Comparison of three ordinal logistic regression methods for predicting person's self-assessed health status with functional, haemodynamic covariates

Master's thesis in statistics

Merri Markkanen
Department of Mathematics and Statistics
University of Jyväskylä
July, 2023

UNIVERSITY OF JYVÄSKYLÄ
Department of Mathematics and Statistics
Name: *Comparison of three ordinal logistic regression methods for predicting person's self-assessed health status with functional, haemodynamic covariates*
Master's thesis in statistics, 39 pages, 19 appendices (58 pages)
July, 2023

Abstract

Surveys, as well as ordinal variable analysis, are commonly used in the medical field. The development of modern technology has also resulted in the development of measurement techniques and the rise of functional data in medicine. It means that functional data is becoming more often used in the analysis of ordinal variables, such as modeling ordinal variables with functional covariates.

In this pro gradu -thesis, three strategies for modeling ordinal variables with functional variables are compared. This is carried out by fitting the models to data obtained from Tampere University Hospital's haemodynamic research group and modeling people's self-assessed health state using functional and haemodynamic variables.

Compared models will be he proportional odds model, partial proportional odds model, and functional ordinal logistic regression model. With the first two models, principal component analysis is applied to haemodynamic variables, and their functionality is ignored. The R program is used for model fitting and analysis.

Based on the results, the partial proportional odds model appears to be best fit for the data, because it does not have as strict assumptions as other models. Worst fit seems to be functional ordinal logistic regression model, which is newer model than others and it seems that it needs more developing, for example in the case of choosing of covariates.

Keywords: ordinal logistic regression, functional data, principal component analysis, proportional odds model, partial proportional odds model, functional ordinal logistic regression model

JYVÄSKYLÄN YLIOPISTO
Matematiikan ja tilastotieteen laitos
Nimi: *Comparison of three ordinal logistic regression methods for predicting person's self-assessed health status with functional, haemodynamic covariates*
Tilastotieteen Pro gradu -tutkielma, 39 sivua, 19 liitettä (58 sivua)
Heinäkuu, 2023

Tiivistelmä

Lääketieteen parissa perinteiset kyselytutkimukset ovat yhä suosittuja, jonka myötä myös järjestysasteikollisten muuttujien analyysia suoritetaan paljon. Modernin teknologian kehittyminen näkyy kuitenkin myös tällä tieteensaralla, kun mittaustekniikoiden kehittyessä funktionaalisen datan määrä on kasvanut. Tämän myötä järjestysasteikollisten muuttujienkin analyysissa yhä useammin hyödynnetään funktionaalista dataa, mm. järjestysasteikollisen muuttujan mallintamisessa.

Tässä pro gradu-tutkielmassa halutaan vertailla kolmea erilaista menetelmää järjestysasteikollisen muuttujan mallintamiseksi funktionaalisten muuttujien avulla. Aineistona käytetään Tampereen yliopistollisen sairaalan hemodynamiikan tutkimusryhmältä saatua aineistoa, josta halutaan mallintaa henkilön itse määrittelemä terveydentila tutkimuksessa mitattujen hemodynaamisten ja funktionaalisten muuttujien avulla.

Sovitettaviksi malleiksi ollaan valittu kumulatiivinen logistinen regressiomalli verrannollisuusoletuksella, osittainen kumulatiivinen logistinen regressiomalli verrannollisuusoletuksella ja funktionaalinen ordinaalinen logistinen regressiomalli. Kahden ensimmäisen mallin kohdalla kovariaattien funktionaalisuus sivuutetaan suorittamalla funktionaalisille muuttujille pääkomponenttianalyysi. Mallien sovitus ja analyysien toteutus tehdään R-ohjelmalla.

Saatujen tulosten perusteella parhaiten aineistoon sopii osittainen kumulatiivinen logistinen regressiomalli johtuen siitä, että sen sisältämät oletukset eivät ole yhtä tiukat kuin kahdella muulla mallilla. Malleista huonoiten aineistoon tulosten perusteella istuu funktionaalinen ordinaalinen logistinen regressio, joka on malleista tuorein ja näyttää vaativan vielä myös kehitystyötä, esim. kovariaattien valinnan suhteen.

Avainsanat: ordinaalinen logistinen regressio, funktionaalinen data, pääkomponenttianalyysi, kumulatiivinen logistinen regressiomalli verrannollisuusoletuksella, osittainen kumulatiivinen logistinen regressiomalli verrannollisuusoletuksella, funktionaalinen ordinaalinen logistinen regressiomalli

Contents

1 Introduction

Ordinal variables are common in many fields of research, including health and social science, where they are often used on surveys. Several methods to analyze ordinal data have been created, and more are being developed. One area of development is analyzing ordinal data when functional data is included. Functional data, as the name implies, is data in the form of a continuous function that is often defined across time. With the advancement of modern technology, functional data has grown more prevalent, and functional data analysis has become a growing field of modern statistics (Mateu and Giraldo, 2022).

In this thesis, the interest is in the case where functional covariates are used to model ordinal variables. There are various approaches to this, including functional and non-functional methods, and some of them have been chosen for comparisons in this thesis. The first model chosen for this thesis is the proportional odds model, in which principal component analysis is done on the functional variables and the corresponding principal component scores are used as covariates in the model instead of using original values.

The proportional odds model is one of the most well-known and widely used ordinal logistic regression models, although it is not always the best fit because it requires a fairly stringent proportional odds assumption to hold true, which often does not with real-world data (Liu, 2022). This is why the second model to be looked at is the partial proportional odds model, which is similar to the proportional odds model but has more relaxed assumptions, and is usually used when the proportional odds assumption does not hold given data.

The functionality of the haemodynamic variables is ignored by using principal component analysis in these two models, but we wanted to present a model that does pay attention to this functionality. Functional ordinal logistic regression, a relatively new model, was chosen for this thesis. Because this model has the same proportional odds assumption as the proportional odds model, the name functional proportional odds model would be more appropriate. The only difference is that this model uses functional variables as they are.

The three methods above are compared by fitting them to data collected by the Tampere University Hospital's haemodynamic research group, which includes the results of various haemodynamic studies. The primary focus of this thesis is on the initial survey question about the subjects' health status and how they responded to it. This ordinal variable will be modeled using functional and haemodynamic variables. We will also examine if these covariates are required for the model, or if BMI, age, and sex are sufficient.

The data used in this thesis are discussed in further depth in the following chapter. Following that, the models to be compared are described in greater detail, as are some of the goodness-of-fit tests that are used and principal component analysis. The results of fitting the models, as well as comparisons between them, are given in chapter 4. The thesis concludes with some discussion on the results.

It should be noted that AI writing tools Wordtune and Quillbot were used to

help with the grammar and paraphrasing of the text.

2 Data

In this section, we describe the data used in the thesis, the research hypothesis, and the response variable and covariates used in the model in greater detail. Let us begin by presenting the data obtained from the Tampere University Hospital's haemodynamics research group. Data were collected during several studies conducted between 2006 and 2021. Only data of healthy participants (751, out of which 353 were women and 368 men), who were in control groups, were chosen for this thesis. The studies themselves included a pre-study doctor visit and successive haemodynamic measurements. Values from 854 variables were collected from the subjects, but only a small number of these were chosen for the thesis.

For this thesis, we chose the subject's health status as the response variable. This variable has four possible values: poor, moderate, good, and excellent. In this thesis, we aim to use ordinal regression to model the probability of research subjects belonging to these various groups. The research question is whether the probability can be modeled using haemodynamic variables, or whether gender, age, and body mass index (BMI) are sufficient covariates on their own. We intend to answer the research question by using ordinal regression, specifically the proportional odds model and various methods derived from it, as detailed in Chapter 3. Covariates will be discussed further in the Section 2.2. But first, we introduce the response variable in greater detail.

2.1 Response

Before taking the haemodynamic measurements, study subjects went through a thorough medical examination. In addition to the basic measurements (weight, blood pressure etc.), the participants had an electrocardiogram done and blood and urine samples were taken. The subjects also completed an initial questionnaire that inquired about their health and lifestyle. One of the initial survey questions asked participants to describe their health status on a scale of poor, moderate, good, and excellent. This variable will be used as a response variable in this thesis.

Despite the fact that there were four possible answers to the question, the subjects in this thesis were divided into three groups based on how they responded to the initial questionnaire question about their subjective health. Because only 14 people considered their health status as poor, they were combined into one group with those who answered "moderate" to the question. Otherwise, the groups remained unchanged, and the groups and their sizes are listed in the Table 1.

As mentioned earlier in Section 2, the research question is whether we can use haemodynamic variables too, or whether gender, age, and BMI are sufficient covariates on their own. These three variables were chosen for comparisons because they are thought to person's self-assessed health status. Self-assessed health status is plotted against the variables age and BMI in Figure 1. The image clearly shows that increasing BMI and age have an effect on how a person answers the question. In contrast to age and BMI for sex, no such correlation can be seen in Figure

Table 1: Number of people in each health status group.

Group	*Size*
Poor/moderate	173
Good	452
Excellent	138

2, where subjective health is plotted against sex. However, we are interested in the interaction of sex with haemodynamic variables, which is why sex is included in the model. The haemodynamic variables used in modeling will be discussed further in the following section.

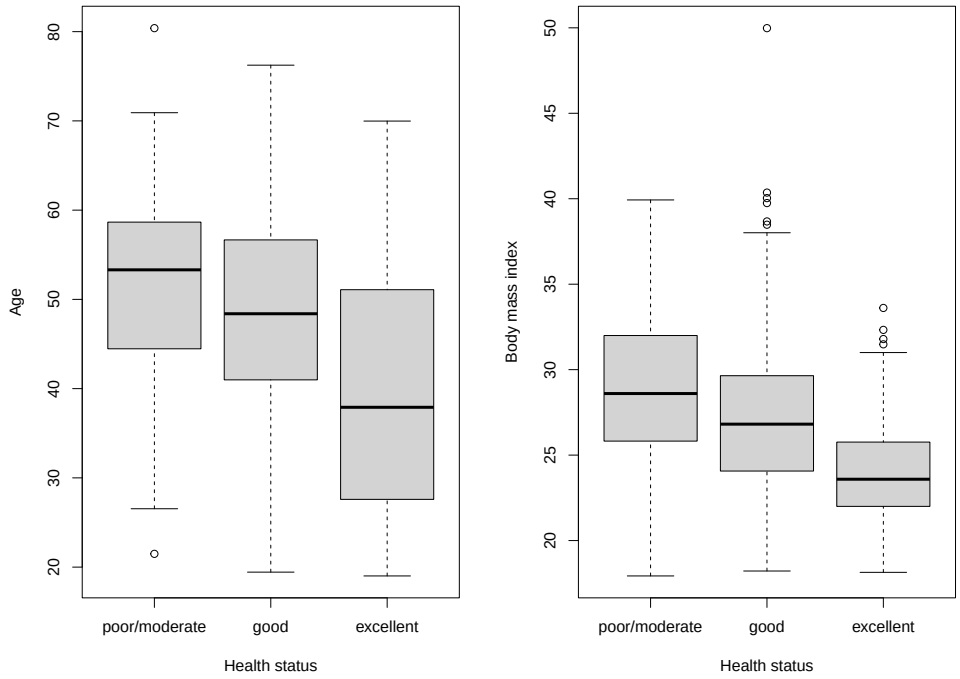


Figure 1: Distributions of age and body mass index for each of the three health status groups.

2.2 Haemodynamic measurements and covariates

Tampere University Hospital's actual study consisted of haemodynamic measurements performed on subjects, the main interest being, how subjects' blood flow was affected by elevation to the upright position. This section goes into greater detail about these measurements as well as haemodynamic covariates used in the model. Although the data come from several studies, the haemodynamic mea-

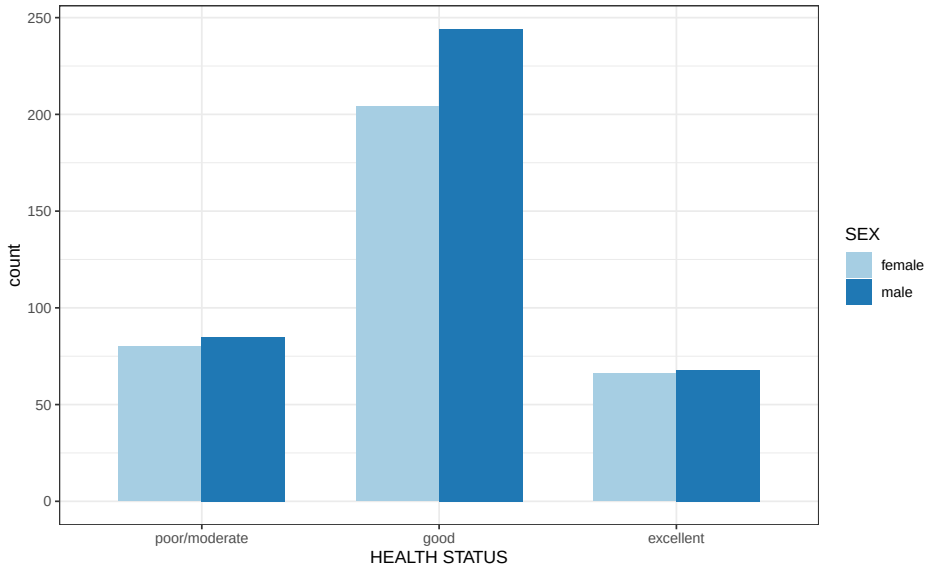


Figure 2: Number of females and males in each of the three health status groups.

surements were all performed using the same protocol, which will be explained in Tahvanainen (2011).

Prior to the measurements, subjects were instructed to refrain from consuming caffeinated products, smoking, or eating a heavy meal for four hours. They also had to abstain from alcohol for 24 hours prior to the experiment. The subjects were placed on a tilt-table with various measuring instruments attached to them for the duration of the study. The measurements were performed by a trained research nurse, and all measurements were completed in a noninvasive manner. The measurement itself took place over a 15-minute period, with the test subject lying horizontally on the tilt-table for the first five minutes. The tilt-table was then raised to a 60-degree angle for five minutes before being returned to a horizontal position for the final five minutes. This was repeated once more to collect the values for comparison. Measurements were taken over the course of several days. This thesis only uses the results of the first ten minutes of the first measurement, i.e, lying and tilted at an angle of 60 degrees.

In Figure 3, there are plotted means of six haemodynamic variables grouped by health status, and one can see how tilting the tilt-table to a 60-degree angle affected them. For example, the average heart rate for all groups starts to rise after the tilt. In the same graph, you can see that the group with health status "good" had higher average value than "excellent"-group before the tilt. However, averages are much closer to each other after the tilt. Group with health status "poor/moderate" average value is higher than other groups both before and after the tilt. Same kind of differences are shown in graphs of other haemodynamic variables, and it seems haemodynamic variables seem to differ by group. However, we cannot be sure that these effects are not just simply due to age and BMI, and that is why we want to examine this more thoroughly.

Table 2: haemodynamic variables that are included in the thesis (abbreviations) and their explanations

Abbr.	*Explanation*
AIX	Augmentation index, which is defined as ratio between augmentation pressure and pulse pressure
SVRI RAD	Systemic vascular resistance from radial arterial pressure. Defined as systemic mean arterial blood pressure minus right arterial blood pressure divided by cardiac output, which is still divided by body surface area
HR CM	Heart rate from CircMon-device
PWV	Pulse wave velocity. In practice, defined as a distance between two measurement sites divided by the traveling time of the pulsewave between them
RAD SAP	Radial systolic arterial pressure
RAD DAP	Radial diastolic arterial pressure
SEVR	Subendocardial viability ratio, which is an index of myocardial oxygen supply and demand
CI	Cardiac index, which is defined as cardiac output divided by body surface area
ECW	Extracellular water volume

The variables' values were continuously measured throughout the measurement, meaning that they are functional variables. However, the values used in this thesis are averages calculated for minute intervals. As a result, the majority of the haemodynamic variables in the data have ten values, e.g, arterial pressures and heart rate. In Figure 4 you can see radial systolic arterial pressure drawn for each subject as continuous variable (line) and minute intervals are marked there with dotten lines. Some variables were only measured when lying down, and the data set contains five values for these variables, e.g, pulse wave velocity. In this thesis, we did not include all measured variables, and the covariates chosen for the thesis are listed in Table 2.

