TECHNISCHE UNIVERSITÄT WIEN

DIPLOMA THESIS

Statistical Modelling of the Relationship between Fatigue and Physiological Parameters of the Cardiovascular System

submitted in satisfaction to the requirements of the degree of

Diplom-Ingenieur

in

Technical Mathematics

to the

Institute of
Analysis und Scientific Computing
TU Wien

supervised by

Ao.Univ.-Prof.i.R. Dipl.-Ing. Dr.techn. Felix Breitenecker

and

Univ.Lektor Dipl.-Ing. Dr.techn. Martin Bachler

by

Ciara Pircher

Registration Number: 01427553

Vienna, on September 5, 2023 _____ _____

Kurzfassung

Müdigkeit führt zu ca. 7% der Unfälle auf Europas Straßen und ca. 13% der unfallbedingten Verletzungen. Trotzdem spielt für die Feststellung von Fahrtüchtigkeit Müdigkeit nur eine untergeordnete Rolle. Ziel dieser Arbeit ist es, im Rahmen des EU-geförderten PANACEA Projekts ein Modell zur Abschätzung von Müdigkeit durch physiologische Parameter des Herz-Kreislaufsystems auf der KSS, der Karolinska-Schläfrigkeitsskala, zu entwickeln.

Herzratenvariabilität (HRV) und Pulswellenform stehen unter Einfluss des autonomen Nervensystems, das Herzrate, Blutdruck und andere Körperfunktionen reguliert und durch Müdigkeit beeinflusst wird. In dieser Arbeit basiert die Modellierung dieses Zusammenhangs auf aufgezeichneten Studiendaten. Jeder der 32 männlichen Studienteilnehmer absolvierte sechs simulierte Fahrten. Elektrokardiographie und Photoplethysmographie wurden jeweils vor und nach jeder Fahrt aufgenommen. Frühere Studien belegen Zusammenhänge zwischen den 23 für diese Arbeit extrahierten HRV- und Pulswellenparametern und Müdigkeit.

Für die Modellierung dieses Zusammenhangs wurde in dieser Arbeit zur Reduktion der Variablen eine Hauptkomponentenanalyse (PCA) mit anschließender multivariater linearer Regression durchgeführt. In einem weiteren Ansatz wurden mit schrittweiser Regression Modelle generiert. Um individuelle physiologische Unterschiede zu erfassen, wurden für manche Modelle individuelle Baselines für die Parameter des Herzkreislaufsystems für jeden Studienteilnehmer definiert.

Zwölf Modelle wurden mit PCA oder schrittweiser Regression erstellt und anhand der Wurzel des mittleren quadratischen Fehlers (RMSE), der Residuen und statistischer Signifikanz von Variablen verglichen. Modelle, die mittels schrittweiser Regression mit Baselines erstellt wurden, liefern, vor allem im Bezug auf Signifikanz, bessere Ergebnisse. Die wichtigsten unabhängigen Variablen sind der Quotient $\frac{\mathrm{LF}}{\mathrm{HF}}$ zwischen niedrigen und hohen Frequenzanteilen der HRV und die Herzrate. Die Pulswellenparameter zeigen zwar insgesamt weniger Zusammenhang, aber die absolute und relative Zeit der Systole sind ebenfalls von Bedeutung. Abgesehen von der Herzrate stützen frühere Studien die berechneten Veränderungen dieser Variablen allerdings nicht.

Das final vorgeschlagene Modell kombiniert die Ergebnisse zweier, mit schrittweiser Regression mit dynamischen Baselines erstellter, Modelle. Es nutzt eine Konstante (0.3155) und fünf Variablen mit den folgenden Koeffizienten: $\frac{\mathrm{LF}}{\mathrm{HF}}$ (0.2751^*), relative Zeit zur dikrotischen Kerbe (8.4928), Verhältnis des zweiten Maximums der Pulswelle zu systolischem Druck (5.383), Zeit der Systole (-0.0160^*) und Pulsankunftszeit (0.0598^*), wobei signifikante Koeffizienten mit * gekennzeichnet sind. Die beiden einfließenden Modelle sind signifikant ($p = 0.0035$ und $p = 0.0143$) und deren Kombination hat ein adjustiertes Bestimmtheitsmaß von 0.17. Diese Arbeit zeigt, dass die Abschätzung von Müdigkeit anhand HRV- und Pulswellenparameter realisierbar ist. Es sollten jedoch weitere Studien mit mehr Messungen an einer vielfältigeren Teilnehmergruppe durchgeführt werden.

Abstract

Around 7% of European road accidents and around 13% accident-related injuries can be linked to driver fatigue. Nonetheless, to date fatigue only plays a minor role when determining a driver's ability to drive. This thesis aims to develop a model to estimate fatigue from parameters of the cardiovascular system on the Karolinska Sleepiness Scale in the context of the EU-funded PANACEA project.

Heart rate variability (HRV) and the pulse wave shape are influenced by the autonomic nervous system, which controls heart rate, blood pressure and other bodily functions and is affected by fatigue. In this thesis, modelling of this relationship is based on recorded study data. In total, 32 male participants completed six driving simulations each. Both electrocardiography and photoplethysmography measurements were recorded before and after driving. Previous studies have shown that the 23 HRV and pulse wave shape parameters that were extracted for this thesis are affected by fatigue.

To reduce the large number of predictors for modelling the relationship in this thesis, a principal component analysis (PCA) and subsequent multivariate linear regression were performed. A second approach using stepwise variable selection was also followed. To account for individual differences in physiology, individual participant baselines for HRV and pulse wave parameters were introduced.

Twelve models using either PCA or stepwise regression were generated and compared with respect to root mean square error (RMSE), residuals and statistical significance of variables. Overall, models created using stepwise regression and any kind of baseline yield better results, especially concerning significance. The most valuable predictors are the ratio $\frac{\mathrm{LF}}{\mathrm{HF}}$ between low and high frequency components of HRV and heart rate. Even though pulse wave shape parameters are less important, absolute and relative systolic time are promising predictors. However, apart from the heart rate, the estimated change in these parameters is not backed by previous studies.

Finally, the proposed model combines the results of two models, both generated using stepwise linear regression with a dynamic baseline. It includes a constant (0.3155) and five variables with the following coefficients: $\frac{\mathrm{LF}}{\mathrm{HF}}$ (0.2751^*), relative to the dicrotic notch (8.4928), quotient of dicrotic wave amplitude and systolic pressure (5.383), systolic time (-0.0160^*) und pulse arrival time (0.0598^*), where significant coefficients are marked with *. Overall, both contributing models are considered significant ($p = 0.0035$ und $p = 0.0143$) and their combination has an adjusted R^2 value of 0.17. These results indicate that fatigue prediction from HRV and pulse wave shape parameters is feasible. However, further research with a higher number of measurements and a more diverse participant group is necessary.

Acknowledgement

First of all, I extend my gratitude to my supervisors, Ao.Univ.-Prof.i.R. Dipl.-Ing. Dr.techn. Felix Breitenecker and Univ.Lektor Dipl.-Ing. Dr.techn. Bakk.techn. Martin Bachler. While both played important roles in shaping this thesis, I am especially grateful for Martin's exceptional guidance throughout the process. I could not have wished for better supervision.

I would also like to thank the Austrian Research Promotion Agency (FFG) for awarding me the FEMtech scholarship for a paid internship at the Austrian Institute of Technology (AIT). This enabled me to dedicate myself fully to this thesis while reducing financial concerns at the same time. I greatly benefited from the months I spent working at the AIT. I am especially grateful for the welcoming atmosphere, the experience I was able to gain, and the input from colleagues concerning my research question.

My parents also offered me a great deal of financial support throughout my entire studies. I greatly appreciate their contribution to my academic journey, as their support facilitated reaching my educational goals.

I want to express my appreciation to my friends and colleagues for their moral support and companionship during my time at university. While sometimes the countless coffee breaks went a long way in maintaining my motivation, other times you helped me stay disciplined and focussed. In particular, I want to thank Nico for guiding me towards invaluable literature that significantly enriched my research and enabled me to reach a better understanding of the subject matter.

Most importantly, I am immensely grateful for everything Michi has done for me in the past few years. His unwavering moral support, as well as his assistance with everyday tasks in stressful times, were the cornerstones of my journey. Your belief in me surpassed my own, and your encouragement motivated me to soldier on, even in the most challenging moments.

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Diplomarbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Wien, am 5. September 2023

Ciara Pircher

Contents

HRV	heart rate variability
PW	pulse wave
AIT	Austrian Institue of Technology
VTI	Swedish National Road and Transport Research Institute
KSS	Karolinska sleepiness scale
EEG	electroencephalography
EOG	electrooculography
PPG	photoplethysmography
ECG	electrocardiography
HR	heart rate
bpm	beats per minute
PAT	pulse arrival time
ANS	autonomic nervous system
SNS	sympathetic nervous system
PNS	parasympathetic nervous system
RRI	R-peak to R-peak interval
NNI	normal-to-normal interval
PWA	pulse wave analysis
SBP	systolic blood pressure
DBP	diastolic blood pressure
TPD	total pulse duration
BAC	blood alcohol content
RSS	residual sum of squares
RSE	residual squared error
RMSE	root mean square error
AIC	Akaike information criterion
BIC	Bayesian information criterion
PC	principal component
PCA	principal component analysis
PCR	principal component regression
EDF	epirical distribution function

1 Introduction

To date, when assessing fitness to drive, predominantly impairment due to alcohol consumption or drug use is measured. However, fatigue can majorly affect a driver's vigilance and lead to loss of attention or falling asleep at the wheel. According to a study conducted in 19 different European countries, 17% of drivers report having fallen asleep while driving. Approximately 7% of European road accidents as well as 13.2% of accident-related injuries and 3.6% of fatalities can be linked to fatigue. In comparison to other European countries, Austria has a high rate of falling asleep while driving. Therefore, any assessment system, which can alert a driver to possible high fatigue, has the potential to reduce the number of accidents, injuries and deaths on Europe's roads. [19]

In the context of the EU-funded PANACEA project, which aspires to develop a more comprehensive system to assess fitness to drive, this thesis aims to develop a predictive model to estimate fatigue from cardiovascular parameters. The PANACEA assessment system is intended for use in the field of commercial driving, as with bus or lorry drivers. Although it will also take other factors such as alcohol, drugs, stress and distraction into account, this thesis focusses only on fatigue assessment through heart rate variability (HRV) and pulse wave shape parameters.

The variability in time between successive heartbeats is strongly connected to the activity of the autonomic nervous system (ANS) [46]. As, with rising fatigue, the balance between the sympathetic and parasympathetic branches of the ANS shifts, certain HRV parameters are affected and could therefore be useful in prediction [14]. The connection between pulse wave shape parameters and fatigue is less clear, but there are studies, that show that effects of rising fatigue can be linked to changes in the pulse wave shape [27].

The characteristics of HRV and the pulse wave can be extracted from electrocardiography (ECG) and photoplethysmography (PPG) measurements. In this context, some background is given on the physiology of the cardiovascular system and the functionality of the measurement methods in use. Data is measured using the Austrian Institute of Technology's (AIT) proprietary device, the SmartPWA, and processed by previously existing code in MATLAB (The MathWorks Inc., Natick, USA). Since all characteristics are then output in MATLAB, this thesis made use of MATLAB R2022b for all necessary modelling and programming.

Within the scope of this thesis, the available cardiovascular parameters were analysed and in further course used as variables in multivariate regression models. The number of parameters computed from ECG and PPG measurements is too large to feasibly include all in a predictive model, which raises the question, which parameters attribute most to accurate prediction of fatigue. The modelling and statistical methods applied in this thesis are largely based on the books *An Introduction to Statistical Learning* (James et al., [21]) and *The elements of statistical learning: data mining, inference, and prediction* (Hastie et al., [20]). MATLAB's predefined functions for the dimension reduction technique prin-

cipal component analysis (PCA) as well as stepwise variable selection were put to use to determine the most valuable predictors and generate multivariate regression models.

Due to various anatomical factors, there can be large differences between individuals in HRV and pulse wave parameters [24, 25]. Since these parameters are used as predictors in the generated regression models, this can have a huge effect on the generality of the models. Therefore, two different versions of an individual baseline were pursued: a fixed baseline, with one reference measurement per participant, and a dynamic baseline, with changing reference measurements.

Twelve different models were generated using various combinations of the aforementioned modelling methods and different types of individual baselines. The models are evaluated using the F-test and residual analysis and compared with respect to quality of fit. The root mean square error (RMSE) and adjusted R^2 statistic are the main characteristics of goodness of fit computed for analysis.

The most important parameters in fatigue prediction are extracted from the generated models and compared to literature. The differences between the expected effect of fatigue on HRV and pulse wave parameters and the effect suggested by the generated models are highlighted and discussed.

While an effect of fatigue on HRV parameters is well established, the influence on pulse wave shape parameters is unclear. Even though many papers have been published on fatigue assessment, most of which use classification algorithms, no example of a predictive model mapping cardiovascular parameters to the Karolinska sleepiness scale (KSS) was found. This thesis aims to use fresh data to validate previous results from other studies concerning the effect of fatigue on HRV parameters, while also determining the value of pulse wave shape parameters in fatigue prediction. Finally, after considering different approaches, a model for predicting fatigue on the KSS is proposed.

2 Background

This chapter is dedicated to establishing a basis in the physiology connected to fatigue, the cardiovascular system and its parameters derived for analysis. First, the PANACEA Project, within which this thesis has been conducted, will be presented. The following sections contain a short introduction to the physiology of the cardiovascular system and an explanation of established characteristics used for its analysis. The basic functionality of measurement methods through which they are obtained, i.e. cardiography and photoplethysmography, will also be explained. Lastly, the definition and measurement methods of fatigue used in the PANACEA Project will be given.

2.1 The PANACEA Project

The PANACEA Project, short for "practical and effective tools to monitor and assess commercial drivers' fitness to drive", is concerned with defining and measuring professional drivers' fitness to drive. It is funded by the European Union as part of the research and innovation funding programme Horizon 2020, which ran from 2014 to 2020 [57]. The duration of the project is set to be three years, from May 2021 to April 2024. Within PANACEA, 16 partners from all across Europe, including the Austrian Institue of Technology (AIT) in Vienna, Austria, and the Swedish National Road and Transport Research Institute (VTI) in Linköping, Sweden, collaborate to achieve a higher standard of road safety. [35]

The main goals of the PANACEA project are to develop ways to assess the driving ability of commercial drivers, such as bus, taxi or lorry drivers, as well as defining appropriate counter-measures and policy recommendations. Although this thesis focusses solely on fatigue as a possible cause for driving impairment, throughout the project the influence of other factors, such as alcohol, drugs, stress and distraction, on one's ability to drive are also considered. Once a comprehensive monitoring and assessment system is developed, its sensitivity, specificity, effectiveness and operability should be estimated. Finally, the usefulness and acceptance of the developed methods should also be evaluated. [35]

2.2 Heart

This section will give an overview of cardiac anatomy and function in the context of the circulatory system. Also, the electric properties and conduction system of the heart will be described in order to understand the methods of ECG, which is used to measure heart activity. Finally, heart rate variability, derived metrics and their interpretation will be discussed.

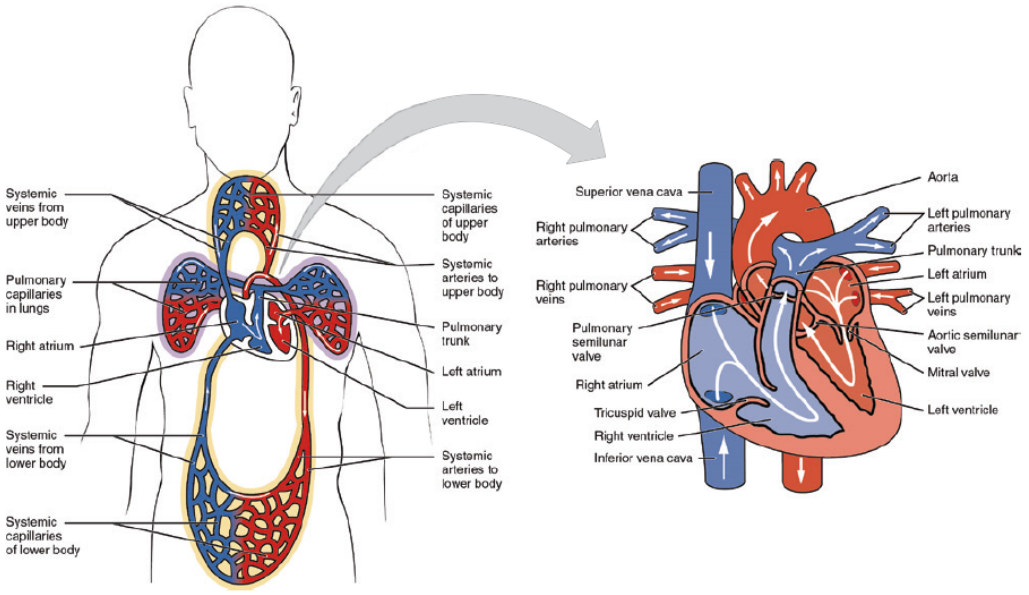

Figure 2.1: The heart's place within the circulatory system can be seen on the left. On the right, a more detailed view of the heart is given. Image adapted from [10, Ch. 7] under the license [13].

2.2.1 Physiology of the Heart

One could imagine the heart as a pump, which keeps the blood circulating within our bodies. A healthy adult heart displaces approximately 5 litres of blood every minute. In principle, the heart fills with blood and then, through muscular contraction, ejects the blood into the vascular system, through which it is transported all over the body. In general one distinguishes between veins, blood vessels leading to the heart, and arteries, blood vessels carrying blood away from the heart. [10, Ch. 7]

The heart is divided into four chambers, the right and left atria, which sit above the right and left ventricles, respectively. Figure 2.1 shows the heart with all corresponding annotations. It should be noted that, although the heart is usually depicted from the perspective of someone looking at a person, right and left are generally considered from the opposite perspective, as if one were speaking of their own heart. [10, Ch. 7]

The right and left sides of the heart are physically separated by a wall. While the right side of the heart is responsible for the movement of deoxygenated blood through the lungs, the left side distributes the oxygenated blood across the rest of the body. A detailed description of blood flow through the heart follows. [9, Ch. 26]

The right atrium collects blood low in oxygen from the veins leading to it. As it contracts, the blood goes through a valve to the right ventricle, from which it is pushed through the pulmonary valve into the artery leading to the lungs. The blood, now rich in oxygen, returns from the lungs and fills the left atrium. When the left atrium contracts, the blood passes through another valve into the left ventricle. From there it is pressed through the aortic valve into the aorta, the biggest artery of the body. It is then circulated to all other organs. An overview of the blood flow through the heart and circulatory system is indicated

by the white arrows in Figure 2.1. [9, Ch. 26]

The valves between the four chambers are in fact very important. They assure, that current can only flow in one direction. This allows a constant influx to the heart through the veins while, at the same time, there is a regular output of blood to the arteries. It is important to keep in mind, that the blood flow from the heart is not completely steady, but echoes the rhythmic pulse of the heart. [10, Ch. 7]

2.2.2 Electric Properties of the Heart

The previous section has made clear, that the contraction of the heart muscles is essential for the transport of oxygen through the circulatory system. To this end the muscles of the heart are electrically stimulated around 60 times per minute [32, Ch. 2]. This section will give an outline of the electrical activity that keeps our heart at the correct pace.

First of all, it is helpful to understand what is meant by the aforementioned stimulation. All cells, also the muscular cells of the heart, are electrically charged. When a cell is at rest it has a negative potential. It is considered to be polarised. Stimulation is the process of electrical change, to the depolarised state, in which the potential level within the cell becomes positive. This stimulation causes muscle contraction. [48, Ch. 1]

A small group of cells, called the sinus node or sinoatrial node, in the right atrium are the main source of electrical stimulation. The frequency, in which impulses are generated in the sinus node, can be influenced through the autonomic nervous system to adapt to certain needs. The sinus node along with the rest of the heart's conduction system is depicted in figure 2.2. [32, Ch. 2]

In a healthy heart the stimulation is triggered by the sinus node and spreads to neighbouring cells from there. The left atrium is depolarised shortly after the first impulse in the right atrium. In the cardiac conduction system, the ventricles are connected to the atria through the atrioventricular node, but are otherwise isolated. The impulse can only pass through this node and therefore the ventricles contract only after both atria. The conductors split into branches for the left and right ventricle after the atrioventricular node. The Purkinje fibres, which seperate from the left and right branches, carry the stimulation through the musculature. [32, Ch. 2]

Should the sinus node fail, the atrioventricular node as well as the Purkinje fibres are capable of generating impulses, albeit at a lower frequency and without any connection to the autonomic nervous system. [32, Ch. 2]

2.2.3 Electrocardiography

The heart displays a distinct, repeating pattern of electrical activity that can give a lot of insight into its current state of health. Naturally, measurement of the heart's electric processes has been of great interest for some time and methods to quantify them have been developed. This section gives a short introduction to current methods and the resulting graph, the electrocardiogram (ECG).

During depolarisation, when some cells are already stimulated, i.e. have a positive potential, whereas others have not yet experienced any electrical changes, there is a difference in potential between cells. This results in an electric current from one cell to another [32,

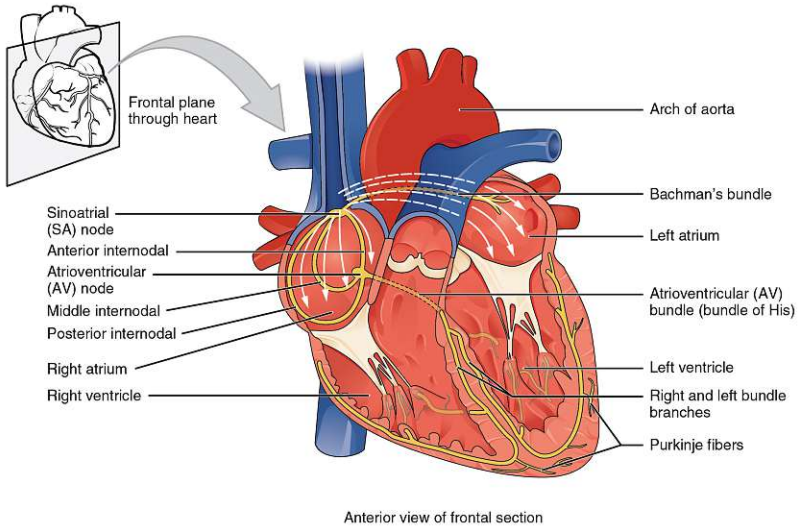

Figure 2.2: The conduction system of the heart is schematically depicted in yellow. Depolarisation starts in the sinoatrial node, spreads through the atria as indicated by the white arrows and then crosses to the ventricles through the atrioventricular node. The impulse is taken to the ventricular muscles through the Purkinje fibres. Image taken from [33] under the license [13].

Ch. 1]. It is the sum of these electric currents within the heart, or the sum of differences in potential, that is depicted in the electrocardiogram. [32, Ch. 3]

The differences in potential between cells can be treated as vectors [32, Ch. 3]. Summarising these partial voltages, we obtain the cardiac vector, a vector with varying direction and magnitute, which is defined to be fixed in the centre of the heart. The cardiac vector represents the collective electric activity of the heart. [48, Ch. 1]

Willem Einthoven was the first to suggest measuring cardiac electric activity during stimulation. He postulated, that the human body is a conductor and the activity of the heart could be detected on the skin. In his first experiments, he attached electrodes to both forearms and the left lower leg and measured the difference in potential between each combination of two of the three electrodes. He called each pair of electrodes a lead. The resulting curve represents the magnitude and direction of the cardiac vector in relation to time, projected onto the corresponding lead used in measurement. This means each lead can be attributed to a certain viewpoint, from which cardiac activity is monitored. For an illustration see figure 2.3. [32, Ch. 4]

Einthoven termed the measurement using both forearms lead I. He named the combination of electrodes on the right forearm and left lower leg lead II and the combination of the left lower leg and left forearm lead III. [32, Ch. 4]

Figure 2.4 shows the segment of an electrocardiogram which corresponds to a single heartbeat. The first wave of the pattern, the P wave, is caused by the depolarisation of the atria. As the ventricles become depolarised, the cardiac vector changes direction twice and also reaches its maximum in magnitude at the R peak, indicating the maximum of electric activity in the heart. This means ventricle depolarisation is seen in the QRS interval. The T wave is generated by the repolarisation, that is the return to the state of rest, of the

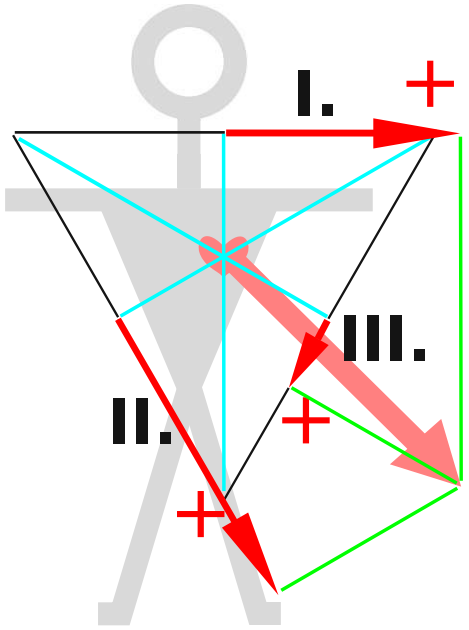

Figure 2.3: The image shows the three leads proposed by Einthoven in black. The light red arrow represents the cardiac vector, the sum of electrical activity in the heart. The projection of the cardiac vector onto each lead is shown in red. An Einthoven ECG measurement depicts the magnitude of the corresponding projection. Image used under the license [13].

ventricles. Although the ECG signal could look quite different when using another lead, the depicted signal is representative for the signals acquired in this thesis. [48, Ch. 1]

Since Einthoven's first experiments, there have been many developments in electrocardiography. Today, ECGs of a clinical standard are usually recorded using twelve leads. However, due to its simplicity, Einthoven's procedure is still in use in many wearable or portable ECG monitors. [48, Ch. 1]

2.2.4 Heart Rate Variability and Connected Metrics

The human heart does not beat in completely regular intervals. The ANS, which keeps the body in homoeostasis, a state of stable physical conditions, is constantly monitoring and correcting the heart rate to the needed pace via the sinus node. This means certain fluctuations in the time between two successive heartbeats are in fact healthy. This variance indicates that our bodies can quickly adapt to environmental change or stressors and shows a degree of resilience. [46]

The ANS is divided into two branches, the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS). A rise in sympathetic nervous activity is considered to be a "fight or flight" response to a stressful situation. It leads to higher heart rates and blood pressure, increased muscle strength and faster breathing [37]. The PNS on the other hand, is associated with rest, sleepiness and lower heart rates. The variability in time from one heartbeat to another gives insight to the balance in SNS and PNS activity. [11]

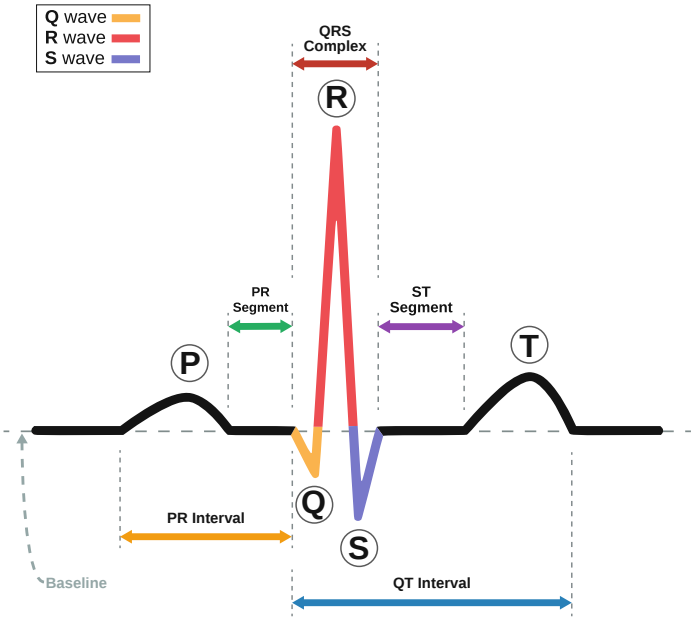

Figure 2.4: Segment of an ECG signal corresponding to one heartbeat. Depolarisation of the atria results in the P wave, while depolarisation of the ventricles corresponds to the QRS complex. The repolarisation of the ventricles can be seen in the T wave. Image used under the license [13].

For further reference, HRV is defined as follows:

Definition 2.1. ***Heart Rate Variability (HRV)*** *is the fluctuation of time intervals between consecutive heartbeats [46].* ***HRV Analysis*** *is the extensive analysis of beat-to-beat intervals in both time and frequency domains [48, Ch. 1].*

Since an R-peak in the ECG corresponds to a heartbeat, the time between consecutive R-peaks, the R-peak to R-peak interval (RRI), is of special interest for the analysis and quantification of HRV. Unfortunately, this is not a one to one correspondence. In preprocessing, R-peaks resulting from so-called ectopic beats, heartbeats that do not originate from the sinus node and therefore are not influenced by the ANS, as well as those considered to be outliers are removed. After preprocessing, the remaining intervals are referred to as the normal-to-normal intervals (NNIs) while the remaining heartbeats are called normal heartbeats. [28, 46]

Definition 2.2. ***Normal heartbeats*** *are heartbeats originating from the sinus node which are controlled by the ANS. The* ***normal-to-normal intervals (NNIs)*** *are given by the time differences between two successive normal heartbeats.*

The NNIs are usually measured in milliseconds (ms) so, obviously, the instantaneous heart rate, usually measured in beats per minute (bpm), can be directly calculated from the NNIs. We call all HRV measures calculated directly from the fluctuations in NNIs over time the time-domain measures of HRV. Table 2.1 gives a detailed description of all characteristics in the temporal domain used in this thesis as well as their physiological interpretations. [48, Ch. 1]

Definition 2.3. ***Time-domain measures of HRV*** *are measures calculated from the series of time differences between NN intervals. The time-domain measures used in this work are defined in table 2.1*

Table 2.1: Overview of time-domain HRV parameters derived from ECG data. A description and physiological interpretation of the different metrics is also given. Of course all measures are dependent on the recording time. Units are given in parenthesis. Adapted from [48] and [37] with additional information from [46].

HRV Parameter	Description	Physiological Interpretation
mean HR (bpm)	mean of HRs throughout a recording	mean heart rate
mean NN (ms)	mean of NNIs throughout a recording	mean duration between two normal heart beats
SDNN (ms)	Standard deviation of NNIs (i.e. the square root of variance of NNIs)	estimates PNS activity, especially that related to changes in HR caused by respiration
RMSSD (ms)	Root mean square of successive differences between normal heartbeats	estimates high frequency variability in the HR induced by the PNS and is correlated to pNN50
pNN50 (%)	Percentage of successive NNIs, that differ by more than 50 ms	pNN50 is correlated to high frequency HRV, PNS activity and RMSSD

For further analysis the NNI data can be examined in the frequency domain. One can observe the power spectrum of NNIs, which depicts the variance in NNIs as a function of frequency. Characteristics derived from the power spectrum of NNIs are referred to as frequency-domain measures. As before, all characteristics available for this thesis are summarised in table 2.2. A detailed description and corresponding physical interpretation is given:

Definition 2.4. ***Frequency-domain measures of HRV*** *are measures calculated from the power spectrum of the series of time differences between NN intervals. Frequency-domain measures used in this work are defined in table 2.2.*

The power spectrum can of course be obtained using the Fourier transformation, but it is important to note, that multiple different algorithms are in use to determine the power spectrum of NNIs. Unfortunately, their results and therefore also frequency-domain measures calculated by different means are not comparable. Regrettably, the means of calculation are mostly not included in publications, which makes the comparison of measured values hard.

Usually SDNN is considered an indicator for overall heart rate variability. Along with the physiological interpretation given in 2.1, this coincides with the suggestion that higher HRV is associated with higher parasympathetic activity while lower HRV is associated with stress. As seen in table 2.2, the high frequency component (HF) of HRV is connected to the PNS. Even though the low frequency (LF) component has been shown to reflect both sympathetic and parasympathetic activity, it is mostly considered to be influenced by the SNS. [8]

Table 2.2: Overview of frequency-domain HRV parameters derived from ECG data. A description and physical interpretation of the different metrics is also given. Units for the parameters are given in parenthesis. Adapted from [48] and [37] with information from [46] and [9, Ch. 38].

HRV Parameter	Description	Physiological Interpretation
LF (ms^2)	Absolute power in the low frequency band ($0.04 - 0.15$ Hz)	Influenced mostly by baroreceptor activity, an essential part of the ANS which keeps mean blood pressure constant. For slow breathing rates PNS influence is greater, SNS has a larger influence for faster breathing.
LFnorm (unitless)	absolute LF power divided by absolute power of LF+HF	
HF (ms^2)	Absolute power in the high frequency band ($0.15 - 0.4$ Hz)	Influenced by PNS and HR fluctuations related to breathing (at a respiration reate of 9-24 breaths per minute). Also, highly correlated to RMSSD and pNN50.
HFnorm (unitless)	absolute HF power divided by absolute power of LF+HF	
LF/HF ratio (unitless)	ratio of low frequency and high frequency power	Indicates the balance between SNS and PNS activity. The lower the ratio, the higher the parasympathetic influence compared to sympathetic activity.
TP (ms^2)	Absolute power in HF, LF and also very low frequency ($0.003 - 0.04$ Hz)	Total power across the spectrum (incl. ultra- and very low frequencies, which can only be extracted from longer recordings [28]).

According to the standards of measurements set by [28], all HRV measures introduced above can be extracted from a short term ECG recording of 2 to 5 minutes. However, it is important to compare only recordings of the same length, as some statistical measures, such as SDNN, are only well-defined in this case. This is due to the fact that a longer recording would also lead to a higher total variance. [28]

It should also be noted, that HRV parameters can differ greatly from person to person, as they are affected by multiple factors, such as age, height, weight and, of course, general health. All time domain indices as well as TP, LF and HF tend to be higher in males than females, whereas the mean HR seems to be higher for women. Height has been shown to correlate with lower heart rates while a greater weight has been associated with lower normalised HF, higher normalised LF and therefore also a higher $\frac{LF}{HF}$ ratio. Needless to say, this poses some challenges when trying to compare measurements from different people. [24, 25]

2.3 Blood Circulation System

We have already reviewed the basic action of the circulatory system while discussing the function of the heart within it in section 2.2.1. This section will focus on the movement of blood after it has left the heart. The form of the pulse wave and common characteristics in its shape will be described. Finally, we will also explain how to measure the pulse wave using photoplethysmography.

As seen in figure 2.1, blood is pumped from the heart into arteries, which divide into smaller blood vessels, the arterioles. Both of these vessels have flexible walls to accommodate the changes in pressure caused by the unsteady flow of blood into the circulatory

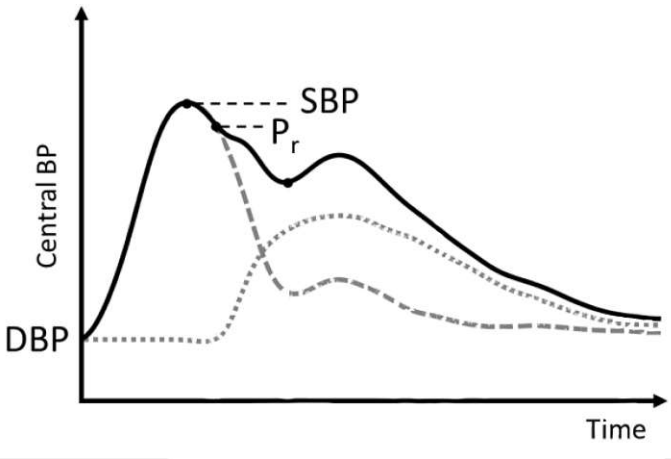

Figure 2.5: The wave caused by ventricular ejection is shown as a grey dashed line, while the reflected wave is illustrated as a grey dotted line. P_r represents the point of wave reflection. The resulting pulse wave, depicted as a solid black line, is given by the superposition of the pulse pressure wave and its reflections. Image adapted from [47].

system. Arterioles divide into a multitude of tiny blood vessels, the capillaries, which supply the surrounding tissue with oxygen. Afterwards, vessels merge again into venules and then veins to carry the blood back to the heart. [10, Ch. 8]

2.3.1 Pulse Wave Analysis

When the left ventricle ejects blood into the aorta, the sudden increase of blood volume causes it to expand and a pressure pulse is sent through the arterial system [3]. The pressure wave is partly reflected at branching points of vessels as it propagates through the system. The reflected wave travels back towards the heart. The main interest in pulse wave analysis (PWA) is to examine the shape of the pressure pulse wave, which is comprised of a superposition of the wave caused by ejection from the heart and its reflection. A wave like this can be seen in figure 2.5. [47]

Obviously, the distance the wave has travelled alters its shape, since the influence of reflections changes. As it moves further from the heart, the amplitude rises and the trough between the two peaks of the wave, that can be seen in figure 2.5, widens. Figure 2.6 shows the different waveforms at multiple points of measurement. [9, Ch. 33]

Nevertheless, pulse waves measured at the same location are still not comparable. The elasticity as well as size of arteries also have an effect on wave velocity. In stiffer vessels the wave travels faster and therefore the reflection returns at an earlier stage, which, of course, changes the waveform. Since age influences both artery stiffness and size, it has a considerable effect on the wave shape. The heart rate determines the ejection time, so obviously a change in HR also influences the pulse wave. Additionally, some other factors, such as physical fitness and body height have been shown to have an effect on the wave form. Differences in gender on the other hand can mostly be explained by differences in height. [34]

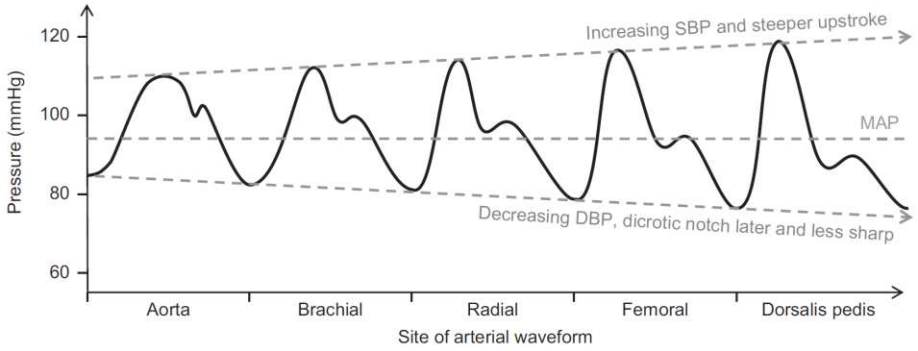

Figure 2.6: Shape change of the pulse pressure wave as it propagates through the system. From left to right the approximate wave shapes of measurements at the aorta, arm, hand, leg or foot can be seen. The amplitude of the wave becomes larger, while the mean arterial pressure (MAP) always stays the same. Image taken from [9, Ch. 33].

