University of the Western Cape

Blood Tests That Effectively Predict Excessive Alcohol Consumption in Patients with Liver Disorders.

A Report submitted in fulfillment of the requirements for the STA221 module GROUP 12

By:

Bennedict Mkhonto 4226279

Griffiths Moshoeshoe 4314422

Liyema Mngxekeza 4133533

Thaakirah Mosoval 4314422

Elethu Molose 4366125

Ayabulela Carol 4385096

Supervisor:

Mr. Matthew Wayne Valentine

DUE: 27 September 2024

# 

[**Abstract 3**](#_heading=h.3y2ehuhv8ndo)

[**Research Questions 3**](#_heading=h.l5e6n7w03pdz)

[**Literature Review & Introduction 4**](#_heading=h.jskynxfnafii)

[Introduction 4](#_heading=h.yg27jfqwl6d3)

[**Methodology 6**](#_heading=h.b3ai4ocsqulr)

[**Analysis Section of Research Report 9**](#_heading=h.22tpg5f0r5e5)

[**Appendix 19**](#_heading=h.aq6klpv9bcms)

[**Timeline Plan 19**](#_heading=h.wmmaekqkdcg3)

[Regression analysis 22](#_heading=h.i9w7k4ffbrk7)

[**Bibliography 49**](#_heading=h.aczhq1hsyfru)

# 

# Abstract

Predicting high alcohol intake among patients living with liver disorders is important for better diagnosis and care because it can help prevent more liver damage and consequences, including cirrhosis and alcoholic liver disease (ALD). The research explores the predictive role of blood test results, specifically GGT levels, in predicting more alcohol intake, using machine learning techniques to classify liver disorders. The primary research question is, how can blood test results be utilized to predict excessive alcohol consumption in patients with liver disorders? The proposed research question is, how does alcohol consumption influence the level of gamma-glutamyl transferase (GGT) in individuals with liver disorders? The research findings aim to improve early detection of alcohol-related liver disorders, enabling timely actions and contributing to public health strategies to reduce alcohol-related risks, offering a tool that could improve clinical intervention strategies and public health efforts. In basic, this research provides new ideas into the connection between alcohol use and liver function, with the potential to reduce the levels of liver disorders.

# Research Questions

**Main Research Question:**

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

**Proposed Research Questions:**

* How does consuming alcohol affect the level of gamma-glutamyl transferase (GGT) among individuals with liver disorders?
* Which blood test variable is the best predictor of excessive alcohol consumption in patients with liver disorder among the ones provided?

# Literature Review & Introduction

## Introduction

Excessive consumption of alcohol is a worldwide health concern that usually affects individuals suffering from liver disorder. Long-term alcohol consumption eventually harms the liver, which is responsible for metabolizing alcohol, and usually leads to illnesses. Blood tests, particularly those that measure liver enzymes like gamma-glutamyl transferase (GGT), provide a non-invasive way to track alcohol intake of alcohol consumers and its effects on liver health. This literature review will mainly focus on how blood test results, especially GGT, can be utilized to predict excessive alcohol consumption in patients who have liver disorders. Studies show that GGT is an effective biomarker for identifying liver damage caused by alcohol use. Nevertheless, its accuracy is restricted, as increased levels can also be caused by different liver conditions and factors such as metabolic syndrome. Research indicates a clear connection between high alcohol intake and GGT levels, which are usually more sensitive to alcohol consumption compared to liver enzymes such as SGOT and SGPT. Disagreements in research happen because of differences in methods and personal characteristics like age, gender, and alcohol consumption history, impacting the accuracy of GGT as a predictor. Analysis focusing on GGT as a predictor of alcohol consumption will be conducted in this literature review, evaluating important results and methods used. This will better our understanding of GGT's function and highlight ways to enhance its predictability in managing liver disorders.

The current research lacks comprehensive studies that combine GGT with other biomarkers and address issues in predicting alcohol use in individuals who do not have a history of drinking. The efficacy of GGT as a useful indicator can be affected by different variables. This review examines the advantages and drawbacks of GGT, evaluates present research patterns, and proposes enhancements for its clinical application.

**The relationship between the four blood tests GGT, SGOT, SGPT, MCV and Alcohol intake in patients with liver diseases**

Is GGT a good predictor of excessive alcohol in patients with liver disorders?

When people drink a lot of alcohol, their GGT levels usually go up, However, this connection is not straightforward and can be influenced by factors like age, gender, and how much alcohol someone has consumed in the past. The precise process of how GGT is elevated remains unknown. Enzymes may be released due to hepatic cell damage or activation following alcohol exposure. In alcoholic liver disease, part of the rise in levels can be attributed to both hepatocyte cholestasis and hepatocyte injury. It rises following five weeks of consuming over 50g daily. It typically rises to three times the maximum reference level (Rosman AS, 1992)

Some people who misuse alcohol do not always have high GGT levels, and in some cases, chronic alcohol abusers see their initially elevated GGT levels decrease even with continued drinking. Incorrect results can be observed in conditions such as non-alcoholic liver diseases, obesity, diabetes, and heart failure. The fluctuating responsiveness of the marker makes it unsuitable for standalone alcohol testing in people with liver disorders. The pattern of alcohol abuse differs significantly from that of healthy volunteers and non-alcoholic liver disease (Sillanaukee P, 1996)

Are SGOT, SGPT good predictors of excessive alcohol intake in patients with liver disorders?

Both SGOT and SGOT are found in hepatocytes, but SGOT is found in skeletal and myocardial cells. In alcohol related liver damage, the SGOT is triggered more than SGPT. This is the reverse of the normal pattern in acute hepatocellular disease where the SGPT exceeds the SGOT. Incorrect results are found in non-alcoholic liver disease, muscle damage and myocardial damage. Despite these results, the specificity is high at 90% or more (Conigrave et al.1995).

Is MCV a good predictor of excessive alcohol in patients with liver disorders?

It rises following six weeks of alcohol misuse but in view of the half-life, it can remain triggered for up to three months and thus it has a limited use in monitoring alcohol intake. Its sensitiveness is higher in women (86.3%) than in men (63.0%) (Morgan et al. 1981). Incorrect results are found in hypothyroidism, Vitamin B12, non- alcoholic liver disease and in some patients who smoke (Sillanaukee P, 1996).

GGT association with alcohol in patients with liver disorder

Levels of GGT rise in response to alcohol consumption up until a certain extent, the response is different individuals based on their drinking history. GGT levels correlate at a moderate level with alcohol consumption (r=0.30-0.40) in men while in women r=0.15-0.30). GGT does not respond to a single dose of alcohol consumption unless you have an excessive alcohol consumption history. In experiment with volunteers, 60 g of ethanol daily for 3 weeks produced more than a 15% increase in enzyme levels while 5 weeks produced almost doubling of mean levels in young volunteers from 27 U/I to 52 U/I, this proves that a person with an excessive history, the GGT is very sensitive on those people (Sillanaukee P, 1996).

AST and ALT association with alcohol in patients with liver disorder

Just like GGT, AST and ALT are not increased by a single episode of excessive alcohol consumption, a study of eight young healthy male volunteers was conducted to see the sensitiveness of AST and ALT on alcohol consumption, consumption of 60g excessive alcohol per day for 5 weeks produced a slightly rise in AST and ALT. A longer duration of drinking will increase the sensitiveness of ALT and AST. While AST values correlate highly with GGT values (r=0.61-0.68), they do not correlate as highly with alcohol consumption (r=0.24-0.34) (Sillanaukee P, 1996).

**Tests that were done to prove which test is the best when predicting excessive alcohol in patients with liver disorders:**

Blood samples were collected and medical history, including alcohol intake history and laboratory results, from 1578 patients that either have the liver cirrhosis due to alcohol or with similar alcohol intake but no signs of any liver disease (Whitfield et al. 2015). Whitfield made it clear that the data will help them determine which test is the best at distinguishing between those who do not have cirrhosis and those who have it, because of long-term excessive alcohol intake? Another reason is to determine which test is the best in identifying continuing alcohol use among patients with liver disorder? (Whitfield et al. 2015

Comparison of two means test and statistical analysis

Comparison of two means tests was performed between the groups, test results showed a positive skewed distribution (ALT, AST, and GGT). For ROC curve analysis, test results where the control mean was higher that the case mean had the assumption of higher results indicating abnormalities. Statistical analyses were performed using SPSS (Kamath and Kim, 2007)

The test means for patients with excessive drinking history and those with liver disorder shown that only AST and GGT have shown the significant effect of abstinence. These two tests also showed significant control by abstinent or non-abstinent. The result came out and showed that abstinence was associated with lower AST and GGT in cases but not in controls (Whitfield et al, 2015).

In conclusion, GGT, SGOT, SGPT and MCV have shown problems in detection particularly in patience with, with GGT having a moderate correlation with the excessive drinking alcohol in patients with liver disorders, it becomes the best when we compare it to the other test.Gender predisposition, age and gender are all

factors that could influence the outcomes of these tests, leading the

particularity and sensitivity to differ between populations. According to Sanjiv,

Victor and Harris, the results showed that alcohol intake also the activities of

SGOT & GGT were comparable in both genders, but the connection between

alcohol consumption and activities of SGOT & SGPT was mainly influenced

by males- they also discovered that males were higher for all markers. The

informative effects of age were that as age increased, SGPT reduced and

SGOT & GGT increased. Although the three blood tests, namely GGT,SGOT

and SGPT can detect alcohol consumption while MCV is the blood test

we are not sure about if it can detect excessive alcohol or not, with GGT being

the very sensitive test, they all have limitations, some of those limitations are

age, gender and the period of intake of alcohol. With these results we can

clearly see that GGT is the best blood test to use when testing for excessive

alcohol consumption in patients with liver disorders.

# 

# Methodology

**Introduction**

The methodology section of this research report outlines the specific analytical methods that will be utilized in analyzing the 5 blood tests that are thought to be sensitive to liver disorders. The main research question at hand is to study **how blood test results can predict excessive alcohol consumption in patients with liver disorders**. The research further addresses two proposed questions on **how consuming alcohol affects the levels of gamma-glutamyl transferase (GGT) among individuals with liver disorders** and **Which blood test variable is the best predictor of excessive alcohol consumption in these patients.** To efficiently address the research question at hand, a combination of descriptive analysis, correlation analysis, and multiple regression modeling will be employed in the research.

Descriptive analysis will be utilized to examine the distribution of blood test variables, providing a clear overview of the data, while correlation analysis will assess the strength and direction of relationships between alcohol consumption and specific markers like gamma-glutamyl transferase (GGT). Furthermore, multiple regression modeling will be applied to identify the best predictors of excessive alcohol consumption among the 5 provided blood tests. By the effective use of these methods, this report will guarantee accurate findings that will contribute to a comprehensive understanding and a well-informed prediction of which blood test a doctor should take into account when diagnosing alcohol-related liver disorders.

**Descriptive analysis**

We are going to use the proc univariate procedure to determine the type of distribution of our data which is the most crucial part of analyzing data since different statistical tests and models assume specific distribution types, so it is important to check if it is normally distributed or not, we are going to assess the skewness of the and the shape of the distribution from normality using the proc univariate procedure and lastly it is very difficult to analyze raw data, so we are going to use print procedure to make it easy for us to read the raw data and understanding the patterns or relationships between variables.

We are going to perform the proc means procedure on SAS which will create a tubular or a graphical summary of the data which will help us with a better understanding of the data, this procedure will also provide us with the mean, std, min, max, etc. Our variable to analyze will be alcohol and we will categorize the data using the selector. We are going to use the proc univariate procedure in SAS which will use to create a plot that we help us observe whether the data is approximately normally distributed in the form of a histogram, again our dependent variable or a variable to analyze will be alcohol and we will categorize our data using the selector.

6

**Correlation analysis**

The measure of the linear relationship between two quantitative variables on the same subject is what the proc correlation procedure is all about, and our goal for this research is to determine the relationship between the tests and the alcohol consumed, so this procedure will help us identify how strong the relationship is between the two variables if there is a relationship between them. The correlation p ranges from (-1 to 1) where p=1 and p=1, are perfect positive correlation and perfect negative correlation. p=o indicates that there is no linear relationship.

