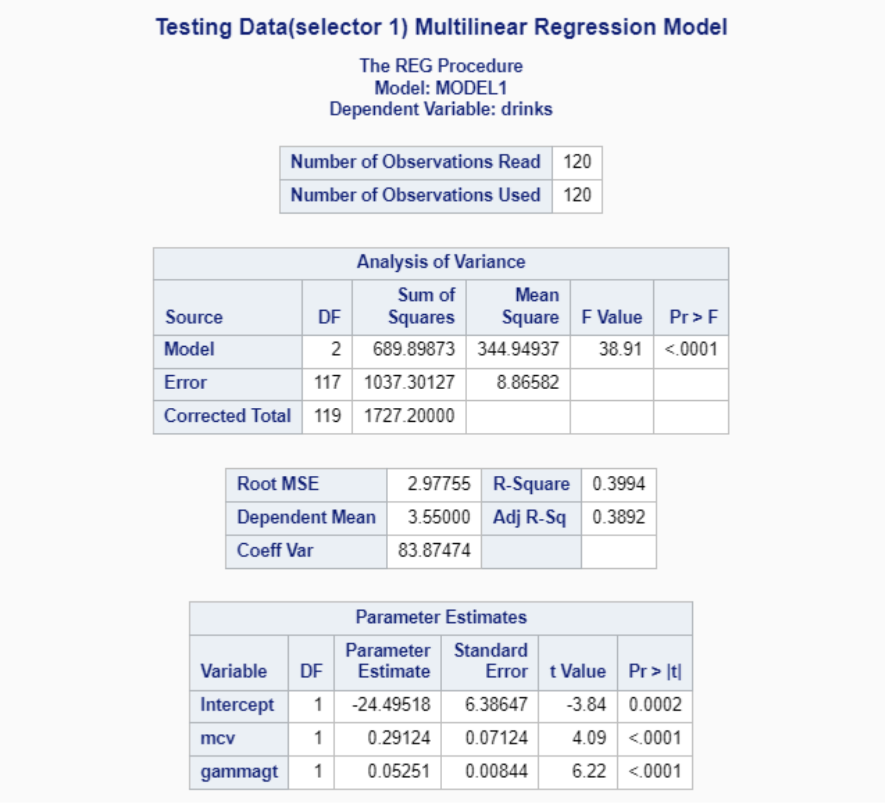
Model: MODEL1
Dependent Variable: drinks



Training data (selector 2) multilinear regression model-Stepwise selection for training data with 0.05 Significance level



Multiple Linear Regression Model of Testing data set Selector 1



Testing Data(selector 1) Multilinear Regression Model

The REG Procedure
Model: MODEL1
Dependent Variable: drinks

Number of Observations Read	120
Number of Observations Used	120

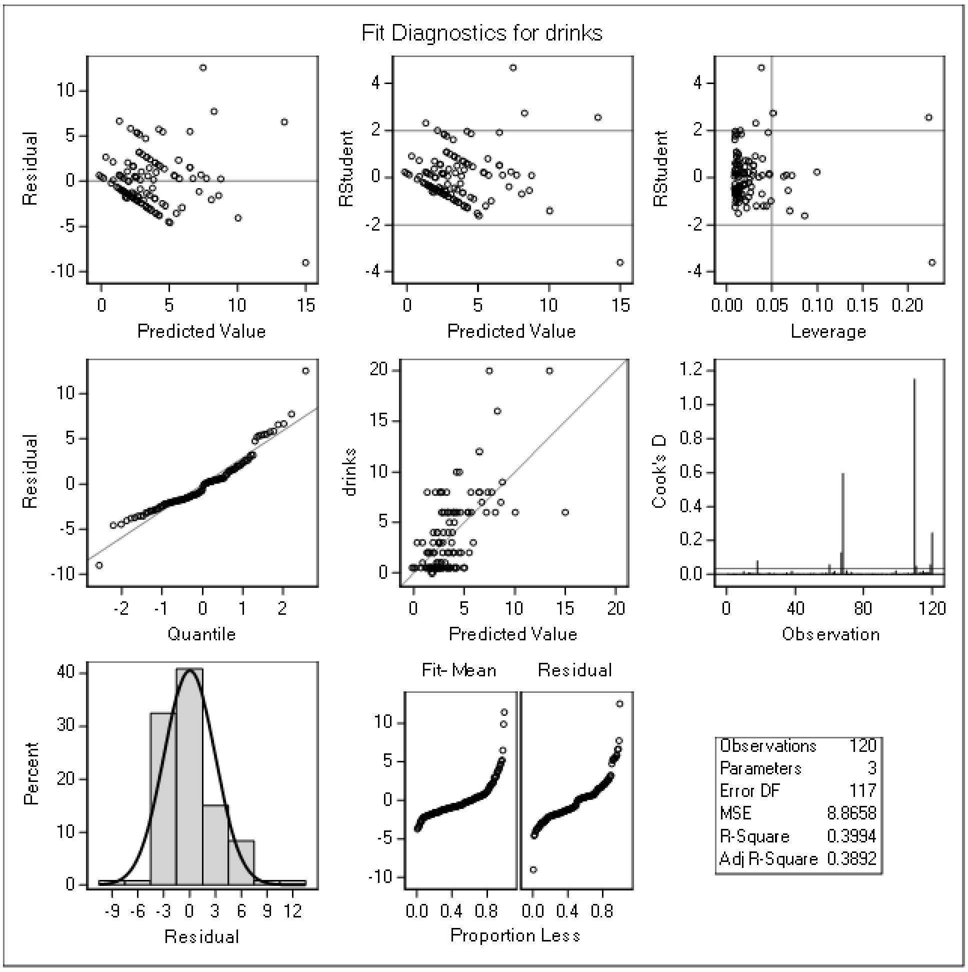
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	689.89873	344.94937	38.91	<.0001
Error	117	1037.30127	8.86582		
Corrected Total	119	1727.20000			

