

STATISTICAL QUANTIFICATION OF PARAMETERS OF HEART RATE VARIABILITY POWER

ANA STANOJCIC

Bachelor's thesis
2024:K18



LUND UNIVERSITY

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

Statistical Quantification of Parameters of Heart Rate Variability Power

Ana Stanojic

Abstract

Heart rate variability (HRV) describes the variations in time intervals between heartbeats. The signals are usually recorded from electrocardiogram (ECG) beat detection. In this thesis, HRV signals of 97 participants were analyzed using spectral and statistical analysis. The signals are non-stationary, however, the data was treated approximately short time stationary. Using spectral analysis, power of HRV is expressed using exponential decay model from which parameters (a - multiplier for the exponential decrease and c - amplitude multiplier) are extracted. The effects of the different factors (*age, sex, SMBQ, length, weight, BMI, state and trait of anxiety*) on extracted parameters a and c , are statistically analyzed.

The Welch's t -test for the difference of two means of parameter c shows that there is no significant difference in the average values of this parameter for female and male. The Welch's t -test for the parameter a shows significant difference between the mean values of this parameter for female and male, indicating the faster decay of the power of HRV (PHRV) for males. Multiple linear regression (MLR) showed that out of 8 tested independent variables, only the factor *age* is significant for the parameter c . MLR for parameter a shows no significance for any of the tested factors. Also, the MLR model with interaction term for the two factors (*age* and *state*) shows no significant interaction of the two factors for the parameter c .

Future research could be directed to investigating the regression models for the two parameters on the samples that include different age categories separately for female and male participants as well as exploring how the parameters of the model of PHRV relate if non-linear methods for quantifying HRV are used.

Populärvetenskaplig sammanfattning

Denna forskning fokuserar på hjärtfrekvensvariation (HRV), vilket är fluktuationer i tidsintervall mellan följande hjärtslag. Dessa variationer samlas in genom användandet av ett elektrokardiogram (EKG). Genom att analysera HRV data från 97 deltagande, syftar denna studie till att upptäcka om olika faktorer som ålder, kön, ångest, och kroppsmått påverkar HRV. Att förstå HFV är väsentligt eftersom det är kopplat till diverse olika hjärt-kärlsjukdomar, där bättre insikt kring HFV kan hjälpa tidigt uppfångande och hantering av hjärt-kärlsjukdomar.

Studien visade att kraften av HRV minskar snabbare hos män jämfört med kvinnor, vilket tyder på en skillnad i hur män och kvinnors hjärtan reagerar över en längre tid. Ålder framstod som en betydande faktor av en specifik parameter för HRV, därav indikerar detta att när människor åldras finns det en anmärkningsvärd påverkan på hjärtfrekvensvariation. Förvånansvärt visade inte faktorerna vikt, längd, "body mass index" (BMI) och ångestnivåer påverkan på HRV i denna studie.

Dessa fynd åskådliggör vikten av att ha i åtanke kön likväl ålder vid bedömning av hjärthälsa, vilket leder till en mer skraddarsydd behandlingsplan. Studien föreslår att framtida forskning bör undersöka HRV med fler faktorer samt större provtagning i syfte att ådagalägga mer kring HFV. Genom mer precis data kring HRV kan individer gynnas av förbättrad diagnostik och personifierade behandlingsplaner.

Denna studie upplyser faktorer som påverkar hjärtfrekvensvariation, med emphasis på ålder och kön. Dessa insikter är essentiella för framsteg inom personligt skraddarsydd vård och framtida forskning inom området kan leda till bättre hälsa-ingrepp för alla åldrar och kön.

Dedication

To my mother who always believed in me.

Acknowledgements

This thesis would not have been possible without the guidance and the help of several individuals who in one way or another contributed with their valuable assistance.

I would like to express my sincere thanks to my supervisor Prof. Maria Sandsten who provided me with encouragement and patience throughout my work on this thesis. Her guidance and help with the programming aspect in MATLAB enabled me to complete my work efficiently.

Special thanks to Prof. Selena Vasic who generously provided her constructive comments and suggestions.

Lastly, I am grateful to my family. Their support and belief in me kept my spirit and motivation high during the years of my bachelor's degree studies and time spent on working on this thesis.

Contents

1	Introduction	1
1.1	Background Information	2
1.2	Methods for quantifying HRV	4
1.2.1	Time-domain methods	4
1.2.2	Frequency-domain methods	4
1.2.3	Non-linear Methods	5
1.3	Aim of the thesis	5
2	Mathematical concepts	6
2.1	Spectral Analysis	6
2.1.1	Power Spectral Density (PSD)	7
2.1.2	Periodogram	7
2.1.3	Modifying the Periodogram	9
2.1.4	Spectrogram	10
2.2	Statistical Analysis	10
2.2.1	Welch's t -Test for the Difference Between Two Means	10
2.2.2	Multiple Linear Regression	11
3	Data Processing	14
3.1	Description of Data	14
3.2	Spectral Analysis	15
4	Statistical Analysis	18
4.1	Descriptive Statistics	18
4.2	Welch's t -test	19
4.2.1	Parameter c	19
4.2.2	Parameter a	22
4.3	Multiple Linear Regression	24
4.3.1	Parameter c	24
4.3.2	Parameter a	29
4.4	MLR with interaction term	30
5	Conclusion	35
6	Bibliography	37

List of Figures

1.1	QRS complex pattern recorded using ECG	3
3.1	Samples of Respiratory and HRV signal	15
3.2	Sample of Respiratory Spectrogram and HRV Spectrogram	16
3.3	Sample of the PHRV over time plot	16
3.4	Sample Exponential Model of PHRV	17
4.1	Summary metrics for numerical factors	18
4.2	Histograms of parameters a and c of the exponential decay model	19
4.3	Distribution of parameter c by <i>sex</i>	20
4.4	Descriptive statistics for female and male for the parameter c	20
4.5	Distribution of parameter a by <i>sex</i>	22
4.6	Descriptive statistics for female and male and the results of the Welch t -test for the parameter a	23
4.7	Distribution and diagnostic plots of the MLR model	25
4.8	Diagnostic graphs for the assumptions for the logarithmic correction	26
4.9	Table of VIF values for the two MLR models	26
4.10	Diagnostic graphs for the assumptions for the reduced MLR model	27
4.11	Summary statistics of the parameter c coefficients, p -values, (R^2) value and adjusted (R^2) value in the reduced MLR model	28
4.12	Summary statistics of the coefficients of the parameter a , p -values, R^2 value and adjusted R^2 value in the reduced MLR model	29
4.13	Histograms of factors <i>age</i> and <i>state</i> of the exponential decay model	30
4.14	Diagnostic graphs for the logarithmic correction model with interaction term	31
4.15	Summary of the MLR analysis for factors <i>age</i> , <i>state</i> and their interaction .	32
4.16	Scatter plot of c against <i>age</i> , colored by <i>state</i>	33

Chapter 1

Introduction

Heart Rate Variability (HRV) represents the physiological phenomenon of variation in the time interval between consecutive heartbeats. It is a critical marker for understanding the function of an autonomic nervous system and overall cardiovascular health. Using spectral and statistical analysis of HRV data, researchers can gain profound insights into the underlying mechanisms that influence heart rhythms, assess stress levels, and predict potential health issues. Hence, investigating and understanding HRV could indicate changes in psychological and physiological states of a person. It is of interest to look into different factors which could induce changes in the heart rate variation, in order to assess the health of an individual and further recommend different treatments. [1]

In this thesis, different methods of spectral analysis will be utilized to extract the parameters of the power of HRV (parameter c that represents the PHRV multiplier, and parameter a - the exponential decay factor). Statistical analysis will be performed to explore the relationship between these parameters and different factors of HRV.

The data are obtained from the experiment conducted on a sample of 97 participants whose breathing was observed in different daily situations, investigating the relationship between the respiratory frequency and HRV. The model for the power of HRV (PHRV) is developed with two parameters (PHRV multiplier and the exponential decay factor). Data are statistically analysed. Welch's t -test for the difference of two means is conducted to check if there is significant difference between the mean values of the parameters c and a of power of HRV for females and males. Multiple Linear Regression (MLR) is

performed with 8 available factors (*age, sex, Shirom-Melamed Burnout Questionnaire (SMBQ), length, weight, BMI, state of anxiety, and trait of anxiety*) to investigate if there is a linear relationship between parameters of HRV and each of the factors. Results of the statistical analysis in this thesis allow for conclusions about HRV that can contribute to the advancement of personalized medicine, improve diagnostic accuracy, and enhance therapeutic interventions.

1.1 Background Information

Human heart's contractions are initiated by the sinoatrial node (SA-node). The SA-node sends an electrical stimulus 60 to 100 bpm (beats per minute) under normal resting conditions depending on age of the individual [2]. These signals trigger the contractions of the atria of the heart, which in response starts up the contraction of the ventricles. HRV is a physiological phenomenon that describes slight time fluctuations between two consecutive heartbeats. HRV is highly non-stationary, and this becomes more profound when HRV data are measured in naturally non-stationary environments, for example, social stress. In general, HRV data measured in such situations are more difficult to analyze than those measured in stationary environments.

HRV can be measured in different ways, but the most convenient and efficient method is electrocardiography (ECG). ECG produces electrocardiogram which describes the pattern of the heart beats as changes in voltage of electrical impulses, which are measured through the electrodes attached to the skin. The typical ECG pattern contains three main parts, the first peak (P-wave), second peak (QRS complex) and the last peak (T-wave). P-wave shows how the impulse spreads across the atria causing them to contract and pump blood into ventricles and then relax. Then, the impulse causes the ventricles to contract as seen in QRS complex. Lastly, throughout the T-wave, ventricles relax as the impulse stops spreading (Figure 1.1). In this sense, HRV is the variation in time between the two peaks of consecutive QRS complexes, known as the RR interval (measured in milliseconds) [3].

