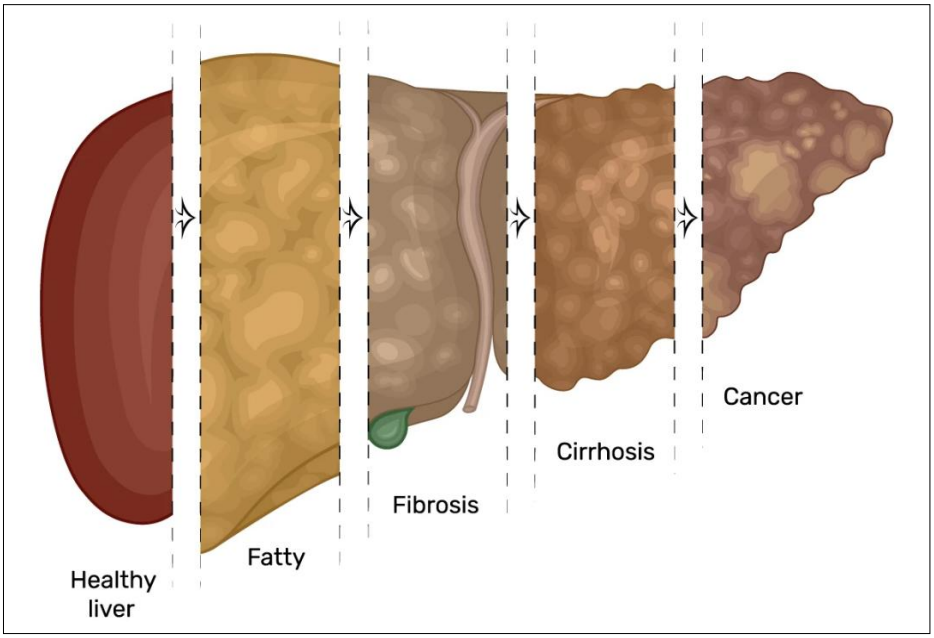
# Multivariate Statistical Analysis of Liver Disease Determinants

**S/19/819**
**STA4053 – Multivariate Methods II**

---

## 1. Introduction

Liver disease poses a significant public health challenge, with multiple lifestyle, genetic, and clinical factors influencing its onset and progression.




The primary objective of this study is to apply a range of multivariate statistical techniques to a liver disease dataset in order to uncover patterns, relationships, and underlying structures within the data. By systematically analyzing these factors, the study aims to enhance understanding of liver health determinants, support predictive modeling, and inform effective healthcare strategies.

# 2. Methodology

## 2.1. Dataset Description

The dataset comprises 1,700 observations and 11 variables, capturing demographic, lifestyle, and health-related information relevant to liver disease. The variables are categorized as follows:

- **Continuous variables:**
  Age, BMI, AlcoholConsumption, PhysicalActivity, LiverFunctionTest
- **Categorical variables:**
  Gender, Smoking, GeneticRisk, Diabetes, Hypertension, Diagnosis

## 2.2. Data Preprocessing

Before applying multivariate techniques, the dataset underwent the following preprocessing steps to ensure data quality and compatibility with the statistical methods:

1. **Handling Missing Values:**
   Missing values were checked and imputed as needed, though the final dataset had no missing values.
2. **Standardization of Continuous Variables:**
   Continuous variables were standardized using z-score normalization. This step ensures that variables with different units and scales contribute equally to the analysis.
3. **Encoding Categorical Variables:**
   Categorical variables were transformed into numerical format using one-hot encoding, creating binary indicator variables for each category, which is necessary for techniques that require numerical input.
4. **Preparation for Specific Analyses:**
   For Principal Component Analysis (PCA) and Factor Analysis (FA), only the standardized continuous variables were used, as these methods assume continuous data.