We are going to use hypothesis testing to prove the correlation, where the null hypothesis will be that there is no linear relationship between the two variables, and the alternative hypothesis will oppose the null hypothesis and assume that there is a linear relationship between the two variables. So we are going to do the proc correlation on SAS, the proc correlation procedure is divided into two parts the Spearman correlation procedure which we use in a distribution that is not normally distributed, and the Pearson correlation procedure which is used when the data is assumed to be normally distributed and this is where the proc univariate procedure plays a role since we are using it to see if the data is normally distributed or not.

We must identify the dependent variable (alcohol), the categorical variable(s)/grouping variable(s) which in this case is the selector, we also need to identify the WITH variable which obtains correlation between the dependent variables and the WITH variable. It is also a good idea to plot a graphical display which will also help us visualize the relationship of the two variables, we have options of graphing a scatter plot, histogram, or matrix. So on the results of SAS, we are going to analyze the p-value, if the p-value is less than 0.05 we reject the null hypothesis and say that the is a linear relationship between the two variables in this case the blood test and the alcohol, if the p-value is greater than 0.05 we do not reject the null hypothesis we say that the is a linear relationship between the blood test and the alcohol.

**Multiple Regression Model**

The measure of linear relationship between two or more variables is obtained using the regression procedure. In this case, since we are comparing alcohol and the blood tests, we are comparing the measure of the linear relationship

between 1 dependent variable with multiple independent variables. The goal of doing the multiple linear regression model is to know if there is any relationship between the alcohol level and the blood tests and to find out which blood test is

the most significant in determining the alcohol level using the selection techniques.

In this procedure we are going to use a hypothesis test to determine our linear relationship if there is any, the null hypothesis will assume that beta which is the independent variable, in this case, which is the blood test, is equal to zero, meaning that the independent variable is not included in our regression model,there is no predictive linear relationship between the dependent (alcohol) and the independent variable (blood test). The alternative hypothesis will oppose the null hypothesis and assume that beta is not equal to 0, the blood test is included in the linear regression model, this is a predictive linear association between the blood level and alcohol.

So, we are going to perform the proc regression procedure on SAS to determine whether we need to determine all the independent variables (all the tests) and compare them with the dependent variable (alcohol level). After writing our code, we will run it and analyze the results. So, on the results we are going to look at the p-value for each blood test, if the p-value is less than 0.05 then we reject the null hypothesis, instead, we accept the alternative hypothesis, and that means that there is a predictive linear relationship between that test and the alcohol level, the blood test is included in the regression model, in other words, beta is not equal to zero, but if the p-value is greater than 0.05 we do not reject the null hypothesis instead we accept it, which means that beta is equal to zero and that proves that there is no predictive linear relationship between the blood test and the alcohol, thus the test will not be included in the regression model.

Another value that we will analyze is the R-square value which is the measure of the distinct association between alcohol (dependent variable) and the blood test (independent variable). We will convert it to percentage, and it will give us the level of variability between the alcohol and the blood test. So, our regression model formula is given by y=b0+b1x+b2x+b3x+b4x where b is beta and bxi represents the blood tests depending on whether the blood test is included in the regression model or not.

So when it comes to determining the significant variable, we have to perform a selection technique which is the backward selection technique, this is a technique that starts with the full modeling containing all the independent variables and then removes the least significant variable until we reach the one most significant independent variable step by step, and this is done using the p-value, the variable with the biggest p-value is the least significant variable and the one with the lowest p-value is the most significant variable. This technique will help us answer our research question: Which blood test variable is the best predictor of excessive alcohol consumption in patients with liver disorder among the ones provided? By finding the most significant variable, we found the best predictor of excessive alcohol consumption in patients with liver disorders.

In conclusion, to determine how alcohol consumption affects the Gamma Gt, mcv, alkphos, and the Sgot among individuals with liver disorders, we will use the following procedures which include the univariate procedure which will help us determine whether the distribution is normal or not and will help us determine the skewness of the distribution, the correlation procedure will provide us with the measure of the linear relationship between the level of alcohol and the blood test. Lastly, we will perform the multiple linear regression procedure which will help us in determining the linear relationship between the level of alcohol and the blood test, also it will provide us with the measure of variability between the alcohol level and the blood test. These procedures will

help us in finding the relationship between alcohol and the flour blood tests and by doing this we will be able to answer our research questions, how does the level of alcohol consumption affect the Gammagt, MCV, alkphos, and the Sgot among individuals with liver disorders, we are going to use the backward selection in the regression procedure to determine which blood test has a higher relationship with the alcohol consumed which is the main goal of this research.

**Selection Techniques**

Besides stepwise regression, various other selection methods will be utilized to make sure that the most important blood test variables are incorporated in the final model, thus improving the regression analysis. The process of backward elimination starts with including all blood test variables in the model and then removing the least significant variables one at a time until the final model is achieved. At each iteration, the variable with the greatest p-value, which shows the least impact on the model, is removed until only variables with significance remain. This method assists in creating a concise model that emphasizes key predictors and is especially useful for pinpointing variables that have significant effects on the outcome variable (Miller, 2002). On the contrary, backward selection begins with all variables in the model and removes the least statistically significant one at a time. This approach guarantees that only blood markers that are highly relevant are included in the regression analysis. Moreover, the best subset selection method examines every possible variable combination to choose the one that most accurately aligns with the data, utilizing criteria like the AIC or BIC. This method is particularly advantageous for dealing with a large number of variables, as it determines the best combination of predictors for alcohol consumption.

Through the utilization of these selection methods, the analysis will be strong and productive, guaranteeing that the ultimate model includes the most important blood test factors in forecasting alcohol intake. Utilizing both backward and forward selection techniques helps in retaining important predictors and eliminating unnecessary variables to achieve the research objective of identifying the most effective indicators of excessive alcohol consumption among individuals with liver disorders. Implementing these methods guarantees a sophisticated, evidence-based strategy, as highlighted by Burnham and Anderson (2002), allowing for the recognition of crucial predictors while upholding model simplicity and precision.

# Analysis Section of Research Report

**Introduction**

**Research Overview**

This study looks into the ability of blood tests to anticipate high levels of alcohol intake in individuals with liver illnesses. The research uses the UCI BUPA Liver Disorder dataset to analyze the relationship between alcohol intake and specific blood test variables. Our goal is to provide valuable insights into clinical diagnostics by exploring how blood test results can predict excessive alcohol consumption in patients with liver disorders. Frequent alcohol intake can worsen liver conditions, and identifying them early using blood markers can enhance patient outcomes and care.

**Created Research Question**

To delve further into the inquiry, a research question was formulated: "What effect does alcohol intake have on the GGT levels in individuals with liver conditions?" This particular inquiry centers on GGT, an enzyme frequently utilized as an indicator of liver injury, specifically from alcohol. By exploring this issue, we can comprehend how alcohol intake directly affects GGT levels and how it relates to other blood test factors, enhancing the development of diagnostic methods for liver ailments. We also took into consideration all other variables to come up with which variable could be the best predictor.

**Aims and Objectives**

The main goal of this study is to determine which blood test factors are most effective in foreseeing high levels of alcohol consumption in individuals with liver conditions. The main goals consist of Evaluating the impact of alcohol intake on GGT levels and other enzymes related to the liver. Identifying the blood test variables that show the strongest correlation with excessive alcohol consumption.Utilizing statistical techniques like regression models in order to establish a dependable predictive structure.

**Structure of Analysis**

This study is organized to investigate the connection between crucial blood test indicators (MCV, Alkphos, SGPT, SGOT, GGT) and alcohol intake. Every part is intended to systematically assess the interaction of these factors, beginning with Descriptive Statistics to offer a summary of the dataset. This part will provide an overview of the fundamental statistical characteristics of the variables, including the average, variability, and spread, giving a broad insight into the distribution of the data. Following that, Correlation Analysis will evaluate the degree and orientation of the connections between alcohol consumption and different blood markers using Pearson and Spearman correlation coefficients. These investigations will assist in recognizing the markers that are most closely linked to alcohol intake.

After completing the correlation analysis, we will use Hypothesis Testing to investigate if the recorded correlations are statistically significant, to identify the markers that can consistently signal alcohol consumption. Moreover, a Regression Analysis will be performed to investigate how these markers collectively influence alcohol consumption, helping us pinpoint the most influential predictors in a multivariate context. In the Results and Discussion section, the study will compare its findings to previous research and discuss how they align or differ, while also exploring the clinical significance of the results in diagnosing alcohol-related health problems. This all-encompassing strategy guarantees a detailed examination of the connection between blood test factors and alcohol intake.

**Results and Discussion**

**Introduction**

The objective of the Analysis

This study aims to examine the correlation between blood test variables (MCV, Alkphos, SGPT, SGOT, GGT) and alcohol intake (measured in drinks). The main objective is to determine if these blood test markers can be dependable signs of alcohol consumption, which is important for both clinical and public health purposes. Healthcare professionals can make well-informed decisions in diagnosing and treating alcohol-related health issues by comprehending these relationships.Furthermore, we will employ regression analysis to assess the collective predictive ability of these blood markers on alcohol intake. This will give a broader understanding of how various factors collectively influence alcohol consumption, as opposed to considering them individually. The main focus of the research is to identify the blood test variables significantly linked to alcohol intake and the level of these connections. This examination will: Respond to the main research inquiry on the relationship between blood test factors and alcohol consumption. Examine which specific markers (MCV, SGPT, SGOT, etc.) offer the most reliable indication of alcohol consumption as stated in the research question. Examine how these variables interact using regression analysis to uncover the strongest predictors of alcohol consumption. A summary of basic statistical characteristics, such as mean and standard deviation, will be provided for each variable to present an overall understanding of the dataset. Analysis of correlation: Using Pearson and Spearman correlation coefficients, we will examine the strength and direction of the association between each blood test variable and alcohol intake.Hypothesis Testing: Hypothesis tests will be performed for every blood marker to establish the significance of the correlations that have been observed. This will aid in determining if each variable is a reliable indicator of alcohol intake. Regression Analysis: Our goal is to construct a multiple regression model.

**Body**

When sharing the analysis results, it is crucial to establish connections between the findings and pertinent literature. If past research has shown a significant connection between GGT levels and alcohol intake, referencing these studies when discussing similar findings will enhance the credibility of the results. This agreement supports the idea that the data aligns with previous studies, strengthening the credibility of the findings. Furthermore, any differences between the results of this research and existing literature can be pointed out and analyzed to offer fresh perspectives or challenge current beliefs.

Literature has been crucial in determining which variables to analyze. Markers like MCV and GGT are commonly acknowledged in research as significant indicators of alcohol intake. The latest study affirms their importance, suggesting they must be incorporated into the framework because of their strong associations with alcohol consumption. However, variables such as Alkphos, which show weak or insignificant correlations with alcohol intake, may not be appropriate for incorporation into a final predictive model. Thoroughly reviewing the literature guarantees that the model is grounded in evidence and useful in addressing the research question, leading to valuable outcomes for clinical and public health purposes.

**1. Descriptive Analysis**

**Procedures**

1. **Descriptive Statistics**

**Dataset: MCV**

* + **Mean**: 90.15580
  + **Median**: 90
  + **Mode**: 91
  + **Standard Deviation**: 4.57117
  + **Skewness**: -0.46676 (slight negative skew)
  + **Kurtosis**: 2.72095 (platykurtic)

**Interpretation**

* **Mean Score**: The average mcv of red blood cells is 90.16 and indicates the tendency
* **Standard Deviation**: A standard deviation of 4.57 indicates some variability, but values are generally close
* **Skewness**: Negative skewness of -0.47 means that there are a few low values that are pulling the mean down
* **Kurtosis**: A kurtosis of 2.72 suggests that the data has a platykurtic distribution with lighter tails than a normal distribution and fewer outliers.