Root MSE	2.97755	R-Square	0.3994
Dependent Mean	3.55000	Adj R-Sq	0.3892
Coeff Var	83.87474		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-24.49518	6.38647	-3.84	0.0002
mcv	1	0.29124	0.07124	4.09	<.0001
gammagt	1	0.05251	0.00844	6.22	<.0001

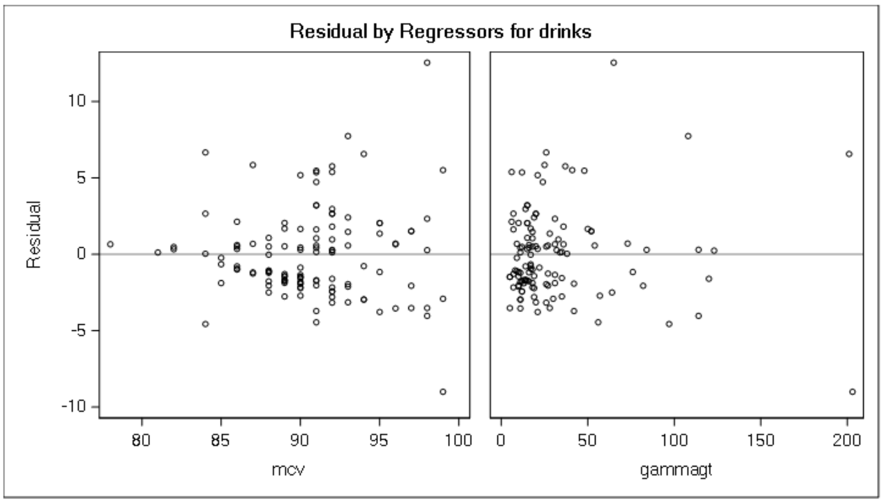
Model 2

Testing Data(selector 1) Multilinear Regression Model



Observations	120
Parameters	3
Error DF	117
MSE	8.8658
R-Square	0.3994
Adj R-Square	0.3892

Testing Data(selector 1) Multilinear Regression Model



Multiple Linear Regression Model (Reduced Model)

CODES USED TO PRODUCE THE MUTLIPLE LINEAR REGRESSION (REDUCED MODEL)

```
libname mylib '/home/u63312069/';

run;

ods html;

proc reg data="/home/u63312069/group_6_train (2).sas7bdat";

model drinks =Mcv alkphos sgpt sgot Gammagt;

Title "blood test analysis using proc reg”.

run;

ods html close;

ods html;

proc reg data="/home/u63312069/group_6_train (2).sas7bdat";

model drinks =Mcv alkphos sgpt sgot Gammagt/

selection=forward

slentry=0.05;

Title "Producing Forward Selection for Reduced Model";

run;

ods html close;

ods html;

proc reg data="/home/u63312069/group_6_train (2).sas7bdat";

model drinks =Mcv alkphos sgpt sgot Gammagt/

selection=backward

slentry=0.10;

run;

ods html close;
```

blood test analysis using proc reg

The REG Procedure

Model: MODEL1

Dependent Variable: drinks

Number of Observations Read	276
Number of Observations Used	276

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	648.90854	129.78171	15.13	<.0001
Error	270	2315.82696	8.57714		
Corrected Total	275	2964.73551			