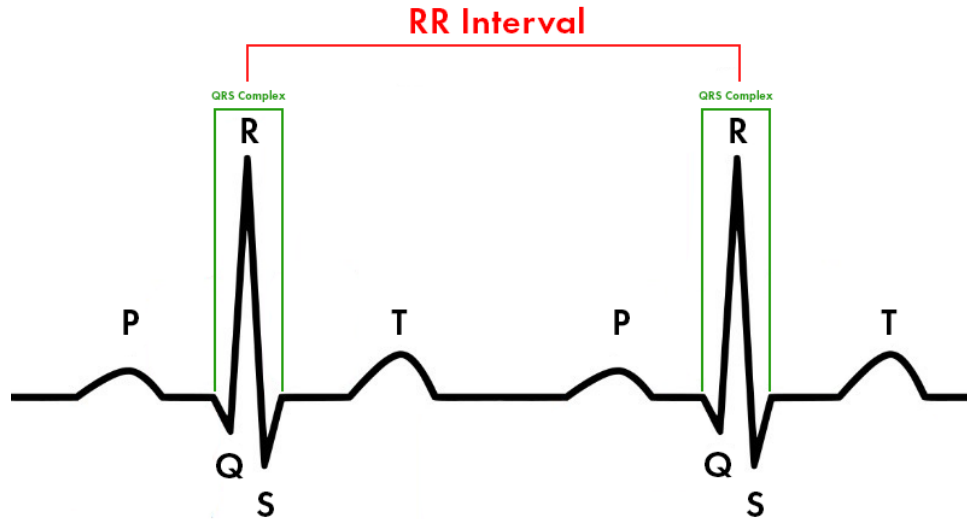


Figure 1.1: QRS complex pattern recorded using ECG

The autonomic nervous system (ANS) is a part of the peripheral nervous system that regulates involuntary physiologic processes such as blood pressure, heart rate, respiration, digestion, and sexual arousal. It consists of three divisions: sympathetic (SNS), parasympathetic (PNS), and enteric. The first two are of interest to us. HRV is controlled by ANS, therefore, HRV is a good indicator of the state of a section of ANS responsible for managing cardiac activity. Essentially, the heart rate is controlled by the balance between SNS and PNS. SNS is in-charge of guiding the state of overall elevated activity and attention. More adrenaline in the SA node leads to increased blood pressure and heart rate. PNS as the opposite function from SNS, is governing the resting and digesting processes, such as lowering the heart rate and blood pressure or restarting the gastrointestinal digestion .[4]

HRV signals are considered as non-stationary as they contain both linear and non-linear contributions. They can be analyzed in frequency domain. SNS is usually associated with the "low frequency band" (LF) that ranges from 0.04 to 0.15 Hz. On the other hand, PNS is associated with "high frequency band" (HF) ranging from 0.15 to 0.4 Hz. Throughout this thesis, we will be focusing mainly on the HF interval [5].

There are certain parameters that have a stronger impact on HRV, such as *sex*, *age*, *psychological conditions* or *disorders*. Previous conclusions indicate that females have lower HRV compared to males, however, this difference decreases over time. *Age* is another factor that has shown noticeable influence on HRV, where the HRV values decrease with increase in age [5]. We will focus on extracting parameters of recorded HRV values using frequency-domain methods, and then analysing them using statistical methods. The base for the HRV analysis is the ECG recording from which the HRV time series can be extracted.

1.2 Methods for quantifying HRV

Based on the ECG recording, the series of consecutive RR intervals is analyzed to produce quantitative outputs, which can be used as variables for research. HRV can be analyzed using three different methods: 1) time-domain, 2) frequency-domain, and 3) non-linear methods. The detailed mathematical description of each method exceed the scope of this thesis, but the main aspects of the most commonly used methods for quantifying HRV will be outlined.

1.2.1 Time-domain methods

Time-domain methods are the simplest and most intuitive for perform HRV analysis. They represent direct mathematical operations to the series of successive RR interval values. These parameters include the standard deviation (SDNN) of RR intervals, the mean heart rate, mean value of RR intervals. [6]

1.2.2 Frequency-domain methods

In the frequency-domain methods, a power spectral density (PSD) estimate can be calculated for the RR interval series. The most important periodic component of HRV is the “respiratory sinus arrhythmia” which ranges from 0.15 Hz to 0.4 Hz and is considered an “high frequency” oscillation. “Low frequency” oscillations are usually ranging from 0.04

Hz to 0.15 Hz. The frequency-domain measures which are extracted from the PSD estimate, include absolute and relative power for each frequency band, as well as the LF/HF power ratio and peak frequencies for each band.

1.2.3 Non-linear Methods

The analysis of complex processes and measures have been developed in the field of non-linear dynamics, describing the underlying structure of non-stationary, non-periodic but deterministic series of data. The non-linear features of HRV have can be investigated using novel methods such as detrended fluctuation analysis (DFA) [7] and adaptive fractal analysis (AFA) [8] that will not be used in this thesis but are commonly used in the scientific research.

1.3 Aim of the thesis

The aim of this thesis is to complete statistical analysis of the relationship between the parameters extracted from HRV data and the factors gathered in the experiment. The goal is to first obtain the spectrograms for recorded signals and then the plot of power of HRV. The exponential model of extracted power of HRV at each time point will be created. Using the data from the model, a set of parameters can be estimated and their effect on HRV discussed. Several analysis will be used among which are Welch's t-test for the difference of two means, multiple linear regression (MLR) and MLR with interaction term. They will give more insight into possible relationship between the parameters of PRHV and the observed factors of HRV. These findings can help to understand the difference in statistical properties of HRV for different frequencies of breathing of participants (female and male).

Chapter 2

Mathematical concepts

In this section, theoretical outline of different methods of spectral analysis used to process the collected HRV data and methods for statistical analysis to be used to analyze the obtained data, are presented.

2.1 Spectral Analysis

Spectral Analysis is one of the most commonly used methods in both theoretical and applied analysis of stationary stochastic processes. It shows the analysis of continuous or discrete time-series with respect to frequency-domain descriptions and provides more insight in characteristics of a time-series, compared to time-domain analysis. [9]

One of the main drawbacks of spectral analysis is that it is a linear method. It considers HRV time series as a linear combination of the independent oscillatory components, where interactions between those components are neglected. Discrete Fourier Transform (DFT) is often used as a method of transforming samples taken at regular time intervals into amplitudes of scaled *sine* and *cosine* functions (frequency domain). Usually, DFT takes in \mathbf{n} , number of time domain signal samples ($\mathbf{x}[\]$) and decompose them into two frequency domain ($\mathbf{X}[\]$) arrays of $(\mathbf{n}/2)+1$ samples, each describing the amplitudes of sine and cosine functions. [10]

A DFT for time-series, $x(t), t = 0, +1, +2, \dots, n - 1$ is given as follows:

$$X(k) = \sum_{t=0}^{n-1} x(t)e^{-i2\pi\frac{kt}{n}},$$

for $f = \frac{k}{n}$ in the interval $(-0.5 \leq f \leq 0.5)$. Non-parametric spectrum methods involve Fourier transform to calculate the power spectrum.

2.1.1 Power Spectral Density (PSD)

Power Spectral Density (PSD), denoted R_x , shows how a time-series signal's power varies in the frequency domain. The PSD for a discrete signal has two definitions. The first definition says that covariance function $r(\tau)$ with condition : $\sum_{-\infty}^{\infty} |r(\tau)| \leq \infty$ will have symmetric, non-negative, and integrable Fourier transform:

$$R_x(f) = \sum_{\tau=-\infty}^{\infty} r(\tau)e^{-i2\pi f\tau}.$$

The second definition looks at the power of the signal as a limit of an energy signal :

$$R_x(f) = \lim_{n \rightarrow \infty} \frac{|X(f)|^2}{n}.$$

Using PSD analysis which utilises FFT (version of DFT that requires a smaller, finite number of samples where $n = 2^s$, s being an integer) we describe oscillations in the RR and transform them into discrete frequencies that help our understanding of the physiological mechanisms responsible for HRV [11].

2.1.2 Periodogram

Periodogram is a non-parametric, direct way to estimate the PSD under assumption of signal stationarity, and it is a tool used for discovering dominant periods or frequencies in the observed series [12]. It can be found as:

$$\hat{R}_x(f) = \frac{1}{n}|X(f)|^2,$$

for every frequency, and relating to the second definition of PSD. We can rewrite this expression using the estimate of the covariance function as $\hat{r}_x = \frac{1}{n} \sum_{t=0}^{n-1-|\tau|} x(t)x(t+|\tau|)$, into a form that resembles the first definition of PSD:

$$\hat{R}_x(f) = \sum_{\tau=-n+1}^{n-1} \hat{r}_x(\tau) e^{-i2\pi f\tau}.$$

To assure the validity of the estimate, it should have approximate unbiasedness and low variance. We are interested in whether periodogram would be considered as unbiased estimator of PSD, so we check the expected value of its definition: $\mathbb{E}[\hat{R}_x(f)] \approx R(f)$:

$$\mathbb{E}[\hat{R}_x(f)] = \hat{R}_x(f) = \sum_{\tau=-n+1}^{n-1} \mathbb{E}[\hat{r}_x(\tau)] e^{-i2\pi f\tau} = \sum_{\tau=-n+1}^{n-1} \left(1 - \frac{|\tau|}{n}\right) r_x(\tau) e^{-i2\pi f\tau}.$$

If we let $n \rightarrow \infty$, then $\sum_{\tau=-\infty}^{\infty} r_x(\tau) e^{-i2\pi f\tau} = R_x(f)$, meaning that periodogram is indeed asymptotically unbiased estimate of PSD. However, for smaller values of n , we use a rewritten formula for estimation of PSD. In frequency domain, this formula represents convolution between Fejér Kernel ($K_n(f)$) and spectral density:

$$\mathbb{E}[\hat{R}_x(f)] = \int_{-0.5}^{0.5} R_x(u) K_n(f-u) du.$$

This means that when we convert a signal from time domain into frequency domain and applying the FFT, we encounter an issue known as "leakage". Leakage is energy spilled over into adjacent frequencies from the true frequency of the signal [13]. It is caused by inconsistencies in period properties when applying FFT, resulting in height of the signal's amplitude being lower compared to the true amplitude height of the signal.

