NATIONAL TECHNICAL UNIVERSITY OF ATHENS

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

MASTER'S DEGREE IN DATA SCIENCE AND MACHINE LEARNING

DATA ANALYSIS USING MIXED INTERACTION AND CHAIN GRAPH MODELS

THESIS

GOURGOURA KONSTANTINA

Student ID number: 03400128

Supervisor Professor: C. Caroni-Richardson

Committee

C. Caroni-Richardson	V. Papanikolaou	K. Chrysafinos
Professor NTUA	Professor NTUA	Professor NTUA

Athens, 2023

Abstract

Statistical analysis is a science that empowers us to process, present and interpret data which come from various sources, including engineering, biology, economics, and psychology. This thesis includes two parts: a theoretical part and a practical part. The first part consists of two chapters, providing a theoretical background for different graphical model techniques in statistical analysis. The second part consists of a third chapter, applying these techniques to real-world data and comparing their outcomes.

A graph serves as a valuable tool for the visual representation of relationships between data. Through a graph, it is easier to communicate complex information not only to experts of the field, but also people who have no knowledge of the theory behind it. Furthermore, this approach allows us to discern patterns and relationships that otherwise would not be observed in the raw data.

In the first chapter, we present three types of undirected graphical models. The first technique centers around log-linear models, which find application when datasets consist of discrete variables. The second technique is about Gaussian graphical models, utilized specifically with datasets comprising continuous variables. Lastly, the third technique delves into mixed interaction models, which are designed for datasets consisting of both continuous and categorical variables. These graphical models aim to gain insights into the hidden associations and present then into the form of a diagram, where each edge represents an association between the variables it connects.

In the second chapter, we present one of the most important directed graphical models. This technique centers around chain graph models and it can be utilized when dealing with datasets consisting of both continuous and categorical variables. Chain graph models use blocks to separate variables based on their role within the model. They share similarities with the mixed interaction models, but their difference lies in the regression process. In chain graph models each variable is regressed based on the variables present in all prior blocks, while in the mixed interaction models the regression considers the variables from all prior blocks and the current block.

In the third chapter, we fit Gaussian graphical models, mixed interaction models and chain graph models for statistical analysis of three sets of data using the statistical program R. Specifically, we analyzed post-COVID data from a study following patients who were hospitalized for COVID-19. The survey included various data, such as anthropometric, hospitalization and psychological data which were obtained though questionnaires. After each application, an explanation of the results is provided, shedding light on the meaning of the outcomes. Furthermore, at the end of all three applications the findings are compared and discussed.

Key – words: Gaussian graphical model, mixed interaction model, chain graph mode, AIC, BIC, graph, maximized log likelihood, blocks, directed edges, undirected edges.

Acknowledgments

I would like to express my gratitude to my supervisor C. Caroni for her guidance, expertise, and support during my research. Her insightful comments and encouragement have been valuable in shaping the direction and scope of my thesis. Also, I would like to thank my committee members V. Papanikolaou, and K. Chrysafinos for their contribution to this journey.

I extend my appreciation to Professor E. Stanghellini for her valuable mentorship, in-depth knowledge, and constant encouragement. Her perspectives, as well as her guidance have a great importance and I am truly fortunate to have had her by my side. Moreover, I would like to thank Professor F. Bartolucci, Dr. G. Pucci, and Dr. P. Rivadeneyra for their significant assistance to this study. Without their collaborative efforts none of this would have been possible.

It is important to acknowledge that this thesis was based on the research grant entitled "Multidimensional statistical analysis of databases relating to patients affected by "long-Covid" in order to characterize their manifestations, identify their risk factors and identify their therapeutic trajectories.". This grant was provided and funded by the University of Perugia, in Italy.

I would also like to express my gratitude to the staff and faculty at the National Technical University of Athens, who have managed to create a supportive academic environment for the students. Additionally, the help of the administrative staff has been crucial for the coordination of all the necessary procedural processes.

Last by not least, I am deeply grateful to my family and friends, who have been very supportive throughout my studies. They consistently provided me with love, care, and a listening ear in times of need. Their belief in my abilities has always been inspiring and motivating me to improve myself and achieve my goals.

Contents

Tables

Figures

Chapter 1- Undirected Graphical Models

In this chapter, we will introduce the fundamentals of undirected graphs. Specifically, we will discuss how they relate to log-linear models, Gaussian graphical models, and mixed interaction models, which combine elements of both. By understanding the basics of undirected graphs and their applications in different types of models, we can better analyze and interpret complex data.

1.1 Graph

A graph is a collection of points, which are interconnected by lines. The points and lines are called nodes and edges respectively. Each edge represents a relationship between the nodes it connects.

In the context of a model, the nodes can be used to represent variables, and the edges can represent the relationships between these variables. By this representation, we can gain insight into how different factors influence one another and how changes in one variable can affect others.

Consider two nodes denoted by A and B. In undirected graphs, the edge between the two nodes is denoted either by $[AB]$ or $[BA]$, as the edge does not have a specific direction (Edwards, 2000).

One common way to represent a graph visually is by using a diagram.

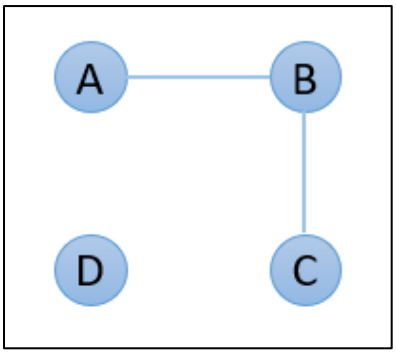


Figure 1: Undirected graph

$Figure\ 1$ depicts an undirected graph, where the collection of nodes is composed of $A, B, C,$ and D and the collection of edges consists of $[AB]$ and $[BC]$.

1.1.1 Definitions

In accordance with the definitions provided in Edwards's (2000) book, we will now present some key terms and concepts.

These definitions will help us establish a common vocabulary and understanding of the concepts related to graph theory and its applications. Having a clear understanding of key terms and concepts is crucial in any field, as it allows us to communicate ideas effectively and efficiently. By defining these terms at the outset, we can ensure that everyone is on the same page and has a shared understanding of the material.

When there is an edge connecting two nodes, we refer to them as adjacent. In mathematical notation, we write $A \sim B$, where A and B are both nodes and $\sim$ denotes adjacency. In $Figure\ 1$, A and B are adjacent, but B and D are not, as there is no edge connecting them.

A graph is considered complete if every pair of nodes is connected by an edge. In other words, there are no vertices in the graph that are not adjacent to each other. The graph shown in $Figure\ 1$, is not complete. The complete version of it, is shown in $Figure\ 2$.

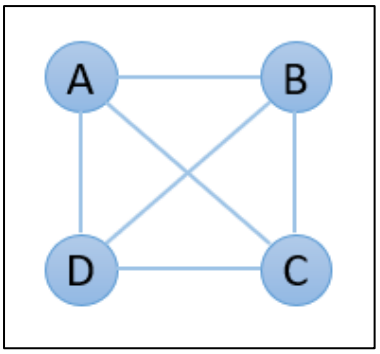


Figure 2: Undirected Complete graph

A clique is a subset of nodes in an undirected graph, where every two distinct nodes are adjacent to each other. In simpler terms, a clique forms a complete subgraph, with all nodes in the clique being pairwise connected by edges. A clique is considered maximal when it satisfies the conditions of being a clique, and adding any additional nodes to it would violate this property (Hastie et al., 2009). In $Figure\ 2$, the nodes A, B, C and D form a maximal clique, as every pair of distinct nodes are connected by edges $[AB], [AC], [AD], [BC], [BD]$ and $[CD]$, and no additional edge can be added while preserving the clique property.

A path is a sequence of k nodes, where each consecutive pair of nodes is adjacent. This path has length $k-1$. $Figure\ 1$ provides an example where A, B, C is a path of length 2 between the nodes A and C.

A k-cycle is a path consisting of k nodes, forming a closed loop. A 3-cycle example can be found in $Figure\ 2$, where the path initiates at node B, then proceeds to C and D, and eventually terminates at node B.

A chordless cycle is a k-cycle where the k nodes are distinct, and only the pairs of vertices that are adjacent are those with a difference in their indices of 1 or $k-1$. In other words, there are no edges connecting non-adjacent vertices within the cycle. $Figure\ 2$ displays a chordless 4-cycle, where the path initiates at node A, then proceeds to B, C, D, and terminates at node A.

A graph is considered triangulated if it does not contain any chordless cycles of length greater than or equal to four. This means that every cycle in the graph must have at least one chord, which is a non-cycle edge connecting two vertices within the cycle. The graph, shown in $Figure\ 2$, is triangulated since it does not contain any cycles of length four or more that are chordless.

The number of edges in which a node is involved is called the node's degree. The highest degree among all nodes in a graph is referred to as the graph's degree (Koller & Friedman, 2009).

1.2 Log-Linear Models

Log-linear models are statistical models that enable the analysis of the relationships between categorical variables. They calculate the likelihood of observing a specific combination of variable categories in the model by assuming that the logarithm of the likelihood is a linear function of the variables. This property makes it possible to model intricate associations and interactions between multiple categorical variables. Due to this feature, log-linear models are beneficial in many areas such as epidemiology, social sciences, and market research.

1.2.1 Log-linear, Graphical & Decomposable Models

Log-linear models can be classified as hierarchical or nonhierarchical. The type of log-linear models that we are focusing on in this thesis are hierarchical models, which are also the most common models. In a log-linear model, the term "hierarchical" refers to the inclusion of all lower-order terms that can be obtained from the variables present in a higher-order term (Gauraha, 2017). The higher-order terms are called generators. A hierarchical log-linear model is said to be graphical, if its generators correspond to the cliques of the undirected graph, where the nodes represent the variables, and the edges represent the two-factor terms included in the model. A graphical log-linear model is considered decomposable if its graph is triangulated. In accordance with the definition provided in subsection $(1.1.1)$, a graph is triangulated when it does not contain cycles of length four or more without a chord (Maathuis et al., 2018).

1.2.2 Hierarchical Model

Let $X = (X_1, X_2, \dots, X_k)$ be k discrete random variables of a dataset D and let d be a subset of this dataset, $d \subseteq D$. Furthermore, let I be the set of all possible combinations of values for these variables. A cell of I is denoted as $i = (i_1, i_2, \dots, i_k)$ and it is an observation of the values of the discrete variables, where i_j is the value of the j-th variable in the cell.

A hierarchical log-linear model is given by the formula,

$$\log p_i = \sum_{d \subseteq D} u_i^d \tag{1.1}$$

where,

- $\log p_i$ is the logarithm of the probability of observing cell i,
- $\Sigma_{d \subseteq D}$ is the summation of all possible interaction terms in the model,
- u_i^d is an interaction term, which corresponds to cell i and depends only on the variables involved in the d-th subset,

(Edwards, 2000).

To provide better clarity, we will give an example.

Suppose we have three categorical variables: gender (G) with two levels (male, female), age (A) with three levels (young, middle-aged, old), and smoking (S) with two levels (smoker, nonsmoker). We are interested in modeling the joint distribution of these variables using a Log-Linear model.

The saturated model assumes that the logarithm of the probability of each possible combination of levels of the three variables is a linear function of the variables. Using equation (1.1), we derive the following,

$$\log p\big(i = (g, a, s)\big) = u + u_{i_g}^G + u_{i_a}^A + u_{i_s}^S + u_{i_g, i_a}^{GA} + u_{i_g, i_s}^{GS} + u_{i_a, i_s}^{AS} + u_{i_g, i_a, i_s}^{GAS}$$

where,

- $p(i)$ is the probability of the outcome i,
- u is the intercept term,
- $u_{i_g}^G, u_{i_a}^A, u_{i_s}^S$ represent the main effects,
  - $u_{i_g}^G$ is the effect of gender on the probability of the outcome, and i_g=male, female,
  - $u_{i_a}^A$ is the effect of age on the probability of the outcome, and i_a=young, middle-aged, old,
  - $u_{i_s}^S$ is the effect of smoking on the probability of the outcome, and i_s=smoker, nonsmoker,
- $u_{i_g, i_a}^{GA}, u_{i_g, i_s}^{GS}, u_{i_a, i_s}^{AS}$ represent the interaction effect,
  - u_{i_g, i_a}^{GA} is the interaction effect between gender and age,
  - u_{i_g, i_s}^{GS} is the interaction effect between gender and smoking,
  - u_{i_a, i_s}^{AS} is the interaction effect between age and smoking,
- u_{i_g, i_a, i_s}^{GAS} is the three-way interaction effect between gender, age, and smoking.

The terms in this model represent how the probability of the outcome is related to each variable and their interactions, using a log-linear relationship. The model allows for the effects of each variable to be conditional on the other variables, and for their interactions to be included in the model as well.

By considering this example, we can identify three distinct types of dependencies that can be modeled using a log-linear model.

- The marginal independence between two discrete variables (e.g., gender and age), where the probability of one variable taking any value does not depend on the value of the other variable, and vice versa.
- The conditional independence between two variables, given a third variable. For example, gender and age are conditionally independent given smoking if

$$P(G, A|S) = P(G|S)P(A|S)$$

  In other words, once we know the value of smoking, knowing the value of gender does not give us any additional information about the value of age, and vice versa.
- The higher-order interaction involves all three variables.

1.2.3 Log-Likelihood

The likelihood function of a hierarchical log-linear model expresses the probability of observing a set of data given the parameters of the model. The likelihood function is derived from the joint distribution of the observed data and the model parameters, and it is given by the formula,

$$L = c \prod_{i \in I} p(i)^{n_i} \tag{1.2}$$

where,

- i is the i-th cell in the contingency table,
- n_i is the observed count of i,
- $p(i)$ is the predicted probability of i, and
- c is a constant term.

The maximized log-likelihood of the model can be obtained by maximizing the likelihood function (1.2).

$$\hat{l} = c + \sum_{i \in I} n(i) \log \hat{p}(i) \tag{1.3}$$

where $\hat{p}(i)$ are the maximum likelihood estimates (MLEs) of the model (Højsgaard et al., 2012).

1.2.4 Goodness of fit

The deviance is a measure of how well the model fits the data. It measures the difference in the log-likelihood between the fitted model m and the saturated model s, and it is scaled by a factor of 2. This is a method to assess the goodness of fit of a model, as well as to compare different models. A deviance value of 0 indicates a perfect fit, while greater values indicate a poorer fit.

The deviance for a Log-Linear model is given by the formula,

$$D = 2\left(\hat{l}_s - \hat{l}_m\right) \tag{1.4}$$

where l_s, l_m are the maximized log-likelihood function of the saturated model, and the fitted model respectively. Utilizing the equation (1.3),

$$\hat{l}_s = c + \sum_{i \in I} n(i) \log \hat{p}_s(i) = c + \sum_{i \in I} n(i) \log\left(\frac{n(i)}{N}\right) \tag{1.5}$$

where N is the number of observations, and

$$\hat{l}_m = c + \sum_{i \in I} n(i) \log \hat{p}_m(i) \tag{1.6}$$

By combining $(1.4), (1.5), (1.6)$,

$$D = 2\sum_{i} n(i)\log\left(\frac{n(i)}{\hat{m}(i)}\right) \quad \sim \quad X_k^2 \tag{1.7}$$

where,

- n_i is the observed count of the i-th cell, and
- $\hat{m}(i)$ is the expected cell count, $\hat{m}(i) = N\hat{p}_m(i)$.

The deviance can be used to perform hypothesis tests on the model, using the chi-squared distribution with degrees of freedom k, equal to the difference in dimension between the saturated model s (the log-linear model with every interaction term) and the alternative model m (the log-linear model of interest) (Højsgaard et al., 2012).

1.2.5 Model Selection

Model selection is important in statistical modeling and data analysis. The purpose of model selection is to identify the most appropriate model that accurately captures the underlying patterns and relationships in the data while avoiding overfitting or underfitting.

One way to select the best model among models fitted based on the maximized log-likelihood function is to apply the Akaike's Information Criterion (AIC), which is defined as

$$AIC = -2\ln L + 2d$$

where,

- lnL is the maximized log-likelihood function and
- d is the number of parameters of the model under consideration.

This criterion is used by calculating and comparing the AIC values for models with different numbers of variables and selecting the model with the lowest AIC each time, until there is no other model with a lower AIC value (Καρώνη & Οικονόμου, 2017).

In a similar manner to the AIC criterion, we apply the BIC criterion (Bayesian Information Criterion), which is defined as

$$BIC = -2\ln L + d\ln N$$

where,

- lnL is the maximized log-likelihood function,
- d is the number of parameters of the model under consideration and
- N is the number of observations.

As the number of observations approaches infinity, the probability of BIC accurately selecting the best model increases significantly and tends to one, in contrast to AIC criterion which has the tendency to favor more complex models in such circumstances (Hastie et al., 2009).

1.3 Gaussian Graphical Models

Gaussian graphical models are statistical models used to represent the conditional independence relationships among variables in a multivariate Gaussian distribution. These models are used to analyze continuous data, where the variables follow a multivariate Gaussian distribution. The graph shows the dependencies between variables, with nodes representing variables and edges representing conditional dependence relationships. The edges in a Gaussian graphical model indicate partial correlations, which are the conditional correlations between variables after accounting for the other variables in the model. Gaussian graphical models are commonly used in various fields such as statistics, biology, and finance.

1.3.1 Gaussian Graphical Model

The formula of a Gaussian graphical model can be represented using a precision matrix, also known as an inverse covariance matrix, or a concentration matrix.

Let $X = (X_1, X_2, \ldots, X_k)$ be a multivariate Gaussian random vector with k variables and let Σ denote the covariance matrix of X. The concentration matrix Ω, which is the inverse of Σ, represents the conditional independence relationships among the variables in the Gaussian graphical model.

The density function of X can be expressed as

$$f(x) = (2\pi)^{-\frac{k}{2}} |\Omega|^{\frac{1}{2}} exp\left(-\frac{1}{2}(x-\mu)^T \Omega (x-\mu)\right), x \in \mathbb{R}^k \qquad (1.8)$$

where,

- $f(x)$ represents the joint probability density function of the k-dimensional Gaussian distribution with concentration matrix Ω,
- $(2\pi)^{-\frac{k}{2}}$ is a normalization constant,
- $|\Omega|$ represents the determinant of the concentration matrix Ω ($det\Omega = det\Sigma^{-1}$),
- μ is the k-dimensional vector of means of the Gaussian distribution $N(\mu, \Sigma)$,
- Σ is the $k \times k$ covariance matrix of the Gaussian distribution $N(\mu, \Sigma)$,
- $(x-\mu)^T$ is the transpose of $(x-\mu)$, and
- $(x-\mu)^T \Omega (x-\mu)$ is the quadratic form of $(x-\mu)$ with respect to the concentration matrix Ω,

(Maathuis et al., 2018)

In a Gaussian graphical model, the partial correlations between variables can be obtained from the elements of the concentration matrix Ω,

$$p_{ij \mid V\setminus\{i,j\}} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}, \quad i,j = 1,2,\dots,k \tag{1.9}$$

where,

- $p_{ij \mid V\setminus\{i,j\}}$ represents the partial correlation between X_i and X_j, given all the other variables in set of nodes V except i and j,
- ω_{ij} refers to the element located in the i-th row and j-th column of the matrix Ω.

A zero entry in Ω indicates conditional independence between the corresponding variables, while a non-zero entry indicates a conditional dependence.

The graphical structure of such a model, can be inferred from the sparsity pattern of the precision matrix Ω, where non-zero entries in Ω correspond to edges between variables in the graphical model.

To provide better clarity, we will give an example.

Suppose we have three continuous variables X, Y and Z, and let the precision matrix to be,

$$\Omega = \begin{bmatrix} \omega_{XX} & \omega_{XY} & \omega_{XZ} \\ \omega_{YX} & \omega_{YY} & 0 \\ \omega_{ZX} & 0 & \omega_{ZZ} \end{bmatrix}. \tag{1.10}$$

Based on this concentration matrix (1.10), and the conditional independence relationships described earlier, it can be concluded that there is a conditional independence relationship between the variables Y and Z, as $\omega_{YZ} = \omega_{ZY} = 0$.

If we consider each variable as a node and each edge as a partial correlation between the two variables it connects, the graph of this Gaussian graphical model illustrates the conditional dependence relationships among the variables. It's important to note that the partial correlation reflects the conditional correlation between two variables after accounting for the other variables in the model. $Figure\ 3$ displays the graph corresponding to the model having the concentration matrix as presented in equation (1.10).

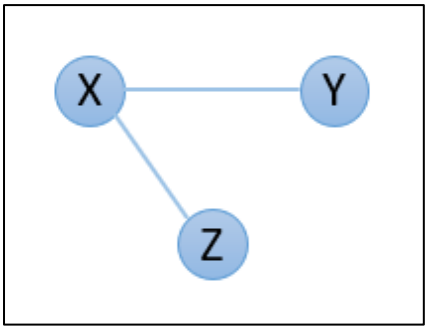


Figure 3: Undirected graph of a Gaussian graphical model.

1.3.2 Log-Likelihood

In a Gaussian graphical model, the log-likelihood function is used to estimate the parameters of the model. The log-likelihood function is the natural logarithm of the joint probability density function of the multivariate Gaussian distribution. The joint probability density function is a function of the mean vector μ and the concentration matrix Ω.

By utilizing equation (1.8), the log-likelihood function is given by,

$$L(\mu, \Omega) = -\frac{n}{2} k \log 2\pi + \frac{n}{2} \log|\Omega| - \frac{n}{2} tr(S\Omega) - \frac{n}{2} (\bar{x} - \mu)^T \Omega (\bar{x} - \mu) \qquad (1.11)$$

where,

- n is the number of observations,
- $|\Omega|$ denotes the determinant of the concentration matrix Ω,
- S is the sample covariance matrix,

$$S = \frac{1}{n} \sum_{i=1}^{n} \left(x^i - \bar{x}\right) \left(x^i - \bar{x}\right)^T$$

, and $\bar{x}$ the sample mean,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x^i ,$$

- $tr(S\Omega)$ denotes the trace of the product of S and Ω.

The maximized log-likelihood function is obtained by maximizing the log-likelihood function (1.11) with respect to the parameters μ and Ω. The maximization occurs at

$$\hat{\mu} = \bar{x} \tag{1.12}$$

and as the matrices $\hat{\Sigma}$ and S have different entries where $\omega_{ij} = 0$, we can deduce the following result,

$$tr(\hat{\Omega}S) = tr(\hat{\Omega}\hat{\Sigma}) = k\,. \tag{1.13}$$

By substituting equations (1.12) and (1.13) into equation (1.11), we obtain,

$$\hat{l} = -\frac{n}{2}k\log 2\pi + \frac{n}{2}\log|\Omega| - \frac{n}{2}k \tag{1.14}$$

(Edwards, 2000).

1.3.3 Goodness of fit

The deviance of a Gaussian graphical model is a measure of the goodness of fit of the model. It is calculated as the difference between the estimation of the maximized log-likelihood of the saturated model and the estimation of the maximized log-likelihood of a reduced model.

The formula for the deviance of a Gaussian graphical model is,

$$D = 2(\hat{l}_s - \hat{l}_m) \tag{1.15}$$

where l_s, l_m are the maximized log-likelihood function of the saturated model, and the reduced model respectively.

Utilizing equation (1.14), the estimation of the log-likelihood of the saturated model is,

$$\hat{l}_s = -\frac{n}{2}k\log 2\pi - \frac{n}{2}\log|S| - \frac{n}{2}k \tag{1.16}$$

where $\log|\Omega| = -\log\left|\hat{\Sigma}\right| = -\log|S|$ which is the sample covariance matrix of the data.

Similarly, the estimation of the log-likelihood of the reduced model is,

$$\hat{l}_m = -\frac{n}{2}k\log 2\pi - \frac{n}{2}\log\left|\hat{\Sigma}\right| - \frac{n}{2}k\,. \tag{1.17}$$

By combining equations (1.15), (1.16) and (1.17), we obtain,

$$D = n\log\frac{\left|\hat{\Sigma}\right|}{|S|}\,. \tag{1.18}$$

To further discuss the deviance, it has the potential to be utilized for the likelihood ratio test, which compares two nested models, one of which is a simplified or reduced version of the other. In the case of Gaussian graphical models, the LRT can be utilized to determine if a more intricate model with more edges (i.e., more nonzero elements in the concentration matrix) provides a better fit to the data than a simpler model with fewer edges.

The LRT statistic is based on the difference in deviance between the two models. From equation (1.18), it can be expressed as,

$$LRT = n\log\frac{\left|\hat{\Sigma}_0\right|}{\left|\hat{\Sigma}_1\right|} \;\sim\; X_p^2$$

where $\hat{\Sigma}_0$ and $\hat{\Sigma}_1$ represent the estimation of the covariance matrix of the simpler M_0 and more complex M_1 model, respectively.

The LRT statistic follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the two models. A large LRT value suggests that the more complex model fits the data significantly better than the simpler model, while a small value indicates that the simpler model is sufficient (Højsgaard et al., 2012).

1.3.4 Model Selection

As stated earlier, the significance of model selection lies in its ability to determine the most suitable model for a given dataset. Through the comparison of various models, we can determine the one that exhibits the best fit to the data and offers the most accurate predictions. This process simplifies the model by eliminating redundant variables. This enhances interpretability and reduces the potential for overfitting.

For selecting the best Gaussian graphical model, we employ the AIC and BIC criteria, which are also used in log-linear models.

The AIC (Akaike's Information Criterion) value is given by the equation,

$$AIC = -2\ln L + 2d\,,$$

while the BIC (Bayesian Information Criterion) value is determined by

$$BIC = -2\ln L + d\ln N$$

where,

- lnL is the maximized log-likelihood function,
- d is the number of parameters of the model under consideration and
- N is the number of observations.

When comparing multiple models, the models with lower AIC values are preferred. A lower AIC indicates a better trade-off between model fit and complexity. Thus, the model with the lowest AIC is considered the best model among the alternatives.

1.4 Mixed Interaction Models

Mixed interaction models are statistical models that combine both categorical and continuous variables to capture complex relationships between variables. These models are often represented using undirected graphical models, which provide a visual representation of the categorical associations and continuous dependencies between variables. By incorporating categorical and continuous variables, mixed interaction models combine the strengths of both log-linear and Gaussian graphical models.

1.4.1 Mixed Interaction Model

Let $D = (D_1, D_2, \dots, D_d)$ be d discrete variables and $C = (C_1, C_2, \dots, C_c)$ be c continuous variables of a dataset V of N observations. Furthermore, let I be the set of all possible combinations of values for the discrete variables. A cell of I is denoted as $i = (i_1, i_2, \dots, i_d)$ and it is an observation of the values of the discrete variables, where i_j is the value of the j-th variable in the cell. An entire row in the dataset (observation) is denoted as $(i, y) = (i_1, i_2, \dots, i_d, y_1, y_2, \dots, y_c)$. Additionally, the multivariate Gaussian distribution $N(\mu_i, \Sigma)$ represents the conditional distribution of the continuous variables C, when the discrete variables D are restricted to fall within a particular cell i.

The density function, also known as Conditional Gaussian density, can be expressed as

$$f(i, y) = p_i (2\pi)^{-\frac{c}{2}} \det(\Sigma)^{-\frac{1}{2}} exp\left(-\frac{1}{2}\,(y - \mu_i)^T \Sigma^{-1} (y - \mu_i)\right) \qquad (1.19)$$

where,

- $f(i, y)$ represents the density function, which calculates the probability density of the continuous variables C given the discrete variables D falling in cell i,
- p_i is the probability of the discrete variables falling in cell i,
- $(2\pi)^{-\frac{1}{2}}$ is a normalization constant,
- y represents the vector of continuous variables,
- μ_i represents the mean vector associated with cell i. It denotes the expected values of the continuous variables given the discrete variables falling in that cell, and
- Σ represents the covariance matrix,

(Wermuth & Lautitzen, 1990).

A key observation here is that, unlike the mean μ_i of the multivariate Gaussian distribution, the covariance matrix Σ remains constant and does not vary with different values of discrete variables. This property characterizes these models as homogeneous.

The Conditional Gaussian density function can also be expressed using equation (1.20). The parameters (p_i, μ_i, Σ) presented in equation (1.19) are referred to as moment parameters, and they can be converted into canonical parameters (g_i, h_i, K) by utilizing formulas (1.21), (1.22), and (1.23).

$$f(i, y) = exp\left(g_i + h_i^T y - \frac{1}{2} y^T K y\right) \tag{1.20}$$

$$g_i = \log(p_i) - \frac{c}{2}\log 2\pi - \frac{1}{2}\log\det\Sigma - \frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i \tag{1.21}$$

$$h_i = \Sigma^{-1}\mu_i \tag{1.22}$$

$$K = \Sigma^{-1} \tag{1.23}$$

By utilizing the canonical parameters (g_i, h_i, K), the formula for the mixed interaction model consists of a generating class G, which is divided into three distinct components. Each generator in class G is denoted as $(a, b) \in G$, where $a \in G \cap D$ represents the discrete variables and $b \in G \cap C$ represents the continuous variables (Højsgaard et al., 2012).

$Table\ 1$ presents the three components of the model (discrete, linear, quadratic), along with the corresponding canonical parameter formulas, as well as the generators of the model.

Model component	Canonical parameter formula	Generator
Discrete	$g_i = \sum_{(a,b)\in G} g_{i_a}$	$\max(\{a \mid (a,b) \in G\})$
Linear	$h_i = \sum_{(a,b)\in G} h_{i_a}^b$	$\max(\{a \mid (a,\ b) \in C \wedge c \in b\})$
Quadratic	$K = \begin{pmatrix} k_{c_1c_1} & \cdots & k_{c_1c_n} \\ \vdots & \ddots & \vdots \\ k_{c_nc_1} & \cdots & k_{c_nc_n} \end{pmatrix}$	$(\{b \mid (a,b) \in G\})$

Table 1: Mixed interaction model's formula

To provide better clarity, we will give an example.

Suppose we have four variables, two continuous C_1, C_2 and two discrete variables D_1, D_2. Let $Figure\ 4$ be a graph and formula (1.24) its corresponding model formula.

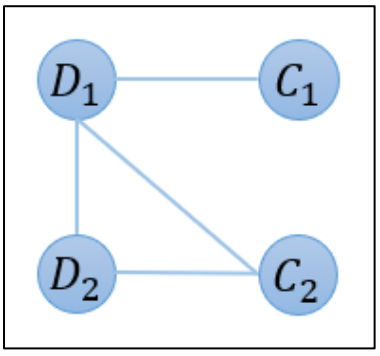


Figure 4: Mixed Interaction Model Graph

$$D_1 * D_2 * C_2 + D_1 * C_1 \qquad (1.24)$$

The model formula (1.24) is a summation of two terms, which are the generators of the generating class G.

$$G = \{(D_1 * D_2 * C_2), (D_1 * C_1\ )\}$$

The first component of the model is called discrete, and it is generated by considering the maximal set of discrete variables within the generating class G. The first generator $D_1 * D_2 * C_2$ gives the set of discrete variables (D_1, D_2), while the second generator $D_1 * C_1$ yields the set (D_1). The appropriate set for the discrete generator is the maximal set among these two sets. Thus, the first set is the correct choice.

The second component is referred to as linear, and it is determined by the maximal set of discrete variables from the generating class G for each continuous variable. In this case, we obtain the set (D_1) from the first generator, and the set (D_1, D_2) from the second generator.

Finally, the third component of the model, known as quadratic, is the sets of the continuous variables belonging to the generating class G. The set of continuous variables obtained from the first generator is (C_2), while the set derived from the second generator is (C_1).

$Table\ 2$ summarizes the results.

Model component	Canonical parameter formula	Generator
Discrete	$g_i = u + u_m^{D_1} + u_l^{D_2} + u_{ml}^{D_1 D_2}$	$\{(D_1, D_2)\}$
Linear	$h_i^{C_1} = v + v_m^{D_1}$ $h_i^{C_2} = v + v_m^{D_1} + v_l^{D_2} + v_{ml}^{D_1 D_2}$	$\{(D_1)\}$ $\{(D_1, D_2)\}$
Quadratic	$K = \begin{pmatrix} k_{c_1 c_1} & 0 \\ 0 & k_{c_2 c_2} \end{pmatrix}$	$\{(C_1), (C_2)\}$

Table 2: MIM's formula Example

1.4.2 Log-Likelihood

In a mixed interaction model, the likelihood function represents the probability of observing the given data under the assumed model. The maximum likelihood estimation approach is used to estimate the moment parameters (p_i, μ_i, Σ), which characterize the model. These parameters offer the best explanation for the observed data. The MLE procedure considers both the categorical and continuous variables, effectively capturing their complex interactions and dependencies.

By utilizing equation (1.19), the log-likelihood function is given by,

$$\log f(i, y) = \log p_i - \frac{c}{2}\log 2\pi - \frac{1}{2}\log\det\Sigma - \frac{1}{2}\left((y-\mu_i)^T\Sigma^{-1}(y-\mu_i)\right) \quad (1.25)$$

The maximized log-likelihood function is obtained by maximizing the log-likelihood function (1.25) with respect to the parameters i and y. The maximization occurs at

$$\hat{p} = \frac{n_i}{N} \quad (1.26)$$

$$\hat{\mu}_i = \bar{y}_i \quad (1.27)$$

$$\hat{\Sigma} = S = \sum_i \frac{n_i S_i}{N} \quad (1.28)$$

By simplifying and substituting equations from (1.26) to (1.28) into equation (1.25), we obtain,

$$\hat{l} = \sum_i n_i \log\left(\frac{n_i}{N}\right) - \frac{1}{2}Nc\ log2\pi - \frac{1}{2}N\log\det\Sigma - \frac{1}{2}Nc \quad (1.29)$$

(Højsgaard et al., 2012).

1.4.3 Goodness of fit

The deviance of a mixed interaction model is a measure of the goodness of fit. It is calculated as the difference between the estimation of the maximized log-likelihood of the saturated model and the estimation of the maximized log-likelihood of a reduced model. The deviance is important in model selection and evaluation within the framework of mixed interaction models. A low deviance value is desired, as it indicates a better fit of the model to the data.

The formula for the deviance of a mixed interaction model is,

$$D = 2\left(\hat{l}_s - \hat{l}_m\right) \tag{1.30}$$

where l_s, l_m are the maximized log-likelihood function of the saturated model, and the reduced model respectively.

Utilizing equation (1.29), the deviance formula (1.30) takes the form,

$$\begin{aligned} D = 2\sum_i n_i \log\left(\frac{n_i}{\hat{m}_i}\right) - N\log\det\left(S\hat{\Sigma}^{-1}\right) + N\left(tr\left(S\hat{\Sigma}^{-1}\right) - c\right) \\ + \sum_i n_i(\bar{y}_i - \hat{\mu}_i)^T\hat{\Sigma}^{-1}(\bar{y}_i - \hat{\mu}_i) \end{aligned} \tag{1.31}$$

(Højsgaard et al., 2012).

