UNIVERSITY of the WESTERN CAPE

UNIVERSITY of the WESTERN CAPE

Natural Science
Practical Assignment
STA221

MEMBERS:
Imaan Adams
Dateh Joseph Ambe
Bernard Bladergroen
Mariam Baker
Masingita Baloyi
Simphiwe Bala

DUE DATE: 10 October 2025

Plagiarism Declaration:
We declare that this assignment/submission is our own, honest work/idea/creation and we have not cheated, collaborated or colluded with anyone nor been deceptive in completing it. Where we do refer to others' work/ideas/creations, we have referenced them and the source.

Member	Signature
Imaan Adams	IA
Dateh Joseph Ambe	DA
Bernard Bladergroen	BB
Mariam Baker	MM
Masingita Baloyi	MB
Simphiwe Bala	SB

UNIVERSITY of the WESTERN CAPE

Table of Contents

UNIVERSITY of the WESTERN CAPE

Appendix

UNIVERSITY of the WESTERN CAPE

Identifying key biomarkers of alcohol consumption using multiple statistical techniques

Authors

Imaan Adams, Dateh Joseph Ambe, Mariam Baker, Bernard Peter Bladergroen, Masingita Baloyi, Simphiwe Bala

Abstract

We present in this paper the correlation of alcohol intake to biochemical blood markers in order to determine which biochemical marker is the most reliable indicator of excessive alcohol consumption. The study centers on the relationship between changes in certain biomarkers- Mean Corpuscular Volume (MCV), Alkaline Phosphatase (ALKPHOS), Serum Glutamate Pyruvate Transaminase (SGPT/ALT), Serum Glutamate Oxaloacetate Transaminase (SGOT/AST), and Gamma-Glutamyl Transferase (GGT) and the amount of alcohol consumed on a daily basis. These biomarkers were tested individually; however, no single study has been able to test the overall comparative prediction of these biomarkers as a result of the choice of statistical modelling. The paper fills that gap by constructing a multiple regression model to estimate the diagnostics strength of such blood markers in explaining alcohol consumption. Analysis of biomarker data was performed by a quantitative method, SAS correlational and regression analysis. The findings revealed that MCV was more likely to follow a normal distribution and more likely to be constant, whilst SGPT, SGOT and GGT, specifically, were skewed positive and had large tails suggesting that these were more responsive to changes in the degree of alcohol intake. Correlation and regression analysis indicated statistically significant correlation of MCV and GGT with alcohol consumption, where the combination of the two variables explained 21% of the variance with GGT being the most sensitive biochemical variable. The authors conclude that a hematological (MCV) and enzymatic (GGT) biomarker combination offers a strong model to detect the effects of alcohol and augments diagnostic capabilities of the measures in clinical and population health practice.

Literature Review & Introduction

Introduction

Alcoholic liver disease (ALD) is among top ten causes of global mortality. While the consequences of liver toxicity caused by alcohol are well documented and understood by the public, with patients self-reporting their alcohol consumption being unreliable it is essential to ensure the objective accuracy of this consumption. Thus, in order to find this objective measurement physicians, make use of biomarkers that screen for chronic alcohol consumption and liver damage. However, the usefulness of discrete blood markers- Mean corpuscular volume (MCV), Alanine aminotransferase (SGPT or ALT), Aspartate aminotransferase (SGOT or AST), Alkaline phosphatase (ALKPHOS), and Gamma-glutamyl transpeptidase (GAMMAGT or GGT)- as forecasts of alcohol intake, although sensitive, remain best for initial investigations. Though some of these biomarkers are potentially valuable, available literature shows that there is a need to conduct additional studies to ascertain the reliability and the constraining factors of such biomarkers.

The relationship between these liver biomarkers and the consumption of alcohol is thus critically evaluated in this review on the application of regression analysis to measure the relationship. This aims at evaluating the efficacy of the biomarkers as predictors of drinking behavior in the clinical practice setting.

In the current study, it is argued that liver function tests, especially MCV and gammagt, are of value as predictors of alcohol intake but that the success of such measures depends on the nature of the assay used as well as the type of analysis done.

The recent epidemiological and biomedical data imply that liver functional tests, notably serum AST, ALT, and MCV, can be utilized as sound biological indicators of measuring alcohol exposure. Experimental studies have found these biomarkers to be consistently related to alcohol-related pathology and are therefore useful proxies of alcohol consumption. With this, this review targets such research to review their validity as markers of alcohol consumption, focusing on effectiveness across various populations and situations.

Excessive alcohol consumption as a factor in liver disease

Effects of liver disease

The significance of diagnosing alcoholic liver disease stems from the need to inform patients as this disease can have fatal consequences if left unchecked. According to Devarbhavi et al. (2023), liver disease is one of the main causes of over two million deaths annually, with approximately two-thirds of these deaths being men. Most of these deaths are due to complications with the advanced liver diseases cirrhosis and hepatocellular carcinoma (HCC), thus it is necessary to investigate the lethality of it worldwide. It was found that cirrhosis alone is currently the tenth-leading cause of death in Africa; ninth-leading cause in Southeast Asia and Europe and the fifth-leading cause of death in the Eastern Mediterranean. Additionally, the mortality risk of cirrhosis is increased due to complications, such as kidney dysfunction and infections, and liver failure.

Risk of developing alcoholic liver disease (ALD)

With a focus on patients specifically, it is important to outline the causes and effects of alcohol that eventually lead to alcoholic liver disease (ALD). Sinha and Sinha's (2024) argue the liver plays an important role in the body by regulating blood pressure, secreting clotting factors and storing minerals such as iron, folic acid and vitamin B-12 which are important in the development of red blood cells. They outlined that excessive alcohol consumption is a common factor of liver disease and causes direct bone marrow suppression which has a toxic effect on blood cell lines and causes iron to not be properly incorporated into hemoglobin molecules. Maner et al.'s findings (2019) agree with this idea and adds that chronic alcoholism leads to alcoholic cirrhosis, prevents folate absorption and prevents the liver from storing these minerals properly, creating a deficit and causes abnormalities in red blood cell development, which may lead to anemia. However, this idea is not completely accurate as in Singal et al.'s (2018) view only 10-20% of individuals that heavily consume alcohol develop liver disease and cirrhosis, thus other diseases and factors could be contributing alongside alcohol. These factors include being overweight for 10 years, already existing diseases such as chronic hepatitis B or an infection, smoking cigarettes and genetics. Thus, in order to diagnose ALD, there must be documentation of the patient's chronic heavy alcohol consumption, especially over a long period of time, and the exclusion of any other causes of liver diseases.

Excessive alcohol & other factors

Defining the category of excessive alcohol consumed allows for a more concrete idea that can be used during tests instead of an arbitrary statement.

As stated earlier by Singal et al. (2018), in order to diagnose ALD, there needs to be documentation on heavy alcohol consumption. They define excessive alcohol consumption as, in males, more than 5 drinks consumed over two hours. Additionally, males that consume more than three drinks per day are advised that they are at a higher risk of developing liver disease.

Diagnosis of liver disease

Significance of tests for diagnosis

In order to make an accurate diagnosis, the tests used to make the diagnosis must be reliable and more than one test must be used to make an informed decision. Singal et al. (2018) argues that patients often inaccurately report their alcohol intake. Thus, alongside questionnaires, it is best to also use biomarker tests or do a liver biopsy to check for alcohol-induced liver damage. It was found that the biochemical tests mean corpuscular volume (MCV), aminotransferases (SGOT and SGPT) and γ-glutamyl transferase (GGT) are sensitive tests but are not specific in patients that have cirrhosis.

The usefulness of the blood tests

Mean Corpuscular Volume (MCV)

The liver's direct link to the development of red blood cells underlines the importance of testing these cells in order to find out what is affecting the liver. Maner et al. (2019) defines mean corpuscular volume (MCV) as a laboratory value that measures the average size and volume of a red blood cell, with it gradually increasing with age and is typically higher in men than women. This value is mainly used to determine the underlying cause of anemia. The value MCV is measured in the metric femtoliters (fL) and normally ranges between 80 and 100 fL. MCV, alongside hemoglobin and haematocrit, classifies anemia into three main categories. These categories are microcytic, normocytic and macrocytic anemia, where MCV is below, within and above normal range respectively. Macrocytic anemia (>100 fL) is split into two subcategories called megaloblastic and non-megaloblastic, whereby megaloblastic anemia is a result of chronic alcoholism due to the folate deficiency in the body. Kumara et al.'s (2020) findings agree with this as their result of MCV ≥ 100 fL/L shows a strong bond with ALD and thus concludes that MCV would be a useful test for detecting ALD.

Alkphos & Gammagt (GGT)

The blood test alkphos is best when used together with gammagt (GGT) and gives an idea of whether or not the patient has ALD. According to Thapa and

Walia (2007), when the liver is damaged, it is unable to excrete alkaline phosphate (alkphos) that is typically made in bones, intestines and the liver, creating a lack of naturally produced alkphos in the body. This makes using alkphos a good blood test for detecting ALD as chronic alcoholism targets those areas in the body however this elevation in the test creates a challenge. Alkphos can only narrow the illness' source down to the liver and bones but cannot distinguish which one it is. Through investigation, it was found that it is necessary to use GGT alongside alkphos in order to distinguish between the two, as GGT only increases in "cholestatic" disorders but not in bone diseases. While GGT should be used when using alkphos, Kumara et al. (2020) adds that it can also be used on its own as it has a strong bond with ALD shown when GGT is greater or equal to 25 IU/L.