After looking into the expectation of the periodogram, some well-known conclusions regarding its variance can be stated. As the number of data points, n asymptotically increases ($n \rightarrow \infty$), the variance of periodogram:

$$V[\widehat{R}_x(f)] \approx \begin{cases} R_x^2(f), & 0 < |f| < \frac{1}{2} \\ 2R_x^2(f), & f = 0, \pm\frac{1}{2} \end{cases}$$

Hence, periodogram cannot be seen as consistent estimator of spectral density as it does not decrease as $n \rightarrow \infty$. The conclusion is that periodogram is a biased estimate of spectral density for a finite number of data samples. In order to correct this, in the next section, we will introduce the modification of the concept called the *lag window* $k_n(\tau)$ (Inverse Fourier Transform of Fejér's Kernel).

2.1.3 Modifying the Periodogram

When attempting to estimate the PSD, we run into issues with bias which is caused by the height of the side-lobes appearing in the Fejer's Kernel. Therefore, the techniques of bias reduction should be applied which involve modifying the periodogram, called *data windowing* [14]:

$$\widehat{R}_w(f) = \left| \sum_{t=0}^{n-1} x(t)w(t)e^{-i2\pi ft} \right|^2.$$

The idea of windowing is to reduce leakage bias by lowering the heights of side-lobes.

One of the most common technique is the method called the *Hanning window*, that introduces a new function $h(t)$:

$$h(t) = \frac{1}{2} - \frac{1}{2} \cos \frac{2\pi t}{n-1},$$

$t=0, \dots, n-1$, that can be normalized to fit the format of $w(t)$:

$$w(t) = \frac{h(t)}{\sqrt{\sum_{t=0}^{n-1} h^2(t)}}.$$

The modified periodogram using the Hanning window [15] will have lower side-lobes and wider main-lobe, compared to Feijer's Kernel. This should result in reduced leakage that will increase the accuracy of estimation of PSD compared to the regular periodogram.

2.1.4 Spectrogram

Spectrograms provide valuable insights into HRV by revealing frequency components and their changes over time. They offer detailed time-frequency analysis for non-stationary signals but are complex for interpretation and have resolution trade-offs between time and frequency domains where a higher time-resolution means lower-frequency resolution and vice versa. In HRV analysis, a spectrogram can show how the power of different frequency components changes over time. Overall, one can observe how the frequency components change over time by examining a spectrogram which provides insights into ANS activity and how well it responds to various stimuli. [16]

2.2 Statistical Analysis

2.2.1 Welch's t -Test for the Difference Between Two Means

Welch's t -test is a method of hypothesis testing when the means of two independent groups are being compared. The two populations do not have the same variance [17].

If $X_1 \in N(\mu_1, \sigma_1^2)$ and $X_2 \in N(\mu_2, \sigma_2^2)$, we will briefly define the null and alternative hypothesis in the Welch's t -test test:

$$H_0 : \mu_1 = \mu_2,$$

$$H_1 : \mu_1 \neq \mu_2$$

When working with Welch's t -test it is important to understand how the test statistic:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

and degrees of freedom are calculated:

$$df \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}.$$

In these expressions \bar{X}_1 , \bar{X}_2 are the sample means for the two groups, s_1^2 and s_2^2 denote the sample variances for the two groups, while n_1 and n_2 represent the sample sizes for the two groups.

Lastly, we will briefly mention which assumptions must hold for performing of the Welch's t -test. Data must be continuous and collected using a random sample. Each investigated group should have more than 15 observations or the sample data should follow a normal distribution. [18]

2.2.2 Multiple Linear Regression

Multiple linear regression (MLR) refers to a statistical technique that uses two or more independent variables to predict the outcome of a dependent variable. The technique determines the variation of the model and the relative contribution of each independent variable in the total variance. MLR model is a fit to represent a set of sample data, with **dependent variable (response)** y that depends on k **independent (regressor) variables**, x_1, x_2, \dots, x_k :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$$

This equation represents a MLR model with one response y , regressor variables x_j and **regression coefficients** β_j , where $j = 0, 1, 2, \dots, k$. The regression coefficients measure the expected change in y per unit change in the corresponding x_j when all remaining regressor variables $x_i (i \neq j)$ are held constant. β_0 is the y -intercept, and ϵ is the model's random error term: $\epsilon_i \approx IID(0, \sigma^2)$. By using MLR, we can predict behaviour of one variable based on known information about several other variables. To apply MLR model the following assumptions should hold:

- 1) **Independently** and **randomly** selected observations (y_i) from the population.
- 2) **Linearity**: A linear relationship between the dependent variable and the independent variables. This can be checked by looking for any clear patterns or systematic structures in the residuals. For example, a Residuals vs Fitted plot can be created. In this plot, each

point represents a combination of the factor levels and replicates, and the residuals should be spread randomly around the horizontal line ($y=0$) with no clear pattern.

3) **No multicollinearity**: Multicollinearity exists when several independent variables in a model are correlated. Detecting multicollinearity is important because even though it does not reduce the explanatory power of the model, it reduces the statistical significance of the independent variables. A common measure of multicollinearity among the independent variables in a MLR model is *Variance Inflation Factor (VIF)* which can be considered for each parameter. VIF describes how severe multicollinearity is in the least square regression analysis, by measuring the number of inflated variances caused by multicollinearity. A high VIF of an independent variable indicates a highly collinear relationship with other variables. Typical way of interpreting the value of VIF is: VIF = 1 (variables not correlated); $1 < \text{VIF} < 5$ (moderately correlated); $\text{VIF} > 5$ (highly correlated); VIF above 10 is a cause for concern [19].

4) **Normally distributed residuals** (errors): In this context, residuals represent the difference between the created line of best fit and the actual data values:

$$\epsilon_i = y_i - \hat{y}_i.$$

The normality of residuals [20] affects the accuracy of the hypothesis tests, the confidence intervals for the regression coefficients, as well as the validity of the model selection criteria: coefficient of determination R^2 and adjusted R^2 . We will use QQ-plots to compare the quantiles of the residuals to the theoretical quantiles of a normal distribution. We expect the residuals to follow a diagonal straight line.

5) **Homoscedasticity**: Homoscedasticity means that the residuals have constant variance across all levels of the fitted values [21]. Homoscedasticity can be checked by plotting scale-location plot. If there is no funnel looking shape or pattern denoting increasing or decreasing spread of residuals as the fitted values increase, then the model does not suffer from systematic errors related to the levels of the factors. Hence the variance of the residuals remains constant, satisfying the assumption of homoscedasticity.

When working with MLR we are interested in the coefficient of determination (R^2), a statistical measure that represents the proportion of the variance for a dependent variable that is explained by an independent variable or variables in a regression model [22]. Essentially, R^2 explains to what extent the variance of one variable explains the variance of the second variable. So, if the R^2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.

Another coefficient to be considered is adjusted R^2 , a modified version of R^2 that has been adjusted for the number of predictors in the model. Adjusted R^2 addresses a limitation of R^2 and is used to determine how reliable the correlation is and how much it is determined by the addition of independent variables. R^2 will stay the same when adding more predictors, even if they are not contributing meaningfully. It increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected. Typically, the adjusted R^2 is positive and always lower than the R^2 .

So far, we have described how to build a multiple linear regression model with multiple factors of HRV. The MLR equation, also known as additive model, investigates only the main effects of predictors. It assumes that the relationship between a given predictor variable and the outcome is independent of the other predictor variables. This assumption might not be true. For example, the age factor may affect some other factor which is referred to as an interaction effect.

The multiple linear regression equation, with interaction term ($\beta_3(x_1x_2)$) between two predictors x_1 and x_2 , can be written as follows:

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3(x_1x_2) + \epsilon.$$

The coefficient β_3 represents the increase/decrease in effect of factor x_1 for one unit increase of factor x_2 , and equally the increase in effect of factor x_2 for one unit increase of factor x_1 .

Chapter 3

Data Processing

3.1 Description of Data

Data used in this thesis are secondary and obtained through an experiment conducted at the Department of Laboratory Medicine, Division of Occupational and Environmental Medicine, Lund University [23,24]. There were 97 participants in total, 47 females and 50 males, of ages ranging from 20 to 61. Respiration rate and the HRV measurements were taken in two instances, as participants were in resting state and in controlled breathing state. By the controlled breathing state we consider individual's breathing following the frequency of the metronome pattern starting at 0.12 Hz and slowly increasing the pace to 0.35 Hz. The data was collected over 5-minute intervals with the sampling frequency of 4Hz. None of the participants were taking any medicine that is known to affect the heart rate patterns or were suffering from any cardiovascular diseases. Physiological parameters such as *age*, *BMI*, *sex* and psychological parameter describing stress levels in form of *SMBQ* and *State and Trait of anxiety* measurement were recorded for each participant. Data collection was conducted as a combination of two medical techniques. Using a strain gauge attached to a strap that is wrapped over participant's torso measured changes in the pressure caused by the inflation/deflation of the torso during breathing, therefore, providing the information about respiratory signals of the individual [25]. The other technique was the use of ECG, showing the pattern of heart rate signals. Here, a sample of respiratory and HRV signal is presented (Figure 3.1).