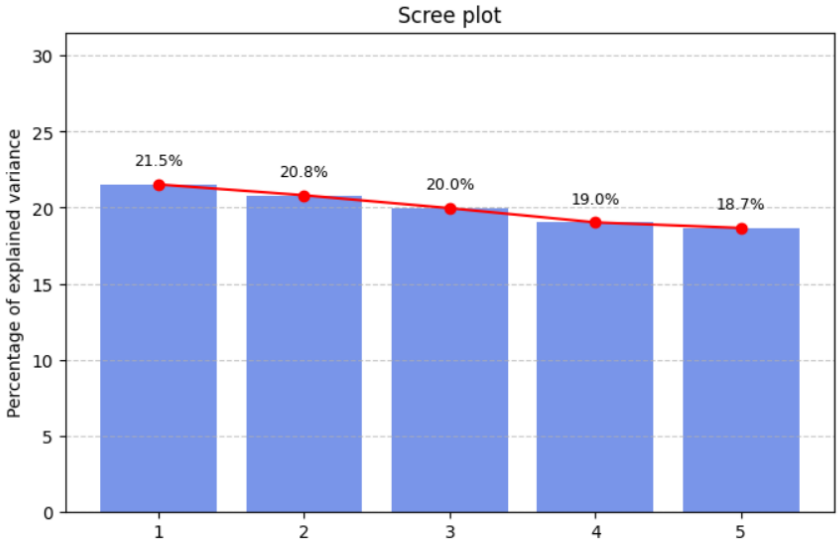
## 2.3. Statistical Methods Employed

- **Principal Component Analysis:**
  Used to reduce dimensionality and identify key components explaining variance in continuous variables.
- **Factor Analysis:**
  Used to detect latent factors underlying observed correlations among continuous variables.

- **Discriminant Analysis:**
  Used to classify subjects based on predictors and identify variables that best separate liver disease presence.
- **Canonical Correlation Analysis:**
  Used to explore relationships between two sets of continuous variables.

# 3. Results and Discussion

## 3.1. Principal Component Analysis (PCA)




The first five PCs explain the following percentages of variance respectively,

- PC1: 21.53%
- PC2: 20.82%
- PC3: 19.96%
- PC4: 19.03%
- PC5: 18.66%.

The cumulative variance explained by the first four PCs is approximately **81.3%**, which exceeds the commonly used threshold of 70%. Therefore, the first four PCs were retained.

- **PCA Loadings:**

The loadings represent the correlation between the original variables and the principal components.

| | PC1 | PC2 | PC3 | PC4 | PC5 |
|---|---|---|---|---|---|
| Age | 0.142532 | 0.642743 | 0.528061 | -0.270527 | 0.463177 |
| BMI | 0.589468 | 0.330854 | 0.161827 | 0.442837 | -0.566366 |
| AlcoholConsumption | -0.511236 | 0.411125 | -0.182574 | 0.702889 | 0.205495 |
| PhysicalActivity | -0.032965 | -0.533364 | 0.710716 | 0.418651 | 0.184528 |
| LiverFunctionTest | 0.608082 | -0.154650 | -0.395617 | 0.247768 | 0.623232 |

- **PC1** is strongly influenced by BMI and LiverFunctionTest (positive), and AlcoholConsumption (negative), suggesting this component represents a contrast between body composition/ liver function and alcohol use.
- **PC2** is dominated by Age (positive) and PhysicalActivity (negative), reflecting an age–activity axis, where older individuals tend to have lower physical activity.
- **PC3** is mainly defined by Physical Activity (positive) and Age (positive), suggesting it reflects variation in physical activity and age, somewhat independent of other variables.
- **PC4** shows strong positive loading for Alcohol Consumption and Physical Activity, and moderate positive loading for BMI, possibly reflecting a lifestyle component combining drinking and activity patterns.

PCA reduced the dimensionality of the data while preserving key patterns. These results suggest that liver health in this dataset is influenced by a combination of body composition, liver function, alcohol use, and physical activity, with different principal components capturing different aspects of these relationships.

## 3.2. Factor Analysis

According to the Kaiser criterion (eigenvalues > 1) and the visible "elbow" in the plot, **two factors** were selected for retention.