There are a few different options for measuring pulse waves, but the following section will focus only on photoplethysmography (PPG), which actually measures fluctuations in blood volume. The PPG wave form is usually measured some distance away from the aorta (e.g. finger or ear) and seems to be influenced by multiple factors in the arterial, cardiac, respiratory and autonomic systems [2]. That being said, PPG is widely used for pulse wave analysis and it has been established that the PPG signal measured on a finger shows the same physiological characteristics as the pressure waveform discussed above [47].

2.3.2 Photoplethysmography

Photoplethysmography (PPG) is a measurement technique used to determine oxygen saturation of blood and visualise the pulse wave. Obviously the latter is of interest in this context. The PPG was first introduced by Alrick Hertzman in 1937. It was him who named it a photoplethysmograph, incorporating the Greek word "plethysmos", which translates to fullness. Even in its beginnings the PPG was believed to measure blood volume changes. Although modern methods are more sophisticated, this belief turned out to be quite accurate. [2]

To be exact, a photoplethysmograph measures changes in light intensity. Biological tissue absorbs light that is transmitted to it. Some light however, still travels through the tissue or is scattered and reflected. This remaining light is registered by the photoplethysmograph. Most blue wavelengths are absorbed by biological tissue, while infrared and near-infrared easily pass through. Light in the green-yellow regions is mainly absorbed by red blood cells. In general, blood has a higher absorption rate than other tissue, which leads to a higher measured light intensity when blood volume decreases. [48, Ch. 6]

The pulse wave can be divided into two components: a fluctuation component, called AC after the alternating electric current, and background absorption, termed DC for direct current. The DC component corresponds to the light absorbed by bones, skin, venous blood and even some part of arterial blood. It is important to note, that this component is not constant. It changes slowly, largely influenced by breathing. The AC component

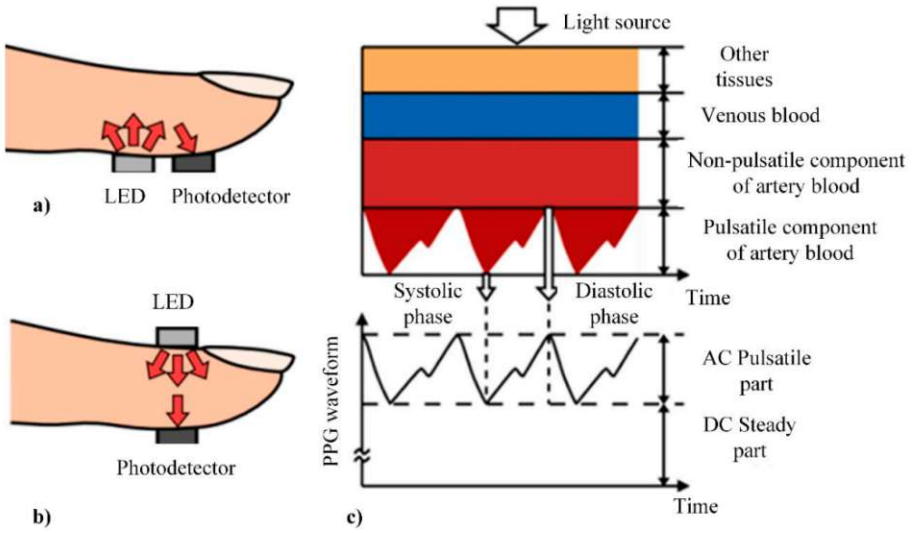

Figure 2.7: a) reflection PPG measurement at the finger. **b)** transmission PPG measurement at the finger. **c)** Emitted light, that is not absorbed by any blood components or other tissue, such as skin or bones, forms the steady part of the PPG. The pulsatile component of the PPG results from light absorbtion of the pulsatile component of arterial blood. This image was taken from [15] (original source: [49]).

corresponds to the light absorbed by the pulsating part of arterial blood. This AC part leads to the characteristic shape of the pulse wave. As can be seen in figure 2.7, the light intensity, which actually decreases as blood volume increases, does not directly correspond to the pulse wave. The wave is inverted to bear a greater resemblance to the arterial pressure wave form. [2, 48, Ch. 6]

The PPG contains a source emitting light into tissue, some of which is absorbed. A photodetector is then used to quantify how much light returns to the device. In general, there are two common options for PPG measurements: reflection and transmission, both of which are depicted in figure 2.7. [39]

For PPGs using reflection, the photodetector and the light source must be attached side by side. The sensor measures the light that is scattered back from tissue at the place of measurement. This represents an evaluation of superficial blood flow. This method can be used almost anywhere on the body, but is also more susceptible to motion artefacts due to movement of the chosen area. The waveform, especially the amplitude, could also be influenced by pressure on the sensor. Due to greater absorption rate in red blood cells, green wavelengths are usually used for reflection PPGs. [39, 48, Ch. 6]

For transmission, the light source and photodetector must be placed on opposite sides of the tissue under inspection. This poses an additional restriction to the measurement location, as at least some light must be detectable after transmission through the body. Many possible sensor placements, such as the cheek or tongue, are not normally used since they are quite uncomfortable. The fingers or ears are usually the best candidates for measuring a transmission PPG, even though they are easily affected by environmental conditions, such as cold temperatures. This method mostly utilises infrared or near-infrared

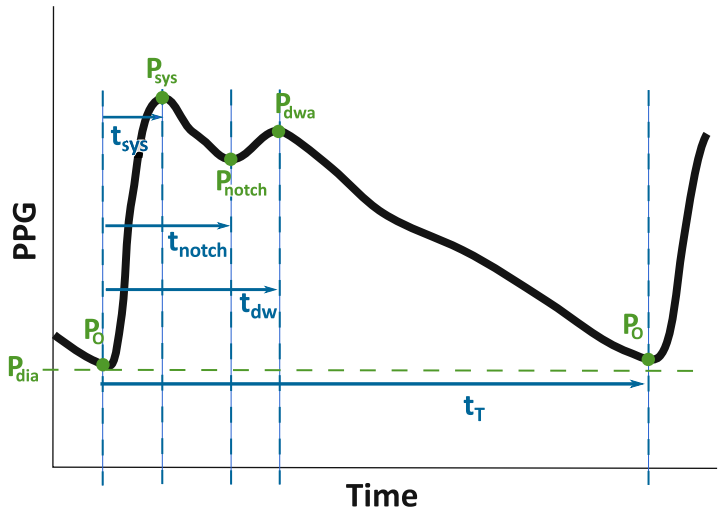

Figure 2.8: The image shows all characteristic points of the pulse wave (labelled in green) and all significant time durations (labelled in blue) defined in this section.

wavelengths, which pass through tissue more easily, and can be used for measurement of deeper tissue [39]. [48, Ch. 6]

Usually any PPG measurement is subject to post-processing before showing up on a screen. In most cases a band-pass filter is used to eliminate gradual changes as well as spikes. This eliminates the slow fluctuations due to respiration. Additionally, some sort of smoothing is generally also applied. [2]

There are many advantages to photoplethysmography, no matter whether reflection or transmission is chosen. It is a relatively cost-effective, non-invasive method of obtaining pulse wave measurement. Moreover, it is simple enough to facilitate unsupervised measurements. Although this may still be some time away, there is also some promising research into the use of PPG in determining blood pressure. However, measured pulse waves are not comparable for different individuals. There are many factors that influence the signal, some being skin colour, the amount of fat and muscle in the tissue and pressure on the sensor. [2, 47]

2.3.3 Pulse Wave Shape Parameters

Although pulse waves can look quite different for separate individuals and measurement locations, they mostly have similar main features. As images so far suggest, usually one can distinguish a prominent first peak in the wave followed by a dip and a second, less prominent peak. This section will explain the significance of these points and give an overview of the most important pulse wave shape parameters used in this thesis. Figure 2.8 gives a visual representation of all characteristic points and time durations defined in this section.

Characteristic Points

As suggested above, due to the difference between individuals, trying to characterise the pulse wave in absolute values is pointless. Hence, the only way to find characteristic points is by defining them in relation to their surrounding points. Of course, maxima and minima are easy to extract and are therefore good candidates for such points.

As already stated, a pulse wave has a very prominent peak at the beginning of the wave, which is connected to the rise in pressure and the expansion of the aorta as blood flows into it. The level of the first peak in the pulse wave corresponds to the maximum pressure after ventricle contraction. This is termed systolic blood pressure (SBP). In healthy individuals, when at rest, SBP does not rise over 120 mmHg [48, Ch. 4].

Definition 2.5. ***Systolic blood pressure (SBP)*** P_{sys} *corresponds to the first peak in the pulse wave. Usually this is also the maximum of the pulse wave over one heartbeat.[47]*

Diastolic blood pressure (DBP) on the other hand, corresponds to the lowest point of the pulse wave, e.g. the wave's onset. It describes the arterial pressure just before ventricular contraction, when there is no blood flow into the aorta. The DBP of a healthy person at rest is usually below 80 mmHg [48, Ch. 4].

Definition 2.6. ***Diastolic blood pressure (DBP)*** P_{dia} *corresponds to the minimal value in the pulse wave over one heartbeat [47].*

The trough between the first and the second peak also has a significant meaning. When the aortic valve closes, there is a short fall and rise in pressure in the aorta which is seen as a small dip between the two peaks of the pulse wave. This called the dicrotic notch and marks the end of the systole, the ejection period of the heart. [10, Ch. 8].

Definition 2.7. *The* ***Dicrotic Notch*** P_{notch} *is the (most prominent) local minimum between the first and the second peak of the pulse wave [10, Ch. 8].*

Lastly, the second peak of the pulse wave is caused by the wave reflections discussed in 2.3.1, which become more prominent as the wave propagates. This is also referred to as the dicrotic wave. Therefore, the second peak is called the dicrotic wave amplitude [22]. Sometimes the term diastolic peak is also used, since it occurs during diastole, the period when the heart does not eject blood. It is important not to confuse it with DBP.

Definition 2.8. *The* ***dicrotic wave amplitude*** P_{dwa} *corresponds to the second peak of the pulse wave over one heartbeat. It is sometimes also called the* ***diastolic peak****. [22, 47]*

Before moving on to time-related characteristics of the pulse wave, we need one more crucial definition. Physiologically, the onset of the wave is just before the aortic valve opens, i.e. before the ejection period of the heart begins. In contrast to the maxima and minima discussed up to now, it is not defined, which exact point on the curve represents the onset. The exact choice of the point and details of calculation during processing of data in this thesis are outlined in section 3.4.3.

Definition 2.9. *The* ***onset*** P_O *of a pulse wave is a point in time, before the ventricles contract and blood is ejected from the heart. There is no rise in blood pressure before the onset.*

Time-Related Parameters

Using the onset P_O of the pulse wave, as defined in 2.9, as a reference point, some characteristic time durations can be defined. One value of interest is the total duration of a pulse, since it gives insight to the time, that passes between two pulses, and allows us to regard time durations between characteristic points in proportion to the total length of the pulse.

Definition 2.10. *The **total pulse duration (TPD)** t_T is the time from the onset of a pulse wave to the the onset of the successive pulse wave [22].*

The time from the onset P_O of a pulse wave to the point P_{sys}, where SBP is reached, is called the crest time. As higher artery stiffness increases wave velocity, the time at which the maximum in pressure is reached shortens with higher stiffness of blood vessels and therefore also with age. [47]

Definition 2.11. *The **time to the systolic peak** t_{sys} is the time from the onset of a pulse wave to the point of systolic blood pressure. It is also sometimes referred to as the **crest time**. [22, 47]*

Similarly, the time from the onset P_O to the dicrotic notch P_{notch} is defined:

Definition 2.12. *The **time to dicrotic notch** t_{notch} is the time from the onset of a pulse wave to the dicrotic notch.*

The last time duration of interest is the dicrotic wave time, the time from the onset P_O of the wave to the point of dicrotic wave amplitude P_{dwa}. With increasing arterial stiffness the reflections that make up this part of the pulse wave occur earlier and therefore dicrotic wave time shortens. The ratio of body height h and the difference between crest time t_{sys} and dicrotic wave time has even been investigated as an index of artery stiffness. [47]

Definition 2.13. *The **dicrotic wave time** t_{dw} is the time from the onset of a pulse wave to the point where dicrotic wave amplitude is reached [22].*

2.3.4 Pulse Arrival Time

The pulse arrival time (PAT) is technically not a pulse wave shape parameter, but it is defined by the pulse wave. The calculation of PAT requires simultaneous ECG and PPG measurements, since it describes the time duration between the R-peak in an ECG and the onset of the corresponding pulse wave at the location of PPG measurement. PAT is representative for the time it takes a pulse wave, originating in the heart, to reach a certain point, e.g. the finger or the ear, in the body. All information in this short subsection is taken from [48, Ch. 4].

Definition 2.14. ***Pulse arrival time (PAT)** is defined as the time it takes pulse waves to reach a certain location in the body. More exactly, it is the time from an R-peak in an ECG to the onset of the corresponding pulse wave measured in a PPG at said location.*

PAT is highly related to systolic blood pressure and there is ongoing research into a form of blood pressure measurement which utilises the pulse arrival time, mainly because it is quite simple to measure. The details of the calculation of PAT used within this thesis are outlined in section 3.4.3.

2.4 Fatigue

In literature, there is no uniform definition for fatigue. It can refer to both mental and physical tiredness. Within the PANACEA Project however, partners decided to concentrate on the physical aspect and use the following definition:

Definition 2.15. ***Fatigue*** *here also relates to sleepiness and rest. Fatigue is defined as the biological drive for recuperative rest [55], with sleepiness as a special case referring to accumulated sleep debt, prolonged wakefulness or troughs in the circadian rhythm. The KSS, which is explained below, serves as a reference for the definition of fatigue[40, 41].*

2.4.1 Approaches to Measuring Fatigue

Studies on fatigue in the context of driving use a variety of different techniques to estimate the driver's level of fatigue. This section is based on a recent review (Doudou et al., [14]), examining current driver drowsiness measurement technologies and available products. Roughly, one can differentiate four different approaches: monitoring driving behaviour, the driver's activities, the use of physiological signals and subjective fatigue estimates.

Monitoring Driving Behaviour

Steering wheel movement, vehicle deviation and speed can all be connected to the driver's fatigue. A drowsy driver is likely to need more steering corrections and will have more difficulties keeping lane position. All these indicators require a variety of sensors in the car.

Monitoring the Driver's Activities

Although facial expressions and head position can indicate a driver's fatigue, the most used metrics focus on the eyes. Studies usually examine the percentage of eyelid closure and eye movements. Generally, as sleepiness occurs, fast eye movements will become slower and the percentage of eyelid closure will rise. This approach requires at least one camera filming the driver's face, sometimes realised through wearable cameras on glasses.

Analysis of Physiological Signals

Since the effect of fatigue on cardiac activity, measured using an ECG, is one of the main topics of this thesis, it will be discussed in more detail further on. Apart from ECGs, electroencephalography (EEG) and electrooculography (EOG) measurements are also affected by fatigue.

An EEG measures brain activity through electrodes placed on the scalp. The effects of drowsiness or sleep can be detected by a change in frequency and amplitude of the signal.

An EOG measures eye movement and blink patterns through electrodes placed around the eye. As mentioned above, indicators for fatigue are slower eye movements and a higher percentage of eye closure. However, some results indicate that eye movement alone is not sufficient for detecting drowsiness.

Subjective Fatigue Estimates

Unfortunately, the necessary equipment, i.e. electrodes, cables, cameras or other measurement devices, limits the use of the methods mentioned above in a real-life driving context. For fast and simple assessment of fatigue, mostly subjective scales based on questionnaires are used. Although there is a multitude of different scales, here we will highlight the Karolinska sleepiness scale (KSS). It is a well-established method for obtaining subjective fatigue values and focusses on sleepiness, rather than mental fatigue, which is why it was chosen for this experiment.[40]

Definition 2.16. *The **Karolinska sleepiness scale (KSS)** is a nine-point scale, where 1 indicates low sleepiness, while 9 is the highest level of sleepiness. A user is asked to assess their own sleepiness in the preceding 5 minutes and marks the corresponding KSS level [40]. The nine levels of the KSS are:*

Table 2.3: Levels of the Karolinska sleepiness scale (KSS) [40]

Level	Description
1	Extremely alert
2	Very alert
3	Alert
4	Rather alert
5	Neither alert nor sleepy
6	Some signs of sleepiness
7	Sleepy, but no effort to keep awake
8	Sleepy, some effort to keep awake
9	Very sleepy, great effort to keep awake, fighting sleep

The KSS was first designed and tested using a sleep deprivation experiment and EEG as well as EOG measurements in 1990 by Torbjörn Åkerstedt and Mats Gillberg [41]. In its original version, odd steps were not labelled. However, a later study shows that results using the fully labelled scale are comparable to results obtained using the partly labelled scale [30].

According to a review (Åkerstedt et al., [40]) investigating KSS as an indicator for sleep deprivation and impaired waking function, the subjective rating is highly sensitive to periods of prolonged wakefulness. Higher KSS values correlate with difficulties staying awake and findings are consistent across indiviudal subjects. Additionally, the review's discussion of some articles, specifically on driving at night or after sleep deprivation, indicates an increased risk of accidents for KSS levels 8 or 9.

2.4.2 Effects of Fatigue on the Cardiovascular System

As established in section 2.2.4, heart rate variability (HRV) is strongly connected to the activity of the autonomic nervous system (ANS). Remember, higher HRV and more power in the HF component mostly indicate parasympathetic activity while more power in the LF component coincides with higher sympathetic activity. Therefore, the $\frac{LF}{HF}$ ratio, which

is considered to be a marker of balance between PNS and SNS, decreases with rising PNS activity.

As the body prepares for sleep the heart rate decreases, allowing for more variability between beats, and the PNS becomes dominant [37]. This leads us to believe that as fatigue arises, SDNN and HF should increase due to higher parasympathetic activity, whereas LF, heart rate and the $\frac{LF}{HF}$ ratio should decrease, showing lower sympathetic activity [8].

A review of multiple studies on the effect of fatigue on HRV, most of them in the context of driving, mostly confirms the expected trend. The measured change to HRV parameters and basic information concerning the trials as well as sources are given in table 2.4. In addition to the trends stated previously, NN50 and total power in the frequency domain were shown to increase with fatigue. According to Awais et al. [4], normalised HF and LF show similar behaviour to their non-normalised counterparts. One study also shows RMSSD to decrease significantly for sleepy drivers.

As for the effect of fatigue on the pulse wave (PW) parameters, studies investigating the connections are much harder to find. Only one publication linking fatigue to a change in PW characteristics is included in table 2.4. The evaluation of sleepiness while flying, rather than driving, showed a significant increase in pulse arrival time, crest time and diastolic time, which is the time from wave onset P_O to the diastolic peak P_{dwa} [27].

Even though the $\frac{LF}{HF}$ ratio is a rather appealing parameter for fatigue assessment due to its physiological interpretation, it tends to show some controversy. Most studies, as reflected in table 2.4, show a negative trend for rising sleepiness, but some research, such as [43], has inferred the opposite or no significant change [1]. However, it should also be noted that multiple different procedures are used to estimate the power spectrum and the applied method is often not clarfied. This could be one reason for the discrepancies in frequency-domain measures in table 2.4.

One main reason for the inconsistencies in findings is the variance of HRV and pulse wave parameters between individuals. Most studies concerning sleepiness while driving are also conducted with only few participants. Furthermore, it should not be forgotten, that there is a multitude of factors that influence HRV, therefore changes seen in parameters may not always be due to fatigue [37].

Moreover, it can be hard to collect high quality data for the investigation of changes in cardiovascular parameters in the context of sleepiness behind the wheel. ECG and PPG equipment is obtrusive while driving and surrounding electromagnetic noise or driver movements can cause artefacts in the measurement [1]. For these reasons as well as safety concerns, most trials are completed in simulators. Although it has been suggested that fatigue can be studied equally in real and simulated environments [38], it has also been considered, that increased stress due to risk in real driving compared to simulated driving could increase sympathetic activity and cover up sleep-related changes to HRV [1].

Table 2.4: Some results on the influence of fatigue on HRV and pulse wave parameters. The direction of change in each parameter for rising fatigue is stated in the trend column.

Parameter	Trend	Alert	Sleepy	p	Summary	Source
$\frac{LF}{HF}$ ratio	negative	1.008 ± 0.122	0.994 ± 0.157	$p < 0.01$	trial simulating driving at night (0:00 - 2:00)	[45]
		1.006 ± 0.156	0.9979 ± 0.147		trial simulating driving at night (3:00 - 5:00)	
	negative	1.8 ± 1.15	1.2 ± 0.87	$p < 0.01$	simulated driving tasks with sleep-deprived truck drivers	[36]
	none	2.1 ± 1.5	2.1 ± 0.9	none	trial conducted in Sweden on real roads (no simulation)	[1]
	positive	3.18 ± 1.58	4.33 ± 2.27	$p < 0.05$	trial on real roads with professional drivers	[43]
	negative	2.01 ± 0.98	1.39 ± 0.59	$p < 0.01$	trial in a simulated driving environment	[4]
RMSSD	negative	-28% with increasing fatigue		not given	simulated driving at night-time in a simulator; trial used sensitivity test instead of hypothesis test	[26]
mean RR	positive	688.7 ± 84 ms	753.9 ± 103 ms	$p < 0.05$	see above	[1]
	positive	899 ± 122 ms	927 ± 132 ms	$p < 0.05$	see above	[43]
	positive	0.89 ± 0.05 s	0.97 ± 0.09 s	not given	simulated driving and other tests during sleep deprivation	[18]
	positive	781.87 ± 29.48 ms	958.77 ± 91.74 ms	$p < 0.01$	trial with trained pilots using an aviation training simulator at night-time (2:00-6:00)	[27]
SDNN	positive	40.8 ± 17	53.2 ± 23	$p < 0.05$	see above	[1]
	positive	63.6 ± 21.1	73.7 ± 24.3	$p < 0.005$	see above	[43]
NN50	positive	39 ± 47	52.8 ± 48	$p < 0.05$	see above	[1]
LF	positive	222.4 ± 191	449.5 ± 365	$p < 0.05$	see above	[1]
	positive	732.9 ± 434.44	1057.5 ± 637.42	$p = 0.009$	driving simulation and other tasks during day-time	[56]
HF	positive	127.2 ± 121	241.2 ± 212	$p < 0.05$	see above	[1]
	negative	859.03 ± 77.4	626.18 ± 54.1	$p = 0.039$	see above	[56]
LFnorm	negative	0.54 ± 0.1	0.46 ± 0.08	$p < 0.01$	see above	[4]
	negative	0.74 ± 0.11	0.21 ± 0.02	not given	see above	[18]
HFnorm	positive	0.32 ± 0.08	0.37 ± 0.06	$p < 0.05$	see above	[4]
TP	positive	373.4 ± 302	741.4 ± 584	$p < 0.05$	see above	[1]
PAT	positive	407.66 ± 9.85 ms	443.509 ± 14.07 ms	$p < 0.05$	see above	[27]
diastolic time	positive	no values given		$p < 0.01$	see above	[27]
crest time	positive	no values given		$p < 0.05$	see above	[27]

3 Methods

This chapter covers the theoretical basis for data analysis and generating statistical models. First, a short recap of descriptive statistics, with a focus on the methods applied, is given. Then, an overview of multivariate linear regression, the assessment of model quality and some computation algorithms are presented. Before going into detail on the data collection and processing in this thesis, some attention is given to principal component analysis and its use in linear regression.

3.1 Hypothesis Testing in Descriptive Statistics

Descriptive statistics are simply a collection of numerical and graphical techniques, with the goal of organising and describing data characteristics. Descriptive statistics yields no additional information about causality or the sampled population. These tasks require inferential statistics. Unless stated otherwise, the introduction to this section is based on [16] and [21, Ch. 13].

In descriptive statistics, three levels of measurement are distinguished: nominal (or categorical), ordinal and continuous. Their characteristics are detailed in table 3.1. Techniques for data analysis vary according to the level of measurement. In general, a measure of central tendency and the spread of scores is determined.

There are three common measures for central tendency. The **mode**, which is simply the category or value with the highest number of cases, can be used for unordered or nominal data. If an ordered list of data can be produced, the **median** is the score in the middle of this list. In case there is an even amount of values, the average of the two scores closest to the middle of the list is taken as the median. Additionally, for continuous data, the **mean** can be calculated as the average of all scores.

For measuring dispersion on a nominal level, it is common to determine the **frequency distribution**, which corresponds to the number of cases in each category. If data can be ordered, the **range**, the distance between the highest and lowest value, can be calculated. The ***k*-th percentile**, is the score, such that $k\%$ of all scores are below the percentile. For continuous data, the **interquartile range**, the difference between the quartiles (25-th and 75-th percentile), and the **standard deviation**, the average difference of scores to the mean, can be calculated.

A common graphical representation of the frequency distribution is the histogram, a bar chart showing the number of cases per category. For continuous data, intervals can be defined to take the place of categories. Mean and standard deviation are usually used for data coming from a normal distribution. If data is skewed, median and interquartile range are mostly preferred as measures. For non-normal data, histograms or boxplots, illustrating median, quartiles and range, are used for graphical representation. Statistical analysis of data in this thesis will be performed in accordance with these conventions.

Table 3.1: Levels of measurement for data in descriptive statistics

Level	Description	Examples
Nominal	scoring into broad categories with no order	sex, species
Ordinal	scoring into ordered categories	age group, fatigue levels
Continuous	data on a continuous scale	temperature, height

Whether data follows a normal distribution, or not, needs to be taken into account when performing descriptive statistics, but does not have a huge effect on the methods used. However, for parametric methods of inferential statistics, normality of data is an indispensable assumption. If data does not come from a normal distribution, non-parametric statistics must be used. So, naturally, a method for checking data for normality is needed. Hypothesis tests represent such a procedure for decision-making.

Hypothesis tests provide answers to "yes-or-no" questions about data in a statistical context. They are therefore a useful tool for checking the distribution of data. First, a **null hypothesis H_0**, usually an assumption of no change, is defined. For example, in a clinical trial, the null hypotheses could be, that there is no difference in a certain measured value between two treatment groups. The null hypothesis H_0 is tested against the **alternative hypothesis H_1**, that H_0 is untrue. Hence, there are only two possibilities: H_0 is true or H_0 is false.

It is important to understand, that while a rejection of H_0 indicates that H_1 may hold, failing to reject H_0 gives less information. There could be multiple reasons for this beside the fact that H_0 is true, e.g. very small sample size.

To assess the hypothesis using the given data, a **test statistic T** is constructed. The test statistic is designed to measure the extent to which the data is consistent with H_0. It therefore strongly depends on the type of hypothesis to be tested.

The test statistic T can not be directly interpreted to conclude the hypothesis test. The ***p*-value** helps quantify the values of T. It is defined as the probability p of observing an equal or more extreme value of T, under the assumption that H_0 is true. For example, for a test statistic T, for which small values support the null hypothesis H_0, and a random data sample X, this can be formulated as

$$p = \mathbb{P}(\exists \tilde{X} : T_{\tilde{X}} \geq T_X | H_0 \text{ is true}).$$

Therefore, a small p-value indicates that it is unlikely to observe the given data under the H_0 and speaks for its rejection. [21, Ch. 13]

Lastly, the question remains how small the p-value should be in order to reject H_0. This boundary, usually called the **significance level α**, is somewhat flexible. Typical significance levels are $\alpha = 0.05$ or $\alpha = 0.01$. Respectively, this corresponds to a 5% or 1% chance of observing the calculated test statistic even though H_0 is true.

The significance level needs to be chosen with the application in mind. One differentiates between **type I errors**, rejecting H_0 if it were actually true, and **type II errors**, failing to reject H_0 when it is actually false. Depending on the field in question, significance levels considerably smaller than 0.01 could be necessary, if avoiding type I errors is of utmost

importance. Due to the fact that H_0 is only rejected, if $p < \alpha$, it also holds that the type I error is bound by α.

The probability of correctly rejecting the null hypothesis H_0, i.e. when H_1 holds, is defined as the **power of the hypothesis test against H_1** [23, Ch. 3]. Obviously, test power also corresponds to the probability of not making an error of type II.

3.1.1 Testing for Normality

As stated previously, it is often of interest whether observations are likely to come from a normal distribution. This is a family of continuous distributions for real-valued random variables. It is given by the density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$

where $\mu \in \mathbb{R}$ is the mean and $\sigma \in \mathbb{R}$ is the standard deviation of the distribution. Even though the distribution of a random variable can not be determined with complete confidence, hypothesis tests can be used to estimate the goodness of fit. Most information on tests of normality was taken from [23, Ch. 14]. Any other sources used are referenced separately.

Denote by $X_1, \ldots, X_n$ independent identically distributed real-valued observations with the cumulative distribution function F. The empirical distribution function of the sample $X_1, \ldots, X_n$ can be defined as:

Definition 3.1. *The **empirical distribution function (EDF)** of $X_1, \ldots, X_n$ is the step function given by*

$$\hat{F}_n(t) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{X_i < t}.$$

The EDF plays a crucial role in a well-known class of hypothesis tests, the **Emipirical Distribution Function (EDF) tests**. The basic idea of EDF-tests is to measure the discrepancy between the EDF $\hat{F}_n$ and the believed cumulative distribution function of the data, which will be called F_0. It can be shown, that the EDF $\hat{F}_n$ uniformly tends to the true distribution function F with probability one (Glivenko-Cantelli theorem, [23, Ch. 11]). Therefore, the hypotheses of an EDF test can be formulated as

$$H_0 : F = F_0 \tag{3.1}$$
$$H_1 : F \neq F_0 \tag{3.2}$$

where, when testing for normality, F_0 must be the cumulative distribution function of the normal distribution.

Any metric d on the space of distribution functions fulfils the purpose of measuring discrepancy between F and F_0. Hence, $d(\hat{F}_n, F_0)$ could be used to derive the test statistic for an EDF test. Using the **Komogorov-Smirnov metric** $d_K(F, G) = \sup_t |F(t) - G(t)|$ as the measure of distance one obtains the Kolmogorov-Smirnov test.

Definition 3.2. *The **Kolmogorov-Smirnov Test** is an empirical distribution function (EDF) test. It tests the null hypothesis, that $X_1, \ldots, X_n$, with EDF $\hat{F}_n$, comes from a distribution with cumulative distribution function F_0, of which the parameters need not be known. The test statistic is given by*

$$T_n = \sup_t n^{1/2} |\hat{F}_n(t) - F_0(t)| = \sup_t n^{1/2} d_K(F_n, F)$$

where d_K is the Kolmogorov-Smirnov metric.

By this procedure, multiple other tests of normality can also be defined. Among others, the Anderson-Darling test is another common choice of EDF test.

Definition 3.3. *The **Anderson-Darling Test** is an empirical distribution function (EDF) test. It tests the null hypothesis, that $X_1, \ldots, X_n$, with EDF $\hat{F}_n$, comes from a distribution with cumulative distribution function F_0, of which the parameters need not be known. The test statistic is based on the squared differences of F_0 and $\hat{F}_n$ and is given by*

$$T_n = n \int_{-\infty}^{\infty} (\hat{F}_n(x) - F_0(x))^2 [F_0(x)(1 - F_0(x))]^{-1} dF_0(x).$$

While the Kolmogorov-Smirnov-test is widely used, numerical power comparisons discussed in [23, Ch. 14] have shown the Anderson-Darling-test to outperform the Kolmogorov-Smirnov-test. Another numerical power comparison [42] came to the same conclusion, especially for small sample sizes. Since, in this case, the amount of data available is rather small, the Anderson-Darling-test seems to be the better choice and will be used for all tests of normality throughout this thesis.

3.1.2 Wilcoxon Signed Rank Test

The Wilcoxon signed rank test is a useful test for comparing the measures for central tendency of two samples. It is a non-parametric alternative to the paired t-test, which compares the means of two samples from normal distributions. This subsection is largely based on [54] and references to this source will be omitted. Any information taken from other sources is cited accordingly.

Non-parametric methods do not need to estimate parameters of the underlying distribution, as would be the case in parametric tests such as the paried t-test. Therefore, non-parametric tests often require less assumptions to be made on the data.

Using the Wilcoxon signed rank test, a sample from a non-normal distribution can be compared with a single hypothesised value. It can therefore be used to test an assumption of the median value. Furthermore, median values of two samples can be compared by testing the differences against zero.

The test consists of six simple steps, that are outlined in table 3.2. Let $X_1, \ldots, X_n$ be a sample of the random variable X with an absolutely continuous cumulative distribution function. First, an assumption about the median M is made, e.g. that it is equal to the value M_0, and the null hypothesis is formulated:

$$H_0: \quad M = M_0. \quad [6]$$

Table 3.2: Steps of Wilcoxon signed rank test [54]

Step	Details
1	Define the null hypothesis H_0 and the hypothesized value M_0, with which the data should be compared.
2	Rank all observations by the magnitude of their absolute deviation from M_0 from small to large. Any observation taking the exact value M_0 should be ignored. Two observations with identical absolute deviation are ranked by the average of the two corresponding ranks.
3	For each observation, determine whether the deviation from M_0 is positive or negative
4	Calculate: R_+ = sum of all ranks for observations with positive deviations R_- = sum of all ranks for observations with negative deviations.
5	Define the test statistic R as the smaller value: $R = \min(R_+, R_-)$.
6	Calculate the p-value.

Next, the differences

$$D_i = X_i - M_0 \quad i = 1, \ldots, n$$

are computed and ordered by their absolute value from small to large. The new, ordered, list will be indexed using $j = 1, \ldots, n$. If any X_i corresponds to the assumed median M_0, i.e. $D_i = 0$, this value is omitted from the list. The ordered list items are then given the ranks $R_j = j$ according to their position in the list. Hence, the higher the absolute deviation from the assumed median, the higher the rank. Should two differences be exactly the same, the value is only recorded once and the rank is replaced by the average of both values' ranks. [6]

For the calculation of R_+ and R_-, the original signs need to be represented in the resulting ordered list. This can be done, by replacing the list elements D_j with

$$\text{sgn}(D_j)|D_j| \quad j = 1, \ldots, n.$$

Denoting $D_+ = \{j|\text{sgn}(D_j) > 0\}$, the calculation of R_+ can be written as

$$R_+ = \sum_{j \in D_+} R_j.$$

The calculation of R_- can be obtained in an analogue fashion. Finally, the test statistic R can be defined as $R = \min(R_+, R_-)$.

If H_0 holds, the median M of X is equal to M_0 and therefore D_i are expected to be distributed symmetrically around 0. Hence, the expected values of R_+ and R_- should be equal. If $M > M_0$ or $M < M_0$, then $\mathbb{E}(R_+) > \mathbb{E}(R_-)$ or $\mathbb{E}(R_+) < \mathbb{E}(R_-)$, respectively [6]. Since $R_+ + R_- = \frac{n(n+1)}{2}$ and $R = \min(R_+, R_-)$, it follows, under the additional condition

that H_0 is fulfilled, that

$$\mathbb{E}(R_+) = \mathbb{E}(R_-) \implies \mathbb{E}(R) = \frac{n(n+1)}{4} =: N$$
$$\mathbb{E}(R_+) \neq \mathbb{E}(R_-) \implies \mathbb{E}(R) < N.$$

Finally, the Wilcoxon signed rank test rejects the null hypothesis H_0 with significance level α, if

$$R \leq T_\alpha(n),$$

where $T_\alpha(n)$ is the smallest number to fulfil $\mathbb{P}(R \leq T_\alpha(n)) \leq \alpha$. [6]

3.1.3 Correction for Multiple Testing

When performing a single test, a suitable threshold for interpreting a tests p-value as being significant is usually $\alpha = 0.05$ or $\alpha = 0.01$. However, when performing multiple tests, determining an appropriate significance level is more difficult. The significance level needs to be adapted to account for the number of tests being conducted. This is termed a correction for multiple testing and will be explained in detail in this subsection. It is based entirely on information from [52].

When each test is performed with the significance level α, this means each test has a chance of α to produce a false conclusion, i.e. a significant p-value even when there is, in fact, no real effect. Therefore, the probability of drawing at least one false conclusion in a series of multiple tests increases considerably with the number of tests performed. Multiple testing without any correction can lead to false rejections of H_0, i.e. type I errors.

Imagine, for example, that 15 hypotheses are to be tested at a significance level of $\alpha = 0.05$. The probability of the tests yielding a significant result on data, where there is no real effect, can be calculated as

$$\begin{aligned}\mathbb{P}(\text{at least one significant result}) &= 1 - \mathbb{P}(\text{no significant results}) \\ &= 1 - (1 - 0.05)^{15} \\ &\approx 0.54.\end{aligned}$$

This means there is a probability of 0.54, that the tests yield at least one false conclusion.

There are many different approaches for correcting for multiple testing, one of which is the simple and well known Bonferroni correction.

Definition 3.4. *The **Bonferroni correction** is a method for decreasing the chance of type I errors due to multiple testing. When performing m test at the significance level α, the m tests are simply performed at the adapted significance level $\frac{\alpha}{m}$.*

This method limits the probability of making at least one false rejection. However, the correction comes at a cost. Using the adapted significance level, only very strong effects will be detected as being significant. Some true effects will not be interpreted as such by the test and therefore a Bonferroni correction leads to more type II errors.

A different approach is to control the proportion of significant test results, that are in fact false rejections. This is less restrictive than limiting the probability of making a single type I error, especially with a high number of tests.

Definition 3.5. *Controlling the **false discovery rate (FDR)** is a method for decreasing the expected proportion of type I errors among all significant results when performing multiple independent or positively correlated tests. For m tests with the null hypothesis H_i, $i \in I = 1, \ldots, m$, the p-values are sorted in increasing order: $p_1 \leq p_2 \leq \ldots \leq p_m$. Set*

$$k = \max\{i \in I | p_i \leq \frac{\alpha}{m} i\}, \tag{3.3}$$

then reject all hypothesis $H_1, \ldots, H_k$.