**Dataset: Alkphos**

* + **Mean**: 69.92391
  + **Median**: 67
  + **Mode**: 62
  + **Standard Deviation**: 18.04804
  + **Skewness**:0.73452 (moderate positive skew)
  + **Kurtosis**: 0.69085 (platykurtic)

**Interpretation**

* **Mean Score**: The average alkphos level of 69.92 indicates central tendency
* **Standard Deviation**: A standard deviation of 18.05 indicates moderate variability
* **Skewness**: Positive skewness of 0,73 shows some higher values pushing the mean up
* **Kurtosis**: A kurtosis of 0.69 suggests lighter tails and fewer outliers than a normal distribution.

**Dataset: SGPT**

* + **Mean**: 30.34783
  + **Median**: 25.5
  + **Mode**: 17
  + **Standard Deviation**: 19.89633
  + **Skewness**: 3.21713 (strong positive skew)
  + **Kurtosis**: 14.97404 (leptokurtic)

**Interpretation**

* **Mean Score**: The average sgpt level of 30.35 indicates central tendency
* **Standard Deviation**: A standard deviation of 19.90 indicates significant variability
* **Skewness**: Positive skewness of 3.22 shows that a few extremely high values are pulling the mean up
* **Kurtosis**: A kurtosis of 14.97 suggests heavier tails and more outliers than a normal distribution.

**Dataset: SGOT**

* + **Mean**: 24.56884
  + **Median**: 22
  + **Mode**: 20
  + **Standard Deviation**: 10.56438
  + **Skewness**: 2.39092 (strong positive skew)
  + **Kurtosis**: 8.28326 (leptokurtic)

**Interpretation**

* **Mean Score**: The average sgot level of 24.56 indicates central tendency
* **Standard Deviation**: A standard deviation of 10.56 indicates moderate variability
* **Skewness**: Strong positive skewness of 2.39 suggests that a few higher values are pulling the mean up
* **Kurtosis**: A kurtosis of 8.28 indicates a sharper peak, heavier tails, and more outliers than a normal distribution.

**Dataset: GAMMAGT**

* + **Mean**: 36.94928
  + **Median**: 24
  + **Mode**: 11
  + **Standard Deviation**: 38.81243
  + **Skewness**: 3.11633 (strong positive skew)
  + **Kurtosis**: 12.41223 (leptokurtic)

**Interpretation**

* **Mean Score**: The average gammagt level of 36.95 indicates central tendency
* **Standard Deviation**: A standard deviation of 38.81 shows a significant level of variability
* **Skewness**: A strong positive skewness of 3.12 suggests that there are a few extremely high values pulling the mean up
* **Kurtosis**: A kurtosis of 12.41 indicates a leptokurtic distribution with a sharper peak, heavier tails, and more outliers than a normal distribution.

**Dataset: DRINKS**

* + **Mean**: 3.40942
  + **Median**: 3
  + **Mode**: 0.5
  + **Standard Deviation**: 3.28342
  + **Skewness**: 1.61193 (strong positive skew)
  + **Kurtosis**: 4.31777 (leptokurtic)

**Interpretation**

* **Mean Score**: The average of 3.41 drinks indicates a moderate average level of consumption
* **Standard Deviation**: A standard deviation of 3.28 shows a relatively high level of variability because it is almost the same as the mean.
* **Skewness**: Strong positive skewness of 1.61 suggests that a few high consumption levels are pulling the mean up
* **Kurtosis**: A kurtosis of 4.32 indicates a leptokurtic distribution with a sharper peak, heavier tails, and more outliers than a normal distribution.

**Dataset: SELECTOR**

* + **Mean**: 1.5665217
  + **Median**: 2
  + **Mode**: 2
  + **Standard Deviation**: 0.49663
  + **Skewness**: -0.28456 (slight negative skew)
  + **Kurtosis**: -1.94415 (platykurtic)

**Interpretation**

* **Mean Score**: The average of 1.57 for the selector variable suggests a general tendency towards the lower end of the scale being used
* **Standard Deviation**: A standard deviation of 0.5 shows low levels of variability
* **Skewness**: Weak negative skewness of -0.28456 suggests that a few low values are pulling the mean down
* **Kurtosis**: A kurtosis of -1.94 indicates a platykurtic distribution with a flatter peak, lighter tails, and fewer outliers than a normal distribution.

**Dataset: \_DATAOBS\_**

* + **Mean**: 171.7790
  + **Median**: 171.5
  + **Mode**: No Mode
  + **Standard Deviation**: 100.38693
  + **Skewness**: -0.00697 (weak negative skew)
  + **Kurtosis**: -1.224597 (platykurtic)

**Interpretation**

1. **Mean Score**: The average of 171.78 suggests a central tendency around this value
2. **Standard Deviation**: A standard deviation of 100.39 suggests considerable variability in the data
3. **Skewness**: Weak negative skewness of -0.01 indicates that the data is nearly symmetrical.
4. **Kurtosis**: A kurtosis of -1.22 indicates a platykurtic distribution with a flatter peak, lighter tails, and fewer outliers than a normal distribution.
5. **Distribution Analysis**
   * **Skewness**: Measures the asymmetry of the data distribution. Positive skewness indicates a longer tail on the right side.
   * **Kurtosis**: Measures the 'tailedness' of the distribution. High kurtosis indicates heavy tails.
6. **Frequency Tables**
   * Create tables showing how frequently each value occurs.

**Example**

**Dataset**: Test scores of students

* **Mean Score**: 75
* **Standard Deviation**: 10
* **Skewness**: 0.5 (positive skew)
* **Kurtosis**: 3.2 (slightly heavy tails)

**Interpretation**

* **Mean Score**: The average test score is 75, suggesting a central tendency.
* **Standard Deviation**: A standard deviation of 10 indicates moderate variability in scores.
* **Skewness**: Positive skewness means there are a few higher scores that might be pulling the average up.
* **Kurtosis**: A kurtosis of 3.2 suggests that the data has slightly heavier tails than a normal distribution.

[**2. Correlation Analysis**](#_heading=h.lm5byqaeuonm)

**Procedure**

MCV and drinks:r=0.35770 (p < .0001) **moderate positive correlation**

alkphos and drinks:r=0.11519 (p = .0580) **Weak positive correlation**

sgpt and drinks: r=0.18880 (p = .0016) **weak positive correlation**

sgot and drinks: r=0.25817 (p < .0001) **moderate positive correlation**

gammagt and drinks: r=0.38516 (p < .0001) **This is the strongest positive correlation of all five variables.**

**Interpretation**

Gammagt shows the highest correlation with alcohol consumption (both Pearson and Spearman). This implies that gammagt levels play a crucial role in determining alcohol intake in patients with liver disorders.MCV also demonstrates a moderate relationship with alcohol intake, suggesting that it could be another significant indicator.SGOT and SGPT also have statistically significant correlations but they are not as strong as the correlation with Gammagt.

The Pearson correlation between GGT levels and alcohol consumption is r = 0.38516, showing a moderate to strong positive relationship of all the five variables. This indicates that there is a strong connection between GGT levels and alcohol intake, which is consistent with previous research addressed in the above literature review highlighting GGT as a marker for alcohol consumption.

**Interpretation:**

**Regression analysis**

H0: beta is equal to zero.

Ha: beta is not equal to zero.

For mcv the p-value<0.0001 which is less than the alpha value of 0.05, so we reject the null hypothesis and accept the alternative hypothesis, and we conclude that mcv is included in the regression model. Alkphos, the p-values<0.0001 which is less than the alpha value of 0.05, so we reject the null hypothesis and accept the alternative hypothesis. We conclude that alkphos is included in the regression model.

For sgpt the p-value=0.3062 which is greater than a=0.05, so we do not reject the null hypothesis, we conclude that sgpt is not included in the regression. Sgot, the p-value=0.3543 which is greater than the alpha value=0.05, so we do not reject the null hypothesis, we conclude that sgot is not included in the regression model. For gammagt, the p-value<0.05, we reject the null hypothesis, and we conclude that gammagt is not included in the regression model.

**Univariate analysis:**

Looking at the histogram of the mcv, the histogram is skewed to the left, so it is not normally distributed. The histogram of the alkphos test is skewed to the right, so the distribution of alkphos is normal. The histogram of sgpt is skewed to the right, so sgpt is not normally distributed. The histogram of sgot is skewed to the right, which proves that the distribution of sgot is not normal. The histogram of gammagt is skewed to the right which proves that gammagt is not normally distributed.

**Means analysis:**

Mcv has 276 observations, mean=90.16, standard deviation=4.57, a minimum=65 and a maximum= 102. Alkphos has 276 observations, mean=69.92, standard deviation=18.48, a minimum=4 and a maximum= 138. Sgpt has 276 observations, mean=90.16, standard deviation=19.90, a minimum=4 and a maximum= 155. Sgot has 276 observations, mean=10.56, standard deviation=10.56, a minimum=5 and a maximum= 82 and gammagt has 276 observations, a mean of 36.95, a standard deviation=38.81, a minimum of 5 and a maximum of 297

**Conclusion**

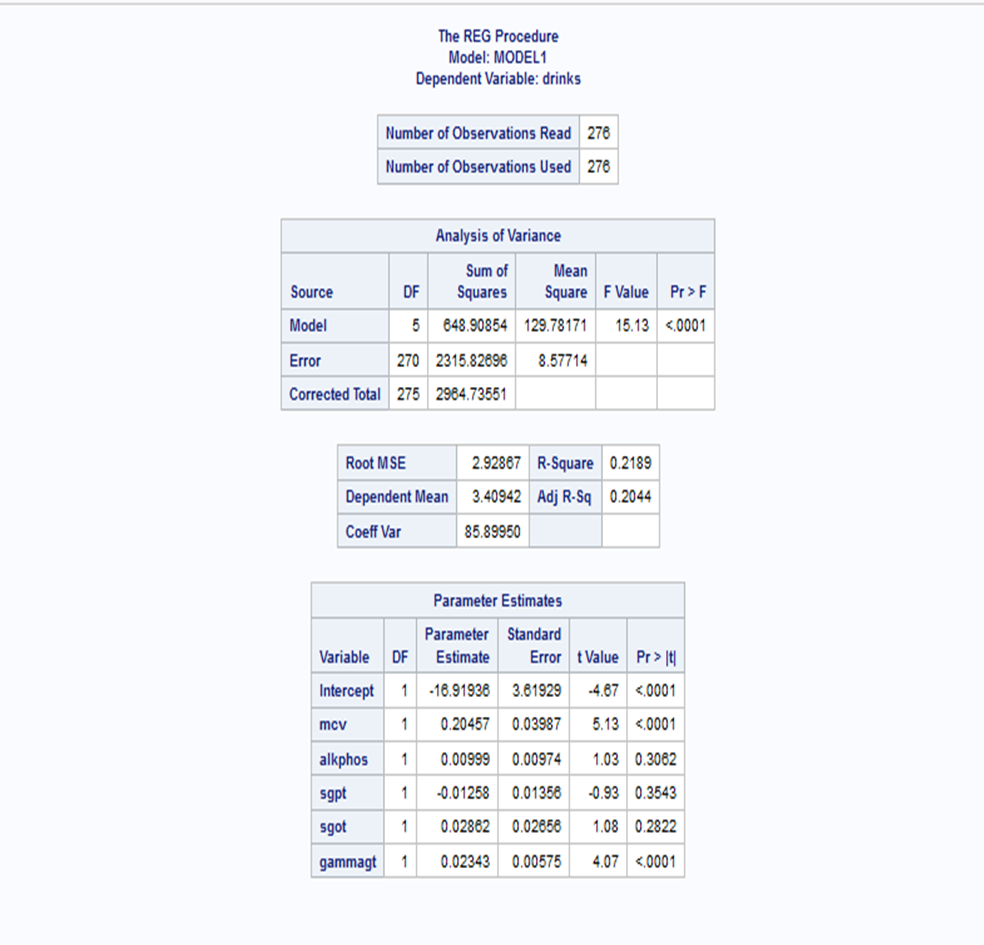
In conclusion, analyzing different data sets on red blood cell metrics, liver enzymes, and alcohol consumption offers important information on the average values and spread of these factors. The findings show notable differences in the data sets, especially with SGPT, SGOT, and GAMMAGT, which display pronounced positive skewness and leptokurtic distributions. This shows that numerous outliers and extreme values have an impact on the average measurements. Furthermore, the links between these factors indicate that high GAMMAGT levels are most closely related to alcohol consumption, highlighting its importance as a key indicator for evaluating alcohol intake in patients with liver issues.