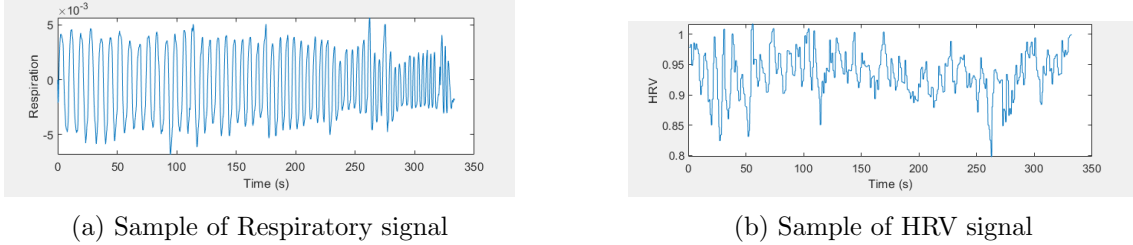


Figure 3.1: Samples of Respiratory and HRV signal

3.2 Spectral Analysis

The methods explained in the Mathematical concepts chapter, assume the stationarity or approximate stationarity of the signals, while the data collection resulted in recorded non-stationary respiratory and HRV signals of 5-minute lengths. Therefore, the first course of action was to cut the recordings into shorter, 16 second intervals, and make the data appear stationary. Additionally, the mean of the data was also adjusted to zero.

As our data was recorded in time-domain, it was needed to be transformed into frequency-domain, for us to look at the power of the time-series signal. This transformation of the HRV signal was done by using periodograms as a single-window method. First, the periodogram of the respiratory signal and periodogram of the HRV signal were created for every patient by applying the Fourier transform on sectioned segments of the signals. In attempt to reduce the effect of leakage in periodograms, they were modified using Hanning window with length of 64.

After creating the periodograms of the HRV signals for each participant, a spectrogram of HRV and respiratory signals was made (Figure 3.2) to further observe the behaviour of HRV and extract the time-varying PHRV.

We begin this extraction of data by finding the maximum values of respiratory spectrogram at each time point. Then, the frequencies of those maximum values were extracted for each time point as we can see in Figure 3.2 (a). Next, we used the extracted frequencies as midpoints to create time-dependent frequency band on the HRV spectrogram.

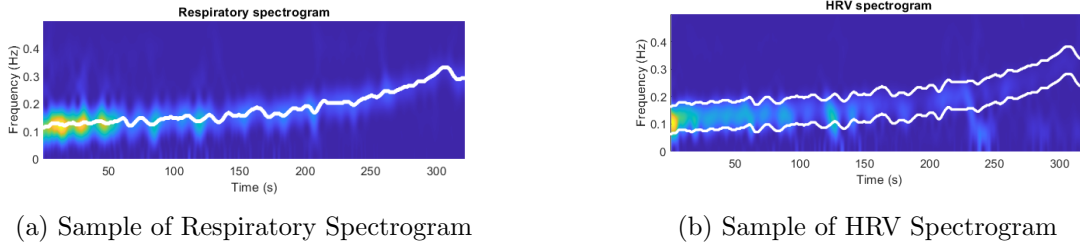


Figure 3.2: Sample of Respiratory Spectrogram and HRV Spectrogram

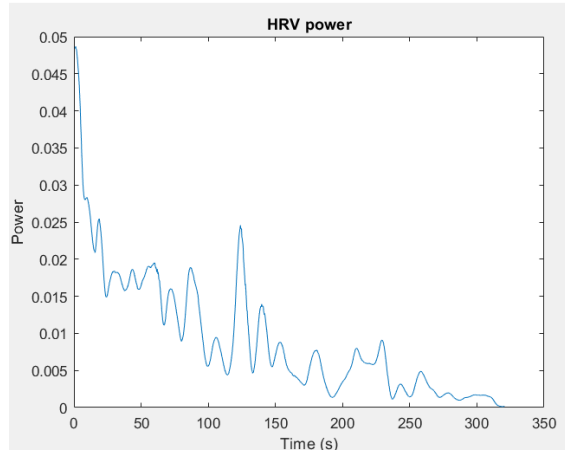


Figure 3.3: Sample of the PHRV over time plot

This frequency band was created individually for each time point and can be seen as the white continuous line on the spectrogram as seen in Figure 3.2 (b). Lastly, PHRV was extracted as the average power in the frequency band at each time point and its plot was created to show its variation over time (Figure 3.3).

Approximately on all PHRV plots of the participants, we can notice that in the first 50 seconds, there is a notable high power. There is also a general trend of decreasing power, suggesting a reduction in the amplitude of HRV fluctuations. Lastly, the power levels off at a low value towards the end of the period (beyond 250 seconds), indicating minimal HRV fluctuations. It is of interest to further investigate the relationship between PHRV and some physiological and psychological parameters using statistical methods. In order to proceed with such analysis we needed to extract parameters of PHRV. This was done by fitting an exponential model in form of:

$$y(t) = ce^{at},$$

to the graph of PHRV (Figure 3.4). The results were values for a , the decay rate parameter for each participant, and c , scaling factor that describes the multiplier of the whole PHRV function.

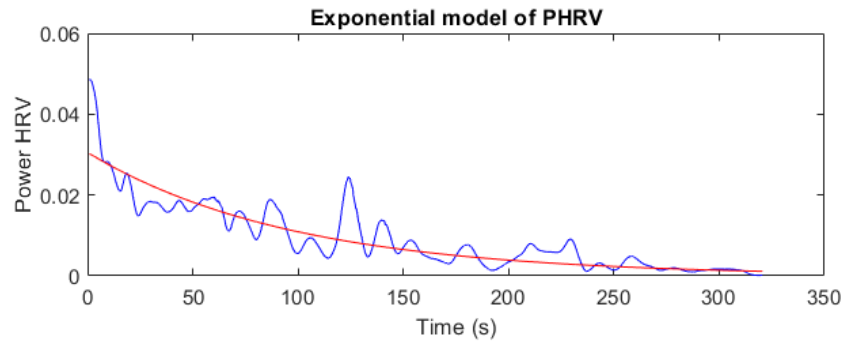


Figure 3.4: Sample Exponential Model of PHRV

The exponential (red) curve was chosen for the model as it follows the decreasing trend of the PHRV (blue) well.

However, it is important to mention that the PHRV shows fluctuations around the exponential model. These oscillations are not captured by the exponential model, which only represents the general trend. This approximation still provides a solid model for the further analysis.

Chapter 4

Statistical Analysis

4.1 Descriptive Statistics

We will start the statistical analysis with a simple descriptive statistics to visualize the data and get the idea of its spread. Figure 4.1 below shows the summary of descriptive statistics for the *factors of HRV* and *two parameters* of the exponential PHRV model.

The mean age of the participants is 29.6. The distribution of age is positively skewed with somewhat larger standard deviation of 9.46 years. The distributions of factors *SMBQ*, *length*, *weight*, *BMI* and *state* and *trait of anxiety* are all quite symmetrical (the values of the mean and the median are quite similar).

statistic	age	SMBQ	length	weight	BMI	state	trait	a	c
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1 mean	29.6	3.18	175.	71.6	23.3	33.1	38.0	-0.00987	0.0517
2 median	26	2.91	175	70	22.9	31	38	-0.0097	0.0406
3 sd	9.46	1.20	8.97	12.5	3.12	7.59	10.1	0.00392	0.0478
4 min	20	1	155	48	17.0	20	20	-0.0233	0.0007
5 max	61	5.77	193	100	30.5	54	64	-0.0004	0.225

Figure 4.1: Summary metrics for numerical factors

The histograms (Figure 4.2) show the distribution of the two parameters a and c of the exponential model for the PHRV:

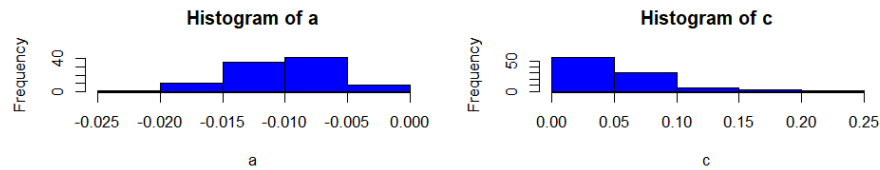


Figure 4.2: Histograms of parameters a and c of the exponential decay model

We can see that the distribution of the values of the parameter a is negatively-skewed with majority of data still falling into the interval around the mean value of -0.00987 . On the other hand, the distribution of the parameter c is positively-skewed with only a few values less than the mean of 0.0517 . So, the mean of the data for parameter c (the amplitude of the PHRV) is bigger than the median indicating that the mean overestimates the majority of the values in the distribution.

In the following section, the analysis of the parameters c and a of the exponential decay model of the PHRV will be performed and the relationship between these parameters and the factors of HRV investigated.

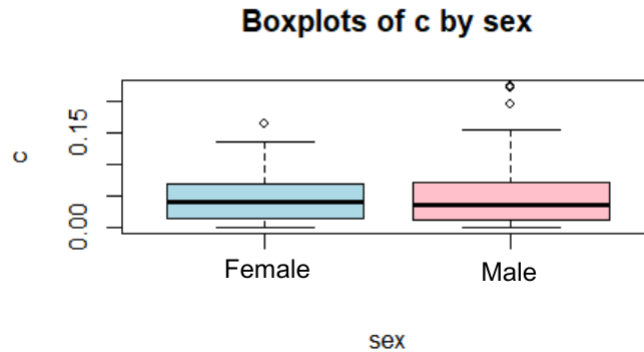
4.2 Welch's t-test

In the observed sample of 97 participants with approximately equal number of female (47) and male (50) participants, it is of interest to check whether there is a significant difference between the mean values of the parameters c and a of PHRV for female and male participants.