SGOT & SGPT

While other tests are useful in the detection of whether or not it's the liver that is afflicted, it is necessary to also find out what stage the liver disease has developed to. Nyblom (2004) puts forward the idea that the blood tests SGOT and SGPT are typically used as a ratio when detecting liver disease. Upon investigation, it was discovered that a high SGOT/SGPT ratio above 1 was found in patients with advanced liver diseases while a high ratio below 1 indicates high alcohol consumption but without severe liver disease. It was concluded however that it is not the alcohol dose that flags these blood tests but rather other factors. Thapa and Walia (2007) clarifies this measure of elevation in SGOT and SGPT to be moderate (3-20 times) in advanced liver disease and adds that SGOT is typically higher than SGPT in chronic liver disease.

Constraints of the blood tests

Although Singal et al. (2018) notes that these blood tests are best as an initial test, Liangpunsakul et al. (2010) argues that these blood tests, both when they are used alone or in combination, produces predictive values that are too low to predict heavy drinkers. Savage et al. (2000) counterargues as their findings found that, in using MCV where the result falls into the macrocytic category, drug therapy is the leading cause with alcohol ranking second.

Conclusion:

This literature review has addressed the importance of diagnosing liver disease and biomarkers tests needed for diagnosis. With the fatality of this disease being at an all-time high, it is all the more important to have a reliable test that can accurately diagnose it. From the different investigations and views, we can

conclude that MCV and GGT are the most powerful blood tests in confirming the presence of alcoholic liver disease (ALD), followed by SGOT and SGPT in detecting advanced liver disease. Although alkphos is good for narrowing the patient's ailment down to a few areas of the body, it is less significant without GGT to target one specific area. As it is necessary to consider the constraints of these blood tests, it is important to keep in mind that these tests are best used as an initial investigation into whether the patient has developed liver disease or not.

UNIVERSITY of the WESTERN CAPE

Research Questions

Main Research Question - Which Blood tests are good predictors that a physician might use to inform their diagnosis?

Created Research Questions:

1. Is MCV a strong predictor for chronic alcohol consumption, and is it useful in the diagnosis of alcoholic liver disease?

Methodology:

Introduction:

Research methodology offers a structure on how the research will be conducted and details the procedures, principles, and rationale that will be followed in the process of addressing the key questions of the research. This paper uses a quantitative method to examine the correlation between the results of routine blood tests and alcohol use based on UCI BUPA Liver Disorders data. The dataset, which is actively utilised in medical and computational studies, offers biochemical cues which can be statistically modelled to be used as notifications of alcohol consumption patterns.

The subject is of importance both in medical and data science. Biomarkers of alcohol-related liver conditions are used to diagnose alcohol-related liver conditions by clinicians, whereas data analysts and computer scientists usually use this dataset to test classification algorithms. Nevertheless, one of the variables in the dataset has been misinterpreted in many previous studies, which have given incorrect results about the alcohol consumption. This study rectifies that fallacy and leads to better clinical interpretation as well as more precise analytical modelling.

The methods selected will focus on answering the three research questions of the study: What relationship exists between particular blood test results and reported consumption of alcohol? Are such blood tests accurate predictors of the extent of alcohol consumption of a person? What can be understood based on correcting the usual variable misconceptions that have been committed in past research?

The methodology follows a systematic procedure: (1) Descriptive Analysis summarizes the patterns in the data; (2) Correlation Analysis measures the linear relationship; (3) Multiple Regression predicts alcohol consumption; (4) ANOVA evaluates the variance explained; (5) Hypothesis Testing evaluates the statistical significance; (6) Model Selection models the predictors; and (7) Model Assessment measures the overall performance of the model. All these measures possess a detailed quantitative model of which biomarkers are most effective predictors of alcohol consumption.

Describing methods

Descriptive Analysis of Biomarkers

In order to summarize and comprehend the central tendencies, variability, and distributional properties of each biomarker and the dependent variable, DRINKS, descriptive analysis was done. The given stage answers the first research question by recognizing whether there is any trend in blood test measures that are in line with alcohol intake as was proposed in the previous studies.

Mean, standard deviation, skewness, and kurtosis were determined to describe the data distribution and outliers could be viewed graphically through histograms and boxplots. DRINKS frequency groupings were also developed in order to determine the patterns of consumption.

The review showed that the stability of biomarkers was not similar. The variable of MCV had low variation and almost normal distribution and the enzymatic biomarkers (SGPT, SGOT, and GGT) had positive skewness and heavy-tailed distributions. These data agree with Alatalo et al. (2008), who reported that moderate drinking with increased BMI is likely to increase GGT and ALT, and Agarwal et al. (2016), who found out that GGT is the most useful biomarker of alcohol consumption.

The descriptive analysis found that MCV is an insensitive hematological variable, but enzymatic markers, especially GGT, have a significant variability and skewness in relation to alcohol consumption. These findings will serve as a basis to the further correlation analysis as they will show which biomarkers should be subject to more intensive inferential testing

Correlation Analysis

After the descriptive analysis, the correlation analysis was used to establish the strength and direction of linear correlations between alcohol consumption (DRINKS) and each biochemical biomarker. This discussion is directly relevant to Research Question 1, which is whether an increase in the level of enzymes in the blood is linked to an increase in alcohol consumption.

To measure the relations between DRINKS and the biomarkers (MCV, ALKPHOS, SGPT, SGOT and GGT), Pearson correlation coefficients were calculated.

It is indicated that GGT has the biggest correlation with alcohol consumption (Agarwal et al., 2016), whereas MCV has moderate but significant associations. This trend was validated through the correlation in this dataset since the

positive correlation of GGT to DRINKS was the highest, followed by SGPT and SGOT. The relationship of ALKPHOS was weaker or an inverse relationship, which is likely to be in agreement with the fact that the substance reduces marginally as the level of alcohol consumption increases.

Correlation analysis has found GGT as the variable that has the greatest association with alcohol consumption making it a worthy predictor in future regression modelling. This analysis therefore reduces the area of study to biomarkers that have statistically significant linear relationships.

Full Model Multiple Regression

Based on the results of the correlation, multiple regression was used to forecast the alcohol consumption through all biomarkers together. This subtopic reacts to Research Question 2, which evaluates the idea of whether combinations of the biomarkers can be used to predict the drinking levels as effectively as the individual variable.

The model developed with DRINKS as the dependent variable and the five biomarkers as the predictors can be seen in equation 2. The coefficients were estimated, and the proportion of variance explained (R-squared) was determined using this complete model. The residual was checked to determine linearity, independence, and normality.

Further to explain the contribution of each variable, the analysis of variance components was calculated and the results were the Sum of Squares of Error (SSE), Total (SSTO) and Regression (SSR). The ratio SSR/SSTO was used to show the percentage of the variance in DRINKS that the model accounts. Liangpunsakul et al. (2010) argue that such a breakdown is informative on the contribution of every biomarker to alcohol consumption variance.

The entire regression equation brought out that a combination of GGT and MCV explained about 21 percent of the variance in DRINKS. This proves the superiority of GGT as a predictor variable and that MCV is still a supportive and stabilizing factor. The results of an analysis of variance data confirm the significance of the model and establish the procedure to test and revise the choice of the predictors.

Analysis of Variance (ANOVA)

The statistical evaluation of differences in model performance concerning variable interpretation and classification accuracy was done through the Analysis of Variance (ANOVA) technique. This test helps to address Research

Question 3 by testing the impact of methodological decisions, like, using the correct or incorrect variable, on the outcomes.

To compare the accuracy of the classification between the studies which appropriately used the dichotomized drinks variable (x6) and those that used incorrectly the selector field (x7), a one-way ANOVA was made. The dependent variable used was the highest classification accuracy of each study. The ANOVA used by Welch was necessitated by unequal sample sizes, so that heteroscedasticity is not compromised. Effect size was measured by eta-squared where an alpha level of 0.05 was utilized to test significance.

The findings verified a strong difference in the mentioned accuracy of the reported models, provided that the applications were correct and incorrect, and this fact gives a certain importance to methodological accuracy in selecting the variables.

The ANOVA test confirmed that wrong use of variables has a significant impact on distorting classification accuracy. This validates the intention of the current paper, which is to rectify these methodological errors and give a valid model of interpreting liver disorder data.

Hypothesis Testing

Testing of hypotheses was conducted to see whether any of the predictors or the entire regression model were statistically significant. This will allow verification of the predictive strength of the model and answer Research Question 2 with the goal of determining which biomarkers can statistically significantly influence alcohol consumption.

Single coefficients were tested with T-tests and overall significance of the model was tested with F-tests. The null hypothesis was that the blood tests did not have any effect on alcohol consumption, and the alternative hypothesis was one or more (Bevans, 2020). The p-values of below 0.05 accepted the predictors as significant contributors.

Skewness and kurtosis were also available to test the importance of distributional shape. The Jarque-Bera test was used to test the deviations of normality and skew and kurtosis were combined into a two-degree chi-square test. The identification of non-normality was useful in making decisions concerning data transformations or nonparametric alternatives.

The significance of GGT and MCV as predictors was statistically significant and the normality testing proved the importance of close interpretation of highly

skewed variables. The results of these findings guarantee that the results of regression model are statistically and interpretively sound.

Selections and Multiple Regression (Reduced Model) Methods

Once the important predictors were identified, variable selection methods were used to parsimonize and to ensure the regression model. This part is aimed at developing a smaller model that leaves only the most powerful predictors.

Forward selection, backward elimination and stepwise regression were used. Such repetitive processes eliminated the insignificant predictors and retained the ones with a good explanatory capacity. The lower model narrowed down to biomarkers with predictive consistency—mainly GGT and MCV.

This is a refined approach that is supported by previous studies. Alatalo et al. (2008) stated that both BMI and GGT are significant factors contributing to alcohol-related variance, whereas Agarwal et al. (2016) affirmed that GGT remains sensitive even in the moderate intake levels.