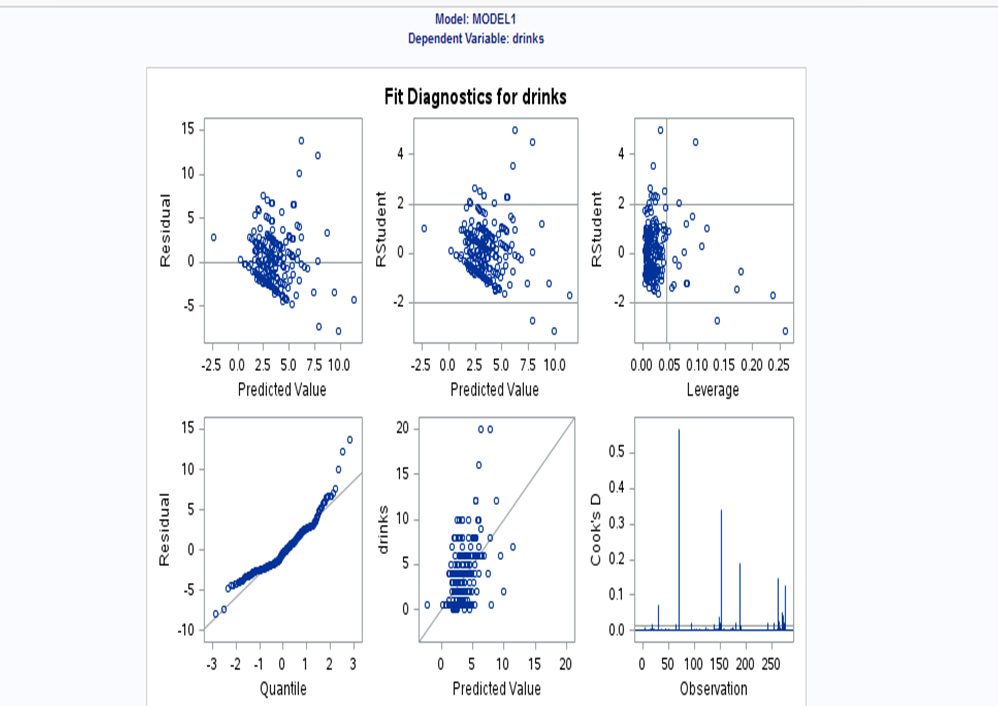
The results from the regression analysis support these conclusions by including MCV and Alkphos levels in the model, but not SGPT and SGOT which did not show statistical significance as predictors. The examination of one variable revealed that some variables did not follow a normal distribution, emphasizing the complexity of the data. In general, this research highlights the significance of employing different statistical methods to fully grasp the connections between biochemical markers and alcohol intake, which could help steer future studies and clinical evaluations in this field.

Future studies should concentrate on delving deeper into the biochemical markers associated with alcohol intake, particularly with a bigger sample size to validate the strength of the current results. In addition, examining how age, gender, and existing liver conditions influence these connections would offer a more thorough insight into the effects of alcohol on liver enzymes and red blood cell measures. Including longitudinal data in this research would also aid in establishing causal links instead of just correlations.

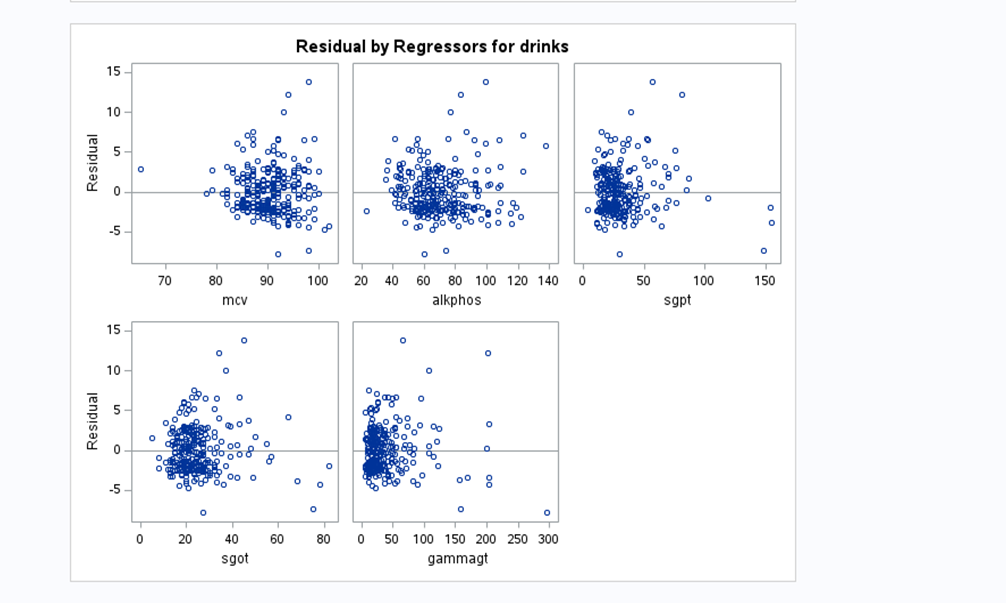
Various constraints in this research could impact the understanding and applicability of the results. A major drawback is the dependency on cross-sectional data, which hinders our ability to determine the cause-and-effect relationship between alcohol intake and fluctuations in liver enzyme levels. Furthermore, the overall statistical results may have been affected by the skewness and existence of outliers in important variables like SGPT, SGOT, and GAMMAGT. Another constraint is the absence of monitoring potential factors that could distort results, such as diet, medication use, or genetic predispositions, all of which could affect liver enzyme levels and alcohol breakdown.

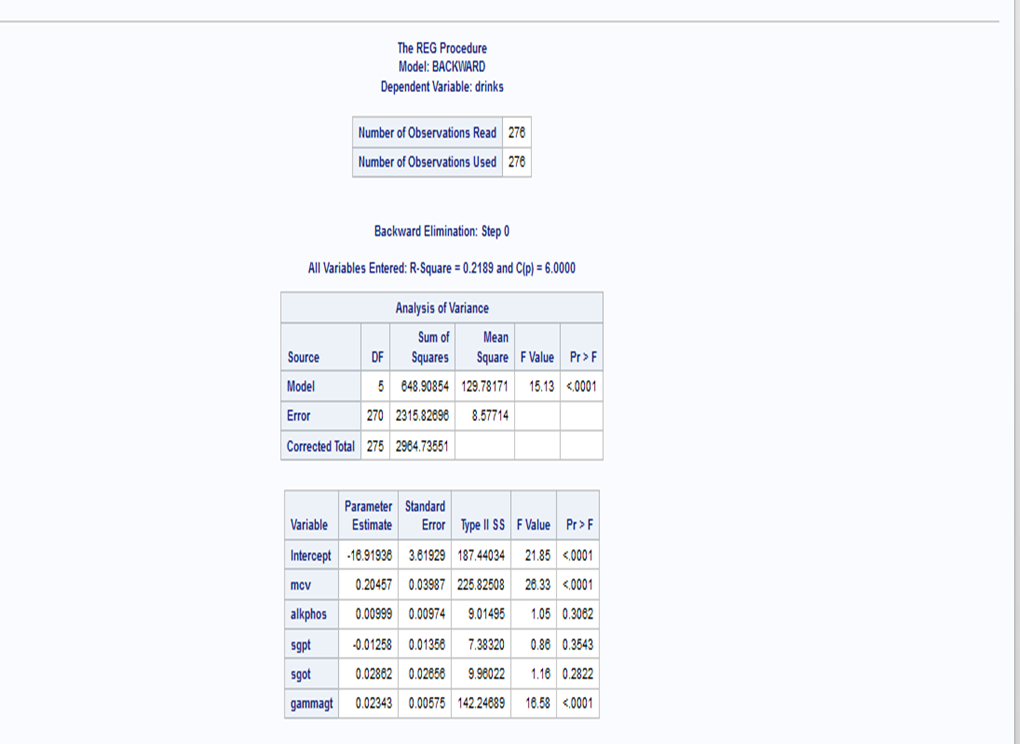
The results of this research have significant consequences for clinical practice, especially in identifying and assessing patients with presumed alcohol-induced liver injury. Increased levels of GAMMAGT, specifically, ought to be seen as an important marker during the evaluation of alcohol consumption in patients, and healthcare providers could include this enzyme in a standard liver function test for individuals who drink heavily. Additionally, recognizing the atypical distribution of specific markers (such as SGPT, SGOT) in patients can assist doctors in acknowledging the drawbacks of relying on standard reference values and promoting personalized evaluations.