1.4.4 Model Selection

Model selection is an important aspect of statistical analysis. It can help us to determine the most appropriate model for a given dataset. By evaluating and comparing different models, we can find the one that demonstrates the strongest alignment with the data and provides the most reliable predictions. The selection process simplifies the model by eliminating variables, resulting in improved interpretability and reduced risks of overfitting. In the context of mixed interaction models, the criteria AIC and BIC can be utilized. These criteria serve as valuable tools for evaluating the goodness of fit and complexity of models.

As already stated in paragraph $1.3.4$, the AIC (Akaike's Information Criterion) value takes the following form,

$$AIC = -2\ln L + 2d\,,$$

while the BIC (Bayesian Information Criterion) value is obtained through the equation,

$$BIC = -2\ln L + d\ln N$$

where,

- lnL is the maximized log-likelihood function,
- d is the number of parameters of the model under consideration and
- N is the number of observations.

As previously mentioned, during the comparison of multiple models, those with lower AIC or BIC values are selected. Generally, a lower value signifies a balance between model fit and complexity. Thus, the model that exhibits the lowest AIC or BIC is regarded as the best choice among the available alternatives.

Chapter 2- Directed Graphical Models

In this chapter, we will introduce chain graph models, an important subset of directed graphs. Directed graphical models serve as valuable tools for representing and analyzing complex dependencies. In particular, chain graph models find their strength in cases where variables can be grouped into distinct blocks. The features that set them apart are their reliance on the ordering between the distinct blocks of variables, while they do not permit ordering within each of these blocks.

2.1 Chain Graph Models

Chain graph models are graphical models that represent conditional dependencies between variables using a graph structure. These models allow for both directed and undirected edges, capturing a more flexible range of relationships. An additional feature of these models is that they can be used when the given dataset consists of both continuous and discrete variables. They have various applications in fields such as genetics and social sciences.

2.1.1 Chain Graph Model

$Figure\ 5$ provides a simple example of a chain graph model consisting of two blocks. Similarly, to undirected graphical models, each variable of the model is represented by a circle or node. The connections among the nodes are represented by edges providing a clear visualization of the conditional relationships between the variables. By analyzing and interpreting the edges, we can gain insights into the potential influence of one variable on another.

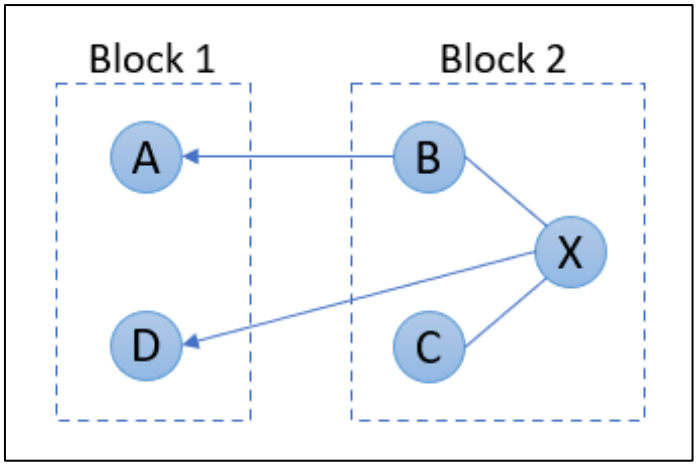


Figure 5: Chain Graph Model

To construct a chain graph, we utilize a block grouping mechanism. Variables are divided into different blocks based on their role in the model. Within each block the variables are considered concurrent, and each variable is regressed on all the variables which belong to all blocks located to the right. In $Figure\ 5$, variables B, C, X belong in block 2, thus they are concurrent, and variable A, which belongs in block 1, is regressed on all previous variables (the variables belonging to all prior blocks), B, C, X.

Furthermore, the block located to the left of the chain graph contains the purely response variables, while the block located to the right contains the purely explanatory variables. All the intermediate variables belong to the in between blocks.

It is important to mention that valuable information is provided not only by present, but also by missing edges. The absence of an edge between two variables indicates their conditional independence.

2.1.2 Density Function

In a chain graph model, the joint density function is factorized based on the connected components of the underlying graph. The connected components of such a graph are the subgraphs which remain after eliminating all arrows (directed edges) of the initial graph.

Consider a graph of a chain graph model, and a set of its connected components $g_1, g_2, \dots, g_n$. The joint density function is given by the following formula,

$$f = \prod_{i=1}^{n} f_{g_i|g>i}$$

where,

- $f_{g_i|g>i}$ represents the conditional density function of the variables in the connected component g_i given all previous connected components g_j, for $j > i$.

(Wermuth & Sadeghi, 2012).

It is worth highlighting that the form of the conditional density function depends on the type of variables involved in the model and each connected component (continuous, discrete), as well as the modeling assumptions (Cox & Wermuth, 1993).

2.1.3 Model Selection

The process of selecting the most appropriate model is crucial, as it allows us to identify the model which best aligns with the dataset and as a result it can provide more accurate predictions. By thoroughly evaluating and comparing various models, redundant variables are removed of the studied model, offering on one hand better interpretability, and on the other hand minimized overfitting risk. Similar to the prior models discussed, in the case of chain graph models we can also utilize the model selection criteria AIC and BIC.

The AIC (Akaike's Information Criterion) criterion retains the form,

$$AIC = -2\ln L + 2d\,,$$

And corresponding formula of the BIC (Bayesian Information Criterion) criterion is,

$$BIC = -2\ln L + d\ln N$$

where,

- lnL is the maximized log-likelihood function,
- d is the number of parameters of the model under consideration and
- N is the number of observations.

As stated earlier, the objective is to identify the model that exhibits the lowest value in either of these two criteria.

Chapter 3- Data Analysis

In this chapter, we perform statistical data analysis using R. We focus on a specific dataset, and we utilize it to apply three different techniques to study and compare their results. First, an undirected Gaussian graphical model is applied to the dataset, then a mixed interaction model is employed, and finally, we conduct statistical analysis using a chain graph model.

3.1 Dataset Presentation

In this thesis, we conducted an analysis of post-COVID data obtained from a study following patients who were hospitalized for COVID-19 in Italy. This survey was carried out in the year 2021. The dataset comes from the research grant entitled "Multidimensional statistical analysis of databases relating to patients affected by "long-Covid" in order to characterize their manifestations, identify their risk factors and identify their therapeutic trajectories.", which was provided and funded by the University of Perugia, in Italy. The primary objective of this study is to investigate and gain insights into the long-term predictors of the health consequences of this disease. The dataset used in this analysis consists of 108 observations and 106 variables, providing information for our investigation.

The survey included various types of data,

- Demographic data: Age, gender
- Anthropometric data: Height, weight, body circumference (arm, waist, hip, neck)
- Symptoms: Fatigue, rash, sleep, diarrhea, etc.
- Medical history & treatments: Diabetes, insulin, etc.
- Hospitalization data: Admission, follow-up visit, respiratory support, etc.
- Habits: Smoking
- Laboratory data: Blood sample (Ferritin, DDIMER, fibrinogen, etc.)
- Pulmonary & cardiac function: CT scan, spirometry, echocardiogram, etc.
- Cardiorespiratory test: the ratio of minute ventilation to carbon dioxide production, Lactate threshold, etc.
- Psychological data: stress, depression, etc.

Given the extensive number of variables compared to the number of observations, a selection process was needed to focus on the most important factors. We identified and prioritized the most relevant variables from the above data categories. This approach ensures that our analysis is able to capture the essential aspects of the study. The selected variables are presented in $Table\ 3$.

Explanatory Variable	Type	Description	Source Data
BMI	Continuous	Body mass index $\left(\frac{kg}{m^2}\right)$	Anthropometric data
TIME_UNTIL_FOLUP	Continuous	Duration between the patient's discharge and the follow-up visit (days)	Hospitalization data
AGE	Continuous	Patient's age (years)	Demographic data
GENDER	Categorical	1: Female 2: Male	Demographic data
SMOKING	Categorical	1: Nonsmoker 2: Smoker	Health Habit data
Response Variable	Type	Description	Source Data
VE.VCO2.SLOPE	Continuous	Ratio of minute ventilation to carbon dioxide production	Cardiorespiratory test
VO2	Continuous	maximal oxygen consumption	Cardiorespiratory test
LT	Continuous	Lactate threshold	Cardiorespiratory test
FEV1	Continuous	Max exhaled air volume in 1 second after max inhalation	Spirometry test
FERRITIN	Continuous	Cellular iron-storing protein	Laboratory data
DDIMER	Continuous	Blood clotting test	Laboratory data
IES.R.TOTAL	Continuous	Impact of Event Scale-Revised Self-assessment of post-traumatic stress symptom severity	Psychological data
DEPRESSION	Categorical	Self-assessment 1: Normal 2: Moderate 3: Severe	Psychological data
ICU	Categorical	Intensive Care Unit 1: No 2: Yes	Hospitalization data
FATIGUE	Categorical	Self-assessment 1: No 2: Yes	Symptoms
HRTC	Categorical	High Resolution Computed Tomography 1: Normal 2: Lesions 3: Severe Lesions	CT scan
HRTC_2cat	Categorical	1: merge HRTC's categories 2: HRTC's NAs	CT scan

Table 3: Post-Covid Dataset

3.2 Preprocessing Data

3.2.1 Create New Variables

By utilizing R and the given initial dataset we created the explanatory variables AGE, BMI and $TIME\_UNTIL\_FOLUP$, as well as the response variable $HRTC\_2cat$. These variables were created as they often appear to be important in various research.

- The explanatory variable AGE was derived by calculating the difference between each patient's admission date to the hospital and their birth date.
- For the BMI variable, we used the available data on the patients' weight and height in the formula,

$$\frac{Weight}{Height^2} \quad \left(\frac{kg}{m^2}\right)$$

- The variable $TIME\_UNTIL\_FOLUP$ was created by calculating the interval between the admission date to the hospital and the subsequent date of the follow-up visit.
- The response variable $HRTC\_2cat$ was determined by the given variable $HRTC$. By merging $HRTC$'s three categories into one, we created $HRTC\_2cat$'s first category, and the second category of $HRTC\_2cat$ consists of all $HRTC$'s missing values.

3.2.2 Define Subsets of Data

It is important to mention that for research purposes in each technique we used a slightly different subset of the variables shown in $Table\ 3$. The following $Tables\ 4, 7$ and 10 present these subsets. Along with them, we provide $Tables\ 5, 8$ and 11 giving information about the dimensions of each dataset, the missing values, and the duplicated rows. Additionally, we present $Tables\ 6, 9$ and 12 with a more detailed exploration of the missing values for each variable.

UNDIRECTED GAUSSIAN GRAPHICAL MODEL			
Explanatory Variable	Type	Description	Source Data
BMI	Continuous	Body mass index $\left(\frac{kg}{m^2}\right)$	Anthropometric data
TIME_UNTIL_FOLUP	Continuous	Duration between the patient's discharge and the follow-up visit (days)	Hospitalization data
AGE	Continuous	Patient's age (years)	Demographic data
Response Variable	Type	Description	Source Data
VE.VCO2.SLOPE	Continuous	Ratio of minute ventilation to carbon dioxide production	Cardiorespiratory test
VO2	Continuous	maximal oxygen consumption	Cardiorespiratory test
LT	Continuous	Lactate threshold	Cardiorespiratory test
FEV1	Continuous	Max exhaled air volume in 1 second after max inhalation	Spirometry test
FERRITIN	Continuous	Cellular iron-storing protein	Laboratory data
DDIMER	Continuous	Blood clotting test	Laboratory data
IES.R.TOTAL	Continuous	Impact of Event Scale-Revised Self-assessment of post-traumatic stress symptom severity	Psychological data

Table 4: Dataset subset - Undirected Gaussian Graphical Model

Rows	Columns	Missing values	Duplicated rows
108	10	181	2

Table 5: Dataset basic information - UGGM

Variable	Type	Na count	Na %
AGE	numeric	2	1.85%
BMI	numeric	12	11.11%
TIME_UNTIL_FOLUP	numeric	13	12.04%
DDIMER	numeric	28	25.93%
FERRITIN	numeric	17	15.74%
FEV1	numeric	17	15.74%
VO2	numeric	23	21.3%
VE.VCO2.SLOPE	numeric	23	21.3%
LT	numeric	22	20.37%
IES.R.TOTAL	numeric	24	22.22%

Table 6: Missing values information for each variable - UGGM

MIXED INTERACTION MODEL			
Explanatory Variable	Type	Description	Source Data
BMI	Continuous	Body mass index $\left(\frac{kg}{m^2}\right)$	Anthropometric data
TIME_UNTIL_FOLUP	Continuous	Duration between the patient's discharge and the follow-up visit (days)	Hospitalization data
AGE	Continuous	Patient's age (years)	Demographic data
GENDER	Categorical	1: Female 2: Male	Demographic data
SMOKING	Categorical	1: Nonsmoker 2: Smoker	Health Habit data
Response Variable	Type	Description	Source Data
VE.VCO2.SLOPE	Continuous	Ratio of minute ventilation to carbon dioxide production	Cardiorespiratory test
VO2	Continuous	maximal oxygen consumption	Cardiorespiratory test
LT	Continuous	Lactate threshold	Cardiorespiratory test
FEV1	Continuous	Max exhaled air volume in 1 second after max inhalation	Spirometry test
FERRITIN	Continuous	Cellular iron-storing protein	Laboratory data
DDIMER	Continuous	Blood clotting test	Laboratory data
IES.R.TOTAL	Continuous	Impact of Event Scale-Revised Self-assessment of post-traumatic stress symptom severity	Psychological data

Table 7: Dataset subset - Mixed Interaction Model

Rows	Columns	Missing values	Duplicated rows
108	12	195	2

Table 8: Dataset basic information – MIM

Variable	Type	Na count	Na %
AGE	numeric	2	1.85%
BMI	numeric	12	11.11%
TIME_UNTIL_FOLUP	numeric	13	12.04%
GENDER	factor	0	0%
SMOKING	factor	14	12.96%
VO2	numeric	23	21.3%
DDIMER	numeric	28	25.93%
FERRITIN	numeric	17	15.74%
FEV1	numeric	17	15.74%
IES.R.TOTAL	numeric	24	22.22%
VE.VCO2.SLOPE	numeric	23	21.3%
LT	numeric	22	20.37%

Table 9: Missing values information for each variable – MIM

CHAIN GRAPH MODEL			
Explanatory Variable	Type	Description	Source Data
BMI	Continuous	Body mass index $\left(\frac{kg}{m^2}\right)$	Anthropometric data
TIME_UNTIL_FOLUP	Continuous	Duration between the patient's discharge and the follow-up visit (days)	Hospitalization data
AGE	Continuous	Patient's age (years)	Demographic data
GENDER	Categorical	1: Female 2: Male	Demographic data
SMOKING	Categorical	1: Nonsmoker 2: Smoker	Health Habit data
Response Variable	Type	Description	Source Data
VO2	Continuous	maximal oxygen consumption	Cardiorespiratory test
FEV1	Continuous	Max exhaled air volume in 1 second after max inhalation	Spirometry test
DDIMER	Continuous	Blood clotting test	Laboratory data
DEPRESSION	Categorical	Self-assessment 1: Normal 2: Moderate 3: Severe	Psychological data
ICU	Categorical	Intensive Care Unit 1: No 2: Yes	Hospitalization data
FATIGUE	Categorical	Self-assessment 1: No 2: Yes	Symptoms
HRTC	Categorical	High Resolution Computed Tomography 1: Normal 2: Lesions 3: Severe Lesions	CT scan
HRTC_2cat	Categorical	1: merge HRTC's categories 2: HRTC's NAs	CT scan

Table 10: Dataset subset - Chain Graph Model

Rows	Columns	Missing values	Duplicated rows
108	13	199	2

Table 11: Dataset basic information – CGM

Variable	Type	Na count	Na %
GENDER	factor	0	0%
ICU	factor	12	11.11%
SMOKING	factor	14	12.96%
AGE	numeric	2	1.85%
BMI	numeric	12	11.11%
TIME_UNTIL_FOLUP	numeric	13	12.04%
FATIGUE	factor	14	12.96%
DEPRESSION	factor	24	22.22%
HRTC	factor	40	37.04%
HRTC_2cat	factor	0	0%
VO2	numeric	23	21.3%
FEV1	numeric	17	15.74%
DDIMER	numeric	28	25.93%

Table 12: Missing values information for each variable – CGM

3.2.3 Dealing with Missing Values

To handle missing values in all three sub datasets, we followed an approach which involves two steps.

The first step is filtering the dataset. With the help of R, firstly we identify and then remove the rows which have a high percentage of missing values, specifically those with less than 70% of complete values. This step reinforces the reliability of the remaining data for a meaningful analysis.

The second step is filling in the rest of the missing values in the dataset. We use the function $knnImputation()$ which performs the k-nearest neighbors (KNN) algorithm. This algorithm takes into consideration the values of neighboring data points (the default number of neighbors is 10) and imputes the missing values based on them. The imputation based on the complete observations maintains the integrity and completeness of the dataset.

This approach provides us with an appropriate dataset for analysis and modeling.

3.2.4 Preliminary Analysis

Before we delve into the applications of graphical model techniques, a preliminary analysis will be conducted. The following analysis centers on the pairwise examination of the variables listed in $Table\ 3$. The analysis's steps are presented in $Table\ 13$ and through them we can gain insights into the relationships and dependencies between variables.

Covariate \ Response	Quantitative	Qualitative
Quantitative	1. **Scatter plots** • y axis: response var • x axis: covariate • nonparametric regression line 2. **Cause-effect relationship** t-test Simple Linear Regression • what is the effect of covariate on response? 3. **Spearman's Correlation test**: rank-based non-parametric correlation that does not depend on the distribution of the data (avoid the effect of outliers) • How are response var and covariate related?	1. **Box plots** (y axis: covariate, x axis: response var) 2. **Comparison**: Density plots • if Gaussian and ○ 2 groups: Independent t-test ○ 3+ groups: ANOVA • If not Gaussian and ○ 2 groups: Wilcoxon rank-sum (Mann-Whitney U) ○ 3+ groups: Kruskal-Wallis test • Is there a difference between response's groups based on the covariate?
Qualitative	1. **Box plots** (y axis: response var, x axis: covariate) 2. **Comparison**: Density plots • if Gaussian and ○ 2 groups: Independent t-test ○ 3+ groups: ANOVA ○ What is the difference in response for patients from different covariate's groups? • If not Gaussian and ○ 2 groups: Wilcoxon rank-sum (Mann-Whitney U) ○ 3+ groups: Kruskal-Wallis test ○ Is there a difference between covariate's groups based on response?	1. **Bar plots** 2. **Correlation**: Chi square test of independency (non-parametric) • How are response var and covariate related?

Table 13: Pairwise Preliminary Analysis

3.2.4.1 Quantitative explanatory variables vs Quantitative Response variable

1. **Scatter plots**
   - Is the non-parametric regression line close to linear? Is the slope positive, negative, or close to zero?
   - The blue line represents the non-parametric regression line.

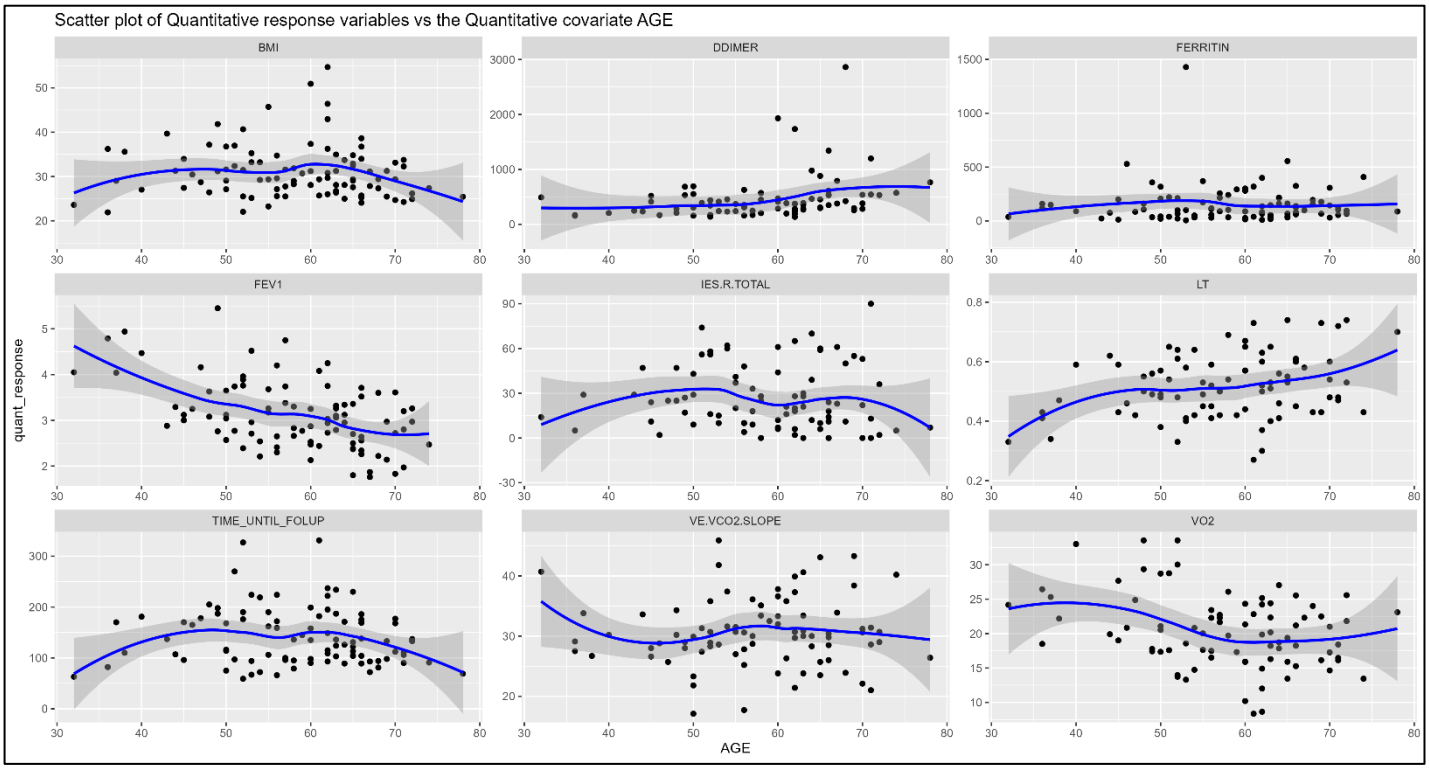


Figure 6: Scatter plots of Quantitative response variables vs the covariate AGE

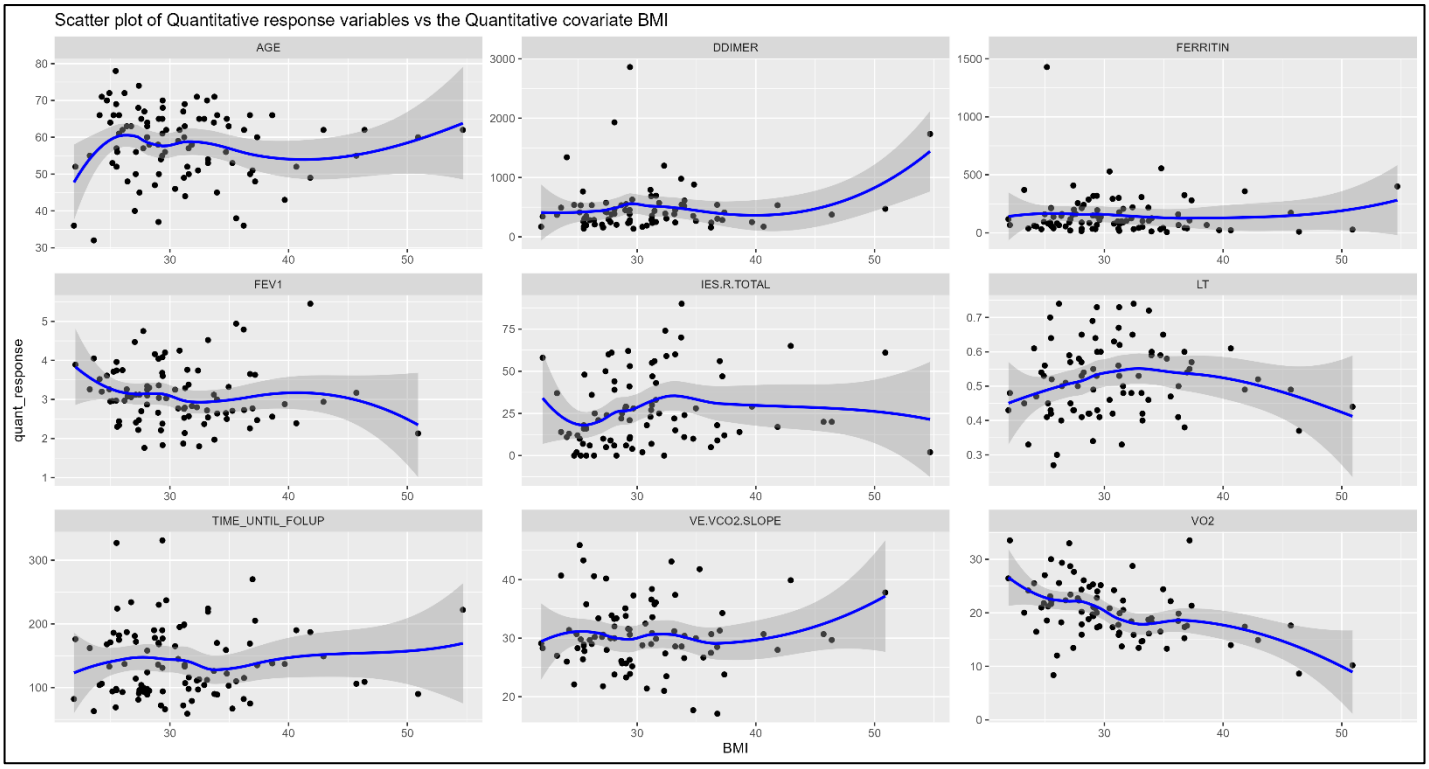


Figure 7: Scatter plots of Quantitative response variables vs the covariate BMI

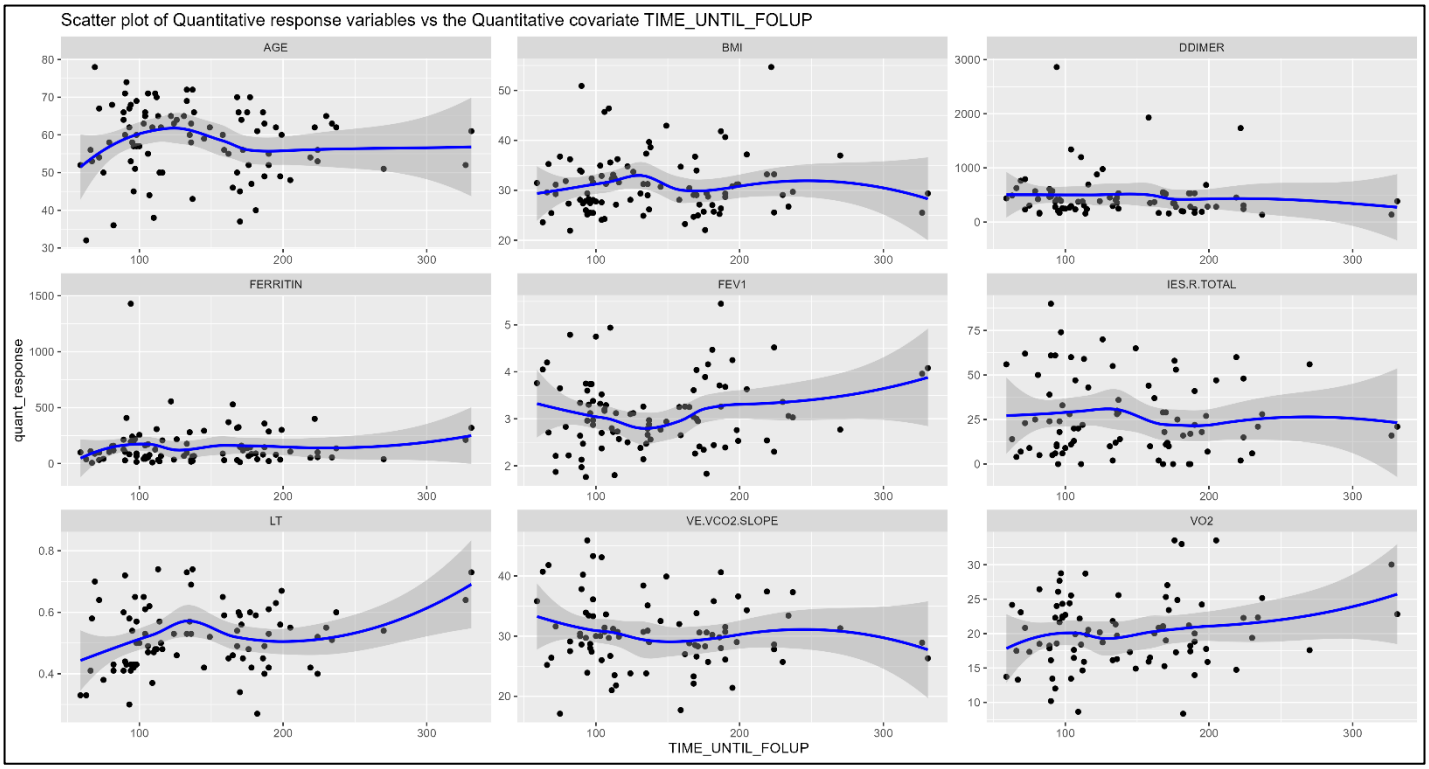


Figure 8: Scatter plots of Quantitative response variables vs the covariate TIME_UNTIL_FOLUP

Results

$Table\ 14$ presents three columns, one for each explanatory variable, and the rows show the response variables that appear to have linear relationship with the corresponding covariate based on Scatterplots in $Figures\ 6, 7$ and 8.

The symbol " $+$ " means that the non-parametric regression line has a positive slope and " $-$ " indicates a negative slope.

AGE	*BMI*	*TIME_UNTIL_FOLUP*
- VO2	- VO2	+ VO2
- FEV1	+ DDIMER	- VE.VCO2.SLOPE
+ LT	- FEV1	+ FEV1
+ DDIMER	+ VE.VO2.SLOPE	+ LT

Table 14: Scatterplots' comments

2. Simple Linear Regression **t-test**

- What is the effect of the covariate on the response variable?
- If $\boldsymbol{p_{value} < 0.05}$, then we reject the $H_0: \beta = 0$. This indicates that the covariate's value does not affect the value of the response variable.
- In $Table\ 15$ we present the p_{value} for each t-test.

AGE	*p-value*	*BMI*	*p-value*	*TIME_UNTIL_FOLUP*	*p-value*
VO2	***0.004***	**VO2**	***<0.001***	**VO2**	*0.11*
LT	***0.005***	**LT**	*0.7*	**LT**	*0.071*
FEV1	***<0.001***	**FEV1**	*0.4*	**FEV1**	*0.2*
FERRITIN	*0.9*	**FERRITIN**	*0.8*	**FERRITIN**	*0.8*
DDIMER	***0.009***	**DDIMER**	*0.2*	**DDIMER**	*0.4*
IES.R.TOTAL	*0.7*	**IES.R.TOTAL**	*0.2*	**IES.R.TOTAL**	*0.5*
VE.VCO2.SLOPE	*0.8*	**VE.VCO2.SLOPE**	*>0.9*	**VE.VCO2.SLOPE**	*0.4*

Table 15: t-test results

3. **Spearman's Correlation test**: rank-based non-parametric correlation
   - How are response variable and covariate related?
   - Significance level **0.2**

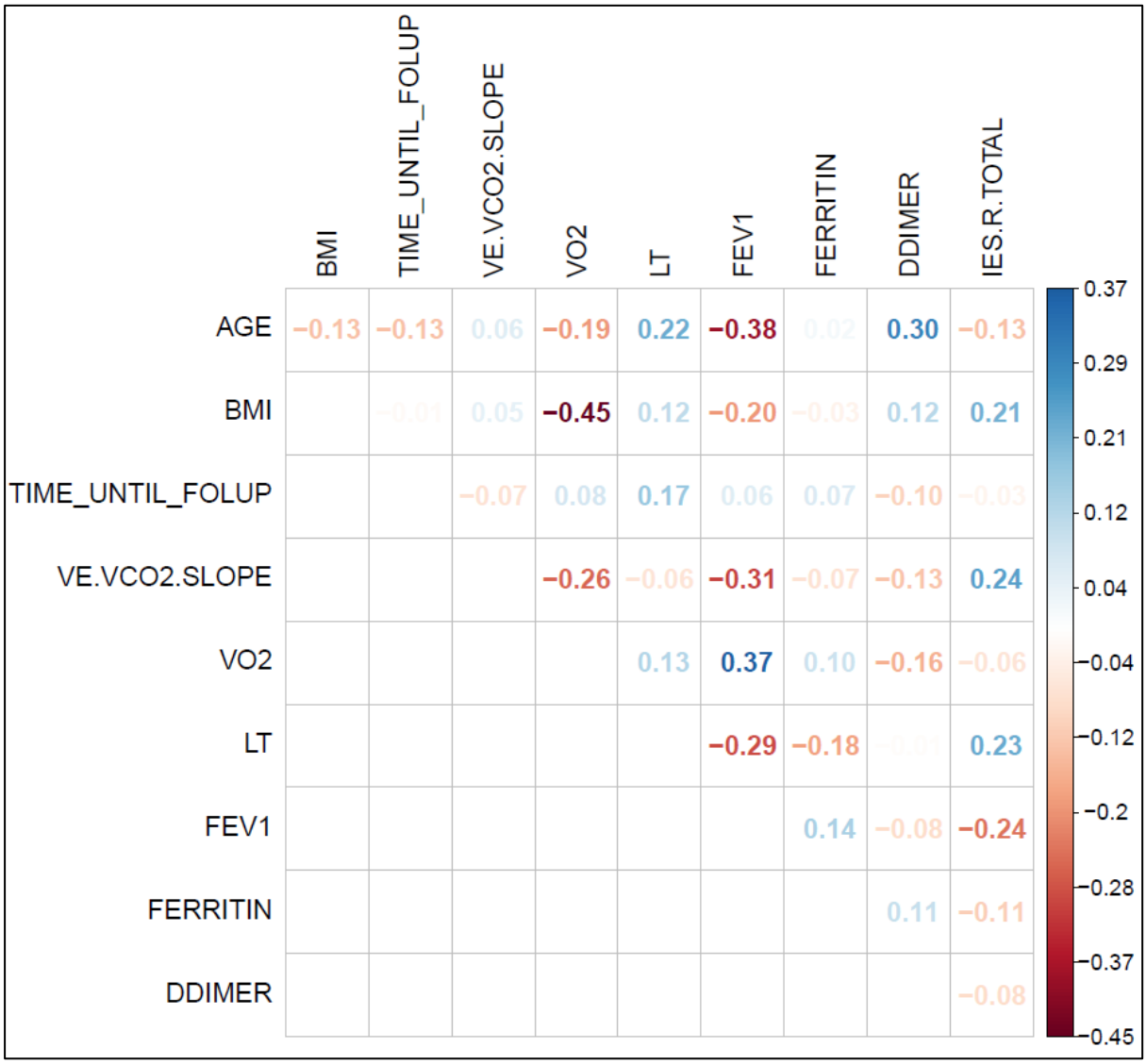


Figure 9: Correlation matrix

Summary

In $Table\ 14$, we present the pairs (covariate, response) that occur to have a "strong" relationship in the *Scatter plots*, the *t-test* and the *Spearman's correlation test* (in all three).