This means, that for larger p-values, the corrected significance level is also higher. For the highest p-value fulfilling the requirement (3.3), and all p-values that are smaller, the null hypotheses are rejected. The advantage of FDR compared to the Bonferroni correction, is that it allows to control the type I errors, while also comparatively reducing the amount of type II errors. The advantages of FDR are especially pronounced for a large number of tests.

Depending on the application, particularly the severity of penalty for type I errors, it may be wise to choose a highly restricting correction, such as the Bonferroni method. If, however, a certain proportion of type I errors does not affect the overall interpretation of test results, a less restrictive approach, such as the false discovery rate, may be favourable. In the context of this thesis, the bonferroni correction is likely to be too restrictive. Therefore, when correction for multiple testing is necessary, the false discovery rate will be controlled at the level $\alpha = 0.05$.

3.2 Multivariate Linear Regression

The goal of this section is to understand the theoretical background of linear regression with multiple variables. Unless stated otherwise, the entire section is based on [20, Ch. 2-3]. First, it is helpful to go over some basic terminology used to characterise variables.

In general, a regression model geared for prediction can be interpreted as a function with one or more input variables that influence a certain output variable. Often the goal is to estimate an unknown output value for a given set of input values. In regression, variables are distinguished by their function within the model as well as their nature.

Definition 3.6. *The inputs to a regression model are called **independent variables** or **predictors**, while the output is referred to as a **dependent variable** or **response**.*

While some variables may be comparable in size, for example those that take values in the real numbers, others might not be comparable in any way. They rather indicate a certain state, such as a colour or species.

Definition 3.7. ***Quantitative variables** have a certain order and are comparable in size. **Qualitative variables** have no order or relation between different values. They are mostly descriptive.*

With all variable types, creating a predictive model comes down to finding a function that accurately estimates a response from given predictors. Usually the prediction of quantitative variables using a model is referred to as regression. When predicting qualitative variables, the term classification is often used.

Of course, data is necessary to derive any linear model. This is usually called the training data. Before going into detail on how such a model is obtained, some notation needs to be clarified. In general, bold letters are used for matrices and vectors, in upper and lower case respectively, and non-bold symbols describe scalars. Let $\boldsymbol{X}$ denote the data matrix, where each column $\boldsymbol{x}_j$ corresponds to the data observed for the j-th input variable. Each line of $\boldsymbol{X}$ contains the data of one measurement. Thus, x_{ij} is the measured value of the j-th variable in the i-th measurement. Let $\boldsymbol{y}$ be the response vector, where each entry y_i is the observed response of the i-th measurement. The predicted output for any given data is denoted by $\hat{\boldsymbol{y}}$. Thus, for n measurements and p variables, under the assumption they all take values in $\mathbb{R}$, a model can be described as the map

$$\begin{aligned} \mathcal{M}: \quad & \mathbb{R}^{n\times p} \to \mathbb{R}^n \\ & \boldsymbol{X} \longmapsto \hat{\boldsymbol{y}}. \end{aligned}$$

3.2.1 Linear Regression

A regression model is called linear, if it is linear in all independent variables. Thus the model has the following form:

Definition 3.8. *A **linear regression model** $\mathcal{M}$ maps a data matrix $\boldsymbol{X}$ to an estimated response $\hat{\boldsymbol{y}}$ such that*

$$\mathcal{M}(\boldsymbol{X}) = \beta_0 + \sum_{j=1}^{p} \boldsymbol{x}_j \beta_j = \hat{\boldsymbol{y}}, \tag{3.4}$$

*where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$. The **model coefficients** are denoted by β_j, and β_0 is called the **intercept**.*

For the rest of this section, let us assume, that the model $\mathcal{M}$ has p inputs and the training data consists of n measurements, i.e. $\boldsymbol{X} \in \mathbb{R}^{n\times p}$. The coefficients β_j as well as the intercept β_0 need to be estimated from the training data. Often β_0 is included in the coefficient vector $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ and a column of ones is added to the data matrix $\tilde{\boldsymbol{X}} = (\boldsymbol{1}, \boldsymbol{X})$. Then (3.4) can be written as the matrix-vector multiplication:

$$\hat{\boldsymbol{y}} = \tilde{\boldsymbol{X}}\beta. \tag{3.5}$$

Although (3.4) describes a linear function, it is also possible to fit certain non-linear functions using linear regression. The columns of the data matrix $\boldsymbol{X}$ need not exactly correspond to the measured variable data. Additional variables can be derived from measured data. To fit a function that is quadratic in $\boldsymbol{x}_j$, for example, one must derive the additional variable, $\boldsymbol{x}_j^2$, by taking the element-wise square of the original data. Then coefficients can be fitted using linear regression with $p+1$ variables. The problem does not become more complex, since the model function remains linear in the coefficients to be estimated. By this

procedure, arbitrary polynomial functions can be fitted and interactions between variables, modelled as the element-wise product of two variables, can be included in the model.

Ordinary Least Squares Method

The ordinary least squares method is by far the most used procedure for determining the coefficients β_j, for $j = 0, \dots, p$, in equation (3.4). The basic idea is quite simple: the difference between the measured response $\boldsymbol{y}$ and the estimated response $\hat{\boldsymbol{y}}$ should be as small as possible. This is realised by choosing the coefficients $\beta = (\beta_0, \beta_1, \dots, \beta_p)^\top$ such that they minimise the sum of squared differences.

Definition 3.9. *The **residual sum of squares (RSS)** is the sum of squared differences between measured and estimated response:*

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - \mathcal{M}(\boldsymbol{x}_i))^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2, \tag{3.6}$$

*where $\boldsymbol{x}_i$ and y_i correspond to the i-th row of the data matrix and response, respectively, β is the coefficient vector and $\mathcal{M}$ is the linear model from definition 3.8. In some cases the term **sum of squared errors (SSE)** is also used.*

Now we can derive a formula for the coefficient vector β by minimising the RSS. We will do this under the assumption, that a column of ones has been added to the data matrix and the model is described by (3.5). We will denote each line of $\tilde{\boldsymbol{X}}$ by $\tilde{\boldsymbol{x}}_i$. Then the residual sum of squares (3.6) can be rewritten as the vector product

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{n} (y_i - \underbrace{(1, \boldsymbol{x}_i)}_{\tilde{\boldsymbol{x}}_i} \cdot \beta)^2 \\ &= \sum_{i=1}^{n} (y_i - (\tilde{\boldsymbol{X}} \cdot \beta)_i)^2 \\ &= \sum_{i=1}^{n} ((\boldsymbol{y} - \tilde{\boldsymbol{X}} \cdot \beta)_i)^2 \\ &= (\boldsymbol{y} - \tilde{\boldsymbol{X}} \cdot \beta)^\top \cdot (\boldsymbol{y} - \tilde{\boldsymbol{X}} \cdot \beta) \end{aligned}$$

where the last equality holds because the vector $(\boldsymbol{y} - \tilde{\boldsymbol{X}} \cdot \beta) \in \mathbb{R}^n$ multiplied with its transpose is the sum of all squared elements. A point is a minimum of a function in multiple variables without constraints, if the value of the Jacobi-Matrix at this point is 0 and the Hesse-Matrix at this point is positive definite. The following theorem summarises the sufficient conditions for such a minimum.

Theorem 3.10. *The function $f : \mathbb{R}^m \to \mathbb{R}$ has a local minimum at $\boldsymbol{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m$, if*

1. *the Jacobi matrix takes the value $\mathbf{0}$ at $\boldsymbol{x} \in \mathbb{R}^m$:*

$$J_f(\boldsymbol{x}) = \left(\frac{\partial f}{\partial x_i}(\boldsymbol{x})\right)_{i=1,\ldots,m} = \left(\frac{\partial f}{\partial x_1}(\boldsymbol{x}), \ldots, \frac{\partial f}{\partial x_m}(\boldsymbol{x})\right)^\top = \mathbf{0}$$

2. *the Hesse matrix $H_f(\boldsymbol{x}) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x})\right)_{i,j=1,\ldots,m}$ is positive definite:*

$$\boldsymbol{v}^\top H_f(\boldsymbol{x})\boldsymbol{v} > 0 \qquad \text{for all} \quad \boldsymbol{v} \in \mathbb{R}^m \backslash \mathbf{0}$$

The Jacobi and Hesse matrices of the residual sum of squares function, described by

$$\begin{aligned} RSS: \quad & \mathbb{R}^{p+1} \to \mathbb{R} \\ & \beta \longmapsto (\boldsymbol{y} - \tilde{\boldsymbol{X}}\beta)^\top (\boldsymbol{y} - \tilde{\boldsymbol{X}}\beta), \end{aligned}$$

can be calculated as

$$\begin{aligned} J_{RSS} &= \frac{\partial RSS}{\partial \beta} = -2\tilde{\boldsymbol{X}}^\top (\boldsymbol{y} - \tilde{\boldsymbol{X}}\beta) \\ H_{RSS} &= \frac{\partial^2 RSS}{\partial^2 \beta} = -2\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}. \end{aligned} \tag{3.7}$$

Let us assume that $\tilde{\boldsymbol{X}}$ has full rank. Then $\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}$, and therefore also the Hesse matrix of the RSS (3.7), is positive definite. As for fulfilling the first requirement in theorem 3.10, the coefficient estimates $\hat{\beta}$ must simply be calculated by setting

$$\begin{aligned} -2\tilde{\boldsymbol{X}}^\top (\boldsymbol{y} - \tilde{\boldsymbol{X}}\hat{\beta}) &= 0 \\ \Leftrightarrow \tilde{\boldsymbol{X}}^\top \boldsymbol{y} &= \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}\hat{\beta} \\ \Leftrightarrow \hat{\beta} &= (\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^\top \boldsymbol{y}. \end{aligned} \tag{3.8}$$

Then, by substitution in (3.5), the predicted response for the training data $\tilde{\boldsymbol{X}}$ can be calculated as

$$\hat{y} = \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^\top \boldsymbol{y}, \tag{3.9}$$

whereas the predicted response to any vector of new input values $\tilde{\boldsymbol{x}}_0 = (1, \boldsymbol{x}_0)$, $\boldsymbol{x}_0 \in \mathbb{R}^{1\times p}$, can be calculated by $\hat{y_0} = \tilde{\boldsymbol{x}}_0 \hat{\beta}$. The difference between the measured response and the estimate generated by a model is called the residual.

Definition 3.11. *The **residuals** for the linear model (3.4) are calculated as the difference between the measured response $\boldsymbol{y}$, from the training data, and the response $\hat{\boldsymbol{y}}$, estimated by the model:*

$$\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\beta}. \tag{3.10}$$

Residuals can be useful for evaluating the goodness of fit of a model. However, before reviewing criteria that can help with model selection, some characteristics and details of computation of the least squares fit are discussed.

The Gauss-Markov Theorem

The Gauss-Markov theorem, in rough words, states that the least squares estimates $\hat{\beta}$ of the coefficients β have a smaller variance than any other linear, unbiased estimate. So, to some extent, this famous result in statistics justifies the use of the least squares fit. In this subsection, the theorem is presented in the context of linear regression.

It is important to understand that, no matter how good a model is, there will always be an error in prediction. Denoting the true coefficients, which are unknown and approximated by the least squares fit, by β, it still holds that

$$\boldsymbol{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}, \tag{3.11}$$

where $\boldsymbol{X}$ is a set of predictors, $\boldsymbol{y}$ is the corresponding response and $\boldsymbol{\epsilon}$, with $\epsilon_i > 0$, is the error term. Such an error can be caused by influence of unknown factors, that were not included in the model, or unmeasurable variance. For such a model the Gauss-Markov Theorem is formulated as follows:

Theorem 3.12. *In the context of the model* (3.11)*, let the data matrix* $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ *have full rank and the the error* $\boldsymbol{\epsilon}$ *fulfill*

(i) the expected error is 0, i.e. $\mathbb{E}(\boldsymbol{\epsilon}) = 0$

(ii) the error has constant variance, i.e. $\boldsymbol{V}(\boldsymbol{\epsilon})_{jj} = 0$ *where* $\boldsymbol{V}(\boldsymbol{\epsilon})$ *is the covariance matrix of* $\boldsymbol{\epsilon}$

(iii) the errors are uncorrelated, i.e. $\boldsymbol{V}(\boldsymbol{\epsilon})_{ij} = 0$ *for* $i \neq j$

Then the least squares estimate $\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$ *has the minimum variance among all estimators of the form* $\boldsymbol{a}^\top \boldsymbol{y}$ *for arbitrary* $\boldsymbol{a} \in \mathbb{R}^n$*. [7]*

The assumptions (i) - (iii) are essential for estimating model prediction accuracy. Due to these additional assumptions concerning the error $\boldsymbol{\epsilon}$, it holds that the coefficient estimates $\hat{\beta}$ have constant variance and their expected value corresponds to the true coefficients β. Some results, e.g. whether a coefficient contributes significantly to the model, could be inaccurate, if $\boldsymbol{\epsilon}$ does not fulfill the requirements. Fortunately, results can still be trusted, if variance and expected value of errors do not deviate too far from the assumption. Additionally, it follows, that the residuals, as defined in (3.10), should display the the properties (i) - (iii). This will be discussed in section 3.2.2 in more detail.

Computation by QR Decomposition

In models with many variables or a large training data set, the direct computation of $\hat{\beta}$, as in (3.8), may not be efficient. There are more suitable methods of computation. According to MATLAB's documentation [50], the main fitting algorithm for built-in regression functions `fitlm` and `stepwiselm` uses QR factorisation. First, the definition and main characteristics of QR factorisation are summarised:

Definition 3.13. *The **QR factorisation** of a matrix $\boldsymbol{A}$ is its decomposition into the product of the matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$ with the following properties:*

- *$\boldsymbol{Q}$ is an **orthogonal matrix**, i.e. $\boldsymbol{Q}^\top \cdot \boldsymbol{Q} = \boldsymbol{Q} \cdot \boldsymbol{Q}^\top = I$*
- *$\boldsymbol{R}$ is an **upper triangular matrix**, i.e. $r_{ij} = 0 \quad \forall i > j$.*

Theorem 3.14. *For any matrix $\boldsymbol{A} \in \mathbb{R}^n$ one can find an orthogonal matrix $\boldsymbol{Q}$ and an upper triangular matrix $\boldsymbol{R}$ such that $\boldsymbol{A} = \boldsymbol{QR}$. This QR factorisation is unique, if $\boldsymbol{A}$ is invertible and $\boldsymbol{R}$ is restricted to only positive values in its diagonal.*

For easier understanding, let us at first consider a univariate linear model with no intercept, i.e. (3.4) with $p = 1$ and $\beta_0 = 0$. The data matrix $\boldsymbol{X}$ therefore corresponds to the vector $\boldsymbol{x} = (x_1, \dots, x_n)^\top$. The coefficient estimate, as calculated in (3.8), is then

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i^2 \right)^{-1} \cdot \sum_{i=1}^{n} x_i y_i = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\langle \boldsymbol{x}, \boldsymbol{x} \rangle},$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of $\mathbb{R}^n$, and the residuals are given by

$$\boldsymbol{r} = \boldsymbol{y} - \hat{\beta} \cdot \boldsymbol{x} = \boldsymbol{y} - \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\langle \boldsymbol{x}, \boldsymbol{x} \rangle} \cdot \boldsymbol{x}. \tag{3.12}$$

Now let us assume a multivariate linear model with p variables in $\mathbb{R}^n$ and an intercept. This means the data matrix has the form $\boldsymbol{X} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \dots, \boldsymbol{x}_p)$, where $\boldsymbol{x}_j \in \mathbb{R}^n$ and $\boldsymbol{x}_0 = \boldsymbol{1}$ denotes the vector of ones added to include the intercept. Remember, we assumed that $\boldsymbol{X}$ has full rank, which means that all $\boldsymbol{x}_j$ are linearly independent. Therefore the columns of $\boldsymbol{X}$ span an $(p+1)$-dimensional subspace in $\mathbb{R}^n$. This is often called the **column space of $\boldsymbol{X}$**.

While the independent variable $\boldsymbol{y}$ could be any element of $\mathbb{R}^n$, the estimated response $\hat{\boldsymbol{y}}$ is given by a linear combination of $\boldsymbol{x}_j$, $j = 0, \dots, p$, and is therefore an element of the subspace spanned by the columns of $\boldsymbol{X}$. Hence, the model output $\hat{\boldsymbol{y}}$ can be interpreted as a projection of $\boldsymbol{y}$ onto the column space of $\boldsymbol{X}$. The residuals, given by equation (3.10), are orthogonal to this subspace. This orthogonality is expressed in equation (3.8), which determined the choice of $\hat{\beta}$. Altogether, one can interpret the response estimate $\hat{\boldsymbol{y}}$ as the orthogonal projection of $\boldsymbol{y}$ onto the column space of $\boldsymbol{X}$.

To simplify the calculation of the coefficients $\hat{\beta}$, we can make the additional assumption, that the columns of $\boldsymbol{X}$ are orthogonal, i.e. $\langle \boldsymbol{x}_j, \boldsymbol{x}_k \rangle = 0$ for $j \neq k$. This is not a major restriction, since a procedure for orthogonalisation, such as the Gram-Schmidt-process, can be applied.

Definition 3.15. *The **Gram-Schmidt process** is a method for orthogonalising a set of vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k \in \mathbb{R}^n$. The orthogonal vectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_k \in \mathbb{R}^n$ are given by*

$$\begin{aligned} \boldsymbol{u}_1 &= \boldsymbol{v}_1 \\ \boldsymbol{u}_{j+1} &= \boldsymbol{v}_{j+1} - \sum_{i=1}^{j} \frac{\langle \boldsymbol{u}_j, \boldsymbol{v}_{j+1} \rangle}{\langle \boldsymbol{u}_j, \boldsymbol{u}_j \rangle} \boldsymbol{u}_j \qquad \text{for} \quad j = 1, \dots, k-1. \end{aligned} \tag{3.13}$$

Notice the similarity between the computation of the residuals for a univariate model (3.12) and the Gram-Schmidt process (3.13). An orthogonal set of vectors can be derived from any number of vectors and will span the same subspace in $\mathbb{R}^n$ as the original set of vectors. Therefore a basis can always be transformed to an orthogonal basis of the same vector space.

Theorem 3.16. *For any basis $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k)$ of a subset of $\mathbb{R}^n$, where $n, k \in \mathbb{N}$, there exists a basis of orthogonal vectors $(\boldsymbol{u}_1, \dots, \boldsymbol{u}_k)$ of the same subset. The vectors $(\boldsymbol{u}_1, \dots, \boldsymbol{u}_k)$ can be obtained using the Gram-Schmidt process.*

Returning to the computation of coefficients for a multivariate model, under the given restrictions, formula (3.8) can be simplified. Each element of the matrix $\boldsymbol{X}^\top \boldsymbol{X}$ is given by

$$(\boldsymbol{X}^\top \boldsymbol{X})_{ij} = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \qquad \text{for all} \quad i = 1, \dots, n, \ j = 0, \dots, p.$$

Hence, due to the orthogonality assumption, $\boldsymbol{X}^\top \boldsymbol{X}$ is a diagonal matrix. In consequence, the inverse is given by

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \operatorname{diag}\left(\frac{1}{\langle \boldsymbol{x}_0, \boldsymbol{x}_0 \rangle}, \frac{1}{\langle \boldsymbol{x}_1, \boldsymbol{x}_1 \rangle}, \dots, \frac{1}{\langle \boldsymbol{x}_p, \boldsymbol{x}_p \rangle} \right).$$

Since $\boldsymbol{X}^\top \boldsymbol{y} = (\langle \boldsymbol{x}_0, \boldsymbol{y} \rangle, \langle \boldsymbol{x}_1, \boldsymbol{y} \rangle, \dots, \langle \boldsymbol{x}_p, \boldsymbol{y} \rangle)^\top$, each element of the coefficient vector $\hat{\beta}$ is given by

$$\hat{\beta}_j = ((\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y})_j = \frac{\langle \boldsymbol{x}_j, \boldsymbol{y} \rangle}{\langle \boldsymbol{x}_j, \boldsymbol{x}_j \rangle} \qquad \text{for all} \quad j = 0, \dots, p. \tag{3.14}$$

The computation of coefficient estimates for a linear model with p variables and an intercept, under the assumption that the data matrix $\boldsymbol{X} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \dots, \boldsymbol{x}_p)$ has full rank, is summarised in Algorithm 1. The algorithm uses the Gram-Schmidt process to obtain an orthogonal basis $\boldsymbol{Z} = (\boldsymbol{z}_0, \boldsymbol{z}_1, \dots, \boldsymbol{z}_p)$ for the column space of $\boldsymbol{X}$, as well as the computations leading up to equation (3.14).

Since the procedure uses the Gram-Schmidt process, the columns of $\boldsymbol{Z}$ form an orthogonal basis of the column space of $\boldsymbol{X}$. Hence, the regression coefficients computed correspond to the projection onto the same subspace and therefore also to the least squares fit. The orthogonal basis is, in fact, comprised of the residuals of the regression of $\boldsymbol{x}_j$ on all previous orthogonal basis vectors $\boldsymbol{z}_0, \dots, \boldsymbol{z}_{j-1}$. This is expressed in (3.15). Of course, $\boldsymbol{x}_j$ can also be rewritten as a linear combination of $\boldsymbol{z}_0, \dots, \boldsymbol{z}_{j-1}$.

Since $\boldsymbol{z}_p$ only appears in $\boldsymbol{x}_p$ with a non-zero coefficient, we are assured that $\hat{\beta}_p$, the coefficient of the regression of $\boldsymbol{y}$ on $\boldsymbol{z}_p$, is, in fact, also the multiple regression coefficient corresponding to the regression of $\boldsymbol{y}$ on $\boldsymbol{x}_p$. By rearranging variables it becomes clear that the same holds for any other variable $\boldsymbol{x}_j$, $j = 0, \dots, p$. Therefore the multiple regression coefficient of a variable $\boldsymbol{x}_j$ is always given by the univariate regression coefficient of regressing the response $\boldsymbol{y}$ on $\boldsymbol{z}_j$, the residual after regressing $\boldsymbol{x}_j$ on all other variables. The Elements of Statistical Learning (Hastie, Tibshirani, Friedman, [20]) gives the following interpretation: '*The multiple regression coefficient $\hat{\beta}_j$ represents the additional contribution of $\boldsymbol{x}_j$ on $\boldsymbol{y}$, after $\boldsymbol{x}_j$ has been adjusted for all other variables $\boldsymbol{x}_0, \dots, \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, \dots, \boldsymbol{x}_p$*'.

Algorithm 1 Regression by Successive Orthogonalisation

Require: $\boldsymbol{X} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, $\boldsymbol{y}$

Initialise $\boldsymbol{z}_0 = \boldsymbol{x}_0 \quad (= \mathbf{1})$ ▷ step 1 of the G.-S. process

for $j = 1, 2, \ldots, p$ **do** ▷ steps 2 to p+1 of the G.-S. process

get coefficients $\hat{\gamma}_{lj}$ of the linear model $\sum_{i=0}^{j-1} \boldsymbol{z_i}\hat{\gamma}_{ij} \approx \boldsymbol{x}_j$:

$$\begin{aligned} \hat{\gamma}_{lj} &= \frac{\langle \boldsymbol{z}_l, \boldsymbol{x}_j \rangle}{\langle \boldsymbol{z}_l, \boldsymbol{z}_l \rangle} \qquad \text{for} \quad l = 1, \ldots, j-1 \\ \boldsymbol{z}_j &= \boldsymbol{x}_j - \sum_{i=0}^{j-1} \hat{\gamma}_{ij} \boldsymbol{z_i} \end{aligned} \tag{3.15}$$

end for

get the regression coefficients $\hat{\beta}_p = \frac{\langle \boldsymbol{z}_p, \boldsymbol{y} \rangle}{\langle \boldsymbol{z}_p, \boldsymbol{z}_p \rangle}$

Algorithm 1 can be represented using matrices. Denote the data matrix by $\boldsymbol{X} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, its orthogonalisation by $\boldsymbol{Z} = (\boldsymbol{z}_0, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_p)$ and

$$\boldsymbol{\Gamma} = \begin{pmatrix} 1 & \hat{\gamma}_{01} & \hat{\gamma}_{02} & \cdots & \hat{\gamma}_{0(p-1)} & \hat{\gamma}_{0p} \\ 0 & 1 & \hat{\gamma}_{12} & \cdots & \hat{\gamma}_{1(p-1)} & \hat{\gamma}_{1p} \\ \vdots & 0 & 1 & \ddots & \vdots & \vdots \\ \vdots & \vdots & 0 & \ddots & \hat{\gamma}_{p(p-1)} & \vdots \\ \vdots & \vdots & \vdots & \ddots & 1 & \hat{\gamma}_{(p-1)p} \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix},$$

then (3.15) in matrix form is given by

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Gamma}.$$

Introducing the diagonal matrix $\boldsymbol{D} = \operatorname{diag}(\langle \boldsymbol{x}_0, \boldsymbol{x}_0 \rangle, \langle \boldsymbol{x}_1, \boldsymbol{x}_1 \rangle, \ldots, \langle \boldsymbol{x}_p, \boldsymbol{x}_p \rangle)$, we get the QR decomposition of $\boldsymbol{X}$

$$\boldsymbol{X} = \underbrace{\boldsymbol{Z}\boldsymbol{D}^{-1}}_{\boldsymbol{Q}} \underbrace{\boldsymbol{D}\boldsymbol{\Gamma}}_{\boldsymbol{R}}.$$

The orthogonality of $\boldsymbol{Q}$ can be verified easily and $\boldsymbol{R}$ is an upper triangular matrix by definition. This makes the calculation of the least squares estimate significantly easier. The coefficient and response estimate, when substituting the QR decomposition in (3.8) and (3.9), are given by

$$\begin{aligned} \hat{\beta} &= \boldsymbol{R}^{-1}\boldsymbol{Q}^\top \boldsymbol{y} \\ \hat{\boldsymbol{y}} &= \boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{y}. \end{aligned}$$

The question remains, how restrictive assuming a data matrix $\boldsymbol{X}$ with full rank is. If $\boldsymbol{X}$ did not have full rank, $\boldsymbol{X}^\top\boldsymbol{X}$ would be singular. In consequence, the system of linear equations for the coefficients $\hat{\beta}$ (3.8) would not have a unique solution. The fact that the response estimate $\hat{\boldsymbol{y}}$ is a projection of $\boldsymbol{y}$ onto the column space of $\boldsymbol{X}$, however, remains. Hence, the final estimate of $\boldsymbol{y}$ stays unchanged.

Rank deficiencies can usually be avoided by checking the chosen predictors for redundancy. Variables, that can be formed as a linear combination of other variables, can be omitted without losing information. Mostly, regression software packages implement strategies for removing such redundancies. Another case, in which $\boldsymbol{X}$ may not have full rank, is where the number of predictors p exceeds the number of measurements for training. This may occur in some cases, but all in all, assuming a data matrix with full rank is usually plausible.

3.2.2 Assessing Model Quality

All models depicting a given system will have some inherent error. Especially for predictions, a model requires a certain generality for the application of findings in training data to new data to be reasonable. The better a model generalises, the smaller the expected prediction error on an independent test sample.

This generality is in opposition with the flexibility of a model to adapt to the underlying structure of the training data. A quadratic or higher order polynomial model, for example, is considerably more flexible than a linear model. Increasing model flexibility, however, does not always yield better results. Too much flexibility when fitting a model function can lead to the detection of patterns, that are actually random effects instead of true model function characteristics. This is called **overfitting**. The model seems to produce great results on training data, but when confronted with previously unseen test data, performance deteriorates. Some of the formerly detected, maybe random, patterns are not present in the new data. [21, Ch. 2]

Figure 3.1 shows an example of overfitting. The smoothing spline fit with higher flexibility, depicted in green, performs well on the training data set, but has a high error on test data. The fitted function, which is clearly overfit, has no similarity to the true relationship between data and response, shown in black. [21, Ch. 2]

This section will outline some of the most used characteristics for evaluating the goodness fit of a model and is, unless otherwise indicated, based on [21, Ch. 2-3]. One key feature, the residual sum of squares, was already introduced in definition 3.9. It indicates the overall fit of the model to all training data, but does not take model flexibility into account and can also be used to deduce further interesting characteristics.

In general, it is always advisable to consider multiple indicators of model performance when choosing a model, since following one characteristic alone, can easily lead to overfitting.

Mean Square Error

The mean squared error (MSE) is a measure for how well predictions match observed data. It compares model response estimates with the corresponding observed response values.

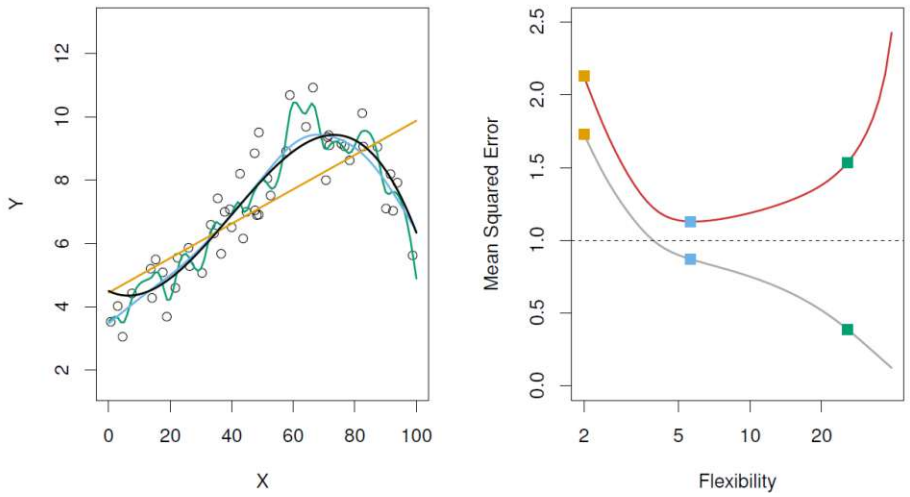

Figure 3.1: Left: The dots represent data simulated from the function f, depicted in black. Different estimates of f are shown: the linear regression line (orange) and two smoothing splines fits (blue and green) with different levels of smoothness. **Right:** The behaviour of mean squared error (MSE) on training data (grey) and test data (red) with changing model flexibility. The dotted line indicates the minimum possible MSE. The squares indicate the MSEs corresponding to the model function estimates shown on the left. Image taken from [21, Ch. 2].

Definition 3.17. *The **mean squared error (MSE)** is a measure for prediction accuracy of a model and is calculated as*

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\boldsymbol{x}_i))^2, \tag{3.16}$$

where $\boldsymbol{x}_i$ denotes the measured data, y_i the corresponding observed response and $\hat{f}$ is the estimated model function, that maps a set of inputs to a response estimate.

The MSE is small if predicted responses are close to observed responses of the test data. The further predictions are from the true values, the larger MSE will be. Since the square root is a monotone function, the root of the mean squared error shows exactly the same behaviour. Hence, the MSE and its square root, the root mean squared error (RMSE), are comparable model characteristics.

Definition 3.18. *The **root mean squared error (RMSE)** is computed as the square root of MSE and is also an indicator for prediction accuracy in a model.*

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\boldsymbol{x}_i))^2}. \tag{3.17}$$

One must distinguish between training and test RMSE. When computing (3.17) using training data, this is called the training RMSE, this gives no indication of accuracy of model predictions from previously unknown data. The test RSME, however, is calculated

using only previously unseen test data and therefore gives information on the prediction accuracy for new cases. Hence, RMSE should mainly be observed on test data sets. As seen in figure 3.1, a large difference between RMSE on test and training data sets can be an indicator for overfitting.

Bias-Variance Trade-Off

There are two important, but conflicting, properties that must be taken into account when choosing a model: bias and variance. Bias refers to the error made when trying to depict a complicated real-world system in a structurally much simpler model. Variance, on the other hand, describes the dependency of the model function on the training data.

Remember that any linear regression model, as in (3.11), will have an error in prediction. A more general model can be written as

$$\boldsymbol{y} = f(\boldsymbol{X}) + \epsilon, \tag{3.18}$$

where f is the true model function, X the data matrix, y the response and ϵ denotes the error. In general, of course, we do not even know the true model function f, but an approximation $\hat{f}$. The expected error in prediction for a fixed training data set $\boldsymbol{X}$ and model function estimate $\hat{f}$ with $\hat{f}(\boldsymbol{X}) = \hat{y}$ can be decomposed into

$$\begin{aligned} \mathbb{E}(\boldsymbol{y} - \hat{\boldsymbol{y}})^2 &= \mathbb{E}[f(\boldsymbol{X}) + \epsilon - \hat{f}(\boldsymbol{X})]^2 \\ &= \underbrace{[f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X})]^2}_{\text{reducible}} + \underbrace{\operatorname{Var}(\epsilon)}_{\text{irreducible}} . \end{aligned} \tag{3.19}$$

The first term in (3.19), the reducible error, can be influenced by the choice of model, while the second term, the irreducible error, is already present in (3.18) and can not be eliminated.

The expected MSE, as defined in (3.16), for any fixed data point x and its corresponding response y, can be further decomposed into bias and variance components

$$\mathbb{E}(y - \hat{f}(x)) = \operatorname{Var}(\hat{f}(x)) + [\operatorname{Bias}(\hat{f}(x))]^2 + \operatorname{Var}(\epsilon). \tag{3.20}$$

Again, ϵ denotes the irreducible error. Note that variance, as the expected value of squared deviation, is always positive. Therefore all terms in (3.20) are greater than zero and, to minimise the expected error in prediction, a model must achieve low bias and low variance simultaneously. This, however, can be tricky. As seen in figure 3.1, a more flexible model reduces bias, since it better adapts to the training data. However, this comes at a cost: the model loses generality and has a high variance. If we choose to minimise the variance, the model may not have the flexibility to find structures of interest in the training data. Hence, bias and variance need to be carefully balanced to achieve the best prediction results. Figure 3.2 depicts the relationship between bias, variance and total error. The bottom of the trough in the total error, shown in black in figure 3.2, would be a good candidate for such a balance. Comparing training and test MSE can be very helpful, when considering the bias-variance trade-off.

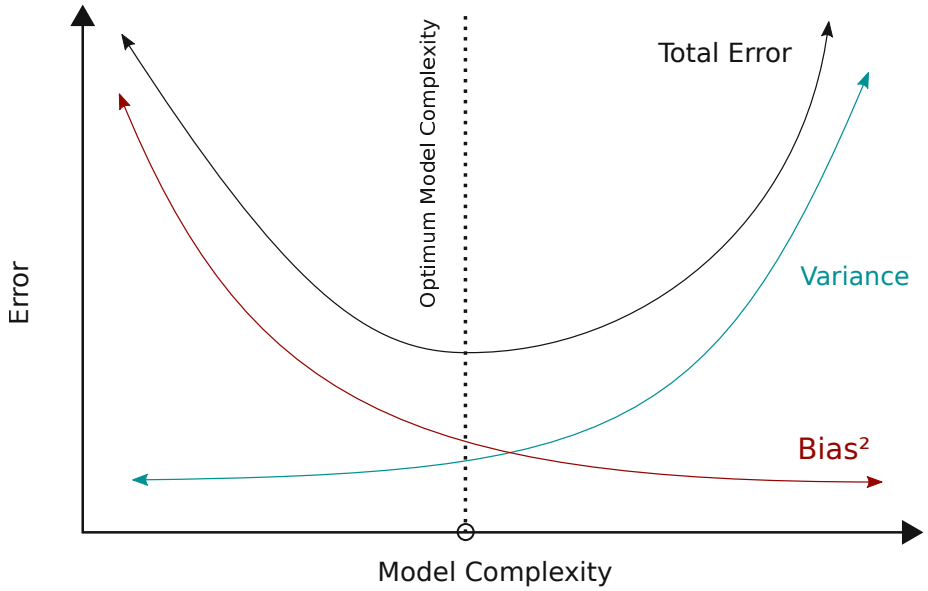

Figure 3.2: The variance component of the expected error is shown in blue, the squared bias component in red. As flexibility, or as it is called here, complexity, in the model increases, so does the variance. As complexity decreases, so does the bias. The black curve shows the total expected error and he dotted line indicates a good choice of model complexity. Image used under the license [13].

Residual Squared Error

The residual squared error (RSE) measures the average amount that the actual response $\boldsymbol{y}$ will vary from the regression line with known coefficients. In other words, the RSE estimates the standard deviation of ϵ in equation (3.18). The RSE is derived from the residual sum of squares, as defined in (3.6).

Definition 3.19. *The **residual squared error (RSE)** measures the standard deviation in the error between model response and response estimate. It is calculated by*

$$RSE = \sqrt{\frac{1}{n-p-1} RSS}.$$

The RSE gives the average deviation from the regression line in units. This is a good indicator for the model fit, since one can expect more accurate predictions for models with lower RSE. However, it can sometimes be difficult to interpret. How many units deviation is acceptable is often highly dependent on the context.

R Squared

The R^2 statistic is a measure of lack of fit of a model. It is derived from the residual squared error. The R^2 statistic is a relative measure rather than an absolute measure, as RSE would be, and is therefore sometimes easier to interpret.

The value of R^2 indicates the proportion of variance in the response $\boldsymbol{y}$ that can be explained by the variables $\boldsymbol{X}$. It is independent of the scale of $\boldsymbol{y}$ and always assumes a value between 0 and 1. For the calculation of R^2 an additional definition is required.

Definition 3.20. *The **total sum of squares (TSS)** is a measure for the total variance in the response $\boldsymbol{y}$. It is calculated as*

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

where $\bar{y}$ is the mean of all measured responses y_i, $i = 1, \ldots, n$.

The R^2 statistic is then computed using RSS, describing the variance in the response that can't be explained by the model, and TSS, the total variance in the response. Hence, TSS − RSS corresponds to the variance in the independent variable explained by regression.

Definition 3.21. *The **R^2 statistic** measures the proportion of variance in the response $\boldsymbol{y}$ that can be explained using the predictors $\boldsymbol{X}$. It is computed as*

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}.$$

While the value of R^2 is always within a given range and easier to interpret than RSE, what value is good still depends greatly on the application. For example, in physics it might be known, that data has a true linear relationship. Then an R^2 value close to 1 can be expected. In biology and psychology however, the true relationship of response and predictors may be considerably more complicated and some influential factors may be unknown. In this case even a very low R^2, the variance in the response explained by the model, may already be considered good.

Even though R^2 is a useful indicator of model fit, maximising its value is not a suitable approach. When adding new variables to a model, R^2 always increases. So, when making decisions, the magnitude of change should also be considered. Also, models that are overfit to the training data can exhibit large R^2 values.