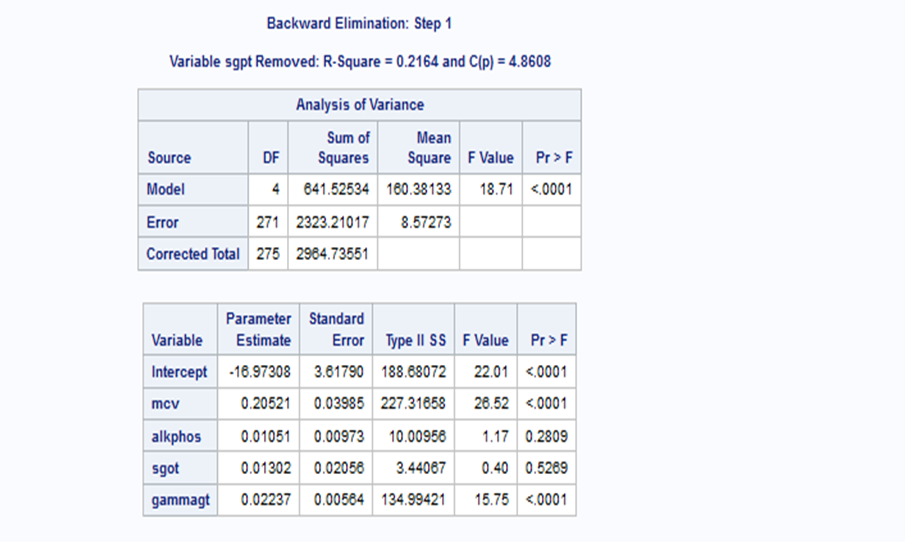
****

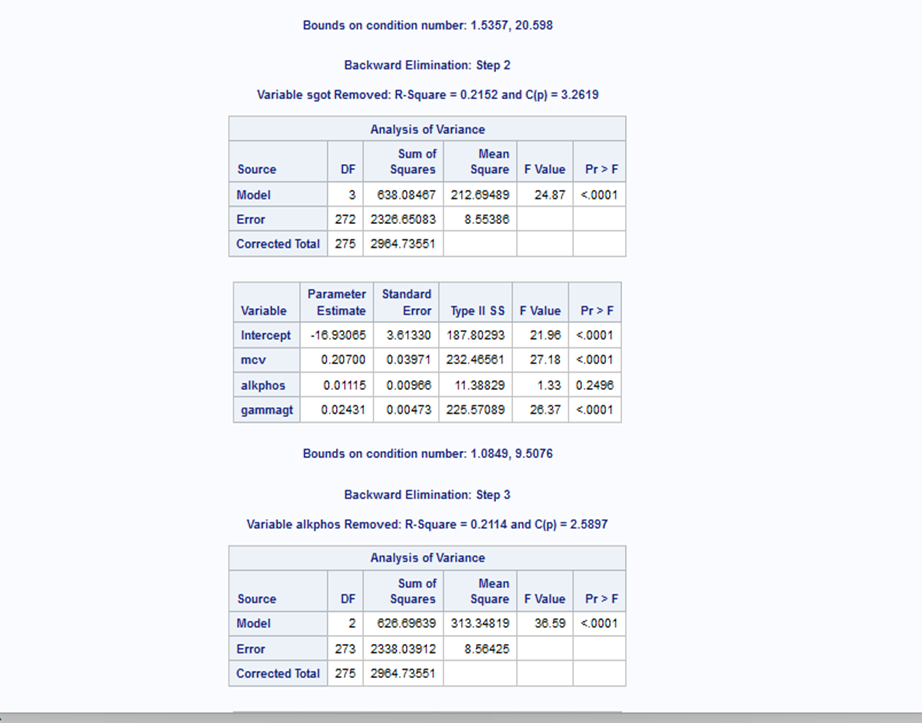
****

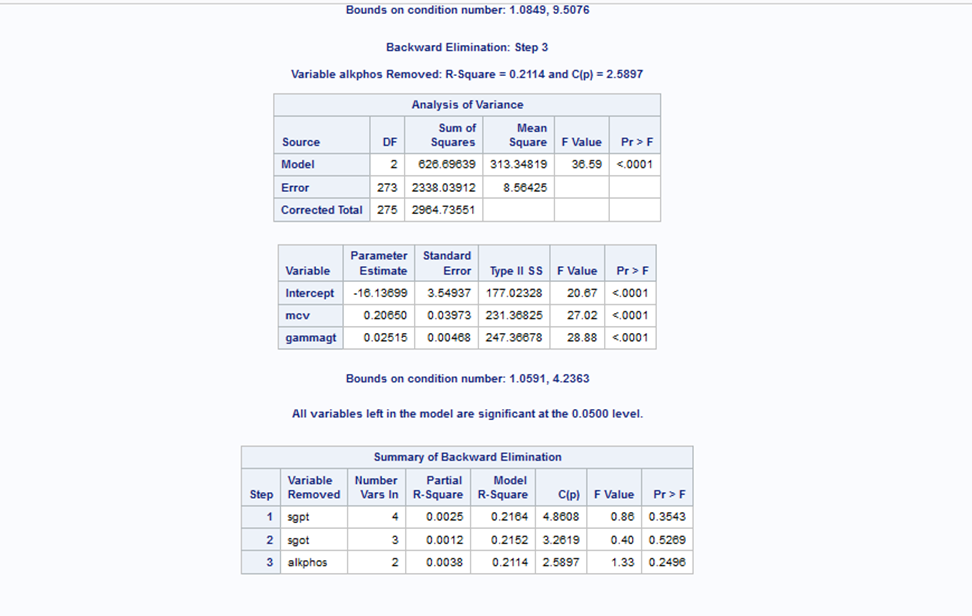
****

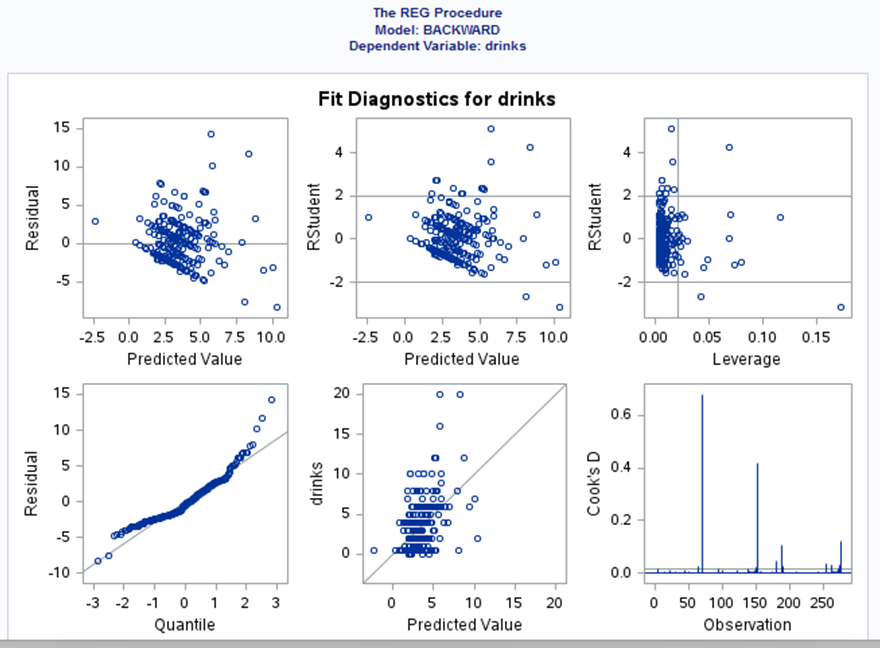
****

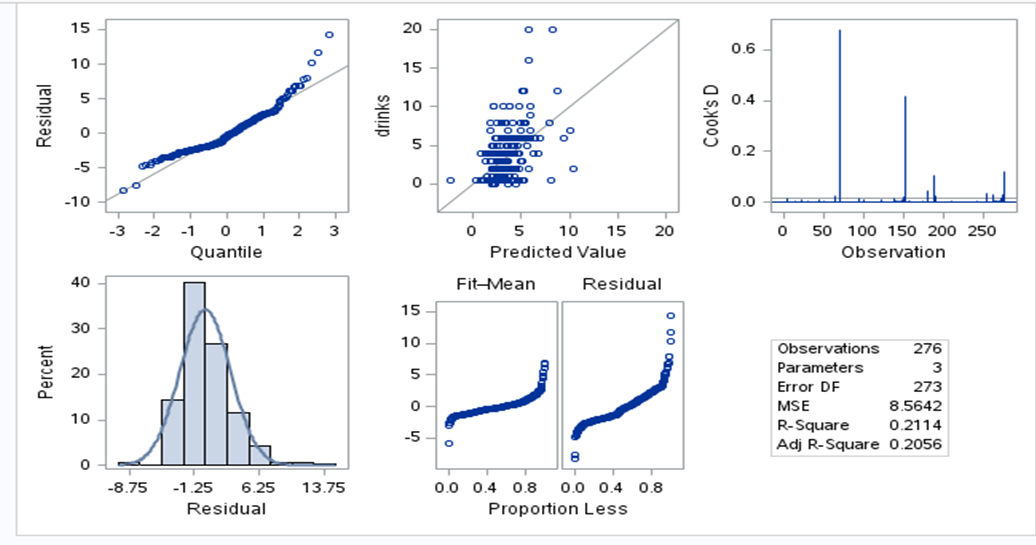
****

****

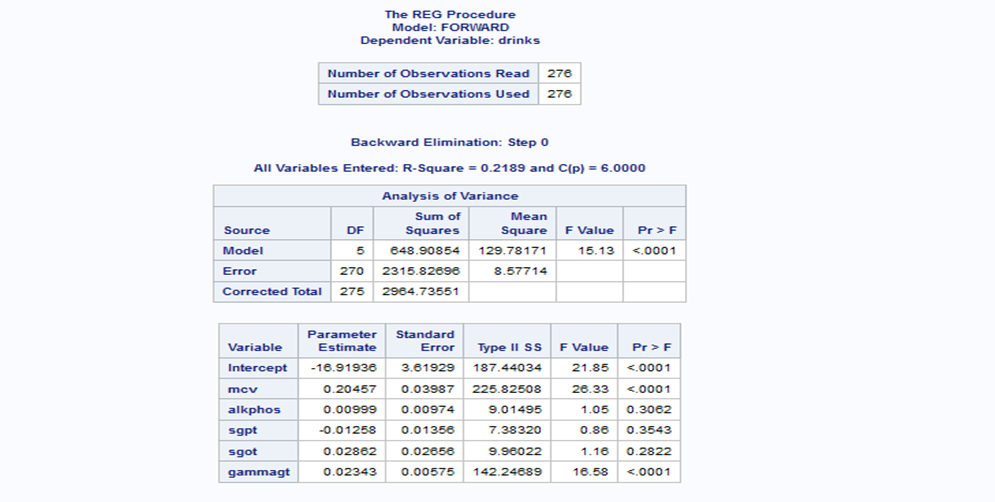
****

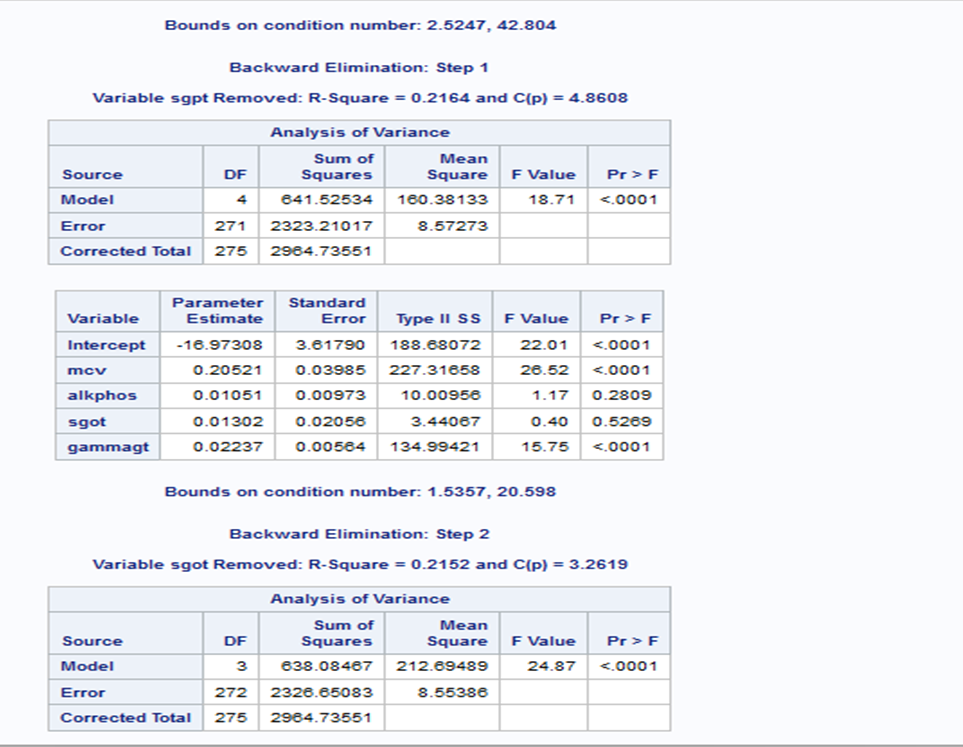
****

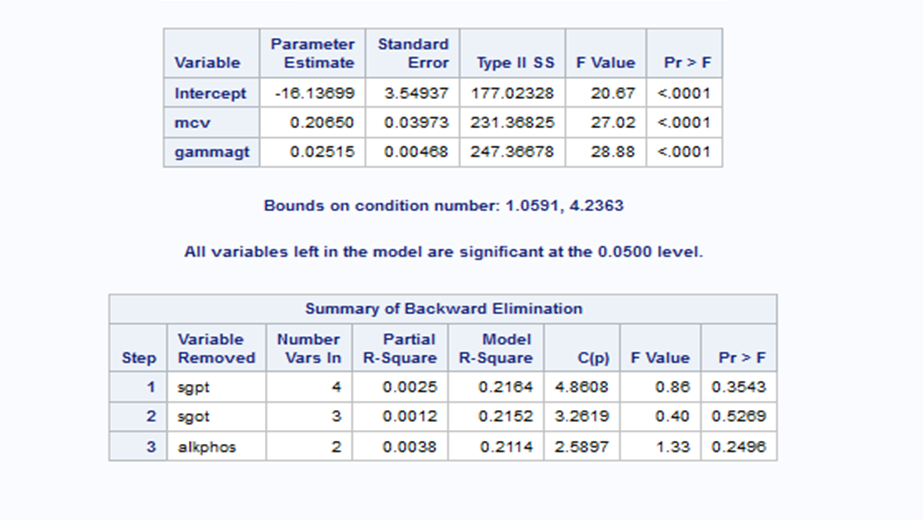
****

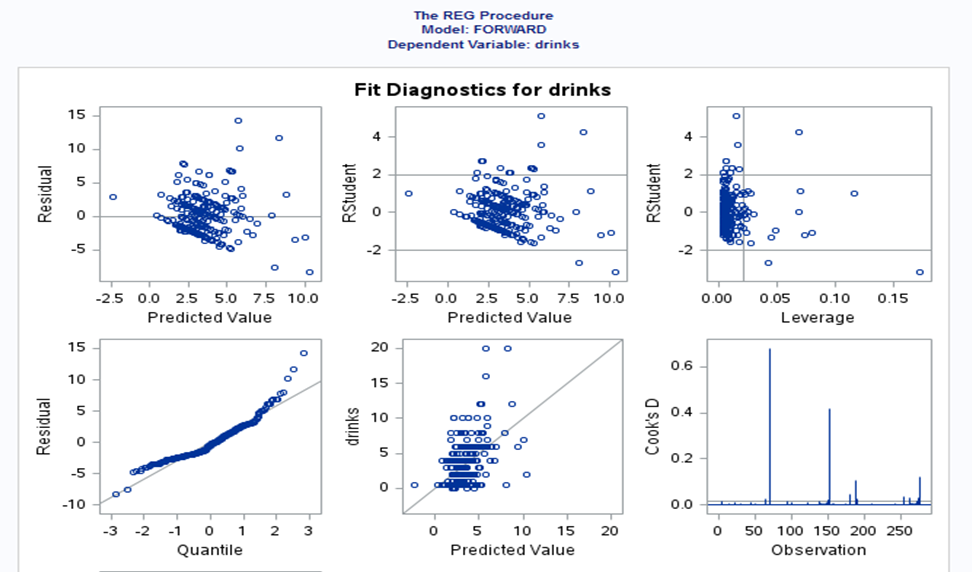
****

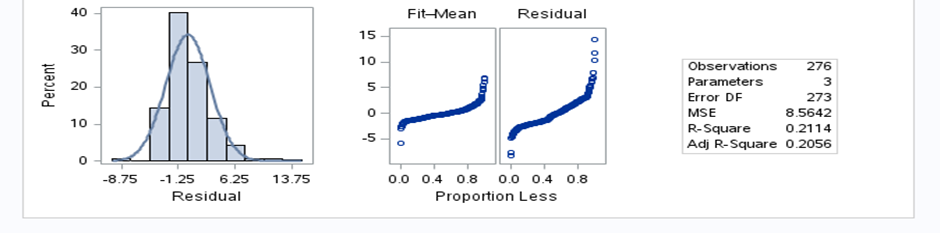
****

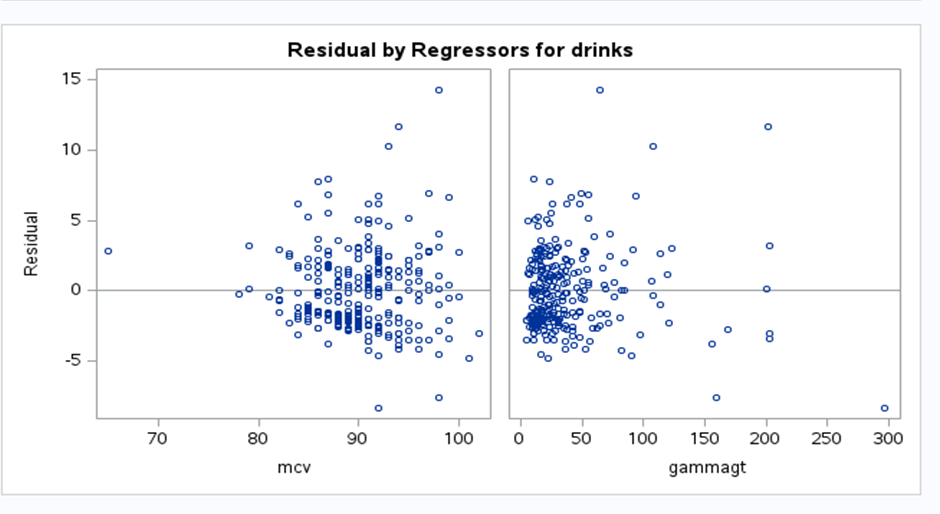
****

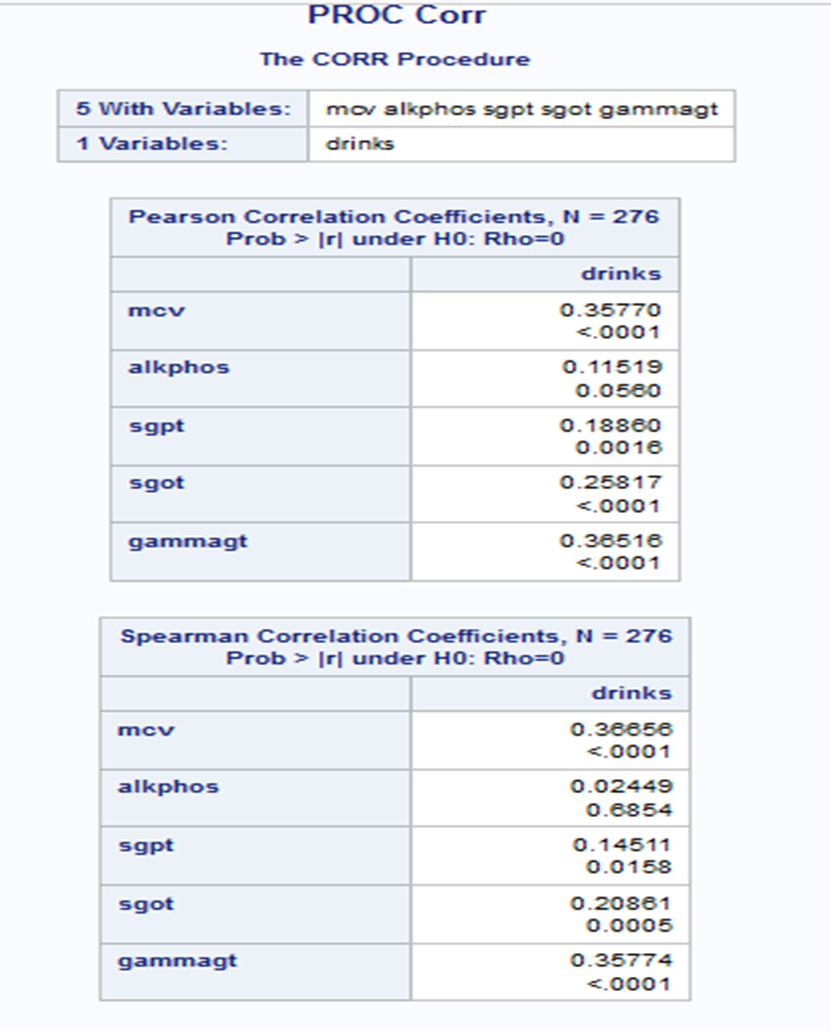
****

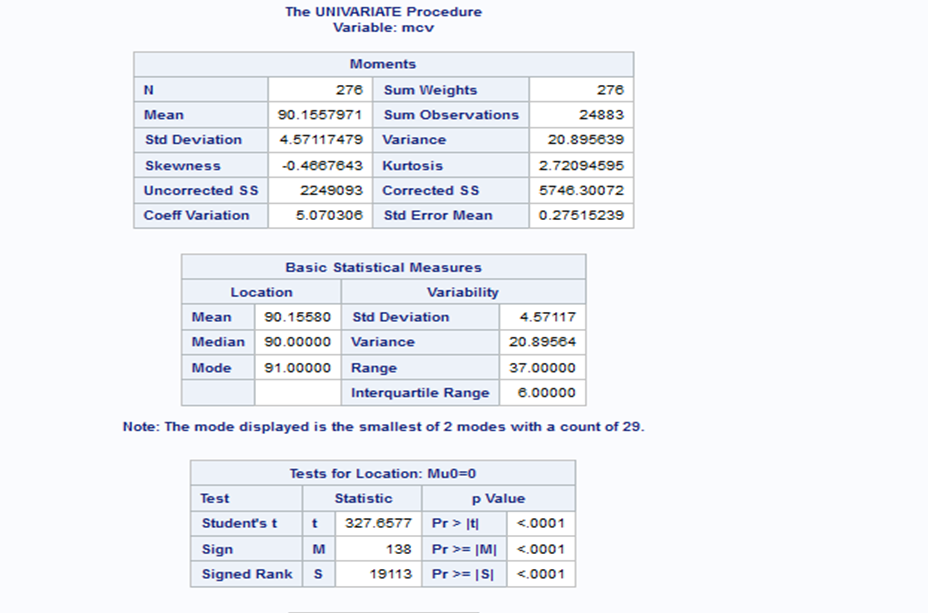
****

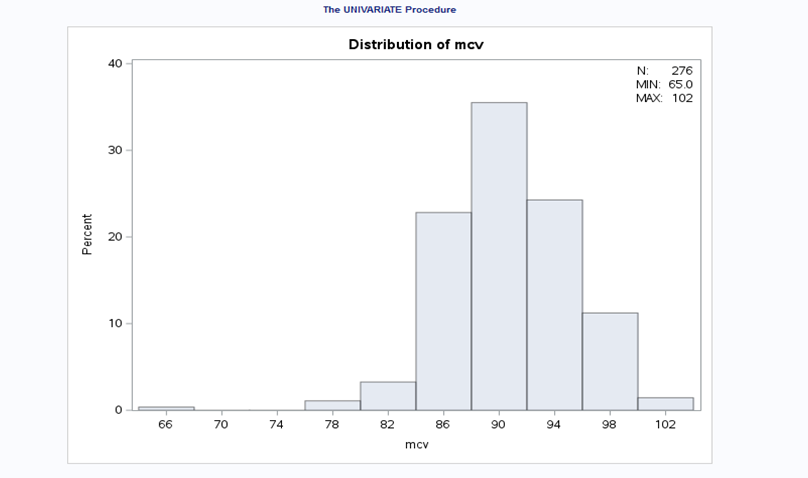
****

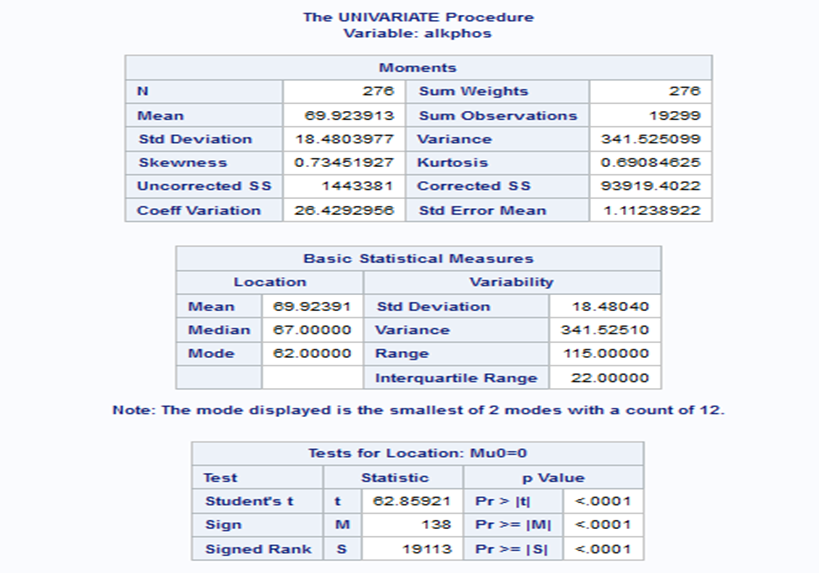
****

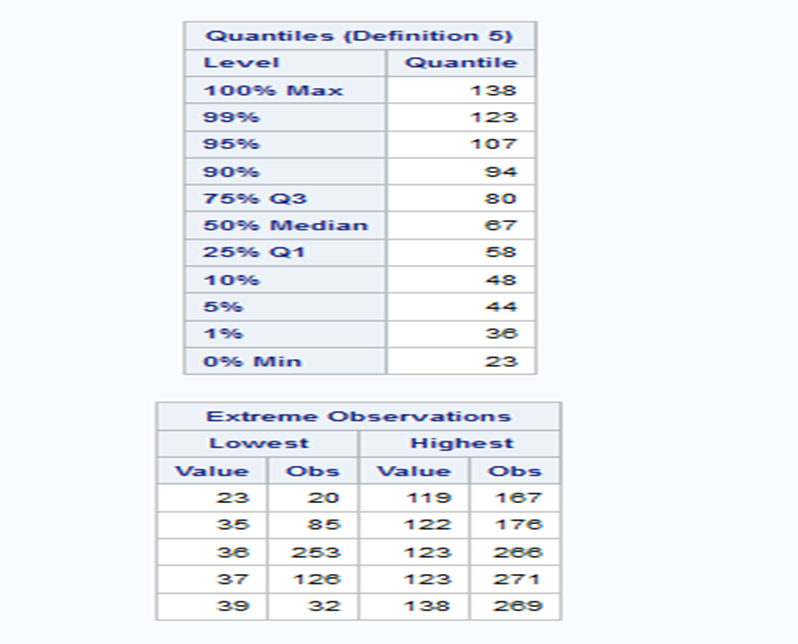
****

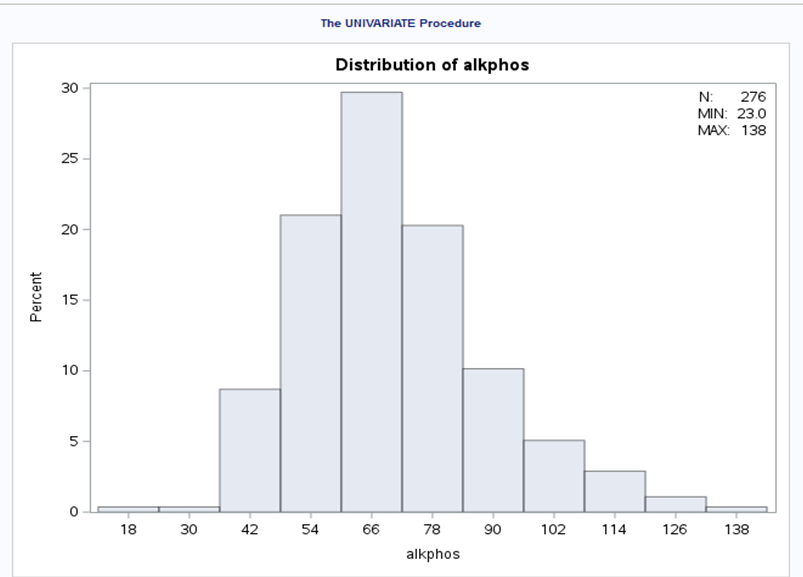
****

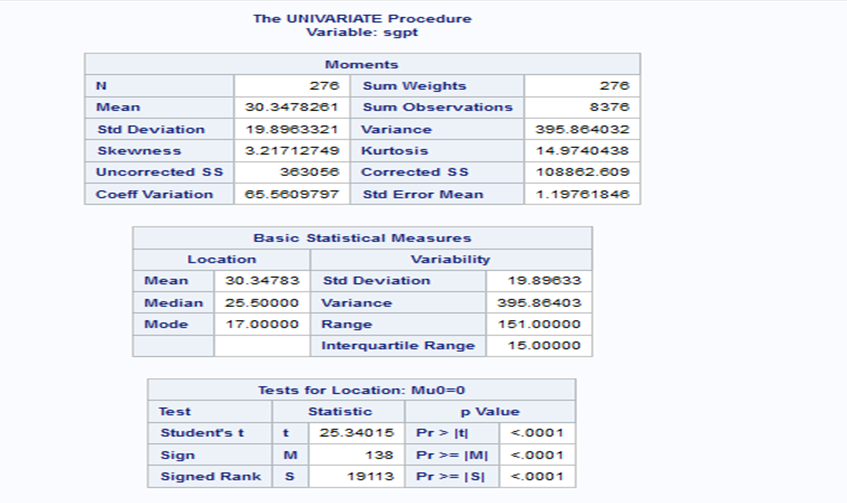
****

****

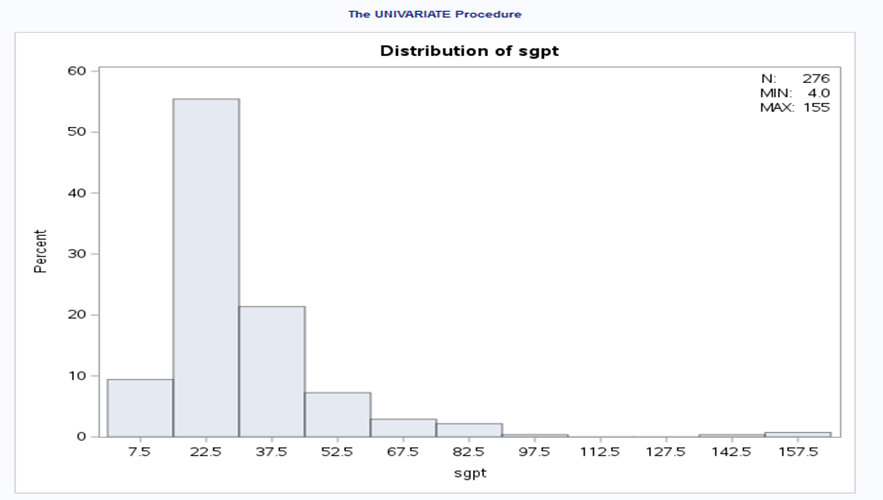
****

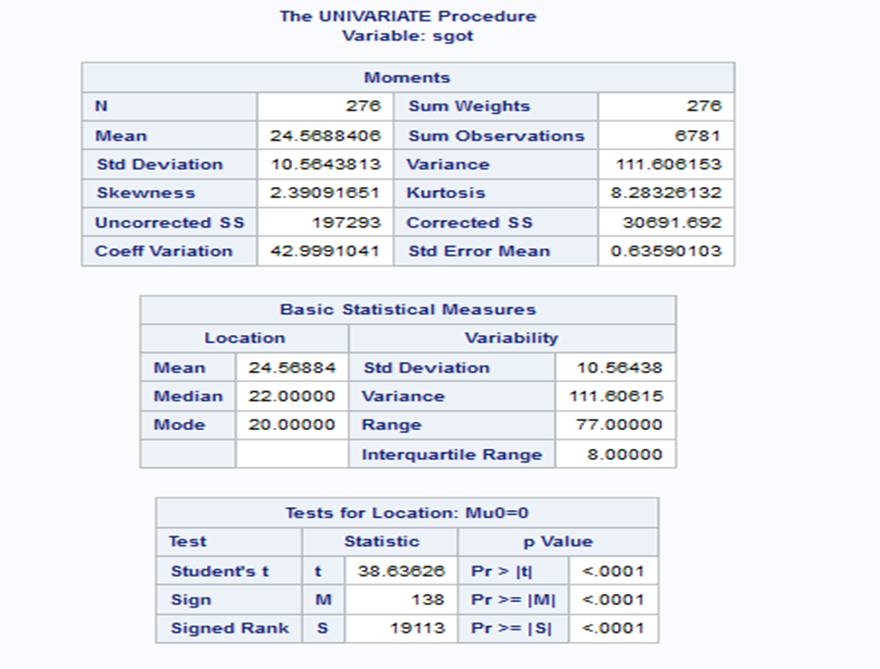
****

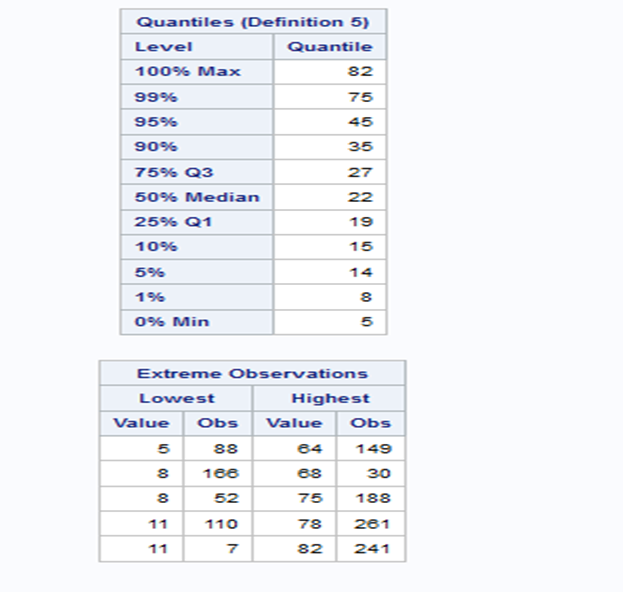
****

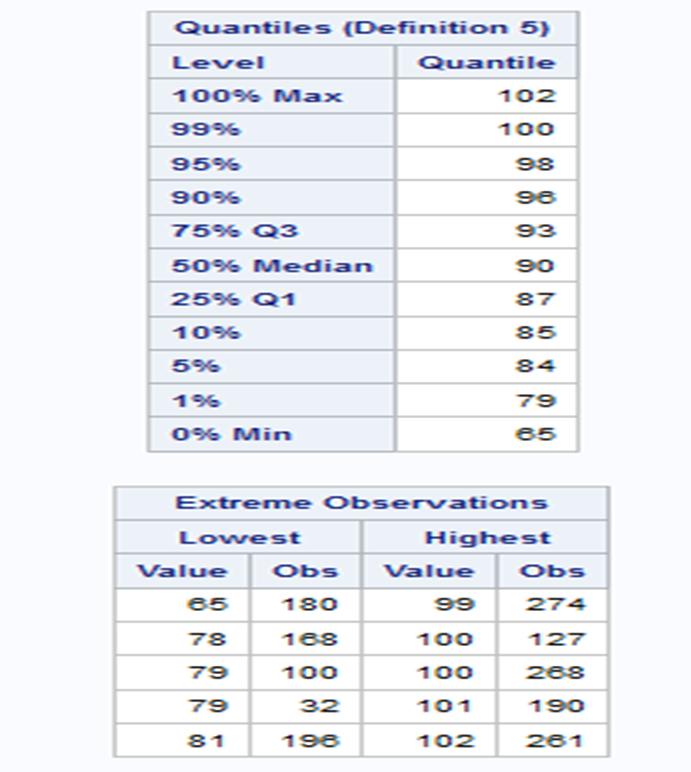
****

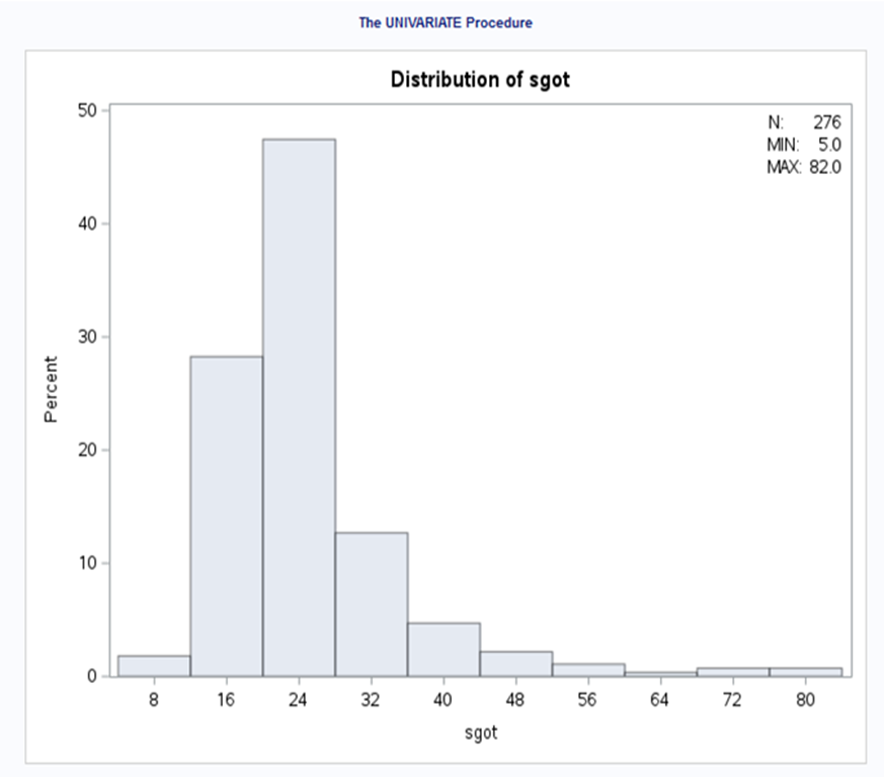
****

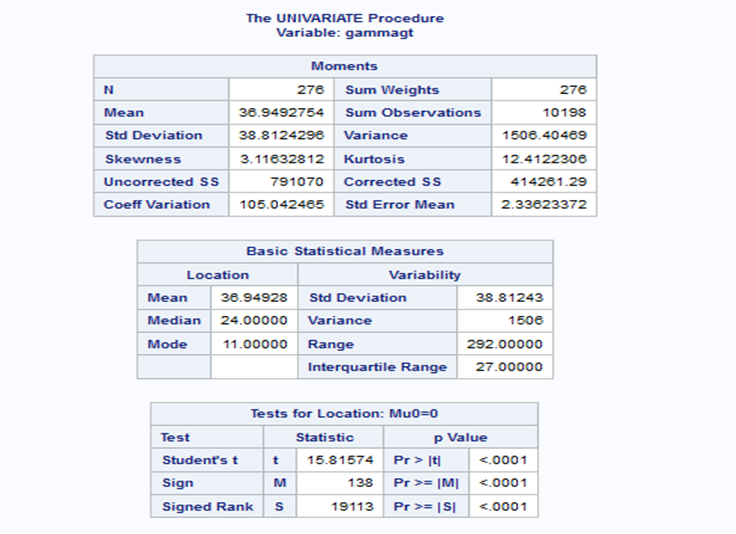
****

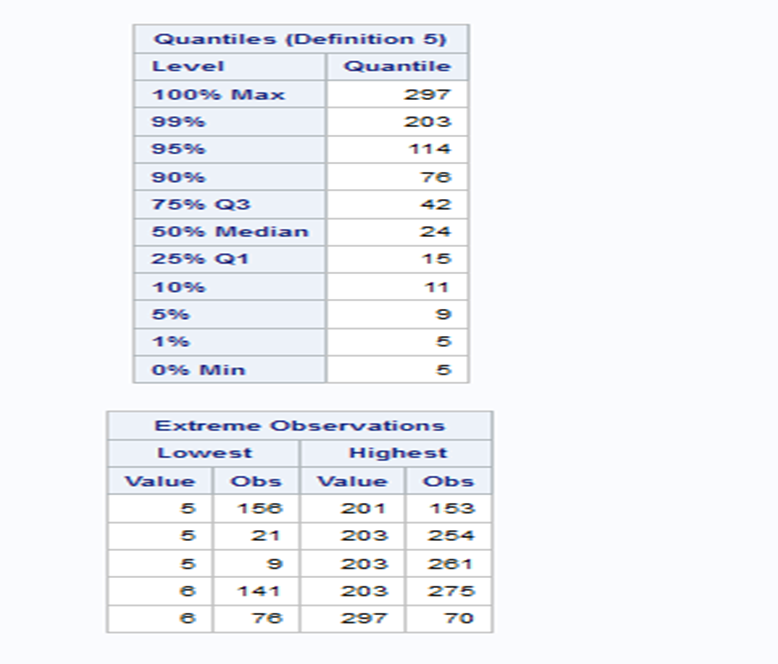
****

****

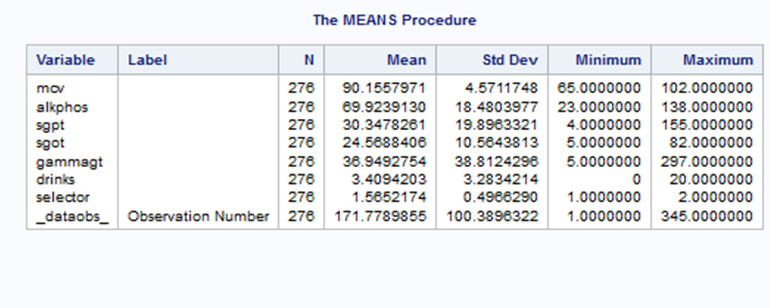
****

****

****

****

****

****

Codes:

Regression analysis

ODS RTF FILE='/home/u63872429/sasuser.v94/REGRESSION\_TEST';

PROC REG DATA= '/home/u63872429/sasuser.v94/group\_12\_train.sas7bdat';

MODEL DRINKS= MCV ALKPHOS SGPT SGOT GAMMAGT;

RUN;

ODS RTF CLOSE;

PROC REG DATA ='/home/u63872429/sasuser.v94/group\_12\_train.sas7bdat';

BACKWARD: MODEL DRINKS= MCV ALKPHOS SGPT SGOT GAMMAGT / SELECTION = BACKWARD SLSTAY=0.05;

run;

PROC REG DATA ='/home/u63872429/sasuser.v94/group\_12\_train.sas7bdat';

FORWARD: MODEL DRINKS= MCV ALKPHOS SGPT SGOT GAMMAGT / SELECTION = FORWARD SLSTAY=0.05;

run;

Correlation procedure

PROC CORR DATA="/home/u63872429/sasuser.v94/group\_12\_train.sas7bdat" Pearson spearman nosimple;

var drinks;