4.2.1 Parameter c

First, we will illustrate the 5-number summary and its visualisation in the form of the box-plot for the parameter c for the female and male participants:

PARAMETER "c"	FEMALES	MALES
MIN	0.0011	0.0007
Q1	0.0165	0.0134
MEDIAN	0.0409	0.0372
Q3	0.0694	0.0723
MAX	0.1649	0.2250

(a) Five-point summary for parameter c for females and males(b) Box-plot for the parameter c for females and malesFigure 4.3: Distribution of parameter c by *sex*

The box-plots for female and male show that the distribution of the parameter c is quite similar for the two sexes. The median values, as well as the first and third quartiles are almost the same. The maximum value of the parameter c for male (0.2250) is somewhat higher than the maximum value for the female (0.1649).

Welch's t -test for parameter c

The following table shows the simple statistics needed to perform the t -test:

PARAMETER "c"	FEMALE	MALE
SAMPLE SIZE	47	50
MEAN	0.049342553	0.054008
VARIANCE	0.001567216	0.002997104

Figure 4.4: Descriptive statistics for female and male for the parameter c

We can see that the variances and sample sizes of the two populations are different. Therefore, we will use the Welch's t -test, that considers the observed constraints.

Definition of the hypotheses:

Null Hypothesis (H_0): There is no significant difference in the average value of the parameter c for females and males.

Alternative Hypothesis (H_1): There is a significant difference in the average value of the parameter c for females and males.

Sample calculations for parameter c :

In order to calculate the t -statistic, the difference between the mean values of the parameter c : $\bar{X}_1 - \bar{X}_2$ was found to be 0.004665447. The **standard error** of the difference between two means is given by:

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{0.000033348 + 0.000059942} \approx 0.0096598.$$

Hence, the corresponding **t -statistic** is:

$$t = \frac{0.004665447}{0.0096598} \approx 0.483.$$

In the Welch t -test, the degrees of freedom (df) does not have to be a whole number, and it is calculated using the Welch-Satterthwaite equation (provided in the Chapter 2) which takes into consideration sample sizes and variances of the two groups. The df is equal:

$$df \approx \frac{8.7014281 \times 10^{-9}}{9.7445736 \times 10^{-11}} \approx 89.25$$

We will round the obtained value df to the nearest integer, so $df \approx 89$ and we look at a t -distribution table for the two-tailed test with 89 degrees of freedom and the level of significance $\alpha=0.05$. For the two-tailed test, the significance level is split into two tails, each tail with $\alpha/2 = 0.025$, meaning that the critical t -value for the test is approximately ± 1.987 . Therefore, the rejection region for this test consists of the intervals of t -values less than -1.987 and greater than 1.987.

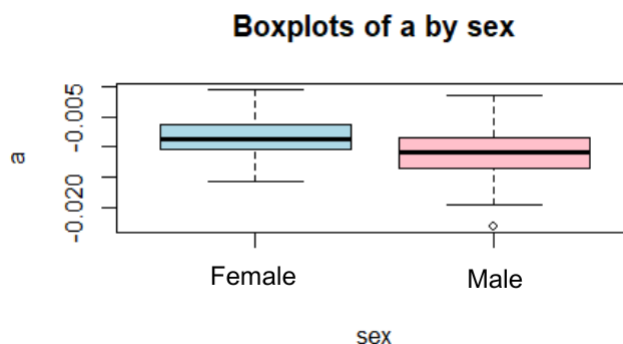
Since the absolute value of the calculated t -value 0.483 is less than the critical t -value 1.987, we fail to reject the null hypothesis. This means that at the $\alpha=0.05$ significance level, there is no significant difference between the mean values of the parameter c of PHRV for the female and male groups. This can be interpreted as there is no significant difference in the amplitude of the PHRV for female and male participants, in this sample. The majority of the participants are younger, and their heart is in general healthier.

4.2.2 Parameter a

The 5-number summary and the box-plot for the parameter a and the sex of the participants are given in the Figure 4.5:

PARAMETER "a"	FEMALES	MALES
MIN	-0.0158	-0.0233
Q1	-0.0104	-0.0136
MEDIAN	-0.0088	-0.0108
Q3	-0.0062	-0.0086
MAX	-0.0004	-0.0014

(a) Five-point summary for parameter a for females and males



(b) Box-plot for the parameter a for females and males

Figure 4.5: Distribution of parameter a by sex

The distributions of the parameter a with respect to female and male participants, are somewhat different. The minimum value for the male group (-0.0233) is smaller than the minimum value for the female group (-0.0158). Also, the median value as well as the max-

imum value for the distribution of the parameter a for the female are both higher than the corresponding values in the distribution for male. To check if there is a significant difference between the mean values of the parameter a in these two groups, we will complete the Welch's t -test. The procedure is same as for the parameter c , so for the purpose of conciseness only the descriptive statistics and the results of the test will be presented in the following table:

PARAMETER "a"	FEMALE	MALE
SAMPLE SIZE	47	50
MEAN	-0.008597872	-0.011064
VARIANCE	0.0000112933	0.0000164575
<i>t</i> -statistic	3.270	
<i>df</i>	94	
<i>Critical t-value</i>	1.986	

Figure 4.6: Descriptive statistics for female and male and the results of the Welch t -test for the parameter a

The mean value of the parameter a for females is -0.008597872 while the mean for males is -0.011064 indicating that the PHRV for the male participants decreases a bit faster than the PHRV of females. The variance of the parameter a for male is slightly higher so the values of the parameter a for males are a bit more scattered.

The accepted level of significance $\alpha = 0.05$ gives the same critical t -value of ± 1.987 . The calculated t -statistic for the parameter a is ($t=3.270$) and falls into the rejection region, so we can conclude that, at $\alpha = 0.05$ significance level, there is significant difference between the mean values of the parameter a for female and male. The test result is in line with the conclusion from the descriptive statistics completed earlier that indicates the steeper decay trend of PHRV for males, and therefore the difference in the mean values of parameter a for females and males.

4.3 Multiple Linear Regression

In this section, we will be developing the MLR models [26] to explore the possible linear relationship between the observed factors of HRV and the parameters c and a of PHRV.

4.3.1 Parameter c

A MLR model is created that takes into account all the 8 independent variables (HRV factors): *age, sex, SMBQ, length, weight, BMI, state and trait of anxiety*, against the parameter c of the PHRV. Prior making predictions based on this model, it is needed to check if the assumptions for the MLR model hold.

We first check the linearity condition. We have created a diagnostic Residuals vs. Fitted Values (Figure 4.7 (b)) plot which indicates that the residuals are fairly evenly spread around the zero line. Some exceptions are visible in a part with bigger fitted values, however, we could still say that the linearity assumption is reasonably satisfied.

We observe the distribution of the residuals. From the histogram (Figure 4.7 (a)), it can be seen that most of the residuals seem to be following the normal distribution, however, this is insufficient to confirm the normality.

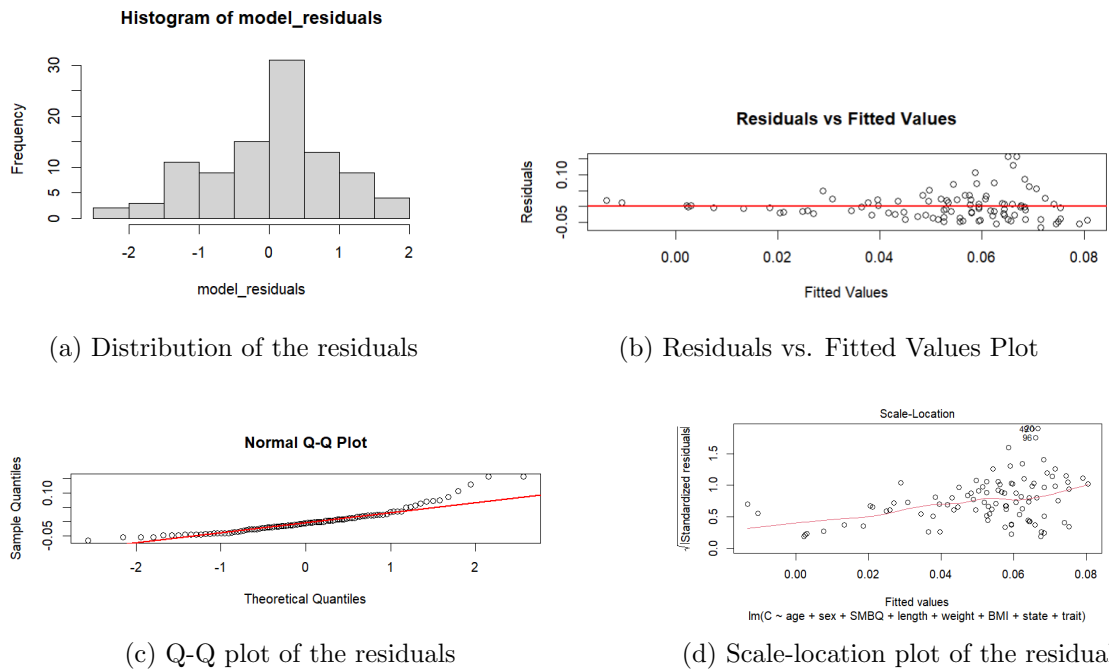


Figure 4.7: Distribution and diagnostic plots of the MLR model

To further explore if the assumption for normality of the residuals holds, Q-Q plot was created (Figure 4.7 (c)). The Q-Q plot shows that the residuals are approximately normal but there are some deviations in the tails. More specifically, the points in the right tail (upper quantiles) deviate more above the line which could indicate outliers.

To check the homoscedasticity assumption, we have created a diagnostic Scale-location plot (Figure 4.7 (d)). The residuals appear to follow the funnel shape across different fitted values. Therefore, we can say that the assumption of homoscedasticity is not satisfied. To deal with this issue, we will linearize our model (transform the dependent variable) by taking a natural logarithm of the both sides of the model.