The symbol " $+$ " at the front of the response variable means that as the value of the covariate increases the value of the response variable increases too, while the symbol " $-$ " indicates that as the value of the covariate increases the value of the response variable decreases.

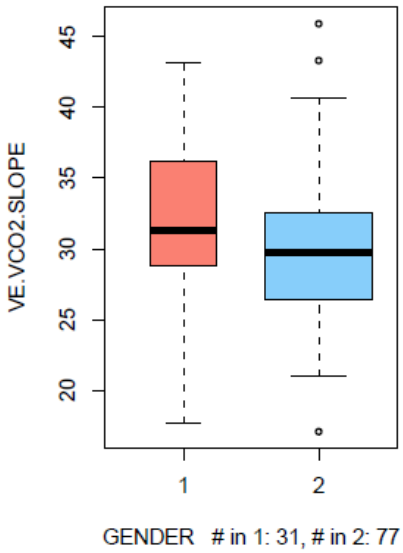
AGE	*BMI*	*TIME_UNTIL_FOLUP*
- VO2	- VO2	+ LT
+ LT		
- FEV1		
+ DDIMER		

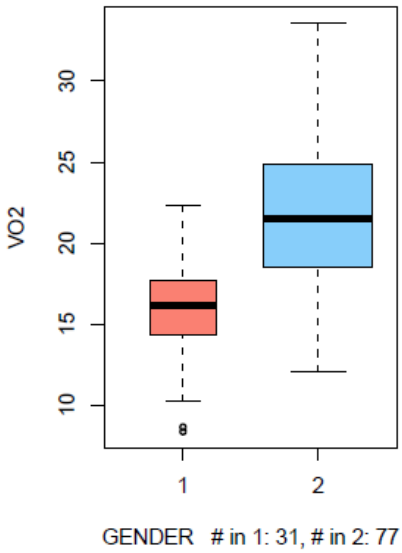
Table 16: Results Quantitative explanatory variables vs subset Quantitative Response variable.

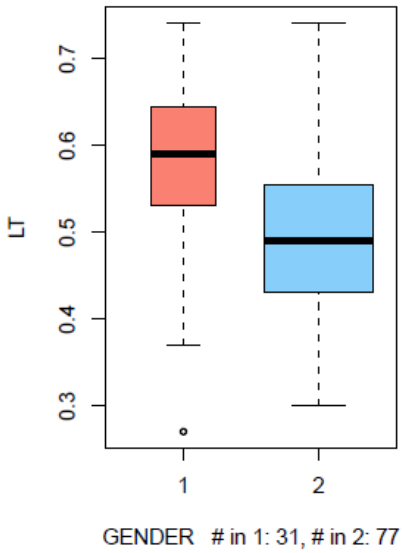
3.2.4.2 Qualitative explanatory variables vs Quantitative Response variable

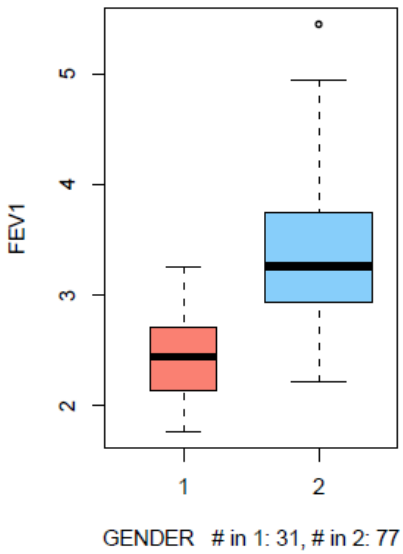
1. **Box plots** (y axis: response var, x axis: covariate)

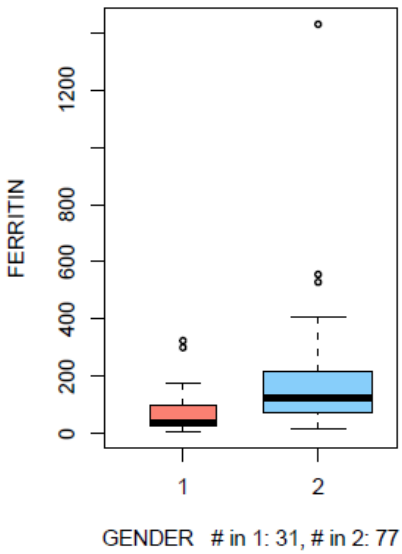
GENDER

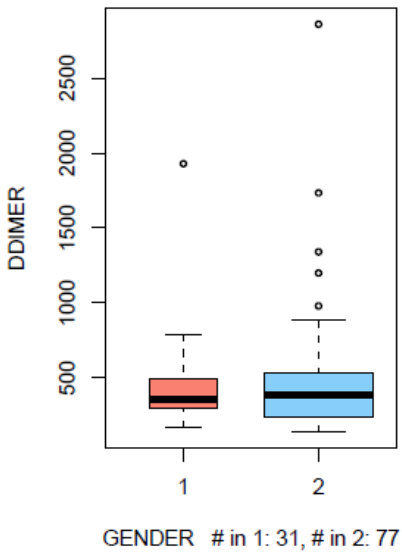


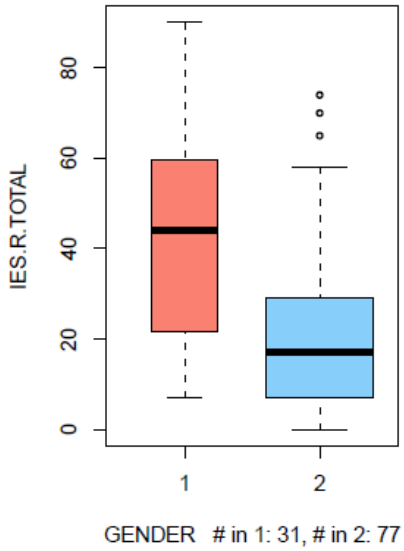




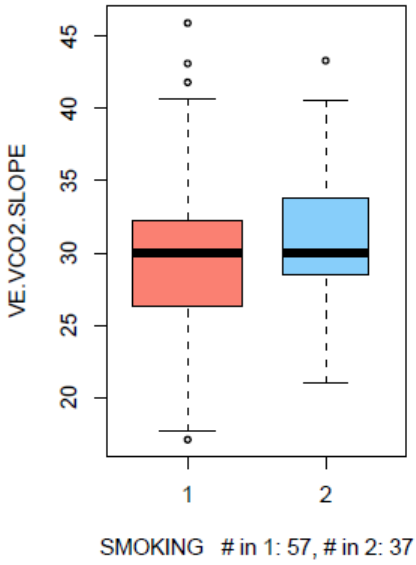


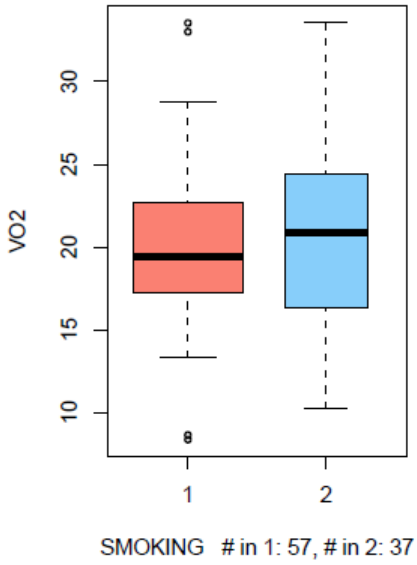


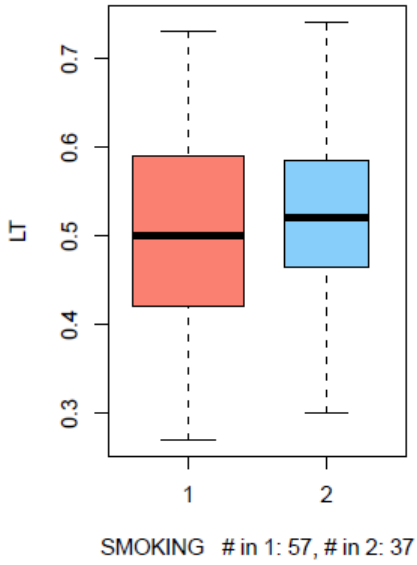


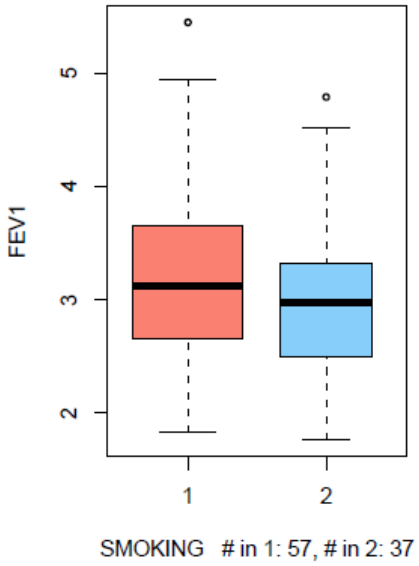

IES.R.TOTAL
0
20
40
60
80
1
2
GENDER # in 1: 31, # in 2: 77

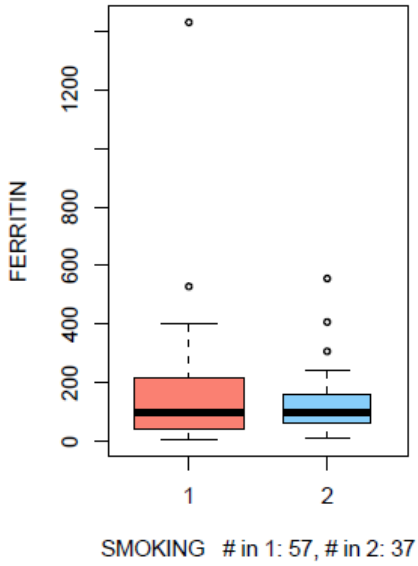
SMOKING

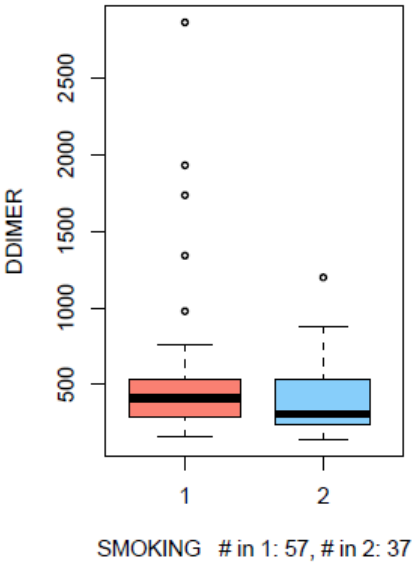

VE.VCO2.SLOPE
20
25
30
35
40
45
1
2
SMOKING # in 1: 57, # in 2: 37

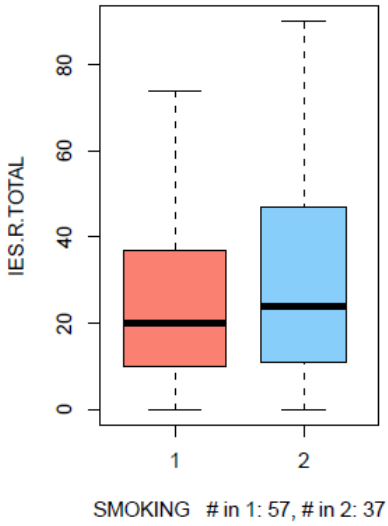

VO2
10
15
20
25
30
1
2
SMOKING # in 1: 57, # in 2: 37


LT
0.3
0.4
0.5
0.6
0.7
1
2
SMOKING # in 1: 57, # in 2: 37









Results

In $Table\ 17$, we have two columns, one for each explanatory variable, and the rows present the response variables that appear to affect the groups of each qualitative covariate based on the Box plots.

GENDER	*SMOKING*
VO2	DDIMER
LT	
FEV1	
FERRITIN	
IES.R.TOTAL	

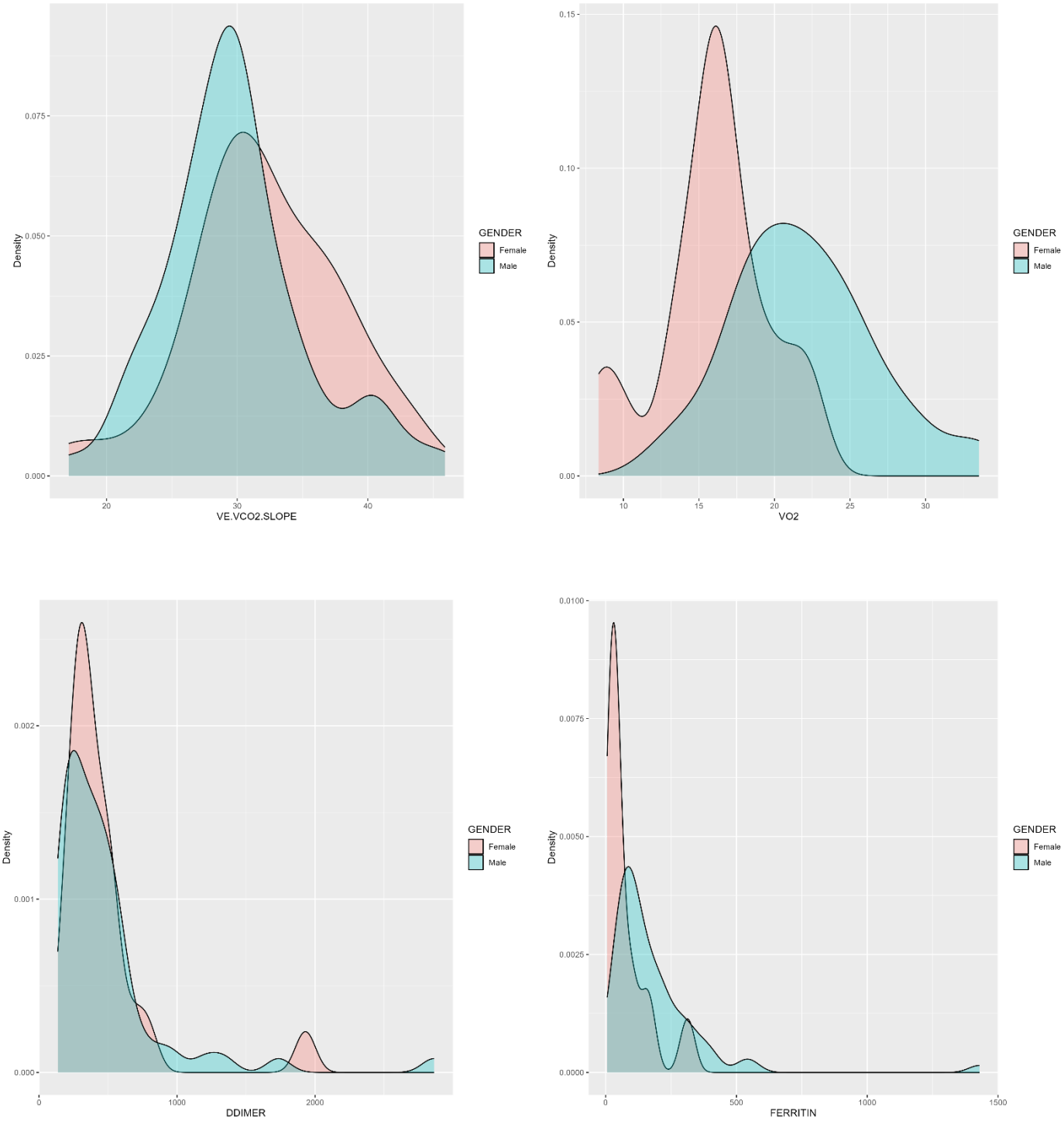
Table 17: Boxplots comments

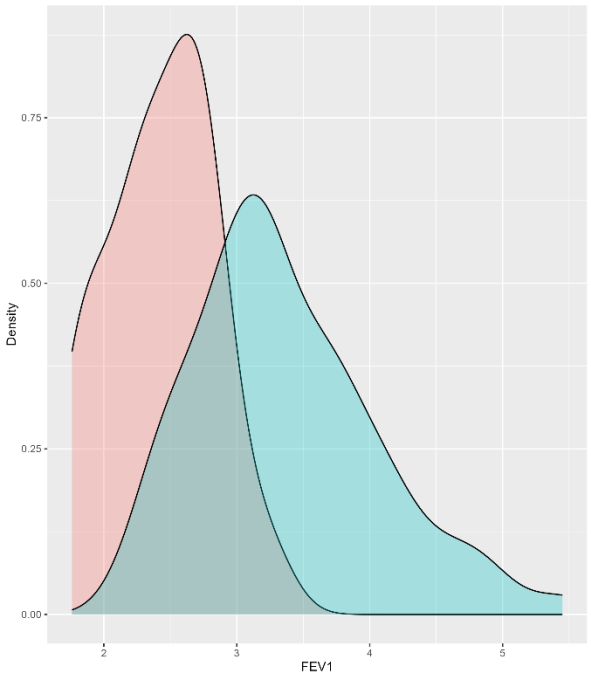
2. Comparison: **Density plots** & **Shapiro-Wilk test**

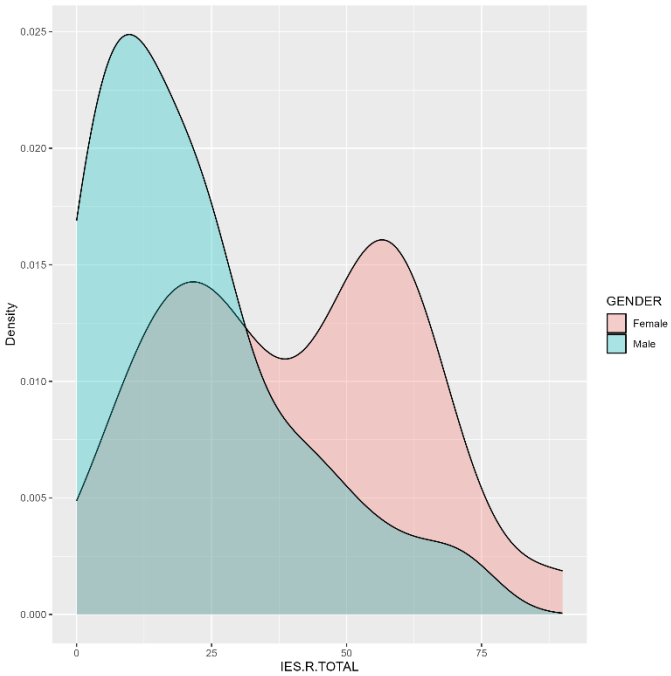
- Shapiro-Wilk test: $\boldsymbol{p\_value < 0.05}$ implies that the distribution of each group's data is significantly different from the normal distribution.

Density Plots

GENDER

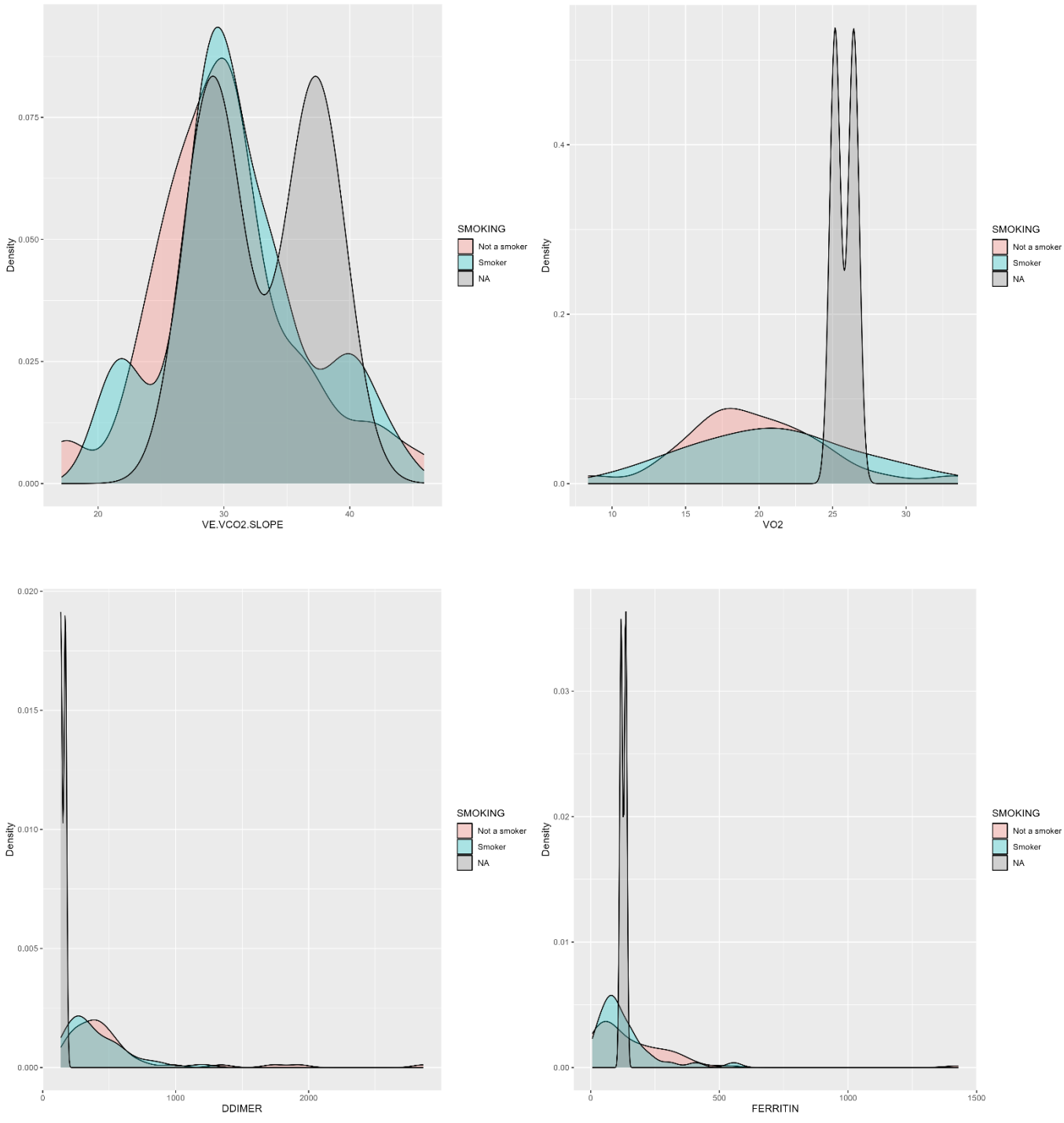


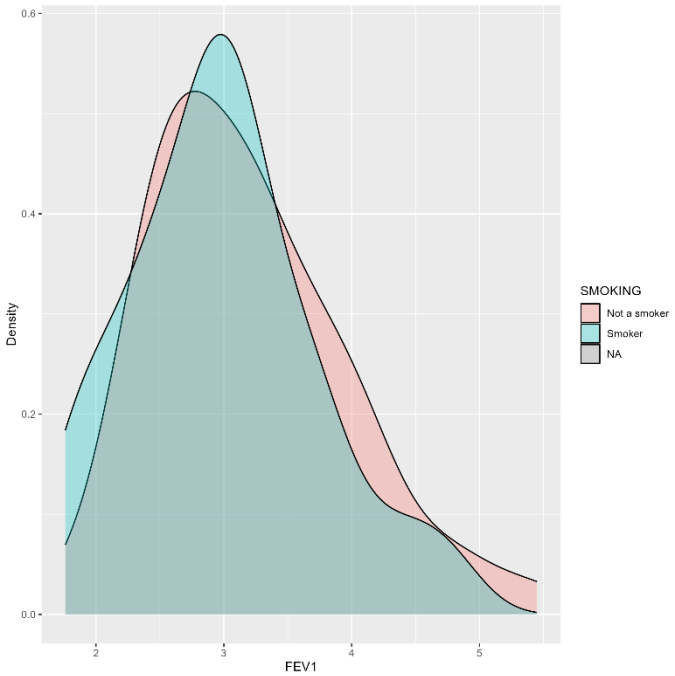

Density
0.75
0.50
0.25
0.00
2
3
4
5
FEV1
GENDER
Female
Male

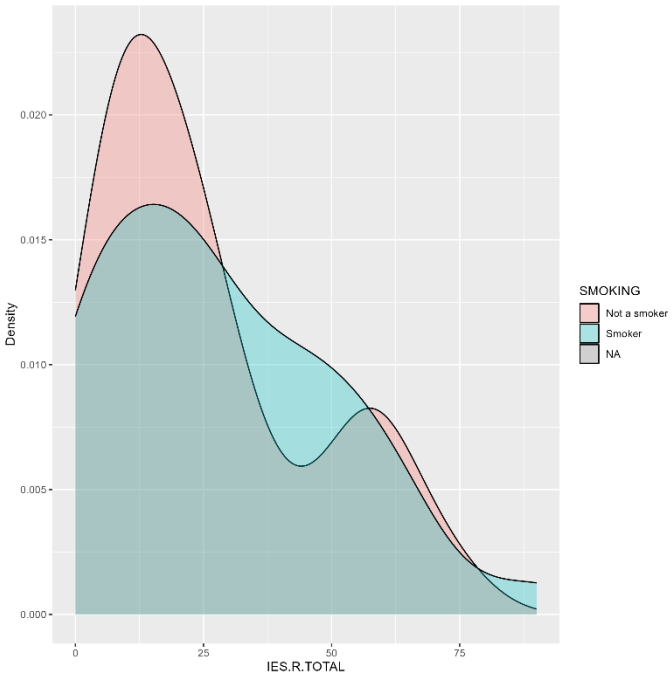

Density
0.025
0.020
0.015
0.010
0.005
0.000
0
25
50
75
IES.R.TOTAL
GENDER
Female
Male

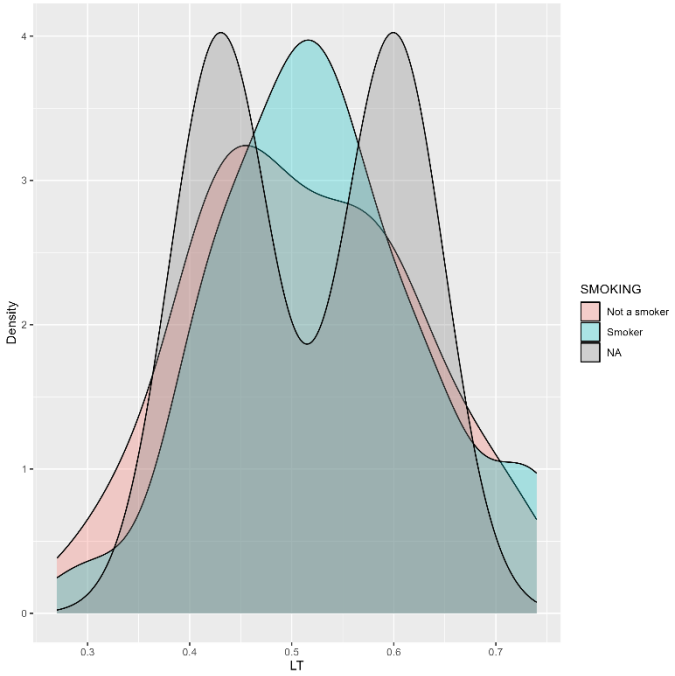

Density
4
3
2
1
0
0.3
0.4
0.5
0.6
0.7
LT
GENDER
Female
Male

SMOKING


Density
0.075
0.050
0.025
0.000
20
30
40
VE.VCO2.SLOPE
SMOKING
Not a smoker
Smoker
NA
Density
0.4
0.2
0.0
10
15
20
25
30
VO2
SMOKING
Not a smoker
Smoker
NA
Density
0.020
0.015
0.010
0.005
0.000
0
1000
2000
DDIMER
SMOKING
Not a smoker
Smoker
NA
Density
0.03
0.02
0.01
0.00
0
500
1000
1500
FERRITIN
SMOKING
Not a smoker
Smoker
NA


0.6
0.4
0.2
0.0
Density
2
3
4
5
FEV1
SMOKING
Not a smoker
Smoker
NA


0.020
0.015
0.010
0.005
0.000
Density
0
25
50
75
IES.R.TOTAL
SMOKING
Not a smoker
Smoker
NA


4
3
2
1
0
Density
0.3
0.4
0.5
0.6
0.7
LT
SMOKING
Not a smoker
Smoker
NA

Summary

In $Table\ 6$, "YES" indicates that the pairwise data for both groups of each covariate (group covariate, response) follow the Gaussian distribution and "NO" otherwise.

For example, the covariate $GENDER$ has two groups (1: Female, 2: Male). The cell (GENDER - Gaussian, VO2) has the label "YES". This means that the pairwise data (group 1 from GENDER, VO2) AND (group 2 from GENDER, VO2) follow the Gaussian distribution. If at least one of them did not, then the label of the cell would be "NO".

Quantitative Response	*GENDER - GAUSSIAN*	*SMOKING - GAUSSIAN*
VE.VCO2.SLOPE	*YES*	*YES*
VO2	*YES*	*YES*
LT	*YES*	*YES*
FEV1	*NO*	*YES*
FERRITIN	*NO*	*NO*
DDIMER	*NO*	*NO*
IES.R.TOTAL	*NO*	*NO*

Table 18: Shapiro-Wilk test results

- if **Gaussian** and
  - **2** groups
    Independent t-test
    $\boldsymbol{p_{value} < 0.05}$ implies that the means of the two groups are different.

GENDER	*p-value*	*SMOKING*	*p-value*
VE.VCO2.SLOPE	*0.126*	**VE.VCO2.SLOPE**	*0.449*
VO2	***<0.001***	**VO2**	*0.321*
LT	***0.009***	**LT**	*0.420*
		FEV1	*0.204*

Table 19: Independent t-test results

- If **not Gaussian** and
  - **2** groups: **Wilcoxon rank-sum** (Mann-Whitney U)
  - Is there a difference between covariate's groups based on response?
  - $p_{value} < 0.05$ implies that the medians of the two groups are different.

GENDER	pvalue	SMOKING	pvalue
FEV1	***<0.01***	**FERRITIN**	*0.645*
FERRITIN	***<0.01***	**DDIMER**	*0.210*
DDIMER	*0.77*	**IES.R.TOTAL**	*0.650*
IES.R.TOTAL	***<0.01***		

Table 20: Wilcoxon rank - sum test results

Summary

In $Table\ 21$ we present the quantitative response variables (rows) that based on them the groups of each covariate differ.

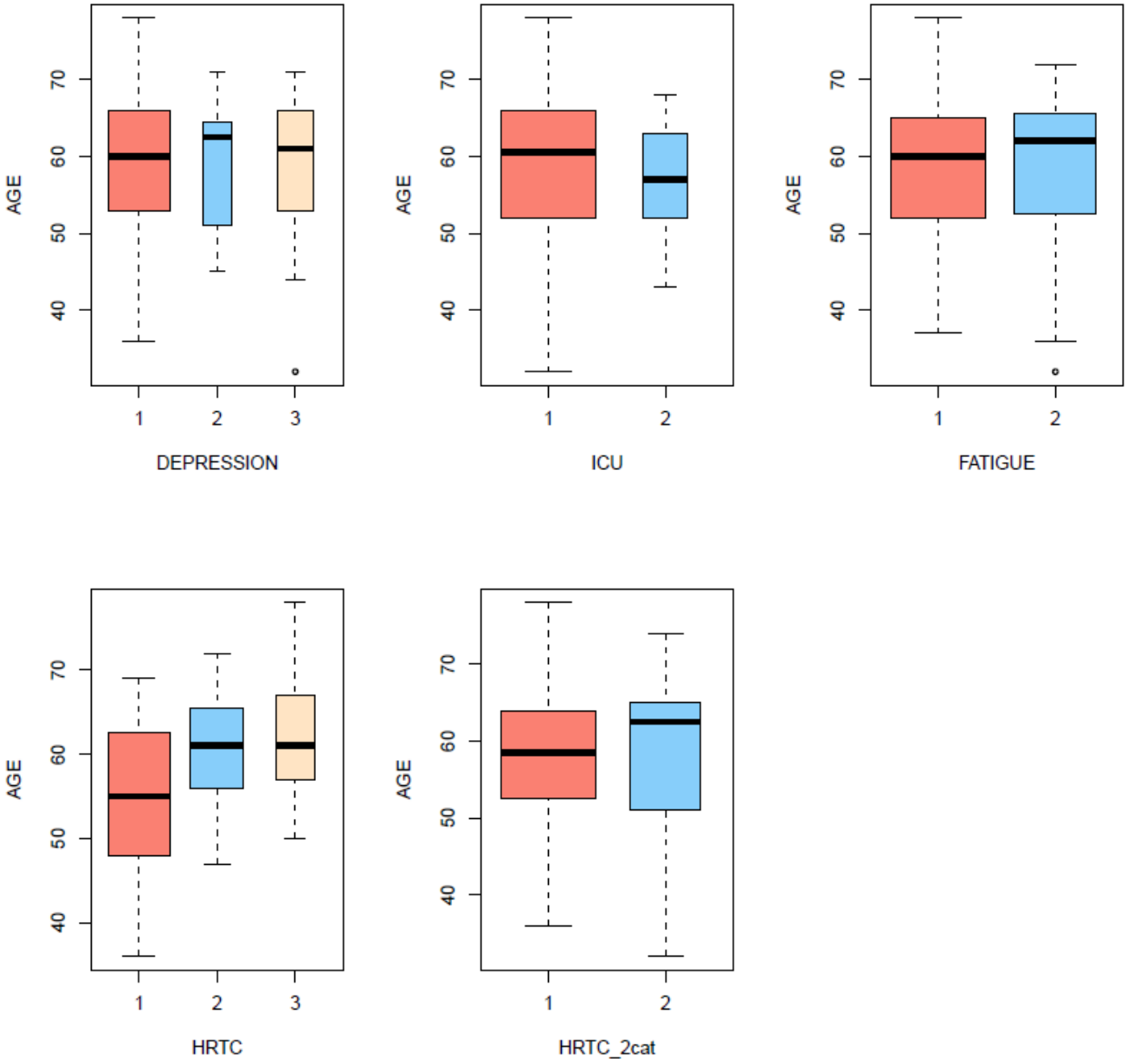
Covariate groups differ based on Response variable	GENDER	SMOKING
	VO2	
	LT	
	FEV1	
	FERRITIN	
	IES.R.TOTAL	

Table 21: Results Qualitative explanatory variables vs Quantitative Response variable.