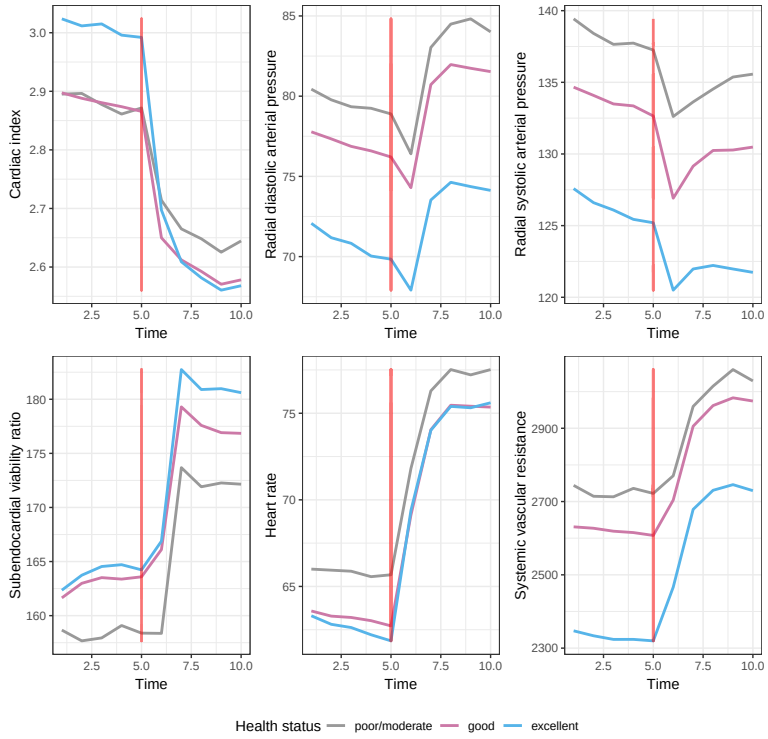


Figure 3: Group-wise mean curves of cardiac index (CI), radial diastolics arterial pressure (RAD_DAP), radial systolic arterial pressure (RAD_SAP), subendocardial viability ratio (SEVR), heart rate (HR_CM) and systemic vascular resistance from radial arterial pressure (SVRI_RAD). Black vertical line indicates the time when the tilt-table was tilted to a 60 degrees angle.

3 Methods

3.1 Proportional odds model

In this thesis, we want to build a regression model for an ordinal response variable. The regression models that are most frequently used, including linear and logistic models, may only be used with response variables that are either continuous or binary. For the case of ordinal responses, there is a family of ordinal regression models, with the proportional odds model being the most well-known member.
The proportional odds model is an extension of binary logistic regression, as are many other ordinal regression models. We use Agresti (2010) as the primary source for this chapter, so the cumulative logit models are introduced first. Following that, the assumptions underlying the proportional odds model are described.

3.1.1 Cumulative logit model

Earlier, we stated that the proportional odds model is an extension of binary logistic regression, which is also the case with the cumulative logit model. Recall

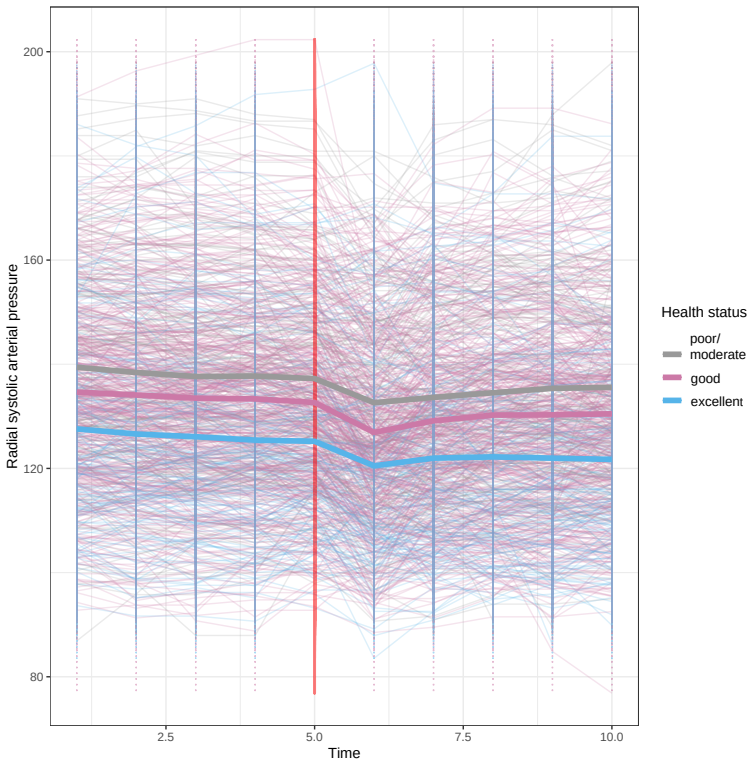


Figure 4: Values of radial systolic arterial pressure (RAD_SAP) plotted for each subject with group means of health status.

that the binary logit is defined as

$$\text{logit}[P(Y_i = 1)] = \log \frac{P(Y_i = 1)}{1 - P(Y_i = 1)},$$

where Y_i is a random variable for subject i taking values either 0 or 1.

Assume now that we have n independent subjects and let Y_i, $i = 1, \ldots, n$. Let us denote values of Y_i as $1, \ldots, k$, where 1 is the lowest group and k is the highest group. To determine cumulative logit for category $j \in 1, \ldots, k$, we use cumulative probabilities. In this instance, we have two possible outcomes: "Y_i belongs to group j or lower group" and "Y_i belongs to a group that is higher than j". The binary variable we now have can be used to obtain binary logit. With the exception of the highest group, we can repeat this process for each category of the response variable. We thus obtain $k - 1$ cumulative logits and they are defined as

$$\text{logit}[P(Y_i \leq j)] = \log \frac{P(Y_i \leq j)}{1 - P(Y_i \leq j)}, \tag{1}$$

where $j = 1, \ldots, k - 1$. Let us now define a cumulative logit model. Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ denote a p-variate covariate vector for subject i. Then write cumulative logit model as follows

$$\text{logit}[P(Y_i \leq j \mid \boldsymbol{x}_i)] = \alpha_j + \boldsymbol{\beta}_j' \boldsymbol{x}_i. \tag{2}$$

In this model, $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp})'$ is p-vector which contains coefficients of covariates for category j. Parameter α_j is the intercept related to category j.

Another way to write this model is using probabilities, that is,

$$P(Y_i \leq j \mid \boldsymbol{x}_i) = \frac{\exp(\alpha_j + \boldsymbol{\beta}_j' \boldsymbol{x}_i)}{1 + \exp(\alpha_j + \boldsymbol{\beta}_j' \boldsymbol{x}_i)}. \tag{3}$$

It is important to note that this model does not compare category j to other categories, but rather compares category j and categories lower than j to categories higher than j. For example, if we have a model with only an intercept α_j, then $P(Y_i \leq j \mid \boldsymbol{x}_i)/(1 - P(Y_i \leq j \mid \boldsymbol{x}_i)) = \exp(\alpha_j)$. In other words, the odds that subject i belongs to a category equal to or lower than j are $\exp(\alpha_j)$-times the odds that subject i belongs to a category higher than j.

In case of interpreting of cumulative logit model's odds ratios, assume we have two subjects 1 and 2, who have two different sets of values of covariates, $\boldsymbol{x}_i^{(1)}$ and $\boldsymbol{x}_i^{(2)}$. Difference between their cumulative logits is

$$\begin{aligned}
\text{logit}\, P(Y_i \leq j \mid \boldsymbol{x}_i^{(1)} &= (x_1^{(1)}, \ldots, x_p^{(1)})') - \text{logit}\, P(Y_i \leq j \mid \boldsymbol{x}_i^{(2)} = (x_1^{(2)}, \ldots, x_p^{(2)})') \\
&= (\alpha_j - \boldsymbol{\beta}_j' \boldsymbol{x}_i^{(1)}) - (\alpha_j - \boldsymbol{\beta}_j' \boldsymbol{x}_i^{(2)}) \\
&= \boldsymbol{\beta}_j' (\boldsymbol{x}_i^{(2)} - \boldsymbol{x}_i^{(1)}).
\end{aligned}$$

Let us calculate the odds ratio from earlier equation:

$$\log \frac{P(Y_i \leq j \mid \boldsymbol{x}_i^{(1)})}{1 - P(Y_i \leq j \mid \boldsymbol{x}_i^{(1)})} - \log \frac{P(Y \leq j \mid \boldsymbol{x}_i^{(2)})}{1 - P(Y \leq j \mid \boldsymbol{x}_i^{(2)})} = \boldsymbol{\beta}_j' (\boldsymbol{x}_i^{(2)} - \boldsymbol{x}_i^{(1)})$$

This is same as

$$\log \frac{P(Y_i \leq j \mid \boldsymbol{x}_i^{(1)})/(1 - P(Y_i \leq j \mid \boldsymbol{x}_i^{(1)}))}{P(Y_i \leq j \mid \boldsymbol{x}_i^{(2)})/(1 - P(Y_i \leq j \mid \boldsymbol{x}_i^{(2)}))} = \boldsymbol{\beta}_j' (\boldsymbol{x}_i^{(2)} - \boldsymbol{x}_i^{(1)})$$

Let us now remove the log from equation so we get odds ratio

$$\frac{P(Y_i \leq j \mid \boldsymbol{x}_i^{(1)})/(1 - P(Y_i \leq j \mid \boldsymbol{x}_i^{(1)}))}{P(Y_i \leq j \mid \boldsymbol{x}_i^{(2)})/(1 - P(Y_i \leq j \mid \boldsymbol{x}_i^{(2)}))} = \exp(\boldsymbol{\beta}_j' (\boldsymbol{x}_i^{(2)} - \boldsymbol{x}_i^{(1)})).$$

We can interpret odds ratio in this case the following way: The odds that subject 1 belongs to a category equal or lower than j, is $\exp(\boldsymbol{\beta}_j' (\boldsymbol{x}_i^{(2)} - \boldsymbol{x}_i^{(1)})$ times the odds that subject 2 belongs to a category equal or lower than j.

If we have two different sets of covariate values, $\boldsymbol{x}_i^{(1)}$ and $\boldsymbol{x}_i^{(2)}$, and the only difference between these two sets is one covariate x_{il}. The value of $\boldsymbol{x}_i^{(1)}$ is denoted as $x_{il}+1$ and differs by one from the value of $\boldsymbol{x}_i^{(2)}$, which is denoted as x_{il}. The odds ratio is $\exp(\beta_{jl})$.

If we want to predict probability that Y_i belongs to specific category, cumulative probabilities are used in this case to describe the probability of a single category in the following manner

$$\begin{aligned} P(Y_i = j \mid \boldsymbol{x}_i) &= P(Y_i \leq j \mid \boldsymbol{x}_i) - P(Y_i \leq j-1 \mid \boldsymbol{x}_i) \\ &= \frac{\exp(\alpha_j + \boldsymbol{\beta}_j' \boldsymbol{x}_i)}{1+\exp(\alpha_j + \boldsymbol{\beta}_j' \boldsymbol{x}_i)} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}_{j-1}' \boldsymbol{x}_i)}{1+\exp(\alpha_{j-1} + \boldsymbol{\beta}_{j-1}' \boldsymbol{x}_i)}. \end{aligned}$$

Above we defined a generalized cumulative logit model, where each category has its own intercept and coefficient vector. However, a more simplified version of this is usually used, where we assume that the different categories all have the same coefficient vector $\boldsymbol{\beta}$, i.e. $\boldsymbol{\beta}_1 = \ldots = \boldsymbol{\beta}_{k-1}$. This model is known as the proportional odds model and i is written as

$$\text{logit}[P(Y_i \leq j \mid \boldsymbol{x}_i)] = \alpha_j - \boldsymbol{\beta}' \boldsymbol{x}_i. \tag{4}$$

Like earlier mentioned, the only difference to model in (2) is $\boldsymbol{\beta}$, which is the same for each category's model. The reason for this is better explained in the next section. Because α_j is the sole parameter that differs across the models in each category, α_j must be equal to or greater than α_{j-1}, because the model uses cumulative probability. It is also worth noting that $\boldsymbol{\beta}'\boldsymbol{x}$ is subtracted from the intercept α_j, where as in (2) it was added.The differences between these two different ways to define the model is that with subtraction, we can assume sign meaning, where if Y is more likely to belong to higher group when x_l, $l \in \{1, \ldots p\}$, increases, then $\beta_l > 0$. The same holds when the model is defined with $P(Y_i > j \mid \boldsymbol{x}_i)$ instead of $P(Y_i \leq j \mid \boldsymbol{x}_i)$. Although (4) is most commonly used to define the proportional odds model, some people prefer defining the model with addition or probability $P(Y_i > j \mid \boldsymbol{x}_i)$ (Harrell, 2001).

3.1.2 Assumptions of proportional odds model

As proportional odds model is a cumulative logit model, the same assumptions as standard logistic regression hold (Harrell, 2001). This essentially means that the covariates are related linearly to the log odds, but not to each other. Then there are two assumptions that distinguish proportional odds models from general cumulative logit models: proportional odds and parallel regression assumption (Long, 1997). These assumptions are frequently referred to as a single assumption, proportionality assumption, which is discussed further at the end of the section. Prior to that, these two assumptions are explained, beginning with the proportional odds assumption.

Let us consider two different sets of values of covariates, $\boldsymbol{x}_i^{(1)}$ and $\boldsymbol{x}_i^{(2)}$. The proportional odds model with those covariates must satisfy

$$\frac{P(Y_i \leq j \mid \boldsymbol{x}_i^{(1)})/(1 - P(Y_i \leq j \mid \boldsymbol{x}_i^{(1)}))}{P(Y_i \leq j \mid \boldsymbol{x}_i^{(2)})/(1 - P(Y_i \leq j \mid \boldsymbol{x}_i^{(2)}))} = \exp(\boldsymbol{\beta'}(\boldsymbol{x}_i^{(2)} - \boldsymbol{x}_i^{(1)})). \tag{5}$$

In other words, the proportional odds models' odds ratios are proportional to the distance of $\boldsymbol{x}_i^{(1)}$ and $\boldsymbol{x}_i^{(2)}$ and the proportionality constant is the same for every odds ratio of the model.

Assume now that the only difference between these two sets of values is one covariate x_{il}. The value of $\boldsymbol{x}_i^{(1)}$ is denoted as $x_{il} + 1$ and differs by one from the value of $\boldsymbol{x}_i^{(2)}$, which is denoted as x_{il}. Following equation (5) odds ratio now reduces to $\exp(-\beta_l)$.

Both of these equations are called proportional odds assumptions. You should note that in the case of the latter, the answer is always $\exp(-\beta_l)$ in every category $j = 1, 2, \ldots, k-1$. We can lead from this second assumption of proportional odds model, which is parallel regression assumption.

To summarize, the parallel regression assumption says that the only difference between models for each category $j = 1, 2, \ldots, k-1$ is the intercept α_j. The model's coefficients for covariates remain same for every category j. In other words, we assume that the coefficients for covariates of the model are independent of Y_i's category j. This assumption is also known as proportionality assumption, and as previously stated, it is frequently mentioned as the only assumption of the proportional odds model. This is due to the fact that proportional odds and parallel regression assumption correspond to each other, and if parallel regression assumption holds, so does the proportional odds assumption. This is also why it is only necessary to test whether the parallel regression assumption is true for the model.

There are numerous approaches to test the assumption. One popular way is to do it graphically (Harrell, 2001), because if this assumption is correct, the slopes of the model do not differ significantly between different groups. Score test is also sometimes recommended, but this test is frequently criticized (Agresti, 2010; Harrell, 2001). The Brant-Wald-test (Brant, 1990) is currently the most recommended test for proportional assumptions. It is also known as the Brant-test, but we will refer to it as the Brant-Wald-test in this thesis because in older literature this test has been referred to often as the Wald-test. The Brant-Wald-test is explained in more detail in Section 3.1.5.