Root MSE	2.92867	R-Square	0.2189
Dependent Mean	3.40942	Adj R-Sq	0.2044
Coeff Var	85.89950		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-16.91936	3.61929	-4.67	<.0001
Mcv	1	0.20457	0.03987	5.13	<.0001
alkphos	1	0.00999	0.00974	1.03	0.3062

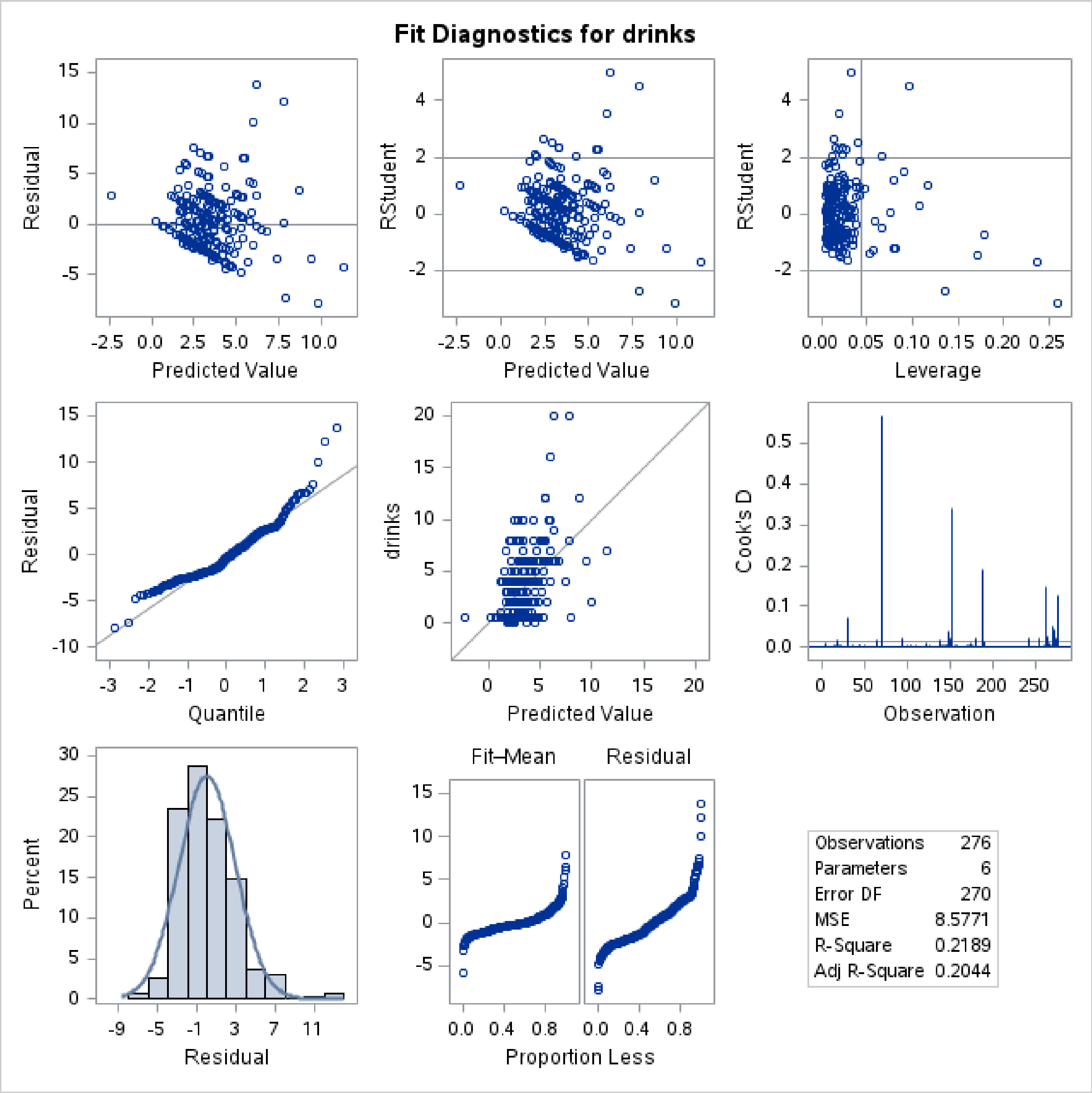
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
sgpt	1	-0.01258	0.01356	-0.93	0.3543
sgot	1	0.02862	0.02656	1.08	0.2822
Gammagt	1	0.02343	0.00575	4.07	<.0001