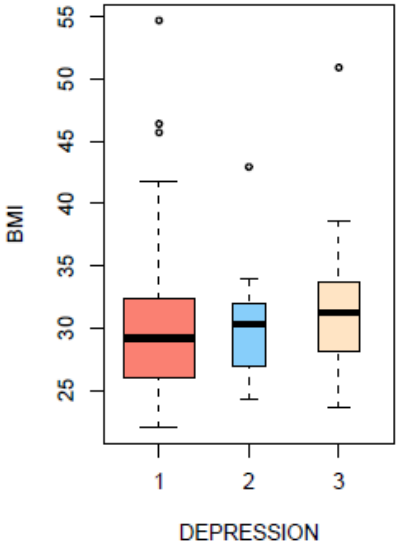
3.2.4.3 Quantitative explanatory variables vs Qualitative Response variable

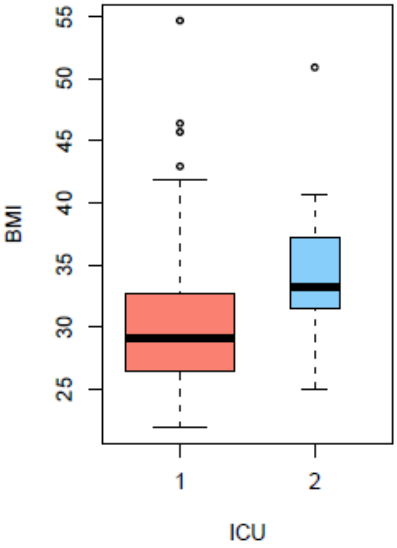
1. **Box plots** (y axis: response var, x axis: covariate)

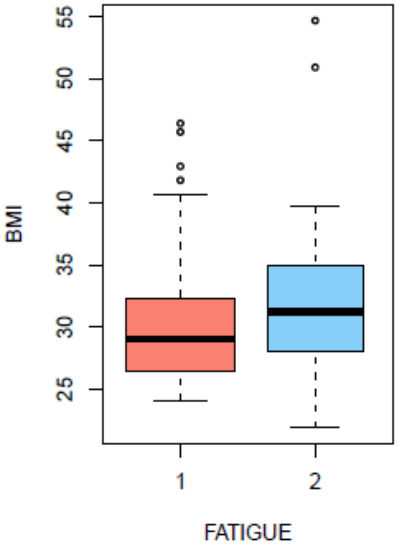
AGE

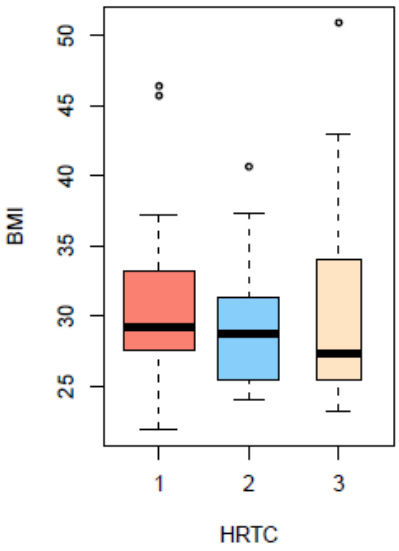


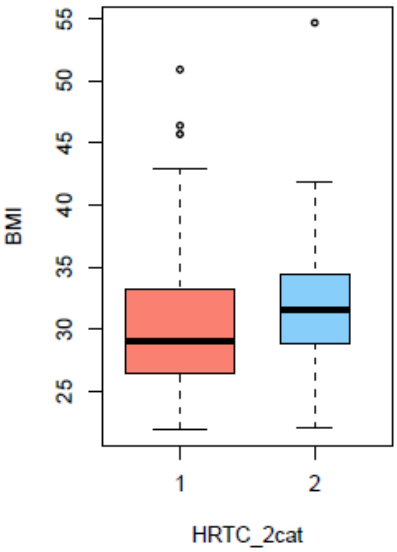
BMI


BMI
25
30
35
40
45
50
55
1
2
3
DEPRESSION

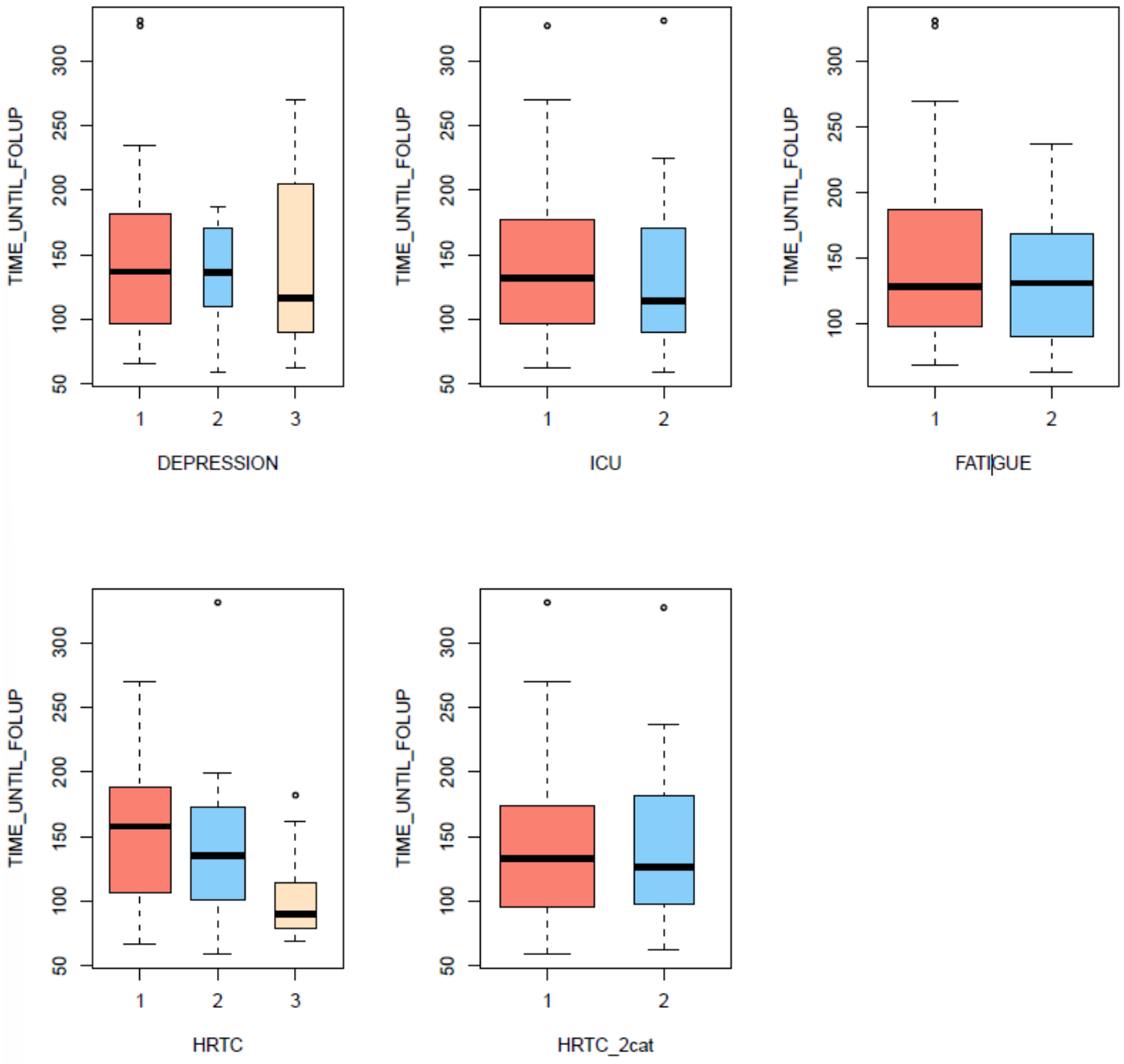

BMI
25
30
35
40
45
50
55
1
2
ICU


BMI
25
30
35
40
45
50
55
1
2
FATIGUE


BMI
25
30
35
40
45
50
1
2
3
HRTC


BMI
25
30
35
40
45
50
55
1
2
HRTC_2cat

TIME_UNTIL_FOLUP



Summary

In $Table\ 22$, we present the qualitative response variables (rows) that seem to affect the groups of each quantitative covariate (columns) based on the Box plots.

AGE	*BMI*	*TIME_UNTIL_FOLUP*
HRTC	ICU	DEPRESSION
	FATIGUE	HRTC
	HRTC	

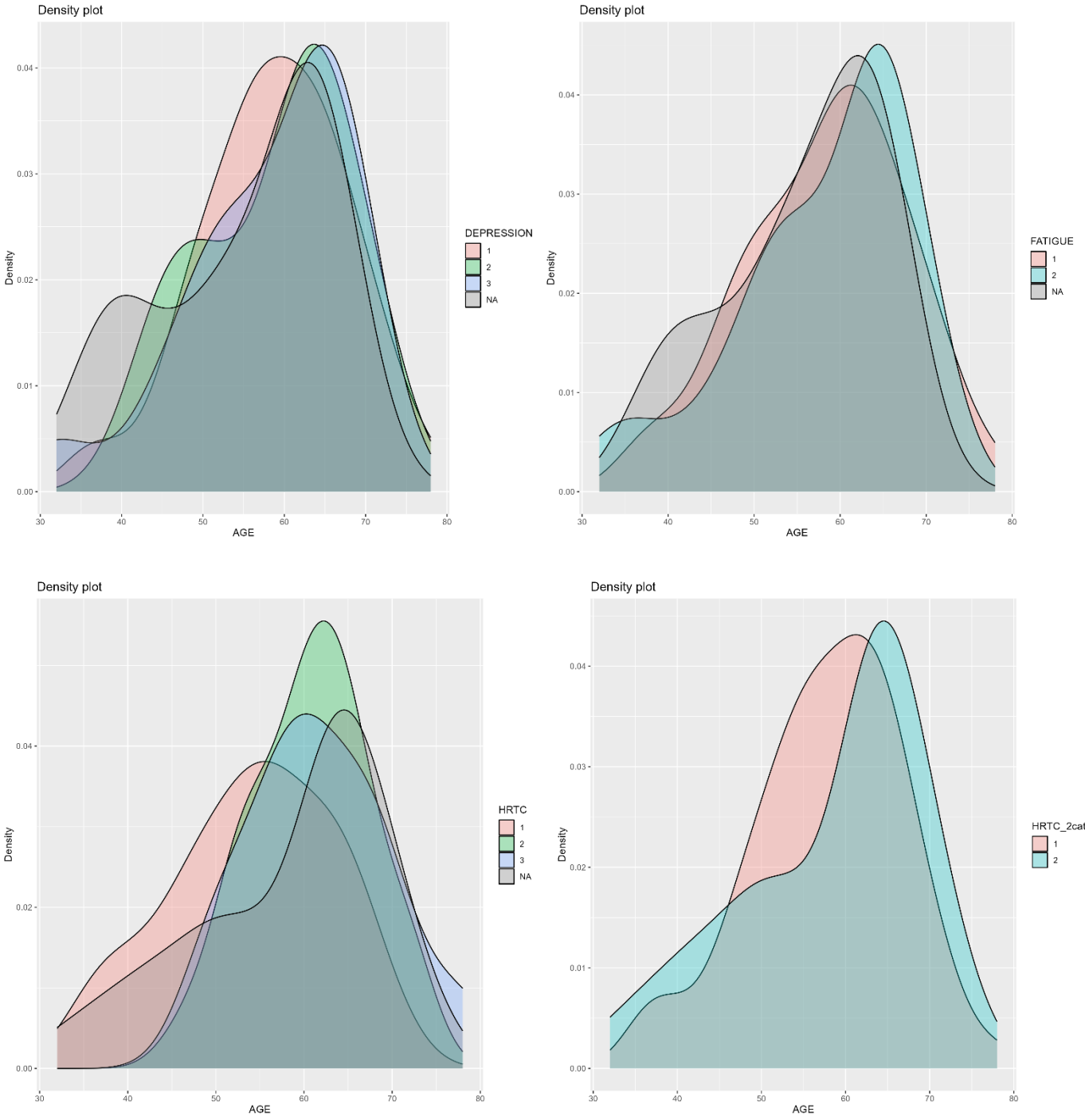
Table 22: Boxplots results

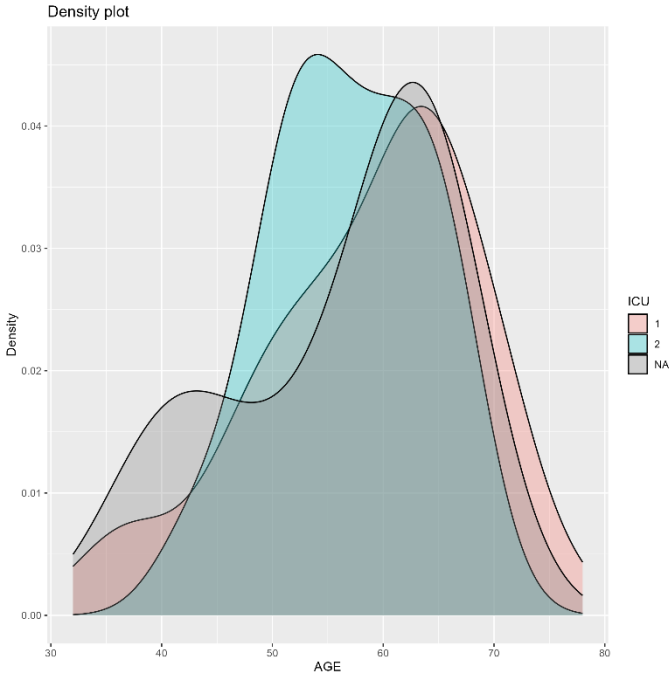
2. Comparison: **Density plots** & **Shapiro-Wilk test**

- Shapiro-Wilk test: $\boldsymbol{p\_value < 0.05}$ implies that the distribution of each group's data is significantly different from the normal distribution.

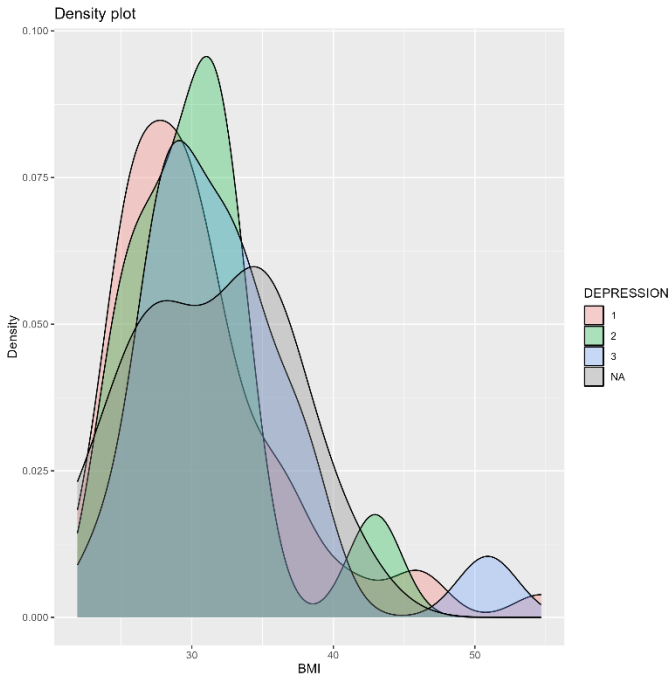
Density Plots

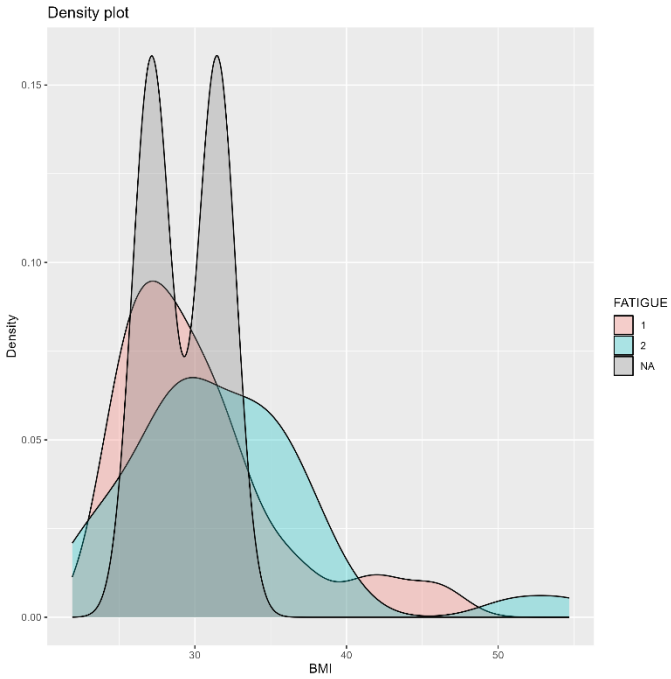
AGE

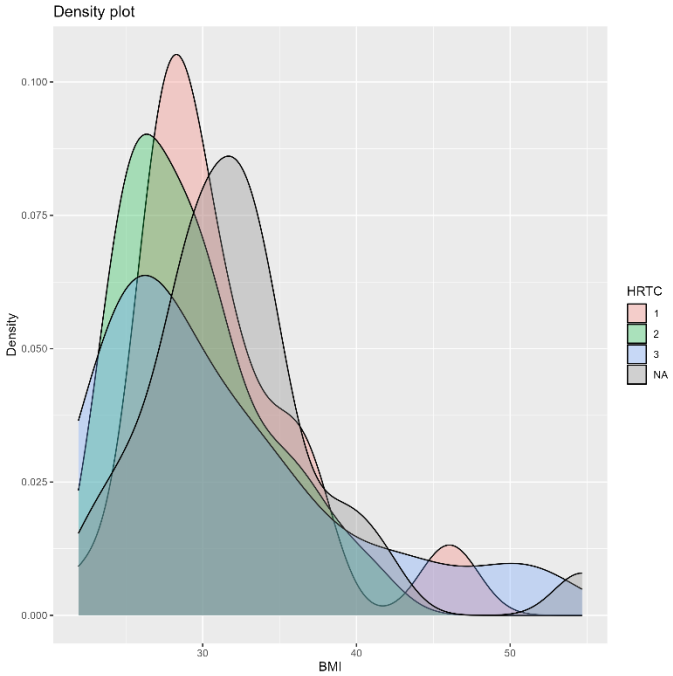


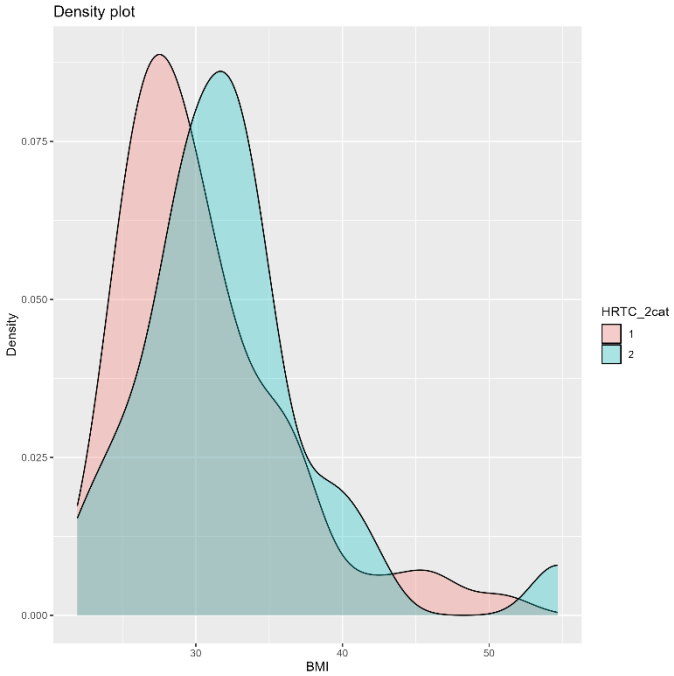

Density plot
Density
AGE
ICU
1
2
NA

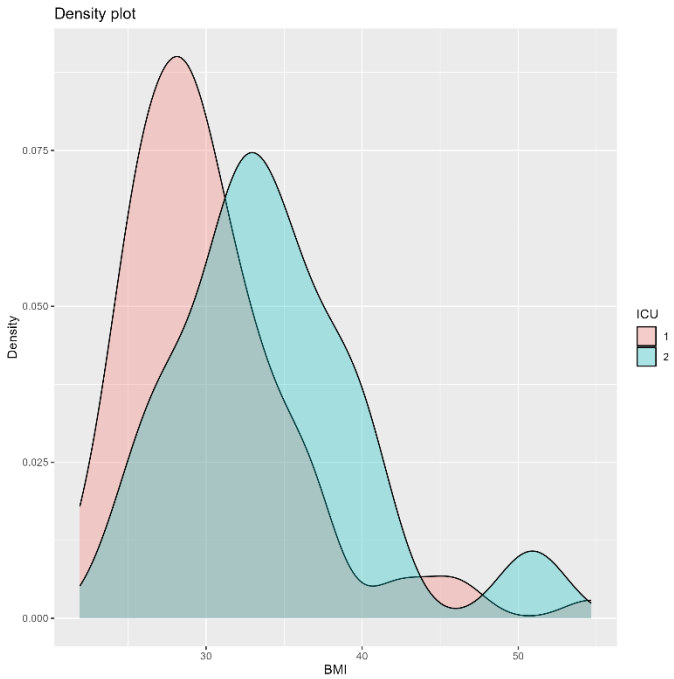
BMI


Density plot
Density
BMI
DEPRESSION
1
2
3
NA

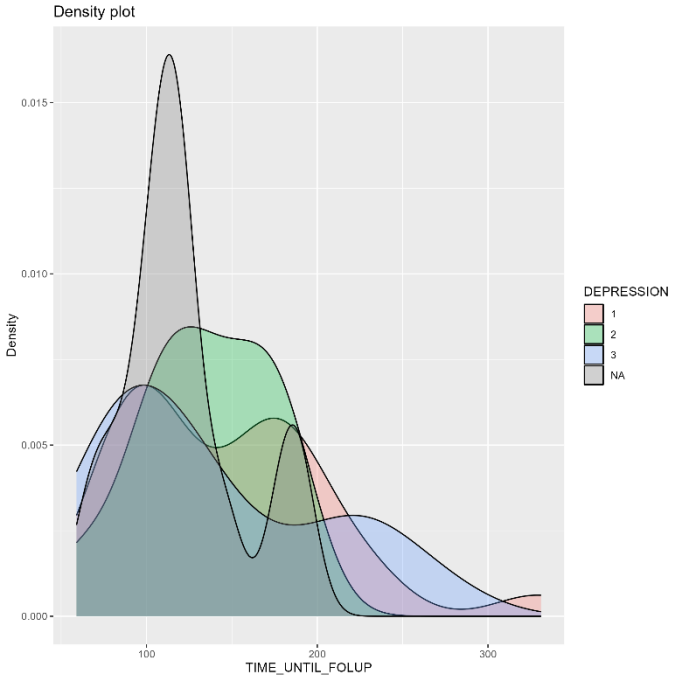

Density plot
Density
BMI
FATIGUE
1
2
NA

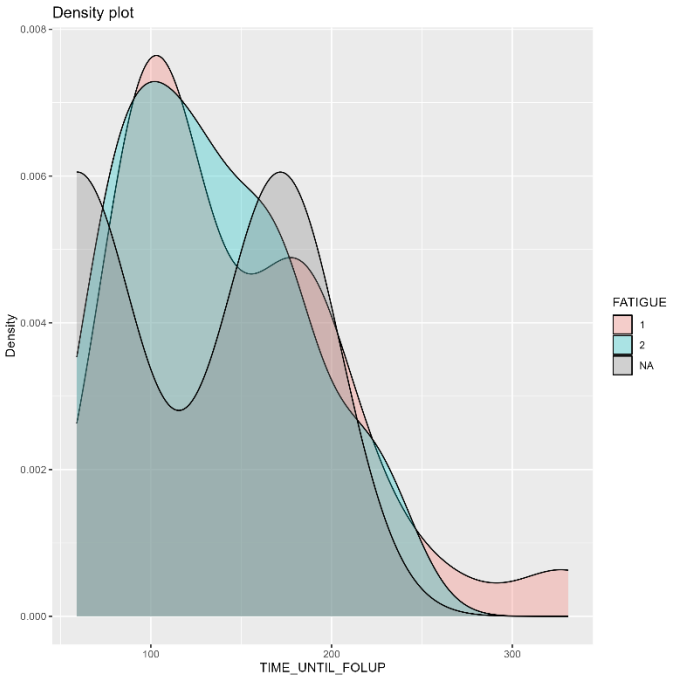

Density plot
0.100
0.075
0.050
0.025
0.000
Density
30
40
50
BMI
HRTC
1
2
3
NA

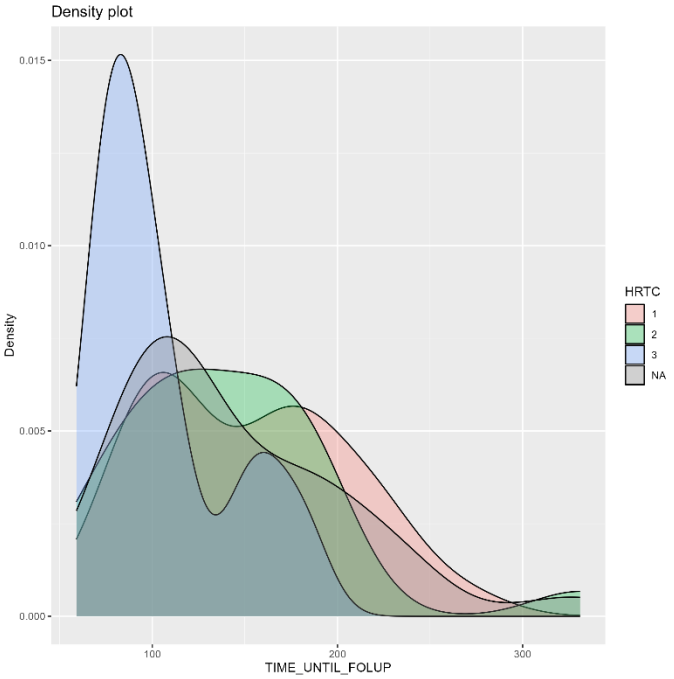

Density plot
0.075
0.050
0.025
0.000
Density
30
40
50
BMI
HRTC_2cat
1
2

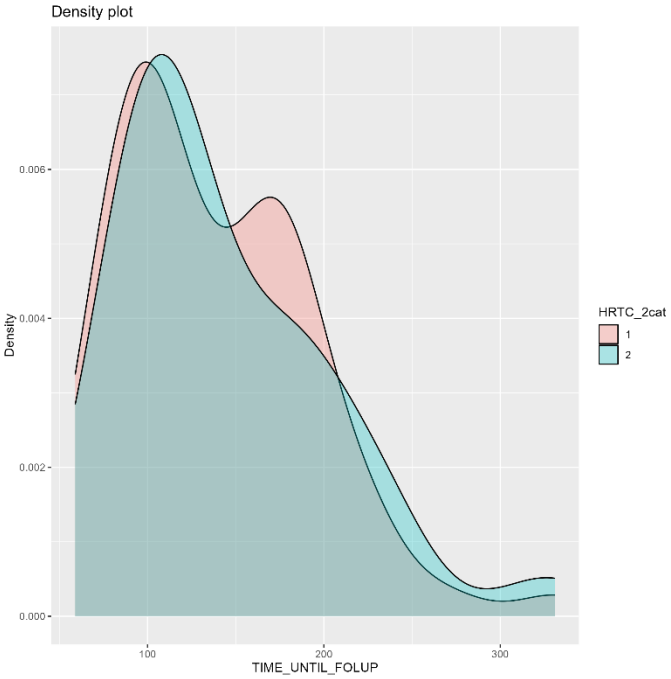

Density plot
0.075
0.050
0.025
0.000
Density
30
40
50
BMI
ICU
1
2

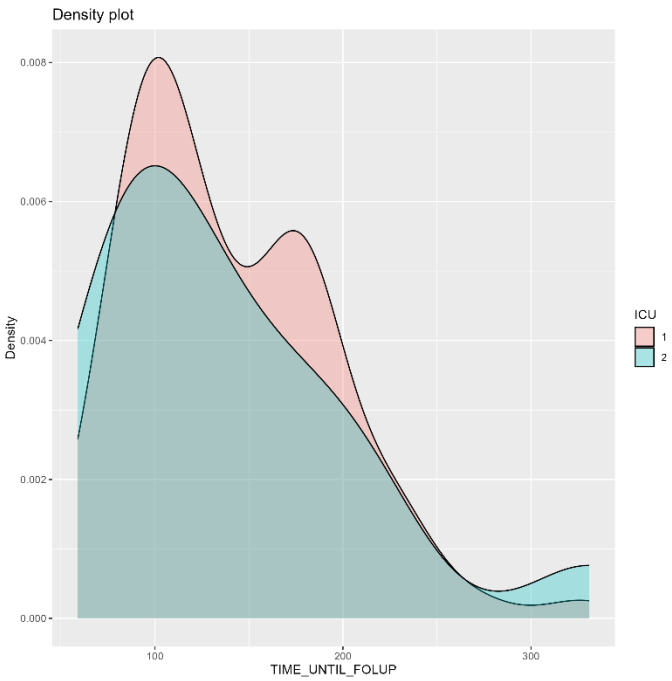
TIME_UNTIL_FOLUP











Summary

In $Table\ 10$, "Yes" indicates that the pairwise data for all the groups of each response variable (covariate, group of response var) follows the Gaussian distribution and "No" otherwise.

For example, the response variable Fatigue has two groups (1: No, 2: Yes). The cell (Age-Gaussian, Fatigue) has the label "No". This means that at least one of the pairwise data (Age, group 1 of Fatigue), (Age, group 2 of Fatigue) did not follow the Gaussian distribution.

Qualitative response	*AGE - Gaussian*	*BMI - Gaussian*	*TIME_UNTIL_FOLUP - Gaussian*
DEPRESSION	*No*	*No*	*No*
ICU	*No*	*No*	*No*
FATIGUE	*No*	*No*	*No*
HRTC	*Yes*	*No*	*No*
HRTC_2cat	*No*	*No*	*No*

Table 23: Shapiro-Wilk test results

- if **Gaussian** and
  - **3+** groups: **Anova**: $p_{value} < 0.05$ implies that the means of the groups are different.

AGE	p-value
HRTC	***0.003***

Table 24: Independent t-test results

- If **not Gaussian** and
  - **2** groups: **Wilcoxon rank-sum** (Mann-Whitney U)
  - Is there a difference between covariate's groups based on response?
  - $p_{value} < 0.05$ implies that the medians of the two groups are different.

AGE	p-value	BMI	p-value	TIME_UNTIL_FOLUP	p-value
FATIGUE	*0.783*	**FATIGUE**	0.112	**FATIGUE**	*0.363*
ICU	*0.356*	**ICU**	***0.006***	**ICU**	*0.634*
HRTC_2cat	*0.438*	**HRTC_2cat**	*0.085*	**HRTC_2cat**	*0.671*

Table 25: Wilcoxon rank-sum results

  - **3+** groups: **Kruskal-Wallis test**
  - Is there a difference between response groups based on the covariate?
  - If $p_{value} < 0.05$, we reject the hypothesis H_0: the mean ranks of the groups are the same.

AGE	p-value	BMI	p-value	TIME_UNTIL_FOLUP	p-value
DEPRESSION	*0.984*	**DEPRESSION**	*0.330*	**DEPRESSION**	*0.763*
		HRTC	*0.338*	**HRTC**	***0.015***

Table 26: Kruskal-Wallis test results

Summary

In $Table\ 27$ we present the qualitative response variables that their groups appear to have differences based on each covariate.

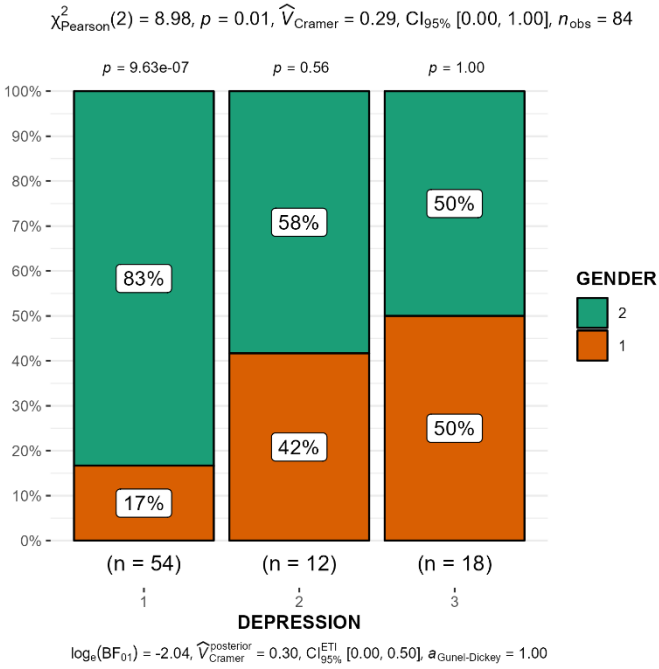
Response variable groups differ based on Covariate	AGE	BMI	TIME_UNTIL_FOLUP
	HRTC	HRTC	HRTC
		ICU	

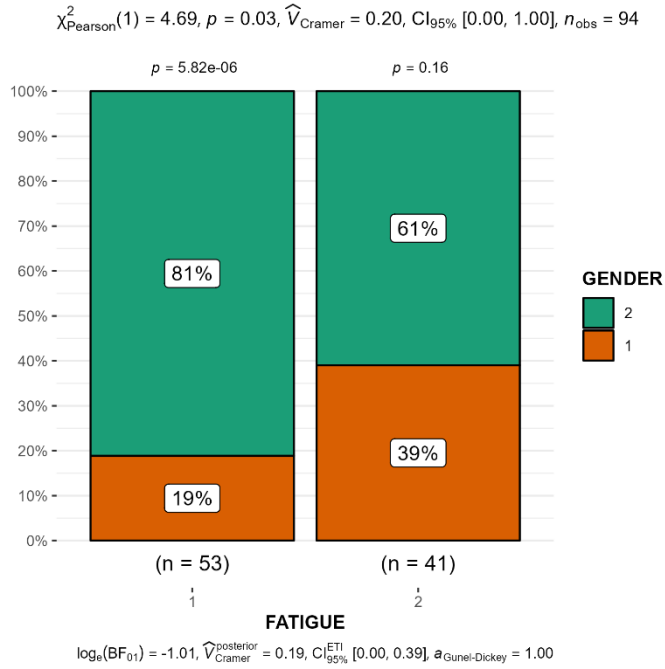
Table 27: Results Quantitative explanatory variables vs Qualitative Response variable.