The simplified model was a succinct and statistically effective alcohol prediction model. It provided a viable but valid predictive model that strikes a compromise between interpretability and accuracy by only retaining GGT and MCV

Model Assessment

Performance of the model in terms of R-squared and adjusted R-squared was measured by the proportion of the variance in DRINKS as explained by the predictors. Large R-squared values indicate a good model fit, whereas comparisons of full and reduced models imply tradeoffs between model complexity and goodness of fit. Since it ascertains the degree to which the biomarkers can be relied on to describe variation in alcohol consumption, this assessment has a direct significance on the research questions. According to the literature, GGT often inflates the contribution to model fits out the rest (Agarwal, Fulgoni, & Lieberman, 2016). Overall, the analysis of the model supports the practical explanatory potential of particular biomarkers.

Altogether, this study design provides a robust approach to understanding the association between alcohol intake and liver biomarkers since it integrates the characteristics of descriptive, correlational, regression, and model testing. The two approaches share strengths; distributional properties can be determined with descriptive analysis, direct relationships can be determined with correlation, the predictive ability of a relationship can be determined with regression and ANOVA, statistical significance can be determined with

hypothesis testing, and the set of predictors can be determined with model selection.

Collectively, these processes will answer the central and formulated research questions so that the research is contextualized within a larger literature that considers GGT and the different enzymes an important marker of alcohol consumption

Multicollinearity

Multicollinearity is a term that is used to describe the phenomenon where the regression model has several independent variables that are significantly correlated not only with the dependent variable but also with each other. The consequences of having multicollinearity present in the model includes: inaccurate and unreliable results when investigating the impact of a single variable on the model, increased/inflated standard error in the model that changes the results of the analysis and an increase in variance of coefficients which makes them unstable. (Shrestha, 2020; Schreiber-Gregory, 2017).

If multicollinearity is present in the regression model that is currently being investigated, it becomes difficult to tell which blood test shows the strongest indicator due to the high correlation. Hence, to get a clearer answer on which blood test(s) is a strong indicator for liver disease it is crucial to test if multicollinearity exists in the model and solve it.

Techniques for Detecting Multicollinearity

- **Pearson Correlation Coefficient**: determine the correlation score and if it has an absolute value close to 0.8, collinearity likely exists
- **Variance Inflation Factor (VIF)**: “measures how much variance of the estimated regression coefficient is inflated if the independent variables are correlated” (Shrestha, 2020). After imposing VIF, tolerance and collinearity measures on the regression model, investigate the variance inflation. If VIF = 1 independent variables are not correlated, 1 < VIF < 5 variables are moderately correlated, 5 <= VIF <= 10 highly correlated, 10 < VIF regression coefficients are "feebly estimated with the presence of multicollinearity" (Daoud, 2017)
- **Tolerance**: the amount of variability in one independent variable that is not explained by the other independent variables. Is the reciprocal of VIF and when investigating the tolerance, if tolerance < 0.1, it indicates multicollinearity
- **Eigenvalue Method**: is the variance of the linear combination of the variables. The sum of eigenvalues must equal the number of independent variables, however since interpreting an eigenvalue close

to 0.05 is difficult, the conditional index can be used instead. If one or more eigenvalues are close to zero and the condition number is large then multicollinearity is present. (Schreiber-Gregory, 2017)

Selection Techniques

In statistical modelling, particularly with multiple predictors, it is crucial to identify a parsimonious model that balances complexity with predictive power. Automated selection techniques provide a structured, algorithm-driven approach to this problem. The three primary methods are:

1. Forward Selection

This is a bottom-up approach that starts with no variables and adds them one at a time.

- **Process:** The algorithm begins with an empty model. It then tests each available predictor by fitting a simple regression model and then identifies the one that provides the most statistically significant improvement to the model's fit (e.g. the one with the smallest p-value, provided it's below a threshold of 0.05). This variable is then added. The process repeats with the remaining variables, testing each one's significance in the presence of the variable already in the model. The algorithm stops when none of the remaining variables are statistically significant enough to add.

2. Backward Selection

This is the inverse, a top-down approach that starts with all variables and removes them one at a time.

- **Process:** The algorithm begins with the "full model" containing every predictor. It then identifies the least statistically significant variable (e.g., the one with the largest p-value, provided it's above a threshold like 0.10) and removes it. The model is refitted without this variable and the process repeats. It stops when all variables remaining in the model are statistically significant.

3. Stepwise Selection

This is a hybrid approach that combines the principles of both forward and backward selection, offering the most flexibility.

- **Process:** The algorithm starts similar to forward selection, adding the most significant variable. However, after each new addition, it performs a check of *all* variables already in the model to see if any have become redundant due to the new relationship. If any variable's significance has

dropped below a removal threshold, it is taken out. This process of adding *and then checking for removal* continues until no more significant variables can be added and no non-significant variables remain in the model.

These techniques help avoid overfitting (a model that is too complex and fits the sample noise) and underfitting (a model that is too simple and misses important relationships). The goal is to produce a final model that is both statistically sound and clinically interpretable, highlighting the most relevant biomarkers for predicting alcohol consumption.

Assessing the Model Fit and Predictive ability

Once an appropriate regression model is identified, it is essential to evaluate how well the model describes the observed data and whether it has meaningful predictive power. Assessing model fit ensures that the regression equation is not only statistically significant but also practically useful for predicting alcohol consumption using blood biomarkers.

Why this method is used:

Model assessment allows us to determine if the reduced regression model provides a good balance between accuracy and simplicity. Without this step there is a risk of drawing misleading conclusions, either by overfitting the model to the sample data or by underfitting and failing to capture relevant relationships. By evaluating mode fit we ensure that the blood test variables identified as predictors can realistically inform clinical practices, which directly answers the research question of which blood tests are good predictors of alcohol intake.

How the method will be applied:

1. Coefficient of Determination (R^2): This statistic quantifies the proportion of variability in alcohol consumption that is explained by the selected blood test predictors. A higher R^2 suggested that the model captured more variance suggesting stronger predictive utility
2. Adjusted R^2: Since additional predictors can inflate R^2, the adjusted version is applied to account for model complexity. This ensures that only predictors contributing meaningful explanatory power are retained.
3. Residual Analysis: Residual plots are inspected to check for patterns that would indicate violations of regression assumptions.
4. Predicted vs Actual Values: Scatterplots of predicted alcohol consumption against observed values provide a visual check of how well the model predictions align with actual data points. Points close to the diagonal line indicate strong predictive ability.

Why this method answers the research question

By systematically evaluating both fit and predictive accuracy this approach verifies whether biomarkers such as MCV, GGT and others can be relied on in a clinical setting. A model that demonstrates strong fit statistics and predictive performance supports the conclusion that specific blood tests can serve as reliable indicators of alcohol intake. On the other hand, poor fit would suggest that these markers alone are not sufficient predictors. Thus, the model assessment provides the critical evidence needed to judge the usefulness of liver function tests as diagnostics tools.

Conclusion:

The methods presented in this methodology have shown that our exploration and analysis of the dataset is justifiable. The first exploration of the data using descriptive analysis, correlation analysis, skewness & kurtosis and ANOVA were used in order to observe the trends of the distributions and decipher if it seemed plausible. In order to test if the predictors were too similar, we tested for multicollinearity before adjusting the dataset to ensure more accurate results. The adjusted data was fitted into a full model regression before the model was assessed in order to judge the fit, goodness of fit and if the model supported initial assumptions. For a more powerful model with more accurate predictions and to narrow down the most significant predictors, the full model is put through selection techniques to form a reduced model. Finally, in order to ensure reliability, the reduced model was used in hypothesis testing in order to ascertain with confidence which blood test predictors can be used by physicians to inform their diagnosis.

Initial Analysis - Results and Discussion

Introduction

This section details the step-by-step analysis we used to answer our research questions. Each technique serves a specific purpose, building upon the last to give a complete picture of what the data tells us.

We start with descriptive statistics to get a basic feel for the data. This helps us summarize the key characteristics of our variables, like the average blood test values, and see how spread out the numbers are. This initial step lets us check if our dataset looks similar to those used in other studies. Next, we use correlation analysis to see how pairs of variables like a specific blood test result and alcohol consumption, move together. This helps us identify initial relationships and see if patterns found in previous research also appear in our data.

We then move to multiple regression modelling , which is the core of our analysis. This allows us to see how all the blood test variables work together to predict alcohol consumption. We will test different versions of this model to pinpoint which biomarkers are the most important predictors. Here, we will constantly compare our results to existing literature to see if our findings make sense based on what is already known.

Finally, we use hypothesis testing and ANOVA to confirm that the patterns we see are statistically significant and not just due to random chance. We also use metrics like R-squared to evaluate how well our models actually explain the variations in alcohol consumption.

The goal of this entire process is to provide solid numerical evidence for our conclusions and to directly connect our findings back to existing research, showing where we confirm previous ideas and where our analysis might offer new insights.

Descriptive analysis

Dataset: Alcohol Consumption and Liver Biomarkers

MCV (Mean Corpuscular Volume)

Mean Corpuscular Volume (MCV) is a hematologic measure of the extent of red blood cells and can often be researched in connection with alcohol consumption. High MCV is linked to prolonged alcohol abuse owing to the impact of alcohol on the erythropoiesis. Within the framework of our

investigation, the study of MCV will give the opportunity to analyse the possibility of utilizing blood cell indexes as trustworthy signals of alcohol use. The question that will guide this analysis is the following, “Does an increase in MCV indicate an increase in alcohol consumption and how does it compare to other biomarkers”

MCV created descriptive statistics. The median was 90, the mean was 90.16 fL and the mode was 91. The variance was 20.90 and the standard deviation was 4.57. The minimum value was 37 and the second quartile (IQR) was 6. Skewness was -0.47, which was a weak left skew and kurtosis was 2.72, which was a moderate heavy tail. The coefficient of variation (CV) was 5.07 with low relative variability. The histogram (Figure 1) appears to be concentrated around the 90 fL point, indicating a normal distribution, whereas the boxplot (Table 1) indicates that there are quite a few outliers at the low end, below 78 fL.