The F-Test

The F-statistic is usually used to determine whether a multiple linear regression model describes a significant relationship. The F-Test is performed by computing the F-statistic and tests the null hypothesis, that all coefficients are zero versus the alternative hypothesis, that at least one coefficient is non-zero. It gives a p-value for the significance of the relationship described by the given model in comparison to a constant model.

The F-statistic for the F-test for a model with $p + 1$ variables is given by

$$F = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)},$$

where TSS denotes the total sum of squares, RSS the residual sum of squares and n is the number of measurements.

The F-statistic can also be adapted to test the significance of a given group of variables. For two models, where $\mathcal{M}_0$ is a smaller model, with $p_0 + 1$ variables, nested in a larger model $\mathcal{M}_1$, with $p_1 + 1$ variables, denote by RSS_i the corresponding RSS. The F-statistic, measuring the change in RSS per additional variable in $\mathcal{M}_1$, is the given by

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(n - p_1 - 1)}. \tag{3.21}$$

The test using (3.21) as its test statistic outputs a p-value indicating the significance of the group of variables in $\mathcal{M}_1$ that are not included in $\mathcal{M}_0$. [20, Ch. 3]

Residuals

As already stated previously, the error properties (i) - (iii) from theorem 3.12 are essential for assessing the quality of a model and should also be represented in the residuals. It is quite simple to check the residuals for these properties, since many problems can be seen instantly by looking at residual plots.

Plotting the residuals against the fitted values, as seen in figure 3.3, can indicate whether a suitable model was chosen. If the true model function is non-linear, the results of a linear regression model can not be trusted. A distinctive pattern in the plot, as in figure 3.3 (a), is a strong indicator of non-linearity in the data. In such a case, other model functions and methods should be considered.

One can also easily judge, whether the residuals are centred around zero, which should be the case if the the expected value of errors fulfils (i). The image (b) in figure 3.3 shows an example of errors with non-constant variance, i.e. a violation of (ii). As the fitted value increases, so does the mean error, which shows in the plot as a funnel shape. Ideally, the plot of residuals against corresponding fitted values shows that the residuals deviate from zero to the same extent for all fitted values. Positive and negative residuals should be equally frequent. Figure 3.3 (c) shows a residual plot, that does not indicate any problems.

Finally, as in theorem 3.12 (iii), there should be no correlation in residuals. The lagged residual plot shown in figure 3.4 is useful for judging whether residuals are uncorrelated. The residual r_t for the t-th measurement x_t is plotted against r_{t-1}, the residual of the previous measurement. Figure 3.4 (a) indicates that, for a positive residual, the following residual is also likely to be positive. They seem to be correlated. Ideally, the lagged residual plot looks roughly circular, as in 3.4 (b).

Of course, the hoped-for residual plots do not indicate, that the corresponding model is actually a good fit. Nevertheless, residual plots are a valuable tool for assessing model accuracy, since they help detect many problems at an early stage.

3.2.3 Stepwise Variable Selection

In some cases, especially in those where there are many different predictors, not all data is actually related to the response. It can be hard to judge which predictors yield the most information. However, a model should include as many variables as necessary, to assure adequate depiction of the relationship, but as little as possible, to avoid redundancies. Variable selection is the process of determining which subset of predictors add valuable

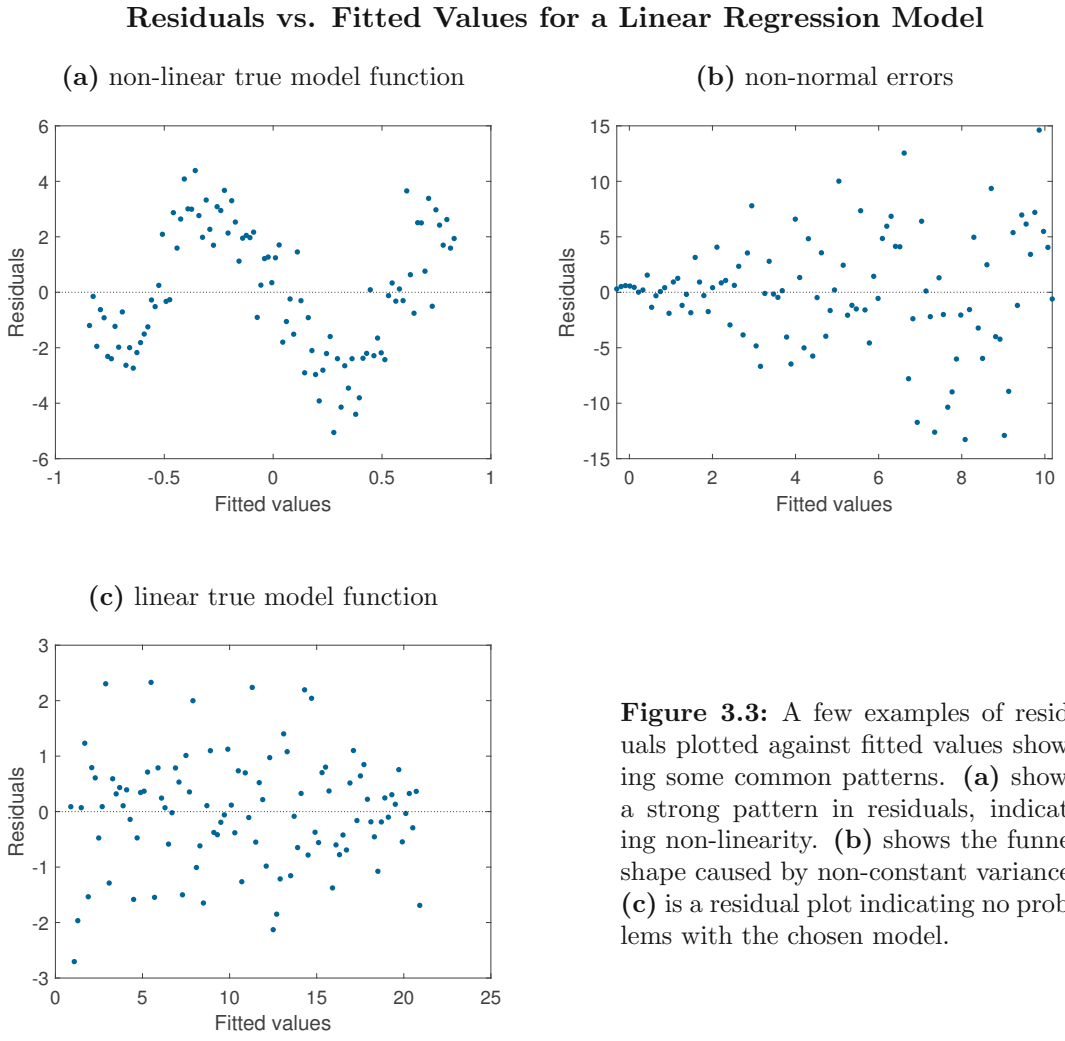

Figure 3.3: A few examples of residuals plotted against fitted values showing some common patterns. **(a)** shows a strong pattern in residuals, indicating non-linearity. **(b)** shows the funnel shape caused by non-constant variance. **(c)** is a residual plot indicating no problems with the chosen model.

information and should be included in the model. The entire subsection on variable selection is based on [21, Ch. 3].

Even for a moderately high number of variables, it is impractical to simply create and compare all models with different variable combinations. This would amount to 2^p models for p variables. Additionally, one might run into a statistical problem: if enough models are generated, it is very likely to find some models, that may look good on training data, but do not describe any real relationship. Two automated, more efficient approaches, forward and backward stepwise selection, are discussed in this subsection.

The question remains, on what basis to decide whether a model, using a certain variable subset, is better than one using another subset. Some criterion is needed for automated evaluation and comparison of models. Five established criteria are presented at the end of this subsection.

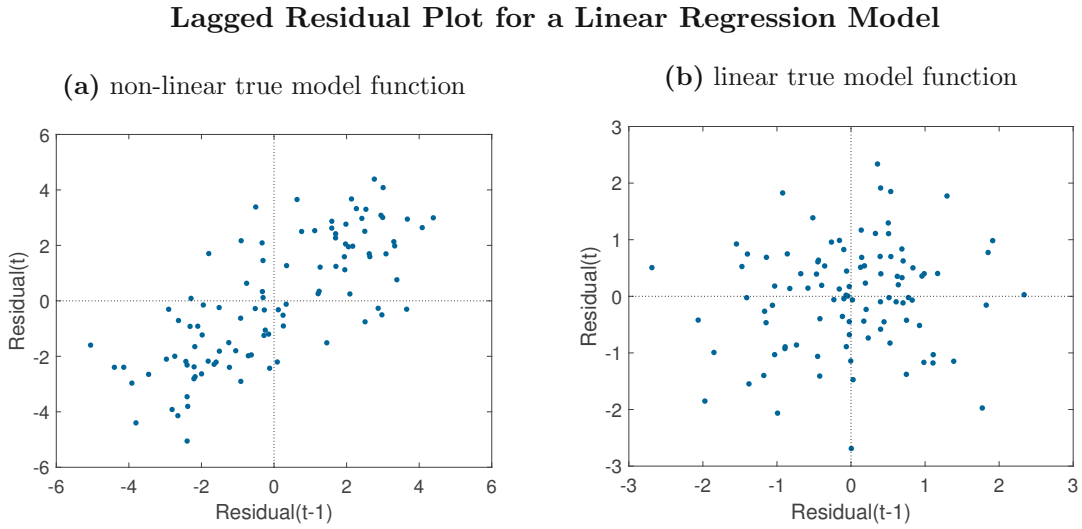

Figure 3.4: A few examples of lagged residual plots, where all residuals are plotted against the residual preceding it in the list. **(a)** shows a strong pattern in residuals, indicating correlation. If a residual is positive, the following residual is likely to have the same sign. **(b)** shows no distinctive correlation pattern in the residuals. The circular shape indicates, that one residual yields no information on the following one.

Forward Stepwise Selection

Stepwise selection always starts with an initial model. Forward stepwise selection starts with the null model, that contains only an intercept, but no predictors. In each step of the algorithm, one predictor is added to the model. To be more exact, the variable that gives the greatest additional improvement of fit is included.

The details of the forward stepwise selection procedure are given in algorithm 2. It only requires generating the null model and $p - k$ models in each step. This amounts to a total of

$$1 + \sum_{k=0}^{p-1}(p-k) = 1 + \frac{p(p+1)}{2}$$

models, which is considerably less than the 2^p models needed for testing all combinations.

Choosing the best model $\mathcal{M}_{k+1}$ in each step of algorithm 2 varies with the choice of criterion, all of which will be explained in detail in the next pages. With RSS as criterion, for example, one simply chooses the model with the lowest RSS.

Even though the forward stepwise selection algorithm has major computational advantages, it does not guarantee finding the best subset of variables from the 2^p combinations. If, for example, the best model with two variables $\mathcal{M}_2$ does not contain the variable chosen for the best model with one variable $\mathcal{M}_1$, then $\mathcal{M}_2$ will not be found by forward stepwise selection.

Algorithm 2 Forward Stepwise Selection

Require: data matrix $\boldsymbol{X}$ with p predictors, criterion $\mathcal{C}$

create the initial null model $\mathcal{M}_0$ ▷ $\mathcal{M}_0$ is a model with no predictors

for $k = 0, \ldots, p-1$ **do**

1. consider each model that adds exactly one predictor to $\mathcal{M}_k$ ▷ $p-k$ models
2. choose $\mathcal{M}_{k+1}$ to be the best of the $p-k$ models according to the criterion $\mathcal{C}$ ▷ possible criteria: R^2, adjusted R^2, RSS, AIC, BIC

end for

select one model from $\mathcal{M}_0, \ldots, \mathcal{M}_p$, which performs best in the chosen criterion $\mathcal{C}$

Backward Stepwise Selection

Backward stepwise selection follows a similar principle to forward stepwise selection, but with different starting conditions. The full least squares model, containing all p variables and an intercept, is used as the initial model. In place of adding variables to achieve better models, in each step the least useful predictor is excluded. The details of the procedure are given in algortihm 3.

Algorithm 3 Backward Stepwise Selection

Require: data matrix $\boldsymbol{X}$ with p predictors, criterion $\mathcal{C}$

create the initial full model $\mathcal{M}_0$ ▷ $\mathcal{M}_0$ is the model containing all p predictors

for $k = p-1, \ldots, 1$ **do**

1. consider each model that removes exactly one predictor from $\mathcal{M}_k$ ▷ k models
2. choose $\mathcal{M}_{k-1}$ to be the best of the k models according to the criterion $\mathcal{C}$ ▷ possible criteria: R^2, adjusted R^2, RSS, AIC, BIC

end for

select one model from $\mathcal{M}_0, \ldots, \mathcal{M}_p$, which performs best in the chosen criterion $\mathcal{C}$

Just as with forward stepwise selection, $1 + \frac{p(p+1)}{2}$ models need to be generated and evaluated. Backward stepwise selection has the same computational benefits as forward stepwise selection. It also remains true, that backward stepwise selection does not necessarily find the best variable combination, as variables needed for an optimal fit with less predictors may be excluded at an early stage.

Variations of Stepwise Selection

In general, forward and backward stepwise selection give similar, but not identical results. The two procedures can be combined to a hybrid approach. In this case, after predictors are added by forward stepwise selection, the procedure may also remove variables, as in backward stepwise selection, if they no longer contribute to the model.

It is also possible to adapt the procedures by defining criteria thresholds, that define what makes a model good enough according to the chosen criterion. In this case the best

model need not be chosen in any step. Predictors are simply added or removed if the change affects the criterion as required by the defined threshold, e.g. decrease in RSS by a given amount.

Criteria

As stated previously, many different criteria can be used to automatically evaluate a model. The **residual sum of squares (RSS)** is already known from definition 3.9. In general, one could say, the lower the RSS, the better a model. Hence, when using the RSS as the criterion $\mathcal{C}$, models with smaller RSS will be favoured. However, it is important to know, that RSS always decreases with an increasing number of predictors. Therefore, it is not particularly suitable for comparing models with different numbers of predictors and, when using a threshold, one should be looking for a substantial decrease.

The $\boldsymbol{R^2}$ **statistic**, from definition 3.21, can also be used as a criterion $\mathcal{C}$. In this case, higher values are better than lower values. The R^2 statistic has similar issues as RSS, it always increases when additional variables are added. Hence, models with higher R^2 will be favoured, but it should not be used for comparing models with different number of variables and, when using a threshold, a considerable increase should be desired.

R^2 can be adjusted for the number of predictors, to counteract these effects. Remember, that R^2 was defined as $1 - \frac{RSS}{TSS}$, where TSS is the total sum of squares.

Definition 3.22. *The* ***adjusted R^2 statistic*** *measures the proportion of variance in the response that is explained by the predictors, irrespective of the number of variables. It is adjusted for the number of predictors used in the model. For a model with p variables and n measurements, it is computed as*

$$\textit{adjusted } R^2 = 1 - \frac{\frac{RSS}{n-p-1}}{\frac{TSS}{n-1}}.$$

As with R^2, when using **adjusted $\boldsymbol{R^2}$** as the criterion $\mathcal{C}$, larger values indicate a better model. Due to the adjustment, this statistic may actually decrease, if an added predictor does not contribute substantially to a model. The addition of unnecessary variables in penalised.

The **Akaike information criterion (AIC)** for the standard linear model, as in definition 3.8, is defined as follows:

Definition 3.23. *The* ***Akaike information criterion*** *for a model with p predictors is computed as*

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2p\hat{\sigma}^2)$$

where n is the number of measurements and $\hat{\sigma}^2$ is an estimate of the variance of the error ϵ associated with the response.

For simplicity, an additive constant is omitted in the formula for AIC. It is easy to see, that the AIC is derived from the RSS, and therefore smaller AIC indicates a better model. However, the AIC also penalises the addition of predictors, that only reduce RSS by a small

value. AIC is intended to estimate test error, rather than training error and is therefore a good choice for the criterion $\mathcal{C}$.

Finally, the **Bayesian information criterion (BIC)** can also be used as the criterion $\mathcal{C}$. It is defined as follows, when omitting some, for this case irrelevant, constants:

Definition 3.24. *The **Bayesian information criterion** for a model with p predictors is computed as*

$$BIC = \frac{1}{n}(RSS + \log(n)p\hat{\sigma}^2)$$

where n is the number of measurements and $\hat{\sigma}^2$ is an estimate of the variance of the error ϵ associated with the response.

Again, the BIC tends to take smaller values for models with less test error. So models with low BIC are considered to be better. BIC tends to penalise the addition of unnecessary predictors more than AIC. The additive term $\log(n)p\hat{\sigma}^2$ in BIC takes the place of $2p\hat{\sigma}^2$ in AIC and for $n > 7$, it already holds that $\log(n) > 2$.

3.3 Principal Component Analysis and Regression

Principal component analysis (PCA) is a dimension reduction technique for regression. A new, smaller, set of features is derived from the data matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \in \mathbb{R}^{n \times p}$. The smaller set of features can later be used as predictors in a regression model. Unless stated otherwise, information in this section was taken from chapters 6 and 12 of [21].

The basic idea is, that, even though all observations are in the p-dimensional column space of $\boldsymbol{X}$, not all of these dimensions hold the same amount of information. In terms of PCA, the amount of information held in a dimension is measured by the variance in the data along this dimension.

Principal components (PCs) are linear combinations of the predictors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$, where the sum of squared coefficients amounts to one. This constraint on the coefficients, also termed normalisation, is necessary. Without it, the variance could be blown up arbitrarily by simply increasing the coefficients. Figure 3.5 shows two-dimensional data and the corresponding principal component directions. Principal components depend only on the given regression data, not the response, and can therefore also be useful for data analysis.

Only a rough outline of the computation of PCs will be given. The first principal component $\boldsymbol{z}_1$ has the form

$$\boldsymbol{z}_1 = \phi_{11}\boldsymbol{x}_1, \ldots, \phi_{1p}\boldsymbol{x}_p \quad \text{where} \quad \sum_{j=1}^{p} \phi_{1j}^2 = 1.$$

Since the first principal component should be the direction of maximum variance, $\phi_{11}, \ldots, \phi_{1p}$ are chosen to maximise the sample variance. Since, without loss of generality, it can be assumed that each column of $\boldsymbol{X}$ is centred around zero, this amounts to

$$\boldsymbol{z}_1 = \max_{\phi_{11},\ldots,\phi_{1p}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{1j} x_{ij} \right)^2 \right\} \quad \text{where} \quad \sum_{j=1}^{p} \phi_{1j}^2 = 1. \tag{3.22}$$

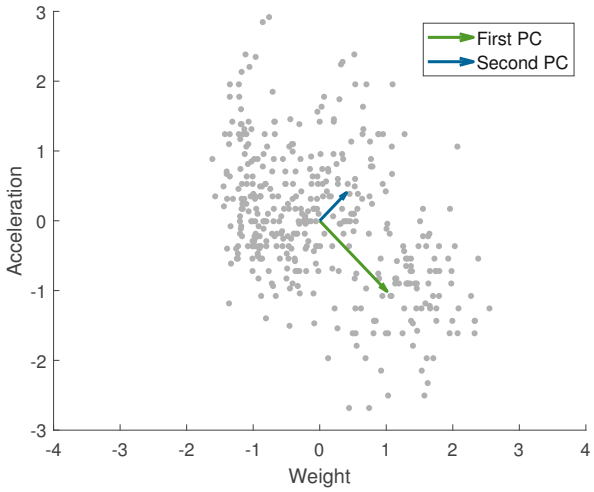

Figure 3.5: The scatter plot shows weight and acceleration data from the MATLAB data set 'carbig' along with its principal components. The arrow in orange shows first associated principal component, the smaller, blue arrow depicts the second PC.

The second principal component $\boldsymbol{z}_2$ can be calculated just like $\boldsymbol{z}_1$ in 3.22 with an additional constraint: $\boldsymbol{z}_2$ and $\boldsymbol{z}_1$ must be uncorrelated. For vectors this simply translates to the vectors being orthogonal to each other.

Equation (3.22) can be solved via the eigen decomposition of the matrix $\boldsymbol{X}^\top \boldsymbol{X}$, which, up to the factor n, corresponds to the sample covariance matrix of the data $\boldsymbol{X}$ [20, Ch. 3]. The eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}$ correspond to the variances of the PCs and the corresponding eigenvectors, ordered by the size of eigenvalues, are the PCs. The terminology is summarised in definition 3.25.

Definition 3.25. ***Principal component analysis (PCA)*** *is a technique, in which new variables, the principal components, are derived from the data matrix* $\boldsymbol{X}$*. The* ***principal components (PCs)*** *correspond to the eigenvectors, ordered by the size of their respective eigenvalues in decreasing order, of the matrix* $\boldsymbol{X}^\top \boldsymbol{X}$*. They are of the form*

$$\boldsymbol{z}_i = \phi_{i1}\boldsymbol{x}_1, \ldots, \phi_{ip}\boldsymbol{x}_p \quad \textit{where} \quad \sum_{j=1}^{p} \phi_{ij}^2 = 1.$$

For the i-th PC, ϕ_{ij}*,* $j = 1, \ldots, p$*, are called the* ***loadings*** *and the the elements of* $\boldsymbol{z}_i$ *are the* ***scores****.*

Due to the properties of the eigen decomposition, the PCs are uncorrelated and form a basis of a subspace. Geometrically, principal component analysis can be interpreted as the projection onto this subspace. For each PC the loadings are unique up to a sign flip, which is consistent with the PC describing a direction. The score vector is also unique up to a sign flip, since $\mathrm{Var}(\boldsymbol{z}) = \mathrm{Var}(-\boldsymbol{z})$.

For **principal component regression (PCR)**, the principal components are constructed and used as predictors in a multivariate regression model. The model can be fit using the least squares method explained in section 3.2.1.

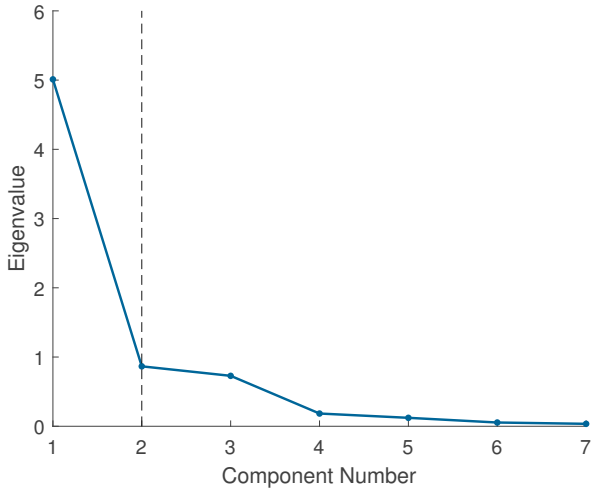

Figure 3.6: The Scree plot shows the eigenvalues, which can be a measure for variance, for each principal component. In this case, using more than two PCs is not advisable, since, in comparison, there is not much information gain.

However, not all principal components are used. This would simply lead to the least squares fit on the data matrix $\boldsymbol{X}$. Instead, to achieve a dimension reduction, only some principal components are included. In theory, as long as the chosen principal components explain a decent amount of the variance in the data, they should hold enough information for a good regression model. Of course, it is desirable to use the smallest set of PCs possible, that achieves this. Also, the question remains, how much explained variance suffices.

There is no fixed amount of variance, that the PCs should explain. In practice, usually more than 80% of the total variance is decent. However, the Scree plot presents a good technique for visual judgement of the number of principal components.

Definition 3.26. *The **Scree plot**, connected to the PCA of a data matrix $\boldsymbol{X}$, depicts the explained variance of data by each principal component. An example of a scree plot is shown in figure 3.6*

A good candidate for the number of PCs to use in regression is given by the point in the scree plot, where the curve flattens. It is often characterised by a sort of elbow shape. Intuitively, this is easy to understand, since all PCs past the point, where the curve flattens, do not add much additional information to the regression model.

PCR utilises the underlying assumption, that those directions in the data matrix $\boldsymbol{X}$, that show the highest variance, are also those, that are connected to the response. This, of course may not be true. Nevertheless, the assumption is reasonable in many cases and often leads to good results.

3.4 Data Collection and Processing

In the following sections the structure and acquisition process of the data, on which this thesis is based, is explained. First, a detailed description of the pilot trial and its partici-

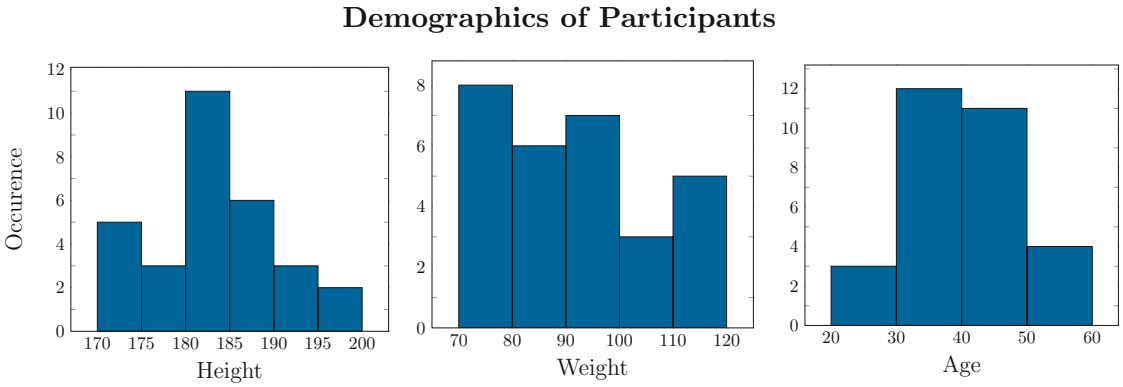

Figure 3.7: Weight, height and age histograms of the 30 participants of the pilot trial.

pants is given. This is followed by an overview of the device used to record data and the subsequent data processing.

3.4.1 Experiment Setup

We received data from a pilot trial within the PANACEA Project, which took place at the VTI in Linköping, Sweden. Before going into detail on driving tasks and data collection in this trial, an overview of participant demographics is given.

Participant Demographics

In total, 32 professional drivers, who are free of motion sickness and sleep disorders, were recruited. In addition, no participants who work nights were accepted. Drivers were asked to complete driving simulation tasks as well as taking ECG and PPG measurements. Two of the participants dropped out before the trial began which brings the number of drivers to 30. The drivers, whose measurements were removed, are also omitted when regarding the demographics of the group.

All recruited participants are male professional drivers between the ages of 25 and 60 of varying stature. Since homogenous groups in trials are preferable, and professional drivers are predominantly male, female drivers were not included. As seen in Figure 3.7, most participants were aged between 30 and 50. Only 7 of the 30 drivers were over 50 or under 30. The median age of drivers was 40 years with an interquartile range of 12 years. The participants had a median height of 183 cm, which is just slightly above the Swedish male average height [31]. The interquartile range in height amounts to 9 cm. The weight of drivers in median was 91 kg with an interquartile range of 21 kg. Although there were less participants at the top end of the scale, their weights were more evenly distributed across the range. A full list of the recorded age, weight and height of all drivers at the time of the trial can be found in Table 3.3. Figure 3.7 shows histograms of the same data.

Table 3.3: Demographic data of all trial participants, excluding the drivers whose data was excluded from the beginning.

Participant	Gender	Height (cm)	Weight (kg)	Age (yrs)
1	Male	178	115	45
2	Male	186	93	29
3	Male	183	75	36
4	Male	180	76	38
5	Male	183	100	39
6	Male	183	110	38
7	Male	173	74	35
8	Male	188	110	59
9	Male	183	79	28
10	Male	174	81	51
12	Male	176	76	47
13	Male	188	100	37
14	Male	197	95	50
15	Male	181	80	45
16	Male	197	85	30
17	Male	190	105	39
18	Male	182	74	37
19	Male	172	96	34
20	Male	190	120	45
21	Male	183	93	46
22	Male	181	78	41
23	Male	185	82	31
24	Male	171	93	49
25	Male	194	91	47
26	Male	188	87	51
27	Male	188	115	26
28	Male	181	91	30
29	Male	183	115	48
30	Male	179	88	47
32	Male	172	78	49

Figure 3.8: Left: One of the two identical driving simulators used in the pilot trial. **Right:** Two examples of the simulated driving environment in rural (top) and urban (bottom) surroundings.

Tasks and Measurement

Each driver was asked to complete six identical simulated driving exercises under varying conditions. Two driving simulators of the same kind were used. Each of them was made up of a car seat, all necessary pedals, a steering wheel and three screens showing the simulated driving environment. The simulation lasted approximately 35 minutes and took the participants through both urban and rural areas.

The drivers completed the six drives on three different days, on each of which two drives were completed back to back. The participants were under no known influence for one of these sets of two simulations. This is referred to as the baseline measurement and is labelled condition C. Another set of drives was completed under the influence of alcohol with a blood alcohol content (BAC) between 0.3‰ and 0.7‰, which is referred to as condition A. Lastly, one set of simulated driving exercises was conducted the day after consuming alcohol. In this case, which will be named condition B, blood alcohol content was not measured. Although the three sets of simulations were not necessarily conducted in this order for all participants, the driving tasks in condition B were always performed the day after those of condition A. The simulations in conditions B and C were conducted in the first half of the day, between 7 a.m. and 1 p.m., while those in condition A took place between 3 p.m. and 9.30 p.m. The basic parameters for each condition are also summarised in table 3.4.

ECG and PPG signals were recorded in the minutes before and after the driving simulation using AIT's smartPWA device, which will be explained in the next section. The measurement requires the use of both hands, therefore its execution was not possible during the simulated drive. Additionally, at the times of measurement, participants were asked to give a subjective rating of their sleepiness on the KSS.

Table 3.4: Summary of basic parameters for the different conditions in which each participant completed the simulated driving exercises.

Condition	Purpose	Influence	No. of Drives	Time of Day
A	Alcohol	BAC 0.3‰- 0.7‰	2	3 p.m. - 9.30 p.m
B	Day After	Residual Alcohol	2	7 a.m. - 1 p.m
C	Baseline	None	2	7 a.m. - 1 p.m

Figure 3.9: The smartPWA as seen during measurement. It is held with both thumbs on the conductive surfaces and the right index finger on the PPG sensor. The signals are immediately processed and visualised on the connected tablet.

3.4.2 Recording Device

All measurements were conducted using the smartPWA (smart pulse wave analysis) device depicted in figure 3.9. This is the AIT's own apparatus, originally developed and built on their premises in Vienna and Wiener Neustadt. Although its original intended purpose is its use in a breathing relaxation exercise, as described in [5], it also constitutes a simple and fast method of obtaining ECG and PPG measurements.

For acquiring these signals, the device must be held with both hands. The surface of the smartPWA has three separate conductive areas, on which the left index finger and both thumbs should be placed. Using these, the device measures a standard Einthoven lead I ECG. The right index finger is placed on an optical sensor, through which a PPG, showing blood volume changes in the finger, can be acquired. [29]

Using Bluetooth low energy, the measured signal is continuously sent to a smartphone or tablet application, where a live display of incoming data is shown [29]. Moreover, the app records the received data and can share it via email to make it available for further analysis [5]. In accordance with the standards of measurement for heart rate variability [28], at least two minutes of data was recorded for each measurement in this trial.

Table 3.5: Overview and explanation of metadata with corresponding given variable names in MATLAB.

Variable	Details
`filename`	Name of the file containing the measured smartPWA data saved as string
`participant`	Participant numbers as listed in table 3.3
`beforeAfter`	Variable showing when the data was taken: `beforeAfter = 1` indicates the data was measured before driving `beforeAfter = 2` indicates the data was measured after driving
`condition`	categorical variable containing A, B or C, corresponding to the condition
`drive`	Variable indicating the simulation in which the measurement was taken: `drive = 1` corresponds to data from the first simulation of the day `drive = 2` corresponds to data from the second simulation of the day
`weight`	Weight of the participant of the corresponding measurement
`height`	Height of the participant of the corresponding measurement
`age`	Age of the participant of the corresponding measurement

3.4.3 Data Structure and Processing

While all metadata as well as KSS values were collected and supplied by the VTI, the physiological data used in regression was obtained by processing the ECG and PPG signals acquired through the smartPWA device.

The provided metadata includes time and condition of measurement, as described in table 3.4. Participant demographics and filenames linking the smartPWA data to the measurements was also handed over. Additionally, the metadata contains a variable called 'drive', taking the value 1 for the first, and 2 for the second drive of a given participant in a certain condition. All metadata was shared in an Excel file and transferred to a table in MATLAB. An overview of all variables relating to metadata can be found in table 3.5.

KSS values were made available in a MATLAB structure array, which also includes corresponding participant numbers and information on the condition and time of measurement, such as the variable 'drive' and whether the KSS value was acquired before or after driving. This information was used to link the KSS results to the metadata and data filenames retrieved from Excel.

The ECG and PPG data was received in text files. Existing code is able to read these files and extract all relevant HRV and pulse wave data. Since the calculation of characteristic numbers from the measured signals is not subject of this thesis, only a rough outline of their computation will be given. For definition and physical interpretation of the HRV and pulse wave parameters see sections 2.2.4 and 2.3.3 in the previous chapter. In concordance with the guidelines for the measurement of HRV (Task Force of the European Society of Cardiology and North American Society of Pacing and Electrophysiology, [28]), all data relating to HRV is calculated using 2 minutes of recorded data. Since signal quality at the beginning of the recording is often lower due to motion artefacts and familiarisation with the device, only the last 2 minutes of each measurement are used.

Since the R-peaks of an ECG correspond to the curve's maxima, their extraction is relatively simple. Remember, the time difference between two R-peaks, also called the RR-Interval, is considered to indicate the heart rate. Ectopic beats and outliers are removed to obtain the normal-to-normal intervals (NNIs). To avoid an effect of short-term variations,

Table 3.6: Overview and explanation of all HRV metrics derived from ECG data, paired with their corresponding given variable names in MATLAB.

Variable	Details
`hr`	heart rate
`rmssd`	root mean square of successive differences of NN intervals
`sdnn`	standard deviation of NN intervals
`pnn50`	percentage of successive NN intervals that differ more than 50 ms
`lfp`	low frequency power
`hfp`	high frequency power
`totalpwr`	total power
`lhratio`	ratio of low and high frequency $\frac{\text{lfp}}{\text{hfp}}$
`lfp_n`	normalised low frequency power $\frac{\text{lfp}}{\text{lfp + hfp}}$
`hfp_n`	normalised high frequency power $\frac{\text{hfp}}{\text{lfp + hfp}}$

the heart rate given by the NNIs is averaged over the whole 2-minute interval. All other HRV parameters in the time domain, such as SDNN, RMSSD and pNN50, can be directly calculated from the series of NN-Intervals over the two minute window. [5]

For HRV measures in the frequency domain, the series of successive NNIs is transformed to the frequency domain using the Lomb-Scargle-periodogram [5]. This algorithm is not subject of this thesis, but more details can be found in [17, 51]. From this, the power in the low frequency band ($0.04 - 0.15$ Hz), the power in the high frequency band ($0.15 - 0.04$ Hz) as well as the total power can be derived. Subsequently, the ratio of low and high frequency powers along with normalised high and low frequency powers can be directly computed as defined in table 2.2. An overview of all variables connecting to HRV measures used is shown in table 3.6.

The PAT is measured as the time difference between an R-Peak, as the source of a pulse wave, and its arrival in the index finger. The onset of the pulse wave is considered to be the point of intersection between the tangent of the rising slope of the wave and a constant baseline, given by the minimum of the signal. Therefore, PAT is given as the time elapsed from any R-Peak in the ECG to the subsequent pulse wave onset in the PPG. For the same reasons as above, PAT is averaged over the last two minutes of each recording. Figure 3.10 visualises the calculation. Values for PAT and HR were removed, if either the heart rate was above 200 bpm or below 30 bpm, or the time intervals between consecutive R-Peaks changed by more than 0.2 seconds. [5]

Finally, to calculate the rest of the pulse wave parameters, the total duration of the wave as well as some critical points are extracted. The first peak in the wave indicates the systolic pressure, whereas the dicrotic wave pressure is given by the second peak. There is a minimum between these two peaks, which corresponds to the dicrotic notch. Details on pulse wave parameters and the algorithm used for extracting dicrotic wave pressure and dicrotic notch are presented in [12]. More detailed descriptions of pulse wave parameters were given in section 2.3.3. Some additional variables were derived from the previously defined pulse wave parameters. An overview of all pulse wave related variables used is given in table 3.7.

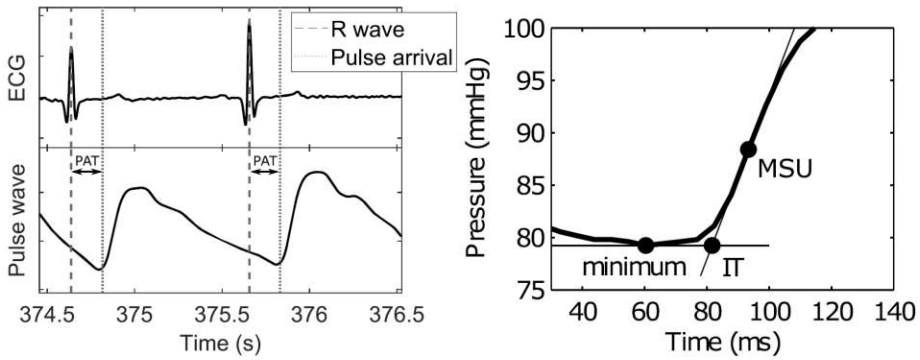

Figure 3.10: In the left image, the time difference corresponding to the PAT is marked in simultaneous ECG and PPG measurements. A visualisation of the calcualtion of the pulse wave onset is shown on the right. The onset of the wave is given by the point of intersecting tangents (IT), passing through the minimum and the maximum systolic upstroke (MSU). Image taken from [5].

Table 3.7: Overview and explanation of all variables derived from measurement of the pulse wave, matched with corresponding given variable names in MATLAB.