The plot of residuals vs fitted values (Figure 4.8 (a)) for the transformed model shows better spread of the points around the x -axis for the assumptions of linearity. We can notice that the homoscedasticity is satisfied from the scale-location plot (Figure 4.8 (c)), as the funnel shape effect is now significantly smaller.

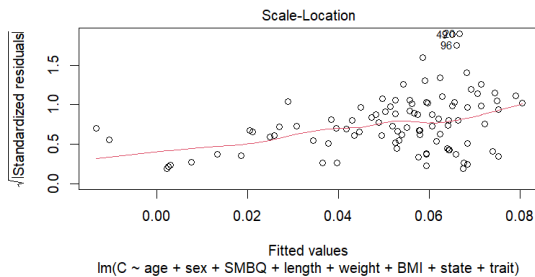
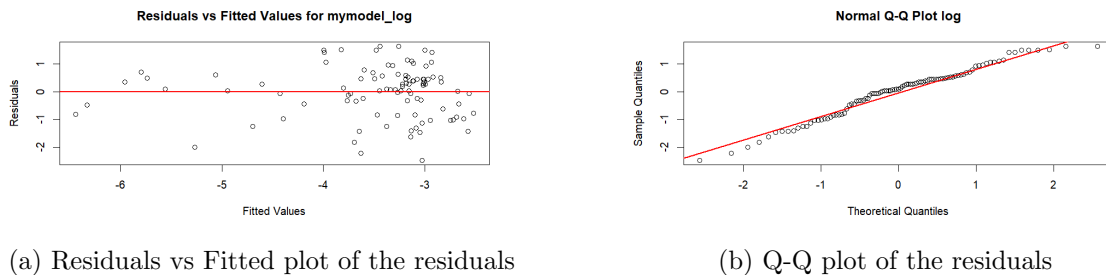


Figure 4.8: Diagnostic graphs for the assumptions for the logarithmic correction

Finally, we observe that the residuals in the Q-Q plot (Figure 4.8 (b)) have a consistent spread, indicating that the normality is satisfied. To check the assumption of multicollinearity of the linearized model VIF measurement was used (Figure 4.9 (a)).

	age	sex	SMBQ	length	age	sex	SMBQ	length
	1.159310	1.978198	3.540138	55.487043	1.145556	1.977001	3.489034	1.859044
	weight	BMI	state	trait	BMI	state	trait	
	162.486405	94.938297	2.793712	4.846022	1.155547	2.772260	4.788222	

(a) Table of VIF values for the logarithmic MLR model

(b) Table of VIF values for the reduced MLR model

Figure 4.9: Table of VIF values for the two MLR models

For the factors *age*, *sex*, *SMBQ*, *state* and *trait* the VIF values range from 1.159 to 4.846 indicating low to moderate multicollinearity, which is generally acceptable. However, the VIF values for predictors *length*, *weight* and *BMI*, are 55.487, 162.486, and 94.938, respectively. These values are extremely high and indicate severe multicollinearity. The VIF values suggest that these three factors are highly correlated with one another or with other predictors in the model which is a cause for concern, as high multicollinearity can lead to unstable coefficient estimates, inflated standard errors and reduced model interpretability. To address this issue, we decided to remove *weight* as a variable from the

linear model, as it had the highest VIF value. Hence, a new linear model was created, named the "reduced MLR model".

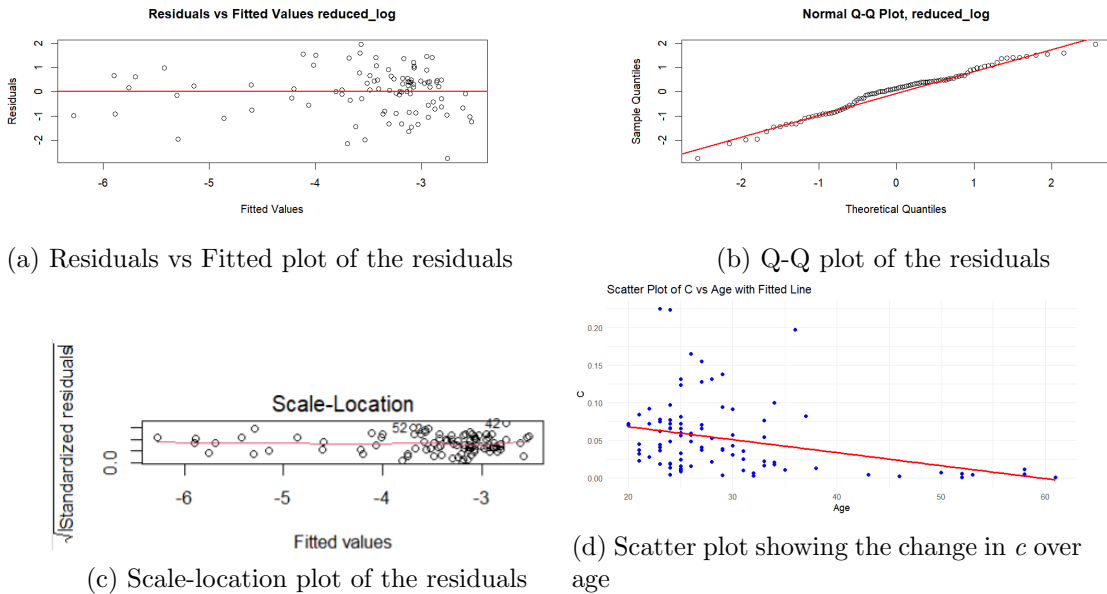


Figure 4.10: Diagnostic graphs for the assumptions for the reduced MLR model

We will briefly check the reduced model for the assumptions of linearity, homoscedasticity and normality. The residuals are fairly randomly scattered around zero (Figure 4.10 (a)), suggesting the linearity assumption is reasonably satisfied. They have a consistent spread along the horizontal line (Figure 4.10 (c)), indicating homoscedasticity is satisfied. The residuals are spread evenly along the line (Figure 4.10 (b)), suggesting normality.

The scatter plot of the parameter c vs factor *age* (Figure 4.10 (d)) shows a moderate negative correlation. Majority of the points follow the negative linear trend, meaning that the older the participant, the lower the value of the parameter c . There are a few higher values of c among younger participants, three of which could be seen as outliers. This can be explained by the potential individual associated diseases of these participants and are beyond the scope of our analysis.

After the reduction of the weight factor, a new table of VIF values was created and we can see that new values ranged from approximately 1.15 to 4.79 (Figure 4.9 (b)). Hence, the issue of multicollinearity has been resolved by removing *weight*. None of the remaining predictors have VIF values that indicate severe multicollinearity. This correction made it easier to interpret the effect of each predictor on the dependent variable and improved the accuracy of *p*-values, helping to determine the significance of each predictor.

Now that the assumptions have been checked, we will analyze how the *p*-values, R^2 and adjusted R^2 values of the reduced model affect the *c* parameter of PHRV (Figure 4.11).

```

Residuals:
    Min       1Q   Median       3Q      Max
-2.7450 -0.6803  0.1220  0.5302  1.9428

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.52833    2.86438  -1.232   0.221
age          -0.08672    0.01111  -7.810 1.07e-11 ***
sexMan       -0.06639    0.27470  -0.242   0.810
SMBQ         -0.04481    0.15265  -0.294   0.770
length       0.01704    0.01491   1.143   0.256
BMI          -0.00815    0.03383  -0.241   0.810
state        -0.02908    0.02153  -1.351   0.180
trait         0.02473    0.02117   1.168   0.246
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9616 on 89 degrees of freedom
Multiple R-squared:  0.4383,    Adjusted R-squared:  0.3941
F-statistic: 9.922 on 7 and 89 DF,  p-value: 4.251e-09

```

Figure 4.11: Summary statistics of the parameter *c* coefficients, *p*-values, (R^2) value and adjusted (R^2) value in the reduced MLR model

It can be seen that the p -value $< 4.251 \times 10^{-09}$, indicates a highly significant relationship. Therefore, at least one of the predictor variables is significantly related to the outcome variable, parameter *c* of PHRV. When looking at the *p*-values in the reduced model, the only significant predictor of PHRV is *age*. Parameter *state* has a second smallest *p*-value = 0.180, and, although technically insignificant, might be interesting to be investigated further. The value of the coefficient of determination $R^2=0.4383$ indicates that approximately 43.83% of the variability in PHRV is explained by the predictors. The adjusted

R^2 is 0.3941, so approximately 39.41% of the variability in PHRV is explained by the predictors (after adjusting for the number of predictors). The adjusted R^2 value is lower than the R^2 value, so some of the predictors in the model may not be contributing much to explain the variance in HRV.

4.3.2 Parameter a

Now, we will investigate a MLR model that represents the relationship between the parameter a of PHRV and the factors in "reduced MLR model" (Figure 4.12).

```

Residuals:
      Min       1Q   Median       3Q      Max
-0.0121732 -0.0024831 -0.0001178  0.0025794  0.0097116

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.132e-03  1.110e-02   0.823   0.4127
age          8.732e-07  4.302e-05   0.020   0.9838
sexMan      -1.309e-03  1.064e-03  -1.230   0.2219
SMBQ        7.599e-04  5.913e-04   1.285   0.2021
length     -1.047e-04  5.776e-05  -1.813   0.0732
BMI         3.831e-05  1.310e-04   0.292   0.7707
state      -5.606e-05  8.340e-05  -0.672   0.5032
trait      -3.876e-05  8.201e-05  -0.473   0.6376
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.003725 on 89 degrees of freedom
Multiple R-squared:  0.1618,    Adjusted R-squared:  0.09584
F-statistic: 2.454 on 7 and 89 DF,  p-value: 0.0239

```

Figure 4.12: Summary statistics of the coefficients of the parameter a , p -values, R^2 value and adjusted R^2 value in the reduced MLR model

The p -values in the reduced model indicate that no factor is statistically significant predictor for parameter a . The R^2 value is 0.1618 so, only 16.18% of the variability in PHRV is explained by the predictors. The low value of the R^2 might also indicate that the linear regression gives a poor model, and the relationship between the factors of HRV and the parameter a in the PHRV model are not necessarily in linear relationship. The adjusted R^2 is 0.09584, so approximately only 9.584% of the variability in PHRV is explained by the predictors. The adjusted R^2 is lower than the R^2 value, so some of the predictors may not be contributing to explain the variance in PHRV.