The findings reveal that MCV is relatively steady, and its skew is not very high with just significant variability. This is in line with Staufer and Yegles (2016) who found that MCV is not sensitive but stable relative to enzymatic biomarkers. In our dataset, MCV does not have excessive variation, nevertheless, it is a helpful supportive variable, but not a predictor of excessive alcohol consumption.

ALKPHOS (Alkaline Phosphatase)

Alkaline Phosphatase (ALKPHOS) is a liver enzyme which is involved in bile metabolism. It is not the most urgent signifier of the alcohol consumption, yet still it can provide some data about the health and functioning of the liver as a whole. High ALKPHOS levels might reflect a liver metabolic stress, which might indirectly result in alcohol consumption. In this case, the research question is as follows: Does the ALKPHOS differ significantly with the alcohol intake, and the distribution of ALKPHOS regarding other biomarkers?

The ALKPHOS descriptive statistics showed that the mean was 69.92 U /L, the median was 67, and the mode was 62. It had a standard deviation of 18.48 and a variance of 341.53. It was 115 with a minimum and 22 as the interquartile range (IQR). The skew of the data was 0.73, and it was moderately right skewed. The kurtosis was 0.69, which was lower than that of a normal distribution, meaning tails. (Figure 2) histogram states the distribution about the mean, which is continuous at higher values, whilst (Table 2) states the outliers at higher values.

The distribution of the statistics reveals that ALKPHOS is skewed slightly to the right with few extreme values, although most of the values are within a normal physiological range. This tendency is comparable to the ones of Agarwal et al.

(2016) and Staufer and Yegles (2016) where ALKPHOS was not identified as a sensitive biomarker of alcohol intake, but it still offered inter-group variability. Thus, ALKPHOS is another intervention that will be implemented in the case of alcohol-related liver shift of enzymes.

SGPT (ALT)

Serum Glutamate Pyruvate Transaminase (SGPT/ALT) is one of the liver enzymes that are most commonly utilized in a clinical assessment. ALT is highly sensitive to the damage of the hepatocellular and alcohol damage. The question of the research is: Is there a rise in the ALT among the heavy or moderate drinkers and can it be reliably predicted as a predictor of alcohol consumption?

The median of 25.50 and mode 17, mean ALT was 30.35U/L. Standard deviation was 19.90 with variance 395.86 and range 151 (max = 155, min= 4). This was 15 as an interquartile range (IQR). The skewness is 3.22 and is a strong right skewness and the kurtosis is 14.97 that is heavy tails. Normality was rejected ($p < 0.01$) in all tests of normality (Kolmogorov-Smirnov, Anderson-Darling, Cramer-von Mises). The histogram and images in (Figure 3 , Table 3 , Table 4) show that skewness is negative, and extreme outliers are beyond 100 U/L.

These results validate that ALT is highly skewed and with extreme outliers such as was the case with Agarwal et al. (2016) and Alatalo et al. (2008); ALT levels were so high in moderate drinkers and more so with high BMI. This makes ALT sensitive biomarker of alcohol consumption, but variations and non-normality should be considered during prediction model.

SGOT (AST)

The other significant transaminase that is also linked to liver health is Serum Glutamate Oxaloacetate Transaminase (SGOT/AST). Now, AST is a less sensitive test, but it increases with hepatocellular damage, and it could be caused by stress with alcohol. The research question is: How does the AST distribution correlate with the alcohol consumption patterns of other enzymes?

The median and mode of the AST were 22 and 20 with an average of 24.57 U/L. The standard deviation of 10.56 with the range of 77 (min = 5, max = 82) was the variance. The IQR was 8. Its skew was 2.39, which was a moderate value of right skewness, and kurtosis was 8.28, which indicated heavy tails. The K-S and Anderson-Darling tests revealed that the deviation from normality

is significant ($p < 0.01$). The skewness and the outliers above 50 U/L are approved by (Figure 4 , Table 5, Table 6).

AST was biased towards the positive and heavily tailed, hence indicating that the vast majority of the values were average, with a subgroup containing higher values, which showed that they were stressed with alcohol. These results are indicative of Agarwal et al. (2016) and Alatalo et al. (2008) who found that AST was a poorly sensitive and moderately responsive test in comparison with GGT. AST consequently plays a supportive role of anticipating liver damage as a result of alcohol.

GGT (Gamma-Glutamyl Transferase)

Gamma-Glutamyl transferase (GGT) is the most sensitive alcohol consumption biomarker that increases even in the case of moderate alcohol consumption. The clinical importance of it is that it is sensitive to the exposure to alcohol. The research question will be: Is GGT open to extreme variability as is its position as the most powerful alcohol biomarker?

The median GGT was 24, mode 11 and mean was 36.95 U/L. This was between 292 (maximum=297), standard deviation= 38.81 and variance=1506.40. The IQR was 27. The skew was 3.12 showing that skewness was very strong to right skewness, and the kurtosis was 12.41, indicating very heavy tails. Nor was normality (Kolmogorov-Smirnov, Anderson-Darling, Cramer-von Mises) ($p < 0.01$). The skewness and outliers are illustrated in (Figure 5, Table 7), particularly in the outliers that are above 200 U/L.

GGT had extreme skew and kurtosis that justified the sensitivity and variability of GGT being the best biomarker. This is affirmed in literature (Agarwal et al., 2016; Alatalo et al., 2008) GGT has been pointed out to be the strongest predictor of alcohol consumption. It is a significant yet volatile indicator of regression equations because it is highly responsible and generally has to be converted to be examined appropriately.

DRINKS (Daily Alcohol Intake)

The dependent DRINKS variable is the number of drinks that participants report to take every day. To comprehend the issue of drinking behaviour in the dataset and identify heavy drinkers, its distribution is critical to examine it. The research

question will be as follows: Is there skewed distribution of the reported alcohol intake with a small group of heavy drinkers creating variability?

The median of DRINKS was 3.00 with the average of 3.41 per day and the mode of 0.50. The range was 20 and the standard deviation was 3.28. The skewness was 1.61, which means moderate skew towards the right including kurtosis of 4.32, which means heavy tails. According to (Figure 6, Table 8, Table 9), most participants had a low number of drinks/day (less than 5) and also a small number of participants had significantly more. (The frequency table) ensured that the respondents were majorly moderate and light drinkers, with a very minor percentage of respondents exceeding 5 drinks/day.

DRINKS was skewed, with the highest number of the respondents being light-to-moderate drinkers and a few subsets that were heavy drinkers. This is this same way with Agarwal et al. (2016), where high-medical-risk heavy drinkers were in the minority. The DRINKS variable, therefore, makes a good candidate as the predictive model's dependent measure.

Conclusion

The descriptive analysis shows that MCV and ALKPHOS are less unstable and sensitive and ALT, AST, and, especially, GGT have high skew and wide tails and extreme values, which are linked to the increased alcohol consumption. These were echoed on the dependent variable DRINKS where a minor subgroup of heavy drinkers generated the skewness in the data. As in other studies done before (Puuka, 2007; Alatalo et al., 2008; Agarwal et al., 2016; Staufer and Yegles, 2016), GGT and ALT were found to be the most predictive factors of alcohol use. These results make them justify their use in the further correlation and regression analysis.

Correlation Analysis

To narrow down which blood tests can potentially be used to predict males' alcohol consumption, we used correlation analysis. In order to conduct this analysis, we used Pearson Correlation Coefficient which is represented by r. The Pearson Correlation coefficient was also used to prove that these blood tests are reliable, as blood tests that have a strong relationship with drinks have a stronger correlation. (Bagiella, 2012, pp. 4-5)

The procedure for using this method is by measuring the linear relationship between variables which ranges from -1 to 1. The r-value will then be

interpreted using the conventional approach and cutoff points. (Schober, Boer & Schwarte, 2018)

The following Pearson Correlation Coefficient matrix contains the results of the correlation between the explanatory variables and response variable, as well as the relevant p-values. It is extremely clear from the table that the correlation between drinks (the response variable) and the explanatory variables is not strong.

The results of the Correlation matrix indicates that while the relationships between the blood tests and drinks are mostly statistically significant, the correlation descriptive label for these relationships are not stronger than a "weak or lower correlation" as seen in these relationships: sgot and drinks ($r=0.26$, $p<0.0001$), MCV and drinks ($r=0.36$, $p<0.0001$), gammagt and drinks ($r=0.37$, $p<0.0001$). It is clear that sgot, MCV and gammagt are variables that are positively related to drinks, but a single blood test is not a strong predictor of alcohol consumption. Hence, it is necessary to include these variables in a multiple regression model to have stronger predictive power. (Refer to Table 10)

The selection of blood tests with a strong correlation with drinks ensures a stricter test for measuring drinks consumed which allows physicians to come to a more accurate diagnosis of diseases. A high accuracy in diagnosis is important in a field that handles many patients' health, and in turn, life.

It is crucial to note that the relationship between blood tests is statistically significant and have a moderate/strong correlation. This can be seen in these relationships: sgpt and sgot ($r=0.74$, $p < 0.0001$, strong correlation), sgpt and gammagt ($r=0.53$, $p < 0.0001$, moderate correlation), sgot and gammagt ($r=0.57$, $p < 0.0001$, moderate correlation). A high correlation between explanatory variables could be a sign of multicollinearity which will be investigated in the multicollinearity section.

The evidence of the correlation results shows that the individual variables sgot, MCV and gammagt have a stronger positive correlation with alcoholic drinks consumed, compared to other variables. However, the strong correlation between the blood tests sgpt and sgot needs to be investigated for multicollinearity.