- **Eigenvalues of the first five factors:**
  1.0763, 1.0407, 0.9982, 0.9514, 0.9331

| Communalities: | |
|---|---|
| Age | 0.017850 |
| BMI | 0.139057 |
| AlcoholConsumption | 0.068285 |
| PhysicalActivity | 0.007840 |
| LiverFunctionTest | 0.031969 |

Loadings:

| | 0 | 1 |
|---|---|---|
| Age | -0.013540 | 0.132915 |
| BMI | 0.224190 | 0.297986 |
| AlcoholConsumption | -0.243357 | 0.095197 |
| PhysicalActivity | 0.037341 | -0.080282 |
| LiverFunctionTest | 0.178596 | 0.008477 |

- **Communalities:**
  These values indicate how much variance in each observed variable is explained by the two retained factors.

- **Factor Loadings:**
  The loadings show the correlation between each variable and the two factors.

**Factor 1** is most strongly associated with BMI (positive), Liver Function Test (positive), and Alcohol Consumption (negative). This suggests that Factor 1 may represent a dimension contrasting body composition and liver function against alcohol use.

**Factor 2** is mainly related to BMI (positive) and Age (positive), possibly reflecting a general health or demographic factor.

**Low communalities** across variables indicate that the two-factor model explains only a small portion of the variance in each variable. This suggests that the underlying factor structure among these variables is weak, and more factors or additional variables may be needed to better capture the relationships.

## 3.3. Discriminant Analysis (DA)

Linear Discriminant Analysis (LDA) was used to identify which variables best distinguish between individuals with and without liver disease. The model was trained using all standardized continuous variables and encoded categorical variables as predictors, with the binary diagnosis as the target.

```
LDA coefficients:
 Age                  0.580299
BMI                  0.583626
AlcoholConsumption   1.354197
PhysicalActivity    -0.385585
LiverFunctionTest    1.315218
Gender_1             1.346210
Smoking_1            1.721030
GeneticRisk_1       -0.056998
GeneticRisk_2        2.233406
Diabetes_1           0.891357
Hypertension_1       1.458579
```

**Positive coefficients** indicate that higher values of the variable are associated with a greater likelihood of being classified as having liver disease.

- The strongest positive predictors are AlcoholConsumption, LiverFunctionTest, GeneticRisk_2, Smoking_1, and Hypertension_1
- Gender_1 and BMI also contribute positively.

**Positive coefficients** (**PhysicalActivity**, **GeneticRisk_1**) suggest that higher values are associated with a lower likelihood of liver disease.

**PhysicalActivity** shows a negative relationship, indicating that greater activity may be protective.

The LDA results highlight that lifestyle factors (alcohol consumption, smoking, physical activity), clinical measures (liver function), and certain genetic and health indicators (hypertension, diabetes, genetic risk) are important in distinguishing between those with and without liver disease.

The direction and magnitude of the coefficients provide insight into the relative influence of each factor, supporting the importance of both behavioral and biological determinants in liver disease classification.

## 3.4. Canonical Correlation Analysis (CCA)

Canonical Correlation Analysis was performed to explore the relationship between two sets of continuous variables.

- Set 1: Age, BMI, AlcoholConsumption
- Set 2: PhysicalActivity, LiverFunctionTest

The canonical correlations measure the strength of association between the linear combinations of variables from each set.

```
Canonical correlation 1: 0.0642
Canonical correlation 2: 0.0272
```

```
First 5 rows of X_c:
 [[ 0.37821147  1.6881838 ]
 [-0.91512183  0.12348557]
 [ 1.73375526 -0.12788786]
 [ 1.04777931 -1.30931031]
 [ 0.05164612 -1.74295548]]

First 5 rows of Y_c:
 [[ 1.09971663  1.30289712]
 [-0.02368643  1.21331207]
 [-0.59229247 -1.64028615]
 [-0.25244599 -0.16580977]
 [-0.63357274  0.67650959]]
```

The first canonical correlation (0.0642) indicates a very weak positive relationship between the first pair of canonical variates from the two sets.

The second canonical correlation (0.0272) is also very weak, suggesting almost no linear association for the second pair.

These results suggest that, in this dataset, the linear combinations of Age, BMI, and AlcoholConsumption are largely independent from the linear combinations of PhysicalActivity and LiverFunctionTest.