3.1.3 Fitting of proportional odds model

Maximum likelihood modeling is the most common method for fitting the proportional odds model, and it is used by most statistical programs. To define the likelihood function for the proportional odds model, let us define first a binary indicator of the response. Let y_{ij} equal 1 if y_i falls in category j, and 0 otherwise. Now we can write the likelihood function for the proportional odds models as

$$
\begin{aligned}
L(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid y_{ij}, \boldsymbol{x}_i) &= \prod_{i=1}^{n} \Big[\prod_{j=1}^{k} P(Y_i = j \mid \boldsymbol{x}_i)^{y_{ij}} \Big] \\
&= \prod_{i=1}^{n} \Big[\prod_{j=1}^{k} (P(Y_i \leq j \mid \boldsymbol{x}_i) - P(Y_i \leq j-1 \mid \boldsymbol{x}_i))^{y_{ij}} \Big] \qquad (6) \\
&= \prod_{i=1}^{n} \left[\prod_{j=1}^{k} \Big[\frac{\exp(\alpha_j - \boldsymbol{\beta}'\boldsymbol{x}_i)}{1 + \exp(\alpha_j - \boldsymbol{\beta}'\boldsymbol{x}_i)} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}'\boldsymbol{x}_i)}{1 + \exp(\alpha_{j-1} - \boldsymbol{\beta}'\boldsymbol{x}_i)} \Big]^{y_{ij}} \right],
\end{aligned}
$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{k-1})'$. Note that in this equation $j = 1, \ldots, k$, and when $j = k$, $P(Y_i \leq j \mid \boldsymbol{x}_i) = 1$, and when $j = 1$, $P(Y_i \leq j - 1 \mid \boldsymbol{x}_i) = 0$. To obtain maximum likelihood estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ for parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, for example Fisher scoring algorithms can be used. This method is covered in more detail in McCullagh (1980) and Agresti (2010).

The maximum likelihood method has some drawbacks, the most common of which is that the estimate it provides for the coefficient can occasionally be infinite (Agresti, 2010). This is usually due to a small sample size, unbalanced data, or a large number of model parameters. One solution to this is to simplify the model by removing a covariate. However, the model may no longer be as well-fitting as the model with infinite coefficients. In this case, one should consider whether the model with infinite coefficient fits and functions properly. But how to determine how well the models fit? This is explained further in the next section.

3.1.4 Goodness-of-fit tests for proportional odds model

As previously stated in Section 3.1.1, the proportional odds model is an extension of a logistic regression model. This means that the majority of the goodness-of-fitness statistics for logistic regression model can be applied to the proportional odds model (Harrell, 2001). In this thesis, we use pseudo-R^2, which is further explained below using Long (1997) as source.

Pseudo-R^2 is based on the R^2 statistic, which is used to assess the fit of the ordinary least squares estimator for linear regression models. But unlike R^2, pseudo-R^2 cannot be used to estimate how much covariates explain the variation of the response variable. They are instead used to demonstrate an improvement in model likelihood over the null model. For other pseudo-$R2$ properties, a similar interpretation as with R^2 is sought. As a result, their range, like R^2, is generally [0,1], and their limits can be interpreted similarly - if the value obtained is close to zero, the model is thought to be a poor fit, and if it is close to one, the model is thought to be a good fit.

Because most pseudo-R^2's are similar and interpreted similarly, we will only cover the ones used in this thesis, beginning with McFadden's pseudo-R^2 (McFadden, 1974). This is one of the earliest pseudo-R^2's and one of the most commonly

used. Let L denote the likelihood-function from (6) for the model whose goodness-of-fit we want to examine and $L^{(null)}$ for the model without any covariates and only with the intercept. McFadden's pseudo-R^2 is now defined as

$$R^2_{MF} = 1 - \left(\frac{\log L^{(null)}(\hat{\boldsymbol{\alpha}})}{\log L(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})} \right)^{2/n}, \tag{7}$$

where n denotes the sample size. Popularity of R^2_{MF} is based on its similar properties with R^2-statistics (Windmeijer, 1995). However, it has its own problems: the upper limit of it is not precisely one and R^2_{MF} increases when new variables are included in the model. The latter problem is solved using the adjusted Mcfadden's pseudo-R^2, which is defined as

$$R^2_{adjMF} = 1 - \left(\frac{\log L^{(null)}(\hat{\boldsymbol{\alpha}}) - p}{\log L(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})} \right)^{2/n},$$

where p is the number of covariates.

Another popular pseudo-R^2 also used in this thesis is Cox's and Snell's pseudo-R^2. This statistic is defined as

$$R^2_{CS} = 1 - \left(\frac{L^{(null)}(\hat{\boldsymbol{\alpha}})}{L(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})} \right)^{2/n},$$

where L and $L^{(null)}$ are defined the same way as in equation 7. R^2_{CS}'s upper limit is not precisely one either: rather it is $1 - L^{(null)2/n}$ (Hu et al., 2006). This issue is resolved in Nagelkerke's pseudo-R^2 (Nagelkerke, 1991), which is defined as

$$R^2_N = \frac{1 - \left(\frac{L^{(null)}(\hat{\boldsymbol{\alpha}})}{L(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})} \right)^{2/n}}{1 - L^{(null)}(\hat{\boldsymbol{\alpha}})}.$$

There is no particular rule for determining which pseudo-R^2 to use, although studies have been done to compare different pseudo-R^2 indices (see e.g. Windmeijer, 1995). It is important to note that these goodness-of-fit statistics, like others for logistic regression models, do not test the proportionality assumption of the proportional odds model. This was mentioned earlier in the Section 3.1.2 and the Brant-Wald-test, which is commonly used to test it, is explained in the next subsection.

3.1.5 Brant-Wald test

The Brant-Wald test was first introduced by Brant (1990) in his article. This is cited as the main source in this Section along with Long (1997).

The Brant-Wald-test is an omnibus test, which means that it can be used to test the proportionality assumption for the entire model. It can also be used to test the proportional assumption for a single covariate's coefficient, which is why it is popular. Score test, for example, can only be used as an omnibus test for proportional assumption.

Assume now that $\hat{\boldsymbol{\beta}}_j$ is the ML-estimate of $\boldsymbol{\beta}_j$ of the generalized cumulative logit model and write $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}_j)$ for the estimate of its asymptotic covariance matrix. Then write $\hat{\pi}_j$ for $P(Y_i \geq j \mid \boldsymbol{x}_i)$,

$$\text{logit}[\hat{\pi}_j] = -\hat{\alpha}_j + \hat{\boldsymbol{\beta}}_j' \boldsymbol{x}_i$$

to calculate and estimate it. Now we can estimate the asymptotic covariance between $\hat{\boldsymbol{\beta}}_j$ and $\hat{\boldsymbol{\beta}}_i$. To do this, we use formula

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\beta}}_i) = (\boldsymbol{X}'\boldsymbol{W}_{jj}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{W}_{ji}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{W}_{ii}\boldsymbol{X})^{-1}$$

where $i, j = 1, 2, \ldots, k-1$ and $\boldsymbol{W}_{ij}$ is a $n \times n$ diagonal matrix with ith element being $\hat{\pi}_i - \hat{\pi}_j\hat{\pi}_i$. $\boldsymbol{X}$ is $n \times (p+1)$ design matrix that contains ones in the first column and covariates in the remaining columns.

After this we can define $\hat{\boldsymbol{\beta}}^* = (\hat{\boldsymbol{\beta}}_1', \ldots, \hat{\boldsymbol{\beta}}_{k-1}')'$ and

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}^*) = \begin{bmatrix} \widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}_1) & \ldots & \widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_{k-1}) \\ \vdots & \ddots & \vdots \\ \widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}_{k-1}, \hat{\boldsymbol{\beta}}_1) & \ldots & \widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}_{k-1}) \end{bmatrix}$$

We can now construct the Wald test for the model with hypothesis $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \ldots = \boldsymbol{\beta}_{k-1}$. Another way to write this hypothesis is $H_0 : \boldsymbol{D}\boldsymbol{\beta}^* = \boldsymbol{0}$, where $\boldsymbol{D}$ is a $p(k-2) \times p(k-1)$ contrast matrix and is defined as

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{I} & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{I} & \boldsymbol{0} & -\boldsymbol{I} & \ldots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \ldots & -\boldsymbol{I} \end{bmatrix},$$

where $\boldsymbol{I}$ is $p \times p$ identity matrix and $\boldsymbol{0}$ is $p \times p$ matrix full of zeros. Now we have defined everything needed for Wald test statistic written as

$$X^2 = (\boldsymbol{D}\boldsymbol{\beta}^*)'(\boldsymbol{D}\widehat{\text{Cov}}(\boldsymbol{\beta}^*)\boldsymbol{D}')^{-1}(\boldsymbol{D}\boldsymbol{\beta}^*). \tag{8}$$

Under the null hypothesis $X^2 \sim \chi^2_{(k-2)p}$.

We can do this test for single variables by selecting only rows and columns of $\boldsymbol{D}$, $\hat{\boldsymbol{\beta}}^*$ and $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}^*)$ that corresponds with the coefficients tested. In this case, if x_l is the covariate we want to test, then $\hat{\boldsymbol{\beta}}^* = (\hat{\beta}_{1l}, \ldots, \hat{\beta}_{(k-1)l})'$ and

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}^*) = \begin{bmatrix} \widehat{\text{Cov}}(\hat{\beta}_{1l}) & \ldots & \widehat{\text{Cov}}(\hat{\beta}_{1l}, \hat{\beta}_{(k-1)l}) \\ \vdots & \ddots & \vdots \\ \widehat{\text{Cov}}(\hat{\beta}_{(k-1)l}, \hat{\beta}_{1l}) & \ldots & \widehat{\text{Cov}}(\hat{\beta}_{(k-1)l}) \end{bmatrix}$$

.

We define now a $(k-2) \times (k-1)$ contrast matrix $\boldsymbol{D}$ as

$$\boldsymbol{D} = \begin{bmatrix} 1 & -1 & 0 & \ldots & 0 \\ 1 & 0 & -1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \ldots & -1 \end{bmatrix}.$$

and we can calculate X^2 statistics using (8). Now $X^2 \sim \chi^2_{(k-2)}$

Usually, the main focus of the test is the omnibus value. However, if the model fails the omnibus test, examining single covariates with this test is a good idea. In some cases, only a few covariates fail to satisfy the model's proportional assumption. If this occurs, one should think about using a different model, and in the case mentioned earlier, a partial proportional odds model may be better than a proportional odds model. This model is further explained in the following section.

3.2 Partial proportional odds model

In this chapter, the partial proportional odds model is defined. Like its name suggests, it is an extension of the proportional odds model which was introduced in the previous chapter. It was first introduced by Peterson and Harrell (1990) and this article will be used as the main source for this chapter. There are a couple ways to write the partial proportional odds model and first we write it in a way that fits equation (2). In this case, let us assume we have p covariates and q ($q < p$) of them violates the proportional assumption. Now $\boldsymbol{x}_i^{(1)}$ is q-sized column vector of values of the covariates violating the proportional assumption and $\boldsymbol{x}_i^{(2)}$ is $(p-q)$-sized column vector of values of the covariates fulfilling the assumption. The partial proportional odds model can be written as

$$\text{logit}[P(Y \leq j \mid \boldsymbol{x})] = \alpha_j + \boldsymbol{\beta}_{1j}' \boldsymbol{x}_i^{(1)} + \boldsymbol{\beta}_2' \boldsymbol{x}_i^{(2)}, \tag{9}$$

where $\boldsymbol{\beta}_{1j}$ is q-sized vector of coefficients of $\boldsymbol{x}_i^{(1)}$ for category j and $\boldsymbol{\beta}_2$ is $(p-q)$-sized vector of coefficients for $\boldsymbol{x}_i^{(2)}$. In this case, the intercept α_j and $\boldsymbol{\beta}_{1j}$ are different

for each category j, but $\boldsymbol{\beta}_2$ remains same. Another way to write this model is as Peterson and Harrell (1990) do it, in which case the equation is written as

$$\text{logit}[P(Y \leq j \mid \boldsymbol{x})] = \alpha_j + \boldsymbol{\beta}'\boldsymbol{x} + \boldsymbol{\gamma}_j'\boldsymbol{u}, \tag{10}$$

where $\boldsymbol{\beta}$ is as in the case of the proportional odds model. Parameter $\boldsymbol{u}$ is q-vector ($q \leq p$) and is a subset of $\boldsymbol{x}$, which contains covariates that do not follow the proportional assumption. Vector $\boldsymbol{\gamma}_j$ is a q-vector which contains regression coefficients associated with $\boldsymbol{u}$ for category j. In other words, $\boldsymbol{\gamma}_j'\boldsymbol{u}$ is an increment for category j.

3.3 Principal component analysis

As described in Section 2.2, the haemodynamic variables used are functional variables, but the data contain averages of one-minute intervals that were either 5 or 10 pieces, depending on the variable. In this thesis we want to use these variables as covariates for models of Sections 3.1 and 3.2, but in these models non-functional covariates are used. Even though we ignore the functionality of covariates and only use observed values of haemodynamic variables, it also would be problematic because every haemodynamic variable contains 5 or 10 observations. We solve this problem by using principal component analysis, which ignores the functionality of the variable, but can be used to reduce the dimensions of multivariate variables. In this case, we use principal component analysis to replace each individual haemodynamic covariate with their principal component scores in the regression model. The purpose of this chapter is to go through this method in greater detail, starting with a definition and then going through how to select the main components. The primary source here is Jolliffe (2002).

3.3.1 Definition

The idea of principal component analysis is that it reduces the dimension of the multivariate dataset yielding a new set of variables known as the principal components. The principal component scores are uncorrelated and arranged so that the first components contain the majority of the variation in the original variables. Next is explained how this will be done in practice.

Let $\boldsymbol{x} = (x_1, \ldots, x_M)'$ denote a random M-vector. We define linear combinations for the vector as

$$\begin{aligned}
\boldsymbol{a}_1'\boldsymbol{x} &= a_{11}x_1 + \ldots + a_{1M}x_M = \sum_{m=1}^{M} a_{1m}x_m \\
\boldsymbol{a}_2'\boldsymbol{x} &= a_{21}x_1 + \ldots + a_{2M}x_M = \sum_{m=1}^{M} a_{2m}x_m \\
&\vdots \\
\boldsymbol{a}_M'\boldsymbol{x} &= a_{M1}x_1 + \ldots + a_{MM}x_M = \sum_{m=1}^{M} a_{Mm}x_m,
\end{aligned}$$

where vector $\boldsymbol{a}_m = (a_{m1}, a_{m2}, \ldots, a_{mM})'$ $(m = 1, \ldots, M)$ contains M constants.

Let us define combinations so that they are uncorrelated and combinations' variances $Var(\boldsymbol{a}_m'\boldsymbol{x}) = \boldsymbol{a}_m'\boldsymbol{\Sigma}\boldsymbol{a}_m$, where $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{x}$, are maximized under the normalization constraint $\boldsymbol{a}_m'\boldsymbol{a}_m = 1$. Now we have defined the principal components for the dataset, where mth principal component is defined as $z_m = \boldsymbol{a}_m'\boldsymbol{x}$. The M-vector containing all principal components is denoted as $\boldsymbol{z} = (z_1, \ldots, z_M)'$.

Coefficient matrix $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_M)'$, which is also called as the loading matrix, can be solved using eigenvector-eigenvalue decomposition $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'$, where $\boldsymbol{U}$ is an orthogonal matrix which contains eigenvectors of the covariance matrix $\boldsymbol{\Sigma}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix which contains eigenvalues of the covariance matrix. We can now define the loading matrix as $\boldsymbol{A} = \boldsymbol{U}'$.

Above we defined all possible principal components for variable with M observations, but to reduce dimension size, we want to find l $(l \leq M)$ first principal components that contain the majority of the variation. The principal components have been defined in such a way that the first main components explain the majority of the variation in the dataset, but it is up to the analyst to determine how much variation is sufficient, i.e. how many of the first main components are preferred to be used. The most common method of determining this is presented next, by using cumulative percentages.

The method of selecting the main components based on cumulative percentages is quite simple: we define a cumulative percentage for the total variation that we want the selected principal components to explain. In other words, we want to know what percentage the variances of the selected principal components make up of the total variance. The first l component explain

$$\frac{Var(z_1) + \ldots + Var(z_l)}{Var(z_1) + \ldots + Var(z_M)} \cdot 100\%$$

of the total variance.

The method does not define any benchmark values for the total variance of the selected principal components components, and it is up to the analyst to determine

which percentage is acceptable. The selected value is usually between 70% and 90%, and the selection should take into account factors such as the number of all principal components M. As the amount of M grows, the value of the chosen cumulative percentage should decrease. In contrast, if the first couple of main components account for the majority of the variation, the value can be defined larger than 90 % .

Examining the scree plot of the principal components is another popular method. The variances of the principal components are plotted against the corresponding number of the component in the scree plot, and the plot can be used to estimate which components contain enough of the variance. This is accomplished by looking for the "elbow" in the scree plot, or looking for the principal component after which the graph is deemed not "steep" enough. Based on this, the "elbow" principal component and the principal components before it are selected for use. There is an example of a screeplot and how to find "elbow" in Figure 5. It can be seen from the screeplot thatt after the second principal component, the plots for following principal components do not change much, meaning that the "elbow" of the scree plot is the second principal component.