Reduced Multiple Regression Techniques

Model Selection and Specification

To identify the most parsimonious and statistically robust model for predicting alcohol consumption using liver biomarkers, automated stepwise selection procedures (forward, backward, and stepwise) were employed. Consistent with the findings in the literature review, which highlighted the particular diagnostic value of Mean Corpuscular Volume (MCV) and Gamma-Glutamyl Transpeptidase (GGT), all three selection techniques converged on an identical reduced model. The final model excludes the predictors alkphos, sgpt, and sgot, retaining only MCV and gammagt as significant predictors.

Refer to Table 10 and Table 11 for results.

The Estimated Reduced Model

Refer to **Error! Reference source not found.** for the estimated regression equation.

This model indicates that, holding the other variable constant: For every one-unit increase in MCV, alcohol consumption is predicted to increase by 0.21 units and for every one-unit increase in Gammagt, alcohol consumption is predicted to increase by 0.03 units. The negative intercept is a reference point for the model's calculation and has no practical meaning in this context, as biomarker values cannot be zero.

Statistical Significance and Model Fit

The summary table of the final model confirms the high significance of both predictors:

MCV: ($F(1, 342) = 27.02$, $p < .0001$) **mmagt:** ($F(1, 342) = 28.88$, $p < .0001$)

The stepwise selection process, detailed in the summary table, reveals how the model's explanatory power was built:

Step 1: gammagt was entered first, explaining **13.33%** of the variance in alcohol consumption (Model $R^2 = 0.1333$). This initial step was highly significant ($F = 42.16$, $p < .0001$).

Step 2: MCV was added, contributing an additional **7.80%** to the explained variance. The final reduced model, with both predictors, has a combined R-Square of **0.2114**.

This means that the two biomarkers, MCV and Gammagt, together account for approximately **21.14%** of the variance in alcohol consumption within the

sample. The low Mallows' C(p) value of 2.59, which is very close to the number of parameters in the model (including the intercept), indicates that the model has a good balance of bias and variance, confirming it as a superior choice over the full model with non-significant predictors.

Interpretation and Clinical Relevance

The results of the statistical analysis provide strong empirical support for the thesis and literature review. The finding that MCV and Gammagt are the sole significant predictors in a multivariate context aligns directly with prior clinical research:

- It validates the assertion by Papoz et al. (1981) that the combined use of MCV and GGT provides a powerful and accurate prediction of alcohol consumption.
- The significance of Gammagt confirms its well-established role as a sensitive marker for chronic alcohol intake (Puuka, 2007; Fakhari & Waszkiewicz, 2023).
- The retention of MCV underscores its utility as an indicator of alcohol-induced macrocytosis (Liangpunsakul et al., 2010; Savage et al., 2000).

The exclusion of sgot and sgpt from the final model is equally insightful. It suggests that while these enzymes are important general markers of liver cell damage, their unique contribution to predicting the specific behaviour of *alcohol consumption* is not statistically significant when MCV and GGT are already accounted for. This finding emphasizes the importance of using targeted biomarker panels rather than general liver function tests for assessing alcohol intake.

In conclusion, the reduced multiple regression model offers a statistically sound, simplified tool for prediction. It enhances clinical interpretability by focusing on the two most relevant biomarkers, thereby providing a credible, evidence-based foundation for screening and assessment related to alcohol consumption.

Multicollinearity

The method of detecting multicollinearity was used in order to determine the accuracy of the blood tests and their ability to predict the number of drinks consumed. The procedure consists of checking three values: tolerance value, variance inflation (VIF) and eigenvalue.

Multicollinearity is present in the dataset if: a single tolerance value falls below 0.1, a single VIF value is above 10, the eigenvalue is close to 0 but the corresponding condition number is large for that variable. (Schreiber-Gregory, 2017)

Identifying presence of multicollinearity

When reviewing tolerance, no value falls below 0.1, so there is no multicollinearity. The maximum variance inflation (VIF) value is 2.525, with all variables' VIF value being moderately correlated (1< VIF < 5) so there is no indication of multicollinearity here either. (Refer to Table 13)

Numbers 3, 4, 5 and 6 have small eigenvalues close to zero with a corresponding condition number that is large. This indicates that multicollinearity is present. It is apparent that the strong correlation found between the variables sgpt and sgot could be the cause of the multicollinearity. However, despite the eigenvalues indicating that these variables cause multicollinearity, the VIF assures that the moderate correlation between them is not a cause for concern. We can conclude from this information it is not necessary to rid the model of multicollinearity but rather take into consideration that the variables SGPT and SGOT would be best used together, proving Nyblom's (2004) idea correct. (Refer to Table 14)

We can conclude that, although multicollinearity is present in the model, it does not threaten its accuracy. Thus, the results produced by the model are confirmed to be reliable.

Summary and Conclusion

This study aimed to determine which blood tests are most effective for predicting alcohol consumption levels. The analysis was conducted in several stages:

Data Exploration

Each blood test result (MCV, GGT, etc.) and the corresponding reported alcohol intake were examined. Measures of central tendency and variability were calculated to assess spread and distribution patterns. The majority of participants displayed values within the normal range, but a small subset had very high results which resulted in the skewness of the graphs. Such skewness is common in medical datasets.

Simple Relationships

Correlations between individual blood tests and alcohol intake were then assessed. While MCV, SGOT and GGT do have statistically significant relationships with alcohol consumption, the individual connections on their own are not reliable as you cannot predict alcohol intake using just one blood test.

Combining Tests (Multiple Regression)

To improve predictive accuracy, multiple regression analysis was conducted to assess how the blood tests performed collectively. The results were clear, only the two tests MCV and GGT were significant predictors of alcohol intake. The other remaining tests (ALKPHOS, SGPT, SGOT) did not add any useful information once MCV and GGT were considered. The resulting model can be seen in equation 2. This model accounts for about 21% of the variation in alcohol consumption. Therefore, this means that MCV and GGT together give a moderate but not comprehensive indication about a person's alcohol intake.

Variable Relationships

Some blood tests like SGPT and SGOT were strongly related to each other, which can sometimes cause problems in the analysis (multicollinearity). Nevertheless, the simple two-variable model (MCV and GGT) remained the clearest and most effective representation of the data.

Conclusion

After thorough testing, the findings indicate that Mean Corpuscular Volume (MCV) and Gamma Glutamyl Transferase (GGT) is the most effective set of blood markers for estimating alcohol consumption. This supports what previous medical studies have suggested.

Another key takeaway is that simplicity is better, a model that uses only these two key tests is more effective and easier to understand than a more complex one that includes all available blood test results. While the model does not fully capture all factors influencing alcohol intake, it provides medical professionals with a strong evidence-based tool for assessment.

Overall, if a doctor wants to use blood tests to screen for alcohol intake, they should focus on the MCV and GGT results, as they offer the most reliable insight when used together.

Bibliography

Agarwal, S., Fulgoni, V. L., & Lieberman, H. R. (2016). Assessing alcohol intake and its dose-dependent effects on liver enzymes by 24-h recall and questionnaire using NHANES 2001–2010 data. *Nutrition Journal, 15*(62), 1-12. doi:https://doi.org/10.1186/s12937-016-0180-y

Alatalo, P. I., Koivisto, H. M., Hietala, J. P., Puukka, K. S., Bloigu, R., & Niemelä, O. J. (2008). Effect of moderate alcohol consumption on liver enzymes increases with increasing body mass index. *American Journal of Clinical Nutrition, 88*(4), 1097-1103. doi:https://doi.org/10.1093/ajcn/88.4.1097

Bevans, R. (2020, February 20). *Multiple Linear Regression | A Quick Guide (Examples)*. Retrieved from Scribbr: https://www.scribbr.com/statistics/multiple-linear-regression/

Boslaugh, S., Rudd, R. E., & Anderson, J. (2008). Sage Research Methods. (S. Boslaugh, Ed.) *Encyclopedia of epidemiology, 2*(1), 7. doi:https://doi.org/10.4135/9781412953948

Daoud, J. I. (2017). Multicollinearity and regression analysis. *Journal of Physics: Conference Series. 949*, pp. 1-7. IOP Publishing. doi:https://doi.org/10.1088/1742-6596/949/1/012009

Devarbhavi, H., Asrani, S. K., Arab, J. P., Nartey, Y. A., Pose, E., & Kamath, P. S. (2023). Global burden of liver disease: 2023 update. *Journal of Hepatology, 79*(2), 516-537. doi:10.1016/j.jhep.2023.03.017

Fakhari, S., & Waszkiewicz, N. (2023). Old and new biomarkers of alcohol abuse: narrative review. *Journal of clinical medicine, 12*(6), 2124. doi:https://doi.org/10.3390/jcm12062124

Fleming, L., Dhingra, R., & Chowdhury, S. (2019). Descriptive statistics in biomedical research: Guidelines for analysis and interpretation. *International Journal of Research in Medical Sciences, 7*(10), 4001-4008. Retrieved from https://www.msjonline.org/index.php/ijrms/article/view/7405

IBM. (2023). *Model assessment.* IBM SPSS Modeler Documentation. Retrieved from https://www.ibm.com/docs/en/spss-modeler/18.6.0?topic=viewer-model-assessment

James, E. (2019, June 18). *What is correlation analysis? A definition and explanation*. Retrieved from FlexMR Blog: https://blog.flexmr.net/correlation-analysis-definition-exploration

JMP Statistical Discovery. (2024, July 29). *Variable selection in multiple regression*. Retrieved from JMP Statistical Knowledge Portal: https://www.jmp.com/en/statistics-knowledge-portal/what-is-multiple-regression/variable-selection

Kumara, H., Krishna, M., & Vishwanath, H. L. (2020). Usefulness of gamma-glutamyl transpeptidase and mean corpuscular volume in alcoholic liver

disease. *International Journal of Clinical Biochemistry and Research, 5*(4), 638-641. doi:10.18231/2394-6377.2018.0135