In conclusion, MLR results indicate that, for parameter c the only statistically significant factor is *age*. It might be interesting to further explore the relationship between parameter c and the factor *state*, that is potentially significant. MLR analysis for the parameter a indicates that none of the observed factors are statistically significant. Therefore, in the next section, additional analysis will be conducted, focused only on the possible effects of the two factors (*age* and *state*) on the parameters a and c of the PHRV model.

4.4 MLR with interaction term

The conclusions of the MLR model, with the factor *age* being the only statistically significant and the factor *state* being insignificant but with the second lowest p -value in the MLR model, prompted further investigation of their relationship with PHRV. Hence, a MLR model for parameter c using factors (*age* and *state*) with their interaction was created. Before looking into the effect of interaction term onto c parameter, we will illustrate the distribution of the factors *age* and *state* (Figure 4.13):

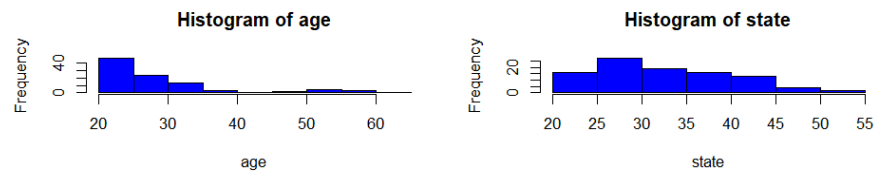
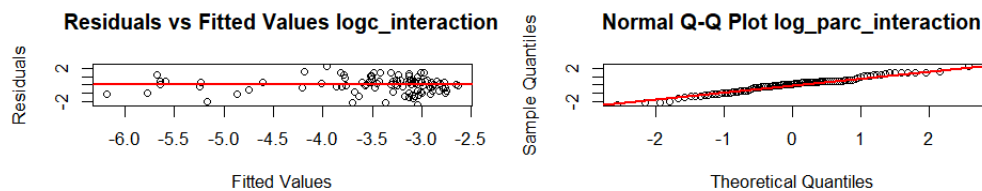


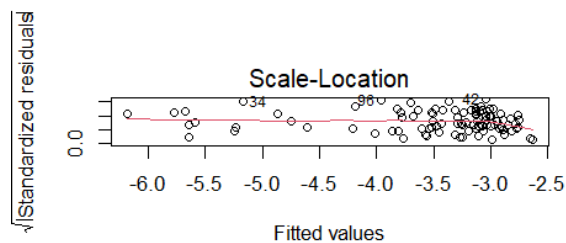
Figure 4.13: Histograms of factors *age* and *state* of the exponential decay model

The distribution of the factor *age* is strongly positively skewed, indicating that the median age is lower than the mean. The sample consists of mostly younger participants (up to 30 years) and several older than 30. This is a possible motivation for some future research that could be focused on specific age groups. Samples can be taken with respect to sex and age groups as they can provide more insight into specific relationships between the parameters of the model for PHRV and the observed factors. The histogram for the distribution of the factor *state of anxiety* is a bit more symmetrical, with a long tail on the right side. Also, *state of anxiety* factor is higher among younger participants.

In the previous section it was shown that taking a natural logarithm of the dependent variable (linearization of the model) provides better approximations. Therefore, we have created a MLR model using the logarithmic correction to look into the main factors (*age* and *state*) and their interaction of the parameter c . Firstly, we shall check the assumptions of linearity, homoscedasticity and normality.



(a) Residuals vs Fitted plot and Q-Q plot of the MLR model with interaction



(b) Scale-location plot of the residuals

Figure 4.14: Diagnostic graphs for the logarithmic correction model with interaction term

The residuals seem randomly scattered around zero on Residuals vs Fitted values, and the residuals are spread evenly along the diagonal on the Q-Q plot suggesting the linearity and normality assumption are satisfied (Figure 4.14 (a)). Residuals show a constant spread (Figure 4.14 (b)), indicating that homoscedasticity is also satisfied.

We will now look into the results of the MLR model with interaction term [27], considering the factors *age* and *state of anxiety* (Figure 4.15).

```

Residuals:
    Min       1Q   Median       3Q      Max
-2.4491 -0.6843  0.1116  0.4674  2.3407

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.2221792   1.3428314  -0.910   0.3651
age          -0.0658867   0.0388431  -1.696   0.0932 .
state         0.0104179   0.0426329   0.244   0.8075
age:state    -0.0007018   0.0013061  -0.537   0.5923
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9605 on 93 degrees of freedom
Multiple R-squared:  0.4144,    Adjusted R-squared:  0.3955
F-statistic: 21.94 on 3 and 93 DF,  p-value: 7.893e-11

```

Figure 4.15: Summary of the MLR analysis for factors *age*, *state* and their interaction

The multiple linear regression equation, with interaction term ($\beta_3(x_1x_2)$) between two predictors: *age* (x_1) and *state* (x_2), can be written as follows:

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3(x_1x_2) + \epsilon.$$

By substituting the obtained values for the coefficients in the MLR model with interaction term, we obtain:

$$y = -1.2221792 - 0.0658867x_1 + 0.0104179x_2 - 0.0007018(x_1x_2) + \epsilon.$$

The coefficient $\beta_3 = -0.0007018$ represents the decrease in age factor for one unit increase of the state factor. The p -values in the interaction model indicate that no factor is statistically significant predictor for parameter c . The R^2 value is 0.4144 so, 41.44% of the variability in PHRV is explained by the predictors. This value also indicates that a linear model is relatively good fit.

The adjusted R^2 is 0.3955, so approximately only 39.55% of the variability in PHRV is explained by the predictors. The adjusted R^2 is lower than the R^2 value, hence, we could potentially remove some predictors which are not contributing to explaining the variance in PHRV.

Finally, the Figure 4.16, shows the color-coded interaction graph which helps visualizing and interpreting the interaction effect when the effect of one independent variable (*age*) on the dependent variable (parameter *c*) depends on the level of another independent variable (*state of anxiety*). The graph shows the parameter *c* vs *age*, colored by the *state of anxiety* variable which is classified into 5 categories, each represented by a distinct color. The colors represent state categories, with category 1 being the lowest range of state values and category 5 being the highest. Each line represents the relationship between *c* and *age* for a specific *state* category.

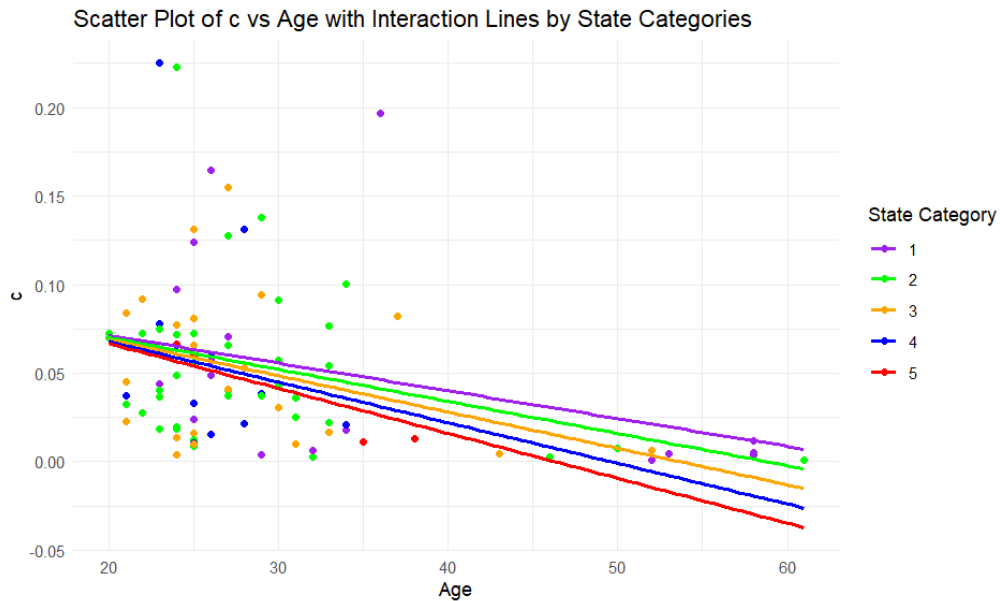


Figure 4.16: Scatter plot of *c* against *age*, colored by *state*

We can observe a low negative correlation between the parameter *c* and the *age* of participants, indicating that the parameter *c* decreases as *age* increases, across all state categories. The *y*-intercepts differ somewhat for each state category meaning that each state category starts at a bit different value of the parameter *c* for the minimum age of 20. The gradients of the lines are relatively consistent across different categories of *state of anxiety* showing a similar rate of decrease, with the overall highest rate of decrease for the highest category (5). The parameter *c* for younger participants (age 20-35) is mainly between 0 and 0.16 where just a few of them have significantly higher values (outliers).

This suggests that other factors might influence the value of the parameter c . On the other hand, older participants (age 45-65) have lower values of the parameter c , (up to 0.025) and a low level of the factor *state of anxiety*, (1-3). This finding makes sense in the context of everyday stress level that younger population is exposed to, while older people often lead relatively stress-less lives.