blood test analysis using proc reg

The REG Procedure

Model: MODEL1

Dependent Variable: drinks



Fit Diagnostics for drinks

Observations	276
Parameters	6
Error DF	270
MSE	8.5771
R-Square	0.2189
Adj R-Square	0.2044

Producing Forward Selection for Reduced Model

The REG Procedure

Model: MODEL1

Dependent Variable: drinks

Number of Observations Read	276
Number of Observations Used	276

Forward Selection: Step 1

Variable Gammagt Entered: R-Square = 0.1333 and C(p) = 27.5647

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	395.32814	395.32814	42.16	<.0001
Error	274	2569.40737	9.37740		
Corrected Total	275	2964.73551			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	2.26799	0.25472	743.45140	79.28	<.0001
Gammagt	0.03089	0.00476	395.32814	42.16	<.0001

Bounds on condition number: 1, 1

Forward Selection: Step 2

Variable Mcv Entered: R-Square = 0.2114 and C(p) = 2.5897

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	626.69639	313.34819	36.59	<.0001
Error	273	2338.03912	8.56425		
Corrected Total	275	2964.73551			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-16.13699	3.54937	177.02328	20.67	<.0001
Mcv	0.20650	0.03973	231.36825	27.02	<.0001
Gammagt	0.02515	0.00468	247.36678	28.88	<.0001

Bounds on condition number: 1.0591, 4.2363

No other variable met the 0.0500 significance level for entry into the model.

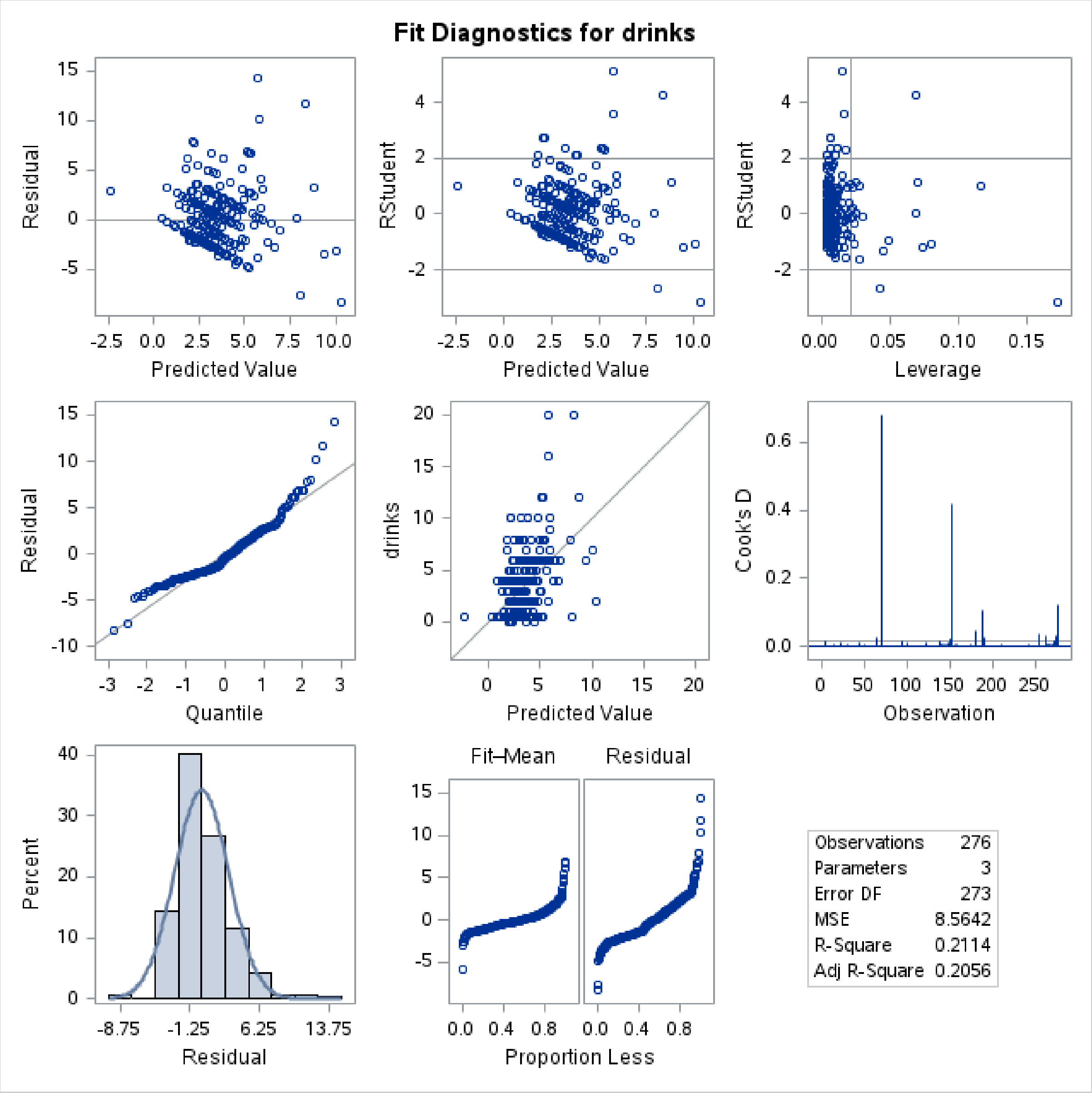
Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Gammagt	1	0.1333	0.1333	27.5647	42.16	<.0001
2	Mcv	2	0.0780	0.2114	2.5897	27.02	<.0001