with mcv alkphos sgpt sgot gammagt ;

title 'PROC Corr';

run;

Univariate procedure

ODS HTML;

PROC PRINT DATA="/home/u63872429/sasuser.v94/group\_12\_train.sas7bdat";

TITLE 'Group 12 Dataset';

RUN;

ODS HTML CLOSE;

ODS HTML;

PROC UNIVARIATE DATA="/home/u63872429/sasuser.v94/group\_12\_train.sas7bdat";

VAR MCV ALKPHOS SGPT SGOT GAMMAGT;

HISTOGRAM / NROWS=2 NCOL=2 CFILL=BLUE PFILL=M3N45;

INSET N='N:' (4.0) MIN='MIN:' (4.1) MAX='MAX:' (4.1)

/ NOFRAME POSITION=NE HEIGHT=2;

RUN;

ODS HTML CLOSE;

Means procedure

ODS HTML;

PROC MEANS DATA= '/home/u63872429/sasuser.v94/group\_12\_train.sas7bdat';

RUN;

ODS HTML CLOSE;

# Appendix

## Timeline Plan

**Group 1:**

Griffiths Moshoeshoe-4381231

Thaakirah Mosoval-4314422

Elethu Molose 4366125

**Group 2**:

Liyema Mngxekeza 4133533

Bennedict Mkhonto 4226279

Ayabulela Carol Moss 4385096

**1st Submission**

Abstract **- group 2**

Timeline Planand Research questions**- group 1**

**Due date:** 5th August 2024

**2nd Submission**

Introduction**-group 1**

Literature review**-group 2**

**Due date:** 16th August 2024

**3rd Submission**

**Methodology**

1. Descriptive Analysis-**group 2**

2. Correlation Analysis-**group 1**

3. Multiple Regression Model (Full Model)-**group 2**

4. Analysis of Variance-**group 1**

5. Hypothesis Testing-**group 2**

6. R-Squared and Coefficient of Variance-**group 1**

7. Multiple Regression Model (Reduced Model)-**Both Groups**

**Due Date:**30th August 2024

**4th Submission**

Analysis**-Both Groups**

**Due date**:20th September 2024

**(Final Submission – 1.5 weeks)**

Final Report-**Both Groups**

**Due date:**7th October 2024

# Bibliography

1. Rosman AS. Utility and evaluation of biochemical markers of alcohol consumption. J Subst Abuse. 1992;4:277–297. [[PubMed](https://pubmed.ncbi.nlm.nih.gov/1458045)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=J+Subst+Abuse&title=Utility+and+evaluation+of+biochemical+markers+of+alcohol+consumption&author=AS+Rosman&volume=4&publication_year=1992&pages=277-297&pmid=1458045&)]