Variable	**Details**
`t_T`	total pulse duration (TPD), i.e. time from wave onset to its end
`t_sys`	time from wave onset to systolic pressure P_{sys} (1^{st} peak)
`t_sys_rel`	time from wave onset to systolic pressure P_{sys} in relation to TPD
`t_notch`	time from wave onset to dicrotic notch P_{notch} (minimum between 1^{st} and 2^{nd} peak)
`t_notch_rel`	time from wave onset to dicrotic notch P_{notch} in relation to TPD
`t_dwp_rel`	time from wave onset to dicrotic wave pressure P_{dwa} (2^{nd} peak) in relation to TPD
`P_dwp_sys`	amplitude of dicrotic wave pressure divided by amplitude of systolic pressure: $\frac{P_{\text{dwa}}}{P_{\text{sys}}}$
`P_notch_sys`	amplitude of dicrotic notch divided by amplitude of systolic pressure: $\frac{P_{\text{notch}}}{P_{\text{sys}}}$
`P_notch_dwp`	amplitude of dicrotic notch divided by amplitude of dicrotic wave pressure: $\frac{P_{\text{notch}}}{P_{\text{dwa}}}$
`pat`	pulse arrival time

4 Approaches to Deriving a Predictive Model

This chapter will guide through the steps of deriving a predictive model from the provided data. A good starting point is a univariate analysis of the data, that is, the analysis of key features in dependence of one single variable. Next, the training data set, from which the model is derived, needs to be defined. Multiple different choices of training data sets and approaches to deriving a predictive model are presented in this chapter.

4.1 Data Analysis

First of all, it is important to keep in mind, that the comparability of the available data is limited. There are differences of unknown extent between measurements of individuals. Additionally, as stated in 3.4, some measurements were taken under the influence of alcohol, which has been shown to have an impact on HRV [44]. However, the exact physiological effects of alcohol consumption in healthy individuals, as well as differences between individual reactions, are unknown.

Even though some measurements could not be conducted or had to be aborted, for the majority of participants, all twelve planned measurements were recorded. Also, in some available measurements, the processing algorithm failed to extract certain values from the measured signals due to quality issues. This is especially profound for the pulse wave shape parameters. The amount of missing values for each variable in available measurements are detailed in table 4.1. From a total of 346 measurements, 77 have at least one missing value while 269 are complete. All together, there are considerably less measurements of satisfying quality, than expected. Especially considering the number of variables to be examined in the model, the lack of high quality data poses challenges.

A closer look at the measured KSS values, reveals that most measurements are around the centre of the scale. The distribution of KSS values is shown in figure 4.1. While 81.5% of all values are between 3 and 6, only 10.98% are lower than this. Measurements with KSS values at the top end of the scale are rare. Only 0.03% of all KSS values are 7 or higher, while 9, the highest value on the scale, was never recorded.

Table 4.1: Missing Values by Variable

Variable	Missing	Variable	Missing	Variable	Missing
hr	1	totalpwr	1	t_dwp_rel	62
lfp	2	lfp_n	2	t_sys	62
hfp	2	hfp_n	2	P_dwp_sys	62
lhratio	2	t_T	62	P_notch_sys	62
rmssd	1	t_notch	62	P_notch_dwp	62
sdnn	1	t_sys_rel	62	pat	7
pnn50	0	t_notch_rel	62	kss	15

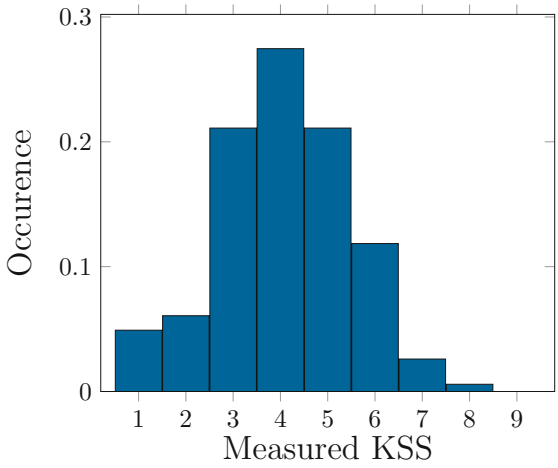

Figure 4.1: Histogram of KSS values given throughout the trial.

As stated in 3.4, some measurements were conducted before driving, while others were taken afterwards. Since driving in itself can be a stressful task and can therefore have an effect on the ANS and HRV, it is worth considering whether there are differences in measurements depending on the time they were obtained. The Wilcoxon signed rank test can be used to conduct such a comparison for each variable separately. Since this amounts to a total of 21 hypothesis tests, a correction for multiple testing is necessary. The false discovery rate was controlled at the level $\alpha = 0.05$. To assure the comparison of only corresponding 'before' and 'after' measurements, the test is conducted using 324 of the 346 recordings.

Restricting the comparison to condition C could avoid unwanted effects due to alcohol. However, this would not leave enough measurements to obtain meaningful results. Therefore, the comparisons are conducted with all available data of all conditions. The results of the comparison are detailed in table 4.2. The null hypothesis, that the difference of compared values comes from a distribution with mean zero is rejected for $\frac{\text{LF}}{\text{HF}}$ ratio, LFnorm, HFnorm, KSS, LF, TP, SDNN, and HR. Therefore, the available data suggests, that the driving task has an impact on the measured values of these variables. For all other variables, the Wilcoxon signed rank test fails to reject the null hypothesis. Nevertheless, one can not conclude, that driving has no effect on their values.

While these results suggest that it may not be wise to mix data from before and after driving, it also can, heuristically, give insight on which variables might be of greater importance with respect to fatigue. The comparison seems to show, that participants felt more fatigued after the driving task. Therefore, other variables for which the null hypothesis was rejected may warrant special attention when deriving a model. Of course, the detected effect could also be due to impacts of driving other than rising fatigue. However, correlations between KSS and other variables, as seen in table 4.3, show similar relationships, at least for $\frac{\text{LF}}{\text{HF}}$ ratio, LFnorm, HFnorm, LF and HR.

While a rise of the $\frac{\text{LF}}{\text{HF}}$ ratio has only been detected in few studies and changes in normalised LF and HF as well as crest time t_{sys} seem to be contradictory to previous results

Table 4.2: Results of the comparison of data recorded before and after driving using the Wilcoxon signed rank test. The result 1 indicates a rejection of the null hypothesis, that the difference between the two times of measurement comes from a distribution with the mean zero. The result 0 indicates failure to reject the null hypothesis. Remember, the null hypothesis is rejected, if the p-value is below the modified significance level, both of which are given in the last columns of the table.

Variable	Result	p-Value	Modified Significance Level (FDR)
$\frac{\text{LF}}{\text{HF}}$ ratio	1	< 0.0001	0.0024
LFnorm	1	< 0.0001	0.0048
HFnorm	1	< 0.0001	0.0071
KSS	1	< 0.0001	0.0095
LF	1	0.0009	0.0119
TP	1	0.0017	0.0143
SDNN	1	0.0045	0.0167
HR	1	0.0054	0.0190
PAT	0	0.0237	0.0214
t_T	0	0.0404	0.0238
$P_{\text{notch}_{\text{dwp}}}$	0	0.1794	0.0262
$P_{\text{notch}_{\text{sys}}}$	0	0.1824	0.0286
t_{notch}	0	0.2151	0.0310
$t_{\text{sys}_{\text{rel}}}$	0	0.5949	0.0333
$P_{\text{dwp}_{\text{sys}}}$	0	0.6936	0.0357
HF	0	0.7262	0.0381
RMSSD	0	0.7666	0.0405
$t_{\text{dwp}_{\text{rel}}}$	0	0.8557	0.0429
$t_{\text{notch}_{\text{rel}}}$	0	0.8709	0.0452
pnn50	0	0.8830	0.0476
t_{sys}	0	0.9201	0.0500

collected in table 2.4, the changes seen in LF, TP and SDNN are in line with the existing research outlined in table 2.4.

4.2 Allocation of Test and Training Sets

This section describes three different ways of dividing the available data into training- and test data sets. Remember, the data can be structured by the condition of measurement, A, B or C, and the time of measurement, before or after driving.

Since the magnitude of the effect of alcohol on HRV is unclear, both immediately after and a day after drinking, it may be advisable to train the model only on data from condition C, where there is no known influence. Also, for sufficiently accurate results, the potential differences between individuals should be taken into account.

It can be difficult, or even impossible, to exclude all unwanted effects, due to alcohol, time of measurement or individual differences, from the training data set. Being too strict

Table 4.3: Correlations between KSS and other variables, computed using the Spearman's correlation coefficient ρ, and corresponding p-values are listed. The result 1 indicates a significant correlation where the FDR was controlled at the level $\alpha = 0.05$. The result 0 indicates no significant correlation. Remember, a result is deemed significant, if the p-value is below the modified significance level, both of which are given in the last columns of the table.

Variable	ρ	Result	p-Value	Modified Significance Level (FDR)
$\frac{\text{LF}}{\text{HF}}$ ratio	0.2123	1	0.0001	0.0025
LFnorm	0.2123	1	0.0001	0.0050
HFnorm	−0.2123	1	0.0001	0.0075
HR	−0.1800	1	0.0010	0.0100
LF	0.1640	1	0.0028	0.0125
$t_{\text{sys}_{\text{rel}}}$	−0.1319	0	0.0306	0.0150
$P_{\text{notch}_{\text{sys}}}$	−0.1269	0	0.0372	0.0175
TP	0.1138	0	0.0388	0.0200
SDNN	0.1076	0	0.0508	0.0225
$P_{\text{notch}_{\text{dwp}}}$	−0.1093	0	0.0736	0.0250
t_{sys}	−0.1023	0	0.0941	0.0275
pnn50	0.0648	0	0.2396	0.0300
HF	0.0592	0	0.2844	0.0325
t_T	0.0629	0	0.3043	0.0350
RMSSD	0.0517	0	0.3492	0.0375
$P_{\text{dwp}_{\text{sys}}}$	−0.0425	0	0.4879	0.0400
$t_{\text{notch}_{\text{rel}}}$	−0.0399	0	0.5148	0.0425
PAT	0.0158	0	0.7764	0.0450
t_{notch}	0.0108	0	0.8607	0.0475
$t_{\text{dwp}_{\text{rel}}}$	0.0087	0	0.8866	0.0500

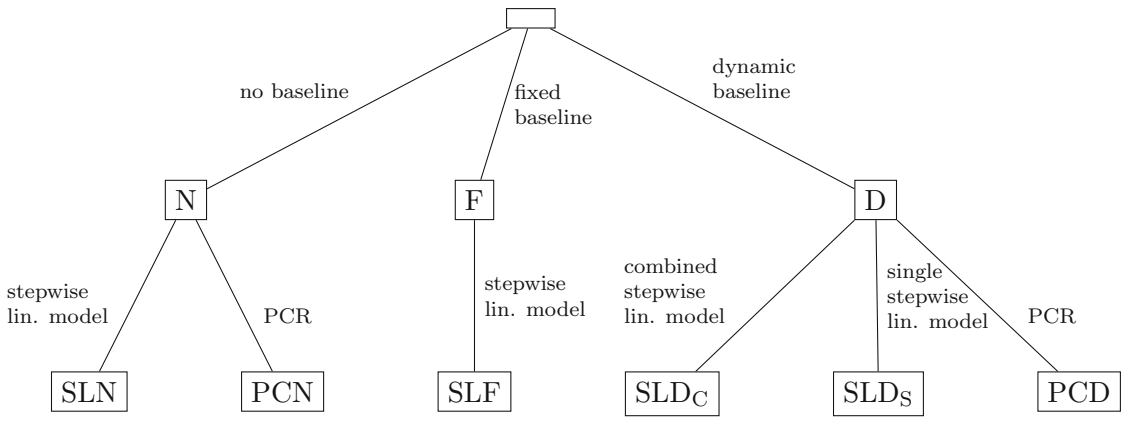

Figure 4.2: The tree represents an overview of different choices of training data sets and modelling approaches for generating a model from each set. The options are explained in detail in this chapter.

on training data requirements could result in a miniscule data set, which is not suitable for training a model. The following three approaches attempt to strike a balance between excluding unwanted effects from training data and retaining enough measurements in the training set to sensibly derive a model.

Figure 4.2 outlines the options for training data sets and models, that can be derived from them. The structural overview shows all different combinations of training data sets and modelling approaches examined in this thesis, which will be explained in detail throughout this chapter. The abbreviations used in figure 4.2 will be used throughout the rest of this thesis.

4.2.1 No Participant Baseline (N)

In a first attempt, all measurements recorded prior to the driving simulation in the baseline condition C, where there were no known influences, are used as training data. This excludes many measurements, and therefore also unknown effects due to driving and alcohol, but does not account for individual differences in HRV. This choice of training data, with no so-called baseline measurement for individual participants, will be abbreviated as N.

After subjecting the data to these training set requirements, only 56 out of the 346 measurements remain. This still includes 16 recordings with missing values, 13 of which are in pulse wave data and 3 of which correspond to missing KSS values. Therefore, the data set that can actually be used for training is comprised of only 40 measurements.

Since this is already a very small data set, none of it was reserved as test data. The models are tested on data from both remaining conditions, A and B, irrespective of the time of recording, as well as data measured in condition C after driving. Due to issues of comparability, model results on each of these test sets will be reviewed separately.

4.2.2 Dynamic Participant Baseline (D)

Similar to the approach in 4.2.1, only data in baseline condition C is included in the training data set. However, in this case, measurements recorded both before and after driving are used. In order to account for individual differences when training a model, the measurement before driving acts as a baseline in values for each participant. The training data then corresponds to the differences from this individual participant baseline, rather than the measured HRV and pulse wave parameters at a given time.

The chosen baseline, the measurement recorded directly before driving, is dynamic in the sense, that for each driving simulation a new participant baseline is set. The training data set is comprised of the difference in measured values before and after completing the driving simulation. This choice of training data will be abbreviated as D.

Of course, this data can only be derived if both measurements, before and after driving, are recorded in adequate quality. From the 346 measurements, 10 need to be excluded because there is no corresponding recording before or after driving. After dividing the data by time of recording and computing the differences, this is reduced to 168 samples, 56 of which are in baseline condition C. This, however, still includes 20 recordings with missing values, 4 of which correspond to missing KSS values and all of which include missing pulse wave data.

As in 4.2.1, the possible training data set in this case is too small to reserve some of it for testing. The models are tested on the before-after difference data of the conditions A and B and model results will be examined on both sets separately.

4.2.3 Fixed Participant Baseline (F)

Lastly, an approach with one fixed participant baseline for all measurements is used. For each individual, the first measurement taken in baseline condition C, before driving, is used as the baseline. Optionally, when two measurements in condition C, recorded before driving, are available, the average of both measurements can be used as the fixed baseline. Then the training data consists of the differences between any other measurement and the allocated participant baseline. This approach will be abbreviated as F.

In this case, due to lack of data, it is unfeasible to use only recordings from baseline condition C to train a model. Instead, for each participant, the two measurements for a given recording time (before or after driving) and a given condition (A,B or C) are randomly divided between the test and training data set. If there is only one measurement, e.g. if the corresponding recording is missing, it is simply randomly allocated to either the test or training set.

With this approach, the training data set may not exclude many of the unwanted effects that could arise in the data. However, setting the first recording as individual baseline has the advantage, that only a single baseline measurement is needed for multiple evaluations, while also resulting in a larger training data set. The training data set consists of 141 recordings, 85 of which have no missing values. There are no missing HRV values, 45 missing pulse wave parameters and 11 KSS values, that were not recorded.

If there are multiple measurements taken in condition C before driving, averaging them for a baseline, while ignoring missing values, leads to an even larger training data set. Only those variables, that are missing their value in both averaged measurements, will result in a missing value in the baseline. This leads to less problems evaluating the difference to the baseline and therefore less missing values. In this case, 103 of the 141 measurements are free of missing values. The distribution of missing values is as follows: 6 KSS values and 32 pulse wave parameters.

4.3 Multivariate Stepwise Linear Models (SL)

One of the main approaches is creating a multivariate stepwise linear regression model from the chosen training data set. Matlab provides a function, `stepwiselm`, for exactly this purpose [50].

Data can be handed to `stepwiselm` as a data matrix and a response vector or in a table, where the last column holds the response data. The function automatically takes care of removing missing values and performs forward and backward selection in a hybrid approach. In the optional arguments of `stepwiselm`, one of five criteria (residual sum of squares (RSS), R^2, adjusted R^2, AIC and BIC) can be set for the selection process. Since there are not too many criteria, all of them were tested when generating stepwise models. The default starting model is the constant model while the default upper limit is a model

including all variable interactions. Upper and lower limits for the generated model can also be adapted through optional arguments. [50]

Even though the Karolinska sleepiness scale measures fatigue on an ordinal scale, rather than continuously, the generated regression models still give real-valued outputs. These are mapped to a KSS prediction by the rounding to the closest integer. Values larger than 9 are mapped to the top end of the scale, while values smaller than 1 are mapped to the lowest KSS value.

All generated models are linear in all variables. Fitting polynomials of higher order simply leads to overfitting. This could suggest, that the underlying relationship of the data is, in fact, linear. However, overfitting using higher order polynomials could also be a result of a small training data set. In either case, it is unwise to attempt fitting higher order polynomials. Hence, this thesis focuses on the linear approach.

4.3.1 Model without Participant Baseline (SLN)

Two models using no participant baseline were generated. As aforementioned, the training set consists of all measurements in condition C recorded prior to driving. The simplest approach, of using stepwise multiple regression on the entire training set, did not yield satisfying results. It mostly lead to constant models with a large prediction error. Since the training set is rather small, especially due to missing values in the pulse wave parameters, the HRV and pulse wave data were modelled separately as an attempt to increase the sample size. The size of the training set for the pulse wave model remains unchanged, but the HRV model can be generated using 53 measurements, instead of 40. When predicting a KSS value from HRV and pulse wave data, both models are evaluated and the mean of both KSS values is output as prediction.

For the first model, which will be called SLN_1, the HRV model was generated by `stepwiselm`, using the constant model as a starting model. A linear relationship between data and response was used as an upper limit. Stepwise selection was performed with the criterion AIC, where variables were removed from the model, if this resulted in a change in AIC greater than 2. The threshold for a variable to be included in the model, was a change of less than 1, i.e. when the addition of a variable results in a decrease of AIC or an increase by less than 1. The model generated from pulse wave data only differs in the thresholds used. For the pulse wave model, a variable is included, if the change in AIC is less than 1.7, while variables resulting in a change larger than 1.8 were removed. The chosen thresholds for AIC are above the standard values for this criterion. Using a constant starting model, the requirements for variables to be added need to be relaxed for any variables to be chosen. The exact values of the thresholds were determined empirically.

A second model, SLN_2, was generated using the same principle. The results of the two models, using HRV and pulse wave data separately, are combined for a final prediction. The two single models are both created using stepwise variable selection with a linear starting model, while the variables are also at most in a linear relationship to the response, i.e. no quadratic terms or interactions. The R^2 statistic was used as a criterion with the default thresholds. This means a variable is included in the model, if it results in an increase of R^2 of at least 0.1. If a variable only leads to an increase of less than 0.05, the variable is removed. These are standard threshold values for the R^2 criterion that work well with a

Table 4.4: Parameters of models created without participant baseline, as well as the starting model, upper limit and variables included by stepwise selection. The results are obtained by a combination of two linear models, generated using either HRV or pulse wave data, as indicated in the column 'Data'. Diff_{AIC} denotes the difference in AIC resulting from adding or removing a variable and Diff_{R^2} denotes the same for R^2.

Model	Data	Method	Model Limits	Thresholds	Selected Variables
SLN_1	HRV	stepwise regression	Start: constant Upper: linear	Enter: $\text{Diff}_{\text{AIC}} < 1$ Remove: $\text{Diff}_{\text{AIC}} > 2$	`age`, `hr`, `lfp`, `lhratio`, `sdnn`, `pnn50`, `totalpwr`
	PW	stepwise regression	Start: constant Upper: linear	Enter: $\text{Diff}_{\text{AIC}} < 1.7$ Remove: $\text{Diff}_{\text{AIC}} > 1.8$	`height`, `age`, `t_T`, `t_sys_rel`, `t_sys`, `P_dwp_sys`, `P_notch_sys`, `P_notch_dwp`, `pat`
SLN_2	HRV	stepwise regression	Start: linear Upper: linear	Enter: $\text{Diff}_{R^2} > 0.1$ Remove: $\text{Diff}_{R^2} < 0.05$	`age`, `hr`, `lhratio`, `sdnn`, `totalpwr`
	PW	stepwise regression	Start: linear Upper: linear	Enter: $\text{Diff}_{R^2} > 0.1$ Remove: $\text{Diff}_{R^2} < 0.05$	`height`, `age`, `t_T`, `t_sys_rel`, `t_sys`, `P_notch_sys`

linear starting model. The main parameters of both models are summarised in table 4.4.

4.3.2 Model using Dynamic Participant Baseline (SLD)

Four different models using a dynamic participant baseline will be discussed in this thesis. As in 4.3.1, for two of these models, SLD_{C1} and SLD_{C2}, HRV and pulse wave data were modelled separately and later combined. The other two models, SLD_{S1} and SLD_{S2}, were generated by stepwise multiple regression using both data sets, HRV and pulse wave, simultaneously. The main parameters of all SLD models are summarised in table 4.5.

SLD_{C1} combines two multivariate stepwise linear models with a small number of predictors. The algorithm was given the constant model as a starting model and the upper limit was defined as linear. When selecting the pulse wave variables, a parameter was included, if the AIC decreased or increased by less than 0.8. Variables were removed if their addition caused an increase of more than 0.81. The AIC thresholds are above the standard values, since, for a constant starting model, standard thresholds do not allow the addition of any variables. The exact values of the thresholds were determined empirically. For the HRV data, residual sum of squares (RSS), also known as SSE (sum of squared errors), as it is termed in MATLAB, was used as a threshold criterion with the default thresholds. It is evaluated by the F-test with the test statistic (3.21), determining the change in RSS in two models of different size. If a variable was significant at the level $p < 0.05$, it was included in the model. A p-value larger than 0.1 leads to removing the variable in question. These thresholds, as well as all other thresholds for SLD models, are the standard values for the corresponding criteria.

When modelling pulse wave and HRV data simultaneously with the same parameters as used for the HRV component of SLD_{C1}, only hrv data is included in the model. No variables concerning the pulse wave are selected. This model, generated by a single multiple regression, only contains one variable, `lhratio`, and will be called SLD_{S2}.

For the second combined model SLD_{C2}, generated from two multiple regression models, both the starting model and the upper limit were set as linear. Pulse wave and HRV data

Table 4.5: Parameters of models created with a dynamic participant baseline, as well as the starting model, upper limit and variables included by stepwise selection. For the models with the index C, the results are obtained by a combination of two linear models, generated using either HRV or pulse wave data, as indicated in the column 'Data'. The models using the index S were generated as a single model from HRV and pulse wave data. Diff_{AIC} denotes the difference in AIC resulting from adding or removing a variable. Diff_{BIC} and Diff_{R^2} denote the same for BIC and R^2, respectively. $\text{Diff}_{p_{\text{SSE}}}$ is the p-value for the F-test of the change in SSE resulting from addition or removal of a variable.

Model	Data	Method	Model Limits	Thresholds	Selected Variables
SLD_{C1}	HRV	stepwise regression	Start: constant Upper: linear	Enter: $\text{Diff}_{p_{\text{SSE}}} < 0.05$ Remove: $\text{Diff}_{p_{\text{SSE}}} > 0.1$	`lhratio`
	PW	stepwise regression	Start: constant Upper: linear	Enter: $\text{Diff}_{\text{AIC}} < 0.8$ Remove: $\text{Diff}_{\text{AIC}} > 0.81$	`t_dwp_rel`, `t_sys`, `P_dwp_sys`, `pat`
SLD_{C2}	HRV	stepwise regression	Start: linear Upper: linear	Enter: $\text{Diff}_{R^2} > 0.1$ Remove: $\text{Diff}_{R^2} < 0.05$	`hr`, `lfp`, `hfp`, `lhratio`, `pnn50`, `hfp_n`
	PW	stepwise regression	Start: linear Upper: linear	Enter: $\text{Diff}_{R^2} > 0.1$ Remove: $\text{Diff}_{R^2} < 0.05$	`t_notch`, `P_dwp_sys`, `t_sys`, `P_notch_sys`, `P_notch_dwp`, `t_sys_rel`,
SLD_{S1}	all data	stepwise regression	Start: constant Upper: linear	Enter: $\text{Diff}_{p_{\text{SSE}}} < 0.05$ Remove: $\text{Diff}_{p_{\text{SSE}}} > 0.1$	`lhratio`
SLD_{S2}	all data	stepwise regression	Start: linear Upper: linear	Enter: $\text{Diff}_{\text{BIC}} < 0$ Remove: $\text{Diff}_{\text{BIC}} > 0.01$	`weight`, `height`, `age`, `hr`, `lfp`, `hfp`, `lhratio`, `pnn50`, `t_notch_rel`, `t_sys`, `P_dwp_sys`

were modelled separately, both using R^2 as a criterion with the standard thresholds, i.e. variables were included in the model, if the change in R^2 was larger than 0.1 and were removed if they lead to a change smaller than 0.05.

Finally, SLD_{S1} represents a larger model in comparison to SLD_{S2}. It is also created by stepwise multiple regression using pulse wave and HRV simultaneously, but with a linear starting model. The upper limit is also set as linear and BIC with default thresholds is used as a criterion. A variable is included in the model, if it leads to a decrease in BIC. If the increase in BIC, caused by the addition of a variable, is larger than 0.01, the parameter is removed from the model.

4.3.3 Model using a Fixed Participant Baseline (SLF)

Two models, SLF_1 and SLF_2, were created using a fixed participant baseline, as described in section 4.2.3. Both models were generated using stepwise multiple regression on HRV and pulse wave data simultaneously. Creating and combining two separate models, as done previously, did not lead to satisfying results. For both models, a constant model was used as a starting model and a linear model as an upper limit. The thresholds used are the standard thresholds for the respective selection criteria. The model parameters of SLF_1 and SLF_2 are summarised in table 4.6.

For the model SLF_1, the baseline was set to be the very first measurement, irrespective of the number of missing values. The thresholds were defined using AIC with the default values. This means a variable is added to the model, if it results in a decrease in AIC,

Table 4.6: Parameters of models created with a fixed participant baseline, as well as the starting model, upper limit and variables included by stepwise selection. The column ´BL' indicates whether the first measurement or the average of both initial measurements was used as the baseline. All models are generated using HRV and pulse wave data. Diff_{AIC} denotes the difference in AIC resulting from adding or removing a variable. $\text{Diff}_{\text{Adj}R^2}$ is the change in adjusted R^2 due to variable addition or removal.

Model	BL	Method	Model Limits	Thresholds	Selected Variables
SLF_1	first	stepwise regression	Start: constant Upper: linear	Enter: $\text{Diff}_{\text{AIC}} < 0$ Remove: $\text{Diff}_{\text{AIC}} > 0.01$	`age`, `lfp_n`, `t_sys_rel`, `t_notch_rel`, `P_dwp_sys`, `P_notch_sys`, `P_notch_dwp`
SLF_2	average	stepwise regression	Start: constant Upper: linear	Enter: $\text{Diff}_{\text{Adj}R^2} > 0$ Remove: $\text{Diff}_{\text{Adj}R^2} < -0.05$	`height`, `age`, `hr`, `lfp`, `lhratio`, `totalpwr`, `lfp_n`, `t_T`

whereas an increase of more than 0.01 leads to its removal.

The model SLF_2 uses the average of both initial measurements as a baseline. The adjusted R^2 value determines, whether a variable is added or removed. Since adjusted R^2 already penalises for adding variables, any variable causing an increase in adjusted R^2 is included. A decrease in adjusted R^2 of more than 0.05 indicates, that the corresponding variable should be removed from the model.

4.4 Principal Component Regression (PC)

The second main approach in this thesis is principal component regression, that is, multivariate regression with the computed principal components (PCs) instead of the measured variables. MATLAB provides a function for the principal component analysis of raw data, `pca`. For a given array containing data, the function computes the corresponding PC vectors, as well as their coefficients and the percentage of the total variance explained by each PC. [50]

The data and response are centred before computing the principal components of the data and conducting regression. The multivariate linear regression is performed using the MATLAB function `fitlm`. No stepwise procedure is used for selection, since the PCA already excludes unnecessary information. The data transformation is later reversed to obtain corresponding non-centred KSS estimates from the principal component regression results.

The method of principal component regression (PCR) was followed both with a dynamic baseline and without any participant baseline. Since these attempts did not yield more accurate predictions than the corresponding stepwise linear models, the procedure was abandoned. No further PCR models, such as models using a fixed baseline, were generated.

4.4.1 Model without Participant Baseline (PCN)

First, PCA is conducted for a model without any participant baseline. The principal components for all data recorded in condition 'C' before driving, excluding the response KSS, are computed and ordered by the size of their eigenvalues, and therefore also their

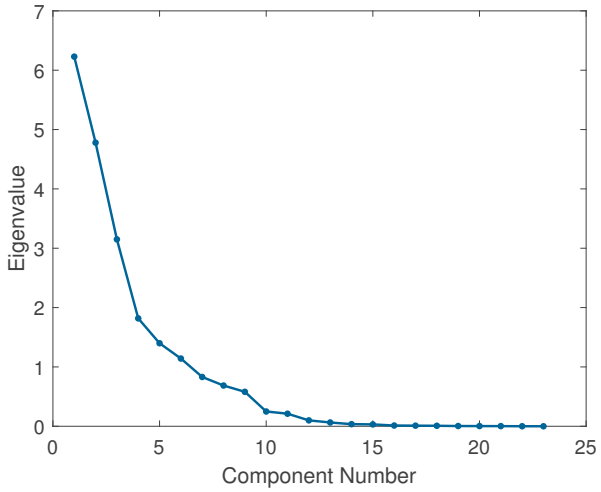

Figure 4.3: The Scree plot indicates the variance explained by each principal component computed from the data recorded in the baseline condition 'C' before driving.

explained variance, in decreasing order. Figure 4.3 shows the corresponding Scree plot, a visual aid for selecting the number of principal components to use in regression.

The choice was made to use either the first 6 or 8 principal components, since they explain 86.74% and 93.85% of the total variance, respectively. For each choice, a principal component regression (PCR) model was then generated using the function `fitlm`. From here on, the model using 6 PCs will be called PCN_1, while PCN_2 refers to the model with 8 PCs.

Table 4.7: The percentage of total variance explained by each principal component (PC) and the five largest contributors to each PC are given. The contribution of any original variable to a PC is determined by the correlation between the computed values of PCs, the scores, and the original variables. The correlation coefficient ρ is given in brackets following the corresponding variable and computed using Spearman's correlation coefficient.

PC	Variance explained	Largest Contributors
1	29.18%	`lfp` (0.94), `totalpwr` (0.83), `sdnn` (0.83), `rmssd` (0.79), `pnn50` (0.78)
2	22.39%	`t_notch` (0.90), `t_sys` (0.75), `pat` (0.73), `hr` (−0.64), `t_T` (0.62)
3	14.75%	`t_T` (−0.76), `hr` (0.75), `t_dwp_rel` (0.74), `P_dwp_sys` (−0.70), `P_notch_sys` (0.68)
4	8.52%	`height` (−0.55), `P_notch_dwp` (0.46), `hfp_n` (−0.45), `lfp_n` (0.45), `lhratio` (0.45)
5	6.55%	`hfp_n` (−0.64), `lfp_n` (0.64), `lhratio` (0.64), `P_notch_dwp` (0.43), `pat` (0.31)
6	5.35%	`height` (0.59), `P_notch_sys` (0.38), `weight` (0.36), `pnn50` (0.27), `P_dwp_sys` (0.26)
7	3.89%	`weight` (0.53), `age` (0.48), `P_dwp_sys` (−0.28), `P_notch_sys` (0.24), `height` (0.17)
8	3.22%	`weight` (0.56), `age` (−0.46), `P_notch_dwp` (0.29), `P_dwp_sys` (−0.28), `pnn50` (0.17)

Table 4.7 illustrates which variables contribute most to each principal component used. Instead of assessing the contribution of a single variable by the size of its coefficient, the

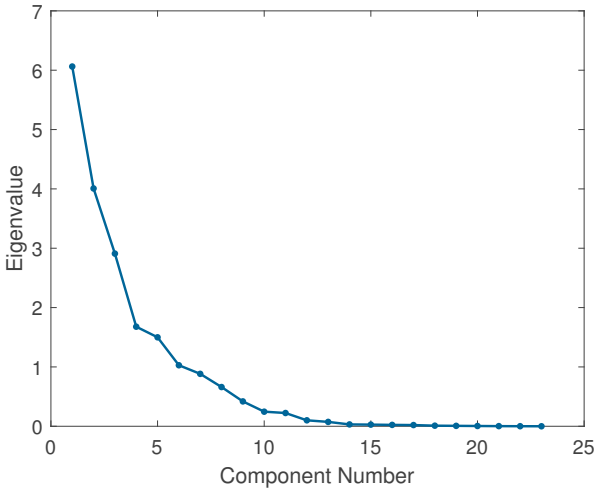

Figure 4.4: The Scree plot indicates the variance explained by each principal component computed from the differences between data recorded before and after driving in the baseline condition 'C'.

correlation between the values of each variable and each PC is computed. Those variables with the largest correlation coefficients also have the highest influence on the PC in question.

4.4.2 Model using Dynamic Participant Baseline (PCD)

Finally, two PCR models using a dynamic participant baseline are discussed. The training data set consists of the differences between recordings before and after driving, in the baseline condition 'C'. The response is separated from the rest of the data, before computing the PCs. As previously, these are ordered by the percentage of total variance they explain in decreasing order. Figure 4.4 depicts the corresponding Scree plot.

Again, models are generated using the function `fitlm` and a fixed number of PCs. In this case models using 5 PCs, explaining 81.11% of the total variance, or 6 PCs, explaining 86.27%, look most promising. The model using 5 PCs will be referred to as PCD_1, the model generated with 6 PCs as PCD_2. Table 4.8 gives insight concerning the influence of the original variables on the PCs used in modelling.

Table 4.8: The percentage of total variance explained by each principal component (PC) and the five largest contributors to each PC are given. The contribution of any original variable to a PC is determined by the correlation between the computed values of PCs, the scores, and the original variables. The correlation coefficient ρ is given in brackets following the corresponding variable and computed using Spearman's correlation coefficient.

PC	Variance explained	Largest Contributors
1	30.43%	`t_sys_rel` (−0.83), `t_notch_rel` (−0.73), `t_sys` (−0.69), `hfp_n` (−0.66), `lfp_n` (0.66)
2	20.11%	`t_dwp_rel` (0.73), `t_notch_rel` (0.68), `t_notch` (0.62), `P_dwp_sys` (−0.59), `t_sys_rel` (0.51)
3	14.61%	`t_T` (0.68), `hr` (−0.67), `rmssd` (0.58), `pnn50` (0.50), `hfp` (0.50)
4	8.42%	`age` (0.71), `lhratio` (−0.53), `hfp` (0.50), `hfp_n` (0.49), `lfp_n` (−0.49)
5	7.53%	`P_notch_dwp` (0.55), `weight` (−0.49), `hr` (−0.4322), `t_T` (0.50), `P_notch_sys` (0.41)
6	5.16%	`height` (0.60), `P_notch_dwp` (0.44), `weight` (0.32), `lfp` (0.30), `P_dwp_sys` (−0.25)

5 Results

The main results for each of the models introduced in chapter 4 are presented in this chapter. Each section is dedicated to one of the models, using the abbreviations introduced in figure 4.2. The coefficients, their significance and characteristics of quality of fit are described in detail.

The F-test is used to determine whether variable coefficient estimates are significantly different from zero. Stepwise regression models are compared to a constant model, that uses only an intercept. The PCR models do not contain an intercept, therefore $y \sim 0$, where y denotes the response, is used as a reference in these cases.

Since many models with different numbers of variables are discussed, adjusted R^2 is used as a measure of goodness of fit rather than R^2. Additionally, the root mean square error (RMSE) of the model's predictions is given. Lastly, two residual plots, one plotting residuals against corresponding fitted values, the other against the preceding residual, are provided for analysis.

5.1 Model SLN_1

The regression model SLN_1 contains 15 variables and is given by the formula

$$\mathrm{KSS}_{\mathrm{est}} = \frac{(\mathrm{KSS}_{\mathrm{HRV}} + \mathrm{KSS}_{\mathrm{PW}})}{2}$$

where the KSS values from HRV and PW data are estimated by separate models described by the formulas

$$\begin{aligned}\mathrm{KSS}_{\mathrm{HRV}} &= K_{\mathrm{HRV}} + c_1\texttt{age} + c_2\texttt{hr} + c_3\texttt{lfp} + c_4\texttt{lhratio} + c_5\texttt{sdnn} + c_6\texttt{pnn50} + c_7\texttt{totalpwr}\\ \mathrm{KSS}_{\mathrm{PW}} &= K_{\mathrm{PW}} + c_8\texttt{height} + c_9\texttt{age} + c_{10}\texttt{t\_T} + c_{11}\texttt{t\_sys\_rel} + c_{12}\texttt{t\_sys} +\\ &\quad c_{13}\texttt{P\_dwp\_sys} + c_{14}\texttt{P\_notch\_sys} + c_{15}\texttt{P\_notch\_dwp} + c_{16}\texttt{pat}.\end{aligned}$$

The computed variable coefficients c_j, $j = 1, \ldots, 16$ are given in table 5.1.

Assessing Model Accuracy

The Model SLN_1 fits the data considerably better than a simple model using only an intercept. The F-statistic of the pulse wave component SLN_1 versus the constant model is computed as 2.21, which corresponds to the p-value $p = 0.0496$. The part of SLN_1 using HRV data only has an F-statistic of 3.08 when comparing it to the constant model, which results in $p = 0.0097$. Therefore, using $\alpha = 0.05$ as the level of significance, the combination of variables chosen for SLN_1 has a significant relationship to the response.

Table 5.1: Coefficients for the variables of the model SLN_1 and their significance in the model.

Variable	Coefficient	Estimate	p-Value
intercept	$\frac{K_{HRV}+K_{PW}}{2}$	−1.5800	$p_{HRV} < 0.0001$ $p_{PW} = 0.3553$
`age`	$\frac{c_1+c_9}{2}$	−0.0915	$p_{HRV} = 0.0022$ $p_{PW} = 0.0079$
`height`	c_8	−0.0413	$p = 0.3174$
`hr`	c_2	−0.0688	$p = 0.0012$
`lfp`	c_3	418.15	$p = 0.1944$
`lhratio`	c_4	0.1858	$p = 0.1089$
`sdnn`	c_5	−0.0967	$p = 0.0093$
`pnn50`	c_6	−2.282	$p = 0.4453$
`totalpwr`	c_7	358.86	$p = 0.2103$
`t_T`	c_{10}	0.0206	$p = 0.0154$
`t_sys_rel`	c_{11}	65.219	$p = 0.0226$
`t_sys`	c_{12}	−0.0690	$p = 0.0306$
`P_dwp_sys`	c_{13}	29.611	$p = 0.1811$
`P_notch_sys`	c_{14}	−32.164	$p = 0.1411$
`P_notch_dwp`	c_{15}	18.749	$p = 0.2606$
`pat`	c_{16}	−0.0142	$p = 0.4339$

However, the statistical significance of single variables, also computed using the F-test, has a larger range. All p-values indicating the significance of single variables with their computed coefficients in the model are given in table 5.1 in the corresponding rows. Of the 15 variables included in the model only 6, `age`, `hr`, `sdnn`, `t_T`, `t_sys_rel` and `t_sys` are considered significant.