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2004). *Applied linear statistical models* (5th ed.). New York: McGraw-Hill/Irwin. Retrieved from https://users.stat.ufl.edu/~winner/sta4211/ALSM_5Ed_Kutner.pdf

Liangpunsakul, S., Qi, R., Crabb, D. W., & Witzmann, F. (2010). Relationship between alcohol drinking and aspartateaminotransferase:alanine aminotransferase (AST:ALT) ratio, mean corpuscular volume (MCV), gamma-glutamyl transpeptidase (GGT), and apolipoprotein A1 and B in the U.S. population. *Journal of Studies on Alcohol and Drugs, 71*(2), 249-252. doi:10.15288/jsad.2010.71.249

Maner, B. S., Killeen, R. B., & Moosavi, L. (2024, July 27). *Mean Corpuscular Volume*. Retrieved from StatPearls: https://www.ncbi.nlm.nih.gov/books/NBK545275/

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis* (6th ed.). John Wiley & Sons, Inc. Retrieved from https://content.e-bookshelf.de/media/reading/L-16125104-1a3a7c5bd1.pdf

Moriles, K. E., Zubair, M., & Azer, S. A. (2024). Alanine Aminotransferase (ALT) Test. Orlando, Florida: StatPearls Publishing. Retrieved from https://www.statpearls.com/point-of-care/56363

Nyblom, H., Berggren, U., Balldin, J., & Olsson, R. (2004). High AST/ALT ratio may indicate advanced alcoholic liver disease rather than heavy drinking. *Alcohol and alcoholism*, 336-339. doi:10.1093/alcalc/agh074

Papoz, L., Warnet, J. M., Pèquignot, G., Eschwege, E., Claude, J. R., & Schwartz, D. (1981). Alcohol consumption in a healthy population. Relationship to gamma-glutamyl transferase activity and mean corpuscular volume. *JAMA, 245*(17), 1748-1751. doi:https://doi.org/10.1001/jama.245.17.1748

Puuka, K. (2007, March 2). Gamma-glutamyl transferase as a marker of alcohol abuse: effects of moderate drinking, obsity and increasing age on reference intervals. Finland: University of Tampere. Retrieved from https://trepo.tuni.fi/bitstream/handle/10024/67694/978-951-6859-9.pdf?sequence=1&isAllowed=y

Savage, D. G., Ogundipe, A., Allen, R. H., Stabler, S. P., & Lindenbaum, J. (2000). Etiology and diagnostic evaluation of macrocytosis. *American Journal of the Medical Sciences, 319*(6), 343-352. doi:10.1016/S0002-9629(15)40772-4

Schober, P., Boer, C., & Schwarte, L. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia, 126*(5), 1763-1768. doi:https://doi.org/10.1213/ANE.0000000000002864

Schreiber-Gregory, D. N. (2017). Multicollinearity: what is it, why should we care and how can it be controlled? Academia. Retrieved from https://www.academia.edu/download/88889537/1404-2017.pdf

Shrestha, N. (2020). Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics, 8*(2), 39-42. doi:https://doi.org/10.12691/ajams-8-2-1

Silvey, S. D. (1969). Multicollinearity and imprecise estimation. *Journal of the royal statistical society series, 31*(3), 539-552. doi:https://doi.org/10.1111/j.2517-6161.1969.tb00813.x

Singal, A. K., Bataller, R., Ahn, J., Kamath, P. S., & Shah, V. H. (2018). ACG clinical guideline: alcoholic liver disease. *The American Journal of Gastroenterology, 113*(2), 175-194. doi:10.1038/ajg.2017.469

Sinha, A. K., & Sinha, A. (2024). The evaluation of hematological and biochemical parameters in chronic liver disease patients. *Indian Journal of Applied Research, 14*(10), 48-49. doi:https://doi.org/10.36106/ijar/6501873

Sreekumar, D. (2023, August 28). *What is research methodology? Definition, types, and examples*. Retrieved from Paperpal Blog: https://paperpal.com/blog/academic-writing-guides/what-is-research-methodology

Staufer, K., & Yegles, M. (2016). Biomarkers for detection of alcohol consumption in liver transplantation. *World Journal of Gastroenterology, 22*(14), 3725-3734. doi:https://doi.org/http://dx.doi.org/10.3748/wjg.v22.i14.3725

Thapa, B. R., & Walia, A. (2007, July). Liver fuction tets and their interpretation. *The Indian Journal of Pediatrics., 74*(7), 663–671. doi:https://doi.org/10.1007/s12098-007-0118-