Chapter 5

Conclusion

In this thesis, the analysis for HRV signals in frequency domain has been presented. HRV is the physiological phenomenon of variation in the time interval between heartbeats. It is measured by the variation in the “RR interval”. The information about the behaviour of HRV is relevant to many cardiovascular and non-cardiovascular diseases, as well as emotional arousal in mental and social aspects.

In the analysis the Power Spectral Density (PSD) for HRV signals is estimated based on Discrete Fourier Transform (DFT). The spectrogram of respiratory and HRV signals is implemented for the estimation of the power of HRV (PHRV) in the HF band and an exponential model of PHRV is developed. Two parameters, c and a of the exponential model of PHRV are extracted: parameter c which represents the PHRV multiplier and parameter a - the exponential decay factor of the PHRV model. The relationship between the two parameters of PHRV and 8 factors (*age, sex, SMBQ, length, weight, BMI, state, and trait of anxiety*) is observed.

The results of the Welch’s t -test for the difference of two means of the parameter c between females and males show that there is no significant difference between the mean values of parameter c for female and male groups in this sample. Possible further research could investigate if there is a significant difference between the mean value of parameter c between female and male in a sample with older participants. Due to other potential age-related health issues, one might expect a more visible difference between the mean values of parameter c for females and males. The Welch’s t -test for the parameter a shows that

there is a difference between the mean values of this parameter for females and males. MLR model showed that out of 8 tested independent factors, only the factor *age* is significant for the parameter *c*. The factor *state of anxiety* is not significant, however, it has the second smallest *p*-value in the MLR model for the *c* parameter. This was the base to develop MLR model with interaction term. The results obtained using this model indicate that the interaction between the factors *age* and *state* is insignificant for the parameter *c*. As far as the parameter *a* is concerned, the MLR model developed at 5% significance level, shows no significant factors for parameter *a* in the PHRV exponential model.

Further research can be based on a variety of different samples, for example, separate younger and older female/male samples as well as consideration of the relationship between parameters of the exponential model of PHRV with respect to other factors that include specific health-issues, such as atrial fibrillation or other conditions.

In this thesis, we have treated the data as stationary, which is the biggest approximation made in the analysis. Applying non-linear methods will allow to visually search for patterns within a time series (a sequence of values from successive measurements). A possible choice of the non-linear methods would be Detrended fluctuation analysis (DFA) and Adaptive fractal analysis (AFA). DFA determines the level of non-stationarity of the data. It extracts the correlations between successive RR intervals over different time scales. Research has shown that DFA might be more precise in predicting fatal arrhythmic events than that based on traditional methods [7]. AFA utilizes an adaptive detrending algorithm to extract globally smooth trend signals from the data and then analyzes the scaling of the residuals to the fit as a function of the time scale at which the fit is computed. A parameter called *Hurst exponent*, (*H*) quantifies the “memory” or serial correlation in a time series. Advantage of AFA compared to DFA is that AFA can deal with arbitrary, strong non-linear trends while DFA cannot [8].

Results obtained using non-linear methods could lead to more precise mathematical models for the trend of the PHRV, and therefore, better analysis of the relationship between the observed parameters of the model and the factors affecting the HRV.

Chapter 6

Bibliography

[1] U.R. Acharya, et al. Heart Rate Variability: A Review, *Medical and Biological Engineering and Computing*, 44(12):1031-51, January 2007,

DOI: 10.1007/s11517-006-0119-0

[2] J. H. Medicine, “Anatomy and function of the heart’s electrical system,” *Johns Hopkins Medicine*, <https://www.hopkinsmedicine.org/health/conditions-and-diseases/anatomy-and-function-of-the-hearts-electrical-system> (accessed May 13, 2024).

[3] A.A. N. Moh et al. Electrocardiogram, March 2021, DOI: 10.13140/RG.2.2.28154.41920

[4] J. A. Waxenbaum, “Anatomy, autonomic nervous system,” *StatPearls [Internet]*., <https://www.ncbi.nlm.nih.gov/books/NBK539845/> (accessed May 13, 2024).

[5] A. Ledell, “Relating Heart Rate Variability and Age Using Spectral Estimation Methods,” thesis, Centre for Mathematical Sciences, Lund, 2019

[6] A. A. Fedotov, S. A. Akulov, and E. V. Timchenko, “Methods of mathematical analysis of Heart Rate Variability,” *Biomedical Engineering*, vol. 54, no. 3, pp. 220–225, Sep. 2020. doi:10.1007/s10527-020-10008-4

[7] J.Y. Chiang et al., “Detrended fluctuation analysis of heart rate dynamics is an important prognostic factor in patients with end-stage renal disease receiving peritoneal dialysis,” *PloS one*, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4734614/> (accessed Jun. 11, 2024).

[8] J. Gao et al., “Multiscale analysis of heart rate variability in non-stationary environments,” *Frontiers*, <https://www.frontiersin.org/journals/physiology/articles/10.3389/fphys.2013.00119/full> (accessed Jun. 11, 2024).

[9] E. Gospodinova et al. Spectral analysis of heart rate variability, International conference on Automatics and Informatics, Sofia, Bulgaria, 2015, Volume: ISSN 1313-1869, DOI: 10.33422/researchconf.2019.12.894

[10] S. W. Smith, “Chapter 8: The Discrete Fourier Transform,” in *The Scientist and Engineer’s Guide to Digital Signal Processing*, First Edition., California Technical Pub, 1997 : url:<https://www.dspguide.com/ch8/3.htm>

[11] B. W. Johnston, R. Barrett-Jolley, A. Krige, I. D. Welters, Heart rate variability: Measurement and emerging use in critical care medicine, *Journal of Intensive Care Society*, 21 (2), June 2019, DOI: 10.1177/1751143719853744

[12] P. Stoica and R. Moses, *Spectral Analysis of Signals*. New Jersey: Prentice Hall, Inc., 2005.

[13] S. Rapuano and F. J. Harris, “An introduction to FFT and time domain windows,” in *IEEE Instrumentation and Measurement Magazine*, vol. 10, no. 6, pp. 32-44, December 2007, doi: 10.1109/MIM.2007.4428580. (accessed Jun. 11, 2024).

[14] D. J. Jwo, W. J. Chang, I. H. Wu, *Windowing Techniques, the Welch Method*

for Improvement of Power Spectrum Estimation, Computers, Materials and Continua, 67(3):3983-4003, January 2021, DOI: 10.32604/cmc.2021.014752

[15] A. E. Aubert et al, The analysis of heart rate variability in unrestrained rats. Validation of method and results, Computer Methods and Programs in Biomedicine, 60(3):197-213, December 1999, DOI: 10.1016/S0169-2607(99)00017-6

[16] J. A. Vila, J. Presedo, M. Fernandez-Delgado and S. Barro, "Spectral analysis of HRV by means of time frequency distributions: application to ischemia analysis," Proceedings of 18th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Amsterdam, Netherlands, 1996, pp. 1628-1630 vol.4, doi: 10.1109/IEMBS.1996.647585

[17] R. M. West, Best practice in statistics: Use the Welch t-test when testing the difference between two groups. Annals of Clinical Biochemistry. 2021;58(4):267-269. doi:10.1177/0004563221992088

[18] J. Frost, "Independent samples t-test: Definition, using & interpreting," Statistics By Jim, <https://statisticsbyjim.com/hypothesis-testing/independent-samples-t-test/> (accessed Jun. 11, 2024).

[19] R. Tay, Correlation, Variance Inflation and Multicollinearity in Regression Model, Journal of the Eastern Asia Society for Transportation Studies, Vol.12, January 2017. DOI: 10.11175/easts.12.2006

[20] A.A. Suleiman, U.A. Abdullahi, U.A. Ahmad, An Analysis of Residuals in Multiple Regressions, International Journal of Advanced Technology in Engineering and Science, Volume No 03, Special Issue No. 01, April 2015

[21] J. W. Osborne, E. Waters, Four Assumptions of Multiple Regression That Researchers Should Always Test, *Practical Assessment, Research and Evaluation*, 8(2), <http://ericae.net/pare/getvn.asp?v=8n=2>

[22] S.W. Kim, H.Y. Park, H. Jung, S.A. Park, and K. Lim, “Development of a heart rate variability prediction equation through multiple linear regression analysis using physical characteristics and heart rate variables,” *Inquiry: a journal of medical care organization, provision and financing*, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10201176/> (accessed Jun. 27, 2024).

[23] R. Anderson, P. Jönsson, and M. Sandsten, “Effects of age, BMI, anxiety and stress on the parameters of a stochastic model for heart rate variability including respiratory information,” Lund University, <https://portal.research.lu.se/en/publications/effects-of-age-bmi-anxiety-and-stress-on-the-parameters-of-a-stoc> (accessed Jun. 11, 2024).

[24] G. Lindgren, H. Rootzen, and M. Sandsten, *Stationary stochastic processes for scientists and engineers*. CRC Press, 2013

[25] R. Pilacho, “Measuring Breathing Patterns through Strain Gauges,” BMEG442 Engineering Exercise and Sports, <https://sites.udel.edu/coengex/2017/04/06/measuring-breathing-patterns-through-strain-gauges/> (accessed Jun. 4, 2024).

[26] “Multiple coefficient of determination,” Multiple Coefficient of Determination — R Tutorial, <https://www.r-tutor.com/elementary-statistics/multiple-linear-regression/multiple-coefficient-determination> (accessed Jun. 11, 2024)

[27] A. Shiroshita, N. Yamamoto, N. Saka, H. Shiba, S. Toki, M. Yamamoto, E. Dohi, Y. Kataoka, Expanding the Scope: In-depth Review of Interaction in Regression Models, *Ann Clin Epidemiol.* 2023 Jun 10;6(2):25-32. doi: 10.37737/ace.24005.

Bachelor's Theses in Mathematical Sciences 2024:K18
ISSN 1654-6229
LUNFMS-4077-2024
Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden
<http://www.maths.lu.se/>