Quality of Fit

The quality of fit is summarised using the adjusted R^2 statistic and the root mean square error. The adjusted R^2 value for the model generated using HRV data is 0.2188 while that of the model using pulse wave parameters is 0.2190. The computed adjusted R^2 statistic of the combined model SLN_1 is 0.2077.

Figure 5.1 shows the model estimates computed using SLN_1 plotted against the measured KSS values. The data is displayed for each condition during driving, A, B or C, and each time of measurement, before or after driving, separately. The root mean square error (RMSE) of predictions, calculated for each of these categories separately, ranges between 1.04 and 1.78. The lowest error is achieved on data in condition C, before driving, on which the model was trained. The exact values are given in table 5.2.

Residual Analysis

Figure 5.2 shows two types of residual plots for each of the models used to obtain a prediction. The residuals are plotted against the corresponding fitted values and each residual is

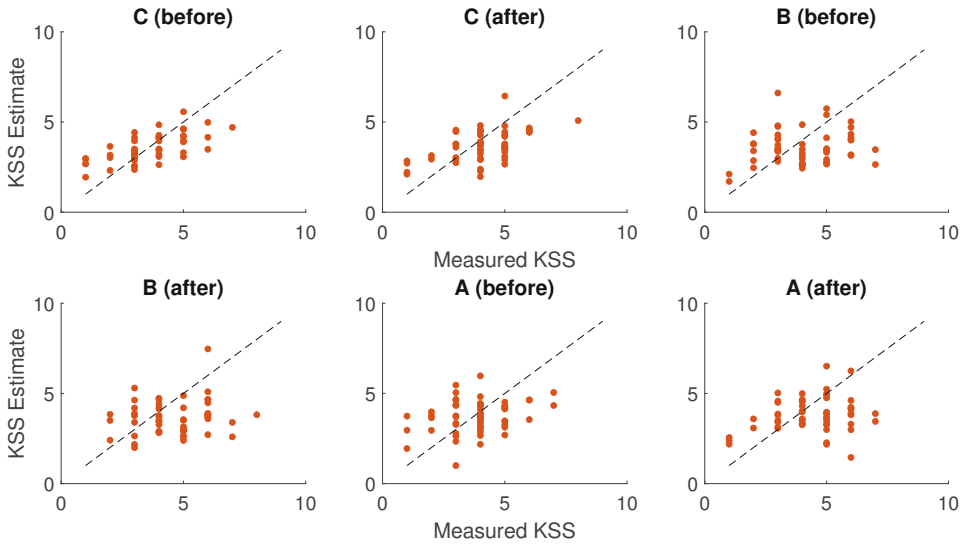

Figure 5.1: KSS values fitted by the model SLN_1 are plotted against the corresponding measured KSS values in each group. The black dotted line indicates where the fitted and measured values are equal.

Table 5.2: RMSE for KSS estimates of SLN, SLF and PCN models. The RMSE is given for each data category separately.

Data	SLN_1	SLN_2	SLF_1	SLF_2	PCN_1	PCN_2
C (before)	1.04	1.11	-	-	1.33	1.26
C (after)	1.21	1.22	1.16	1.08	1.27	1.23
B (before)	1.68	1.52	1.40	1.35	1.42	1.49
B (after)	1.78	1.76	1.58	1.52	1.56	1.65
A (before)	1.35	1.29	1.14	1.03	1.29	1.35
A (after)	1.61	1.60	1.11	1.08	1.64	1.69

plotted against its predecessor in the list of residuals.

The Anderson-Darling test of normality rejects the null hypothesis, that data comes from a normal distribution, if the probability p, of observing a result at least that extreme when the null hypothesis is true, fulfils $p < 0.05$. The test is performed separately for residuals from different measurement conditions as well as time of recording. For SLN_1, all residuals pass the test of normality, indicating no systematic error. The p-values are computed as: $p = 0.5125$ for condition C (before driving), $p = 0.2525$ for C (after driving), $p = 0.1901$ for B (before driving), $p = 0.5338$ for B (after driving), $p = 0.2841$ for A (before driving) and $p = 0.3193$ for A (after driving).

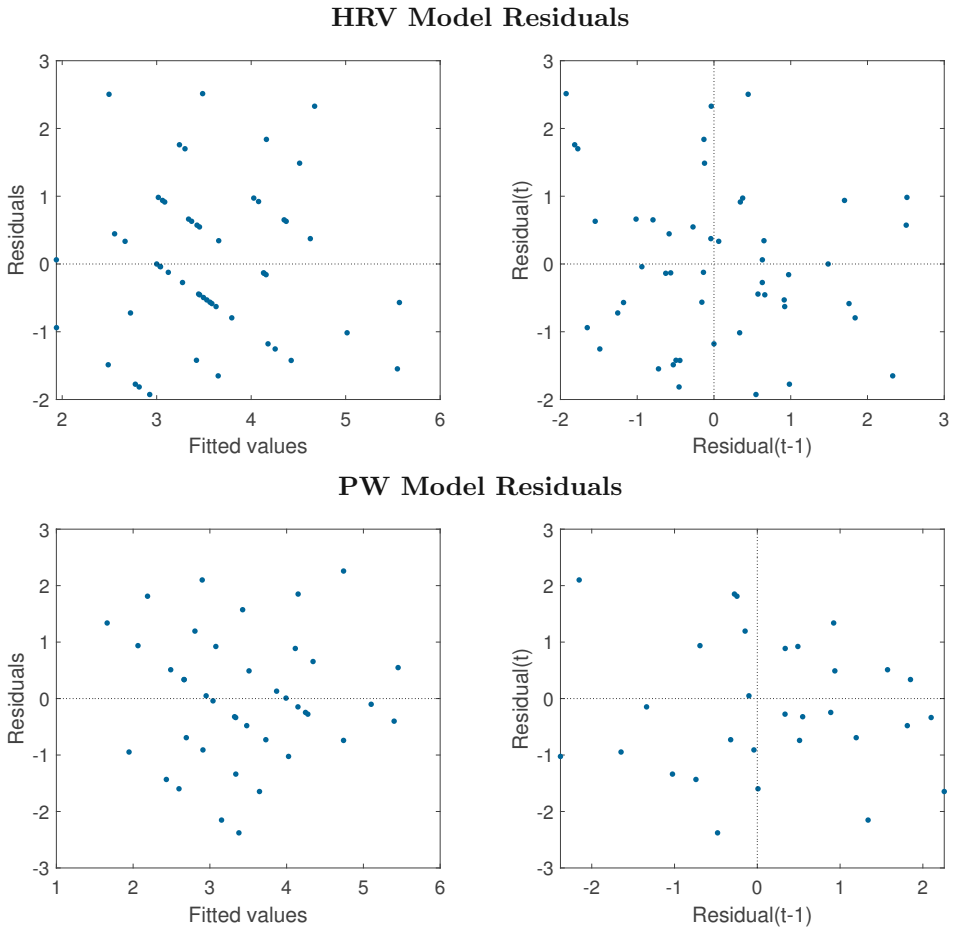

Figure 5.2: Left: Residuals plotted against the values fitted by SLN_1. **Right:** The residuals of the model SLN_1 are plotted against their predecessor in the list of residuals. In both images zero is indicated by black dotted lines.

Table 5.3: Coefficients for the variables of the model SLN_2 and their significance in the model.

Variable	Coefficient	Estimate	p-Value
intercept	$\frac{K_{HRV}+K_{PW}}{2}$	6.9476	$p_{HRV} < 0.0001$ $p_{PW} = 0.9375$
`age`	$\frac{c_1+c_7}{2}$	−0.0790	$p_{HRV} = 0.0042$ $p_{PW} = 0.0134$
`height`	c_6	−0.0514	$p = 0.1153$
`hr`	c_2	−0.0545	$p = 0.0027$
`lhratio`	c_3	0.2280	$p = 0.0440$
`sdnn`	c_4	−0.0984	$p = 0.0066$
`totalpwr`	c_5	587.1	$p = 0.0109$
`t_T`	c_8	0.0205	$p = 0.0144$
`t_sys_rel`	c_9	64.91	$p = 0.0198$
`t_sys`	c_{10}	−0.0703	$p = 0.0238$
`P_notch_sys`	c_{11}	−4.0639	$p = 0.1022$

5.2 SLN_2

The regression model SLN_2 contains ten variables and is given by the formula

$$KSS_{est} = \frac{(KSS_{HRV} + KSS_{PW})}{2}$$

where the KSS values from HRV and PW data are estimated by

$$\begin{aligned} KSS_{HRV} &= K_{HRV} + c_1\texttt{age} + c_2\texttt{hr} + c_3\texttt{lhratio} + c_4\texttt{sdnn} + c_5\texttt{totalpwr} \\ KSS_{PW} &= K_{PW} + c_6\texttt{height} + c_7\texttt{age} + c_8\texttt{t\_T} + c_9\texttt{t\_sys\_rel} + \\ &\quad c_{10}\texttt{t\_sys} + c_{11}\texttt{P\_notch\_sys}. \end{aligned}$$

The computed variable coefficients c_j, $j = 1, \ldots, 11$ are given in table 5.3.

Assessing Model Accuracy

Both models, whose combination amounts to SLN_2, are significantly better than the constant model according to the F-test. The HRV model has an F-statistic of 3.85, or a p-value of $p = 0.0053$. The F-statistic of the model using pulse wave data is computed as 2.76, which corresponds to $p = 0.0274$. Hence, the chosen set of variables has a significant relationship to the response, the KSS values.

The F-statistic for judging the significance of single variables with their computed coefficients is also calculated. The corresponding p-values can be found in table 5.3. Only two of the ten variables, `P_notch_sys` ($p = 0.1022$) and `height` ($p = 0.1153$), as well as the intercept of the pulse wave model ($p = 0.9375$), are not considered to be statistically significant.

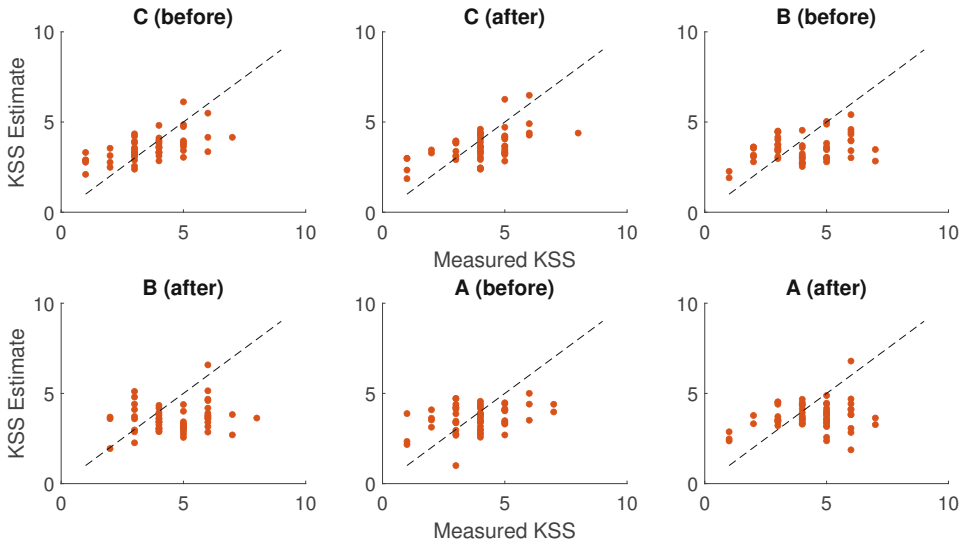

Figure 5.3: KSS values fitted by the model SLN_2 are plotted against the corresponding measured KSS values in each group. The black dotted line indicates where the fitted and measured values are equal.

Quality of Fit

The adjusted R^2 statistic is very similar for both the model based on HRV data and the model based on pulse wave data. The computed adjusted R^2 is 0.2152 and 0.2134, respectively. For SLN_2, the combination of both models, adjusted R^2 is calculated as 0.2091.

Figure 5.3 shows the KSS predictions of SLN_2 plotted against the corresponding recorded KSS values. The data is plotted for each condition of measurement, A,B or C, and each time of measurement, before or after, separately. The RMSE for the model's predictions is calculated for the same categories. The lowest RMSE, 1.11 is achieved on the training data, while the highest RMSE of 1.76 is measured in condition B, before driving. The exact values on all data categories are summarised in table 5.2.

Residual Analysis

Figure 5.4 shows the residuals plotted against corresponding fitted values. Also, each residual plotted against the preceding residual is shown. Both plots are displayed for each of the models, that determine the KSS prediction.

Again, the normality of residuals is checked using the Anderson-Darling test of normality. For the residuals of SLN_2, the null hypothesis is rejected for data measured in condition B before driving ($p = 0.0443$). For all other data categories, the residuals pass the test of normality. The computed p-values for the remaining residuals from data recorded before driving is $p = 0.9252$ for condition C and $p = 0.5716$ for condition A. The residuals from data measured after driving pass the test of normality with $p = 0.2821$ for condition C, $p = 0.7469$ for condition B and $p = 0.5716$ for condition A.

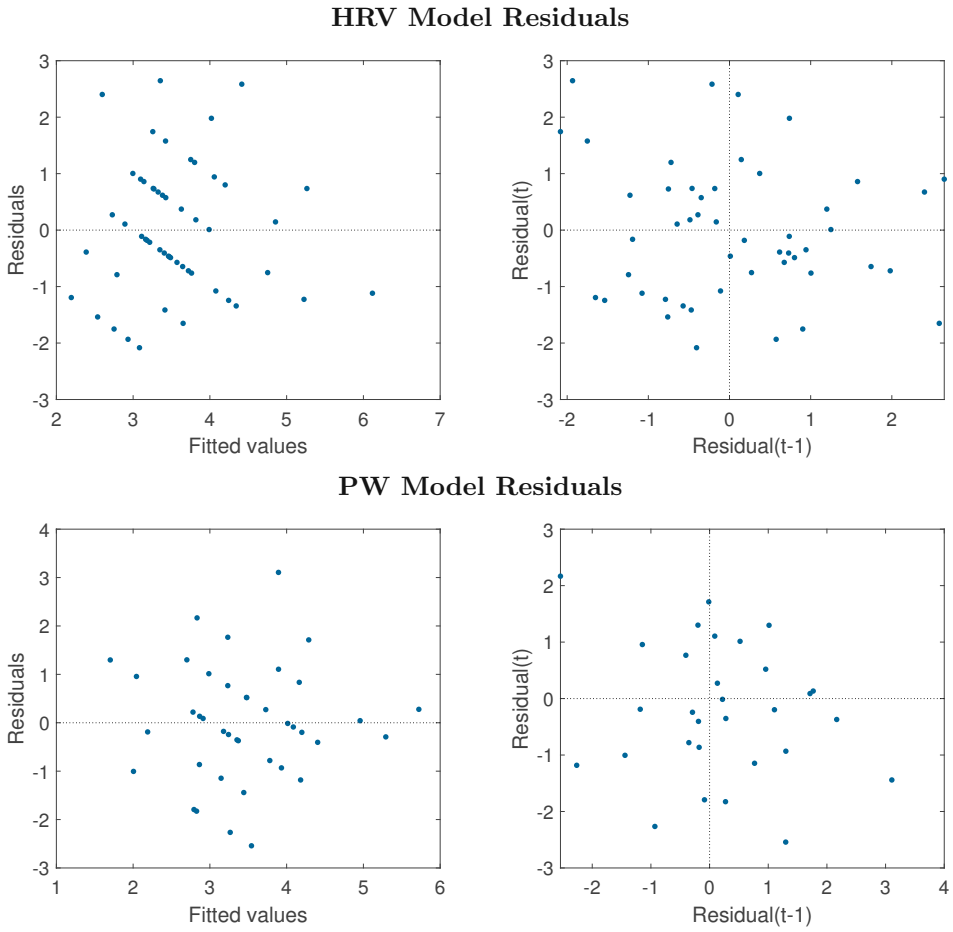

Figure 5.4: Left: Residuals plotted against the values fitted by SLN_2. **Right:** The residuals of the model SLN_2 are plotted against their predecessor in the list of residuals. In both images zero is indicated by black dotted lines.

Table 5.4: Coefficients for the variables of the model SLD_{C1} and their significance in the model.

Variable	Coefficient	Estimate	*p*-Value
intercept	$\frac{K_{\mathrm{HRV}}+K_{\mathrm{PW}}}{2}$	0.3155	$p_{\mathrm{HRV}} = 0.1172$ $p_{\mathrm{PW}} = 0.2354$
`lhratio`	c_1	0.2751	$p = 0.0035$
`t_dwp_rel`	c_2	8.4928	$p = 0.0856$
`t_sys`	c_3	−0.0160	$p = 0.0055$
`P_dwp_sys`	c_4	5.383	$p = 0.0641$
`pat`	c_5	0.0598	$p = 0.0259$

5.3 SLD_{C1}

The regression model SLD_{C1} contains five variables and is given by the formula

$$\mathrm{KSS}_{\mathrm{est}} = \frac{(\mathrm{KSS}_{\mathrm{HRV}} + \mathrm{KSS}_{\mathrm{PW}})}{2}$$

where the KSS values from HRV and PW data are estimated by

$$\begin{aligned}\mathrm{KSS}_{\mathrm{HRV}} &= K_{\mathrm{HRV}} + c_1\texttt{lhratio}\\ \mathrm{KSS}_{\mathrm{PW}} &= K_{\mathrm{PW}} + c_2\texttt{t\_dwp\_rel} + c_3\texttt{t\_sys} + c_4\texttt{P\_dwp\_sys} + c_5\texttt{pat}.\end{aligned}$$

The computed variable coefficients c_j, $j = 1, \ldots, 5$ are given in table 5.4.

Assessing Model Accuracy

The model SLD_{C1} is comprised of two models, one based on pulse wave data, the other on HRV data. According to the F-test, both models are significantly better than the constant model. The F-statistic of the HRV model versus the constant model is 9.41, which corresponds to $p = 0.0035$, while that of the pulse wave model is 3.69, which corresponds to $p = 0.0143$. Therefore, the group of variables chosen in SLD_{C1} has a statistically significant relationship with the response.

The F-test comparing the HRV model to the constant model is equivalent to testing the significance of the variable `lhratio`, since this is the only parameter included in the model. For the pulse wave variables, their significance, computed using an F-test, is given in the corresponding rows of table 5.4. The intercept is not considered as being significant. Of the five variables included in SLD_{C1}, two have a p-value above the level of significance. These are `t_dwp_rel` ($p = 0.0856$) and `P_dwp_sys` ($p = 0.0641$).

Quality of Fit

The adjusted R^2 statistic of the model generated using HRV data is 0.1416, while that of the pulse wave model is 0.2354. The adjusted R^2 statistic of the model SLD_{C1}, which combines both previously mentioned models, is 0.1715.

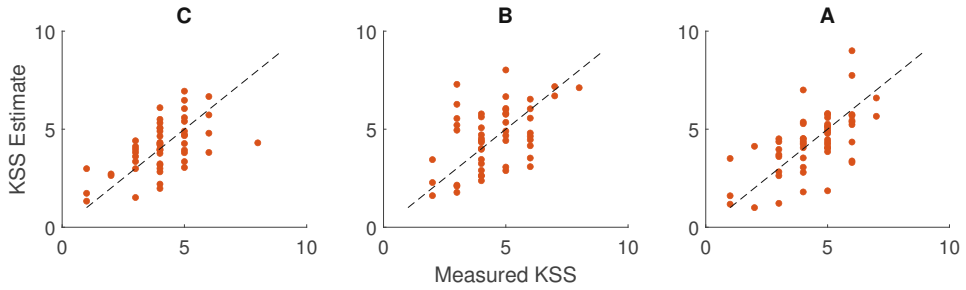

Figure 5.5: KSS values fitted by the model SLD_{C1} are plotted against the corresponding measured KSS values in each group. The black dotted line indicates where the fitted and measured values are equal.

Table 5.5: RMSE for KSS estimates of models using dynamic baselines, i.e. SLD and PCD models. The RMSE is given for each data category separately.

Data	SLD_{C1}	SLD_{C2}	SLD_{S1}	SLD_{S2}	PCD_1	PCD_2
C	1.19	0.93	1.30	0.76	1.19	1.18
B	1.52	1.60	1.29	1.48	1.68	1.54
A	1.30	1.50	1.35	1.97	1.36	1.44

Figure 5.5 shows KSS estimates of the model SLD_{C1} plotted against the respective measured KSS values. Since data recorded both before and after driving is used for a prediction, the data is shown in only three separate categories, each containing recordings from only one of the conditions, A, B or C. The RMSE of the predictions is evaluated on the same categories and is displayed in table 5.5. It ranges between 1.19, on the training data set, and 1.52, for data recorded in condition B.

Residual Analysis

Figure 5.6 shows four residual plots, two for each model used in the KSS prediction. One displays the residuals, plotted against the corresponding fitted values. The other shows the plot of each residual against the previous residual.

The normality of residuals is checked for data from condition A, B and C separately using the Anderson-Darling Test of normality. The p-values calculated are: $p = 0.1503$ for condition C, $p = 0.7392$ for condition B and $p = 0.5734$ for condition A. Therefore all residuals pass the test of normality.

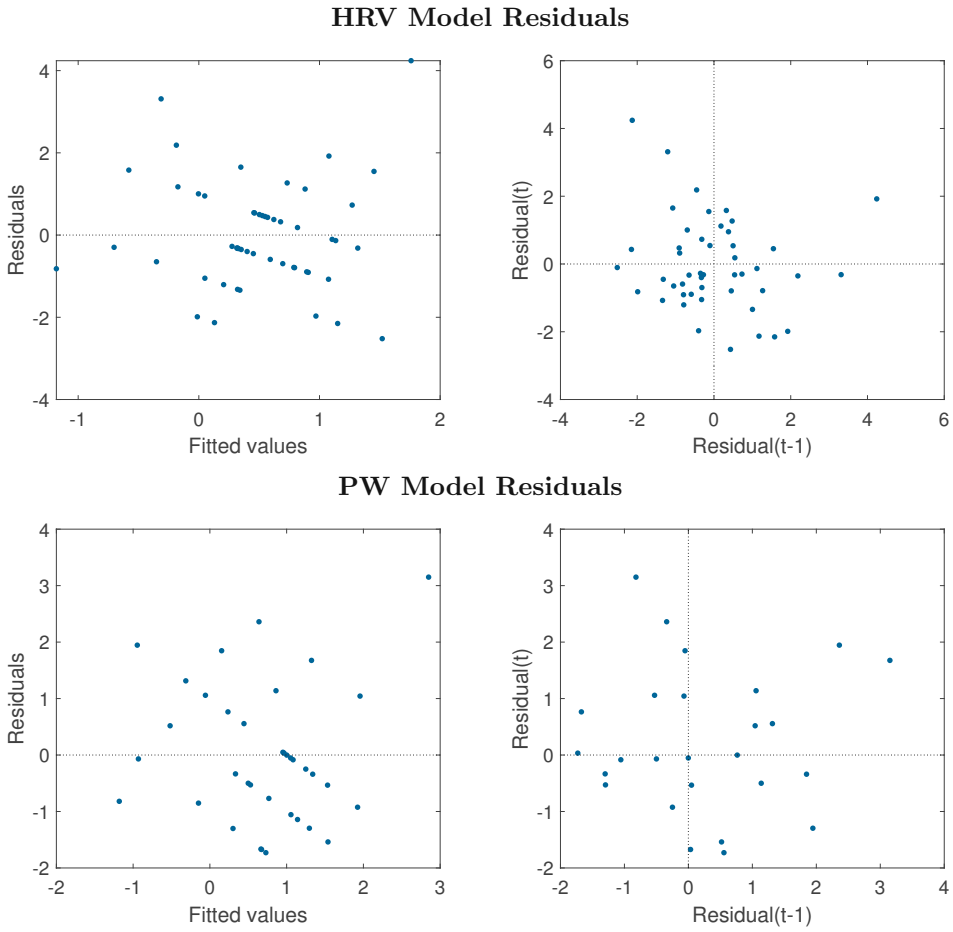

Figure 5.6: Left: Residuals plotted against the values fitted by SLD_{C1}. **Right:** The residuals of the model SLD_{C1} are plotted against their predecessor in the list of residuals. In both images zero is indicated by black dotted lines.

Table 5.6: Coefficients for the variables of the model SLD_{C2} and their significance in the model.

Variable	Coefficient	Estimate	p-Value
intercept	$\frac{K_{\text{HRV}}+K_{\text{PW}}}{2}$	0.2788	$p_{\text{HRV}} = 0.8038$ $p_{\text{PW}} = 0.0158$
`hr`	c_1	-0.0603	$p = 0.0006$
`lfp`	c_2	-813.46	$p < 0.0001$
`hfp`	c_3	3677.9	$p < 0.0001$
`lhratio`	c_4	0.5073	$p = 0.0011$
`pnn50`	c_5	-11.138	$p = 0.0002$
`hfp_n`	c_6	-4.8431	$p = 0.0295$
`t_notch`	c_7	0.0370	$p = 0.0376$
`t_sys_rel`	c_8	28.101	$p = 0.0325$
`t_sys`	c_9	-0.0696	$p = 0.0150$
`P_dwp_sys`	c_{10}	45.601	$p = 0.0137$
`P_notch_sys`	c_{11}	-37.86	$p = 0.0333$
`P_notch_dwp`	c_{12}	29.858	$p = 0.0408$

5.4 SLD_{C2}

The regression model SLD_{C2} contains twelve variables and is given by the formula

$$\text{KSS}_{\text{est}} = \frac{(\text{KSS}_{\text{HRV}} + \text{KSS}_{\text{PW}})}{2}$$

where the KSS values from HRV and PW data are estimated by

$$\begin{aligned}\text{KSS}_{\text{HRV}} &= K_{\text{HRV}} + c_1\texttt{hr} + c_2\texttt{lfp} + c_3\texttt{hfp} + c_4\texttt{lhratio} + c_5\texttt{pnn50} + c_6\texttt{hfp\_n}\\ \text{KSS}_{\text{PW}} &= K_{\text{PW}} + c_7\texttt{t\_notch} + c_8\texttt{t\_sys\_rel} + c_9\texttt{t\_sys} + c_{10}\texttt{P\_dwp\_sys} +\\ &\quad c_{11}\texttt{P\_notch\_sys} + c_{12}\texttt{P\_notch\_dwp}.\end{aligned}$$

The computed variable coefficients c_j, $j = 1, \ldots, 12$ are given in table 5.6.

Assessing Model Accuracy

The F-test is used to determine the accuracy of coefficient estimates. The F-statistic of the model generated from HRV data versus the constant model is 9.34, which corresponds to $p < 0.0001$. The F-statistic of the pulse wave model is 2.99, which corresponds to $p = 0.0214$. Hence, The combination of variables used in SLD_{C2} has a significant relationship the response.

The F-test is also used to determine the statistical significance of single variables within the model. The intercept determined from HRV data has an exceptionally high p-value ($p = 0.8038$). Otherwise all variables are deemed to be significant by the F-test at a significance level of $\alpha = 0.05$. The p-values for all variables are given in table 5.6 in the corresponding rows.

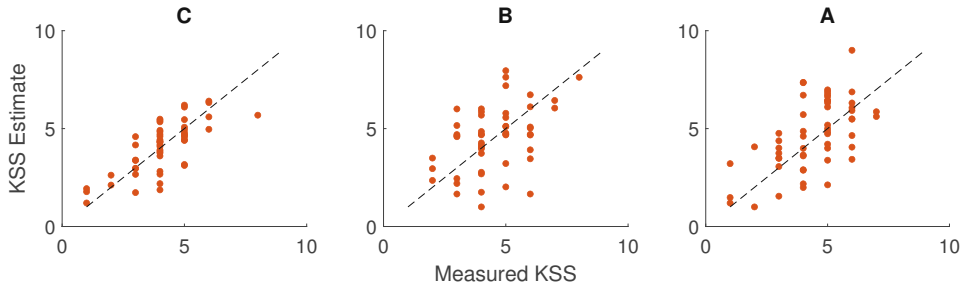

Figure 5.7: KSS values fitted by the model SLD_{C2} are plotted against the corresponding measured KSS values in each group. The black dotted line indicates where the fitted and measured values are equal.

Quality of Fit

The adjusted R^2 statistic of the model generated from HRV data is computed as 0.4953, while that of the model trained on pulse wave data is 0.2542. The combination of both models, SLD_{C2}, has an adjusted R^2 statistic of 0.3983.

Figure 5.7 shows the KSS estimates of SLD_{C2} plotted against the measured KSS values. The data is shown for each condition during driving, A, B or C, separately. The RMSE of predictions, also calculated for each of these categories separately, ranges from 0.93 to 1.50. The RMSE for all conditions is summarised in table 5.5. The lowest error is achieved on condition C, the data the model was trained on. The highest error was calculated for data from condition A.

Residual Analysis

Figure 5.4 shows the residuals plotted against corresponding fitted values. Also, in a second plot, each residual is plotted against the preceding residual. Both residual plots are displayed for each of the models, that determine the KSS prediction.

The normality of residuals is checked using the Anderson-Darling test. The residuals of both the conditions A ($p = 0.9671$) and B ($p = 0.8937$) pass the test of normality. However, for data from condition C, the test rejects the null hypothesis, that it comes from a normal distribution, with $p = 0.0480$.

5.5 SLD_{S1}

The regression model SLD_{S1} contains one variable and is given by the formula

$$\text{KSS} = K + c_1 \texttt{lhratio}.$$

where the KSS values are estimated from only one HRV parameter. The computed variable coefficient for `lhratio` is given by $c_1 = 0.2751$ with a significance of $p = 0.0035$. The intercept is estimated as $K = 0.3123$ with a significance level of $p = 0.1172$.

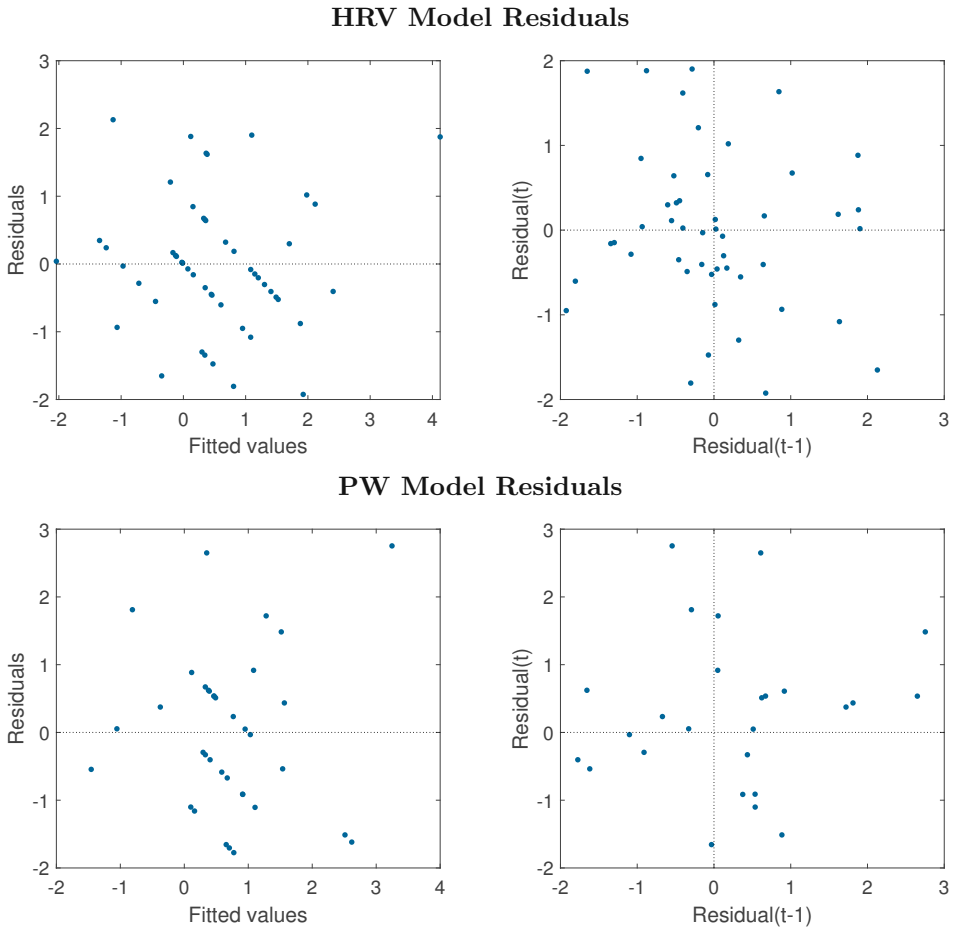

Figure 5.8: Left: Residuals plotted against the values fitted by SLD_{C2}. **Right:** The residuals of the model SLD_{C2} are plotted against their predecessor in the list of residuals. In both images zero is indicated by black dotted lines.

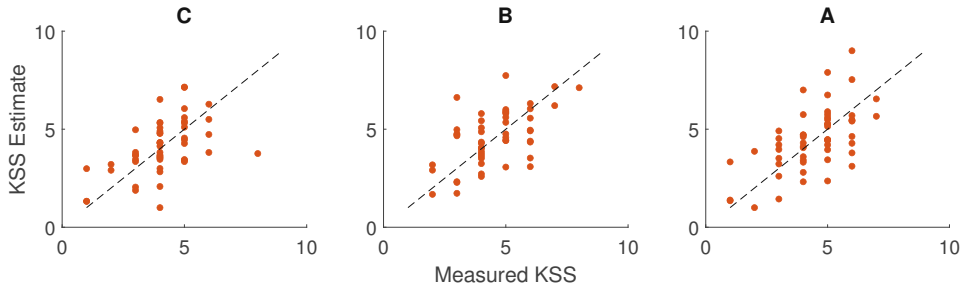

Figure 5.9: KSS values fitted by the model SLD_{S1} are plotted against the corresponding measured KSS values in each group. The black dotted line indicates where the fitted and measured values are equal.

Assessing Model Accuracy

The F-statistic of SLD_{S1} versus the constant model is 9.41 and the corresponding p-value is $p = 0.0035$. Therefore, according to the F-test, the variable included in the model, `lhratio`, has a significant relationship to the response. The intercept K ($p = 0.1172$) has a p-value above the level of significance.

Quality of Fit

The adjusted R^2 statistic of the model SLD_{S1} is computed as 0.1416. The KSS estimates calculated by SLD_{S1} are shown in figure 5.9, plotted against the corresponding measured KSS values. The data is displayed separately for each of the conditions, A, B or C. The RMSE is calculated for the same data categories is shown in table 5.5. The RMSE of all categories is very similar. The lowest RMSE, 1.29 is measured on condition B, while the highest, 1.35, is from condition A.

Residual Analysis

Figure 5.10 shows two residual plots. The residuals are plotted against the corresponding fitted values as well as against their preceding residual. The Anderson-Darling test confirms, that the residuals are likely to follow a normal distribution. The calculated p-values are: $p = 0.1990$ for condition C, $p = 0.8843$ for condition B and $p = 0.6210$ for condition A.

5.6 SLD_{S2}

The regression model SLD_{S2} contains eleven variables and is given by the formula

$$\begin{aligned}\text{KSS} = K &+ c_1\texttt{weight} + c_2\texttt{height} + c_3\texttt{age} + c_4\texttt{hr} + c_5\texttt{lfp} + c_6\texttt{hfp} + \\ &c_7\texttt{lhratio} + c_8\texttt{pnn50} + c_9\texttt{t\_notch\_rel} + c_{10}\texttt{t\_sys} + c_{11}\texttt{P\_dwp\_sys}.\end{aligned}$$

where the KSS values are estimated from HRV and PW data simultaneously. The computed variable coefficients c_j, $j = 1, \ldots, 11$ are given in table 5.7.

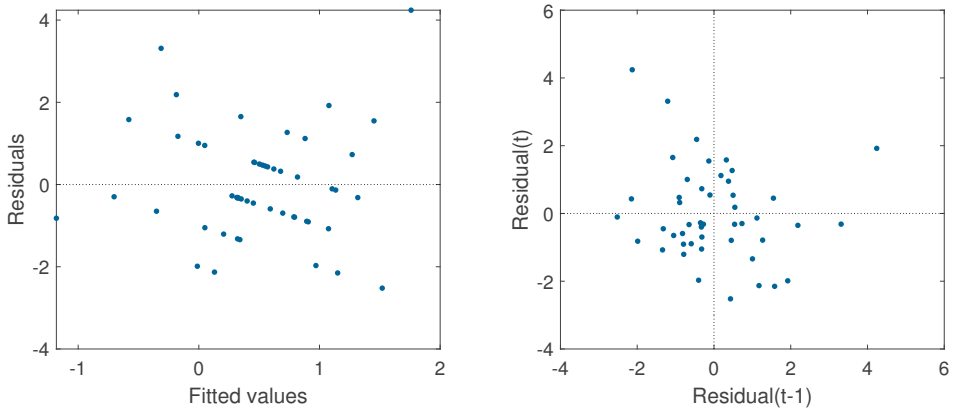

Figure 5.10: Left: Residuals plotted against the values fitted by SLD_{S1}. **Right:** The residuals of the model SLD_{S1} are plotted against their predecessor in the list of residuals. In both images zero is indicated by black dotted lines.

Table 5.7: Coefficients for the variables of the model SLD_{S2} and their significance in the model.