The canonical correlation analysis reveals only a weak relationship between the selected demographic/ lifestyle variables (Age, BMI, AlcoholConsumption) and the health indicators (PhysicalActivity, LiverFunctionTest). This shows that these two sets of variables do not share strong linear relationships in this sample.

# 4. Conclusion and Recommendation

## 4.1. Conclusion

This analysis applied a range of multivariate statistical techniques to a comprehensive liver disease dataset to uncover patterns, relationships, and underlying structures among demographic, lifestyle, and health indicators.

Key findings:

- **Principal Component Analysis (PCA):**
  The first four principal components explained over 80% of the variance among continuous variables, highlighting the importance of BMI, liver function, alcohol consumption, age, and physical activity in liver health.
- **Factor Analysis (FA):**
  Two underlying factors were identified, with BMI and liver function contributing most. However, communalities were generally low, indicating that much of the variance in individual variables could not be explained by a small number of latent factors.
- **Discriminant Analysis (DA):**
  Variables such as alcohol consumption, liver function, smoking, hypertension, and genetic risk were strong predictors for distinguishing between individuals with and without liver disease.
- **Canonical Correlation Analysis (CCA):**
  Only weak linear relationships were found between demographic/lifestyle variables and health indicators, suggesting these sets of variables are largely independent in this dataset.

The analyses revealed that variables such as BMI, liver function, alcohol consumption, age, and physical activity play significant roles in explaining variation among individuals. While techniques like PCA and discriminant analysis successfully identified key predictors and summarized complex patterns, other methods, such as factor analysis and canonical correlation

analysis, indicated that some relationships among variables are relatively weak or not easily captured by linear models.

Overall, the findings highlight both the value and the limitations of multivariate approaches for uncovering the multifaceted nature of liver disease, supporting their use in predictive modeling and informing strategies for improved healthcare decision-making.

### 4.2. Limitations

- Some important factors may not be captured in the dataset, and relationships between variables were often weak or complex.
- The results are specific to this dataset and may not apply to other populations.
- Multivariate methods require careful data preparation and can be difficult to interpret.

### 4.3. Recommendations

- Include more clinical and lifestyle variables in future studies to improve analysis.
- Validate findings with other datasets to check generalizability.
- Consider advanced modeling approaches to better capture complex relationships.

## 5. References


- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage Learning.
- Afifi, A., May, S., Donatello, R., & Clark, V. A. (2019). *Practical multivariate analysis* (6th ed.). Chapman and Hall/CRC.


## 6. Appendices

- Dataset: https://www.kaggle.com/datasets/rabieelkharoua/predict-liver-disease-1700-records-dataset

- Python codes:

*Import Libraries*

```
import pandas as pd
import numpy as np
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler, OneHotEncoder
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
from factor_analyzer import FactorAnalyzer
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.cross_decomposition import CCA
from sklearn.cluster import KMeans
```

*Data Preprocessing*

```
# Load the dataset
df = pd.read_csv('Liver_disease_data.csv')

# Separate continuous and categorical variables
continuous_vars = ['Age', 'BMI', 'AlcoholConsumption', 'PhysicalActivity', 'LiverFunctionTest']
categorical_vars = ['Gender', 'Smoking', 'GeneticRisk', 'Diabetes', 'Hypertension', 'Diagnosis']

#check missing values
missing_values_count = df.isnull().sum()
print('Missing Values : \n' ,missing_values_count)

cat_imputer = SimpleImputer(strategy='most_frequent')
df[categorical_vars] = cat_imputer.fit_transform(df[categorical_vars])

# Standardize continuous variables
scaler = StandardScaler()
df[continuous_vars] = scaler.fit_transform(df[continuous_vars])

# One-hot encode categorical variables (if needed for other techniques)
df_encoded = pd.get_dummies(df, columns=categorical_vars, drop_first=True)

# For PCA, create a separate DataFrame with only standardized continuous variables
df_pca = df[continuous_vars]
```