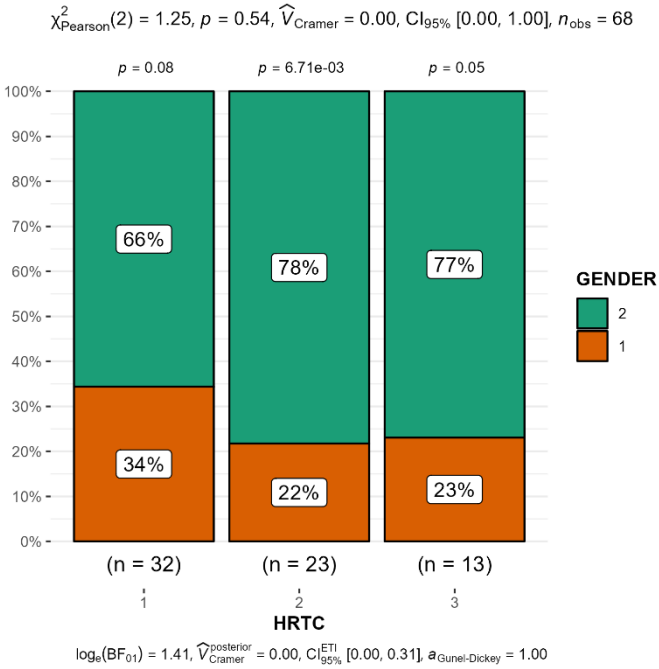
3.2.4.4 Qualitative explanatory variables vs Qualitative Response variable

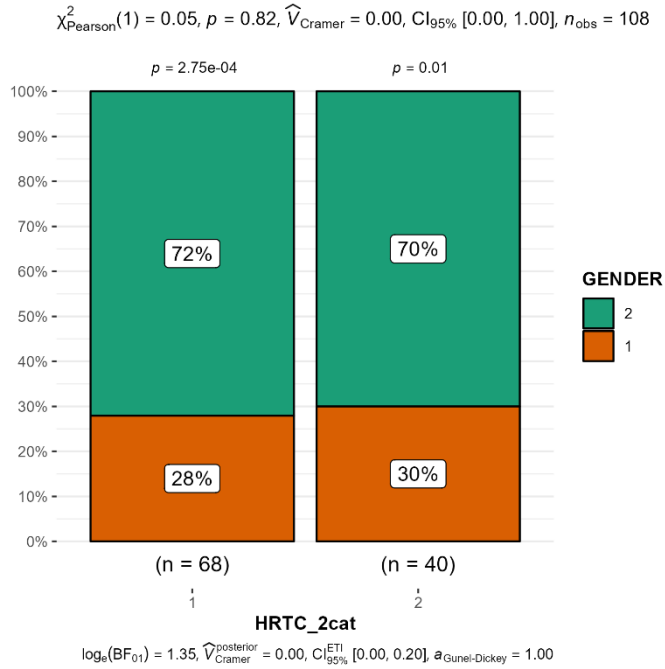
1. Bar plots & Chi-square test

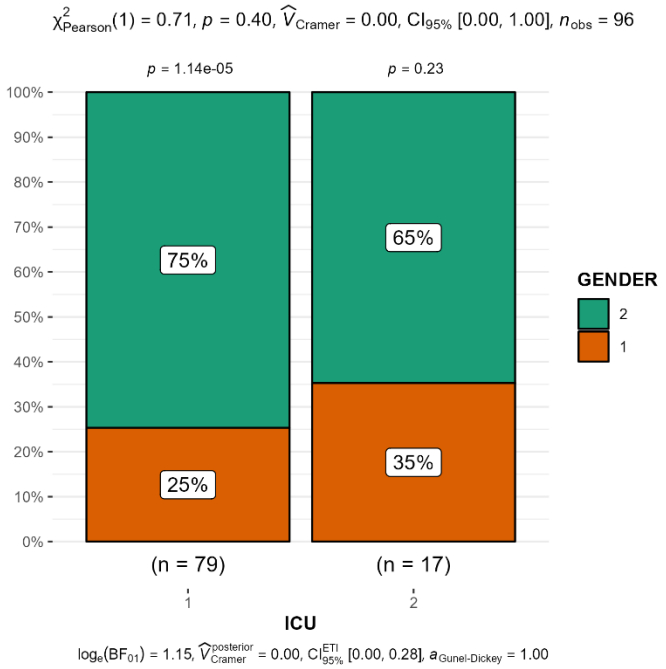
GENDER



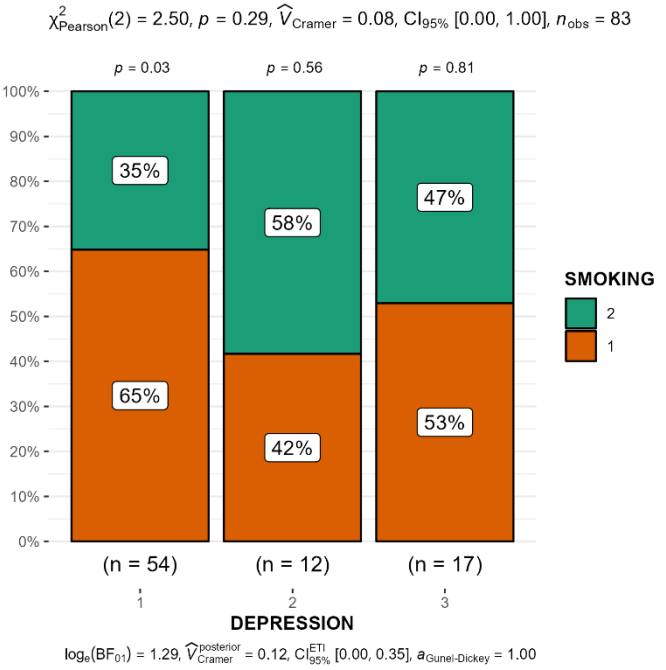


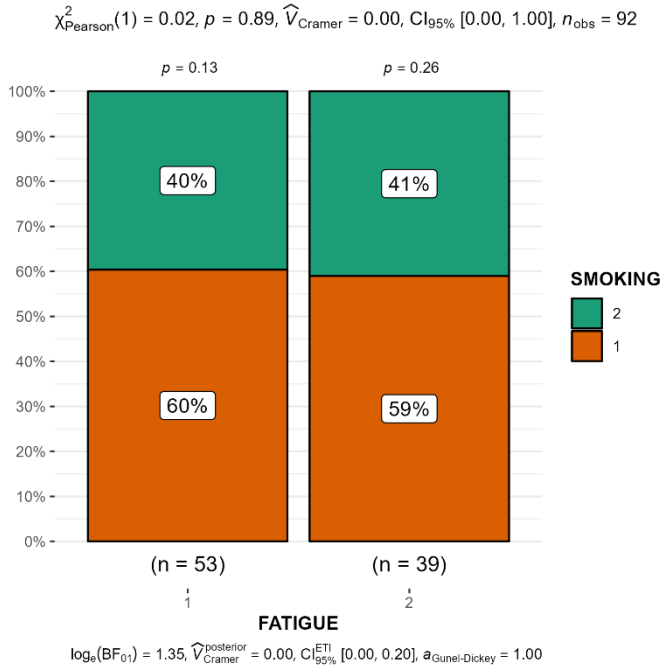


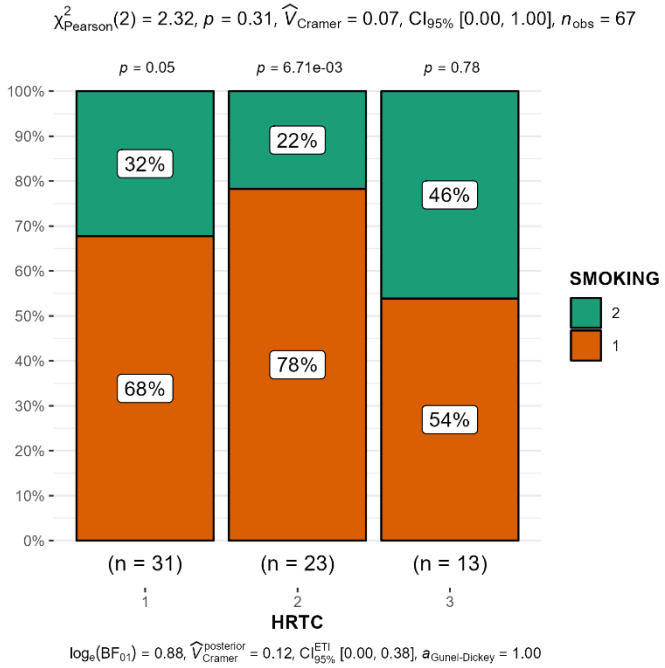


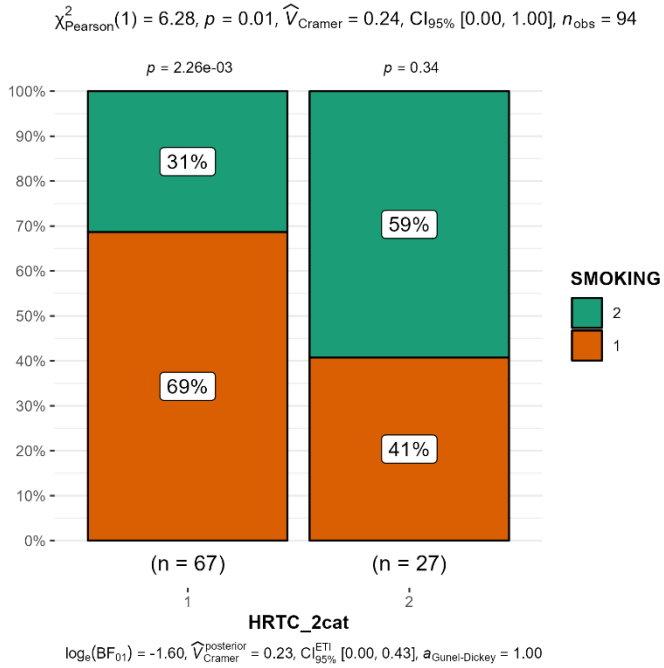

$\chi^2_{Pearson}(1) = 0.71$, $p = 0.40$, $\widehat{V}_{Cramer} = 0.00$, $CI_{95\%}$ [0.00, 1.00], $n_{obs} = 96$
p = 1.14e-05
p = 0.23
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
75%
25%
65%
35%
(n = 79)
(n = 17)
1
2
ICU
GENDER
2
1
$\log_e(BF_{01}) = 1.15$, $\widehat{V}^{posterior}_{Cramer} = 0.00$, $CI^{ETI}_{95\%}$ [0.00, 0.28], $a_{Gunel\text{-}Dickey} = 1.00$

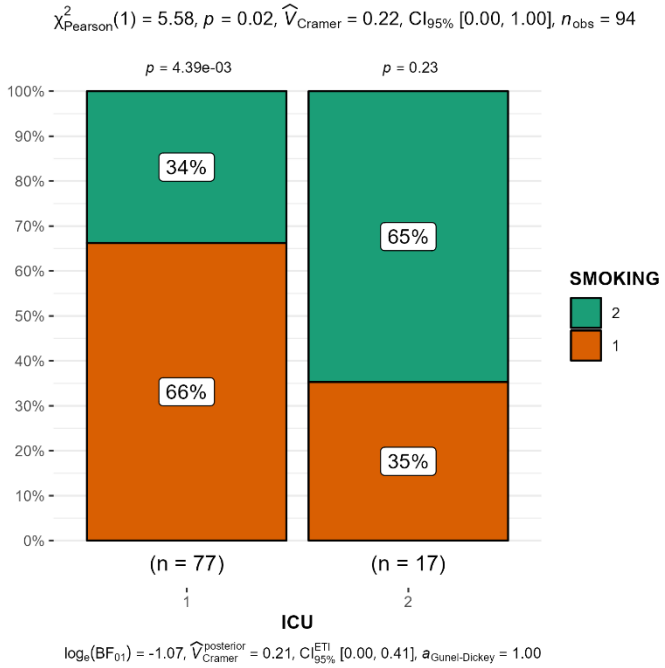
SMOKING


$\chi^2_{Pearson}(2) = 2.50$, $p = 0.29$, $\widehat{V}_{Cramer} = 0.08$, $CI_{95\%}$ [0.00, 1.00], $n_{obs} = 83$
p = 0.03
p = 0.56
p = 0.81
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
35%
65%
58%
42%
47%
53%
(n = 54)
(n = 12)
(n = 17)
1
2
3
DEPRESSION
SMOKING
2
1
$\log_e(BF_{01}) = 1.29$, $\widehat{V}^{posterior}_{Cramer} = 0.12$, $CI^{ETI}_{95\%}$ [0.00, 0.35], $a_{Gunel\text{-}Dickey} = 1.00$


$\chi^2_{Pearson}(1) = 0.02$, $p = 0.89$, $\widehat{V}_{Cramer} = 0.00$, $CI_{95\%}$ [0.00, 1.00], $n_{obs} = 92$
p = 0.13
p = 0.26
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
40%
60%
41%
59%
(n = 53)
(n = 39)
1
2
FATIGUE
SMOKING
2
1
$\log_e(BF_{01}) = 1.35$, $\widehat{V}^{posterior}_{Cramer} = 0.00$, $CI^{ETI}_{95\%}$ [0.00, 0.20], $a_{Gunel\text{-}Dickey} = 1.00$







Plot explanation

Let's consider the last plot as an example. The bar plot ICU vs $SMOKING$ show us that the patients who are smokers ($SMOKING = 2$) have higher likelihood of entering the intensive care unit ($ICU = 2$), while the patients who do not smoke ($SMOKING = 1$) experienced a decreased likelihood of being admitted to the intensive care unit ($ICU = 1$).

2. Correlation: **Chi square test of independency** (non-parametric)
   - How are response var and covariate related?
   - If $\boldsymbol{p_{value} < 0.05}$, then we reject the H_0: the variables are independent and there is no relationship between the two qualitative variables.
   - If two variables are related, the probability of one variable having a certain value is dependent on the value of the other variable.

In $Table\ 28$ we present the results from the **Chi-square test**. Each column corresponds to a qualitative explanatory variable and each row to a qualitative response variable. The bold response variables are those that appear to have a relationship of dependency with the corresponding covariate.

GENDER	*pvalue*	*SMOKING*	*pvalue*
DEPRESSION	***0.013***	**DEPRESSION**	***0.018***
ICU	***0.041***	**ICU**	***0.027***
FATIGUE	*0.557*	FATIGUE	*0.280*
HRTC	*0.574*	HRTC	*0.338*
HRTC_2cat	*0.836*	HRTC_2cat	*1.000*

Table 28: Chi-square test results

Summary

From the Bar-plots and $Table\ 28$ we draw the following conclusions.

i. the value for the qualitative response variable $DEPRESSION$ depends on the value of the qualitative covariate $GENDER$ and $SMOKING$,
ii. the value for the qualitative response variable ICU depends on the value of the qualitative covariate $GENDER$ and $SMOKING$.

3.3 Undirected Gaussian Graphical Model

We begin by importing into the R environment the dataset from $Table\ 4$. This is done with the help of the function $read.xlsx()$. We continue by applying the procedure explained in subchapter $3.2.3$ to deal with the missing values.

The analysis starts by studying the relationships among the continuous variables of interest given in $Table\ 4$ with an undirected Gaussian graphical model (UGGM). We used the function $cmod()$, where c stands for continuous variables in the model. This function is derived from the R package $gRim$.

Next, we perform model selection, which is based on the BIC criterion, using the function $stepwise()$ and its argument for the penalty parameter $k = log(nrow())$. In this case, we prefer BIC over AIC, as is favors a simpler model and consequently makes easier both the interpretation and representation of the result. $Table\ 29$ shows the summary of the saturated model, while $Table\ 30$ presents the summary of the final model after utilizing BIC. As it is shown, the saturated model has 45 possible edges in total and accounting these only 18 are present in the BIC's final output model. In addition, by comparing the two models we observe that both values of AIC and BIC are decreased.

Saturated Model: 10 continuous variables					
-2logL	6919.73	mdim	55	**AIC**	**7029.73**
ideviance	180.85	**idf**	**45**	**BIC**	**7170.77**
deviance	-0.00	df	0		

Table 29: UGGM - Saturated Model's summary

Final Model: 10 continuous variables					
-2logL	6941.31	mdim	28	**AIC**	**6997.31**
ideviance	159.28	**idf**	**18**	**BIC**	**7069.11**
deviance	21.58	df	27		

Table 30: UGGM - Final Model's summary

The final step of this analysis is to examine the relationships between the variables. To achieve this, we calculate the partial correlation matrix through the covariance matrix with the assistance of the function $cov2pcor()$. The output is given by $Table\ 31$. The signs upon the edges in $Figure\ 10$ are derived from the following partial correlation matrix.

Variable	AGE	BMI	TIME_ UNTL_ FOLUP	DDIMER	FERRITIN	FEV1	VO2	VE. VCO2. SLOPE	LT	IES. R. TOTAL
AGE	1.00	-0.30	-0.01	0.33	-0.05	-0.30	-0.34	0.02	0.36	-0.22
BMI	-0.30	1.00	0.07	0.17	-0.01	0.08	-0.53	-0.09	0.23	0.10
TIME_ UNTIL_ FOLUP	-0.01	0.07	1.00	-0.10	0.05	0.08	0.09	-0.03	0.16	-0.09
DDIMER	0.33	0.17	-0.10	1.00	0.11	0.09	0.06	-0.16	-0.07	0.01
FERRITIN	-0.05	-0.01	0.05	0.11	1.00	0.02	0.00	0.24	-0.02	-0.18
FEV1	-0.30	0.08	0.08	0.09	0.02	1.00	0.33	-0.14	-0.27	-0.24
VO2	-0.34	-0.53	0.09	0.06	0.00	0.33	1.00	-0.12	0.41	0.05
VE. VCO2. SLOPE	0.02	-0.09	-0.03	-0.16	0.24	-0.14	-0.12	1.00	-0.18	0.23
LT	0.36	0.23	0.16	-0.07	-0.02	-0.27	0.41	-0.18	1.00	0.26
IES. R. TOTAL	-0.22	0.10	-0.09	0.01	-0.18	-0.24	0.05	0.23	0.26	1.00

Table 31: UGGM - Partial Correlation matrix

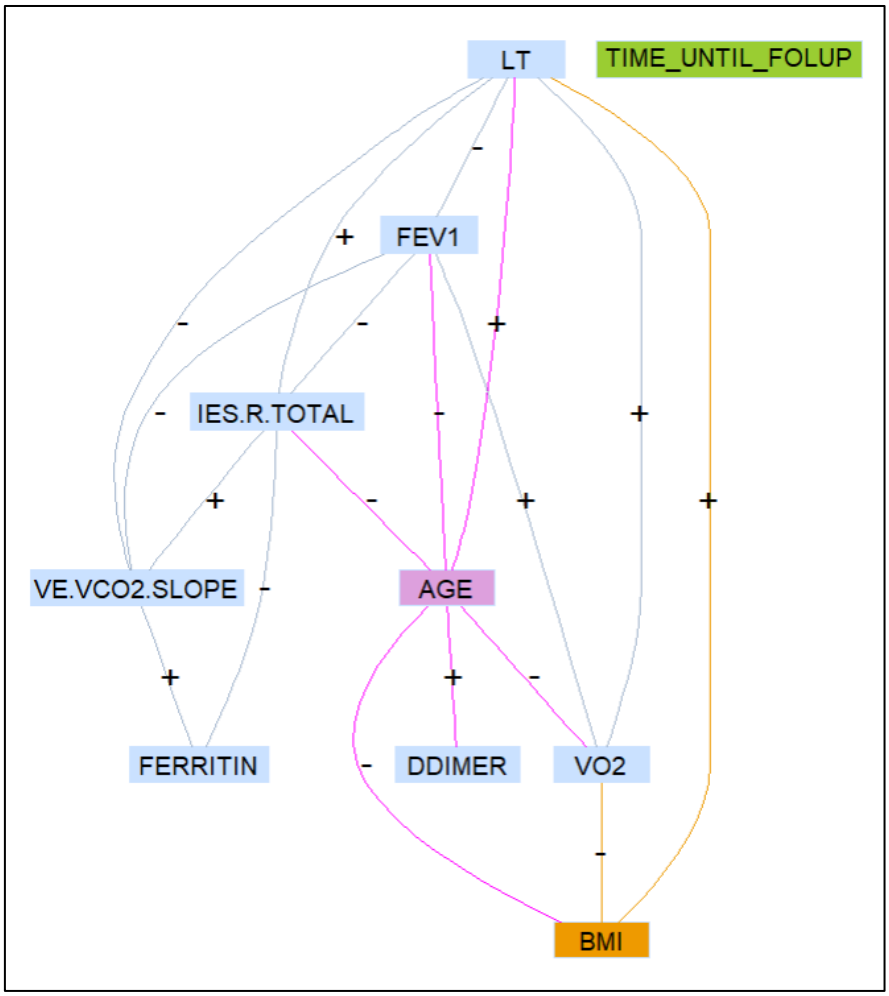


Figure 10: Undirected Gaussian Graphical model using the dataset from Table 4.

The following table explains $Figure\ 10$.

Variables connected by an edge	Sign	Interpretation
AGE ~ LT	+	As age increases, the lactate threshold also increases.
AGE ~ FEV1	-	As age increases, the maximum volume of air exhaled during the first second after a maximum inhalation decreases.
AGE ~ IES.R.TOTAL	-	Older patients tend to be more depressed after being hospitalized due to Covid-19.
AGE ~ BMI	-	As age increases, the body mass index decreases.
AGE ~ DDIMER	+	As age increases, the DDimer value also increases.
AGE ~ VO2	-	As age increases, the maximal oxygen consumption decreases.
BMI ~ VO2	-	As the body mass index increases, the maximal oxygen consumption decreases.
BMI ~ LT	+	As the body mass index increases, the lactate threshold also increases.
LT ~ FEV1	-	As the lactate threshold increases, the maximum volume of air exhaled during the first second after a maximum inhalation decreases.
LT ~ VE.VCO2.SLOPE	-	As the lactate threshold increases, the ratio of minute ventilation to carbon dioxide decreases.
LT ~ IES.R.TOTAL	+	Patients with an increased lactate threshold value, tend to be more depressed.
LT ~ VO2	+	As the lactate threshold increases, the maximal oxygen consumption also increases.
FEV1 ~ IES.R.TOTAL	-	Patients with a higher maximum volume of air exhaled during the first second after a maximum inhalation tend to be less depressed.
FEV1 ~ VE.VCO2.SLOPE	-	Patients with a higher maximum volume of air exhaled during the first second after a maximum inhalation tend to have a decrease in the ratio of minute ventilation to carbon dioxide.
FEV1 ~ VO2	+	Patients with a higher maximum volume of air exhaled during the first second after a maximum inhalation tend to have a higher maximal oxygen consumption.
IES.R.TOTAL ~ VE.VCO2.SLOPE	+	Patients with higher ratio of minute ventilation to carbon dioxide tend to be more depressed.
IES.R.TOTAL ~ FERRITIN	-	Patients with a higher ferritin value tend to be less depressed.
VE.VCO2.SLOPE ~ FERRITIN	+	Patients with higher ratio of minute ventilation to carbon dioxide tend to have an increase to the ferritin value.

Table 32: UGGM interpretation

3.4 Mixed Interaction Model

We begin by importing into R environment the dataset from $Table\ 7$. This is done with the help of the function $read.xlsx()$. We continue by applying the procedure explained in subchapter $(3.2.3)$ to deal with the missing values.

A mixed interaction model (MIM) is applied to analyze the data from $Table\ 7$, using the function $mmod()$, where m stands for mixed variables (continuous and discrete) in the model. This function is derived from the R package $gRim$.

Next, we perform model selection, which is based on the AIC criterion, using the function $stepwise()$ and its argument for the penalty parameter $k = 2$. AIC results in the most appropriate model for the given data. $Table\ 33$ shows the summary of the saturated model, while $Table\ 34$ presents the summary of the final model after utilizing AIC. As is shown, the saturated model has 76 possible edges in total and accounting these only 19 are present in the AIC's final output model. In addition, by comparing the two models we observe that both values of AIC and BIC are decreased.

Saturated Model: 10 continuous variables					
-2logL	4750.51	mdim	98	**AIC**	**4946.51**
ideviance	170.14	**idf**	**76**	**BIC**	**5197.82**
deviance	-0.00	df	0		

Table 33: MIM - Saturated Model's summary

Final Model: 10 continuous variables					
-2logL	4828.59	mdim	41	**AIC**	**4910.59**
ideviance	131.11	**idf**	**19**	**BIC**	**5015.72**
deviance	39.04	df	57		

Table 34: MIM - Final Model's summary

We continue the analysis by studying the relationships between the variables using the moment parameters (p_i, μ_i, Σ) displayed in chapter $(1.4.1)$. Firstly, we investigate the associations between the numeric variables using the partial correlation matrix derived from the covariance matrix Σ ($Table\ 35$). Secondly, we examine the relationships between the numeric and binary variables, by calculating the difference between the means of the groups ($Table\ 36$). Finally, to study the interaction effect of the two binary variables $Gender$ and $Smoking$ on each numeric variable we compare the difference in means between groups when one of the two binary variables remains stable while the other one changes its level ($Table\ 36$). The signs upon each edge were assigned according to the above.

Variable	AGE	BMI	TIME_ UNTL_ FOLUP	VO2	FEV1	IES. R. TOTAL	VE. VCO2. SLOPE	LT	DDIMER	FERRITIN
AGE	1.00	-0.25	0.00	-0.47	-0.42	0.00	0.00	0.44	0.32	0.00
BMI	-0.25	1.00	0.00	-0.41	0.00	0.00	0.00	0.22	0.00	0.00
TIME_ UNTIL_ FOLUP	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VO2	-0.47	-0.41	0.00	1.00	0.00	0.00	-0.14	0.59	0.00	0.00
FEV1	-0.42	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
IES. R. TOTAL	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
VE. VCO2. SLOPE	0.00	0.00	0.00	-0.14	0.00	0.00	1.00	0.00	0.00	0.00
LT	0.44	0.22	0.00	0.59	0.00	0.00	0.00	1.00	0.00	0.00
DDIMER	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
FERRITIN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Table 35: MIM - Partial Correlation matrix

Variable	G12S1[1]	G12S2[2]	G1S12[3]	G2S12[4]	inter_ effect[5]
AGE	-2.09	-1.85	3.53	3.77	0.24
BMI	-4.34	-4.35	-0.21	-0.22	-0.01
TIME_ UNTIL_ FOLUP	0.00	0.00	0.00	0.00	0.00
VO2	5.42	5.38	-0.52	-0.56	-0.04
FEV1	0.94	0.93	-0.12	-0.13	-0.01
IES. R. TOTAL	-18.51	-18.51	0.00	0.00	0.00
VE. VCO2. SLOPE	-1.29	-1.29	0.13	0.13	0.00
LT	-0.07	-0.07	0.01	0.01	0.00
DDIMER	-0.06	-0.05	-0.17	-0.16	0.01
FERRITIN	1.00	1.00	0.00	0.00	0.00

Table 36: MIM - mean differences & interaction effect

$Table\ 36$ columns' explanation:

[1] The difference in outcome variable between individuals with Gender=2 (Male) and Gender=1 (Female), holding Smoking constant at its 1 level.

[2] The difference in outcome variable between individuals with Gender=2 (Male) and Gender=1 (Female), holding Smoking constant at its 2 level.

[3] The difference in outcome variable between individuals with Smoking=2 (Smoker) and Smoking=1 (Nonsmoker), holding Gender constant at its 1 level.

[4] The difference in outcome variable between individuals with Smoking=2 (Smoker) and Smoking=1 (Nonsmoker), holding Gender constant at its 2 level.

[5] The interaction effect of Gender and Smoking on the numeric variable.

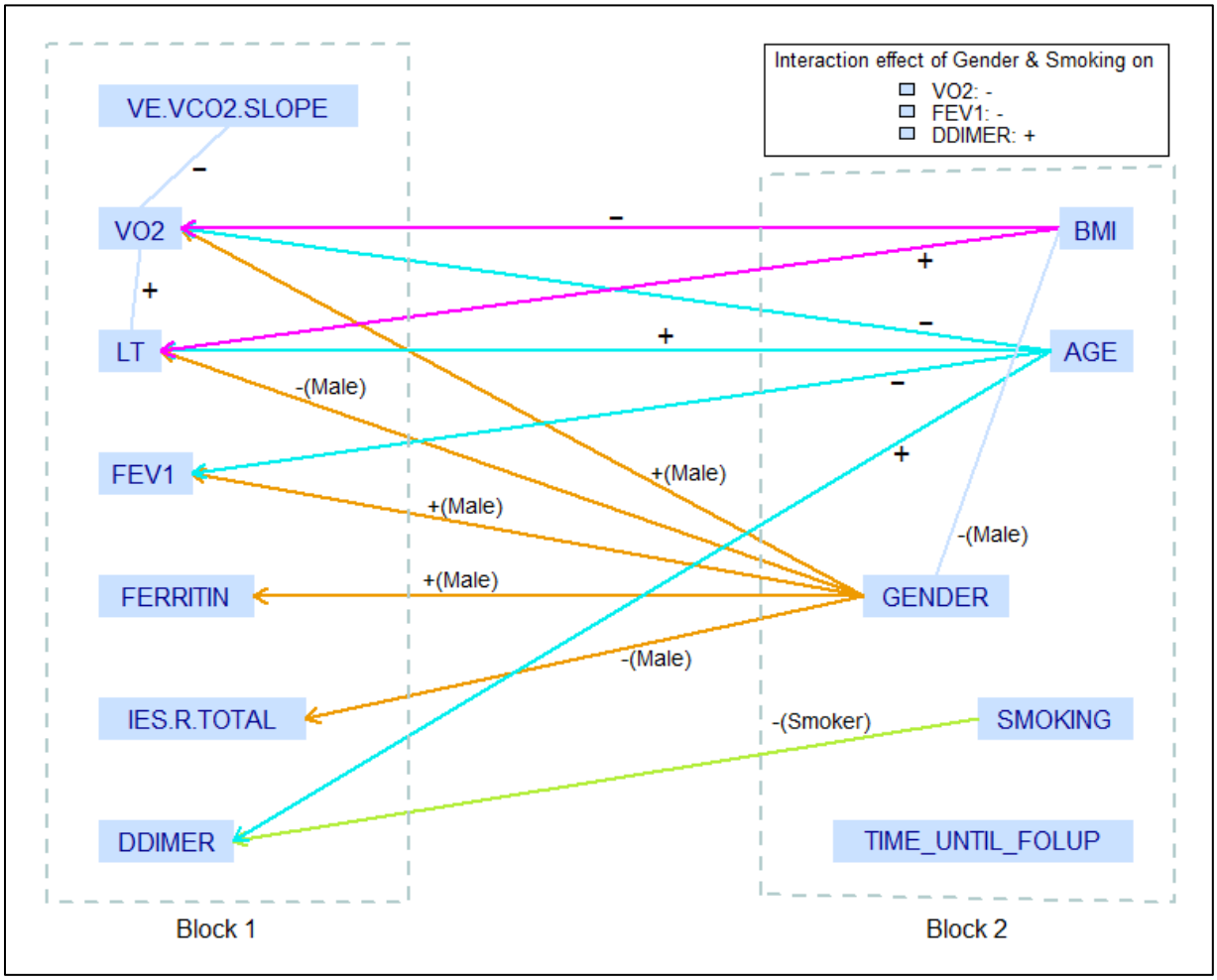


Figure 11: Mixed Interaction Model using the dataset from Table 7.

The mixed interaction model presented in $Figure\ 11$ consists of two blocks. The first block includes the response variables, while the second block includes the explanatory variables.

The following table explains $Figure\ 11$.

Variables connected by an edge	Sign	Interpretation
AGE ~ LT	+	As age increases, the lactate threshold also increases.
AGE ~ FEV1	-	As age increases, the maximum volume of air exhaled during the first second after a maximum inhalation decreases.
AGE ~ DDIMER	+	As age increases, the DDimer value also increases.
AGE ~ VO2	-	As age increases, the maximal oxygen consumption decreases.
BMI ~ VO2	-	As the body mass index increases, the maximal oxygen consumption decreases.
BMI ~ LT	+	As the body mass index increases, the lactate threshold also increases.
BMI ~ GENDER	-(Male)	Males tend to have lower body mass index compared to women.
VO2 ~ GENDER	+(Male)	Males tend to have higher maximal oxygen consumption compared to women.
GENDER ~ IES.R.TOTAL	-(Male)	Males tend to be less depressed than women.
GENDER ~ FERRITIN	+(Male)	Males tend to have higher level of ferritin than women.
GENDER ~ FEV1	+(Male)	Males tend to have higher maximum volume of air exhaled during the first second after a maximum inhalation compared to women.
LT ~ GENDER	-(Male)	Males tend to have lower lactate threshold compared to women.
LT ~ VO2	+	As the lactate threshold increases, the maximal oxygen consumption also increases.
DDIMMER ~ SMOKING	-(Smoker)	Smokers tend to have a decrease to the DDimer value.
VE.VCO2.SLOPE ~ VO2	-	Patients with higher ratio of minute ventilation to carbon dioxide tend to have a decrease to the maximal oxygen consumption.

Table 37: MIM interpretation

3.5 Chain Graph Model

We begin by importing into the R environment the dataset from $Table\ 10$. This is done with the help of the function $read.xlsx()$. We continue by applying the procedure explained in subchapter $(3.2.3)$ to deal with the missing values.

A chain graph model (CGM) is applied to analyze the data from $Table\ 10$, using the function $coxwer()$ from the R package $gRchain$.

Next, we perform model selection, which is based on the BIC criterion, using the function $stepwise()$ and its default argument for the penalty parameter $k = log(nrow())$. The advantage of BIC against AIC is that it favors a simpler model.

A different type of model is fitted for each variable based on its nature.

- An ordinary least squares model is used for continuous variable.
- A binomial logit model is used for binary categorical variables.
- A proportional odds logit model is used for ordinal categorical variables.

The signs assigned to each edge in $Figure\ 12$ are derived from the coefficients obtained from the corresponding fitted models. $Table\ 38$ presents the summary of each final model.