Variable	Coefficient	Estimate	p-Value
intercept	K	-7.2307	$p = 0.1906$
`weight`	c_1	-0.0298	$p = 0.0612$
`height`	c_2	0.0453	$p = 0.1263$
`age`	c_3	0.0379	$p = 0.1261$
`hr`	c_4	-0.0966	$p = 0.0201$
`lfp`	c_5	-1175.5	$p = 0.0015$
`hfp`	c_6	5567.1	$p = 0.0011$
`lhratio`	c_7	1.0477	$p < 0.0001$
`pnn50`	c_8	-13.467	$p = 0.0004$
`t_notch_rel`	c_9	22.762	$p = 0.0214$
`t_sys`	c_{10}	-0.0237	$p = 0.0156$
`P_dwp_sys`	c_{11}	6.3948	$p = 0.0490$

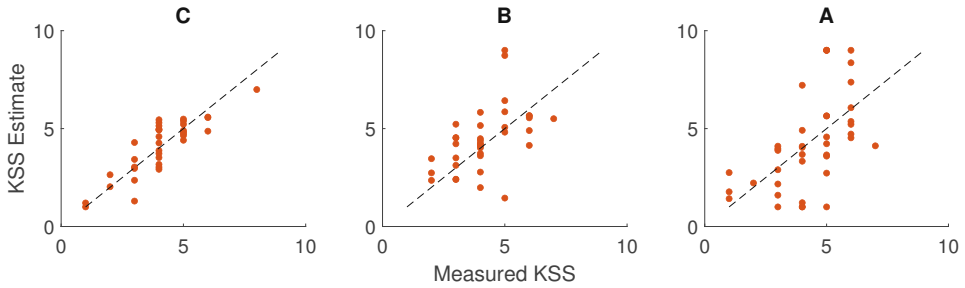

Figure 5.11: KSS values fitted by the model SLD_{S2} are plotted against the corresponding measured KSS values in each group. The black dotted line indicates where the fitted and measured values are equal.

Assessing Model Accuracy

In comparison to a constant model, the F-statistic of SLD_{S2} is 5.83, which corresponds to $p = 0.0002$. Hence, the selected variables with their estimated coefficients have a significant relationship to the response. The F statistic is also used to determine the statistical significance of single variables with their computed coefficient. The coefficient of the intercept K ($p = 0.1906$) is not deemed statistically significant by the F-test. Of the eleven chosen variables, three have p-values above the level of significance $\alpha = 0.05$. This concerns all variables from metadata: weight ($p = 0.0612$), height ($p = 0.1263$) and age ($p = 0.1261$). All p-values for all other variables are below 0.05 and are summarised in table 5.7.

Quality of Fit

The adjusted R^2 statistic of SLD_{S2} is 0.6029. The estimates of SLD_{S2}, plotted against the corresponding recorded KSS values, can be seen in figure 5.11. The values are plotted separately for each of the conditions. The RMSE was also calculated for each of the conditions and is given in table 5.5. The range between the lowest RMSE, 0.76 on the training data fro condition C, and the highest RMSE, 1.97 on condition A, is comparatively large.

Residual Analysis

Figure 5.12 shows two residual plots for the model SLD_{S2}. The first shows the residuals plotted against the corresponding fitted values. The second shows each residual plotted against its predecessor.

All residuals pass the Anderson-Darling test of normality. For condition C, the training data, the test result is $p = 0.9457$. For conditions B and A the results are $p = 0.2228$ and $p = 0.4328$, respectively.

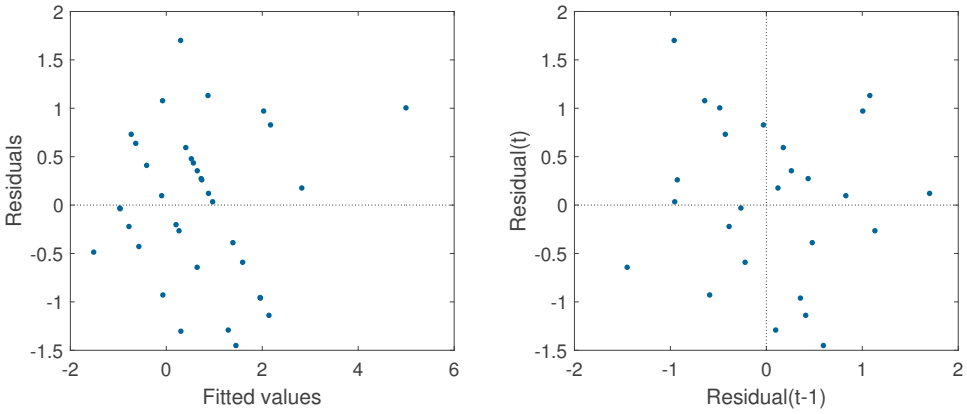

Figure 5.12: Left: Residuals plotted against the values fitted by SLD_{S2}. **Right:** The residuals of the model SLD_{S2} are plotted against their predecessor in the list of residuals. In both images zero is indicated by black dotted lines.

Table 5.8: Coefficients for the variables of the model SLF_1 and their significance in the model.

Variable	Coefficient	Estimate	p-Value
intercept	K	-1.8512	$p = 0.0596$
`age`	c_1	0.0595	$p = 0.0118$
`lfp_n`	c_2	2.6633	$p = 0.0182$
`t_sys_rel`	c_3	-21.868	$p = 0.0034$
`t_notch_rel`	c_4	-21.654	$p = 0.0109$
`P_dwp_sys`	c_5	15.309	$p = 0.0223$
`P_notch_sys`	c_6	-8.3211	$p = 0.1491$
`P_notch_dwp`	c_7	8.2952	$p = 0.0429$

5.7 SLF_1

The regression model SLF_1 contains seven variables and is given by the formula

$$\mathrm{KSS} = K + c_1\texttt{age} + c_2\texttt{lfp\_n} + c_3\texttt{t\_sys\_rel} + c_4\texttt{t\_notch\_rel} + c_5\texttt{P\_dwp\_sys} + c_6\texttt{P\_notch\_sys} + c_7\texttt{P\_notch\_dwp}.$$

where the KSS values are estimated from HRV and PW data simultaneously. The computed variable coefficients c_j, $j = 1, \ldots, 7$ are given in table 5.8.

Assessing Model Accuracy

The F-statistic for SLF_1 versus a constant model is 4.46, which corresponds to $p = 0.0003$. Hence, the variables selected for SLF_1 with their coefficients describe a statistically significant relationship to the response. When using the F test to determine the significance of single variables in the model, the intercept has a result, $p = 0.0596$ just above the level of

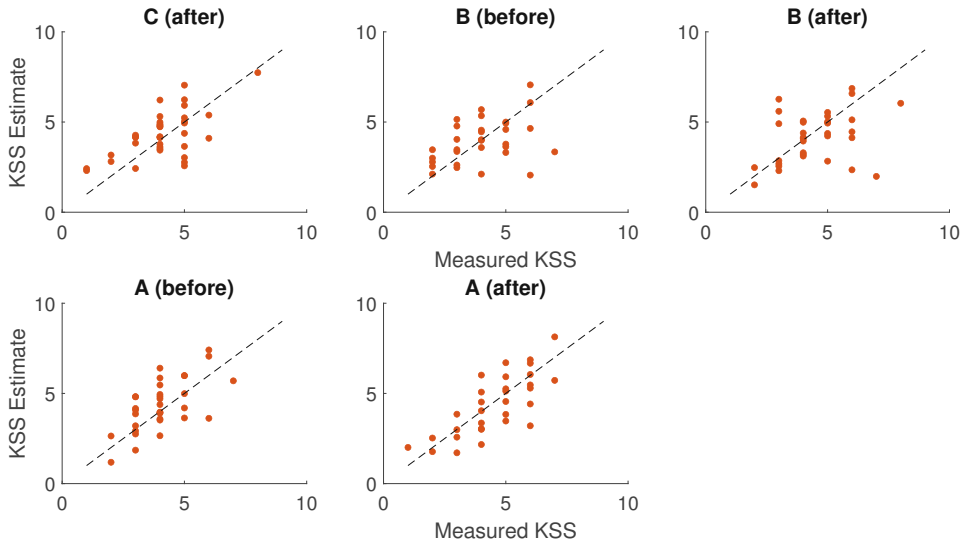

Figure 5.13: KSS values fitted by the model SLF_1 are plotted against the corresponding measured KSS values in each group. The black dotted line indicates where the fitted and measured values are equal.

significance. Only one of the seven selected variables, `P_notch_sys` ($p = 0.1491$) does not pass the F-test. All computed p-values are given in table 5.8.

Quality of Fit

The adjusted R^2 statistic of SLF_1 is calculated as 0.2238. Figure 5.13 shows the KSS estimates by SLF_1 plotted against the corresponding measured KSS values. The data from before and after driving, as well as data from different conditions, is evaluated separately. Since all data from condition C, before driving, is used to determine baselines, this data can not used for evaluations. The RMSE is calculated for the same categories and displayed in table 5.2. The RMSE ranges from 1.11, on data from condition A after driving, to 1.58, on data from condition B after driving.

Residual Analysis

Two residual plots were created for the model SLF_1. They are displayed in figure 5.14. The first shows the residuals plotted against the corresponding fitted values. The second shows each residual plotted against its predecessor.

The residuals are checked for normality using the Anderson-Darling test. For residuals from condition C, after driving, and condition B, after driving, the null hypothesis, that they come from a normal distribution, is rejected with $p = 0.0360$ and $p = 0.0431$, respectively. For residuals from condition B, before driving, the determined p-value is $p = 0.0580$, which is above the level of significance. Residuals from condition A, both before driving ($p = 0.5805$) and after driving ($p = 0.9282$), also pass the test of normality.

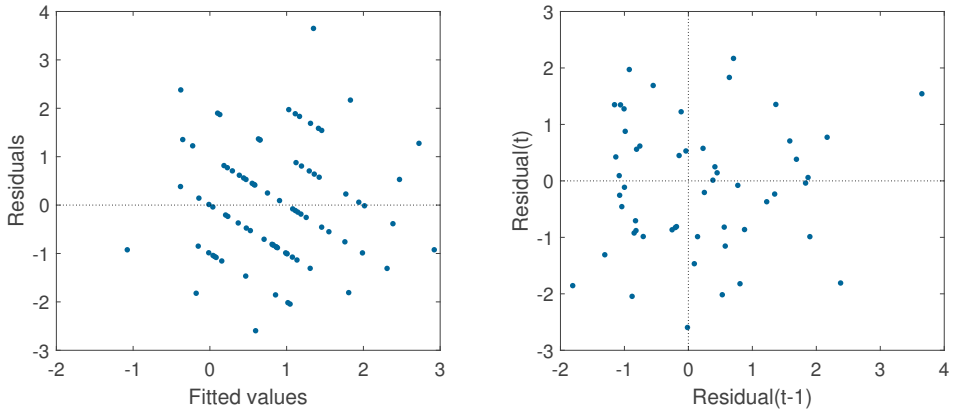

Figure 5.14: Left: Residuals plotted against the values fitted by SLF_1. **Right:** The residuals of the model SLF_1 are plotted against their predecessor in the list of residuals. In both images zero is indicated by black dotted lines.

Table 5.9: Coefficients for the variables of the model SLF_2 and their significance in the model.

Variable	Coefficient	Estimate	p-Value
intercept	K	4.8773	$p = 0.1623$
`height`	c_1	−0.0318	$p = 0.0807$
`age`	c_2	0.0378	$p = 0.0160$
`hr`	c_3	−0.0736	$p = 0.0845$
`lfp`	c_4	−394.46	$p = 0.0613$
`lhratio`	c_5	−0.1314	$p = 0.1413$
`totalpwr`	c_6	325.96	$p = 0.0188$
`lfp_n`	c_7	4.3528	$p = 0.0142$
`t_T`	c_8	−0.0093	$p = 0.0240$

5.8 SLF_2

The regression model SLF_2 contains 8 variables and is given by the formula

$$\begin{aligned}\text{KSS} = K + c_1\texttt{height} + c_2\texttt{age} + c_3\texttt{hr} + c_4\texttt{lfp} + c_5\texttt{lhratio} + \\ c_6\texttt{totalpwr} + c_7\texttt{lfp\_n} + c_8\texttt{t\_T}.\end{aligned}$$

where the KSS values are estimated from HRV and PW data simultaneously. The computed variable coefficients c_j, $j = 1, \ldots, 8$ are given in table 5.9.

Assessing Model Accuracy

The F-test suggests there is a significant relationship between the chosen set of variables with their coefficient estimates and the KSS value. The computed F-statistic of SLF_2 versus a constant model is 3.11, which corresponds to $p = 0.0037$. When testing the statistical significance of single variables in SLF_2 using the F-test, only four of the eight

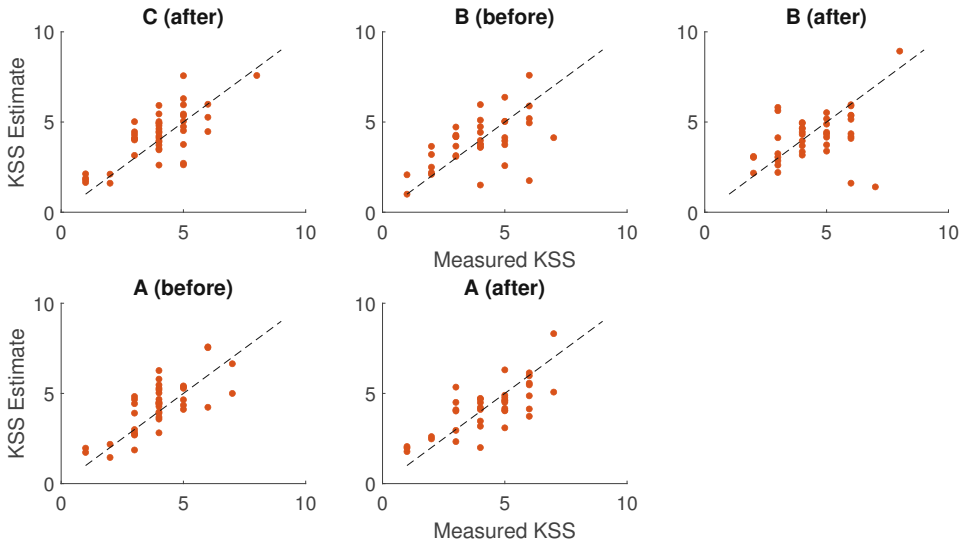

Figure 5.15: KSS values fitted by the model SLF_2 are plotted against the corresponding measured KSS values in each group. The black dotted line indicates where the fitted and measured values are equal.

selected variables pass the test. They are: `age` ($p = 0.0160$), `totalpwr` ($p = 0.0118$), `lfp_n` ($p = 0.0142$) and `t_T` ($p = 0.0240$). All other variables, as well as the intercept, have p-values higher than the level of significance. All p-values are given in table 5.9.

Quality of Fit

The model SLF_2 has an adjusted R^2 statistic of 0.1419. Figure 5.15 shows the model's KSS estimates plotted against the corresponding measured KSS values. The values are plotted separately for data recorded before and after driving, and for each condition. There is no data in condition C, before driving, since this was used to generate baselines. The RMSE is also evaluated separately for different conditions and times of measurement. All RMSE values can be found in table 5.2. The lowest RMSE, 1.03, is recorded on data from condition A, before driving, while the highest RMSE, 1.52, is measured in condition B, after driving.

Residual Analysis

Figure 5.12 shows two residual plots for the model SLF_2. The first shows the residuals plotted against the corresponding fitted values. The second shows each residual plotted against its predecessor.

Residuals are checked for normality using the Anderson-Darling test. Residuals from condition C, after driving, and condition A, both before and after driving, pass the test of normality with $p = 0.2896$, $p = 0.8554$ and $p = 0.6833$, respectively. Residuals from condition B, both before driving ($p = 0.0197$) and after driving ($p = 0.0021$), fail the test of normality.

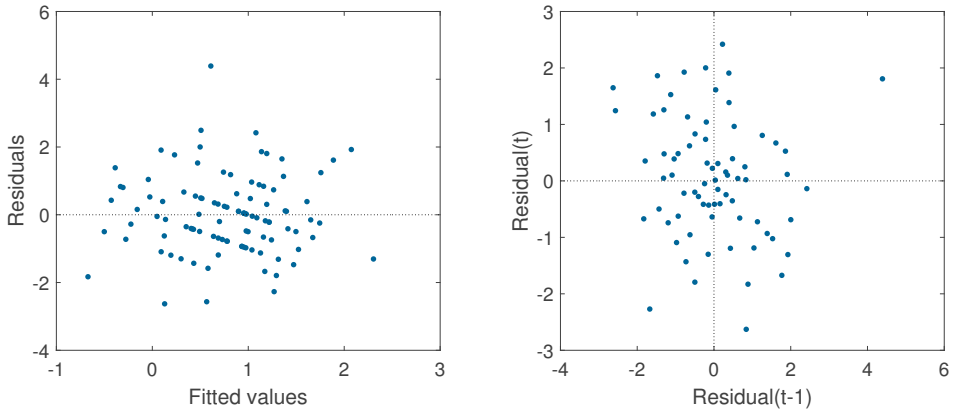

Figure 5.16: Left: Residuals plotted against the values fitted by SLF_2. **Right:** The residuals of the model SLF_2 are plotted against their predecessor in the list of residuals. In both images zero is indicated by black dotted lines.

Table 5.10: Coefficients for the variables of the model PCN_1 and PCN_2, as well as their significance in the corresponding model.

Variable	Coefficient	PCN_1		PCN_2	
		Estimate	p-Value	Estimate	p-Value
PC_1	c_1	0.1579	$p = 0.0973$	0.1704	$p = 0.0701$
PC_2	c_2	0.0635	$p = 0.5405$	0.0692	$p = 0.4963$
PC_3	c_3	0.0227	$p = 0.8596$	0.0079	$p = 0.9502$
PC_4	c_4	-0.0114	$p = 0.9465$	-0.0049	$p = 0.9763$
PC_5	c_5	0.2545	$p = 0.1928$	0.2723	$p = 0.1567$
PC_6	c_6	-0.3018	$p = 0.1775$	-0.3117	$p = 0.1559$
PC_7	c_7	-	-	-0.4634	$p = 0.0769$
PC_8	c_8	-	-	0.1577	$p = 0.5693$

5.9 PCN_1

The regression model PCN_1 is a principal component regression model using 6 PCs without an intercept. It is given by the formula

$$\mathrm{KSS} = \sum_{i=1}^{6} c_i \cdot \mathrm{PC}_i.$$

The computed variable coefficients c_j, $j = 1, \ldots, 6$ are given in table 5.10.

Assessing Model Accuracy

For PCN_1, the F-statistic is returned as 1.07, which corresponds to $p = 0.5726$. Hence, there seems to be no significant relationship between the chosen PCs and the response. The F-test, determining the significance of single variables in the model, gives high p-values for

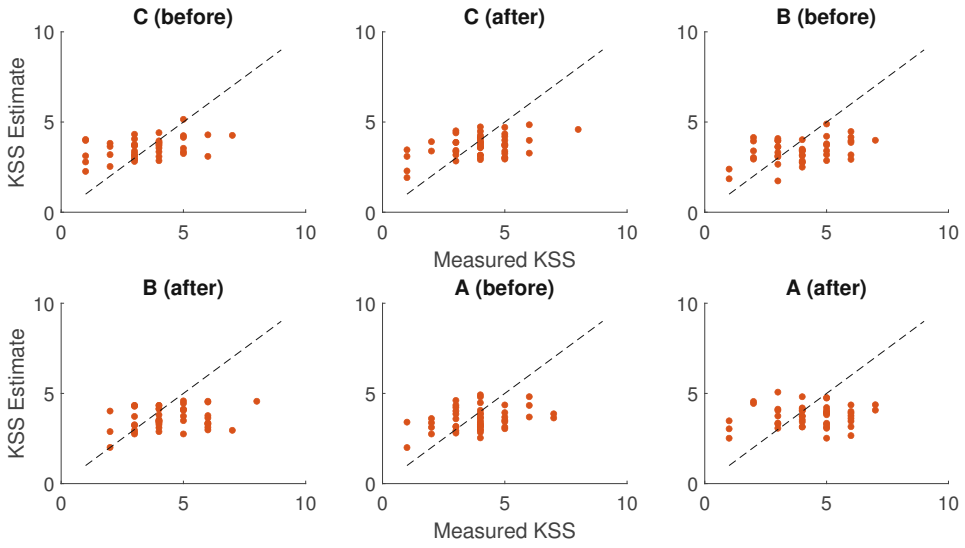

Figure 5.17: KSS values fitted by the model PCN_1 are plotted against the corresponding measured KSS values in each group. The black dotted line indicates where the fitted and measured values are equal.

all variables. Hence, none of the principal components are considered to be statistically significant. The exact p-values are given in table 5.10.

Quality of Fit

The PC model using 6 PCs and no baseline achieves an adjusted R^2 statistic of 0.0295, which is exceptionally low. Figure 5.17 shows the KSS estimates by PCN_1 plotted against the corresponding measured values. The data is displayed separately for each condition, A, B or C, and both times of recording, before or after driving. The RMSE for PCN_1 predictions ranges between 1.27, on data from condition C, after driving, to 1.64, on condition A, after driving. The RMSE for each data category is given in table 5.2.

Residual Analysis

Figure 5.18 depicts the model's residuals plotted against the corresponding fitted values. The same figure also shows a second plot, where each residual is plotted against the preceding residual.

The normality of residuals is checked using the Anderson-Darling test. Residuals from all conditions and times of measurement pass the test of normality. The p-values, by condition and time of recording, are: $p = 0.7817$ and $p = 0.9728$ (C, before and after driving), $p = 0.5431$ and $p = 0.1406$ (B, before and after driving), and lastly $p = 0.7582$ and $p = 0.4985$ (A, before and after driving).

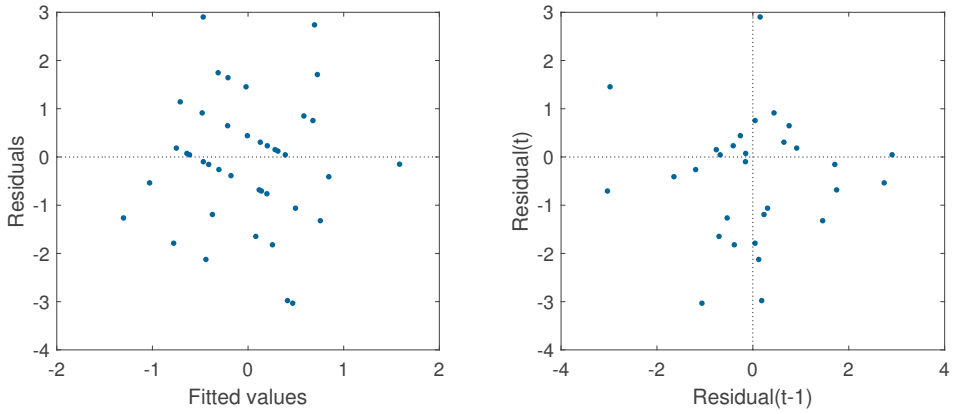

Figure 5.18: Left: Residuals plotted against the values fitted by PCN_1. **Right:** The residuals of the model PCN_1 are plotted against their predecessor in the list of residuals. In both images zero is indicated by black dotted lines.

5.10 PCN_2

The regression model PCN_2 is a principal component regression model using 8 PCs without an intercept. It is given by the formula

$$\text{KSS} = \sum_{i=1}^{8} c_i \cdot \text{PC}_i.$$

The computed variable coefficients c_j, $j = 1, \ldots, 8$ are given in table 5.10.

Assessing Model Accuracy

As with the previous model, that uses less PCs, the combination of variables in the model is not considered to be statistically significant. The F-statistic of PCN_2 is 1.28, which corresponds to $p = 0.4924$. Similar to the previously discussed model, none of the coefficients are deemed statistically significant. The exact p-values returned by the F-test are given in table 5.10.

Quality of Fit

The adjusted R^2 statistic of PCN_2 is computed as 0.0708, which is extremely low. Figure 5.19 shows the KSS estimates by PCN_2 plotted against the corresponding measures values for each condition and time of measurement separately. The model's RMSE ranges from 1.26 to 1.69. The lowest value is achieved on data from condition C, after driving, while the highest RMSE is measured on data from condition A, after driving. The RMSE is shown for each category separately in table 5.2.

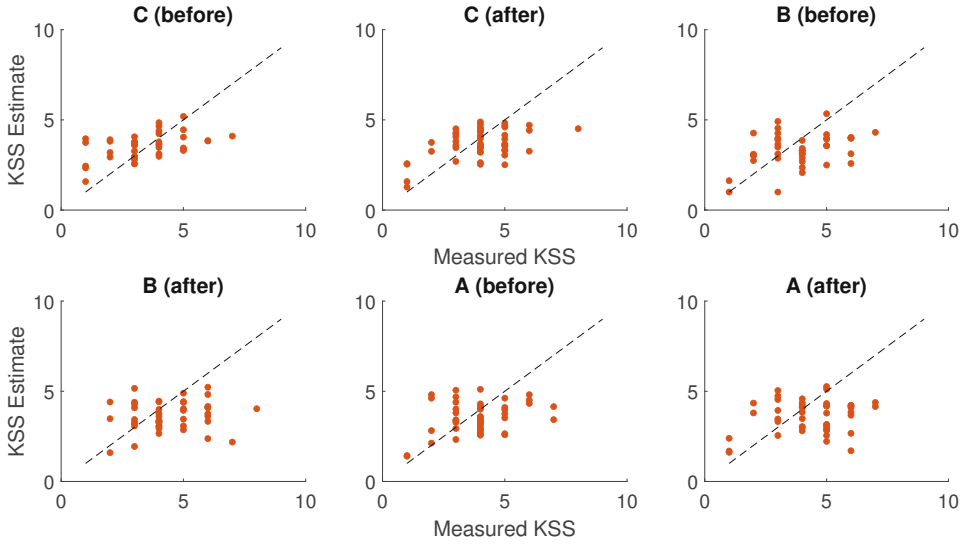

Figure 5.19: KSS values fitted by the model PCN_2 are plotted against the corresponding measured KSS values in each group. The black dotted line indicates where the fitted and measured values are equal.

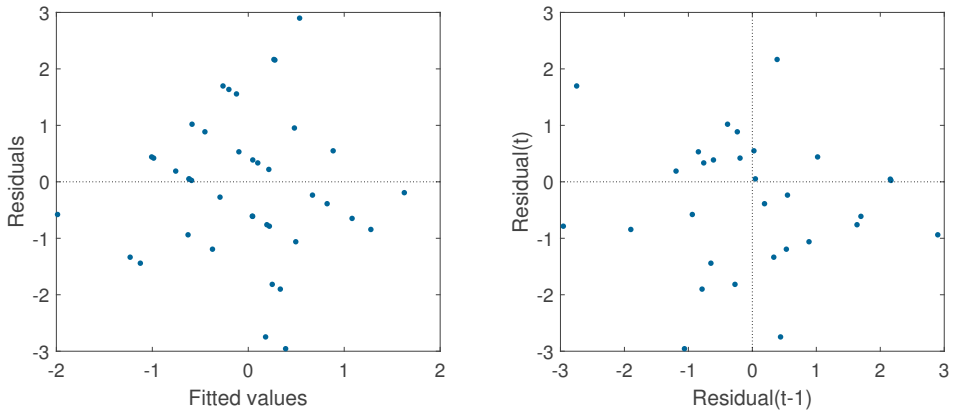

Figure 5.20: Left: Residuals plotted against the values fitted by PCN_2. **Right:** The residuals of the model PCN_2 are plotted against their predecessor in the list of residuals. In both images zero is indicated by black dotted lines.

Table 5.11: Coefficients for the variables of the model PCD_1 and PCD_2, as well as their significance in the corresponding model.

Variable	Coefficient	PCD_1		PCD_2	
		Estimate	p-Value	Estimate	p-Value
PC_1	c_1	0.0687	$p = 0.4765$	0.0728	$p = 0.4518$
PC_2	c_2	−0.1297	$p = 0.2915$	−0.1355	$p = 0.2722$
PC_3	c_3	−0.2866	$p = 0.0294$	−0.2884	$p = 0.0289$
PC_4	c_4	−0.3954	$p = 0.0369$	−0.3925	$p = 0.0386$
PC_5	c_5	0.3400	$p = 0.0743$	−0.3351	$p = 0.0791$
PC_6	c_6	-	-	0.2050	$p = 0.3376$

Residual Analysis

Figure 5.20 shows the residuals of PCN_2 plotted against corresponding fitted values. The plot of each residual against its predecessor is also shown.

According to the Anderson-Darling test of normality, the residuals from all conditions, before and after driving, are likely to be normally distributed. The p-values for condition C are $p = 0.8471$ for data from before and $p = 0.5257$ for data from after driving. Those of condition B are $p = 0.4248$ before and $p = 0.2880$ after driving. In condition A the results are $p = 0.4360$ and $p = 0.3464$, before and after driving, respectively.

5.11 PCD_1

The regression model PCD_1 is a principal component regression model using 5 PCs without an intercept. It is given by the formula

$$\mathrm{KSS} = \sum_{i=1}^{5} c_i \cdot \mathrm{PC}_i.$$

The computed variable coefficients c_j, $j = 1, \ldots, 5$ are given in table 5.11.

Assessing Model Accuracy

The F-test against the constant model suggest, that the combination of PCs is statistically significant. The results are $F = 3.28$ and $p = 0.0235$. The F-test is also used to determine the statistical significance of single variables. Of the five PCs included in the model, only the third ($p = 0.0294$) and fourth ($p = 0.0369$) PC are considered significant. The F-test's results for each PC with its coefficient estimate is given in table 5.11.

Quality of Fit

This principal component model with a dynamic baseline has an adjusted R^2 of 0.2429. The higher quality of fit, compared to models using no baseline, can also be seen in figure 5.21, where the PCD_1 estimates are plotted against the corresponding measured values.

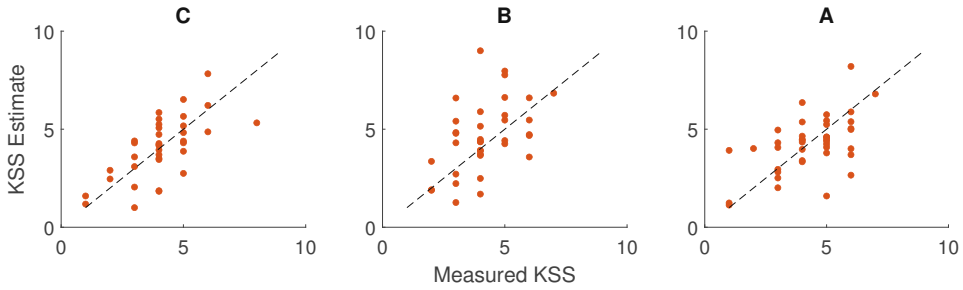

Figure 5.21: KSS values fitted by the model PCD_1 are plotted against the corresponding measured KSS values in each group. The black dotted line indicates where the fitted and measured values are equal.

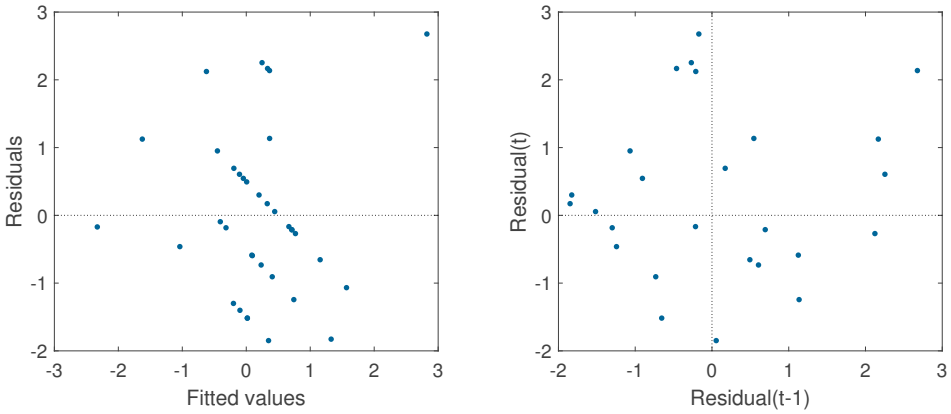

Figure 5.22: Left: Residuals plotted against the values fitted by PCD_1. **Right:** The residuals of the model PCD_1 are plotted against their predecessor in the list of residuals. In both images zero is indicated by black dotted lines.

The data is shown for each condition and time of measurement separately. The lowest RMSE for PCD_1 is 1.18, measured on the training data from condition C. The highest RMSE, 1.68, is measured on data from condition B. The RMSE values for each category are given in table 5.5.

Residual Analysis

Two residual plots are shown in figure 5.22. One shows the residuals plotted against the fitted values, the other shows each residual plotted against its predecessor.

The normality of the residuals is checked using the Anderson-Darling test. The residuals from all three conditions pass the test of normality. The p-values are $p = 0.1724$, $p = 0.3902$ and $p = 0.1396$ for conditions C, B and A, respectively.

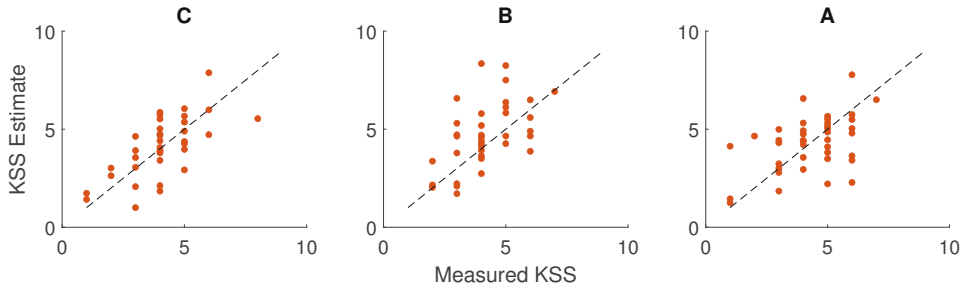

Figure 5.23: KSS values fitted by the model PCD_2 are plotted against the corresponding measured KSS values in each group. The black dotted line indicates where the fitted and measured values are equal.

5.12 PCD_2

The regression model PCD_2 is a principal component regression model using 6 PCs without an intercept. It is given by the formula

$$\mathrm{KSS} = \sum_{i=1}^{6} c_i \cdot \mathrm{PC}_i.$$

The computed variable coefficients c_j, $j = 1, \ldots, 6$ are given in table 5.11.

Assessing Model Accuracy

The model PCD_2 has the F-statistic 2.89, when comparing it to the constant model. The corresponding p-Value, $p = 0.0369$, is under the level of significance. According to the F-test, determining the significance of single variables in the model, only two principal components are statistically significant in the model. The result for the third principal component PC_3 is $p = 0.0294$ and that of the fourth PC_4 is $p = 0.0369$. All other p-values are given with the coefficient estimates in table 5.11.

Quality of Fit

The adjusted R^2 statistic for PCD_2 is 0.2479. Figure 5.23 shows the KSS estimates by PCD_2 plotted against the corresponding measured values. The data is displayed separately for each condition, A, B and C. The RMSE for PCD_2 predictions ranges between 1.18, on data from condition C, to 1.54, on condition B. The RMSE for each data category is given in table 5.5.

Residual Analysis

Figure 5.24 shows the model's residuals plotted against corresponding fitted values. The same figure also shows a second plot, where each residual is plotted against the preceding one.

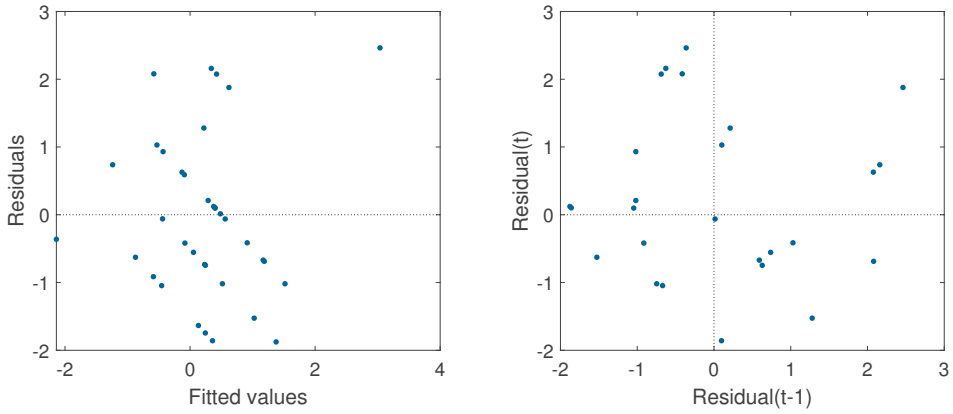

Figure 5.24: Left: Residuals plotted against the values fitted by PCD_2. **Right:** The residuals of the model PCD_2 are plotted against their predecessor in the list of residuals. In both images zero is indicated by black dotted lines.

The normality of residuals is checked using the Anderson-Darling test. According to this test, the residuals from all conditions are normal. The p-values are $p = 0.1605$, $p = 0.4343$ and $p = 0.4864$ for conditions C, B and A, respectively.

5.13 Summary

For analysis and discussion of the results, it can be helpful to see some key indicators of different models side by side. For this purpose this section contains two summary tables.

Table 5.12 shows two of the most important indicators of quality of fit, the root mean square error and the adjusted R^2 statistic. The number of predictors for each model is also included in the table. The RMSE is calculated for two data categories: the training data, which is mostly, but not always, from condition C, and test data, which is comprised of all remaining data, after training data has been removed.

Table 5.13 gives an overview of the variables included in each model. Each variable is indicated by the sign of its coefficient. If the variable was determined to be statistically significant in the corresponding model, it is additionally marked by an asterisk. The table represents an overview of the importance of variables in connection to predicting fatigue, as well as the direction of change with rising fatigue.

The models generated using principal component regression (PCR) are not represented in table 5.13, since, due to the chosen method, all variables contribute to fatigue prediction. For the PCR models, the correlations of the PCs to each variable are given in tables 4.7 and 4.8. The coefficients of the PCs in each of the models are presented in tables 5.10 and 5.11.

Table 5.12: The table shows some key characteristics of quality of fit, the adjusted R^2 statistic and the root mean square error (RMSE) on both training and test data, for each of the models discussed. Additionally, the p-value of the F-test versus a constant model is given for each of the models.

Model	**Adjusted R^2**	**Training RMSE**	**Test RMSE**	**Number of Predictors**	***F*-Test**
SLN_1	0.21	1.04	1.61	15	$p_{HRV} = 0.0097$ $p_{PW} = 0.0496$
SLN_2	0.21	1.11	1.55	10	$p_{HRV} = 0.0053$ $p_{PW} = 0.0274$
SLD_{C1}	0.17	1.19	1.41	5	$p_{HRV} = 0.0035$ $p_{PW} = 0.0143$
SLD_{C2}	0.40	0.94	1.55	12	$p_{HRV} < 0.0001$ $p_{PW} = 0.0214$
SLD_{S1}	0.14	1.30	1.32	1	$p = 0.0035$
SLD_{S2}	0.60	0.76	1.75	11	$p = 0.0002$
SLF_1	0.22	1.18	1.40	7	$p = 0.0003$
SLF_2	0.14	1.14	1.3	8	$p = 0.0037$
PCN_1	0.03	1.33	1.48	6	$p = 0.5726$
PCN_2	0.07	1.26	1.55	8	$p = 0.4924$
PCD_1	0.25	1.19	1.52	5	$p = 0.0235$
PCD_2	0.25	1.18	1.49	6	$p = 0.0369$

Table 5.13: An overview of the variables chosen by each model as well as the sign of the computed coefficient. The sign is marked with an asterisk, if the corresponding variable is statistically significant in the given model.