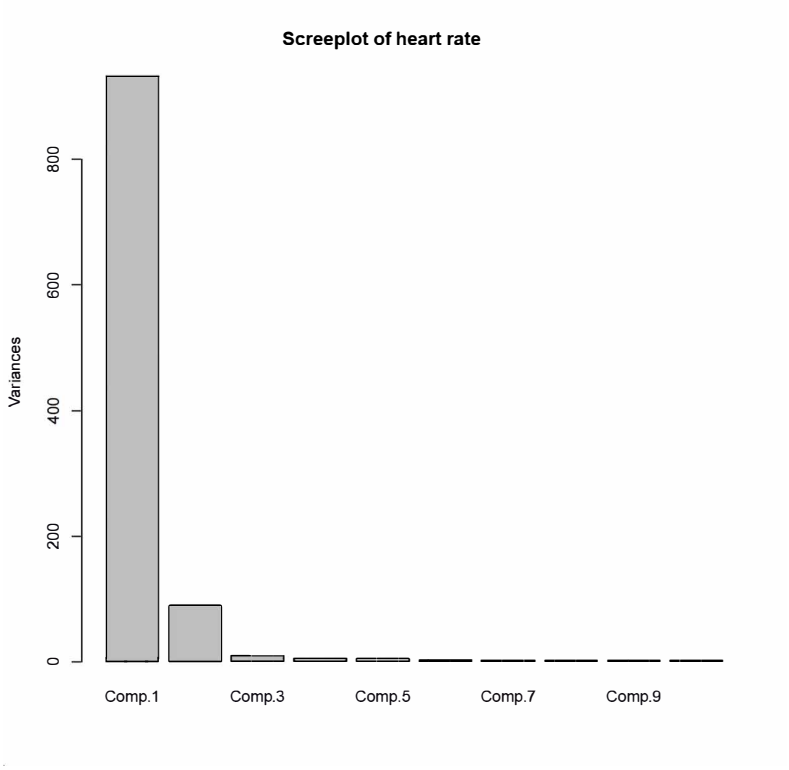


Figure 5: Screeplot of principal components of heart rate (HR).

3.3.2 Functional data and PCA

If the data are functional and can be graphed as a curve, we can use the reduced rank model (James et al., 2000) written as

$$x_i(t) = \mu(t) + a_1(t)\hat{z}_1 + \ldots + a_l(t)\hat{z}_l + \epsilon_i(t),$$

where $i = 1, \ldots, n$, $x_i(t)$ is curve for the ith individual at the time t, $\mu(t)$ is a mean curve, $a_m(t)$ is curve defining the mth principal component and $\hat{z}_m$ is the mth principal component score (sample principal component) and in this model, first l principal component scores are used.

To estimate curves for principal components, we first denote $x_i(t)$ as $\underset{\sim}{z}$ and $\boldsymbol{x} = \underset{\sim}{\boldsymbol{z}} - E(\underset{\sim}{\boldsymbol{z}})$, with $Cov(\boldsymbol{x}) = \boldsymbol{U\Lambda U'}$. Now we want to denote $\boldsymbol{x}$ using principal components. Because principal components analysis can be viewed as rotation of the original data, in this case, we only have to rotate it back, meaning $\boldsymbol{x} = \boldsymbol{zA}$. We can approximate this result by using first l principal components ($l < m$) as

$$\boldsymbol{x} \approx \boldsymbol{a}_1 z_1 + \ldots + \boldsymbol{a}_l z_l.$$

From this equation we can deduce that we can use loadings of the corresponding principal components to estimate the curves for principal components. In Figures 6 and 7 there is an example of reconstructing the original curve of the original variable for a random individual with corresponding principal component scores and loadings of the principal components.

This kind of a reconstruction is used for example in the eigenface approach (Turk and Pentland, 1991) used for face recognition. In this method, instead of the original faces, one calculates with principal component analysis eigenfaces which contain basis features of the original picture. These eigenfaces are then used for face recognition instead of the original face. Similar way we also want to replace original variables with corresponding principal components scores, which contain basis features of the original variables, in the chosen models.

3.4 Functional ordinal logistic regression

The proportional odds model and partial proportional odds model use principal components calculated from interval averages of haemodynamic variables as covariates, but as stated earlier, these variables are actually functional. This cannot be taken into account in previous models, but it is taken into account in the functional ordinal logistics regression. However, before presenting the model, functional data analysis will be presented briefly, which will serve as the model's foundation in addition to the proportional odds model.

3.4.1 Review of functional data analysis

As previously stated, in the functional data analysis data are primarily functional (or the underlying process is assumed to be functional). This means that rather than assuming that the data are made up of a finite set of observations that could be represented as scalars or vectors, data observations are assumed to be functions defined for a set of T. One of the most common types of functional data are data

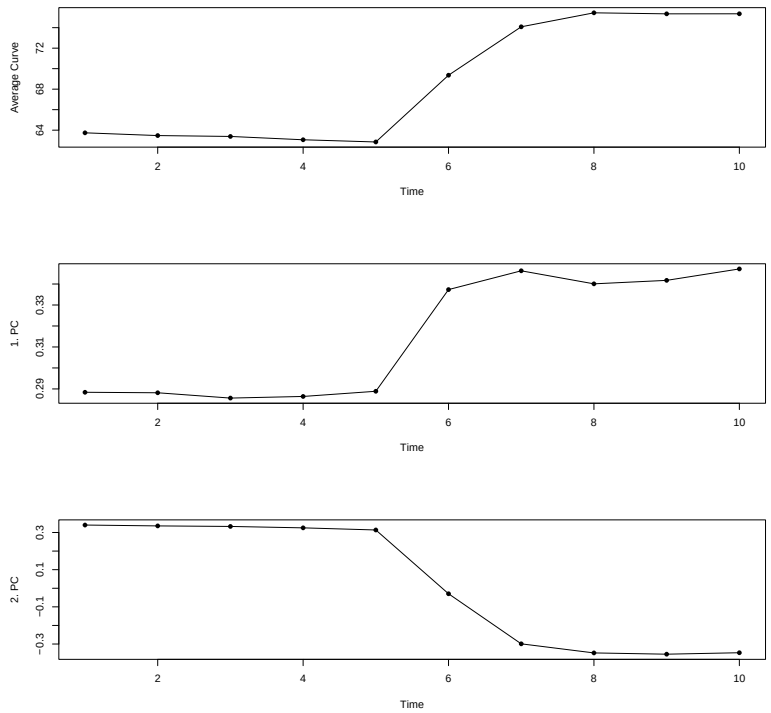


Figure 6: On the first row is the average curve for heart rate (HR). On the second row is the first principle curve and on the third row is the second principal curve for heart rate.

whose observations are assumed to be realizations of the L_2 continuous stochastic process $X = \{X(t), t \in [0, T]\}$, which means that $E(\int_T X^2(t)dt) < \infty)$ has to hold (Wang et al., 2016).

In practice, only a finite number of observation points can be detected from these functions, so functional data analysis is used to approximate these functions. Functional data analysis also seeks to address the challenges posed by functional data's infinite dimensional structure, which often includes in practice dimension reduction like in multivariate analysis, because most often we have only have finite amount of observations from the functional data. One of the most well-known functional data analysis methods is functional principal component analysis, which non-functional version was introduced in Section 3.3. Many other methods are included in functional data analysis, and new methods are constantly being developed. One of the most recent methods is functional ordinal logistics regression, which is presented below.

3.4.2 Functional ordinal logistic regression model

Although both ordinal logistic regression models and functional models have been extensively studied and developed, functional logistic regression models have only been introduced recently. In this section, functional ordinal logistic regression is

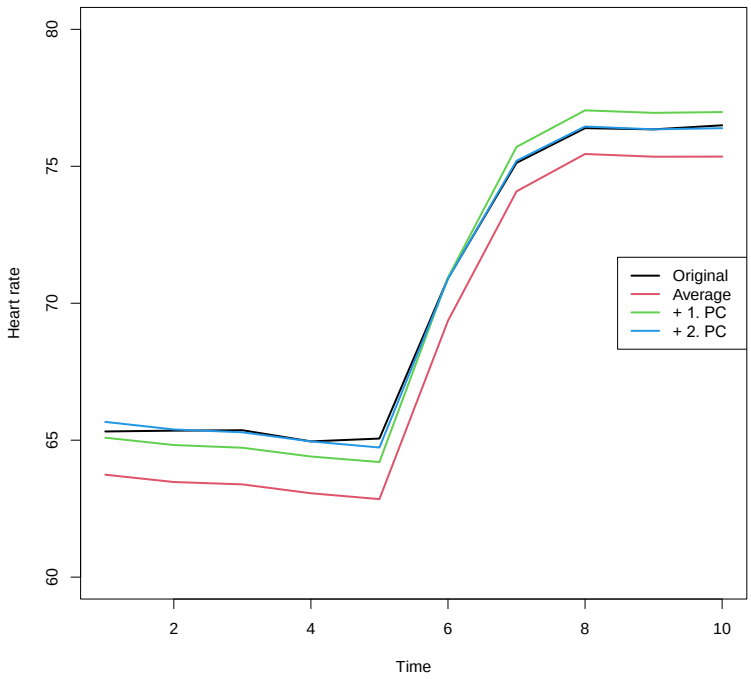


Figure 7: Reconstruction of the original curve of the random individuals' heart rate (HR) with average curve and first two principal curves.

presented using Jacques and Samardzic (2022) and Chiu et al. (2022) as main sources. However, the model introduced in this thesis will be an extended version of the model presented in literature as it includes both non-functional covariates and more than one functional covariate.

Recall that Y_i denotes a random variable for subject i, $i = 1, \ldots, n$. Let us denote values of Y_i as $1, \ldots, k$, where 1 is the lowest group and k is the highest group. Assume now that we have p covariates so that the first q ones are non-functional random variables $\boldsymbol{x} = (x_1, \ldots, x_q)'$, and rest of them are functional random variables $x_j(t)$ $(j = q+1, \ldots, p)$ with values $L_2[0,T], T > 0$, where $L_2[0,T]$ is vector space of all functions $X : (0,T) \to \mathbb{R}$ satisfying $E(\int_T X^2(t)dt) < \infty$. Let $x_j(t)$ be an L_2-continuous process. Now we can write functional logistic regression as

$$\text{logit}\, P(Y_i \leq j | X = \boldsymbol{x_i}) = \alpha_j - \sum_{s=1}^{q} \beta_s x_{si} - \sum_{s=q+1}^{p} \int_0^T \beta_s(t) x_{si}(t)\, dt. \qquad (11)$$

As mentioned earlier, most often we have only knowledge about a finite amount of observed values of $x_i(t)$, which we want to use to approximate function $x_i(t)$. If we assume that there are noise in observed value, one of the recommended methods (Ramsay and Dalzell, 1991) for approximation is one where it is assumed that function $x_i(t)$ can be decomposed into finite amount basis of functions following

way

$$x_i(t) \propto \sum_{r=1}^{R} a_{ir}\phi_r(t) = \boldsymbol{a}_i'\boldsymbol{\phi}(t),$$

where vector $\boldsymbol{\phi}(t) = (\phi_1(t), \ldots, \phi_R(t))'$ contains R basis functions, $\boldsymbol{a_i} = (a_{i1}, \ldots, a_{iR})'$ contains basis expansion coefficients. In this instance, it is recommended (Ramsay and Dalzell, 1991) to use either B-spline basis or Fourier basis as the basis function.

We use the B-spline basis here because the Fourier basis is recommended for data with a repeated pattern and the B-spline basis for other cases. When creating B-spline basis functions (or any other spline function), which are polynomial segments, we must define argument values called knots since we want to connect these segments end-to-end at these knots. The segments are smoothed over the breaks between the knots in a specific fashion for B-spline basis functions. The number of knots used is determined on the shape of the functional data's curve, and usually the amount of knots cannot be higher than the amount of observation points of the function. The number of basis functions to define spline function is determined on the order of used basis functions, which is degrees of freedom + 1, and amount of knots, and is written as order + number of knots - 2.

Usually knots are placed evenly and when using b-spline basis knots have to be placed in the beginning and end of the interval of the function. It is however recommended that if the curvature varies significantly, more knots are required than if the curve does not change. And if the way the curve varies, more knots are placed when the curve changes quickly and fewer knots when it changes slowly (Ramsay and Dalzell, 1991). In Figure 8 is shown an example of choosing knots for a functional variable with ten observation points while using cubic B-spline basis functions with order 4.

B-spline basis function can be used in this case to decompose regression coefficients $\beta(t)$ of functions in a similar way into finite-amount of basis functions

$$\beta(t) \propto \sum_{r=1}^{R} b_r\phi_r(t) = \boldsymbol{b}'\boldsymbol{\phi}(t).$$

After decomposing both $x_s(t)$ and $\beta_s(t)$, (11) can be written as

$$\begin{aligned}
\operatorname{logit} P(Y \leq j | X = \boldsymbol{x}) &= \alpha_j - \sum_{s=1}^{q} \beta_s x_{si} - \sum_{s=q+1}^{p} \int_0^T \sum_{r=1}^{R} b_s r \phi_s r(t) \sum_{r'=1}^{R} a_{sir'}\phi_{sr'}(t)\, dt \\
&= \alpha_j - \sum_{s=1}^{q} \beta_s x_{si} - \sum_{s=q+1}^{p} \sum_{r=1}^{R}\sum_{r'=1}^{R} b_{sr} a_{sir'} \int_0^T \phi_{sr}(t)\phi_{sr'}(t)\, dt \\
&= \alpha_j - \sum_{s=1}^{q} \beta_s x_{si} - \sum_{s=q+1}^{p} \boldsymbol{b}_s'\boldsymbol{\Psi}\boldsymbol{a}_{si},
\end{aligned}$$

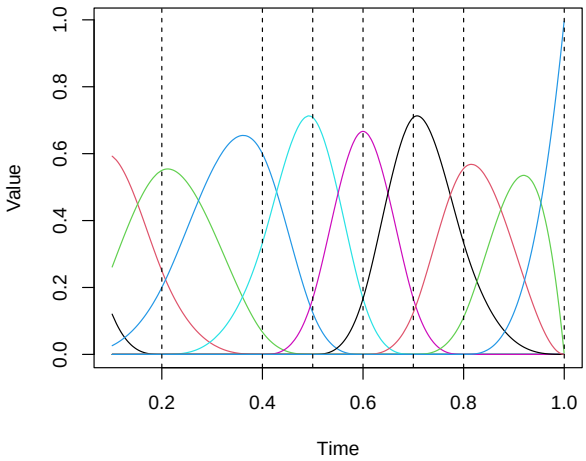


Figure 8: Illustration of cubic B-spline basis functions (df = 3) with two boundary knots placed at 0 and 1, and six interior knots placed at 0.2, 0.4, 0.5, 0.6, 0.7 and 0.8

where $\boldsymbol{\Psi}$ is the $R \times R$ matrix of inner products between basis functions. We can estimate parameters $\boldsymbol{b}$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{j-1})'$ from this function by using maximum likelihood estimation. Because closed form solutions do not exist, an iterative optimization algorithm is used.

4 Results

In this chapter, the models presented in the previous chapter are fitted to data introduced in Chapter 2, after which the results are discussed and a comparison between the three models is made. Before this, principal component analysis (Section 3.3) for haemodynamic variables must be conducted and after that has been done, principal component scores to be used as covariates can be chosen.

4.1 Preparations

In this section, principal component analysis and choosing of haemodynamic covariates to use in model, is explained. This was done using R program, and R code used can be found in Appendix A. It was explained in chapter 2 that haemodynamic variables have ten (or five in the case of ECW and PWV) values. Principal component analysis was used to reduce the dimensions of the mean centered haemodynamic variables, and the corresponding principal component scores were included in the model instead of the original variables. As a result, most of the haemodynamic variables were replaced with their first two principal component scores (and in the case of ECW and PWV, only the first one). This decision was made using screeplots, which are shown in Figure 9.

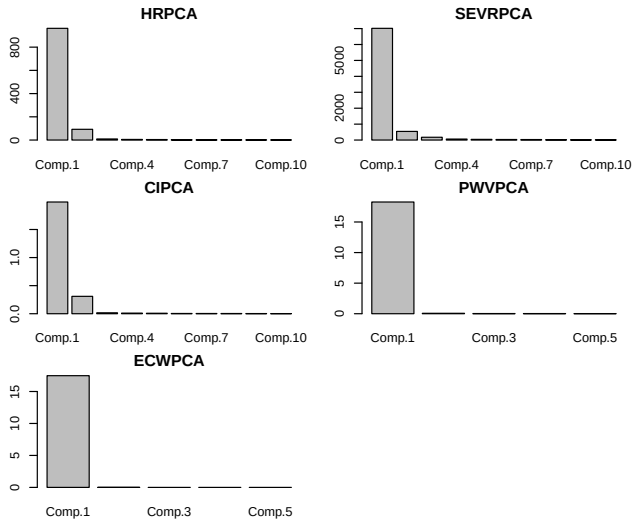


Figure 9: Screeplots of haemodynamic covariates (HR, SEVR, CI, PWP and ECW) used in the model.