Summary for target variable: **HRTC_2cat** – (binomial logit model)					
Model	HRTC_2cat ~ SMOKING				
Deviance Residuals	Min	1Q	Median	3Q	Max
	-1.0891	-0.6485	-0.6485	1.2684	1.8235
Coefficients	**Estimate**	Std. Error	z value	Pr(>|t|)	
(Intercept)	-1.4523	0.3349	-4.336	1.45e-05 ***	
SMOKING2	**1.2409**	0.4676	2.654	0.00795 **	
Dispersion parameter for binomial family taken to be 1					
Null deviance: 115.9 on 95 degrees of freedom					
Residual deviance: 108.6 on 94 degrees of freedom					
AIC: 112.6					
Number of Fisher Scoring iterations: 4					
Summary for target variable: VO2 – (ordinary least squares model)					
Model	VO2 ~ TIME_UNTIL_FOLUP + AGE + BMI + GENDER + TIME_UNTIL_FOLUP:AGE + TIME_UNTIL_FOLUP:GENDER				
Deviance Residuals	Min	1Q	Median	3Q	Max
	-9.4358	-1.9962	-0.1612	2.0727	9.1979
Coefficients	**Estimate**	Std. Error	t value	Pr(>|t|)	
(Intercept)	24.2747093	7.0527834	3.442	0.000882 ***	
TIME_UNTIL_FOLUP	**0.1046366**	0.0489553	2.137	0.035308 *	
AGE	**0.0767692**	0.1095302	0.701	0.485197	
BMI	**-0.3120555**	0.0637743	-4.893	4.39e-06 ***	
GENDER2	**-0.7740464**	2.4119300	-0.321	0.749020	
TIME_UNTIL_FOLUP:AGE	**-0.0020451**	0.0008427	-2.427	0.017245 *	
TIME_UNTIL_FOLUP:GENDER2	**0.0362494**	0.0155825	2.326	0.022275 *	

Dispersion parameter for gaussian family taken to be 12.60276					
Null deviance: 2335.7 on 95 degrees of freedom					
Residual deviance: 1121.6 on 89 degrees of freedom					
AIC: 524.42					
Number of Fisher Scoring iterations: 2					
Summary for target variable: FEV1 – (ordinary least squares model)					
Model	FEV1 ~ AGE + BMI + GENDER + AGE:BMI				
Deviance Residuals	Min	1Q	Median	3Q	Max
	-1.08244	-0.33172	-0.00669	0.26410	1.46549
Coefficients	**Estimate**	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.940637	1.836386	-1.057	0.293414	
AGE	**0.079213**	0.031083	2.548	0.012497 *	
BMI	**0.215477**	0.060545	3.559	0.000594 ***	
GENDER2	**0.856267**	0.117071	7.314	9.82e-11 ***	
AGE:BMI	**-0.003818**	0.001035	-3.691	0.000381 ***	
Dispersion parameter for gaussian family taken to be 0.2429366					
Null deviance: 51.068 on 95 degrees of freedom					
Residual deviance: 22.107 on 91 degrees of freedom					
AIC: 143.47					
Number of Fisher Scoring iterations: 2					
Summary for target variable: DDIMER – (ordinary least squares model)					
Model	DDIMER ~ FEV1 + AGE + GENDER + FEV1:GENDER				
Deviance Residuals	Min	1Q	Median	3Q	Max
	-414.21	-186.99	-57.17	79.36	2198.46
Coefficients	**Estimate**	Std. Error	z value	Pr(>|t|)	
(Intercept)	-1841.416	614.971	-2.994	0.003542 **	
FEV1	**545.447**	192.990	2.826	0.005789 **	
AGE	**16.580**	4.617	3.591	0.000534 ***	
GENDER2	**1290.083**	503.722	2.561	0.012080 *	
FEV1:GENDER2	**-521.158**	196.233	-2.656	0.009341 **	
Dispersion parameter for binomial family taken to be 129402.5					
Null deviance: 14191445 on 95 degrees of freedom					
Residual deviance: 11775630 on 91 degrees of freedom					
AIC: 1409.3					
Number of Fisher Scoring iterations: 2					
Summary for target variable: DEPRESSION – (proportional odds logit model)					
Model	DEPRESSION ~ FATIGUE + GENDER				
Coefficients	Value	Std. Error	t value		
FATIGUE2	**1.570**	0.4837	3.245		
GENDER2	**-1.099**	0.4893	-2.246		
Intercepts	Value	Std. Error	t value		
1|2	0.8171	0.5003	1.6333		
2|3	1.6366	0.5227	3.1311		
Residual Deviance: 140.1466					
AIC: 148.1466					
Summary for target variable: FATIGUE – (binomial logit model)					
Model	FATIGUE ~ DEPRESSION				
Deviance Residuals	Min	1Q	Median	3Q	Max
	-1.8930	-0.8752	-0.8752	1.1774	1.5134
Coefficients	**Estimate**	Std. Error	z value	Pr(>|t|)	

(Intercept)	-0.7621	0.2643	-2.884	0.00393 **	
DEPRESSION2	**0.7621**	0.6350	1.200	0.23002	
DEPRESSION3	**2.3716**	0.6854	3.460	0.00054 ***	
Dispersion parameter for gaussian family taken to be 1					
Null deviance: 131.58 on 95 degrees of freedom					
Residual deviance: 115.42 on 93 degrees of freedom					
AIC: 121.42					
Number of Fisher Scoring iterations: 4					
Summary for target variable: TIME_UNTIL_FOLUP – (ordinary least squares model)					
Model	TIME_UNTIL_FOLUP ~ 1				
Deviance Residuals	Min	1Q	Median	3Q	Max
	-81.219	-43.469	-9.219	36.031	190.781
Coefficients	**Estimate**	Std. Error	t value	Pr(>|t|)	
(Intercept)	140.22	5.68	24.68	<2e-16 ***	
Dispersion parameter for binomial family taken to be 3097.52					
Null deviance: 294264 on 95 degrees of freedom					
Residual deviance: 294264 on 95 degrees of freedom					
AIC: 1047.1					
Number of Fisher Scoring iterations: 2					
Summary for target variable: **ICU** – (binomial logit model)					
Model	ICU ~ BMI + SMOKING				
Deviance Residuals	Min	1Q	Median	3Q	Max
	-1.3396	-0.6449	-0.3918	-0.3058	2.2881
Coefficients	**Estimate**	Std. Error	z value	Pr(>|t|)	
(Intercept)	-5.7716	1.6259	-3.550	0.000385 ***	
BMI	**0.1100**	0.0453	2.428	0.015185 *	
SMOKING2	**1.4230**	0.5982	2.379	0.017369 *	
Dispersion parameter for gaussian family taken to be 1					
Null deviance: 89.653 on 95 degrees of freedom					
Residual deviance: 78.149 on 93 degrees of freedom					
AIC: 84.149					
Number of Fisher Scoring iterations: 5					
Summary for target variable: **AGE** – (ordinary least squares model)					
Model	AGE ~ 1				
Deviance Residuals	Min	1Q	Median	3Q	Max
	-26.188	-6.188	1.812	6.812	19.812
Coefficients	**Estimate**	Std. Error	t value	Pr(>|t|)	
(Intercept)	58.1875	0.9652	60.28	<2e-16 ***	
Dispersion parameter for gaussian family taken to be 89.43816					
Null deviance: 8496.6 on 95 degrees of freedom					
Residual deviance: 8496.6 on 95 degrees of freedom					
AIC: 706.81					
Number of Fisher Scoring iterations: 2					
Summary for target variable: **BMI** – (ordinary least squares model)					
Model	BMI ~ GENDER				
Deviance Residuals	Min	1Q	Median	3Q	Max
	-8.223	-3.988	-1.089	2.158	24.543
Coefficients	**Estimate**	Std. Error	t value	Pr(>|t|)	
(Intercept)	33.307	1.136	29.317	<2e-16 ***	
GENDER2	**-3.170**	1.330	-2.382	0.0192 *	

Dispersion parameter for binomial family taken to be 33.55966					
Null deviance: 3345.1 on 95 degrees of freedom					
Residual deviance: 3154.6 on 94 degrees of freedom					
AIC: 613.69					
Number of Fisher Scoring iterations: 2					
Summary for target variable: **GENDER** – (binomial logit model)					
Model	GENDER ~ BMI				
Deviance Residuals	Min	1Q	Median	3Q	Max
	-1.8469	-1.1102	0.6717	0.7770	1.6206
Coefficients	**Estimate**	Std. Error	Z value	Pr(>|t|)	
(Intercept)	3.70352	1.26707	2.923	0.00347 **	
BMI	**-0.08602**	0.03901	-2.205	0.02747 *	
Dispersion parameter for gaussian family taken to be 1					
Null deviance: 112.14 on 95 degrees of freedom					
Residual deviance: 106.98 on 94 degrees of freedom					
AIC: 110.98					
Number of Fisher Scoring iterations: 4					
Summary for target variable: **SMOKING** – (binomial logit model)					
Model	SMOKING ~ 1				
Deviance Residuals	Min	1Q	Median	3Q	Max
	-1.004	-1.004	-1.004	1.361	1.361
Coefficients	**Estimate**	Std. Error	z value	Pr(>|t|)	
(Intercept)	-0.4229	0.2087	-2.026	0.0428 *	
Dispersion parameter for gaussian family taken to be 1					
Null deviance: 128.89 on 95 degrees of freedom					
Residual deviance: 128.89 on 95 degrees of freedom					
AIC: 130.89					
Number of Fisher Scoring iterations: 4					

Table 38: CGM - Models' summaries

In $Figure\ 12$, it is important to explain the meaning of the blocks presented. The plot introduces four blocks in total.

A chain graph model is fitted to the variables divided into blocks labelled 2, 3 and 4 based on their respective roles in the model. Particularly, block 4 consists of the basic explanatory variables, block 3 includes the intermediate variables and block 2 contains the response variables.

Now, let's turn our attention to block 1. This block consists of a single variable, $HRTC$. This variable corresponds to the initial $HRTC$ variable from the dataset, but it includes its three categories, after removing all the missing values. Its association with the other variables was examined using proportional odds ordinal logistic regression, which is presented in $Table\ 39$.

Summary for target variable: **HRTC** – (proportional odds logit model)				
Model	HRTC ~ GENDER + ICU + SMOKING + AGE + BMI + TIME_UNTIL_FOLUP + FATIGUE + DEPRESSION + VO2 + FEV1 + DDIMER			
Coefficients	**Estimate**	Std. Error	t value	Pr(>|t|)
GENDER2	**1.0441229**	0.5390465	1.9370	0.058 .
ICU2	**1.3120458**	0.4579026	2.8653	0.0059 **
SMOKING2	**-0.1972705**	0.3790765	-0.5204	0.6049
AGE	**0.0689372**	0.0172667	3.9925	2e-04 ***
BMI	**-0.0266518**	0.0232380	-1.1469	0.2565
TIME_UNTIL_FOLUP	**-0.0079339**	0.0033027	-2.4023	0.0198 *
FATIGUE2	**-0.1353469**	0.3624706	-0.3734	0.7103
DEPRESSION2	**-0.8277188**	0.5244957	-1.5781	0.1204
DEPRESSION3	**-0.7012201**	0.5069005	-1.3833	0.1723
VO2	**-0.0796807**	0.0380993	-2.0914	0.0412 *
FEV1	**-0.3383030**	0.2856785	-1.1842	0.2415
DDIMER	**-0.0006524**	0.0004869	-1.3400	0.1858
Intercepts	Value	Std. Error	t value	
1|2	**-0.2831**	0.0510	-5.5529	
2|3	**1.0348**	0.2470	4.1886	
Residual deviance: 106.4238				
AIC: 134.4238				

Table 39: CGM - HRTC – proportional odds logit model's summary

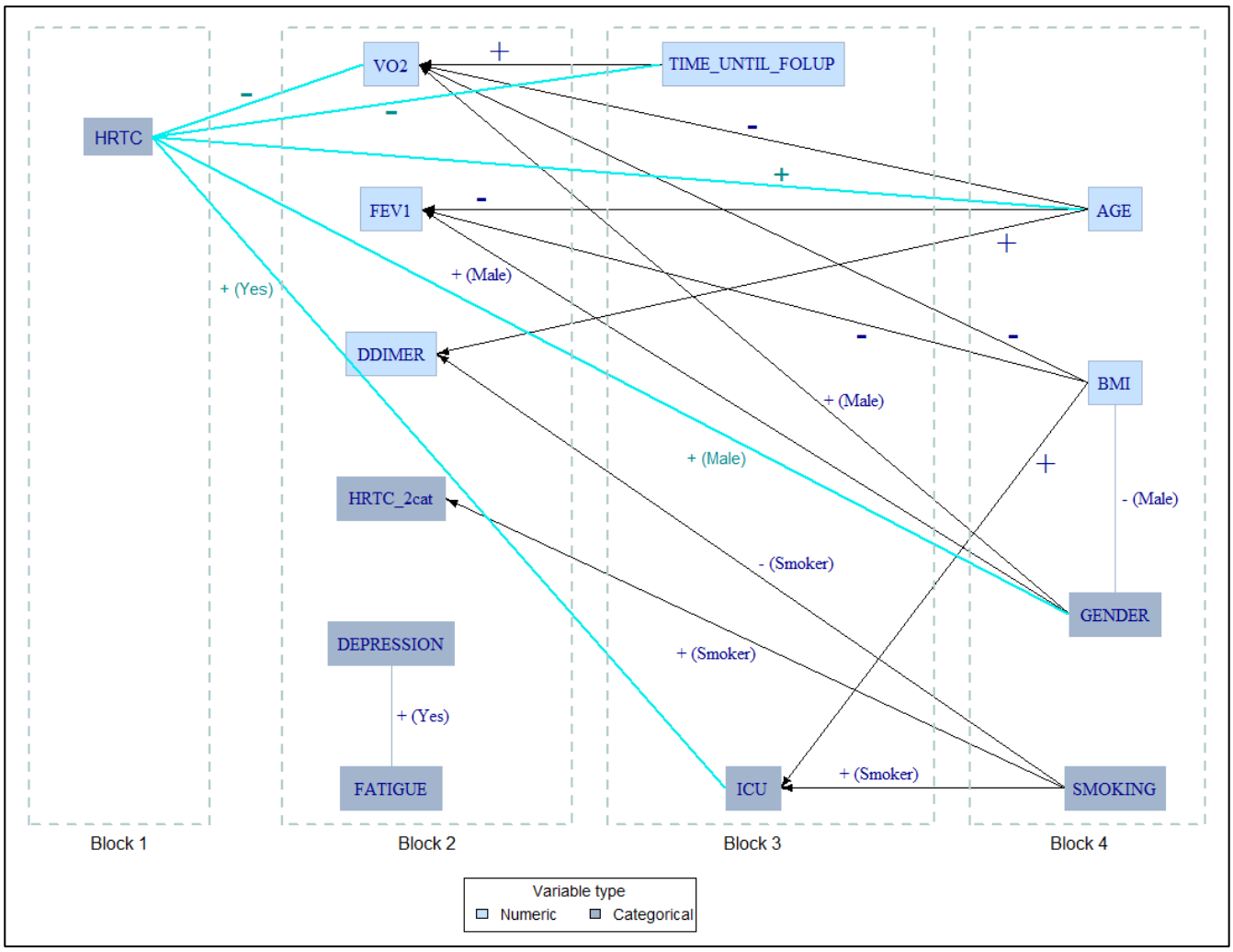


Figure 12: Chain Graph Model using the dataset from Table 10.

Variables connected by an edge	Sign	Interpretation
DEPRESSION ~ FATIGUE	+(Yes)	Patients who felt fatigue tended to be more depressed.
TIME_UNTIL_FOLUP ~ VO2	+	As time passes after the discharge date, the maximal oxygen consumption increases.
AGE ~ FEV1	-	As age increases, the maximum volume of exhaled air during the 1st second after maximum inhalation decreases.
AGE ~ DDIMER	+	As age increases, the DDIMER value also increases.
AGE ~ VO2	-	As age increases, the maximal oxygen consumption decreases.
BMI ~ VO2	-	As the body mass index increases, the maximal oxygen consumption decreases.
BMI ~ FEV1	-	As the body mass index increases, the maximum volume of exhaled air during the 1st second after maximum inhalation decreases.
BMI ~ ICU	+	As the body mass index increases, the likelihood of entering in the intensive care unit also increases.
BMI ~ GENDER	-(Male)	Males tend to have lower body mass index compared to women.
VO2 ~ GENDER	+(Male)	Males tend to have higher maximal oxygen consumption compared to women.
GENDER ~ FEV1	+(Male)	Males tend to have higher maximum volume of air exhaled during the first second after a maximum inhalation compared to women.
DDIMMER ~ SMOKING	-(Smoker)	Smokers tend to have a decrease to the DDimer value.
SMOKING ~ ICU	+(Smoker)	Smokers tend to have an increased likelihood of entering in the intensive care unit.
SMOKING ~ HRTC_2cat	+(Smoker)	Smokers tend to have a decreased likelihood of undergoing a CT scan.
HRTC ~ VO2	-	If a patient has a greater maximal oxygen consumption, then it is more likely to fall into a lower HRTC category. In other words, to have better outcome.
HRTC ~ TIME_UNTIL_FOLUP	-	As time passes after the discharge date, patients tend to have better outcomes from their CT scan.
HRTC ~ AGE	+	As the patient's age increases, it is more likely to have poorer results from their CT scan.
HRTC ~ GENDER	+(Male)	Males tend to have an increases likelihood for having poorer results from the CT scan.
HRTC ~ ICU	+(Yes)	If the patient was admitted to the intensive care unit, then it is more likely to have poorer results from the CT scan.

Table 40: CGM interpretation

Conclusions

In this thesis, we began by introducing the fundamental theory underlying three important types of graphical models: the Undirected Gaussian Graphical model, the Mixed Interaction Model, and the Chain Graph model. Afterwards, we continued to their application on real data, utilizing the statistical capabilities of the programming tool, R. Consequently, we combined theory and application to gain the knowledge and skills for utilizing graphical models for statistical analysis.

In conclusion, we investigated focusing on the associations of objective measurements of physical condition and subjective evaluations obtained through questionnaires. This approach allowed us to gain insights into the long-term health impact of COVID-19. The summarizing of the results from $Figures\ 6, 7$, and 8 and $Tables\ 13, 14$, and 15, revealed that the effects of the virus on individuals were influenced by multiple factors, including the patient's age, the patient's body mass index, as well as the gender of a patient.

One noteworthy general observation is that the chain graph in $Figure\ 8$ shares several edges with the mixed interaction graph in $Figure\ 7$. This consistency is very important and valuable because it reinforces the validity of conclusions coming from these graphs.

We discovered various interesting results, both expected and unexpected, taking into consideration present but also absent associations.

First, we observed differences based on gender. The analysis revealed that males tended to exhibit higher values of maximal oxygen consumption ($VO2$), as well as maximum volume of exhaled air during the first second after maximum inhalation ($FEV1$) compared to females.

Secondly, differences based on age were uncovered. The results suggest that age can influence health outcomes. Particularly, older patients displayed an increase to the level of $DDIMER$, which is a marker of blood clotting. In addition, as age increased, a decline was noticed in both maximal oxygen consumption ($VO2$) and maximum volume of exhaled air during the first second after maximum inhalation ($FEV1$).

One more quite valuable observation is that the self-assessed symptom fatigue was positively associated with depression. But the aspect of interest here lies in the absence of associations between these two subjective assessments and the other objective measurements. This finding highlights the complex relationship between physical and mental health of individuals who are recovering from COVID-19 and shows us the need for support to these people.

To continue with our results, we also observed that the body mass index (BMI) and the admission into the intensive care unit (ICU) played a crucial role to a patient's recovery. Patients with larger body mass index tended to experience a decline in both maximal oxygen consumption ($VO2$) and maximum volume of exhaled air during the first second after maximum inhalation ($FEV1$). Furthermore, individuals with a higher body mass index had an increased likelihood of entering the intensive care unit (ICU) while their hospitalization.

Moreover, we discovered through the graphs an association between the variables $Smoking$ and ICU. More specifically, a higher likelihood of admission in the intensive care unit was observed among smokers.

Last but not least, is the relationship between $VO2$ and $HRTC$. We discovered that those individuals with an increased maximal oxygen consumption ($VO2$) tended to show better outcomes in the High-Resolution Computed Tomography ($HRTC$), while older patients tended to have poorer CT scan outcomes.

In summary, these findings highlight the effectiveness of graphical models in unraveling hidden patterns and understanding complex relationships. Through the application of these sophisticated techniques, we can identify crucial factors that influence health outcomes due to a disease and find more appropriate interventions to mitigate the impacts.

Bibliography

Cox, D.R. & Wermuth, N. (1993). Linear Dependencies Represented by Chain Graphs. *Statistical Science, 8*(3), 204-218. doi:10.1214/ss/1177010887

Cox, D.R. & Wermuth, N. (1996). *Multivariate Dependencies: Models, Analysis and Interpretation.* London: Chapman and Hall.

Edwards, D. (2000). *Introduction to graphical modelling* (2nd ed.). New York: Springer.

Gauraha, N. (2017). Graphical Log-Linear Models: Fundamental Concepts and Applications. *Journal of Modern Applied Statistical Methods, 16*(1), 545-577. doi:10.22237/jmasm/1493598000

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2 ed.). New York: Springer. doi:https://doi.org/10.1007/978-0-387-84858-7

Højsgaard, S., Edwards, D., & Lauritzen S. (2012). *Graphical Models with R.* New York: Springer. doi:10.1007/978-1-4614-2299-0

Koller, D. & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques.* MIT Press.

Lauritzen, S.L & Richardson, T.S. (2002). Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(3), 321-348.

Maathuis, M., Drton, M., Lauritzen, S. & Wainwright, M. (2018). *Handbook of Graphical Models.* CRC Press, Inc.

Tur, I. & Castelo, R. (2011). Learning mixed graphical models from data with p larger than n. *ResearchGate*.

Wermuth, N. & Lauritzen, S.L. (1990). On Substantive Research Hypotheses, Conditional Independence Graphs and Graphical Chain Models. *Journal of the Royal Statistical Society: Series B (Methodological), 52*(1), 21-50. doi:https://doi.org/10.1111/j.2517-6161.1990.tb01771.x

Wermuth, N. & Sadeghi, K. (2012). Sequences of regressions and their independences. *TEST, 21*, 268–273. doi:10.1007/s11749-012-0290-6

Καρώνη, Χ. & Οικονόμου, Π. (2017). *Στατιστικά Μοντέλα Παλινδρόμησης: Με χρήση MINITAB και R* (2 ed.). Αθήνα: Συμεών.

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΜΕΤΑΠΤΥΧΙΑΚΟ: ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕ ΧΡΗΣΗ ΜΟΝΤΕΛΩΝ ΜΙΚΤΗΣ ΑΛΛΗΛΕΠΙΔΡΑΣΗΣ ΚΑΙ ΜΟΝΤΕΛΩΝ ΓΡΑΦΗΜΑΤΩΝ ΑΛΥΣΙΔΑΣ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΓΟΥΡΓΟΥΡΑ ΚΩΝΣΤΑΝΤΙΝΑ

ΑΜ: 03400128

Επιβλέπουσα καθηγήτρια: Χ. Καρώνη-Ρίτσαρντσον

Επιτροπή

Χ. Καρώνη-Ρίτσαρντσον	Β. Παπανικολάου	Κ. Χρυσαφίνος
Ομ. Καθηγήτρια ΕΜΠ	Ομ. Καθηγητής ΕΜΠ	Καθηγητής Ε.Μ.Π

Αθήνα, 2023

Περίληψη

Η στατιστική ανάλυση είναι μια επιστήμη που μας δίνει τη δυνατότητα να επεξεργαζόμαστε, να παρουσιάζουμε και να ερμηνεύουμε δεδομένα που προέρχονται από διάφορες πηγές, συμπεριλαμβανομένης της μηχανικής, της βιολογίας, της οικονομίας και της ψυχολογίας. Η παρούσα διπλωματική εργασία περιλαμβάνει δύο μέρη: ένα θεωρητικό και ένα πρακτικό μέρος. Το πρώτο μέρος αποτελείται από δύο κεφάλαια, τα οποία παρέχουν ένα θεωρητικό υπόβαθρο για διαφορετικές τεχνικές γραφικών μοντέλων στη στατιστική ανάλυση. Το δεύτερο μέρος αποτελείται από ένα τρίτο κεφάλαιο, εφαρμόζοντας αυτές τις τεχνικές σε πραγματικά δεδομένα και συγκρίνοντας τα αποτελέσματά τους.

Ένας γράφος χρησιμεύει ως πολύτιμο εργαλείο για την οπτική αναπαράσταση των σχέσεων μεταξύ των δεδομένων. Μέσω ενός γραφήματος, είναι ευκολότερο να παρουσιάσουμε σύνθετες πληροφορίες όχι μόνο σε ειδικούς του κλάδου, αλλά και σε άτομα που δεν έχουν γνώση της θεωρίας πίσω από αυτές. Επιπλέον, αυτή η προσέγγιση μας επιτρέπει να διακρίνουμε μοτίβα και σχέσεις που διαφορετικά δεν θα παρατηρούνταν στα πρωτογενή δεδομένα.

Στο πρώτο κεφάλαιο, παρουσιάζουμε τρεις τύπους μη κατευθυνόμενων γραφικών μοντέλων. Η πρώτη τεχνική επικεντρώνεται γύρω από log-linear μοντέλα, τα οποία βρίσκουν εφαρμογή όταν τα σύνολα δεδομένων αποτελούνται από διακριτές μεταβλητές. Η δεύτερη τεχνική αφορά τα Gaussian γραφικά μοντέλα, που χρησιμοποιούνται ειδικά με σύνολα δεδομένων που περιλαμβάνουν συνεχείς μεταβλητές. Τέλος, η τρίτη τεχνική εμβαθύνει σε μοντέλα μικτής αλληλεπίδρασης, τα οποία έχουν σχεδιαστεί για σύνολα δεδομένων που αποτελούνται τόσο από συνεχείς όσο και από κατηγορικές μεταβλητές. Αυτά τα γραφικά μοντέλα στοχεύουν στην απόκτηση γνώσης για τις κρυφές συσχετίσεις και προσφέρουν τη δυνατότητα αναπαράστασής τους σε μορφή διαγράμματος, όπου κάθε ακμή αντιπροσωπεύει μια συσχέτιση μεταξύ των δύο μεταβλητών που συνδέει.

Στο δεύτερο κεφάλαιο, παρουσιάζουμε ένα από τα σημαντικότερα κατευθυνόμενα γραφικά μοντέλα. Αυτή η τεχνική επικεντρώνεται γύρω από μοντέλα γραφημάτων αλυσίδας και μπορεί να χρησιμοποιηθεί όταν αντιμετωπίζουμε σύνολα δεδομένων που αποτελούνται τόσο από συνεχείς όσο και από κατηγορικές μεταβλητές. Τα μοντέλα γραφημάτων αλυσίδας χρησιμοποιούν κουτιά για να διαχωρίσουν μεταβλητές με βάση τον ρόλο τους στο μοντέλο. Μοιράζονται ομοιότητες με τα μοντέλα μικτής αλληλεπίδρασης, αλλά η διαφορά τους έγκειται στη διαδικασία παλινδρόμησης. Στα μοντέλα γραφημάτων αλυσίδας κάθε μεταβλητή παλινδρομείτε με βάση τις μεταβλητές που υπάρχουν σε όλα τα προηγούμενα κουτιά, ενώ στα μοντέλα μικτής αλληλεπίδρασης η παλινδρόμηση λαμβάνει υπόψη τις μεταβλητές όχι μόνο από τα προηγούμενα κουτιά, αλλά και από το τρέχον κουτί.

Στο τρίτο κεφάλαιο, προσαρμόζουμε ένα Gaussian γραφικό μοντέλο, ένα μοντέλο μικτής αλληλεπίδρασης και ένα μοντέλο γραφημάτων αλυσίδας χρησιμοποιώντας το στατιστικό πρόγραμμα R. Πιο συγκεκριμένα, αναλύσαμε δεδομένα από μια μελέτη που ακολούθησε ασθενείς που νοσηλεύτηκαν λόγω του ιού COVID-19. Η έρευνα περιλάμβανε διάφορα δεδομένα, όπως ανθρωπομετρικά, νοσηλευτικά και ψυχολογικά δεδομένα τα οποία

προέρχονται από ερωτηματολόγια. Στο κεφάλαιο αυτό, έπειτα από κάθε εφαρμογή, παρέχεται μια επεξήγηση των αποτελεσμάτων και στο τέλος τα ευρήματα συγκρίνονται και σχολιάζονται.

Λέξεις κλειδιά: Gaussian γραφικό μοντέλο, μοντέλο μεικτής αλληλεπίδρασης, μοντέλα αλυσίδας, AIC, BIC, κουτιά, κατευθυνόμενες ακμές, μη κατευθυνόμενες ακμές.

Ευχαριστίες

Θα ήθελα να εκφράσω τις ευχαριστίες μου στην επιβλέπουσα μου Χ. Καρώνη για την καθοδήγηση, και την υποστήριξή της κατά τη διάρκεια της έρευνάς μου. Τα οξυδερκή σχόλια και η ενθάρρυνσή της ήταν πολύτιμα στη διαμόρφωση της κατεύθυνσης και του εύρους της εργασίας μου. Ακόμα, θα ήθελα να ευχαριστήσω τα μέλη της επιτροπής Β. Παπανικολάου και Κ. Χρυσαφινό για τη συμβολή τους.

Επίσης, εκφράζω την εκτίμησή μου στην καθηγήτρια E. Stanghellini για την πολύτιμη καθοδήγηση, τη σε βάθος γνώση και τη συνεχή ενθάρρυνσή της, τα οποία έπαιξαν σημαντικό ρόλο στην εκπόνηση αυτής της εργασίας. Επιπλέον, θα ήθελα να ευχαριστήσω τον καθηγητή F. Bartolucci, τον Dr. G. Pucci και την Dr. P. Rivadeneyra για τη βοήθειά τους σε αυτή τη μελέτη. Χωρίς τη συλλογική τους προσπάθεια τίποτα από όλα αυτά δεν θα ήταν δυνατό.

Είναι σημαντικό να αναγνωρίσουμε ότι η παρούσα διατριβή βασίστηκε στην ερευνητική επιχορήγηση με τίτλο «Multidimensional statistical analysis of databases relating to patients affected by "long-Covid" in order to characterize their manifestations, identify their risk factors and identify their therapeutic trajectories.». Η συγκεκριμένη έρευνα χρηματοδοτήθηκε από το Πανεπιστήμιο της Περούντζια, στην Ιταλία.

Επιπροσθέτως, θα ήθελα να εκφράσω τις ευχαριστίες μου προς το διδακτικό προσωπικό του Εθνικού Μετσόβιου Πολυτεχνείου, που κατάφεραν να δημιουργήσουν ένα υποστηρικτικό ακαδημαϊκό περιβάλλον για τους φοιτητές, όπως επίσης και η βοήθεια του διοικητικού προσωπικού υπήρξε καθοριστική για τον συντονισμό όλων των απαραίτητων διαδικασιών.

Τέλος, είμαι πολύ ευγνώμων στην οικογένεια και τους φίλους μου, που με στήριξαν καθ' όλη τη διάρκεια των σπουδών μου. Μου παρείχαν σταθερά αγάπη, φροντίδα και ένα αυτί που ακούει σε στιγμές ανάγκης. Η πίστη τους στις ικανότητές μου πάντα με ενέπνεε και με παρακινούσε να βελτιώνω τον εαυτό μου και να προσπαθώ να πετύχω τους στόχους μου.

Κεφάλαιο 1 – Μη-κατευθυνόμενα Γραφικά Μοντέλα

Σε αυτό το κεφάλαιο, θα εισάγουμε τις βασικές έννοιες των μη-κατευθυνόμενων γραφημάτων. Συγκεκριμένα, θα συζητήσουμε πώς σχετίζονται με τα log-linear μοντέλα, Gaussian γραφικά μοντέλα και μοντέλα μεικτών αλληλεπιδράσεων, τα οποία συνδυάζουν στοιχεία από τα δύο προηγούμενα. Μέσω της κατανόησης των βασικών μη-κατευθυνόμενων γραφημάτων και των εφαρμογών τους σε διάφορους τύπους μοντέλων, μπορούμε να αναλύουμε και να ερμηνεύουμε καλύτερα πολύπλοκα δεδομένα.

1.1 Γράφος

Ένας γράφος είναι μια συλλογή σημείων, τα οποία συνδέονται μεταξύ τους με γραμμές. Τα σημεία και οι γραμμές ονομάζονται αντίστοιχα κόμβοι και ακμές. Κάθε ακμή αναπαριστά μια σχέση μεταξύ των κόμβων που συνδέει.

Στο πλαίσιο ενός μοντέλου, οι κόμβοι μπορούν να χρησιμοποιηθούν για να αναπαραστήσουν μεταβλητές, και οι ακμές μπορούν να αναπαραστήσουν τις σχέσεις μεταξύ αυτών των μεταβλητών. Μέσω αυτής της αναπαράστασης, μπορούμε να αποκτήσουμε εισαγωγή σε πώς διάφοροι παράγοντες επηρεάζουν ο ένας τον άλλο και πώς οι αλλαγές σε μια μεταβλητή μπορούν να επηρεάσουν άλλες.

Θεωρήστε δύο κόμβους που συμβολίζονται με τα γράμματα Α και Β. Στους μη-κατευθυνόμενους γράφους, η ακμή μεταξύ των δύο κόμβων συμβολίζεται είτε ως $[AB]$ είτε ως $[BA]$, αφού η ακμή δεν έχει μια συγκεκριμένη κατεύθυνση (Edwards, 2000).

Ένας συνηθισμένος τρόπος απεικόνισης ενός γράφου οπτικά είναι μέσω ενός διαγράμματος.

1.2 Log-linear Μοντέλα

Τα log-linear μοντέλα είναι στατιστικά μοντέλα που επιτρέπουν την ανάλυση των σχέσεων μεταξύ κατηγορικών μεταβλητών. Υπολογίζουν την πιθανότητα παρατήρησης ενός συγκεκριμένου συνδυασμού των κατηγοριών των μεταβλητών στο μοντέλο, εκτιμώντας ότι ο λογάριθμος της πιθανότητας είναι γραμμική συνάρτηση των μεταβλητών. Αυτή η ιδιότητα επιτρέπει την μοντελοποίηση περίπλοκων συσχετίσεων και αλληλεπιδράσεων μεταξύ πολλαπλών κατηγορικών μεταβλητών. Λόγω αυτής της ιδιότητας, τα log-linear είναι χρήσιμα σε πολλούς τομείς, όπως η επιδημιολογία, οι κοινωνικές επιστήμες και η έρευνα αγοράς.

1.2.1 Log-linear Μοντέλα, Γραφικά Μοντέλα & Μοντέλα Αποσύνθεσης

Τα log-linear μοντέλα μπορούν να κατηγοριοποιηθούν ως ιεραρχικά και μη-ιεραρχικά. Σε αυτήν την εργασία θα επικεντρωθούμε στα ιεραρχικά μοντέλα, τα οποία είναι και τα πιο κοινά μοντέλα. Σε ένα log-linear μοντέλο, ο όρος "ιεραρχικός" αναφέρεται στο γεγονός ότι συμπεριλαμβάνονται όλοι οι όροι χαμηλότερης τάξης που μπορούν να προκύψουν από τις μεταβλητές που περιέχονται σε κάθε όρο υψηλότερης τάξης (Gauraha, 2017). Οι όροι υψηλότερης τάξης ονομάζονται γεννήτορες. Ένα ιεραρχικό log-linear μοντέλο θεωρείται γραφικό, εάν οι γεννήτορές του αντιστοιχούν στις κλίκες του μη-κατευθυνόμενου γραφήματος, όπου οι κόμβοι αναπαριστούν τις μεταβλητές και οι ακμές αναπαριστούν τους όρους του μοντέλου που περιέχουν δύο μεταβλητές. Ένα γραφικό log-linear μοντέλο θεωρείται αποσυνθέσιμο εάν το γράφημά του είναι τριγωνισμένο. Σύμφωνα με τον ορισμό που παρέχεται στην ενότητα (1.1.1), ένα γράφημα θεωρείται τριγωνισμένο όταν δεν περιέχει κύκλους μήκους τέσσερα ή περισσότερο, χωρίς όμως να υπάρχει χορδή (Maathuis et al., 2018).