Producing Forward Selection for Reduced Model

The REG Procedure

Model: MODEL1

Dependent Variable: drinks



Fit Diagnostics for drinks

Observations	276
Parameters	3
Error DF	273
MSE	8.5642
R-Square	0.2114
Adj R-Square	0.2056

Producing Forward Selection for Reduced Model

The REG Procedure

Model: MODEL1

Dependent Variable: drinks

Number of Observations Read	276
Number of Observations Used	276

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.2189 and C(p) = 6.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	648.90854	129.78171	15.13	<.0001
Error	270	2315.82696	8.57714		
Corrected Total	275	2964.73551			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-16.91936	3.61929	187.44034	21.85	<.0001
Mcv	0.20457	0.03987	225.82508	26.33	<.0001
alkphos	0.00999	0.00974	9.01495	1.05	0.3062
sgpt	-0.01258	0.01356	7.38320	0.86	0.3543
sgot	0.02862	0.02656	9.96022	1.16	0.2822
Gammagt	0.02343	0.00575	142.24689	16.58	<.0001

Bounds on condition number: 2.5247, 42.804

Backward Elimination: Step 1

Variable sgpt Removed: R-Square = 0.2164 and C(p) = 4.8608

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	641.52534	160.38133	18.71	<.0001
Error	271	2323.21017	8.57273		
Corrected Total	275	2964.73551			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-16.97308	3.61790	188.68072	22.01	<.0001
Mcv	0.20521	0.03985	227.31658	26.52	<.0001
alkphos	0.01051	0.00973	10.00956	1.17	0.2809
sgot	0.01302	0.02056	3.44067	0.40	0.5269
Gammagt	0.02237	0.00564	134.99421	15.75	<.0001

Bounds on condition number: 1.5357, 20.598

Backward Elimination: Step 2

Variable sgot Removed: R-Square = 0.2152 and C(p) = 3.2619

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	638.08467	212.69489	24.87	<.0001

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Error	272	2326.65083	8.55386		
Corrected Total	275	2964.73551			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-16.93065	3.61330	187.80293	21.96	<.0001
Mcv	0.20700	0.03971	232.46561	27.18	<.0001
alkphos	0.01115	0.00966	11.38829	1.33	0.2496
Gammagt	0.02431	0.00473	225.57089	26.37	<.0001

Bounds on condition number: 1.0849, 9.5076

Backward Elimination: Step 3

Variable alkphos Removed: R-Square = 0.2114 and C(p) = 2.5897

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	626.69639	313.34819	36.59	<.0001
Error	273	2338.03912	8.56425		
Corrected Total	275	2964.73551			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-16.13699	3.54937	177.02328	20.67	<.0001
Mcv	0.20650	0.03973	231.36825	27.02	<.0001
Gammagt	0.02515	0.00468	247.36678	28.88	<.0001

Bounds on condition number: 1.0591, 4.2363

All variables left in the model are significant at the 0.1000 level.