Appendix

Tables and Figures

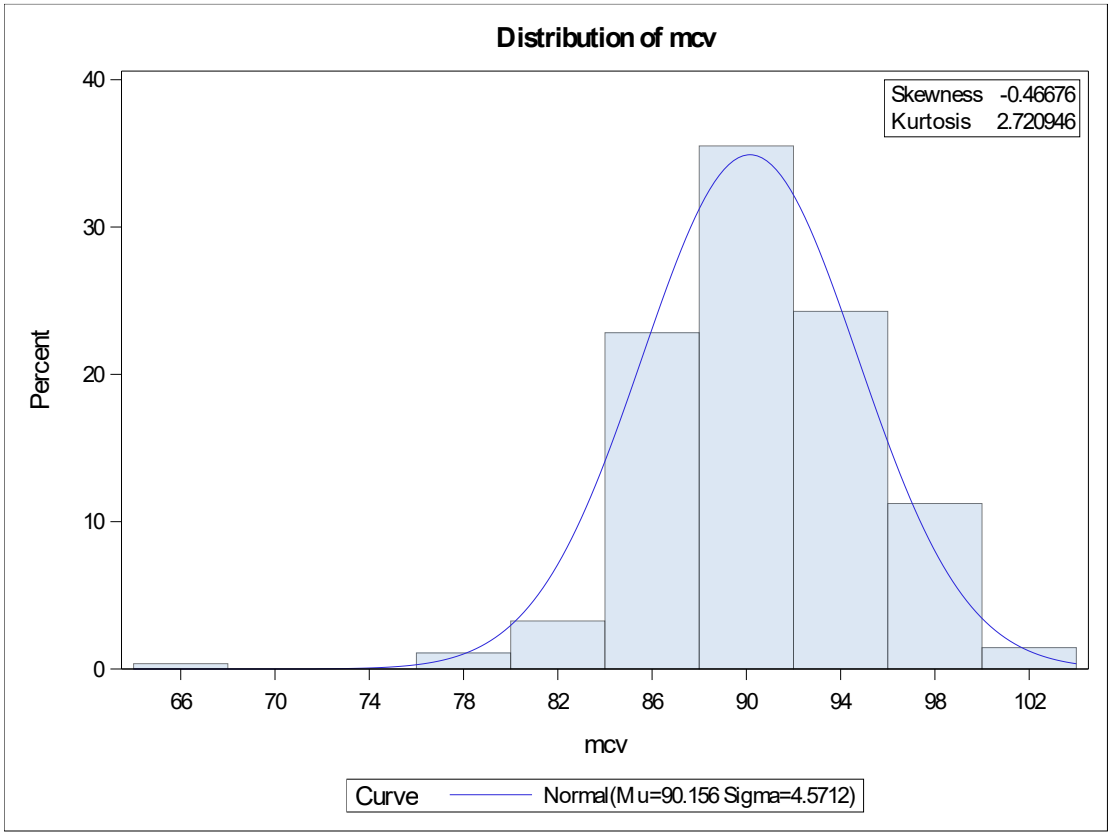

Figure 1: Distribution Curve of MCV

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

Table 1: Summary Statistics of MCV

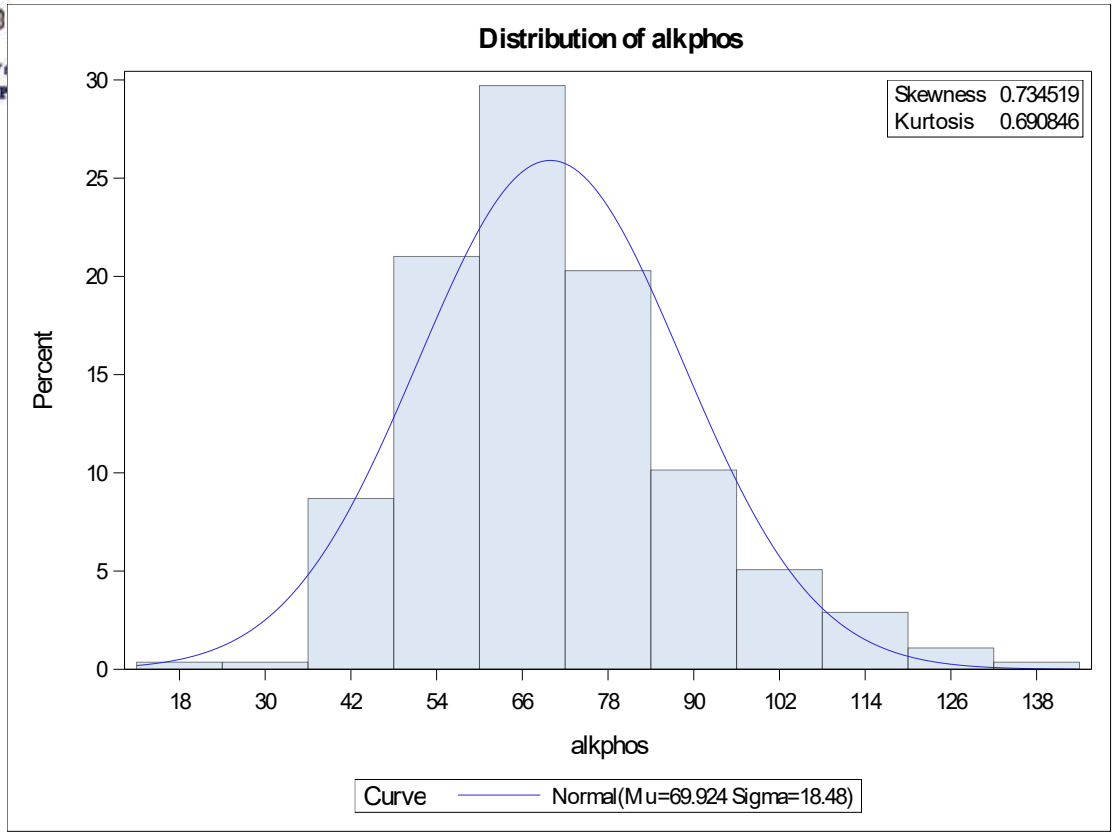

Figure 2: Distribution Curve of Alkphos

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

Table 2: Summary Statistics of Alkphos

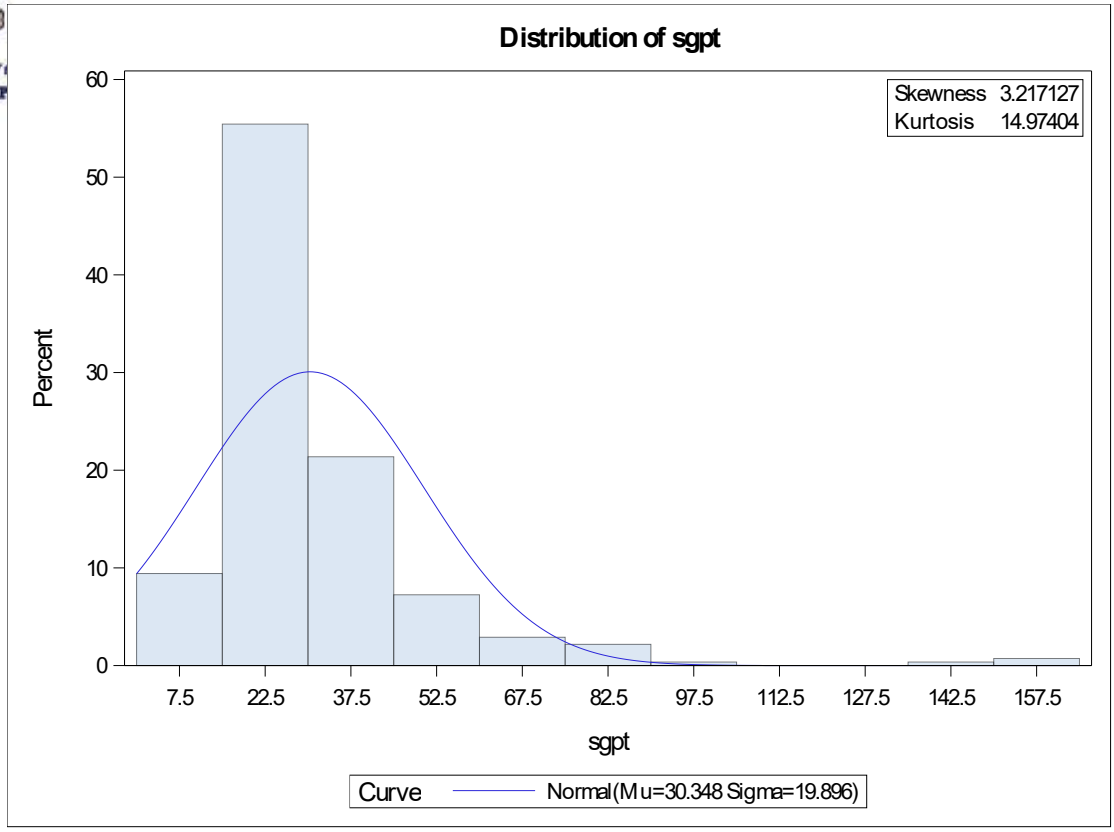

Figure 3: Distribution Curve of SGPT

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.1862009	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	3.3281869	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	18.4827293	**Pr > A-Sq**	<0.005

Table 3: Goodness of Fit Test for 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
Uncorrected SS	363056	**Corrected SS**	108862.609
Coeff Variation	65.5609797	**Std Error Mean**	1.19761846

Table 4: Summary Statistics of SGPT

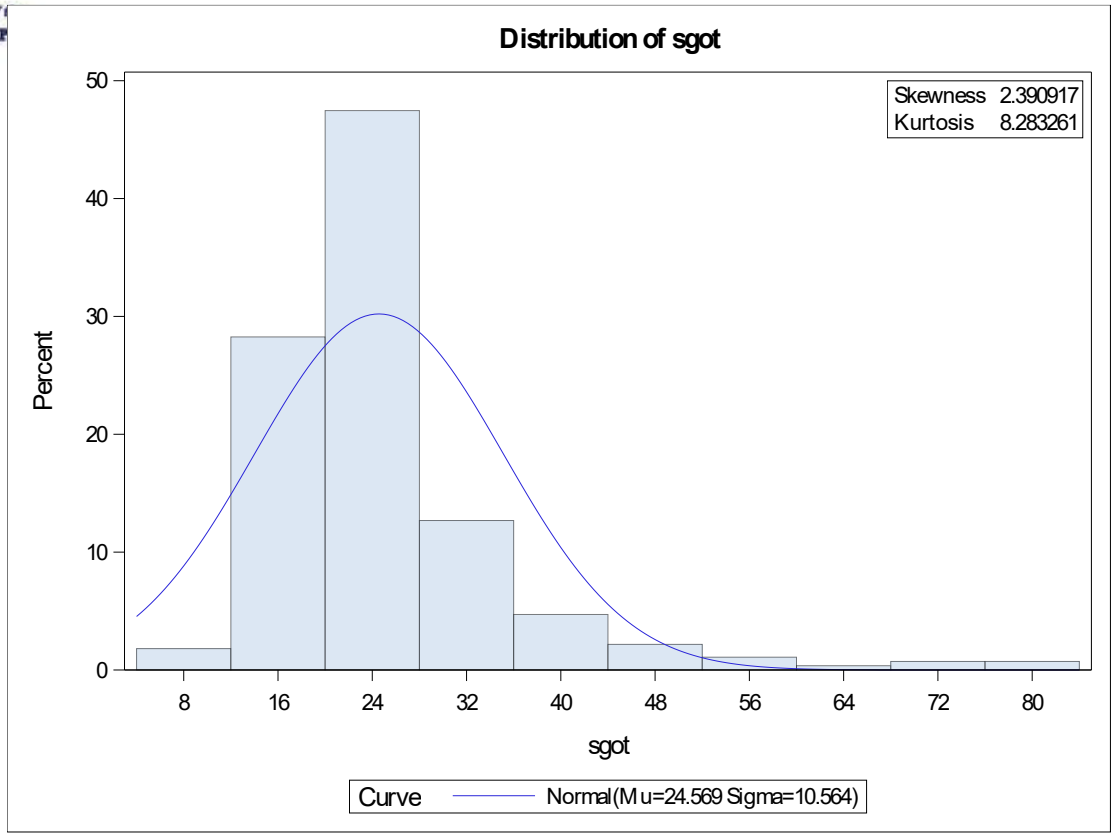

Figure 4: Distribution Curve of SGOT

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

Table 5: Summary Statistics of SGOT

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.1843585	Pr > D	<0.010
Cramer-von Mises	W-Sq	2.4405806	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	13.8250337	Pr > A-Sq	<0.005