The haemodynamic variables used in the model were selected using forward and backward stepwise regression to a proportional odds model that included the principal component scores of all haemodynamic variables. Furthermore, the model included variables such as age, gender, and BMI, and because potential interactions between these variables and haemodynamic variables were suspected, potential interactions were also included in the initial model. The model also

contains the interactions betweeen sex and age, age and BMI, and sex and BMI. The proportional odds model fitted as a result of this is introduced next.

4.2 Fitted proportional odds model

Based on the model with lowest AIC according to forward and backward step regression, we ended up with a model similar to (12), but without 1. principal component score of cardiac index or interaction terms between sex and 2. principal component score of SEVR or age and 2. principal component score of HR. These covariates were introduced to the model to make interpreting and comparing haemodynamic variables easier. This model was fitted with the VGAM-package's vglm-function (Yee et al., 2015). R code used to fit models and comparing them can be found in Appendix B.

$$\begin{aligned}\text{logit}[P(Y_i \leq j \mid \boldsymbol{x}_i)] =& \alpha_j + \beta_1 sex + \beta_2 BMI + \beta_3 age + \beta_4 hr1 \\ &+ \beta_5 hr2 + \beta_6 sevr1 + \beta_7 sevr2 + \beta_8 ci1 \\ &+ \beta_9 ci2 + \beta_{10} pwv + \beta_{11} ecw + \beta_{12} sex \cdot sevr1 \\ &+ \beta_{13} sex \cdot sevr2 + \beta_{14} BMI \cdot pwv + \beta_{15} BMI \cdot ecw \\ &+ \beta_{16} age \cdot hr1 + \beta_{17} age \cdot hr2.\end{aligned} \tag{12}$$

Table 3 displays results of this model. It should be noted that (in all fitted models) values of the regression coefficients are displayed in same manner that odds ratios can be determined using the equation $\exp(\beta_i)$ and they are interpreted in the same way; for example, the odds ratio of sex is $\exp(-0.2579) \approx 0.773$, which can be translated as "The odds that male answers lower health status is 0.773-times lower than the odds that female answers lower health status, when both have the same values for other covariates and both principal component scores of sevr are zero." It should be noted that the majority of covariates are also part of some interaction term, therefore interpreting odds ratios is not always as simple. In the previous example, if both individuals have the same values but both principal component scores of SEVR are zero, the interpreted odds ratio is $\exp(-0.2579 + (-0.0045) + (-0.0056)) \approx 0.765$. In other words, higher values for principle scores of SEVR reduce the odds that male responses lower health status. It should also be noted when interpreting the results of the proportional odds model that the odds ratios are the same for each model version, i.e. model for the cumulative probability that the subject answers health status lower than good and model for the cumulative probability that the subject answers health status lower than excellent.

The interpretation of odds ratios for age and BMI is similar, especially when interaction terms are excluded. When the data are otherwise the same and the principal component scores for HR are zero, the odds that a one-year older person answers lower health status are $\exp(0.0320) \approx 1.032$-times higher than the odds that a younger person answers lower health status. When the values are otherwise

Table 3: Results of the proportional odds model.

Coefficient	*Estimate*	*z-value*	*p-value*
1. intercept	-6.8775	-8.714	< 0.0001
2. intercept	-3.4960	-4.698	< 0.0001
Sex (male)	-0.2579	-1.065	0.2867
BMI	0.1485	6.282	< 0.0001
Age	0.0320	3.521	0.00043
1. PC score of HR	0.0315	2.755	0.0059
2. PC score of HR	0.0402	1.239	0.2152
1. PC score of SEVR	0.0064	2.816	0.0049
2. PC score of SEVR	0.0104	1.740	0.0818
1. PC score of CI	0.0182	0.202	0.8399
2. PC score of CI	-0.5046	-2.230	0.0257
1. PC score of PWV	0.3492	2.595	0.0095
1. PC score of ECW	-0.2874	-2.305	0.0212
Sex * 1. PC score of SEVR	-0.0045	-2.225	0.0261
Sex * 2. PC score of SEVR	-0.0056	-0.796	0.4263
BMI * 1. PC score of PWV	-0.0126	-2.624	0.0087
BMI * 1. PC score of CI	0.0078	1.796	0.0726
Age * 1. PC score of HR	-0.0004	-2.090	0.0366
Age * 2. PC score of HR	-0.0002	-0.268	0.7129

the same and the principal component scores of PWV and ECW are zero, the odds ratio of BMI can be interpreted as "The odds that person with one value higher BMI answers lower health status is $\exp(0.1485) \approx 1.160$-times higher than the odds that person with lower BMI answers lower health status".

These interpretations are compatible with Figure 1, which shows that average BMI and age values for health status groups decrease as health status gets higher. Similarly, previous interpretations of sex support the assumption that males tend to overestimate their health when compared to females.

Because original variables are not included in the model, interpreting the results for haemodynamic variables is more difficult than for age, gender, and BMI. Curves of the principal components are used to interpret the principal component scores of the haemodynamic variables (they can be found in Appendix Figure 10). For example, results of the first principal component score of SEVR are interpreted. When two females with otherwise identical values are compared, the odds ratio can be interpreted "The odds, that female with one value higher 1. principal component score of SEVR answers lower health status, is $\exp(0.0064) \approx 1.006$-times higher than the odds, that person with lower 1. principal component score of SEVR answers lower health status". What this means for the original variable is determined from the curve of 1. principal component of SEVR (Figure 10).

Up until the midpoint of the curve, the 1. principal component of SEVR has

values around 0.330; after that, values range from 0.300 to 0.315. That is why when comparing females with otherwise identical values but different 1. principal component scores of SEVR , the higher value female would have a higher SEVR curve, especially at the beginning, but the difference between the beginning and end values would be lesser. This corresponds to Figure "ref"fig:femalesevr, which displays mean SEVR curves for females grouped by health status. You can see that the mean curve for the excellent-group has a lower starting value than that of the other groups and the largest difference between its beginning and ending values. It should be noted that the excellent-group curve is lowest just at the beginning, which conflicts the interpretation. However, this is because of the fact that we just focused on the SEVR's 1. principal component score, and model also includes 2. principal component score of SEVR.

Figure 12 shows model-predicted probabilities across the range for 1. principal component score of SEVR for 30-year-old female who have BMI of 25 and average values for principal component scores of haemodynamic variables. You can see from this graph that if this person have low 1. principal component score of SEVR, around -200, they are most likely to answer that their health status is excellent with probability of 0.6 and least likely to answer that their health status is poor or moderate with almost zero probability. Respectively when they have high 1. principal component score of SEVR, around 300, they are about as likely to answer for their health status poor or moderate than health status good. The probability that they answer excellent for their health status is near zero.

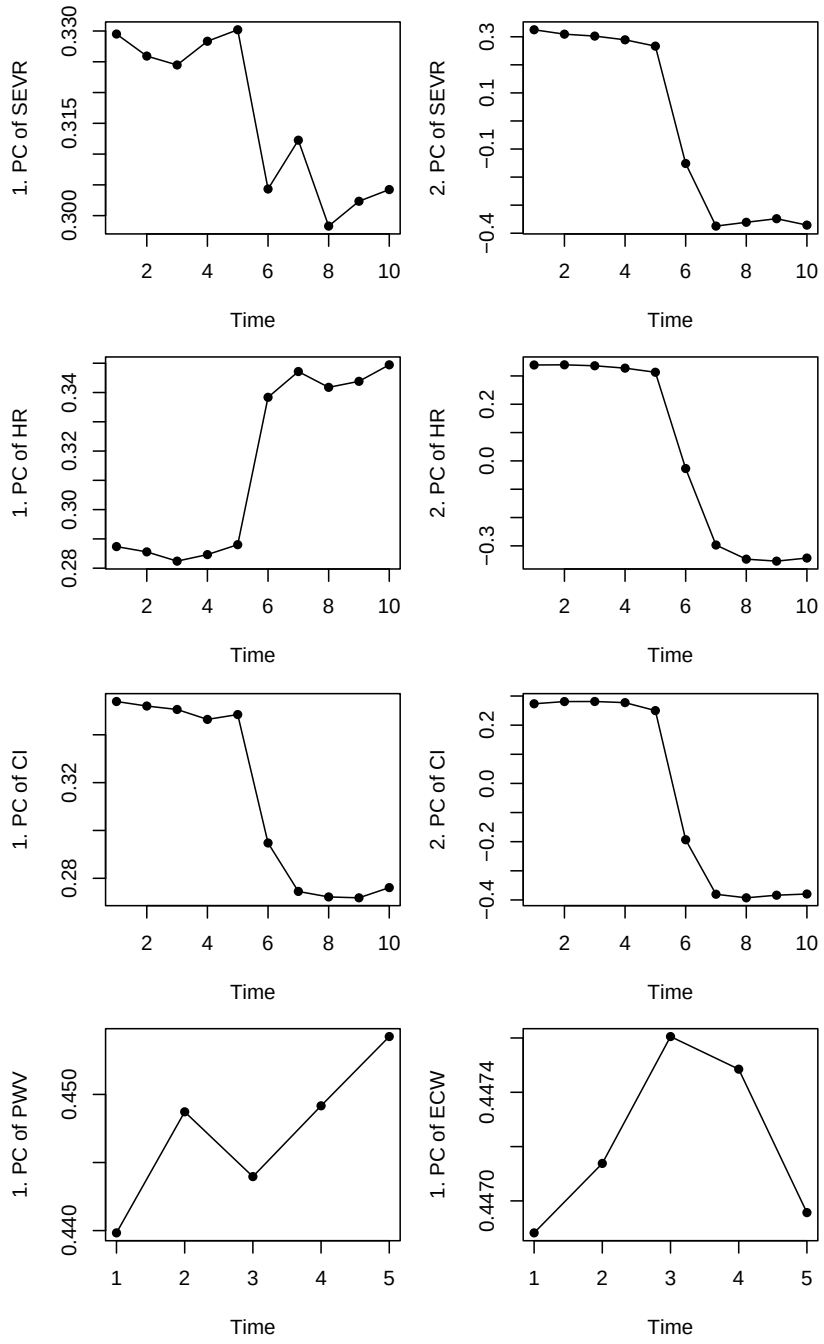
Earlier results are consistent with the boxplot in Figure 13, which shows that the excellent-group has a greater mean and minimum value of 1. principal component score of SEVR when compared to the other groups.

Based on these results, it appears that the model seems to work well with data. However, the proportional odds model has a strict proportional odds assumption that should always be verified. This was done in this thesis by fitting the same model with MASS-package (Venables and Ripley, 2002), so a Brant test could be performed with brant-package (Brant, 1990). The results of this test are shown in Table 4, where you can see that the proportional odds assumption holds with the model's omnibus-value. However, when the test is run on single covariates, it appears that the assumption does not apply to the BMI variable. As a result, we discard this assumption with BMI, while fitting the partial proportional odds model. The next section introduces and discusses this model.

4.3 Fitted partial proportional odds model

Based on previous Brant test results for the fitted proportional odds model, we chose to fit a partial proportional odds model to data where the proportional odds assumption was abandoned with BMI. This model was also fitted with the VGAM-package (Yee et al., 2015), and the model is shown in (13).

Figure 10: Principal component curves corresponding with principal components scores used in the proportional odds model and partial proportional odds model.



$$\begin{aligned}\text{logit}[P(Y_i \leq j \mid \boldsymbol{x}_i)] =& \alpha + \beta_1 sex + \beta_{2j} BMI + \beta_3 age + \beta_4 hr1 \\ &+ \beta_5 hr2 + \beta_6 sevr1 + \beta_7 sevr2 + \beta_8 ci1 \\ &+ \beta_9 ci2 + \beta_{10} pwv + \beta_{11} ecw + \beta_{12} sex \cdot sevr1 \\ &+ \beta_{13} sex \cdot sevr2 + \beta_{14} BMI \cdot pwv + \beta_{15} BMI \cdot ecw \\ &+ \beta_{16} age \cdot hr1 + \beta_{17} age \cdot hr2.\end{aligned} \tag{13}$$

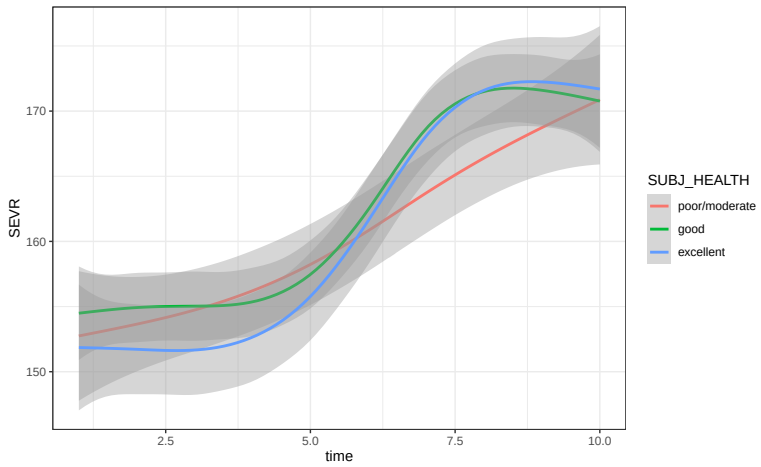


Figure 11: Average mean lines of SEVR for females grouped by health status

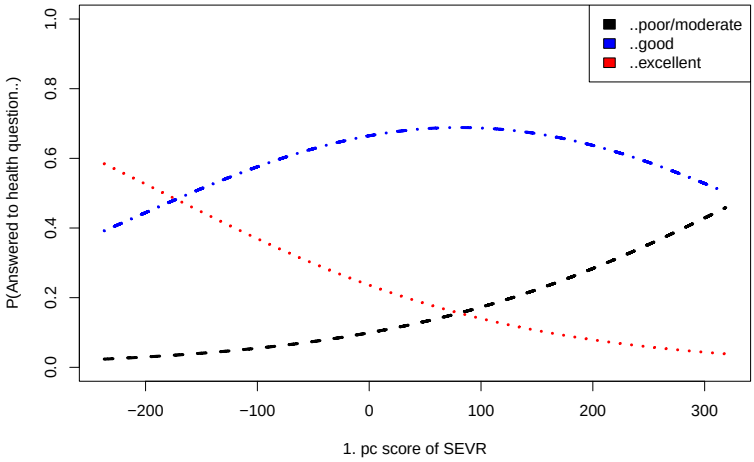
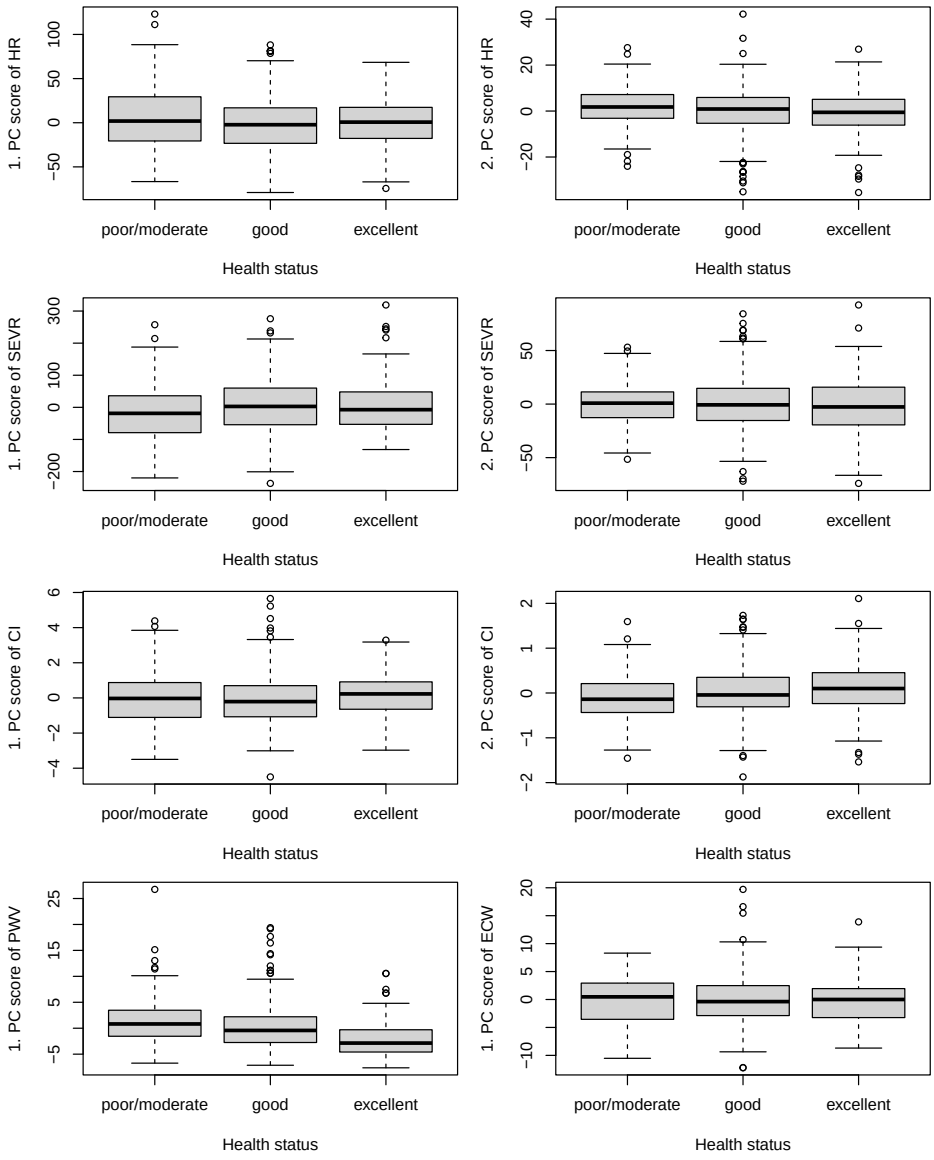


Figure 12: Plots of model-predicted probabilities across the range for 1. principal component score of SEVR for 30-year-old female who have BMI of 25 and average values for principal component scores of haemodynamic variables

Table 5 displays the results of the fitted partial proportional odds model. The results are interpreted similarly to the proportional odds model, and there are not any notable differences in the most of the odds ratios of these two models; for example, the odds ratio of sex in this model is $\exp(-0.2788) \approx 0.757$, whereas in the proportional odds model it was 0.773. This is why we are going to focus on the most notable difference between these two models, the BMI coefficients.