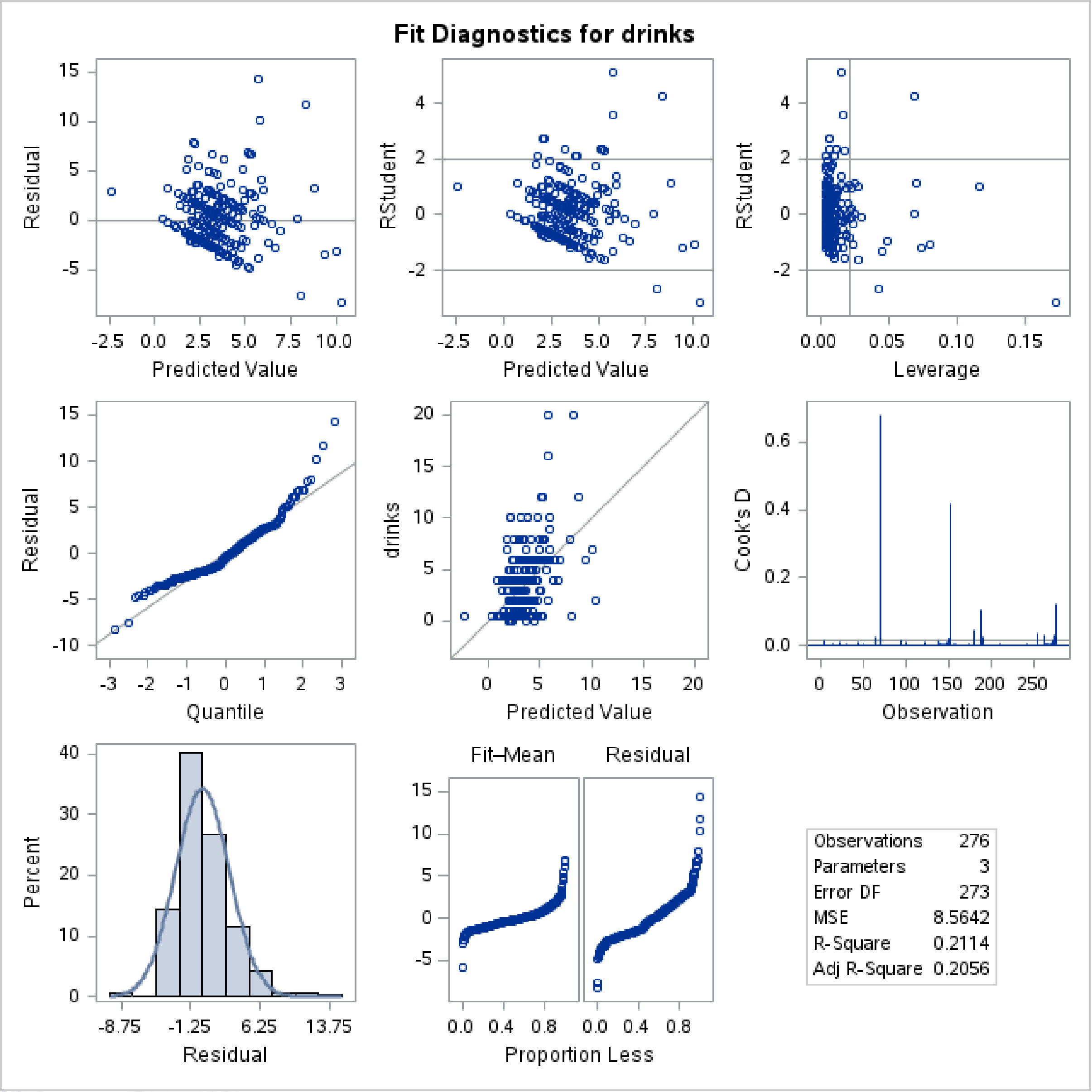
Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	sgpt	4	0.0025	0.2164	4.8608	0.86	0.3543
2	sgot	3	0.0012	0.2152	3.2619	0.40	0.5269
3	alkphos	2	0.0038	0.2114	2.5897	1.33	0.2496

Producing Forward Selection for Reduced Model

The REG Procedure

Model: MODEL1

Dependent Variable: drinks



Observations	276
Parameters	3
Error DF	273
MSE	8.5642
R-Square	0.2114
Adj R-Square	0.2056