## *Principal Component Analysis (PCA)*

```
pca = PCA()
pca.fit(df_pca)

explained_variance = pca.explained_variance_ratio_ * 100
cumulative_variance = explained_variance.cumsum()

plt.figure(figsize=(8, 5))
bars = plt.bar(range(1, len(explained_variance) + 1), explained_variance, alpha=0.7, color='royalblue')
plt.plot(range(1, len(explained_variance)+1), explained_variance, marker='o', color='red', label='Eigenvalues Line')

for i, v in enumerate(explained_variance):
    plt.text(i + 1, v + 1, f"{v:.1f}%", ha='center', va='bottom', fontsize=9)

plt.xlabel('Dimensions')
plt.ylabel('Percentage of explained variance')
plt.title('Scree plot')
plt.xticks(range(1, len(explained_variance) + 1))
plt.ylim(0, max(explained_variance) + 10)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()

loadings = pd.DataFrame(pca.components_.T, columns=[f'PC{i+1}' for i in range(len(df_pca.columns))], index=df_pca.columns)
pca_scores = pd.DataFrame(pca.transform(df_pca), columns=[f'PC{i+1}' for i in range(len(df_pca.columns))])
explained_variance, cumulative_variance, loadings.head(), pca_scores.head()
loadings = pd.DataFrame(pca.components_.T, columns=[f'PC{i+1}' for i in range(len(df_pca.columns))], index=df_pca.columns)

print('Total variances', explained_variance)
print('\n\n')
print(loadings)
```

## *Factor Analysis (FA)*

```
fa = FactorAnalyzer()
fa.fit(df_pca)
ev, v = fa.get_eigenvalues()

plt.figure(figsize=(8, 5))
plt.bar(range(1, len(ev)+1), ev, alpha=0.7, color='royalblue', label='Eigenvalues')
plt.plot(range(1, len(ev)+1), ev, marker='o', color='red', label='Eigenvalues Line')
plt.title('Scree Plot for Factor Analysis')
plt.xlabel('Factors')
plt.ylabel('Eigenvalue')
plt.grid(True)
plt.legend()
plt.show()
```

```
# Select number of factors based on scree plot
n_factors = 2
fa = FactorAnalyzer(n_factors=n_factors, rotation='varimax')
fa.fit(df_pca)

loadings = pd.DataFrame(fa.loadings_, index=df_pca.columns)
communalities = pd.Series(fa.get_communalities(), index=df_pca.columns)
factor_scores = pd.DataFrame(fa.transform(df_pca), columns=[f'Factor{i+1}' for i in range(n_factors)])
variances = fa.get_factor_variance()

print("Eigenvalues:\n", ev)
print("\nFactor Variances (Variance, Proportion, Cumulative):\n", variances)
print("\nCommunalities:\n", communalities)
print("\nLoadings:\n", loadings)
print("\nFactor Scores (first 5 rows):\n", factor_scores.head())
```

### *Discriminant Analysis (DA)*

```
# X: predictors, y: target
X = df_encoded.drop(columns=['Diagnosis_1'])
y = df_encoded['Diagnosis_1']

lda = LinearDiscriminantAnalysis()
lda.fit(X, y)

coefficients = pd.Series(lda.coef_[0], index=X.columns)
predictions = lda.predict(X)

print("LDA coefficients:\n", coefficients)
print("\nPredicted class labels (first 10):\n", predictions[:10])
```

*Canonical Correlation Analysis (CCA)*

```
set1 = ['Age', 'BMI', 'AlcoholConsumption']
set2 = ['PhysicalActivity', 'LiverFunctionTest']

cca = CCA(n_components=2)
X1 = df_pca[set1]
Y1 = df_pca[set2]
cca.fit(X1, Y1)
X_c, Y_c = cca.transform(X1, Y1)

corrs = [np.corrcoef(X_c[:, i], Y_c[:, i])[0, 1] for i in range(2)]

for i, corr in enumerate(corrs):
    print(f"Canonical correlation {i+1}: {corr:.4f}")

print("\nFirst 5 rows of X_c:\n", X_c[:5])
print("\nFirst 5 rows of Y_c:\n", Y_c[:5])
```