In this model, BMI variable has two coefficients, both of which are statistically significant. When comparing two people with otherwise same values and with zero as values for 1. principal component scores of ECW and PWV, the odds ratios can be interpreted as "The odds that person with one value higher BMI answers health status lower than good is $\exp(0.1238) \approx 1.132$-times higher than the odds that person with lower BMI answers health status lower than good" and

Figure 13: Boxplots of principal component scores used in the proportional odds model and partial proportional odds model grouped by health status.



"The odds that person with lower BMI answers health status lower than good is $\exp(0.2073) \approx$ 1.230-times higher than the odds that person with lower BMI answers health status lower than good" What this means in practice can be better understood by comparing graphs of the partial proportional odds model and the proportional odds model, as shown in Figure 14.

As seen in the graphs, higher BMI affects the probability that a person replies poor or moderate rather than good or excellent less in the partial proportional odds model than in the proportional odds model. In the partial proportional odds model, BMI has a greater influence on the probability that a person responds health status less than excellent than in the proportional odds model. It can

Table 4: Brant test results of the proportional odds model.

Test for	X_2	*df*	*p-value*
Omnibus	15.95	17	0.53
Sex (male)	0.37	1	0.54
BMI	5.67	1	0.02
Age	0.21	1	0.65
1. PC score of HR	0.07	1	0.79
2. PC score of HR	2.96	1	0.09
1. PC score of SEVR	0.27	1	0.60
2. PC score of SEVR	0.27	1	0.79
1. PC score of CI	0.07	1	0.91
2. PC score of CI	0.73	1	0.39
1. PC score of PWV	1.36	1	0.24
1. PC score of ECW	0.30	1	0.59
Sex * 1. PC score of SEVR	0.04	1	0.84
Sex * 2. PC score of SEVR	0.01	1	0.92
BMI * 1. PC score of PWV	1.37	1	0.24
BMI * 1. PC score of ECW	0.20	1	0.65
Age * 1. PC score of HR	0.18	1	0.67
Age * 2. PC score of HR	2.49	1	0.11

also be interpreted that a higher BMI makes it more probable that a person will answer excellent in the proportional odds model than in the partial proportional odds model.

4.4 Fitted functional ordinal logistic regression model

In addition to the previous models, a functional ordinal logistic regression model was fitted to the data using the FREG-package (Samardzic, 2022). The haemodynamic covariates are the same variables used in the proportional odds and partial proportional odds models. The model in this case is shown in (14).

Table 5: Results of the partial proportional odds model.

Coefficient	*Estimate*	*z-value*	*p-value*
1. intercept	-6.1370	-7.248	< 0.0001
2. intercept	-4.9600	-4.896	0.0001
Sex (male)	-0.2788	-1.148	0.2511
BMI:1	0.1238	4.783	< 0.0001
BMI:2	0.2073	5.649	< 0.0001
Age	0.0314	3.463	0.0005
1. PC score of HR	0.0314	2.757	0.0058
2. PC score of HR	0.0373	1.153	0.2490
1. PC score of SEVR	0.0065	2.846	0.0044
2. PC score of SEVR	0.0101	1.695	0.0900
1. PC score of CI	0.0240	0.268	0.7888
2. PC score of CI	-0.4814	-2.132	0.0330
1. PC score of PWV	0.2874	2.123	0.0338
1. PC score of ECW	-0.3092	-2.490	0.0128
Sex * 1. PC score of SEVR	-0.0046	-2.249	0.0245
Sex * 2. PC score of SEVR	-0.0055	-0.779	0.4362
BMI* 1. PC score of PWV	-0.0104	-2.156	0.0311
BMI * 1. PC score of ECW	0.0086	1.979	0.0478
Age * 1. PC score of HR	-0.0004	-2.078	0.0377
Age * 2. PC score of HR	-0.0002	-0.304	0.7614

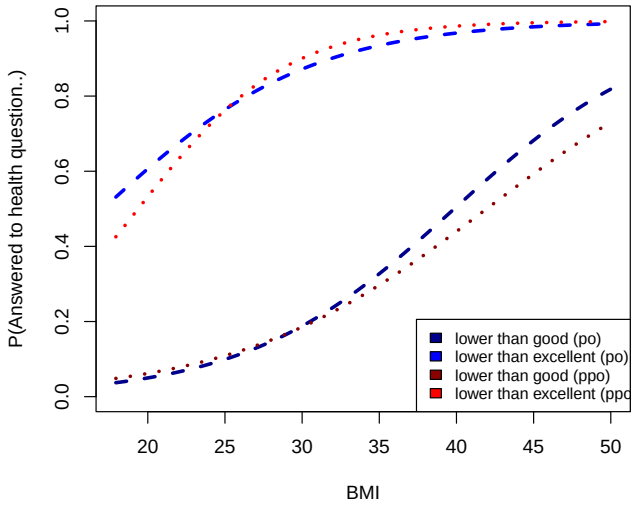


Figure 14: Plots of proportional and partial proportional odds model-predicted cumulative probabilities across range of BMI for female aged 25 with average principal component scores of haemodynamic variables.

$$
\begin{aligned}
\text{logit}[P(Y_i \leq j \mid \boldsymbol{x}_i)] =& \alpha_j + \beta_1 sex + \beta_2 BMI + \beta_3 age + \int_0^1 \beta_4(t) x_{hr}(t)\, dt \\
&+ \int_0^1 \beta_5(t) x_{sevr}(t)\, dt + \int_0^1 \beta_6(t) x_{ci}(t)\, dt + \int_0^1 \beta_7(t) x_{pwv}(t)\, dt \\
&+ \int_0^1 \beta_8(t) x_{ecw}(t)\, dt + \int_0^1 \beta_9(t) sex \cdot x_{sevr}(t)\, dt \\
&+ \int_0^1 \beta_{10}(t) BMI \cdot x_{pwv}(t)\, dt + \int_0^1 \beta_{11}(t) BMI \cdot x_{ecw}(t)\, dt \\
&+ \int_0^1 \beta_{12}(t) age \cdot x_{hr}(t)\, dt.
\end{aligned} \tag{14}
$$

It should be noted that all functions have intervals of [0,1] rather than [1,10] or [1,5], which has no effect on the model.

The FDA-package (Ramsay et al., 2022) was used to define the model's B-spline basis functions. To accomplish this, model knots had to be specified. Because the ECW- and PWV-variables had 5 observation points and the rest of the variables had 10, the maximum number of knots was 5 for ECW and PWV and 10 for the rest of them, when the B-spline basis function used was cubic functions with 3 degrees of freedom, so the maximum number of chosen basis function was 3 and 8. Figure 15 shows that the smoothed means for the ECW and PWV variables are straight, which is why the knots were positioned evenly at the beginning, middle, and finish. Smoothed means for CI, HR, and SEVR appear to be the most "lively" after the fifth observation point, which is why the majority of knots were put in the center and after that. Placements for knots and accompanying B-spline basis functions are shown in Figure 16.

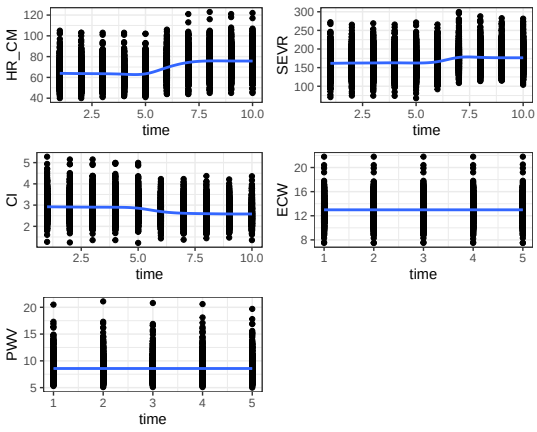


Figure 15: Plots of smoothed means of haemodynamic covariates.

Because coefficients in the functional ordinal logistic regression model are in the form of functions, interpreting beta coefficients of variables is more difficult than

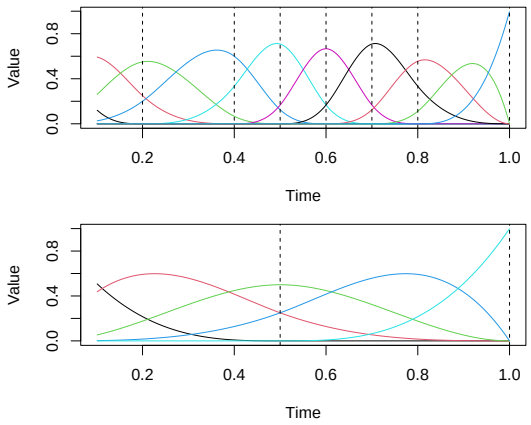


Figure 16: Cubic B-spline basis functions (df=3) used with functional ordinal logistic regression model. In upper plot, the knots used with basis functions are placed at 0.0, 0.2, 0.4, 0.5, 0.6, 0.7, 0.8 and 1.0. In lower plot, the knots are placed at 0.0, 0.5 and 1.0.

in the proportional odds and partial proportional odds models. Even though their coefficients are single numbers, interpreting age, BMI, and sex is also difficult since the model includes interaction terms between them and haemodynamic variables. Interpretation is hard in this case also because, unlike principal component scores of haemodynamic variables, we cannot assume for interpretation that haemodynamic variables have a value of zero (for most of the haemodynamic variables, this means that the subject's heart is not beating). That is why it is easier to plot them and make interpretations about these graphs, which are shown in Figure 17.

According to the graphs, most regression coefficients signs alter fairly evenly between positive and negative (most of the graphs tend to change around the middle, where the tilt happened), meaning that the average effect is low in the case of the majority of haemodynamic variables. This makes determining whether the overall effect is positive or negative for the majority of them difficult. We will interpret the HR's regression coefficient's graph as an example. The effect of HR seems to be positive in overall, but when the table is tilted, the effect briefly turns negative. Therefore, it appears that people who have higher HR at the start and end of the test have a higher odds of answering lower health status than people with lower HR. However, the odds of answering lower health status are lower for person whose HR is higher at the time of the tilt than for person whose HR is lower.

4.5 Model comparisons

We can now assess and compare how well the models fit the data after specifying the models to be fitted. Because there was some discussion about whether haemodynamic variables could be used to predict a health measure, or whether age, gender, and BMI were sufficient predictors of that variable alone, a comparison

Table 6: Coefficients of the sex, BMI and age in functional ordinal logistic regression model.

Covariate	*Coefficient*
Sex (male)	-0.1790
BMI	0.1548
Age	0.0393

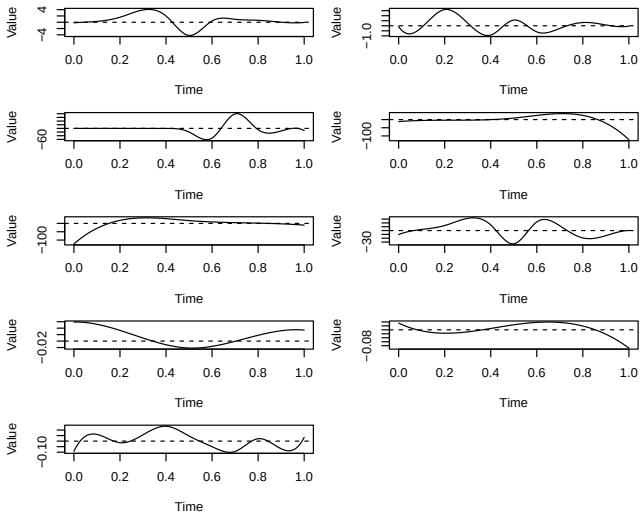


Figure 17: Graphs of the functional ordinal logistic regression model's regression coefficients. From top to bottom on left: HR, CI, ECW, BMI · PWV, age · HR. And on right: SEVR, PWV, sex · SEVR, BMI · ECW

was made with reduced models of earlier models that only included age, gender, and BMI as covariates.

Because the comparison in this situation was between a nested model and a reduced model, the pseudo-R^2 values of the models could be compared. Table 7 compares the proportional odds model to and the comparisons of the partial proportional odds models are shown in Table 8. In both situations, the full model gives higher pseudo-R^2 values, indicating that models with selected haemodynamic variables outperform those without. The same conclusion could be made with model AIC-values, which can be found in Table 9. At least with the proportional odds model and the partial proportional odds model, the full model has lower AIC-values than those without haemodynamic variables.

Because the function used with this model requires at least one functional variable, we could not compare the full functional ordinal logistic regression model to the corresponding model without haemodynamic variables. Instead, we compared the full model to the corresponding model with only one haemodynamic variable, heart rate. In this case, the full model had higher AIC-value than the reduced model, meaning that the chosen model might not be a good fit.

Earlier results seem to indicate that fitted functional ordinal logistic regression model might not be as good fit as proportional odds model and partial proportional odds model. Comparing AIC-values of these three models support this, because functional ordinal logistic regression model has the highest value. Best fit based on AIC-values seems to be the partial proportional odds model.

Table 7: Pseudo-R^2s of the proportional odds models (full model and reduced model with only sex, age and BMI as covariates).

Pseudo-R^2	*Full model*	*Reduced model*
McFadden	0.1201	0.0825
Cox & Snell	0.2024	0.1440
Nagelkerke	0.2387	0.1698

Table 8: Pseudo-R^2s of the partial proportional odds models (full model and reduced model with only sex, age and BMI as covariates).

Pseudo-R^2	*Full model*	*Reduced model*
McFadden	0.1241	0.0888
Cox & Snell	0.2084	0.1540
Nagelkerke	0.2457	0.1816

Table 9: AIC-values of proportional odds model (PO), partial proportional odds model (PPO), and functional ordinal logistic regression model (FOLR) and reduced models with only BMI, age and sex (and HR in case of functional ordinal logistic regression model) as covariates.

Model	*AIC (Full)*	*AIC (Reduced)*
PO	1134.77	1153.59
PPO	1131.814	1147.77
FOLR	1191.21	1150.64

5 Conclusions

Based on the results, it appears that the partial proportional odds model gives the best fit for this data while the functional ordinal logistic regression model gives the worst one. In the proportional odds model and partial proportional odds model, including haemodynamic covariates HR, SEVR, CI, ECW, PWV, and interactions between sex and SEVR, BMI and ECW/PWV, and age and HR gave better results than reduced models with only age, sex, and BMI as covariates. So using haemodynamic variables to model self-assessed health status instead of using only age, BMI and sex, seems to be advisable. However, it should be noted that this kind of comparison could not be done with functional ordinal logistic regression model. Instead the full model was compared to reduced model with age, BMI, sex and HR. The full model gave higher AIC than the reduced model meaning that it gives worse fit than the reduced model.