1.2.2 Ιεραρχικό Μοντέλο

Έστω $X = (X_1, X_2, \dots, X_k)$ να είναι k διακριτές τυχαίες μεταβλητές ενός συνόλου δεδομένων D και έστω d να είναι ένα υποσύνολο αυτού του συνόλου δεδομένων, $d \subseteq D$. Επιπλέον, έστω I το σύνολο όλων των δυνατών συνδυασμών τιμών για αυτές τις μεταβλητές. Ένα κελί του I συμβολίζεται ως $i = (i_1, i_2, \dots, i_k)$ και αποτελεί μια παρατήρηση των τιμών των διακριτών μεταβλητών, όπου i_j είναι η τιμή της j-οστής μεταβλητής στο κελί.

Ένα ιεραρχικό log-linear μοντέλο δίνεται από τον τύπο,

$$\log p_i = \sum_{d \subseteq D} u_i^d \tag{1.1}$$

όπου,

- $\log p_i$ είναι ο λογάριθμος της πιθανότητας παρατήρησης του κελιού i,
- $\Sigma_{d \subseteq D}$ είναι η άθροιση όλων των δυνατών όρων αλληλεπίδρασης στο μοντέλο,
- u_i^d είναι ένας όρος αλληλεπίδρασης, ο οποίος αντιστοιχεί στο κελί i και εξαρτάται μόνο από τις μεταβλητές που εμπλέκονται στο υποσύνολο d,

(Edwards, 2000).

1.2.3 Συνάρτηση Πιθανοφάνειας

Η συνάρτηση πιθανοφάνειας ενός ιεραρχικού log-linear μοντέλου εκφράζει την πιθανότητα παρατήρησης ενός συνόλου δεδομένων δοθέντων των παραμέτρων του μοντέλου. Η συνάρτηση πιθανοφάνειας προκύπτει από την κοινή κατανομή των παρατηρούμενων δεδομένων και των παραμέτρων του μοντέλου και δίνεται από τον τύπο:

$$L = c \prod_{i \in I} p(i)^{n_i} \tag{1.2}$$

όπου,

- i είναι η i-οστή κελί στον πίνακα παρατήρησης,
- n_i είναι ο παρατηρούμενος αριθμός του i,
- $p(i)$ είναι η προβλεπόμενη πιθανότητα του i, και
- c είναι μια σταθερά.

Η εξίσωση (1.3) παρουσιάζει τη λογαριθμοποιημένη συνάρτηση πιθανοφάνειας.

$$\hat{l} = c + \sum_{i \in I} n(i) \log \hat{p}(i) \tag{1.3}$$

όπου $\hat{p}(i)$ είναι οι εκτιμήσεις της μεγιστοποιημένης πιθανοφάνειας (MLEs) του μοντέλου (Højsgaard et al., 2012).

1.2.4 Καταλληλότητα του Μοντέλου

Η απόκλιση (deviance) αποτελεί ένα μέτρο του πόσο καλά ταιριάζει το μοντέλο στα δεδομένα. Μετρά τη διαφορά της μεγιστοποιημένης λογαριθμοποιημένης συνάρτησης πιθανοφάνειας μεταξύ του προσαρμοσμένου μοντέλου m και του κορεσμένου μοντέλου s. Αυτός είναι ένας τρόπος αξιολόγησης της καταλληλότητας του μοντέλου, όπως επίσης χρησιμοποιείται και για τη σύγκριση δύο διαφορετικών μοντέλων. Μια τιμή απόκλισης 0 υποδεικνύει ένα τέλειο προσαρμοσμένο μοντέλο, ενώ μεγαλύτερες τιμές υποδεικνύουν ένα χειρότερη προσαρμογή.

Η απόκλιση για ένα log-linear μοντέλο δίνεται από τον τύπο,

$$D = 2\left(\hat{l}_s - \hat{l}_m\right) \tag{1.4}$$

όπου l_s, l_m είναι η μεγιστοποιημένη λογαριθμοποιημένη συνάρτηση πιθανοφάνειας του κορεσμένου μοντέλου, και του προσαρμοσμένου μοντέλου αντίστοιχα. Χρησιμοποιώντας την εξίσωση (1.3),

$$\hat{l}_s = c + \sum_{i \in I} n(i) \log \hat{p}_s(i) = c + \sum_{i \in I} n(i) \log\left(\frac{n(i)}{N}\right) \tag{1.5}$$

όπου N είναι ο αριθμός των παρατηρήσεων, και

$$\hat{l}_m = c + \sum_{i \in I} n(i) \log \hat{p}_m(i) \tag{1.6}$$

Συνδυάζοντας $(1.4), (1.5), (1.6)$,

$$D = 2 \sum_{i} n(i) \log\left(\frac{n(i)}{\hat{m}(i)}\right) \quad \sim \quad X_k^2 \tag{1.7}$$

όπου,

- n_i είναι ο παρατηρούμενος αριθμός του i-οστού κελιού, και
- $\hat{m}(i)$ είναι ο αναμενόμενος αριθμός κελιών, $\hat{m}(i) = N\hat{p}_m(i)$.

Η απόκλιση μπορεί να χρησιμοποιηθεί για να πραγματοποιήσει ελέγχους υποθέσεων στο μοντέλο, χρησιμοποιώντας την κατανομή X^2 με βαθμούς ελευθερίας k, ίσους με τη διαφορά στη διάσταση μεταξύ του κορεσμένου μοντέλου s και του προσαρμοσμένου μοντέλου m (Højsgaard et al., 2012).

1.2.5 Επιλογή Μοντέλου

Η επιλογή μοντέλου είναι σημαντική στη στατιστική μοντελοποίηση και ανάλυση δεδομένων. Ο σκοπός της επιλογής μοντέλου είναι να αναγνωρίσει το πιο κατάλληλο μοντέλο που αποτυπώνει τα μοτίβα και τις σχέσεις στα δεδομένα, αποφεύγοντας την υπερ-προσαρμογή ή την υπο-προσαρμογή.

Ένας τρόπος για να επιλεγεί το καλύτερο μοντέλο ανάμεσα σε μοντέλα που προσαρμόζονται με βάση τη συνάρτηση πιθανοφάνειας είναι η εφαρμογή του Κριτηρίου AIC, το οποίο ορίζεται ως

$$AIC = -2\ln L + 2d$$

όπου,

- $\ln(L)$ είναι η μεγιστοποιημένη συνάρτηση πιθανοφάνειας και
- d είναι ο αριθμός των παραμέτρων του μοντέλου που εξετάζεται.

Αυτό το κριτήριο χρησιμοποιείται υπολογίζοντας και συγκρίνοντας τις τιμές του AIC για μοντέλα με διαφορετικό αριθμό μεταβλητών και επιλέγεται το μοντέλο με το χαμηλότερο AIC κάθε φορά, μέχρι να μην υπάρχει άλλο μοντέλο με χαμηλότερη τιμή AIC (Καρώνη & Οικονόμου, 2017).

Με παρόμοιο τρόπο με το κριτήριο AIC, εφαρμόζουμε το κριτήριο BIC, το οποίο ορίζεται ως

$$BIC = -2\ln L + d\ln N$$

όπου,

- d είναι ο αριθμός των παραμέτρων του μοντέλου που εξετάζεται και
- N είναι ο αριθμός των παρατηρήσεων.

Καθώς ο αριθμός των παρατηρήσεων πλησιάζει το άπειρο, η πιθανότητα το BIC να επιλέξει με ακρίβεια το καλύτερο μοντέλο αυξάνεται σημαντικά και τείνει προς τη μονάδα, αντίθετα με το κριτήριο AIC που έχει τάση να προτιμά πιο πολύπλοκα μοντέλα σε τέτοιες περιπτώσεις (Hastie et al., 2009).

Τα κριτήρια AIC και BIC βρίσκουν εφαρμογή και στα μοντέλα που παρουσιάζονται στις επόμενες ενότητες.

1.3 Μοντέλα Γκαουσιανής Γραφικής Αναπαράστασης

Τα μοντέλα Γκαουσιανής γραφικής αναπαράστασης είναι στατιστικά μοντέλα που χρησιμοποιούνται για να αναπαραστήσουν τις συνθήκες ανεξαρτησίας μεταξύ μεταβλητών σε μια Γκαουσιανή κατανομή πολλών μεταβλητών. Αυτά τα μοντέλα χρησιμοποιούνται για την ανάλυση συνεχών δεδομένων. Το γράφημα δείχνει τις εξάρτησης μεταξύ των μεταβλητών, με κόμβους που αντιπροσωπεύουν μεταβλητές και ακμές που αντιπροσωπεύουν συνθήκες εξάρτησης. Οι ακμές σε ένα μοντέλο Γκαουσιανής γραφικής αναπαράστασης υποδεικνύουν μερικές συσχετίσεις, οι οποίες είναι οι συνθήκες συσχέτισης μεταξύ των μεταβλητών αφού ληφθούν υπόψιν και οι άλλες μεταβλητές στο μοντέλο. Τα μοντέλα Γκαουσιανής γραφικής αναπαράστασης χρησιμοποιούνται συνήθως σε διάφορους τομείς, όπως η στατιστική, η βιολογία και οι οικονομικές επιστήμες.

1.3.1 Γκαουσιανό Μοντέλο Γραφικής Αναπαράστασης

Η μορφή ενός Γκαουσιανού μοντέλου γραφικής αναπαράστασης μπορεί να αναπαρασταθεί χρησιμοποιώντας έναν πίνακα ακρίβειας, γνωστό και ως αντίστροφος πίνακας συνδιακύμανσης ή πίνακας συγκέντρωσης.

Έστω $X = (X_1, X_2, \ldots, X_k)$ ένα πολυμεταβλητό Γκαουσιανό τυχαίο διάνυσμα με k μεταβλητές και έστω Σ να συμβολίζει τον πίνακα συνδιακύμανσης του X. Ο πίνακας συγκέντρωσης Ω, που είναι ο αντίστροφος του Σ, αναπαριστά τις συνθήκες ανεξαρτησίας μεταξύ των μεταβλητών στο Γκαουσιανό μοντέλο γραφικής αναπαράστασης.

Η συνάρτηση πυκνότητας του X μπορεί να εκφραστεί ως

$$f(x) = (2\pi)^{-\frac{k}{2}} |\Omega|^{\frac{1}{2}} exp\left(-\frac{1}{2}\,(x-\mu)^T \Omega (x-\mu)\right), x \in \mathbb{R}^k \qquad (1.8)$$

όπου,

- $(2\pi)^{-\frac{k}{2}}$ είναι μια σταθερά κανονικοποίησης,
- $|\Omega|$ αναπαριστά τον πίνακα συγκέντρωσης Ω ($det\Omega = det\Sigma^{-1}$),
- μ είναι το k-διάστατο διάνυσμα των μέσων της Γκαουσιανής κατανομής $N(\mu, \Sigma)$,
- Σ είναι ο $k \times k$ πίνακας συνδιακύμανσης της γκαουσιανής κατανομής $N(\mu, \Sigma)$,

(Maathuis et al., 2018).

Μια μηδενική τιμή στον πίνακα συγκέντρωσης Ω υποδηλώνει την υπόθεση της υπό όρους ανεξαρτησίας μεταξύ των αντίστοιχων μεταβλητών, ενώ μια μη μηδενική τιμή υποδηλώνει μια υπό όρους εξάρτηση.

Η γραφική δομή ενός τέτοιου μοντέλου μπορεί να εξαχθεί από το μοτίβο αραιότητας του πίνακα συγκέντρωσης Ω, όπου οι μη μηδενικές τιμές στον Ω αντιστοιχούν σε ακμές μεταξύ μεταβλητών στο γραφικό μοντέλο.

1.3.2 Συνάρτηση Πιθανοφάνειας

Σε ένα Γκαουσιανό μοντέλο γραφικής αναπαράστασης, η συνάρτηση πιθανοφάνειας χρησιμοποιείται για την εκτίμηση των παραμέτρων του μοντέλου. Είναι ο φυσικός λογάριθμος της κοινής πυκνότητας πιθανότητας της πολυδιάστατης Γκαουσιανής κατανομής. Η κοινή πυκνότητα πιθανότητας είναι μια συνάρτηση του διανύσματος μέσης τιμής μ και του πίνακα συγκέντρωσης Ω.

Χρησιμοποιώντας την εξίσωση (1.8), η συνάρτηση πιθανοφάνειας δίνεται από την εξίσωση:

$$L(\mu, \Omega) = -\frac{n}{2} k \log 2\pi + \frac{n}{2} \log|\Omega| - \frac{n}{2} tr(S\Omega) - \frac{n}{2} (\bar{x} - \mu)^T \Omega (\bar{x} - \mu) \qquad (1.11)$$

όπου,

- n είναι ο αριθμός των παρατηρήσεων,
- $|\Omega|$ υποδηλώνει τον πίνακα συγκέντρωσης Ω,
- S είναι ο πίνακας δειγματικής συνδιακύμανσης,

$$S = \frac{1}{n} \sum_{i=1}^{n} \left(x^i - \bar{x}\right) \left(x^i - \bar{x}\right)^T$$

και $\bar{x}$ είναι ο δειγματικός μέσος όρος,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x^i \,,$$

- $tr(S\Omega)$ υποδηλώνει το ίχνος του γινομένου των S και Ω.

Η μεγιστοποίηση της λογαριθμοποιημένης συνάρτησης πιθανοφάνειας προκύπτει στο

$$\hat{\mu} = \bar{x} \tag{1.12}$$

και καθώς οι πίνακες $\hat{\Sigma}$ και S έχουν διαφορετικές καταχωρίσεις όπου $\omega_{ij} = 0$, μπορούμε να συμπεράνουμε ότι

$$tr(\hat{\Omega}S) = tr(\hat{\Omega}\hat{\Sigma}) = k\,. \tag{1.13}$$

Με αντικατάσταση των εξισώσεων (1.12) και (1.13) στην εξίσωση (1.11), έχουμε ότι

$$\hat{l} = -\frac{n}{2}k\log 2\pi + \frac{n}{2}\log|\Omega| - \frac{n}{2}k \tag{1.14}$$

(Edwards, 2000).

1.3.3 Καταλληλότητα του Μοντέλου

Η απόκλιση ενός Γκαουσιανού γραφικού μοντέλου είναι μια μέτρηση της καλής εφαρμογής του μοντέλου. Υπολογίζεται ως η διαφορά ανάμεσα στην εκτίμηση της μεγιστοποιημένης λογαριθμοποιημένη συνάρτηση πιθανοφάνειας του κορεσμένου μοντέλου s και την εκτίμηση της αντίστοιχης συνάρτησης για ένα μοντέλο με λιγότερες μεταβλητές.

Η απόκλιση δίνεται από τη σχέση,

$$D = 2(\hat{l}_s - \hat{l}_m) \tag{1.15}$$

Χρησιμοποιώντας την εξίσωση (1.14), η εκτίμηση του λογάριθμου της συνάρτησης l του κορεσμένου μοντέλου είναι,

$$\hat{l}_s = -\frac{n}{2}k\log 2\pi - \frac{n}{2}\log|S| - \frac{n}{2}k \tag{1.16}$$

όπου $\log|\Omega| = -\log\left|\hat{\Sigma}\right| = -\log|S|$ το οοποίο είναι ο δειγματικός πίνακας συνδιακύμανσης των δεδομένων.

Η αντίστοιχη εκτίμηση για το μοντέλο με λιγότερες μεταβλητές είναι,

$$\hat{l}_m = -\frac{n}{2}k\log 2\pi - \frac{n}{2}\log\left|\hat{\Sigma}\right| - \frac{n}{2}k\,. \tag{1.17}$$

Συνδυάζοντας τις εξισώσεις (1.15), (1.16) και (1.17), έχουμε,

$$D = n\log\frac{\left|\hat{\Sigma}\right|}{|S|}\,. \tag{1.18}$$

1.4 Μεικτά μοντέλα αλληλεπίδρασης

Τα μεικτά μοντέλα αλληλεπίδρασης είναι στατιστικά μοντέλα που συνδυάζουν τόσο κατηγορικές όσο και συνεχείς μεταβλητές για να αποτυπώσουν περίπλοκες σχέσεις μεταξύ αυτών. Τα συγκεκριμένα μοντέλα αναπαρίστανται χρησιμοποιώντας μη κατευθυνόμενα γραφικά μοντέλα, τα οποία παρέχουν μια οπτική αναπαράσταση των κατηγορικών συσχετίσεων και των συνεχών εξαρτήσεων μεταξύ των μεταβλητών. Με την ενσωμάτωση κατηγορικών και συνεχών μεταβλητών, τα μεικτά μοντέλα αλληλεπίδρασης συνδυάζουν τα πλεονεκτήματα τόσο των log-linear γραφικών μοντέλων όσο και των Γκαουσιανών γραφικών μοντέλων.

1.4.1 Μεικτό Μοντέλο Αλληλεπίδρασης

Έστω $D = (D_1, D_2, \dots, D_d)$ να είναι d διακριτές μεταβλητές και $C = (C_1, C_2, \dots, C_c)$ να είναι c συνεχείς μεταβλητές ενός συνόλου δεδομένων V με N παρατηρήσεις. Επιπλέον, έστω ότι I είναι το σύνολο όλων των δυνατών συνδυασμών τιμών για τις διακριτές μεταβλητές. Ένα κελί του I συμβολίζεται ως $i = (i_1, i_2, \dots, i_d)$ και είναι μια παρατήρηση των τιμών των διακριτών μεταβλητών, όπου i_j είναι η τιμή της j-οστής μεταβλητής στο κελί. Μια ολόκληρη γραμμή στο σύνολο δεδομένων συμβολίζεται ως $(i, y) = (i_1, i_2, \dots, i_d, y_1, y_2, \dots, y_c)$. Επιπλέον, η πολυμεταβλητή γκαουσιανή κατανομή $N(\mu_i, \Sigma)$ αναπαριστά την συνθήκη διανομή των συνεχών μεταβλητών C, όταν οι διακριτές μεταβλητές D περιορίζονται να πέφτουν σε ένα συγκεκριμένο κελί i.

Η συνάρτηση πυκνότητας, επίσης γνωστή ως συνάρτηση Conditional Gaussian density, μπορεί να εκφραστεί ως

$$f(i, y) = p_i (2\pi)^{-\frac{c}{2}} \det(\Sigma)^{-\frac{1}{2}} exp\left(-\frac{1}{2}\,(y - \mu_i)^T \Sigma^{-1} (y - \mu_i)\right) \qquad (1.19)$$

όπου,

- p_i είναι η πιθανότητα των διακριτών μεταβλητών να πέσουν στο κελί i,
- $(2\pi)^{-\frac{1}{2}}$ είναι μια σταθερά κανονικοποίησης,
- y αναπαριστά το διάνυσμα των συνεχών μεταβλητών,
- μ_i αναπαριστά το μέσο διάνυσμα που συσχετίζεται με το κελί i. Υποδηλώνει τις αναμενόμενες τιμές των συνεχών μεταβλητών δοθέντων των διακριτών μεταβλητών που πέφτουν σε εκείνο το κελί, και Σ αναπαριστά τον πίνακα συνδιακύμανσης,

(Wermuth & Lautitzen, 1990).

Μία σημαντική παρατήρηση είναι ότι, αντίθετα από τη μέση τιμή μ_i μιας πολυμεταβλητής Γκαουσιανής κατανομής, ο πίνακας συνδιακύμανσης Σ παραμένει σταθερός και δεν ποικίλλει με σε διαφορετικές τιμές των διακριτών μεταβλητών. Αυτή η ιδιότητα χαρακτηρίζει αυτά τα μοντέλα ως ομοιογενή.

Επίσης, η συνάρτηση πυκνότητας της εξίσωσης (1.19) μπορεί να εκφραστεί με τη χρήση της εξίσωσης (1.20). Οι παράμετροι (p_i, μ_i, Σ) που παρουσιάζονται στην εξίσωση (1.19) αναφέρονται ως στιγμιαίοι παράμετροι (moment parameters), και μπορούν να μετατραπούν σε κανονικές παραμέτρους (canonical parameters). Συμβολίζονται ως εξής (g_i, h_i, K) και ισχύουν οι ακόλουθες σχέσεις.

$$f(i, y) = exp\left(g_i + h_i^T y - \frac{1}{2} y^T K y\right) \tag{1.20}$$

$$g_i = \log(p_i) - \frac{c}{2}\log 2\pi - \frac{1}{2}\log\det\Sigma - \frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i \tag{1.21}$$

$$h_i = \Sigma^{-1}\mu_i \tag{1.22}$$

$$K = \Sigma^{-1} \tag{1.23}$$

1.4.2 Συνάρτηση Πιθανοφάνειας

Σε ένα μοντέλο μεικτών αλληλεπιδράσεων, η συνάρτηση πιθανοφάνειας αντιπροσωπεύει την πιθανότητα παρατήρησης των δεδομένων υπό το υποθετικό μοντέλο. Η μέθοδος εκτίμησης μέγιστης πιθανοφάνειας χρησιμοποιείται για να την εκτίμηση των στιγμιαίων παραμέτρων (p_i, μ_i, Σ), οι οποίες χαρακτηρίζουν το μοντέλο. Αυτές οι παράμετροι προσφέρουν την καλύτερη εξήγηση για τα παρατηρούμενα δεδομένα. Η διαδικασία της MLE λαμβάνει υπόψη και τις κατηγορικές και τις συνεχείς μεταβλητές, αποτυπώνοντας αποτελεσματικά τις πολύπλοκες αλληλεπιδράσεις και εξαρτήσεις τους.

Χρησιμοποιώντας την εξίσωση (1.19), η συνάρτηση λογαρίθμου της συνάρτησης πιθανοφάνειας δίνεται από την εξίσωση,

$$\log f(i, y) = \log p_i - \frac{c}{2}\log 2\pi - \frac{1}{2}\log\det\Sigma - \frac{1}{2}\left((y - \mu_i)^T \Sigma^{-1}(y - \mu_i)\right) \tag{1.25}$$

Η μεγιστοποίησή της γίνεται όταν ισχύουν οι εξισώσεις $(1.26), (1.27), (1.28)$.

$$\hat{p} = \frac{n_i}{N} \tag{1.26}$$

$$\hat{\mu}_i = \bar{y}_i \tag{1.27}$$

$$\hat{\Sigma} = S = \sum_i \frac{n_i S_i}{N} \tag{1.28}$$

Απλοποιώντας και αντικαθιστώντας τις εξισώσεις αυτές στην εξίσωση (1.25), προκύπτει ότι

$$\hat{l} = \sum_i n_i \log\left(\frac{n_i}{N}\right) - \frac{1}{2} Nc\ log2\pi - \frac{1}{2} N \log\det \Sigma - \frac{1}{2} Nc \tag{1.29}$$

(Højsgaard et al., 2012).

1.4.3 Καταλληλότητα του Μοντέλου

Η απόκλιση ενός μοντέλου μεικτών αλληλεπιδράσεων είναι μια μέτρηση της προσαρμογής του μοντέλου στα δεδομένα. Υπολογίζεται ως η διαφορά μεταξύ της εκτίμησης της μεγιστοποιημένης λογαριθμοποιημένης συνάρτησης πιθανοφάνειας του κορεσμένου μοντέλου s και της εκτίμησης της μεγιστοποιημένης λογαριθμοποιημένης συνάρτησης πιθανοφάνειας ενός "μειωμένου" μοντέλου m. Η απόκλιση είναι σημαντική στην επιλογή και αξιολόγηση των μοντέλων μέσα στο πλαίσιο των μοντέλων μεικτών αλληλεπιδράσεων. Επιθυμητή είναι μια χαμηλή τιμή απόκλισης, καθώς υποδεικνύει μια καλύτερη προσαρμογή του μοντέλου στα δεδομένα.

Ο τύπος για την απόκλιση ενός μοντέλου μεικτών αλληλεπιδράσεων δίνεται από τη σχέση,

$$D = 2\left(\hat{l}_s - \hat{l}_m\right) \tag{1.30}$$

Χρησιμοποιώντας την εξίσωση (1.29), η τύπος της απόκλισης (1.30) παίρνει τη μορφή,

$$D = 2\sum_i n_i \log\left(\frac{n_i}{\hat{m}_i}\right) - N\log\det\left(S\hat{\Sigma}^{-1}\right) + N\left(tr\left(S\hat{\Sigma}^{-1}\right) - c\right)$$
$$+ \sum_i n_i(\bar{y}_i - \hat{\mu}_i)^T \hat{\Sigma}^{-1}(\bar{y}_i - \hat{\mu}_i) \qquad (1.31)$$

(Højsgaard et al., 2012).

Κεφάλαιο 2- Μοντέλα Κατευθυνόμενων Γραφημάτων

Σε αυτό το κεφάλαιο, θα κάνουμε μια εισαγωγή στα μοντέλα γραφημάτων αλυσίδας, ένα σημαντικό υποσύνολο των κατευθυνόμενων γραφημάτων. Τα κατευθυνόμενα γραφήματα αποτελούν εργαλεία τόσο για την αναπαράσταση, όσο και την ανάλυση πολύπλοκων εξαρτήσεων μεταξύ μεταβλητών. Συγκεκριμένα, τα μοντέλα γραφημάτων αλυσίδας βρίσκουν εφαρμογή σε περιπτώσεις όπου οι μεταβλητές μπορούν να ομαδοποιηθούν σε διακριτά κουτιά. Το χαρακτηριστικό διαφοροποίησης από άλλα μοντέλα είναι η εξάρτησή τους από την ταξινόμηση μεταξύ των διακριτών κουτιών των μεταβλητών, ενώ παράλληλα δεν επιτρέπουν την ταξινόμηση μέσα σε κάθε ένα από αυτά τα κουτιά.

2.1 Μοντέλα Αλυσιδωτών Γραφημάτων

Τα μοντέλα γραφημάτων αλυσίδας είναι γραφικά μοντέλα που αναπαριστούν τις υπό όρους εξαρτήσεις μεταξύ των μεταβλητών χρησιμοποιώντας μια δομή γράφου. Αυτά τα μοντέλα επιτρέπουν τόσο κατευθυνόμενες όσο και μη κατευθυνόμενες ακμές, αντιπροσωπεύοντας έτσι ένα ευέλικτο εύρος σχέσεων. Ένα επιπλέον χαρακτηριστικό αυτών των μοντέλων είναι ότι μπορούν να χρησιμοποιηθούν όταν το δεδομένο σύνολο περιλαμβάνει τόσο συνεχείς όσο και διακριτές μεταβλητές. Έχουν διάφορες εφαρμογές σε πεδία όπως η γενετική και οι κοινωνικές επιστήμες.

2.1.1 Μοντέλο Αλυσιδωτού Γραφήματος

Παρόμοια με τα μη κατευθυνόμενα γραφικά μοντέλα, κάθε μεταβλητή του μοντέλου αναπαρίσταται από έναν κόμβο. Οι συνδέσεις μεταξύ των κόμβων αναπαρίστανται από ακμές που παρέχουν μια καθαρή οπτικοποίηση των υπό όρων εξαρτήσεων μεταξύ των μεταβλητών. Αναλύοντας και ερμηνεύοντας τις ακμές, μπορούμε να αποκτήσουμε γνώση σχετικά με την πιθανή επίδραση μιας μεταβλητής σε μια άλλη.

Για να κατασκευάσουμε ένα γράφημα αλυσίδας, χρησιμοποιούμε έναν μηχανισμό ομαδοποίησης σε κουτιά. Οι μεταβλητές χωρίζονται σε κουτιά βάσει του ρόλου τους στο μοντέλο. Εντός κάθε κουτιού, οι μεταβλητές θεωρούνται ταυτόχρονες και κάθε μεταβλητή υποβάλλεται σε παλινδρόμηση με τις μεταβλητές που ανήκουν σε όλα τα κουτιά που βρίσκονται δεξιά της.

Επιπλέον, το κουτί που βρίσκεται στα αριστερά στο γράφημα αλυσίδας περιέχει αποκλειστικά τις μεταβλητές απόκρισης, ενώ το κουτί που βρίσκεται στα δεξιά περιέχει αποκλειστικά τις επεξηγηματικές μεταβλητές. Οι υπόλοιπες μεταβλητές ανήκουν στα ενδιάμεσα κουτιά.

Είναι σημαντικό να αναφέρουμε ότι χρήσιμες πληροφορίες παρέχονται όχι μόνο από τις παρούσες, αλλά και από τις απουσιάζουσες ακμές. Η απουσία μιας ακμής μεταξύ δύο μεταβλητών υποδηλώνει την υπό όρο ανεξαρτησία τους.

2.1.2 Συνάρτηση Πυκνότητας

Σε ένα γράφημα αλυσίδας, η κοινή συνάρτηση πυκνότητας παραγοντοποιείται βάσει των συνδεδεμένων τμημάτων (components) του γράφου. Τα συνδεδεμένα τμήματα είναι τα μικρότερα κομμάτια του γράφου που παραμένουν μετά την εξάλειψη όλων των κατευθυνόμενων ακμών του αρχικού γράφου.

Ας θεωρήσουμε ένα γράφημα ενός αλυσιδωτού γραφήματος και ένα σύνολο από τα συνδεδεμένα τμήματα του $g_1, g_2, \dots, g_n$. Η συνάρτηση πυκνότητας δίνεται από τον παρακάτω τύπο,

$$f = \prod_{i=1}^{n} f_{g_i|g>i}$$

όπου $f_{g_i|g>i}$ αναπαριστά την υπό όρο συνάρτηση πυκνότητας των μεταβλητών στο συνδεδεμένο τμήμα g_i δεδομένου όλων των προηγούμενων συνδεδεμένων τμημάτων g_j, για $j > i$.

(Wermuth & Sadeghi, 2012).

Αξίζει να τονίσουμε ότι η μορφή της συνάρτησης αυτής εξαρτάται από τον τύπο των μεταβλητών που συμμετέχουν στο μοντέλο και σε κάθε συνδεδεμένο τμήμα (συνεχείς, διακριτές), καθώς και από τις υποθέσεις μοντελοποίησης (Cox & Wermuth, 1993).

Κεφάλαιο 3- Ανάλυση Δεδομένων

Σε αυτό το κεφάλαιο, πραγματοποιούμε στατιστική ανάλυση δεδομένων χρησιμοποιώντας την R. Επικεντρωνόμαστε σε ένα συγκεκριμένο σύνολο δεδομένων και το χρησιμοποιούμε για να εφαρμόσουμε τρεις διαφορετικές τεχνικές για τη μελέτη και σύγκριση των αποτελεσμάτων τους. Αρχικά, εφαρμόζεται ένα Γκαουσιανό γραφικό μοντέλο στο σύνολο δεδομένων, στη συνέχεια χρησιμοποιείται ένα μοντέλο μικτής αλληλεπίδρασης, και τέλος πραγματοποιούμε στατιστική ανάλυση χρησιμοποιώντας ένα γραφικό μοντέλο αλυσίδας.

3.1 Παρουσίαση του Συνόλου Δεδομένων

Σε αυτήν την εργασία, πραγματοποιήσαμε μια ανάλυση των δεδομένων που προήλθαν από μια μελέτη παρακολούθησης ασθενών που νοσηλεύτηκαν για COVID-19 στην Ιταλία. Αυτή η έρευνα πραγματοποιήθηκε το έτος 2021. Το σύνολο δεδομένων προέρχεται από το ερευνητικό πρόγραμμα με τίτλο " Multidimensional statistical analysis of databases relating to patients affected by "long-Covid" in order to characterize their manifestations, identify their risk factors and identify their therapeutic trajectories.", το οποίο χρηματοδοτήθηκε από το Πανεπιστήμιο της Περούτζια, στην Ιταλία. Ο κύριος στόχος αυτής της μελέτης είναι η διερεύνηση και η απόκτηση εμβέλειας στους μακροπρόθεσμους προγνωστικούς παράγοντες των συνεπειών στην υγεία αυτής της ασθένειας. Το σύνολο δεδομένων αποτελείται από 108 παρατηρήσεις και 106 μεταβλητές.

Η έρευνα περιλάμβανε διάφορους τύπους δεδομένων. Ενδεικτικά αναφέρουμε ορισμένους,

- Δημογραφικά δεδομένα: Ηλικία, φύλο
- Ανθρωπομετρικά δεδομένα: Ύψος, βάρος, περίμετρος σώματος (μπράτσο, μέση, γοφός, λαιμός)
- Συμπτώματα: Κόπωση, εξάνθημα, ύπνος, διάρροια κλπ.
- Δεδομένα νοσηλείας: Εισαγωγή, επίσκεψη εξέτασης, αναπνευστική υποστήριξη κλπ.
- Συνήθειες: Κάπνισμα
- Δεδομένα εργαστηρίου: Αίμα (Φερριτίνη, DDIMER, φιμπρινογόνο κλπ.)
- Λειτουργία πνεύμονα και καρδιάς: Αξονική τομογραφία, σπιρομετρία, ηχοκαρδιογράφημα κλπ.
- Ψυχολογικά δεδομένα: άγχος, κατάθλιψη κλπ.

Λόγω του μεγάλου πλήθους μεταβλητών σε σχέση με το πλήθος των παρατηρήσεων, πραγματοποιήθηκε επιλογή μεταβλητών ώστε να επικεντρωθούμε στους πιο σημαντικούς παράγοντες. Δώσαμε προτεραιότητα στις πιο σχετικές μεταβλητές από τις παραπάνω κατηγορίες δεδομένων. Οι επιλεγμένες μεταβλητές παρουσιάζονται στον ακόλουθο πίνακα.