Table 6: Goodness of Fit Test for SGOT

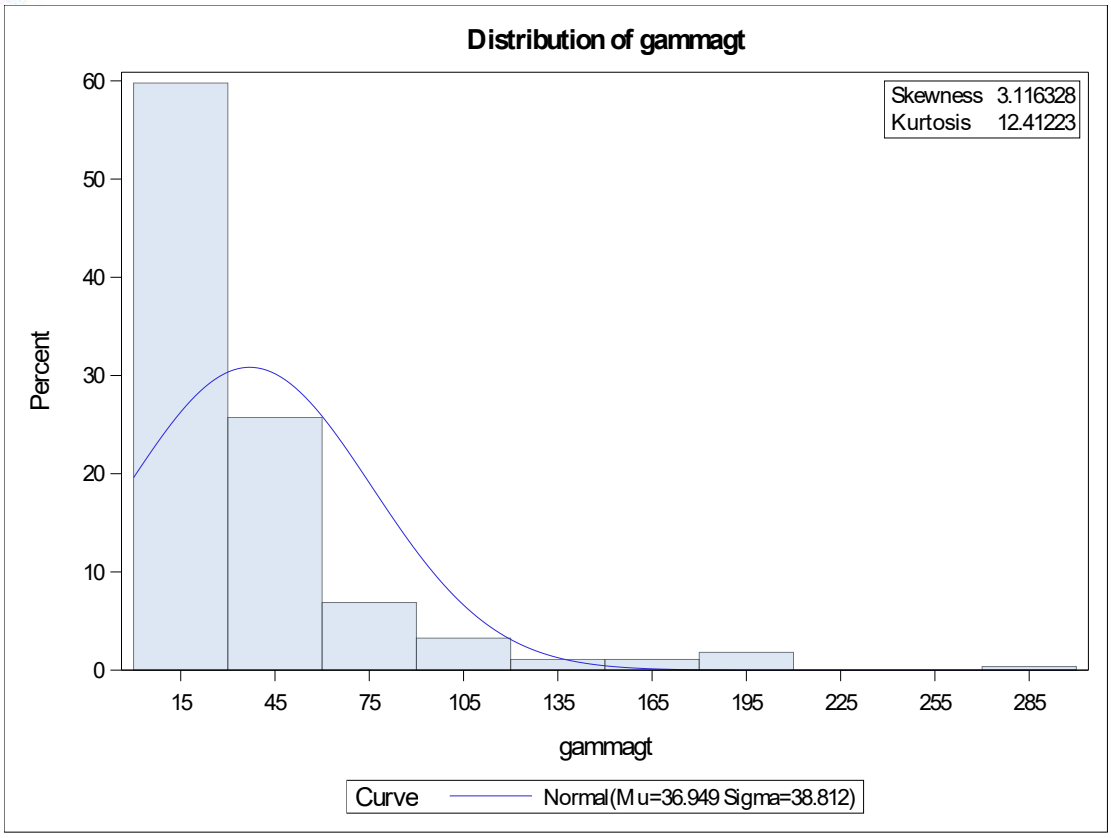

Figure 5: Distribution Curve of Gammagt

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

Table 7: Summary Statistics of Gammagt

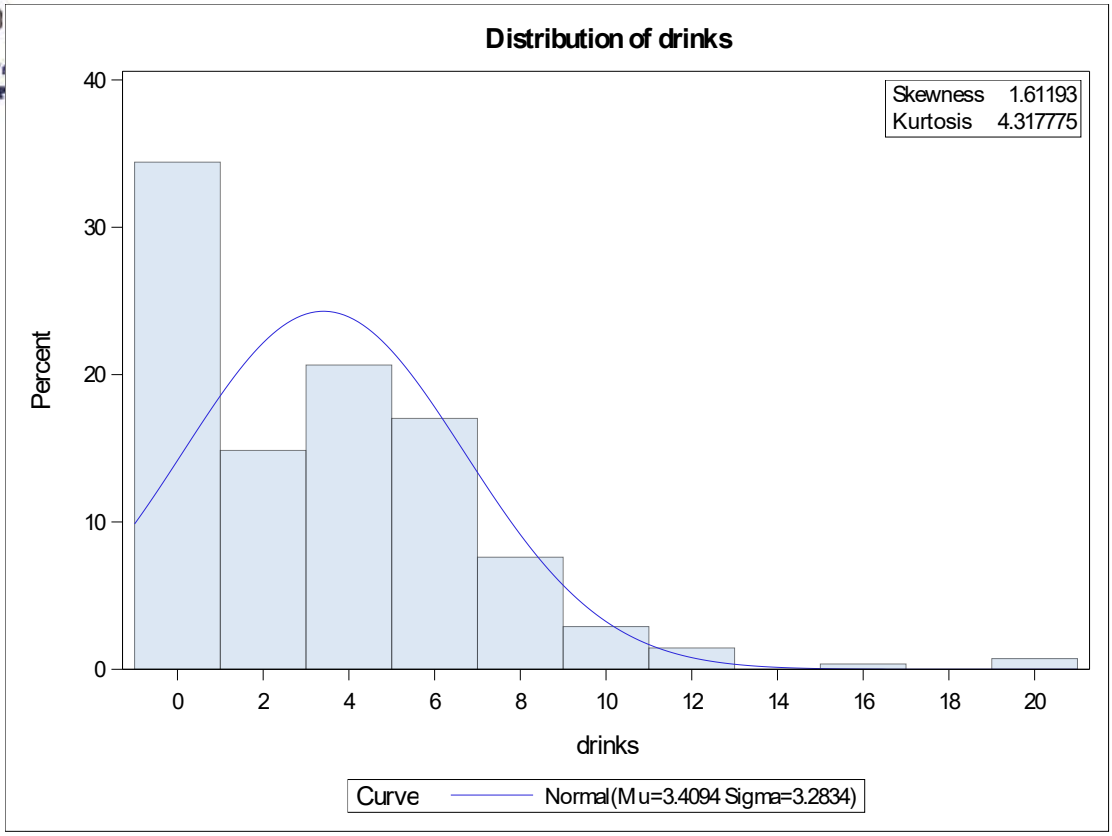

Figure 6: Distribution Curve of Drinks

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.1624211	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	1.8019544	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	11.6337040	**Pr > A-Sq**	<0.005

Table 8: Goodness of Fit Test for Drinks

Moments			
N	276	**Sum Weights**	276
Mean	3.40942029	**Sum Observations**	941
Std Deviation	3.28342145	**Variance**	10.7808564
Skewness	1.61193007	**Kurtosis**	4.31777461
Uncorrected SS	6173	**Corrected SS**	2964.73551
Coeff Variation	96.304391	**Std Error Mean**	0.19763875

Table 9: Summary Statistics of Drinks

Pearson Correlation Coefficients, N = 276 Prob > |r| under H0: Rho=0						
	MCV	**alkphos**	**sgpt**	**sgot**	**gammagt**	**drinks**
MCV	1.00000	0.02636 0.6629	0.15149 0.0117	0.19063 0.0015	0.23618 <.0001	0.35770 <.0001
alkphos	0.02636 0.6629	1.00000	0.10066 0.0951	0.17270 0.0040	0.15608 0.0094	0.11519 0.0560
sgpt	0.15149 0.0117	0.10066 0.0951	1.00000	0.74364 <.0001	0.53273 <.0001	0.18860 0.0016
sgot	0.19063 0.0015	0.17270 0.0040	0.74364 <.0001	1.00000	0.57328 <.0001	0.25817 <.0001
gammagt	0.23618 <.0001	0.15608 0.0094	0.53273 <.0001	0.57328 <.0001	1.00000	0.36516 <.0001
drinks	0.35770 <.0001	0.11519 0.0560	0.18860 0.0016	0.25817 <.0001	0.36516 <.0001	1.00000

Table 10: Pearson Correlation for All Variables

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

Table 11: Final Variables Selected in Forward Selection

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

Table 12: Summary of Forward Selection

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	-16.91936	3.61929	-4.67	<.0001	.	0
MCV	1	0.20457	0.03987	5.13	<.0001	0.93908	1.06487
alkphos	1	0.00999	0.00974	1.03	0.3062	0.96182	1.03970
sgpt	1	-0.01258	0.01356	-0.93	0.3543	0.42860	2.33317
sgot	1	0.02862	0.02656	1.08	0.2822	0.39608	2.52472
gammagt	1	0.02343	0.00575	4.07	<.0001	0.62562	1.59842

Table 13: Tolerance and VIF values of Variables

Collinearity Diagnostics								
Number	Eigenvalue	Condition Index	Proportion of Variation					
			Intercept	MCV	alkphos	sgpt	sgot	gammagt
1	5.27487	1.00000	0.00007966	0.00008091	0.00206	0.00397	0.00210	0.00797
2	0.45482	3.40555	0.00060583	0.00055610	0.01365	0.01869	0.00059471	0.43848
3	0.17838	5.43787	0.00034510	0.00036512	0.02011	0.40300	0.02585	0.46902
4	0.04757	10.52984	0.00000370	0.00002389	0.18182	0.54269	0.85087	0.02342
5	0.04315	11.05612	0.00978	0.01069	0.77095	0.03133	0.11959	0.03031
6	0.00121	66.13396	0.98919	0.98828	0.01141	0.00030915	0.00099652	0.03079

Table 14: Collinearity Diagnostics for Multicollinearity

Equations

Equation 1: Full Model

$$DRINKS = 0.340 + 0.034 \times 1 + 0.027 \times 2 + 0.025 \times 3 + 0.024 \times 4 + 0.023 \times 5$$

Equation 2: Reduced Model

$$Alcohol\ Consumption = -16.14 + 0.21(MCV) + 0.03(GGT)$$

Code

Getting Rid of Selector Variables

```
data mydata.group_1_train;
     set mydata.group_1_train;
     drop selector _dataobs_;
run;
```

Checking the Contents of the Dataset

```
proc contents data=mydata.group_1_train;
run;

proc print data=mydata.group_1_train;
run;
```

Descriptive Analysis

```
proc means data=mydata.group_1_train n mean std min
max maxdec=2;
   var MCV alkphos sgpt sgot gammagt drinks;
run;
```

Skewness and Kurtosis

```
proc univariate data=mydata.group_1_train;
   var MCV alkphos sgpt sgot gammagt drinks;
   histogram / normal;
   inset skewness kurtosis / position=ne;
run;
```

Correlation Analysis

```
proc corr data=mydata.group_1_train plots (only) =
scatter;
     var MCV alkphos sgpt sgot gammagt;
     with drinks;
run;

proc corr data=mydata.group_1_train;
   var MCV alkphos sgpt sgot gammagt drinks;
run;
```

Multi-collinearity

```
proc reg data=mydata.group_1_train;
   model drinks = MCV alkphos sgpt sgot gammagt / vif
tol collin;
run;
```

Simple Regression

```
proc reg data=mydata.group_1_train;
   model drinks = MCV;
run;
```

Hypothesis Test

```
proc reg data=mydata.group_1_train;
   model drinks = MCV alkphos sgpt sgot gammagt;
   test intercept=0, MCV=0, alkphos=0, sgpt=0,
sgot=0, gammagt=0;
run;
```

Assessing model fit and predictive ability

```
proc reg data =mydata.group_1_train;
     model drinks=MCV alkphos sgpt sgot gammagt;
run;
```

UNIVERSITY of the
WESTERN CAPE

Multiple Regression

```
proc reg data=mydata.group_1_train;
   model drinks = MCV alkphos sgpt sgot gammagt;
run;
```

Selection Techniques

```
proc reg data = mydata.group_1_train;
     forward: model drinks = MCV alkphos sgpt sgot
gammagt / selection = forward slentry= 0.05;
run;

proc reg data = mydata.group_1_train;
     backward: model drinks = MCV alkphos sgpt sgot
gammagt  / selection = backward slstay=0.05;
run;

proc reg data = mydata.group_1_train;
     stepwise: model drinks = MCV alkphos sgpt sgot
gammagt  / selection = stepwise;
run;
```