It should be pointed out, however, that the haemodynamic variables used in the models were chosen using a forward backward step regression done on a proportional odds model that contained principal component scores of haemodynamic variables rather than the original values. In other words, the covariates may have been chosen differently, for example, by selecting haemodynamic variables with the lowest AIC for the functional ordinal logistic regression model and using them in all of the models, or by comparing models with the best fit and with different haemodynamic variables. This is howerer more difficult to do because there are no tools available for this in R. This is also why, unlike other models, AIC-values of functional ordinal logistic regression were simply compared, while with others the pseudo-R^2s were also compared. While doing this thesis it became obvious that the functional ordinal logistic regression model is a fresh model with many development opportunities.

It should be noted that, based on the AIC partial and proportional odds model does not seems to differ much with their fit. However, based on the Brant test results of the proportional odds model and partial proportional odds model's results it seems that proportional odds assumption does not hold true for BMI. As previously stated, the proportional odds model does not frequently hold with real-world data, hence the partial proportional odds model having the best fit came as no surprise. Because of this, one obvious area of development for the functional ordinal logistic regression model would be to create a more extended version of it that allows you to relax the proportional odds assumption with single variables (or all of them).

One way to do this comparison differently is to use different goodness-of-fit tests. In this thesis, we use AIC and pseudo-R^2s, however BIC might also have been used. We may have also used a different method than forward backward step regression to select principal component scores for the proportional odds model. For example, Wenbin and Zhang (2007) in their article have recommended using ALASSO penalty.

References

Agresti, A. (2010). *Analysis of Ordinal Data*. John Wiley and Sons, Gainesville, Florida, 2nd edition.

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, 46(4):1171–1178.

Chiu, C.-Y., Wang, S., Zhang, B., Luo, Y., Simpson, C., Zhang, W., Wilson, A., Bailey-Wilson, J. E., Agron, E., Chew, E., Zhang, J., Ziong, M., and Fan, R. (2022). Gene-level association analysis of ordinal traits with functional ordinal logistic regressions. *Genetic Epdemiology*, 46(5-6):234–255.

Harrell, F.E., J. (2001). *Regression Modeling Strategies*. Springer, Charlottesville.

Hu, B., Shao, J., and Palta, M. (2006). Pseudo-R^2 in logistic regression model. *Statistica Sinica*, 16(3):847–860.

Jacques, J. and Samardzic (2022). Analyzing cycling sensors data through ordinal logistic regression with functional covariates. *Applied Statistics. Series C*, 71(4):969–986.

James, G., Hastie, T., and Sugar, C. (2000). Principal component models for sparse functional data. *Biometrika*, 87(3):587–602.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer, UK.

Liu, X. (2022). *Categorical Data Analysis and Multilevel Modeling Using R*. SAGE Publications.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. SAGE Publications, Bloomington, 2nd edition.

Mateu, J. and Giraldo, R. (2022). *Geostatistical Functional Data Analysis*. John Wiley and Sons, 2nd edition.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B*, 40(2):109–142.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behaviour. *Economic Theory and Mathematical Economics. Frontiers in Econometrics*, pages 105–142.

Nagelkerke, N. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692.

Peterson, B. and Harrell, F.E., J. (1990). Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society. Series C*, 39(2):205–217.

Ramsay, J. and Dalzell, C. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society: Series B*, 53(3):505–700.

Ramsay, J. O., Graves, S., and Hooker, G. (2022). *fda: Functional Data Analysis*. R package version 6.0.5.

Samardzic, S. (2022). *FREG: Functional Regression Models*. R package version 1.1.

Tahvanainen, A. (2011). *Whole Body Impedance Cardiography and Continuous Pulse Wave Analysis in the Measurement of Human Haemodynamics during Passive Head-up Tilt*. PhD thesis, University of Tampere.

Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition.

Wang, J.-L., Chiou, J.-M., and Muller, H.-G. (2016). Review of functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295.

Wenbin, L. and Zhang, H. H. (2007). Variable selection for proportional odds model. *Statistics in Medicine*, 26:3771–3781.

Windmeijer, F. A. (1995). Goodness-of-fit measures in binary choice models. *Econometric Reviews*, 14(1):101–116.

Yee, T. W., Stoklosa, J., and Huggins, R. M. (2015). The VGAM package for capture-recapture data using the conditional likelihood. *Journal of Statistical Software*, 65(5):1–33.

A R code for forward backward step regression and choosing haemodynamic covariates for models

```
# Data
library(foreign)
data <- read.spss("2022-02-11_Data_n751.sav",
            to.data.frame=TRUE)

#Subset with variables chosen for thesis
data1 <- data[,c(2,3,6,57,374:393, 504:513, 534:543,
              624:628,639:643, 699:708,679:688, 584:603,
              614:623, 1)]
health <- as.factor(data1$SUBJ_HEALTH)

#omitting rows with missing data
df1 <- na.omit(data1)

# combining poor and moderate groups
levels(df1$SUBJ_HEALTH) <- c("poor/moderate",
                             "poor/moderate","good",
                             "excellent")



# Principal component analysis for haemodynamic variables
# RAD_SAP

#calculating means for 1.-10. measurements of RAD_SAP
keskiarvot <- aggregate(df1[, 5:14], list(df1$SUBJ_HEALTH), mean)
keskiarvot <- keskiarvot[,2:11]
keskiarvot
X <- df1[,(5:14)]
X <- rbind(X, keskiarvot)
X1 <- na.omit(X)
AveSAP<- colMeans(X1)

#mean centering
Xc <- sweep(X1,2,AveSAP,"-")

#Principal component analysis
SAPPCA <- princomp(Xc)
summary(SAPPCA)
```

```
screeplot(SAPPCA)
pairs(SAPPCA$scores[,1:2], col=df1$SUBJ_HEALTH)

#Seperating principal component scores as their own variable
vpca1 <-SAPPCA$loadings
vpca2 <- unclass(vpca1)
vpca3 <- solve(t(vpca2))

#Mean values for health groups were removed, so subset
#would include only principal component scores for
# subjects
SAPPCAscores <- SAPPCA$scores[-c(637:639), ]

#Code is similar with prinicipal component analysis performed
# for other haemodynamic variables, so comments are not
#included with them.
# RAD_DAP
keskiarvot <- aggregate(df1[, 15:24], list(df1$SUBJ_HEALTH), mean)
keskiarvot <- keskiarvot[,2:11]
keskiarvot
X <- df1[,(15:24)]
X <- rbind(X, keskiarvot)
X1 <- na.omit(X)
AveDAP<- colMeans(X1)
Xc <- sweep(X1,2,AveDAP,"-")
DAPPCA <- princomp(Xc)
summary(DAPPCA)
screeplot(DAPPCA)
pairs(DAPPCA$scores[,1:2], col=df1$SUBJ_HEALTH)
vpca1 <-DAPPCA$loadings
vpca2 <- unclass(vpca1)
vpca3 <- solve(t(vpca2))
DAPPCAscores <- DAPPCA$scores[-c(637:639), ]

# SVRI_RAD
keskiarvot <- aggregate(df1[, 65:74], list(df1$SUBJ_HEALTH), mean)
keskiarvot <- keskiarvot[,2:11]
keskiarvot
X <- df1[,(65:74)]
X <- rbind(X, keskiarvot)
X1 <- na.omit(X)
AveSVRI<- colMeans(X1)
Xc <- sweep(X1,2,AveSVRI,"-")
SVRIPCA <- princomp(Xc)
```

```
summary(SVRIPCA)
screeplot(SVRIPCA)
pairs(SVRIPCA$scores[,1:2], col=data1$SUBJ_HEALTH)
vpca1 <-SVRIPCA$loadings
vpca2 <- unclass(vpca1)
vpca3 <- solve(t(vpca2))
cor(X, SVRIPCA$scores)
SVRIPCAscores <- SVRIPCA$scores[-c(637:639), ]

# PWV
keskiarvot <- aggregate(df1[, 45:49], list(df1$SUBJ_HEALTH), mean)
keskiarvot <- keskiarvot[,2:6]
keskiarvot
X <- df1[,(45:49)]
X
X <- rbind(X, keskiarvot)
X1 <- na.omit(X)
AvePWV<- colMeans(X1)
Xc <- sweep(X1,2,AvePWV,"-")
PWVPCA <- princomp(Xc)
summary(PWVPCA)
screeplot(PWVPCA)
pairs(PWVPCA$scores[,1:2], col=data1$SUBJ_HEALTH)
vpca1 <-PWVPCA$loadings
vpca2 <- unclass(vpca1)
vpca3 <- solve(t(vpca2))
cor(X, PWVPCA$scores)
PWVPCAscores <- PWVPCA$scores[-c(637:639), ]


# HR_CM
keskiarvot <- aggregate(df1[, 35:44], list(df1$SUBJ_HEALTH), mean)
keskiarvot <- keskiarvot[,2:11]
keskiarvot
X <- df1[,(35:44)]
X <- rbind(X, keskiarvot)
X1 <- na.omit(X)
AveHR<- colMeans(X1)
Xc <- sweep(X1,2,AveHR,"-")
HRPCA <- princomp(Xc)
summary(HRPCA)
screeplot(HRPCA)
variance = HRPCA$sdev^2 / sum(HRPCA$sdev^2)
variance
```

```
pairs(HRPCA$scores[,1:2], col=data1$SUBJ_HEALTH)
vpca1 <-HRPCA$loadings
vpca2 <- unclass(vpca1)
vpca3 <- solve(t(vpca2))
HRPCAscores <- HRPCA$scores[-c(637:639), ]

# SEVR
keskiarvot <- aggregate(df1[, 25:34], list(df1$SUBJ_HEALTH), mean)
keskiarvot <- keskiarvot[,2:11]
keskiarvot
X <- df1[,(25:34)]
X
X <- rbind(X, keskiarvot)
X1 <- na.omit(X)
AveSEVR<- colMeans(X1)
Xc <- sweep(X1,2,AveSEVR,"-")
SEVRPCA <- princomp(Xc)
summary(SEVRPCA)
screeplot(SEVRPCA)
pairs(SEVRPCA$scores[,1:2], col=data1$SUBJ_HEALTH)
vpca1 <-SEVRPCA$loadings
vpca2 <- unclass(vpca1)
vpca3 <- solve(t(vpca2))
SEVRPCAscores <- SEVRPCA$scores[-c(637:639), ]

# CI
keskiarvot <- aggregate(df1[,75:84], list(df1$SUBJ_HEALTH), mean)
keskiarvot <- keskiarvot[,2:11]
keskiarvot
X <- df1[,(75:84)]
X
X <- rbind(X, keskiarvot)
X1 <- na.omit(X)
AveCI<- colMeans(X1)
Xc <- sweep(X1,2,AveCI,"-")
CIPCA <- princomp(Xc)
summary(CIPCA)
screeplot(CIPCA)
pairs(CIPCA$scores[,1:2], col=data1$SUBJ_HEALTH)
vpca1 <-CIPCA$loadings
vpca2 <- unclass(vpca1)
vpca3 <- solve(t(vpca2))
CIPCAscores <- CIPCA$scores[-c(637:639), ]
```

```
# ECW
keskiarvot <- aggregate(df1[, 50:54], list(df1$SUBJ_HEALTH), mean)
keskiarvot <- keskiarvot[,2:6]
keskiarvot
X <- df1[,(50:54)]
X
X <- rbind(X, keskiarvot)
X1 <- na.omit(X)
AveECW<- colMeans(X1)
Xc <- sweep(X1,2,AveECW,"-")
ECWPCA <- princomp(Xc)
summary(ECWPCA)
screeplot(ECWPCA)
pairs(ECWPCA$scores[,1:2], col=data1$SUBJ_HEALTH)
vpca1 <-ECWPCA$loadings
vpca2 <- unclass(vpca1)
vpca3 <- solve(t(vpca2))
ECWPCAscores <- ECWPCA$scores[-c(637:639), ]

# AIX
keskiarvot <- aggregate(df1[, 55:64], list(df1$SUBJ_HEALTH), mean)
keskiarvot <- keskiarvot[,2:11]
keskiarvot
X <- df1[,(55:64)]
X
X <- rbind(X, keskiarvot)
X1 <- na.omit(X)
AveAIX<- colMeans(X1)
Xc <- sweep(X1,2,AveAIX,"-")
AIXPCA <- princomp(Xc)
summary(AIXPCA)
screeplot(AIXPCA)
pairs(AIXPCA$scores[,1:2], col=data1$SUBJ_HEALTH)
vpca1 <-AIXPCA$loadings
vpca2 <- unclass(vpca1)
vpca3 <- solve(t(vpca2))
AIXPCAscores <- AIXPCA$scores[-c(637:639), ]

# Combining principal component scores and health,
# age, sex and BMI as one data.frame
health <- df1$SUBJ_HEALTH
age <- df1$AGE_CALCULATED
sex <- df1$SEX_01
BMI <-df1$BMI
```

```
BMIgroup <- df1$BMIgroup
agegroup <- df1$agegroup
aix1 <- AIXPCAscores[,1]
aix2 <- AIXPCAscores[,2]
svri1 <- SVRIPCAscores[,1]
svri2 <- SVRIPCAscores[,2]
hr1 <- HRPCAscores[,1]
hr2 <- HRPCAscores[,2]
radsap1 <- SAPPCAscores[,1]
radsap2 <- SAPPCAscores[,2]
raddap1 <- DAPPCAscores[,1]
raddap2 <- DAPPCAscores[,2]
sevr1 <- SEVRPCAscores[,1]
sevr2 <- SEVRPCAscores[,2]
ci1 <- CIPCAscores[,1]
ci2 <- CIPCAscores[,2]
ecw <- ECWPCAscores[,1]
pwv <- PWVPCAscores[,1]

tdata <- data.frame(health, age, sex, BMI, aix1, aix2, svri1,
                              svri2, radsap1, radsap2, raddap1,
                              raddap2, hr1, hr2, sevr1, sevr2,
                              ci1, ci2, ecw, pwv)

#removing missing values (in this case, does not remove
#anything
tdata <- na.omit(tdata)



#Choosing of the haemodynamic covariates

library(MASS)

# Model with all variables and interactions and forward backward
# step regression performed to it
modt1 <- polr(formula = health ~ (sex + BMI + age)*(aix1+
                     hr1 + hr2 + raddap1 + + sevr1 + sevr2 + ci1 +
                     ci2 + pwv + ecw) + sex:BMI + sex:age +
                     age:BMI, data = tdata, Hess = T)
step(modt1, direction = "both")

# Comparing model chosen with step to similar model,
# where all chosen haemodynamic variables have all
```

```
# principal component scores in the model

#Models
vs1 <- polr(formula = health ~ sex + BMI + age + hr1 +
                  hr2 + sevr1 + sevr2 + ci1 + ci2 + pwv +
                  ecw + sex:sevr1 + sex:sevr2 + BMI:pwv +
                  BMI:ecw + age:hr1 + age:hr2, data = tdata,
            Hess = TRUE)
vs2 <- polr(formula = health ~ sex + BMI + age + hr1 +
                  hr2 + sevr1 + sevr2 + ci2 + pwv + ecw +
                  sex:sevr1 + BMI:pwv + BMI:ecw + age:hr1,
                  data = tdata, Hess = TRUE)

#Comparing pseudo-R^2 and AIC values
DescTools::PseudoR2(vs1, which = c("McFadden", "CoxSnell",
                                   "Nagelkerke", "AIC"))
DescTools::PseudoR2(vs2, which = c("McFadden", "CoxSnell",
                                   "Nagelkerke", "AIC"))
```