Επεξηγηματική Μεταβλητή	Τύπος	Περιγραφή	Πηγή Δεδομένων
BMI	Συνεχής	Δείκτης μάζας σώματος $\left(\frac{kg}{m^2}\right)$	Ανθρωπομετρικά δεδομένα
TIME_UNTIL_FOLUP	Συνεχής	Διάρκεια μεταξύ εξιτηρίου και επίσκεψης (ημέρες)	Δεδομένα νοσηλείας
AGE	Συνεχής	Ηλικία (χρόνια)	Δημογραφικά δεδομένα
GENDER	Κατηγορική	1: Γυναίκα 2: Άνδρας	Δημογραφικά δεδομένα
SMOKING	Κατηγορική	1: Μη καπνιστής 2: Καπνιστής	Δεδομένα συνηθειών υγείας
Μεταβλητή Απόκρισης	Τύπος	Περιγραφή	Πηγή Δεδομένων
VE.VCO2.SLOPE	Συνεχής	Αναλογία του συνολικού εισπνεόμενου αέρα προς την παραγωγή διοξειδίου του άνθρακα	Καρδιοαναπνευστικός έλεγχος
VO2	Συνεχής	Μέγιστη οξυγόνωση κατανάλωσης	Καρδιοαναπνευστικός έλεγχος
LT	Συνεχής	Κατώτατο όριο γαλακτικού οξέος	Καρδιοαναπνευστικός έλεγχος
FEV1	Συνεχής	Μέγιστος όγκος αέρα που εκπνέεται σε 1'' μετά από μέγιστη εισπνοή	Σπιρομετρίας
FERRITIN	Συνεχής	Πρωτεΐνη αποθήκευσης σιδήρου στα κύτταρα	Laboratory data
DDIMER	Συνεχής	Blood clotting test	Εργαστηριακά δεδομένα
IES.R.TOTAL	Συνεχής	Αυτοαξιολόγηση της σοβαρότητας των συμπτωμάτων μετά από τραυματικό στρες με την κλίμακα Impact of Event Scale-Revised	Ψυχολογικά δεδομένα
DEPRESSION	Κατηγορική	Αυτοαξιολόγηση 1: Κανονική 2: Μέτρια 3: Σοβαρή	Ψυχολογικά δεδομένα
ICU	Κατηγορική	Μονάδα εντατικής θεραπείας 1: Όχι 2: Ναι	Δεδομένα νοσηλείας
FATIGUE	Κατηγορική	Αυτοαξιολόγηση 1: Όχι 2: Ναι	Συμπτώματα
HRTC	Κατηγορική	High Resolution Computed Tomography 1: Κανονικές 2: Ευρήματα 3: Σοβαρά ευρήματα	CT scan
HRTC_2cat	Κατηγορική	1: Συγχώνευση κατηγοριών HRTC 2: Κενές τιμές HRTC	CT scan

Table 41: Δεδομένα προς ανάλυση

3.2 Προεπεξεργασία Δεδομένων

3.2.1 Δημιουργία Νέων Μεταβλητών

Χρησιμοποιώντας τη γλώσσα προγραμματισμού R και το αρχικό δεδομένων που μας δόθηκε, δημιουργήσαμε τις εξηγηματικές μεταβλητές AGE, BMI και $TIME\_UNTIL\_FOLUP$, καθώς και την αποκριτική μεταβλητή $HRTC\_2cat$. Αυτές οι μεταβλητές δημιουργήθηκαν επειδή συχνά φαίνεται ότι είναι σημαντικές σε διάφορες έρευνες.

Η μεταβλητή AGE προκύπτει από τον υπολογισμό της διαφοράς μεταξύ της ημερομηνίας εισαγωγής του ασθενούς στο νοσοκομείο και της ημερομηνίας γέννησής του.

Για τη μεταβλητή BMI, χρησιμοποιήσαμε τα διαθέσιμα δεδομένα για το βάρος και το ύψος των ασθενών στον τύπο,

$$\frac{Weight}{Height^2} \quad \left(\frac{kg}{m^2}\right)$$

Η μεταβλητή $TIME\_UNTIL\_FOLUP$ δημιουργήθηκε από τον υπολογισμό του χρονικού διαστήματος μεταξύ της ημερομηνίας εισαγωγής στο νοσοκομείο και της επόμενης ημερομηνίας επανεξέτασης.

Η μεταβλητή απόκρισης $HRTC\_2cat$ προέκυψε από τη μεταβλητή $HRTC$. Συγχωνεύοντας τις τρεις κατηγορίες της $HRTC$ σε μία, δημιουργήθηκε η πρώτη κατηγορία της $HRTC\_2cat$, ενώ η δεύτερη κατηγορία της $HRTC\_2cat$ περιλαμβάνει όλες τις απουσιάζουσες τιμές της $HRTC$.

3.3 Μη κατευθυνόμενα Γκαουσιανά Γραφικά Μοντέλα

Ξεκινάμε εισάγοντας το σύνολο δεδομένων στο περιβάλλον της R. Αυτό γίνεται με τη βοήθεια της συνάρτησης $read.xlsx()$. Συνεχίζουμε με τη ανάλυση και τη μελέτη των σχέσεων μεταξύ των συνεχών μεταβλητών με τη χρήση ενός Γκαουσιανού γραφικού μοντέλου. Χρησιμοποιήσαμε τη συνάρτηση $cmod()$, όπου το c αναφέρεται στις συνεχείς μεταβλητές στο μοντέλο. Αυτή η συνάρτηση προέρχεται από το πακέτο $gRim$ της R.

Στη συνέχεια, επιλέγουμε μοντέλο βάσει του κριτηρίου BIC, χρησιμοποιώντας τη συνάρτηση $stepwise()$. Σε αυτήν την περίπτωση, προτιμούμε το BIC αντί του AIC, καθώς ευνοεί ένα πιο απλό μοντέλο και, συνεπώς, διευκολύνει τόσο την ερμηνεία όσο και την αναπαράσταση του αποτελέσματος. Τα αποτελέσματα παρουσιάζονται στους $Πίνακες\ 42, 43$. Όπως φαίνεται, το πλήρες μοντέλο έχει συνολικά 45 πιθανές ακμές και από αυτές μόνο 18 υπάρχουν στο τελικό μοντέλο που προκύπτει από το BIC. Επιπλέον, συγκρίνοντας τα δύο μοντέλα, παρατηρούμε ότι και οι δύο τιμές του AIC και του BIC μειώνονται.

Κορεσμένο Μοντέλο: 10 συνεχείς μεταβλητές					
-2logL	6919.73	mdim	55	**AIC**	**7029.73**
ideviance	180.85	**idf**	**45**	**BIC**	**7170.77**
deviance	-0.00	df	0		

Table 42: UGGM - Κορεσμένο Μοντέλο

Τελικό Μοντέλο: 10 συνεχείς μεταβλητές					
-2logL	6941.31	mdim	28	**AIC**	**6997.31**
ideviance	159.28	**idf**	**18**	**BIC**	**7069.11**
deviance	21.58	df	27		

Table 43: UGGM - Τελικό Μοντέλο

Τέλος, εξετάζουμε τις σχέσεις μεταξύ των μεταβλητών υπολογίζοντας τον πίνακα μερικής συσχέτισης μέσω του πίνακα συνδιακύμανσης με τη βοήθεια της συνάρτησης $cov2pcor()$. ($Πίνακα\ 44$). Τα πρόσημα στις ακμές στο $Σχήμα\ 13$ προέρχονται από τον $Πίνακα\ 44$.

Μεταβλητή	AGE	BMI	TIME_ UNTL_ FOLUP	DDIMER	FERRITIN	FEV1	VO2	VE. VCO2. SLOPE	LT	IES. R. TOTAL
AGE	1.00	-0.30	-0.01	0.33	-0.05	-0.30	-0.34	0.02	0.36	-0.22
BMI	-0.30	1.00	0.07	0.17	-0.01	0.08	-0.53	-0.09	0.23	0.10
TIME_ UNTIL_ FOLUP	-0.01	0.07	1.00	-0.10	0.05	0.08	0.09	-0.03	0.16	-0.09
DDIMER	0.33	0.17	-0.10	1.00	0.11	0.09	0.06	-0.16	-0.07	0.01
FERRITIN	-0.05	-0.01	0.05	0.11	1.00	0.02	0.00	0.24	-0.02	-0.18
FEV1	-0.30	0.08	0.08	0.09	0.02	1.00	0.33	-0.14	-0.27	-0.24
VO2	-0.34	-0.53	0.09	0.06	0.00	0.33	1.00	-0.12	0.41	0.05
VE. VCO2. SLOPE	0.02	-0.09	-0.03	-0.16	0.24	-0.14	-0.12	1.00	-0.18	0.23
LT	0.36	0.23	0.16	-0.07	-0.02	-0.27	0.41	-0.18	1.00	0.26
IES. R. TOTAL	-0.22	0.10	-0.09	0.01	-0.18	-0.24	0.05	0.23	0.26	1.00

Table 44: UGGM - Πίνακας μερικής συσχέτισης

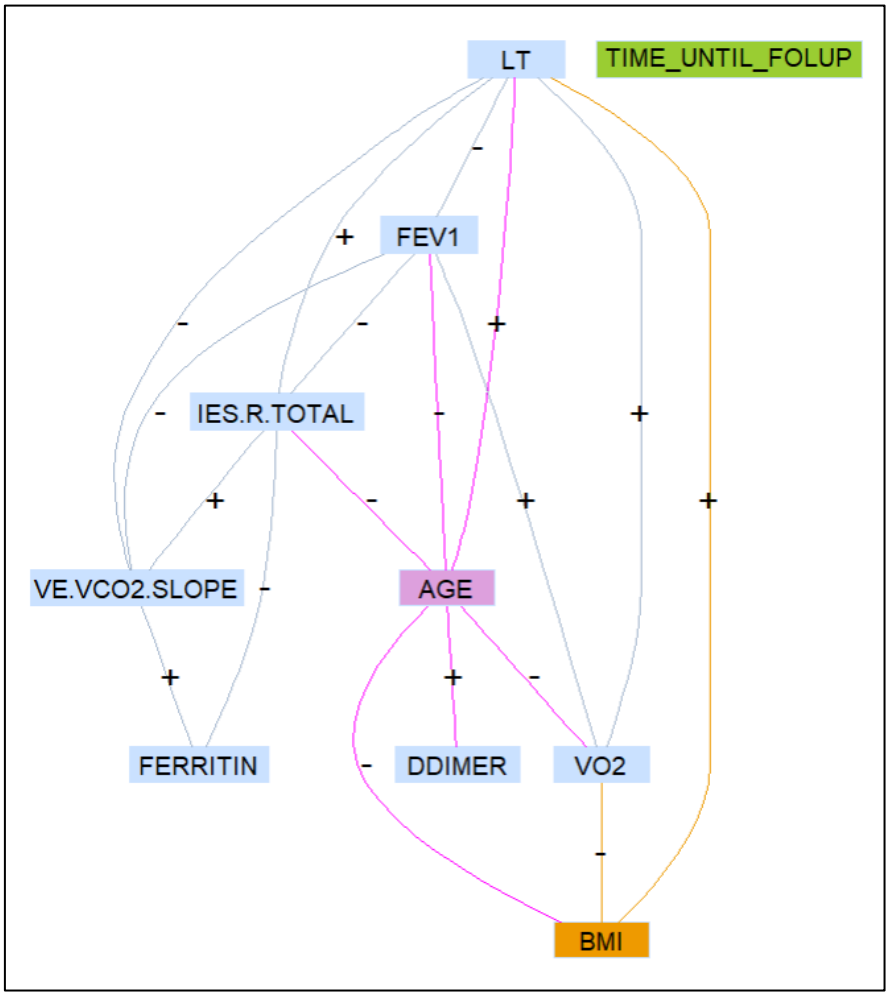


Figure 13: Μη κατευθυνόμενο Γκαουσιανό γραφικό μοντέλο

3.4 Μοντέλο Μικτής Αλληλεπίδρασης

Ξεκινάμε εισάγοντας στο περιβάλλον της R το σύνολο δεδομένων. Αυτό γίνεται με τη βοήθεια της συνάρτησης $read.xlsx()$. Ένα μοντέλο μικτής αλληλεπίδρασης εφαρμόζεται για την ανάλυση των δεδομένων, χρησιμοποιώντας τη συνάρτηση $mmod()$, όπου το m σημαίνει ότι έχουμε μικτές μεταβλητές (συνεχείς και διακριτές) στο μοντέλο. Αυτή η συνάρτηση προέρχεται από το πακέτο $gRim$ της R.

Στη συνέχεια, εκτελούμε την επιλογή μοντέλου, η οποία βασίζεται στο κριτήριο AIC, χρησιμοποιώντας τη συνάρτηση $stepwise()$. Ο $Πίνακας\ 45$ δείχνει τη σύνοψη του κορεσμένου μοντέλου, ενώ ο $Πίνακας\ 46$ παρουσιάζει τη σύνοψη του τελικού μοντέλου μετά τη χρήση του AIC. Όπως φαίνεται, το κορεσμένο μοντέλο έχει 76 πιθανές ακμές συνολικά και από αυτές μόνο 19 υπάρχουν στο τελικό μοντέλο. Επιπλέον, συγκρίνοντας τα δύο μοντέλα παρατηρούμε ότι και οι δύο τιμές του AIC και του BIC μειώνονται.

Κορεσμένο Μοντέλο: 10 συνεχείς μεταβλητές					
-2logL	4750.51	mdim	98	**AIC**	**4946.51**
ideviance	170.14	**idf**	**76**	**BIC**	**5197.82**
deviance	-0.00	df	0		

Table 45: MIM - Κορεσμένο Μοντέλο

Τελικό Μοντέλο: 10 συνεχείς μεταβλητές					
-2logL	4828.59	mdim	41	**AIC**	**4910.59**
ideviance	131.11	**idf**	**19**	**BIC**	**5015.72**
deviance	39.04	df	57		

Table 46: MIM - Τελικό Μοντέλο

Συνεχίζουμε την ανάλυση μελετώντας τις σχέσεις μεταξύ των μεταβλητών χρησιμοποιώντας τις παραμέτρους ροπής (p_i, μ_i, Σ). Αρχικά, διερευνούμε τις συσχετίσεις μεταξύ των αριθμητικών μεταβλητών χρησιμοποιώντας τον πίνακα μερικής συσχέτισης που προέρχεται από τον πίνακα συνδιακύμανσης Σ ($Πίνακας\ 47$). Έπειτα, εξετάζουμε τις σχέσεις μεταξύ των αριθμητικών και των δυαδικών μεταβλητών, υπολογίζοντας τη διαφορά μεταξύ των μέσων όρων των ομάδων ($Πίνακας\ 48$). Τέλος, για να μελετήσουμε την αλληλεπίδραση των δύο δυαδικών μεταβλητών σε κάθε αριθμητική μεταβλητή, συγκρίνουμε τη διαφορά των μέσων όρων μεταξύ των ομάδων όταν η μία από τις δύο δυαδικές μεταβλητές παραμένει σταθερή και η άλλη αλλάζει το επίπεδό της ($Πίνακας\ 48$). Τα πρόσημα σε κάθε ακμή προσδίδονται σύμφωνα με τους $Πίνακες\ 47, 48$.

Μεταβλητή	AGE	BMI	TIME_ UNTL_ FOLUP	VO2	FEV1	IES. R. TOTAL	VE. VCO2. SLOPE	LT	DDIMER	FERRITIN
AGE	1.00	-0.25	0.00	-0.47	-0.42	0.00	0.00	0.44	0.32	0.00
BMI	-0.25	1.00	0.00	-0.41	0.00	0.00	0.00	0.22	0.00	0.00
TIME_ UNTIL_ FOLUP	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VO2	-0.47	-0.41	0.00	1.00	0.00	0.00	-0.14	0.59	0.00	0.00
FEV1	-0.42	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
IES. R. TOTAL	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
VE. VCO2. SLOPE	0.00	0.00	0.00	-0.14	0.00	0.00	1.00	0.00	0.00	0.00
LT	0.44	0.22	0.00	0.59	0.00	0.00	0.00	1.00	0.00	0.00
DDIMER	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
FERRITIN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Table 47: MIM - Πίνακας Μερικής Συσχέτισης

Μεταβλητή	G12S1[1]	G12S2[2]	G1S12[3]	G2S12[4]	inter_ effect[5]
AGE	-2.09	-1.85	3.53	3.77	0.24
BMI	-4.34	-4.35	-0.21	-0.22	-0.01
TIME_ UNTIL_ FOLUP	0.00	0.00	0.00	0.00	0.00
VO2	5.42	5.38	-0.52	-0.56	-0.04
FEV1	0.94	0.93	-0.12	-0.13	-0.01
IES. R. TOTAL	-18.51	-18.51	0.00	0.00	0.00
VE. VCO2. SLOPE	-1.29	-1.29	0.13	0.13	0.00
LT	-0.07	-0.07	0.01	0.01	0.00
DDIMER	-0.06	-0.05	-0.17	-0.16	0.01
FERRITIN	1.00	1.00	0.00	0.00	0.00

Table 48: MIM - Διαφορά μέσων όρων και αλληλεπίδραση

Επεξήγηση στηλών $Πίνακα\ 48$:

[1] Η διαφορά στην έκβαση μεταξύ ατόμων με Φύλο=2 (Άνδρας) και Φύλο=1 (Γυναίκα), που κρατούν σταθερό το κάπνισμα στο επίπεδο 1.

[2] Η διαφορά στην έκβαση μεταβλητή μεταξύ ατόμων με Φύλο=2 (Άνδρας) και Φύλο=1 (Γυναίκα), που διατηρούν το Κάπνισμα σταθερό στο επίπεδο 2.

[3] Η διαφορά στην έκβαση μεταξύ ατόμων με Κάπνισμα=2 (Καπνιστής) και Κάπνισμα=1 (Μη καπνιστής), που κρατούν το Φύλο σταθερό στο 1 επίπεδο.

[4] Η διαφορά στην έκβαση μεταξύ ατόμων με Κάπνισμα=2 (Καπνιστής) και Κάπνισμα=1 (Μη Καπνιστής), κρατώντας το Φύλο σταθερό στο επίπεδο 2.

[5] Η επίδραση αλληλεπίδρασης Φύλου και Καπνίσματος στην αριθμητική μεταβλητή.

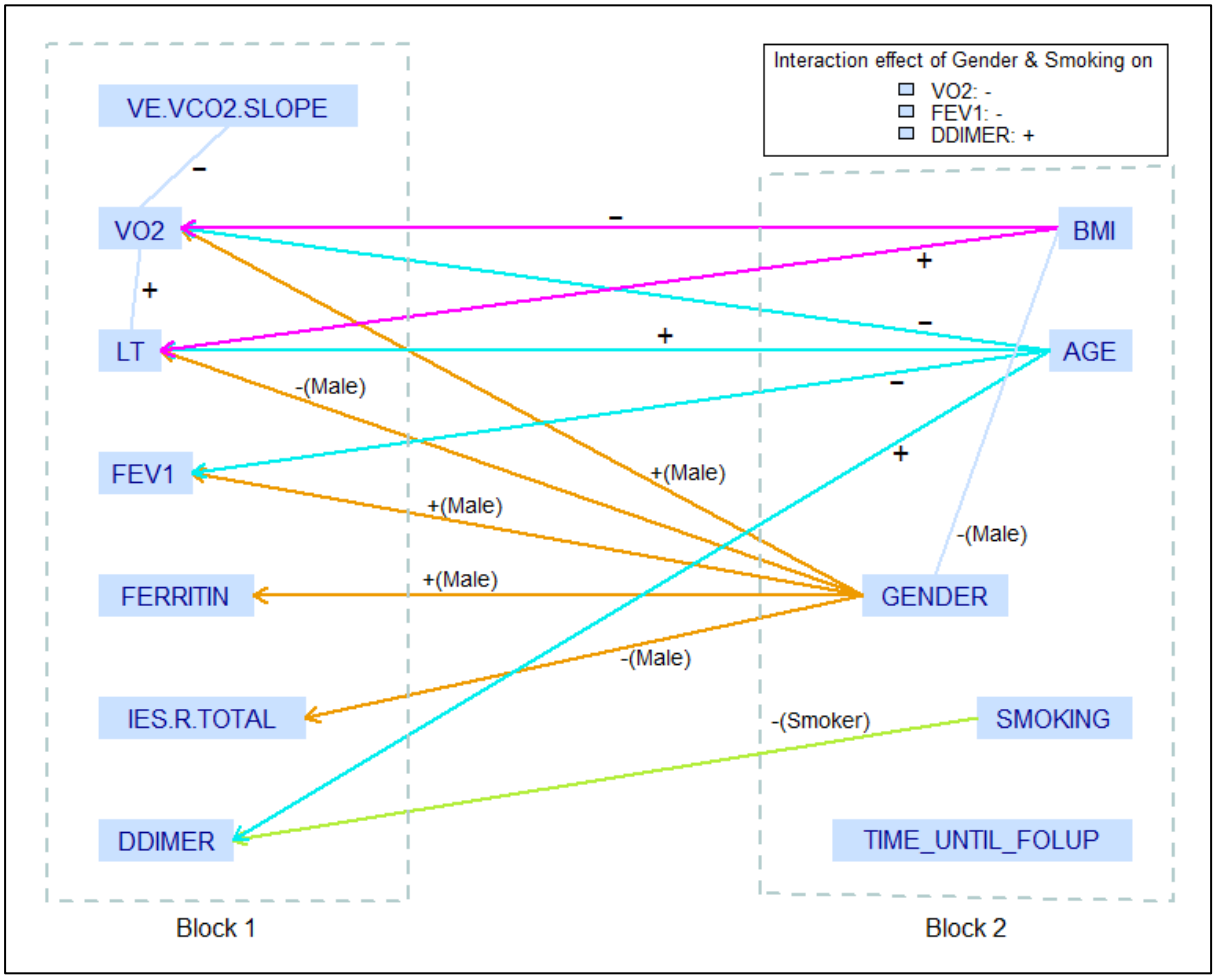


Figure 14: Μοντέλο μεικτής αλληλεπίδρασης

Το μοντέλο μικτής αλληλεπίδρασης που παρουσιάζεται στο $Σχήμα\ 14$ αποτελείται από δύο κουτιά. Το πρώτο κουτί περιλαμβάνει τις μεταβλητές απόκρισης, ενώ το δεύτερο κουτί περιλαμβάνει τις επεξηγηματικές μεταβλητές.

3.5 Μοντέλο γραφήματος αλυσίδας

Ξεκινάμε εισάγοντας στην R το σύνολο δεδομένων. Αυτό γίνεται με τη βοήθεια της συνάρτησης $read.xlsx()$. Ένα μοντέλο γραφήματος αλυσίδας εφαρμόζεται για την ανάλυση των δεδομένων χρησιμοποιώντας τη συνάρτηση $coxwer()$ από το πακέτο $gRchain$ της R.

Στη συνέχεια, εκτελούμε την επιλογή μοντέλου, η οποία βασίζεται στο κριτήριο BIC, χρησιμοποιώντας τη συνάρτηση $stepwise()$. Το πλεονέκτημα του BIC έναντι της AIC είναι ότι ευνοεί ένα απλούστερο μοντέλο.

Ένας διαφορετικός τύπος μοντέλου προσαρμόζεται για κάθε μεταβλητή με βάση τη φύση της.

- Ένα ordinary least squares μοντέλο χρησιμοποιείται για κάθε συνεχή μεταβλητή.
- Ένα binomial logit μοντέλο χρησιμοποιείται για κάθε δυαδική μεταβλητή.
- Ένα proportional odds logit μοντέλο χρησιμοποιείται για διατεταγμένες κατηγορικές μεταβλητές.

Τα πρόσημα που αποδίδονται σε κάθε ακμή στο $Σχήμα\ 15$ προέρχονται από τους συντελεστές που λαμβάνονται από τα αντίστοιχα προσαρμοσμένα μοντέλα.

Στο $Σχήμα\ 15$, είναι σημαντικό να εξηγήσουμε τη σημασία κάθε κουτιού.

Ένα μοντέλο γραφήματος αλυσίδας προσαρμόζεται στις μεταβλητές που χωρίζονται στα κουτιά με ετικέτες 2, 3 και 4 βάσει τους αντίστοιχους ρόλους τους στο μοντέλο. Ειδικότερα, το κουτί 4 αποτελείται από τις βασικές επεξηγηματικές μεταβλητές, το κουτί 3 περιλαμβάνει τις ενδιάμεσες μεταβλητές και το κουτί 2 περιέχει τις μεταβλητές απόκρισης.

Όσον αφορά το κουτί 1, αυτό αποτελείται από τη $HRTC$. Αυτή η μεταβλητή αντιστοιχεί στην αρχική μεταβλητή $HRTC$ από το σύνολο δεδομένων, αλλά περιλαμβάνει μόνο τις τρεις κατηγορίες της, έχοντας αφαιρεθεί όλες οι απουσιάζουσες τιμές. Η συσχέτισή της με τις άλλες μεταβλητές εξετάστηκε χρησιμοποιώντας το proportional odds logit μοντέλο.

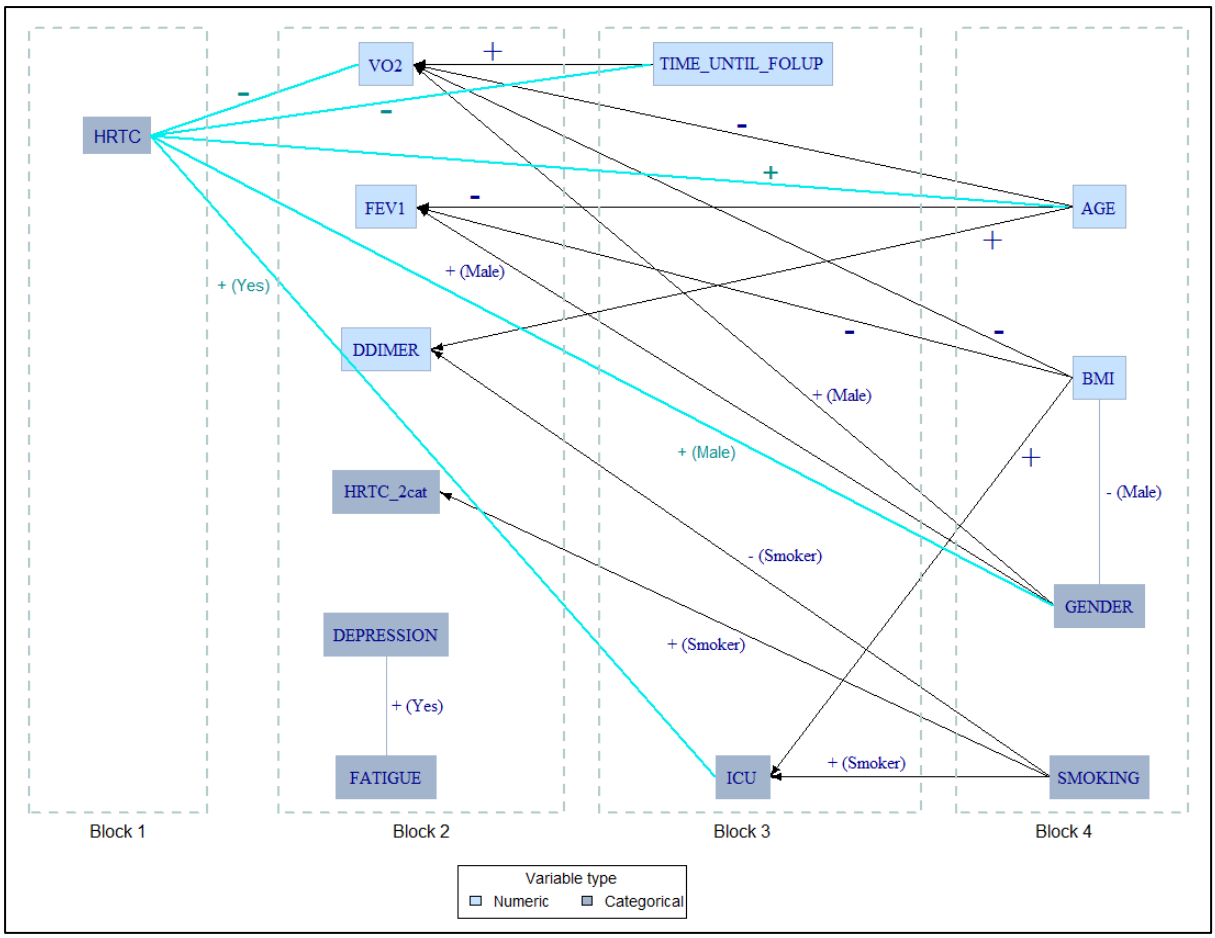


Figure 15: Γραφικό μοντέλο αλυσίδας

Συμπεράσματα

Σε αυτή την εργασία, ξεκινήσαμε εισάγοντας τη θεμελιώδη θεωρία που διέπει τρεις σημαντικούς τύπους γραφικών μοντέλων: το μη κατευθυνόμενο Γκαουσιανό γραφικό μοντέλο, το μοντέλο μεικτής αλληλεπίδρασης και το γραφικό μοντέλο αλυσίδας. Στη συνέχεια, συνεχίσαμε με την εφαρμογή τους σε πραγματικά δεδομένα, αξιοποιώντας τις στατιστικές δυνατότητες του προγραμματιστικού εργαλείου R. Κατά συνέπεια, συνδυάσαμε θεωρία και εφαρμογή για να αποκτήσουμε γνώσεις και δεξιότητες αξιοποίησης γραφικών μοντέλων για στατιστική ανάλυση.

Συμπερασματικά, ερευνήσαμε εστιάζοντας στις συσχετίσεις των αντικειμενικών μετρήσεων της φυσικής κατάστασης και των υποκειμενικών αξιολογήσεων που προέκυψαν μέσω ερωτηματολογίων. Αυτή η προσέγγιση μας επέτρεψε να αποκτήσουμε γνώσεις σχετικά με τον μακροπρόθεσμο αντίκτυπο του COVID-19 στην υγεία. Η σύνοψη των αποτελεσμάτων από τα $Σχήματα$ 6,7 και 8 και τους $Πίνακες$ 13,14 και 15, αποκάλυψε ότι οι επιπτώσεις του ιού στα άτομα επηρεάζονταν από πολλούς παράγοντες, όπως η ηλικία του ασθενούς, ο δείκτης μάζας σώματος του ασθενούς, καθώς και ως το φύλο ενός ασθενούς.

Μια αξιοσημείωτη γενική παρατήρηση είναι ότι το γράφημα αλυσίδας στο $Σχήμα$ 8 μοιράζεται πολλές άκρες με το γράφημα μικτής αλληλεπίδρασης στο $Σχήμα$ 7. Αυτή η συνέπεια είναι πολύ σημαντική και πολύτιμη επειδή ενισχύει την εγκυρότητα των συμπερασμάτων που προέρχονται από αυτά τα γραφήματα.

Ανακαλύψαμε ενδιαφέροντα αποτελέσματα, κάποια από αυτά ήταν αναμενόμενα και κάποια άλλα όχι, λαμβάνοντας υπόψη υπάρχοντες αλλά και απόντες συσχετισμούς.

Πρώτον, παρατηρήσαμε διαφορές με βάση το φύλο. Η ανάλυση αποκάλυψε ότι οι άνδρες έτειναν να παρουσιάζουν υψηλότερες τιμές μέγιστης κατανάλωσης οξυγόνου ($VO2$), καθώς και μέγιστου όγκου εκπνεόμενου αέρα κατά το πρώτο δευτερόλεπτο μετά τη μέγιστη εισπνοή ($FEV1$) σε σύγκριση με τις γυναίκες.

Δεύτερον, αποκαλύφθηκαν διαφορές με βάση την ηλικία. Τα αποτελέσματα δείχνουν ότι η ηλικία μπορεί να επηρεάσει τα αποτελέσματα της υγείας. Πιο συγκεκριμένα, οι ηλικιωμένοι ασθενείς εμφάνισαν αύξηση στο επίπεδο του DDIMER, το οποίο είναι δείκτης πήξης του αίματος. Επιπλέον, καθώς αυξανόταν η ηλικία, παρατηρήθηκε μείωση τόσο της μέγιστης κατανάλωσης οξυγόνου ($VO2$) όσο και του μέγιστου όγκου εκπνεόμενου αέρα κατά το πρώτο δευτερόλεπτο μετά τη μέγιστη εισπνοή ($FEV1$).

Μια ακόμη πολύτιμη παρατήρηση είναι ότι το αυτοαξιολογημένο σύμπτωμα της κόπωσης συσχετίστηκε θετικά με την κατάθλιψη. Αλλά η πτυχή του ενδιαφέροντος εδώ έγκειται στην απουσία συσχετίσεων μεταξύ αυτών των δύο υποκειμενικών εκτιμήσεων και των άλλων αντικειμενικών μετρήσεων. Αυτό το εύρημα υπογραμμίζει τη σύνθετη σχέση μεταξύ της σωματικής και ψυχικής υγείας των ατόμων που αναρρώνουν από τον COVID-19 και μας δείχνει την ανάγκη υποστήριξης αυτών των ανθρώπων.

Για να συνεχίσουμε με τα αποτελέσματά μας, παρατηρήσαμε επίσης ότι ο δείκτης μάζας σώματος (ΔΜΣ) και η εισαγωγή στη μονάδα εντατικής θεραπείας (ΜΕΘ) έπαιξαν καθοριστικό ρόλο στην ανάρρωση του ασθενούς. Οι ασθενείς με μεγαλύτερο δείκτη μάζας σώματος έτειναν να παρουσιάζουν μείωση τόσο της μέγιστης κατανάλωσης οξυγόνου ($VO2$) όσο και του μέγιστου όγκου του εκπνεόμενου αέρα κατά το πρώτο δευτερόλεπτο μετά τη μέγιστη εισπνοή ($FEV1$). Επιπλέον, άτομα με υψηλότερο δείκτη μάζας σώματος είχαν αυξημένη πιθανότητα να εισέλθουν στη μονάδα εντατικής θεραπείας (ΜΕΘ) κατά τη νοσηλεία τους.

Επιπλέον, μέσα από τα γραφήματα ανακαλύψαμε μια συσχέτιση μεταξύ των μεταβλητών $SMOKING$ και ICU. Πιο συγκεκριμένα, υψηλότερη πιθανότητα εισαγωγής στη μονάδα εντατικής θεραπείας παρατηρήθηκε στους καπνιστές.

Τελευταία αλλά όχι λιγότερο σημαντική, είναι η σχέση μεταξύ $VO2$ και $HRTC$. Ανακαλύψαμε ότι τα άτομα με αυξημένη μέγιστη κατανάλωση οξυγόνου ($VO2$) έτειναν να παρουσιάζουν καλύτερα αποτελέσματα στην Υπολογιστική Τομογραφία υψηλής ανάλυσης ($HRTC$), ενώ οι ηλικιωμένοι ασθενείς έτειναν να έχουν χειρότερα αποτελέσματα αξονικής τομογραφίας.

Συνοπτικά, αυτά τα ευρήματα υπογραμμίζουν την αποτελεσματικότητα των γραφικών μοντέλων στην αποκάλυψη κρυφών μοτίβων και στην κατανόηση πολύπλοκων σχέσεων. Μέσω της εφαρμογής αυτών των εξελιγμένων τεχνικών, μπορούμε να εντοπίσουμε κρίσιμους παράγοντες που επηρεάζουν τα αποτελέσματα της υγείας λόγω μιας ασθένειας και να βρούμε πιο κατάλληλες παρεμβάσεις για τον μετριασμό των επιπτώσεων.