2. Sillanaukee P. Laboratory markers of alcohol abuse. Alcohol Alcohol. 1996;**31**:613–616. [[PubMed](https://pubmed.ncbi.nlm.nih.gov/9010553)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Alcohol+Alcohol&title=Laboratory+markers+of+alcohol+abuse&author=P+Sillanaukee&volume=31&publication_year=1996&pages=613-616&pmid=9010553&)]
3. Conigrave KM, Saunders JB, Whitfield JB. Diagnostic tests for alcohol consumption. Alcohol Alcohol. 1995;**30**:13–26. [[PubMed](https://pubmed.ncbi.nlm.nih.gov/7748270)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Alcohol+Alcohol&title=Diagnostic+tests+for+alcohol+consumption&author=KM+Conigrave&author=JB+Saunders&author=JB+Whitfield&volume=30&publication_year=1995&pages=13-26&pmid=7748270&)]
4. Morgan MY, Camilo ME, Luck W, Sherlock S, Hoffbrand AV. Macrocytosis in alcohol-related liver disease: its value for screening. Clin Lab Haematol. 1981;**3**:35–44. [[PubMed](https://pubmed.ncbi.nlm.nih.gov/7226720)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Clin+Lab+Haematol&title=Macrocytosis+in+alcohol-related+liver+disease:+its+value+for+screening&author=MY+Morgan&author=ME+Camilo&author=W+Luck&author=S+Sherlock&author=AV+Hoffbrand&volume=3&publication_year=1981&pages=35-44&pmid=7226720&)]
5. Kamath PS, Kim WR. The model for end-stage liver disease (MELD) *Hepatology.* 2007;45:797–805. <https://doi.org/10.1002/hep.21563>. [[PubMed](https://pubmed.ncbi.nlm.nih.gov/17326206)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Hepatology&title=The+model+for+end-stage+liver+disease+(MELD)&author=PS+Kamath&author=WR+Kim&volume=45&publication_year=2007&pages=797-805&pmid=17326206&)]