Variable	SLN_1	SLN_2	SLD_{C1}	SLD_{C2}	SLD_{S1}	SLD_{S2}	SLF_1	SLF_2
age	$-^*$	$-^*$				$+$	$+^*$	$+^*$
height	$-$	$-$				$+$		$-$
weight						$-$		
hr	$-^*$	$-^*$		$-^*$		$-^*$		$-$
lfp	$+$			$-^*$		$-^*$		$-$
hfp				$+^*$		$+^*$		
lhratio	$+$	$+^*$	$+^*$	$+^*$	$+^*$	$+^*$		$-$
rmssd								
sdnn	$-^*$	$-^*$						
pnn50	$-$			$-^*$		$-^*$		
totalpwr	$+$	$+^*$						$+^*$
lfp_n							$+^*$	$+^*$
hfp_n				$-^*$				
t_T	$+^*$	$+^*$						$-^*$
t_notch				$+^*$				
t_sys_rel	$+^*$	$+^*$		$+^*$			$-^*$	
t_notch_rel						$+^*$	$-^*$	
t_dwp_rel			$+$					
t_sys	$-^*$	$-^*$	$-^*$	$-^*$		$-^*$		
P_dwp_sys	$+$		$+$	$+^*$		$+^*$	$+^*$	
P_notch_sys	$-$	$-$		$-^*$			$-$	
P_notch_dwp	$+$			$+^*$			$+^*$	
pat	$-$		$+^*$					

6 Discussion

In this chapter, the models and results described in chapters 4 and 5 will be analysed and compared. First, the importance of the different variables, when modelling fatigue from HRV and pulse wave data, will be discussed. Then, the results obtained in this thesis will be compared to the results from existing research, which were summarised in chapter 2. Additionally, model results will be compared in the context of quality of fit. The main characteristics to be examined are the adjusted R^2 statistic, the root mean square error (RMSE) and the residuals.

6.1 Model Parameters

While the models generated using principal component regression make use of all measured variables, the linear regression models using a stepwise procedure exclude a large number of variables. This section is dedicated to discussing the contribution of the variables to fatigue prediction. Table 5.13, which gives an overview of the selected variables for each model, can be helpful throughout this section.

6.1.1 Accuracy of Estimates

For all but two models, the group of variables chosen is statistically significant, according to the F-test. This means, that all variable coefficient estimates are simultaneously significantly different from zero. The PCR models without a baseline, PCN_1 and PCN_2, are far from passing the F-test. The contributing variables to the principal components with their corresponding coefficients can not be considered to have a significant relationship to the response. The pulse wave parameters of SLN_1 are barely significant and, overall, the model parameters of PCD_1 and PCD_2 are less significant in comparison to those of other stepwise linear regression models.

Hence, for models based on PCR and models using no baseline, apart from SLN_2, the computed coefficients and chosen variable combinations are considerably less reliable than for other models. That being said, the results of SLN_1, PCD_1 and PCD_2 should not be disregarded completely, since the chosen variables are considered to be significant at the level $\alpha = 0.05$. The parameters of PCN models will not be discussed in further course of this subsection.

These results already suggest, that some kind of individual baseline is of great importance to capture the fatigue-related changes in HRV and pulse wave variables. The principal component analysis (PCA), which, in theory, should lead to better results due to dimension reduction, does not seem to have the desired effect. With respect to parameter significance, the PCR models perform considerably worse than comparable stepwise regression models. This could be caused by the complexity of the cardiovascular system and the variety of

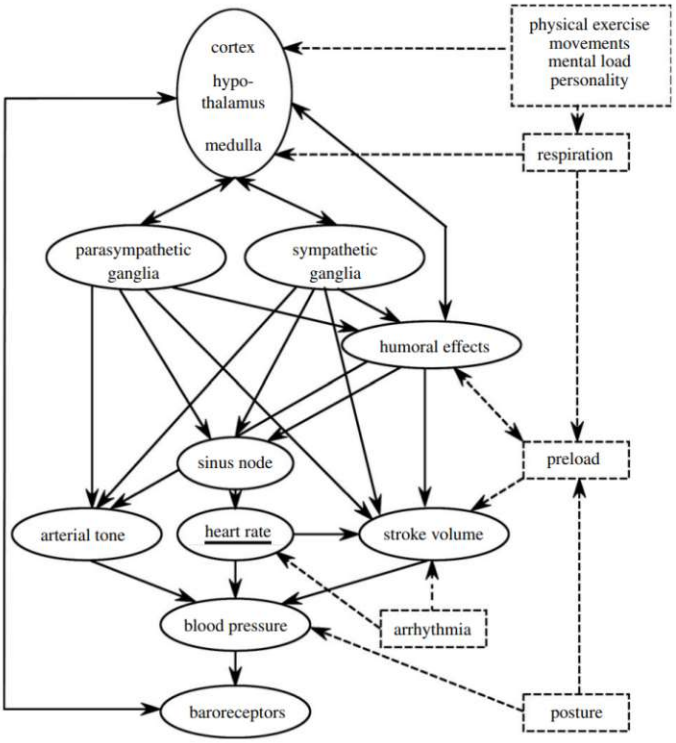

Figure 6.1: The schematic diagram shows a simplified model of heart rate regulation. The arrows indicate dependencies and the dashed boxed show additional factors, that have a considerable influence on the heart rate. [53]

factors regulating the heart rate, a simplification of which is illustrated in figure 6.1. Fatigue is by far not the only factor, that affects HRV and pulse wave parameters. When performing the PCA, the variance of dimensions with the highest variability may be connected to influencing factors other than fatigue, which could explain the lack of significance for most computed PCs.

Regarding the selected HRV parameters, `lhratio` is clearly the most important predictor. The ratio, indicating the balance in the ANS, is chosen in the stepwise variable selection process of seven out of eight models, in five of which it is deemed significant. The heart rate also appears to be valuable predictor. Furthermore, age is also one of the most selected variables, most likely due to dependencies of other parameters on age. Both of these variables is included in five different models, statistically significant in all but one. There are six predictors of moderate importance: `hfp`, `sdnn` and `lfp_n`, which are each selected in two models, `totalpwr` and `pnn50`, which appear in three models and `lfp`, which is included in four models. For each of these variables the coefficients are significant in two different models. The normalised high frequency power `hfp_n` is only selected in one model, where it is deemed significant. Variables of little importance are `height`, `weight` and `rmssd`. Although some of them are included in certain models, the coefficients are not considered to be statistically significant.

In general, the HRV parameters appear to be more sensitive to changes in fatigue than

pulse wave shape parameters. For most models generated from all data simultaneously, a higher number of HRV parameters was selected. The most important variables derived from the pulse wave for fatigue prediction are the absolute and relative time until systolic pressure. Both `t_sys`, which is included in five models, and `t_sys_rel`, which appears in four models, are deemed statistically significant in each of these models. Further important predictors are `P_dwp_sys`, which is selected in five models, in three of which the variable is significant, and `t_T`, which is included in three models with a statistically significant coefficient. `P_notch_dwp` and `P_notch_sys`, which are chosen three and four times, respectively, with only some of the coefficients being significant, can be considered to be less influential in fatigue prediction. The variables `t_notch_rel` and `pat` are each only chosen twice, with opposite signs. Hence, there is no clear contribution to fatigue. Also, `t_notch` and `t_dwp_rel` can be neglected in fatigue prediction.

After PCR with a dynamic baseline (PCD), the third and fourth PC, which explain 14.61% and 8.42% of the variance in the data, are considered statistically significant. The main contributors to the third PC are `hr` with a negative sign, and `t_T`, `rmssd`, `pnn50` and `hfp` with a positive sign. However, for interpretation these signs must be reversed, since both significant PCs have a negative coefficient. The negative contributors to the fourth PC are `lhratio` and `lfp_n`, while the positive contributors are `age`, `hfp` and `hfp_n`. Remember, all main influencing variables for each PC along with their correlation coefficients are given in tables 4.7 and 4.8. Even though the variables with larger influence only partly coincide with those from stepwise regression models, these results emphasise the importance of `hr`, `age` and `lhratio`. It is also worth noting, that the main contributors to the statistically significant PCs do not include any pulse wave parameters, which underlines the importance of HRV in fatigue prediction.

Overall, the finding, that the variables `lhratio` and `hr` are the most important factors in fatigue prediction, while the predictors `lfp_n` and `lfp` can also have a considerable influence, is supported by the data analysis performed in section 4.1. While the correlations between KSS and `totalpwr`, `sdnn`, `t_sys`, `t_sys_rel` or `P_notch_sys`, as seen in table 4.3, are not deemed significant, their low p-values do support their value in predicting fatigue. The data analysis suggests a greater importance of `hfp_n`, than seen in the generated models, and less influence of `hfp` and `pnn50`.

6.1.2 Comparison to Data Analysis and Research

In this section the most influential variables in the generated fatigue prediction models will be discussed with respect to the previously conducted research summarised in chapter 2, especially in table 2.4. When referring to the sign of a variable, this always indicates the coefficient sign, i.e. the direction of change with rising fatigue.

Even though, according to the literature review presented in chapter 2, all measured metadata can have an influence on HRV and pulse wave parameters, this is not seen in the generated regression models. While age is in fact an important parameter, height and weight seem to have a negligible effect.

As for the HRV parameters, the ratio of low and high frequencies $\frac{LF}{HF}$ and the heart rate, which are of great importance in the generated models, are also the most prominent factors in existing research. The rise in the mean RR interval, as noted in table 2.4, can be

equated with a decrease in heart rate. This coincides with the generated stepwise regression models. The PCD models, however, include the heart rate with a positive sign, which is contradictory to previous studies.

The $\frac{LF}{HF}$ ratio is more difficult to interpret, since there are varying results in different studies. While Abtahi et al. [1] measured no significant effect and Rodriguez-Ibañez et al. [43] found a positive change in $\frac{LF}{HF}$, most other studies show a negative effect in $\frac{LF}{HF}$ with rising fatigue. The sign of $\frac{LF}{HF}$ ratio in all SLN and SLD models is positive, which is consistent only with the results of Rodriguez-Ibañez et al. [43]. In SLF_2 the coefficient is negative, which is consistent with the majority of results in existing research. However, in this model the coefficient in question is not significantly different from zero. It is clear, that the $\frac{LF}{HF}$ is of utmost importance for fatigue prediction, but the large discrepancy in results poses challenges in its use as a predictor. As suggested in the standards of measurement for HRV [28], varying results could be the result of varying methods in use to obtain the frequency-domain measures. Since most studies do not clarify the applied method, the exact influence can not be determined. Alternatively, there could be confounding factors that influence the $\frac{LF}{HF}$ ratio in both Rodriguez-Ibañez et al. [43] and the data provided for this thesis, leading to contradictory results in comparison to other research.

The discrepancy between existing research and the generated models does not only concern $\frac{LF}{HF}$, but extends to almost all frequency domain HRV measures. While the low frequency power has been shown to increase with rising fatigue, it is mostly included in models with a negative coefficient. The high frequency power is included with a positive sign. In this case, previous research is inconclusive, showing both statistically significant increase or decrease in HF, which again indicates the difficulties when comparing frequency domain HRV measures from different sources. The results for normalised LF and HF also deviate from those of previous studies. They are included in the generated models with the opposite sign, than could be expected from the literature review in chapter 2. The only frequency domain measure, that shows similar results to existing research, is the total power. As expected, it is included with a positive sign, i.e. the total power increases with rising fatigue.

Even though a negative change in RMSSD has been shown in Mahachandra et al.[26], it was not selected as a variable in a single multivariate model using stepwise variable selection. However, RMSSD is a positively contributing factor to the third principal component of the dynamic baseline models PCD_1 and PCD_2. Taking into account the negative coefficient of PC_3, this is in line with existing research.

It is surprising, that the percentage of NN50, the number of consecutive RR-intervals that differ more than 50 ms, is included in SLD_{C2}, SLD_{S2} and PCD models with a negative coefficient, since Abtahi et al. [1] finds that NN50 increases significantly. Similarly, SDNN has been shown to increase with rising fatigue, but is only included with negative coefficients in stepwise regression models.

The pulse arrival time seems to be considerably less sensitive to changes in fatigue than expected. The study Majumder et al. [1] suggests, that PAT increases with rising fatigue. Nevertheless, PAT is only selected in two different models, with opposite signs. Only the positive coefficient, which is in line with the mentioned study, in SLD_{C2} is statistically significant.

According to Majumder et al. [1], both the crest time, corresponding to `t_sys`, and

the diastolic time, which is calculated as `t_T` − `t_sys`, increase with rising fatigue. Both the total pulse duration and the crest time have been identified as valuable pulse wave parameters for prediction. The crest time is included with a negative coefficient, while the total pulse duration is selected with a positive sign. It stands to reason, that if both crest time and diastolic time increase simultaneously, their sum, the total pulse duration, should also increase. Hence, the variable `t_T` is included as expected in the SLN models, but not in SLF_2 and the principal component analysis with dynamic baseline. As for the crest time, a positive change can be expected. Since this parameter is also essential for computing diastolic time, the change should not surpass the increase in total pulse duration. However, crest time is included with a negative coefficient in all models. This may not coincide with the expected change in crest time itself, but the combined changes in total pulse duration and crest time in the SLN models do coincide with the expected increase in diastolic time.

The models SLN_1, SLN_2 and SLF_2 coincide most with the results found in the literature review of chapter 2. They all include the heart rate and total power as expected. Additionally, both SLN models include the total pulse duration with the expected positive sign and the combination of the coefficients of total pulse duration and systolic time are in line with a rise in diastolic time. Three further models include one variable in the way that research would suggest. These are SLD_{C1}, SLD_{C2} and SLD_{S2}, the first of which includes the pulse arrival time while the others include the heart rate. The models SLD_{S2}, SLF_1 as well as all PCR models are mostly not supported by literature.

6.2 Model Assessment

This section is dedicated to the comparison of the different models, as introduced in chapter 4, with respect to quality of fit. The residuals are checked for unwanted patterns that indicate systematic error, while both the adjusted R^2 statistic and the RMSE are used to judge goodness of fit.

6.2.1 Residual Analysis

Both types of residual plots look acceptable for all models. The plots of residuals versus fitted values do not show any periodic patterns or funnel shapes and therefore to do not indicate systematic error. It is noticeable, that KSS estimates close to the ends of the scale are rare. However, this is most likely to arise from the KSS distribution in the training data, as seen in figure 4.1, rather than systematic error in modelling. The lagged residual plots, in which each residual is plotted against its predecessor, mostly show an even distribution of points around the centre. The lagged residual plot for PCN_1, shown in figure 5.18, shows slight variations in the distribution in each quadrant, but this could still be caused by random effects.

The residuals of the models SLN_1, SLD_{C1}, SLD_{S1}, SLD_{S2} and all principal component models pass the test of normality. Therefore, there is no reason to believe they have systematic error. This also indicates, that a linear model with a dynamic baseline or no baseline is representative of the relationship between the cardiovascular data and fatigue.

For SLN_2 the residuals from recordings before driving in condition B, where measurements were taken the day after drinking, are not normally distributed ($p = 0.0443$). For

the model SLD_{C2}, the residuals of condition C failed the test of normality ($p = 0.0480$). These results could indicate systematic error in SLN_2 and SLD_{C2}. Even though $\alpha = 0.05$ is a well-established level of significance, it is still arbitrary and does not represent a precise boundary. As the residuals of the models SLN_2 and SLD_{C2} only fail the test of normality by a slight margin, these models will not be disregarded completely. Nonetheless, their results should be interpreted with great caution.

Residuals from both models using a fixed baseline, SLF_1 and SLF_2, are likely to come from a non-normal distribution. The normality is violated in condition C and B, after driving, for SLF_1 ($p = 0.0360, p = 0.0431$), and condition B, before and after driving, for SLF_2 ($p = 0.0197, p = 0.0021$). This indicates, that the method using a fixed baseline for each participant leads to systematic error.

6.2.2 Quality of Fit

Since the generated models have a varying number of coefficients, the variance, explained by the different models, can only be compared using the adjusted R^2 statistic, which is summarised along with the RMSE of all models in table 5.12. Even though this is a key indicator of quality of fit, we can not expect any values close to 1. As stated in James et al. [21, Ch. 3], for applications in biology a linear model only constitutes a very rough approximation of the relationship between predictors and response. It is common, that the larger portion of variance in the data is left unexplained by the model. In some situations, even values as low as 0.1 can be considered good. Small values of the adjusted R^2 statistic could be the result of unmeasured factors, that have an effect on fatigue, but are not included as predictors, or factors that influence the HRV and pulse wave parameters, such as stress. In a large interdependent system, such as the cardiovascular system, small adjusted R^2 values must be considered normal.

Since the models are intended for the prediction of fatigue on the nine point KSS scale, ranging from 1 to 9, a root mean square error below 1 would be preferable. Even though a large difference between test and training RMSE clearly indicates a loss of generality and overfitting, a larger difference may be acceptable in this case, since the effects of alcohol, both just after consumption and the day after drinking, are present in most test data and their influence is unclear.

In most cases, for those models that were not based on difference data, predictions are more accurate for data from recordings before driving than for data measured after driving. This is not surprising, as most models were also trained on data recorded before driving, but also supports the assumption, that the driving task itself already has an influence on fatigue and HRV parameters, as the test results in table 4.2 suggested. With some exceptions, most notably the SLD_S models, there is also a clear pattern, that errors in condition B, the day after drinking, are larger than those in condition A, under the influence of alcohol, which in turn, are larger than those in condition C, the training data with no known influence. There are two possible reasons for this, the first simply being that HRV and pulse wave parameters deviate more from the individual baseline one day after drinking than they do while being drunk. Possibly this discrepancy could also be caused by a change in perceived fatigue the day after drinking. Since the KSS scale is a subjective measure, evaluated by each participant for themselves before and after driving, differences could also arise from

over- or underestimated fatigue.

In general, the models without any baseline did not perform as well as those using a baseline. The stepwise regression models SLN_1 and SLN_2 both have an acceptable R^2 value of 0.21 as well as training RMSE slightly above the target of 1. However, the difference between training and test RMSE is rather large, which suggests that the SLN models do not generalise well. Also, the models use 15 and 10 different predictors, respectively, which seems to be more than necessary, since other models achieve similar or better results using less variables. Additionally, figures 5.1 and 5.3 show cleary, that the predictions do not fit the proposed relationship as well as other model options.

The models PCN_1 and PCN_2 have the lowest adjusted R^2, 0.03 and 0.07, respectively. The PCA using no baseline does not adequately capture the variance in the data connected to fatigue. In comparison to other models, the training RMSE of these models are also among the highest. The difference in training and test RMSE is smaller, than with many other models, but this likely due to a higher training RMSE, rather than better generalisation. All in all, both PCN models show a poor quality of fit, which also becomes obvious in the comparison of predictions and recorded KSS values shown in figures 5.17 and 5.19.

The PCR models using a dynamic baseline, the adjusted R^2 of both of which is 0.25, perform better than their counterparts without baseline. While the results of both PCD models are very similar, PCD_2, using more PCs, has a slightly lower error, making it the better fit. The difference of around 0.3 between test and training RMSE is slightly larger than with some other models, but is not large enough to indicate a problem with generality.

Similar to the PCD models, SLF_1 also has a decent goodness of fit. While the adjusted R^2 is only slightly lower and the error on training data is the same, the test RMSE is lower than with the PCD models, indicating a higher generality. The model SLF_2 shows an even smaller difference of only 0.16 in error between test and training data, but also a lower R^2 statistic. However, since the SLF models are the only generated models, where test data does not exclusively include data under the influence of alcohol, these results may not indicate a higher generality of the model, but come from the more heterogeneous grouping of data.

Even though the methods using dynamic baselines are promising, the single models generated from all data simultaneously do not look as promising as other models. SLD_{S1} emphasises the importance of the $\frac{LF}{HF}$-ratio, since it demonstrates that decent results can be achieved using this variable alone. However, in comparison to other models, the test RMSE, 1.3, is high while R^2, 0.14, is comparatively low. While this model has no problems with generality, since test and training RMSE are almost identical, including more variables leads to more accurate predictions. The large difference between RMSE on training and test data and comparatively high R^2 value for SLD_{S2} are a clear sign of overfitting. Hence, using this model for predictions should be discouraged.

The stepwise linear models with a dynamical baseline, that model HRV and pulse wave data separately and then combine the predictions, have promising results. SLD_{C1} has a moderate R^2 value, 0.17, and an RMSE of 1.19 or 1.41, on test and training data respectively. This is comparable with the error of SLF_1 and the training error of the PCD models, but shows a greater generality than the PCD models. The model SLD_{C2} has a high R^2 value, 0.40, and very low RMSE on the training data, 0.94. Even though the difference between test and training error is moderately high, it is not clear, whether this is due to

overfitting. SLD_{C2} achieves a considerably lower training error while keeping the test error on a similar level as other models. It is unclear, whether this model yields meaningful results when applied to conditions other than the baseline condition C. Nevertheless, the low error on data with no known influence is promising. To understand how reliable these results are, the generality of this model should be tested using new data.

All in all, the combined models, using a dynamical baseline (SLD_C), and those models using a fixed baseline (SLF) seem to have the best quality of fit. The model SLD_{C2} achieves the lowest RMSE on data from condition C, while still retaining some generality.

6.3 Summary

First of all, it can not be mentioned often enough, that the data provided for the purpose of generating these regression models is not perfectly suited for the task. Considering the fact, that predicting high KSS values is of most interest in the context of driving, it is unfortunate, that over 90% of all recorded KSS values are below 7. During the entire trial, no participant was tired enough to evaluate themselves at the highest KSS value of 9. This makes it difficult to generate or even test a model, that predicts fatigue accurately at the top end of the scale. Additionally, the training data set is rather small, after removing data influenced by alcohol. Therefore, no data without known influences can be used for model testing, which makes the interpretation of results difficult.

In general, linear regression models appear to be a good choice when modelling fatigue from HRV and pulse wave parameters. However, the principal component analysis and regression did not lead to the expected improvement compared to stepwise linear models, especially with respect to statistical significance. This could be due to the complexity of the cardiovascular system and the multitude of confounding factors, many of which could not be included in the regression data. Due to the large individual differences in HRV and pulse wave data, using a baseline improves model results. In general, the prediction error is highest in condition B, the day after drinking, while it is lowest on data from condition C, on which most models were trained. The most important variables in literature and the regression models generated in this thesis are age, the $\frac{LF}{HF}$-ratio for HRV data, and systolic time and total pulse duration for pulse wave data. However, for the systolic time and most frequency-domain HRV measures, the coefficient signs are the opposite of what could have been expected from previous research. Nonetheless, little research has been conducted on the connection between pulse wave parameters and fatigue and, as some studies demonstrate, there are definitely comparability issues with frequency-domain HRV measures. Contrary to literature, the influence of height, weight, RMSSD and PAT is mostly not seen in the discussed models.

While the SLN models and SLF_2 show the largest similarity to the results in previous research, three other models, SLD_{C1}, SLD_{C2} and SLD_{S2}, at least include one variable as the literature review would suggest. Even though all residual plots look acceptable, the residuals of the models SLN_2, SLD_{C2} and both SLF models fail the test of normality. Therefore their results should be used with caution. The overall best quality of fit, determined by the combination of adjusted R^2 statistic and RMSE on test and training data sets, is seen in the models using a fixed baseline (SLF), which retain a higher generality, and the model

SLD_{C2}, which achieves a low training error.

In the context of predictions, low p-values and RMSE are essential. Therefore, even though SLN_1 seems to score decently in all aspects, it may not be suited for the intended use, since only 6 out of 15 variable coefficients are statistically significant. The combination of all variables is considered significant at the level $\alpha = 0.05$, which is higher than the level of significance achieved by other models. The models SLN_2 and SLD_{C2}, which both have a better statistical significance and quality of fit, are, at least regarding some variables, supported by the literature review. However, both models residuals narrowly fail the test of normality, which could indicate systematic error. The model SLD_{C1} seems to strike a balance, where a decent amount of variance is explained through a small number of variables, while statistical significance, prediction error and the normality of residuals are acceptable. Nevertheless, the results should be interpreted with caution, since this model includes some variables, most notably the $\frac{LF}{HF}$ ratio and systolic time, in a different manner than the majority of previous research.

7 Conclusion

This thesis aimed to quantify the relationship between fatigue and physiological parameters of the cardiovascular system, which are sensitive to changes in fatigue due to their connection to the autonomic nervous system. The main goal was the derivation of a predictive model, that has the ability to accurately predict fatigue on the well-established KSS scale. The model is intended for application in a comprehensive system for assessment of fitness to drive in the commercial context of the EU-funded PANACEA project.

Data from a pilot trial held in Sweden was supplied to train and test statistical models. Key measures of heart rate variability (HRV) in both the time and frequency domain were determined using electrocardiography (ECG), while a number of pulse wave shape parameters and pulse arrival time (PAT) were extracted from photoplethysmography (PPG) measurements. A number of missing values due to lack of signal quality, as well as alcohol as a confounding factor, considerably reduced the number of measurements available for training.

Twelve multivariate linear regression models generated using six different approaches were discussed and evaluated in this thesis. With the goal of reducing the 23 potential predictors to a manageable number, both stepwise variable selection in multivariate linear regression and principal component analysis (PCA) with subsequent linear regression were applied. To account for the considerable individual differences in HRV and pulse wave parameters, two options for an individual baseline, fixed or dynamic, were introduced.

While the linear regression models using stepwise variable selection produced promising results, the principal component analysis (PCA) and subsequent linear regression did not bring the improvement, that could be expected when applying a dimension reduction method to a data set with a large number of different variables. Models using either form of participant baseline generally performed better, than those neglecting the individual variations in physiology.

While, in general, most variables, especially frequency-domain HRV parameters, were not included in the manner the literature review had suggested, the established importance of the $\frac{LF}{HF}$-ratio and heart rate in fatigue prediction is underlined by the generated models. It should be noted, that while a multitude of studies discuss the connection between single HRV measures and fatigue, only one study was found, that evaluates this relationship for pulse wave shape parameters. Furthermore, no previously conducted research on predictive regression models for fatigue based on cardiovascular parameters was found, which makes the results of this thesis difficult to put into context.

Many models, notably SLD_{C2} or SLN_2, that produced good results concerning statistical significance and goodness of fit, have issues with the normality of residuals, which could indicate systematic error. Instead, the further use of the model SLD_{C1} is encouraged, since there are no such issues and it balances a high statistical significance, a decent amount of explained variance (adj. $R^2 = 0.17$) and an acceptable level of generalisation, indicated

by the relation between training and test root mean square error (RMSE) (1.19 and 1.41, respectively). However, it should be mentioned, that the RMSE well above one indicates, that there can be considerable error in KSS prediction. Furthermore, some chosen variables with their respective coefficients are contradictory to previous results in this context.

Even with the restriction of a small data set with avoidable confounding factors, this thesis shows, that the prediction of fatigue on the KSS scale through a regression model using cardiovascular parameters is not only feasible in theory, but also in practice. Of course, the methods and models presented and discussed in this thesis need to be refined, especially by incorporating larger and more heterogeneous data sets, before fatigue assessment for commercial drivers can be used at a large scale. Most importantly, as the trial conducted within the PANACEA Project included only male participants, mainly between the ages of 30 and 50, the models must be generalised and re-evaluated to be applicable to the entire adult population, independently of age or sex.

List of Figures

List of Tables

Bibliography

[1] F. Abtahi, A. Anund, C. Fors, F. Seoane, and K. Lindecrantz. Association of drivers' sleepiness with heart rate variability: A pilot study with drivers on real roads. *EMBEC & NBC 2017*, pages 149–152, 2017.

[2] A. A. Alian and K. H. Shelley. Photoplethysmography. *Best Practice & Research Clinical Anaesthesiology*, 28(4):395–406, 2014.

[3] A. P. Avolio, M. Butlin, and A. Walsh. Arterial blood pressure measurement and pulse wave analysis—their role in enhancing cardiovascular assessment. *Physiological measurement*, 31(1):R1, 2009.

[4] M. Awais, N. Badruddin, and M. Drieberg. A non-invasive approach to detect drowsiness in a monotonous driving environment. *TENCON 2014-2014 IEEE Region 10 Conference*, pages 1–4, 2014.

[5] M. Bachler, W. Sehnert, I. Mikisek, S. Wassertheurer, and T. Mengden. Non-invasive quantification of the effect of device-guided slow breathing with direct feedback to the patient to reduce blood pressure. *Physiological Measurement*, 41(10):104002, 2020.

[6] V. Bagdonavičius, J. Kruopis, and M. S. Nikulin. *Nonparametric tests for complete data*. John Wiley & Sons, 2011.

[7] T. Becker, R. Herrmann, V. Sandor, D. Schäfer, and U. Wellisch. *Stochastische Risikomodellierung und statistische Methoden*. Springer, 2016.

[8] A. Burlacu, C. Brinza, A. Brezulianu, and A. Covic. Accurate and early detection of sleepiness, fatigue and stress levels in drivers through heart rate variability parameters: a systematic review. *Reviews in cardiovascular medicine*, 22(3):845–852, 2021.

[9] D. Chambers, C. Huang, and G. Matthews. *Basic physiology for anaesthetists*. Cambridge University Press, 2019.

[10] M. Chappell and S. Payne. *Physiology for Engineers*. Springer, 2020.

[11] J. Choi and R. Gutierrez-Osuna. Using heart rate monitors to detect mental stress. *2009 Sixth International Workshop on Wearable and Implantable Body Sensor Networks*, pages 219–223, 2009.

[12] P. Chorherr. *Development and Evaluation of an Arterial Pulse Wave Form Analysis Algorithm*. Diploma thesis, Technische Universität Wien, 2019.

[13] Creative Commons. Attribution 4.0 international (cc by 4.0). `https://creativecommons.org/licenses/by/4.0/`. Accessed: 22.12.2022.

[14] M. Doudou, A. Bouabdallah, and V. Berge-Cherfaoui. Driver Drowsiness Measurement Technologies: Current Research, Market Solutions, and Challenges. *International Journal of Intelligent Transportation Systems Research*, 18:297–319, 2020.

[15] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas. Human emotion recognition: Review of sensors and methods. *Sensors*, 20(3):592, 2020.

[16] M. J. Fisher and A. P. Marshall. Understanding descriptive statistics. *Australian critical care*, 22(2):93–97, 2009.

[17] D. S. Fonseca, A. Netto, R. B. Ferreira, and A. M. De Sa. Lomb-scargle periodogram applied to heart rate variability study. *2013 ISSNIP Biosignals and Biorobotics Conference: Biosignals and Robotics for Better and Safer Living (BRC)*, pages 1–4, 2013.

[18] G. D. Furman, A. Baharav, C. Cahan, and S. Akselrod. Early detection of falling asleep at the wheel: A heart rate variability approach. *2008 Computers in Cardiology*, pages 1109–1112, 2008.

[19] M. Gonçalves, R. Amici, R. Lucas, T. Åkerstedt, F. Cirignotta, J. Horne, D. Léger, W. T. McNicholas, M. Partinen, J. Téran-Santos, et al. Sleepiness at the wheel across europe: a survey of 19 countries. *Journal of sleep research*, 24(3):242–253, 2015.

[20] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009.

[21] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013.

[22] D. Korpas, J. Halek, and L. Doležal. Parameters describing the pulse wave. *Physiological research*, 58(4), 2009.

[23] E. L. Lehmann and J. P. Romano. *Testing Statistical Hypothesis*. Springer, 2005.

[24] M. F. Lutfi and M. Y. Sukkar. The effect of gender on heart rate variability in asthmatic and normal healthy adults. *International journal of health sciences*, 5(2):146, 2011.

[25] M. F. Lutfi, M. Y. Sukkar, et al. Relationship of height, weight and body mass index to heart rate variability. *Sudan Med J*, 47(1):14–19, 2011.

[26] M. Mahachandra, I. Z. Sutalaksana, K. Suryadi, et al. Sensitivity of heart rate variability as indicator of driver sleepiness. *2012 Southeast Asian Network of Ergonomics Societies Conference (SEANES)*, pages 1–6, 2012.

[27] S. Majumder, A. K. Verma, C. Wang, A. Mohamud, L. Archer, K. Tavakolian, and N. Wilson. Using photoplethysmography based features as indicators of drowsiness: Preliminary results. *Frontiers in Biomedical Devices*, 41037:V001T09A008, 2019.

[28] M. Malik. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use: Task force of the european society of cardiology and the north american society for pacing and electrophysiology. *Annals of Noninvasive Electrocardiology*, 1(2):151–181, 1996.

[29] T. Mengden, M. Bachler, W. Sehnert, P. Marschall, and S. Wassertheurer. Device-guided slow breathing with direct biofeedback of pulse wave velocity–acute effects on pulse arrival time and self-measured blood pressure. *Blood Pressure Monitoring*, 28(1):52–58, 2023.

[30] A. Miley, G. Kecklund, and T. Åkerstedt. Comparing two versions of the Karolinska sleepiness scale (KSS). *Sleep and Biological Rhythms*, 14:257–260, 2016.

[31] N. R. F. C. (NCD-RisC). A century of trends in adult human height. *eLife*, 5:e13410, jul 2016.

[32] A. Ohly and M. Kiening. *EKG Endlich Verständlich*. Elsevier Health Sciences, 2022.

[33] OpenStax College. Openstax college. `https://openstax.org/`. Accessed: 28.01.2023.

[34] M. F. O'Rourke, A. Pauca, and X.-J. Jiang. Pulse wave analysis. *British journal of clinical pharmacology*, 51(6):507, 2001.

[35] PANACEA Project. PANACEA Project homepage. `https://panacea-project.eu/`. Accessed: 14.09.2022.

[36] M. Patel, S. K. Lal, D. Kavanagh, and P. Rossiter. Applying neural network analysis on heart rate variability data to assess driver fatigue. *Expert systems with Applications*, 38(6):7235–7242, 2011.

[37] A. Persson, H. Jonasson, I. Fredriksson, U. Wiklund, and C. Ahlström. Heart rate variability for classification of alert versus sleep deprived drivers in real road driving conditions. *IEEE Transactions on Intelligent Transportation Systems*, 22(6):3316–3325, 2021.

[38] P. Philip, P. Sagaspe, J. Taillard, C. Valtat, N. Moore, T. Åkerstedt, A. Charles, and B. Bioulac. Fatigue, sleepiness, and performance in simulated versus real driving conditions. *Sleep*, 28(12):1511–1516, 2005.

[39] S. Rajala, H. Lindholm, and T. Taipalus. Comparison of photoplethysmogram measured from wrist and finger and the effect of measurement location on pulse arrival time. *Physiological measurement*, 39(7):075010, 2018.

[40] T. Åkerstedt, A. Anund, J. Axelsson, and G. Kecklund. Subjective sleepiness is a sensitive indicator of insufficient sleep and impaired waking function. *Journal of Sleep Research*, 23(3):240–252, 2014.

[41] T. Åkerstedt and M. Gillberg. Subjective and Objective Sleepiness in the Active Individual. *International Journal of Neuroscience*, 52(1-2):29–37, 1990.

[42] N. M. Razali and Y. B. Wah. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, 2(1):21–33, 2011.

[43] N. Rodriguez-Ibañez, M. A. García-Gonzalez, M. A. F. de la Cruz, M. Fernández-Chimeno, and J. Ramos-Castro. Changes in heart rate variability indexes due to drowsiness in professional drivers measured in a real environment. *2012 Computing in Cardiology*, pages 913–916, 2012.

[44] M. Romanowicz, J. E. Schmidt, J. M. Bostwick, D. A. Mrazek, and V. M. Karpyak. Changes in heart rate variability associated with acute alcohol consumption: current knowledge and implications for practice and research. *Alcoholism: Clinical and Experimental Research*, 35(6):1092–1105, 2011.

[45] A. Sahayadhas, K. Sundaraj, and M. Murugappan. Drowsiness detection during different times of day using multiple features. *Australasian physical & engineering sciences in medicine*, 36(2):243–250, 2013.

[46] F. Shaffer and J. P. Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in public health*, page 258, 2017.

[47] J. Solà and R. Delgado-Gonzalo. *The Handbook of Cuffless Blood Pressure Monitoring*. Springer, 2019.

[48] T. Tamura and W. Chen. *Seamless Healthcare Monitoring*. Springer, 2018.

[49] T. Tamura, Y. Maeda, M. Sekine, and M. Yoshida. Wearable photoplethysmographic sensors—past and present. *Electronics*, 3(2):282–302, 2014.

[50] The Mathworks, Inc. MATLAB statistics and machine learning toolbox user's guide. `https://de.mathworks.com/help/pdf_doc/stats/stats.pdf`. Accessed: 03.02.2023.

[51] J. T. VanderPlas. Understanding the lomb–scargle periodogram. *The Astrophysical Journal Supplement Series*, 236(1):16, 2018.

[52] K. J. F. Verhoeven, K. L. Simonsen, and L. M. McIntyre. Implementing false discovery rate control: increasing your power. *OIKOS*, 108(3):643–647, 2005.

[53] A. Voss, S. Schulz, R. Schroeder, M. Baumert, and P. Caminal. Methods derived from nonlinear dynamics for analysing heart rate variability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367:277–296, 2009.

[54] E. Whitley and J. Ball. Statistics revie 6: Nonparametric methods. *Critical Care*, 6(6):509–513, 2002.

[55] A. Williamson, D. A. Lombardi, S. Folkard, J. Stutts, T. K. Courtney, and J. L. Connor. The link between fatigue and safety. *Accident Analysis and Prevention*, 43:498–515, 2011.

[56] C. Zhao, M. Zhao, J. Liu, and C. Zheng. Electroencephalogram and electrocardiograph assessment of mental fatigue in a driving simulator. *Accident Analysis & Prevention*, 45:83–90, 2012.

[57] Österreichische Forschungsförderungsgesellschaft (FFG). Horizon 2020. `https://www.ffg.at/europa/h2020`. Accessed: 15.09.2022.