B R code used to fit models and compare them

```
# Models are fitted with subset of data
# including only variables used in the models

library(foreign)
data <- read.spss("2022-02-11_Data_n751.sav",
            to.data.frame=TRUE)
data1 <- data[,c(2,3,6,57, 504:513, 534:543, 584:593,
              624:628, 639:644 )]
health <- as.factor(data1$SUBJ_HEALTH)

#removing missing data
df1 <- na.omit(data1)

#combining poor and moderate to one group
levels(df1$SUBJ_HEALTH) <- c("poor/moderate",
                             "poor/moderate","good",
                             "excellent")


#Principal component analysis
```

```
# SEVR
keskiarvot <- aggregate(df1[, 5:14],
                  list(df1$SUBJ_HEALTH), mean)
keskiarvot <- keskiarvot[,2:11]
keskiarvot
X <- df1[,(5:14)]
X
X <- rbind(X, keskiarvot)
X1 <- na.omit(X)
AveSEVR<- colMeans(X1)
Xc <- sweep(X1,2,AveSEVR,"-")
SEVRPCA <- princomp(Xc)
summary(SEVRPCA)
screeplot(SEVRPCA)
pairs(SEVRPCA$scores[,1:2],
         col=data1$SUBJ_HEALTH)
vpca1 <-SEVRPCA$loadings
vpca2 <- unclass(vpca1)
vpca3 <- solve(t(vpca2))
sevrvpca <- vpca3
SEVRPCAscores <- SEVRPCA$scores[-c(663:665), ]

# HR_CM
keskiarvot <- aggregate(df1[, 15:24],
                  list(df1$SUBJ_HEALTH), mean)
keskiarvot <- keskiarvot[,2:11]
keskiarvot
X <- df1[,(15:24)]
X <- rbind(X, keskiarvot)
X1 <- na.omit(X)
AveHR<- colMeans(X1)
Xc <- sweep(X1,2,AveHR,"-")
HRPCA <- princomp(Xc)
summary(HRPCA)
screeplot(HRPCA)
variance = HRPCA$sdev^2 / sum(HRPCA$sdev^2)
variance
pairs(HRPCA$scores[,1:2], col=data1$SUBJ_HEALTH)
vpca1 <-HRPCA$loadings
vpca2 <- unclass(vpca1)
vpca3 <- solve(t(vpca2))
hrvpca <- vpca3
HRPCAscores <- HRPCA$scores[-c(663:665), ]
```

```
# CI
keskiarvot <- aggregate(df1[,25:34],
                    list(df1$SUBJ_HEALTH), mean)
keskiarvot <- keskiarvot[,2:11]
keskiarvot
X <- df1[,(25:34)]
X
X <- rbind(X, keskiarvot)
X1 <- na.omit(X)
AveCI<- colMeans(X1)
Xc <- sweep(X1,2,AveCI,"-")
CIPCA <- princomp(Xc)
summary(CIPCA)
screeplot(CIPCA)
pairs(CIPCA$scores[,1:2],
         col=data1$SUBJ_HEALTH)
vpca1 <-CIPCA$loadings
vpca2 <- unclass(vpca1)
vpca3 <- solve(t(vpca2))
civpca <- vpca3
CIPCAscores <- CIPCA$scores[-c(663:665), ]

# PWV
keskiarvot <- aggregate(df1[, 35:39],
                    list(df1$SUBJ_HEALTH), mean)
keskiarvot <- keskiarvot[,2:6]
keskiarvot
X <- df1[,(35:39)]
X
X <- rbind(X, keskiarvot)
X1 <- na.omit(X)
AvePWV<- colMeans(X1)
Xc <- sweep(X1,2,AvePWV,"-")
PWVPCA <- princomp(Xc)
summary(PWVPCA)
screeplot(PWVPCA)
pairs(PWVPCA$scores[,1:2],
        col=data1$SUBJ_HEALTH)
vpca1 <-PWVPCA$loadings
vpca2 <- unclass(vpca1)
vpca3 <- solve(t(vpca2))
pwvvpca <- vpca3
PWVPCAscores <- PWVPCA$scores[-c(663:665), ]
```

```
# ECW
keskiarvot <- aggregate(df1[, 40:44],
                    list(df1$SUBJ_HEALTH), mean)
keskiarvot <- keskiarvot[,2:6]
keskiarvot
X <- df1[,(40:44)]
X
X <- rbind(X, keskiarvot)
X1 <- na.omit(X)
AveECW<- colMeans(X1)
Xc <- sweep(X1,2,AveECW,"-")
ECWPCA <- princomp(Xc)
summary(ECWPCA)
screeplot(ECWPCA)
pairs(ECWPCA$scores[,1:2],
        col=data1$SUBJ_HEALTH)
vpca1 <-ECWPCA$loadings
vpca2 <- unclass(vpca1)
vpca3 <- solve(t(vpca2))
ecwvpca <- vpca3
ECWPCAscores <- ECWPCA$scores[-c(663:665), ]

par(mfrow=c(3,2))
screeplot(HRPCA)
screeplot(SEVRPCA)
screeplot(CIPCA)
screeplot(PWVPCA)
screeplot(ECWPCA)

# Combining health, age, sex and BMI columns
# with principal component score of haemodynamic
# variables

health <- df1$SUBJ_HEALTH
age <- df1$AGE_CALCULATED
sex <- df1$SEX_01
BMI <-df1$BMI
hr1 <- HRPCAscores[,1]
hr2 <- HRPCAscores[,2]
sevr1 <- SEVRPCAscores[,1]
sevr2 <- SEVRPCAscores[,2]
ci1 <- CIPCAscores[,1]
ci2 <- CIPCAscores[,2]
```

```
ecw <- ECWPCAscores[,1]
pwv <- PWVPCAscores[,1]

tdata <- data.frame(health, age, sex, BMI,
                    hr1, hr2, sevr1, sevr2, ci1, ci2, ecw, pwv)

#removing missing values
tdata <- na.omit(tdata)

# PO-model fitted with polr-function and performing
#  Brant-test

library(MASS)
polrmod <- polr(formula = health ~ sex + BMI + age + hr1 +
                          hr2 + sevr1 + sevr2 + ci1 + ci2 + pwv +
                         ecw + sex:sevr1 + sex:sevr2 + BMI:pwv +
                         BMI:ecw + age:hr1 + age:hr2, data = tdata,
                        Hess = TRUE)
library(brant)
brant(polrmod)

#Fitting of the po-model and ppo-model with vglm-function

library(VGAM)
pomod <- vglm(formula = ordered(health) ~ sex + BMI + age +
                         hr1 + hr2 + sevr1 + sevr2 + ci1 + ci2 +
                         pwv + ecw + sex:sevr1 +  sex:sevr2 +
                         BMI:pwv + BMI:ecw + age:hr1 + age:hr2,
                  family = cumulative(parallel = TRUE,
                                      reverse = FALSE))
ppomod <- vglm(formula = ordered(health) ~ sex + BMI + age +
                           hr1 + hr2 + sevr1 + sevr2 + ci1 + ci2 +
                           pwv +  ecw + sex:sevr1 + sex:sevr2 +
                           BMI:pwv + BMI:ecw + age:hr1 + age:hr2,
           family = cumulative(parallel = FALSE ~ BMI,
                                      reverse = FALSE))
# Reduced models for comparison

povs <- vglm(formula = ordered(health) ~ sex + BMI + age,
             family = cumulative(parallel = TRUE,
             reverse = FALSE))
ppovs <- vglm(formula = ordered(health) ~ sex + BMI + age,
               family = cumulative(parallel = FALSE ~ BMI,
               reverse = FALSE))
```

```
# Results of full models
summary(pomod)
summary(ppomod)

#Comparing pseudo-R^2s and AIC values of full models and reduced mode
library(rcompanion)
nagelkerke(pomod)
nagelkerke(povs)
nagelkerke(ppomod)
nagelkerke(ppovs)
AIC(pomod)
AIC(ppomod)
AIC(povs)
AIC(ppovs)

#Fitting of the FOLR-model

# Basis functions

library(fda)

#Time range for basis functions
times_basis = seq(0.0,1,0.1)

# Location of knots
knots1    = c(0.0, 0.2,0.4, 0.5, 0.6, 0.7, 0.8, 1)
knots2    = c(0,0.5,1)

 #Number of knots
n_knots1   = length(knots1)
n_knots2   = length(knots2)

# order of basis functions: cubic bspline: order = 3 + 1
n_order   = 4

#Creating the basis functions
n_basis1   = n_knots1 + n_order - 2;
n_basis2   = n_knots2 + n_order - 2;
xbasis = create.bspline.basis(c(min(times_basis),
             max(times_basis)),n_basis1,n_order,knots1)
xbasis2 = create.bspline.basis(c(min(times_basis),
               max(times_basis)),n_basis2,n_order,knots2)
time <- sort(runif(636,0.1,1))
```

```
# Functions of haemodynamic variables used in the model
HR_CM <- t(df1[,15:24])
HR_CM.fd = smooth.basis(c(seq(0,1,length=10)),
                  HR_CM,xbasis)$fd
SEVR <- t(df1[,5:14])
SEVR.fd <- smooth.basis(c(seq(0,1,length=10)),
                 SEVR,xbasis)$fd
CI <- t(df1[,25:34])
CI.fd <- smooth.basis(c(seq(0,1,length=10)),
            CI,xbasis)$fd
ECW <- t(df1[,40:44])
ECW.fd <- smooth.basis(c(seq(0,1,length = 5)),
                ECW,xbasis2)$fd
PWV <- t(df1[,35:39])
PWV.fd <- smooth.basis(c(seq(0,1, length = 5)),
                PWV,xbasis2)$fd
# Other covariates
health <- as.factor(df1$SUBJ_HEALTH)
health <- factor(health, levels=rev(levels(health)))
levels(ordered(health))
levels(as.matrix(health))
length(levels(health))
age <- df1$AGE_CALCULATED
sex <- df1$SEX_01
levels(sex) <- c(0,1)
sex <- as.numeric(sex)
BMI <- df1$BMI
SEVRSEX <- SEVR.fd*sex
PWVBMI <- PWV.fd*BMI
ECWBMI <- ECW.fd*BMI
HRAGE <- HR_CM.fd*age

# Fitting and the result

library(FREG)

# olfreg-function of FREG-package, but with a fix
#so the model uses right order of the categories

olfregfix = function(formula, betalist = NULL){

  call = match.call()
```

```
# extract y from formula

y.name = formula[[2]]
# y = get(as.character(y.name)) # search y by name
y = ordered(health)
# easy fix so the model uses order poor/moderate < good <
# excellent instead of the other way around
y.len = length(y)

if(inherits(y, c("numeric", "matrix", "array"), FALSE))
  stop("Y has to be factor")

# extract independent variables from formula

x.var = all.vars(formula)[-1]
x.count = length(x.var)

xfdlist = vector('list', length = x.count)
# stock them in the list
names(xfdlist) = x.var

type = c()
nbasis = c()

df = lapply(x.var, get)
no = which(lapply(df, class)=="fd")
if(length(no)>1){
  range = get(x.var[no[1]])$basis$range
 # take range from the first fd
}else range = get(x.var[no])$basis$range

# if (inherits(get(x.var), what = "fd")){
#   x.fun = get(x.var)
#   range = x.fun$basis$rangeval
# }else stop("Please enter a functional covariate")

bbasis.names = vector('list', length = x.count)
# to stock beta basis names
betalist = vector('list', length = x.count)
for(i in 1:x.count){

  x = get(x.var[i])

  if(inherits(x, what = "fd")){
```

```
  type[i] = x$basis$type
  nbasis[i] = x$basis$nbasis
  x.len = dim(x$coefs)[2]
  #range = x$basis$rangeval
}else if(inherits(x, what = "numeric")){
  cbasis = create.constant.basis(rangeval = range)
  x.len = length(x)
  x = fd(matrix(x,1,y.len),cbasis)
}

if(x.len != y.len)
  stop('The number of observations of ',x.var[i],
       ' is ', x.len,
       ' and is not equal to the number of observations of y ',
         y.len)

if(!class(x) %in% c("fd", "numeric"))
  stop('Variable ', x.var[i],
  ' has to be either fd or numeric')

xfdlist[[i]] = x

# create betalist

if(is.null(betalist)){
  betalist = vector('list', length = x.count)
}
if(is.null(betalist[[i]])){
  if(class(x) %in% "fd"){
    bbasis = with(x, fd(basis = basis,
    fdnames = fdnames))$basis
  }else if(class(x) %in% "numeric"){
    bbasis = create.constant.basis(rangeval = range)
  }

  betalist[[i]] = bbasis

}else if(length(betalist) != length(xfdlist)){

  stop('length(betalist) is ', length(betalist),
       ' but it  must be equal to the number of independent varia
         length(xfdlist))

  betaclass = sapply(betalist, class)
```

```
    wrong = which(betaclass != 'basisfd')

    if(length(wrong) > 0)
      stop('All components of betalist must have class basisfd')
  }

  bbasis.names[[i]] = betalist[[i]]$names
}

# estimation

p = length(xfdlist)
# constant and independent functional variables
y = as.matrix(y)
#N = dim(y)[1]
# number of observations
Z  = NULL
# for any number of covariates
for (i in 1:p) {
  xfdi       = xfdlist[[i]]
  xcoef      = xfdi$coefs
  xbasis     = xfdi$basis
  bbasis     = betalist[[i]]
  basis.prod  = romberg_alg(xbasis,bbasis)
  Z          = cbind(Z,
                 crossprod(xcoef, basis.prod))

}

x = Z
n = nrow(x)
xc = ncol(x)
wt = rep(1, n) # weights
ind_xc = seq_len(xc)
if(!is.factor(y)) y = ordered(y,
                      levels = c("poor/moderate",
                      "good", "excellent"))
# as.factor(y)
ylev = levels(y)
lylev = length(ylev)
q = length(ylev)-1L
ind_q = seq_len(lylev-1L)
y = unclass(y)
```

```
  coefs = rep(0, xc)
  logit = function(p) log(p/(1 - p))
  # qlogis for quantiles
  taby = tabulate(y)
  alphas = cumsum(taby)[-length(taby)]/n
  # space
  initial = logit(alphas)

  start = c(coefs, initial)

  res = optimization(x, y, start, loglik, gradient,
          Hessian)$beta
  beta = res[seq_len(xc)]
  alpha = res[xc + ind_q]
  names(alpha) = paste("y <=",
                           ylev[1:length(ylev)-1])
  names(beta) = paste("X", unlist(bbasis.names),
                          sep = ".")

  # fitted values
  eta = as.vector(x %*% beta)
  cumprob = plogis(matrix(alpha, n, q, byrow = TRUE) - eta)
  fitted.values = as.matrix(t(apply(cumprob, 1,
                         function(x) diff(c(0, x, 1)))))
  colnames(fitted.values) = ylev
  # additional output
  loglik = optimization(x, y, start, loglik, gradient,
              Hessian)$ll
  grd = optimization(x, y, start, loglik, gradient,
            Hessian)$grd
  hessian = optimization(x, y, start, loglik, gradient,
                 Hessian)$hessian
  iteration =  optimization(x, y, start, loglik, gradient,
                     Hessian)$iter

# calculate degrees of freedom and AIC
  loglik = -loglik
# return the actual value of log-likelihood and not
# the negative one
  df = length(alpha) + length(beta)
  AIC = -2*loglik + 2*df

  instance = list()
```

```
  instance$call = call
  instance$no.var = x.count
  instance$xfdlist = xfdlist
  instance$betalist = betalist
  instance$coefficients = beta
  instance$alpha = alpha
  instance$ylev = ylev
  instance$fitted.values = fitted.values
  instance$loglik = loglik
  instance$grd = grd
  instance$hessian = hessian
  instance$df = df
  instance$AIC = AIC
  instance$iteration = iteration

  class(instance) = "olfreg"
  instance

  return(instance)

}

#Fitting of the folr-model
folrmod <- olfregfix(health ~ sex + BMI + age + HR_CM.fd +
                              SEVR.fd +  CI.fd + PWV.fd + ECW.fd +
                               SEVRSEX + PWVBMI +  ECWBMI + HRAGE)

#Summary and AIC value of the folr-model
folrmod$AIC
summary(folrmod)

#you should note that in this model when calculating odds ratios,
# minus sign should be added to value of coefficients, so they
# can be interpreted same way as in po- and ppo-model
exp(-folrmod$coefficients[1:3])

# FREG-packages plot_olfreg-function with fix, so functions of
# coefficients can be interpreted similar way as earlier models
# coefficients

plot_olfregfix = function(object){

  if(inherits(object, "olfreg")){
```

```
    betacoeff = list()
    p = length(object$xfdlist)

    beta = -object$coefficients
    alphas = -object$alpha
    nalphas = length(object$alpha)

   #fix done to the original function,
   #marginals were changed so figure
   #would not be too big
    par(mar = c(4,4,1,1))
    par(mfrow = c(5,2))


    j = 1
    for (i in 4:p){
      betafd = fdPar(object$betalist[[i]],0,0)$fd
      # create and extract fd object of each variable and store it
      nbeta = betafd$basis$nbasis
      # extract the number of basis aka coeffs
      m = j
      j = j + nbeta-1

      betacoef = beta[m:j]
      betafd$coefs = as.matrix(betacoef)

      betacoeff[[i]] = betafd
      plt = plot(betacoeff[[i]],
                 xlab = "Time", ylab = "Value")

    }
    return(plt)

  }else stop('Model has to be of the class olfreg')
}

#Plotting the regression coefficients of folr-model (Figure 16)
plot_olfregfix(folrmod)
```