6. Whitfield JB, Rahman K, Haber PS, Day CP, Masson S, Daly AK, et al. Brief report: Genetics of alcoholic cirrhosis-GenomALC multinational study. *Alcoholism: Clinical and Experimental Research.* 2015;39:836–842. <https://doi.org/10.1111/acer.12693>. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4398999/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/25872595)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Alcoholism:+Clinical+and+Experimental+Research&title=Brief+report:+Genetics+of+alcoholic+cirrhosis-GenomALC+multinational+study&author=JB+Whitfield&author=K+Rahman&author=PS+Haber&author=CP+Day&author=S+Masson&volume=39&publication_year=2015&pages=836-842&)]
7. Sillanaukee, P., Massot, N., Jousilahti, P., Vartiainen, E., Sundvall, J., Olsson, U., Poikolainen, K., Ponnio, M., Allen, J. P. & Alho, H. (2000) Dose–response of laboratory markers to alcohol consumption in a general population. American Journal of Epidemiology.
8. Nemesanszky, E., Lott, J. A. & Arato, M. (2010) Changes in

serum enzymes in moderate drinkers after an alcohol challenge.Clinical Chemistry, 34, 525–527.

1. Belfrage, P., Berg, B., Cronholm, T. et al.(1973) Prolonged administration of ethanol to young, healthy volunteers: effects on biochemical, morphological and neurophysiological parameters, Acta Medica Scandinavica Supplement, 552 , 1–44.Conigrave, K., Degenhardt, L., Whitfield, J. B., Saunders, J. B., Helander, A.,Tabakoff, B. on behalf of the WHO/ISBRA study group. (2002) CDT, GGT and AST as markers of alcohol use: the WHO/ISBRA collaborative project. Alcoholism, Clinical and Experimental Research, 26, 332–339.
2. Agarwal,S., Fulgoni 3rd, V. L., Lieberman, H. R. (2016). Assessing alcohol intake & its dose-dependent effects on liver enzymes by 24-h recall and questionnaire using NHANES 2001-2010 data.Oak Ridge Institute for Science and Education, Belcamp, MD,21017,USA. Henry M Jackson Foundation, Bethesda, MD, 20817, USA. Military Nutrition Division, U.S. Army Research Institute of Environmental Medicine,Natick,MA,01760-5007,USA. 10.1186/s12937-016-0180-y.
3. Patel, R., Mueller, M. (2023). Alcohol Liver Disease. University of Central Florida. University of California San Diego.