Andrea Borenich

Statistical analysis of maximal cardiac exercise stress testing

Master's Thesis

to achieve the university degree of
Master of Science

submitted to
Graz University of Technology

Supervisor
Univ.-Prof.i.R. Dipl.-Ing. Dr.techn. Ernst Stadlober

Co-Supervisor
Mag. Heimo Traninger

Institute of Statistics

Graz, April 2018

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

....

(date) (signature)

Acknowledgement

I would like to express my deepest gratitude towards all the people who have helped and supervised me with this research project. Without their support, patience and guidance, I would never have been able to finish my thesis.

First I want to thank my supervisor Univ.-Prof.i.R. Dipl.-Ing. Dr.techn. Ernst Stadlober for the competent supervision and support while writing this thesis. Furthermore I want to thank the medical institute ZARG and ProDoc, from whom I got the data. I would also like to thank Jenny Chapman and Piyush Gampawar for helping me with scientific written English. Finally, I would like to thank my parents and friends who have always supported, encouraged and continued to believe in me and in all my accomplishments.

Abstract

Maximal cardiac stress exercise testing is one of the common health checks especially for heart diseases. One of the crucial results is the maximal workload which a patient achieves during the examination. The aim of this thesis is to derive normal values of the maximal workload using a regression analysis with variables selection. This is done with the data from the medical institute ZARG and ProDoc, which include in total 28,625 ergometries. The resulting normal values are compared with the normal values, which are currently recommended in the guidelines of Austria. The newly calculated normal values have a better goodness of fit, especially for women, than the currently used normal values. Furthermore, the analysis of in total 1,453 ergometries at entry examinations of stage II and stage III of cardiac out-patient rehabilitation is done to calculate the normal values of maximum workload. Finally, the improvement of maximum workload during stage II and stage III of cardiac out-patient rehabilitation is analyzed.

Zusammenfassung

Die Ergometrie ist eines der Standard-Gesundheitschecks speziell bei Herzerkrankungen. Eine der wesentlichen Resultate ist die maximale Leistung, die ein Patient während der Untersuchung erreicht. Das Ziel dieser Arbeit ist es mit Hilfe einer Regressionsanalyse mit Variablenselektion Normalwerte für die maximale Leistung zu erhalten. Dies wird mit Hilfe der Daten des medizinischen Instituts ZARG gemacht. Insgesamt standen dafür Daten von 28.625 Ergometrien zur Verfügung. Die resultierenden Normalwerte werden mit den in Österreich momentan in den Praxisleitlinien empfohlenen Normalwerten verglichen. Die neu berechneten Normalwerte besitzen eine bessere Anpassungsgüte als die momentan verwendeten Normalwerte, speziell für Frauen. Weiters beinhaltet die Arbeit Normalwerte für die maximale Leistung wie auch Normalwerte für die Verbesserung der maximalen Leistung für Ergometrien während der Eingangsuntersuchung von Phase II und Phase III der ambulanten Rehabilitation. Für diese Analyse wurden insgesamt Daten von 1.453 Ergometrien verwendet.

Contents

List of figures ii

List of tables ii

List of abbreviations xv

1 Introduction 1

2 Medical literature overview 3
2.1 Cardiovascular diseases . . . 3
2.2 Risk factors for cardiovascular diseases . . . 4
2.3 Maximal cardiac exercise stress testing . . . 4
2.4 Normal values for the maximum workload of maximal cardiac exercise stress testing . . . 5
2.5 Medications . . . 6

3 Statistical literature overview 7
3.1 Statistical programming language R . . . 7
3.2 Exploratory data analysis . . . 7
3.2.1 Numerical summary measures . . . 8
3.2.2 Box plot . . . 10
3.2.3 Bar plot . . . 12
3.2.4 Histogram and density plot . . . 12
3.2.5 Plot of the empirical distribution function . . . 15
3.2.6 Quantile-quantile plot . . . 16
3.2.7 Scatter plot . . . 17
3.2.8 Correlation coefficient . . . 18
3.3 Multiple linear regression . . . 20
3.3.1 Multiple linear regression model . . . 21
3.3.2 Estimating the regression coefficients . . . 22
3.3.3 Analysis of variance and the coefficient of determination . 25
3.3.4 Diagnostics of the model . . . 27
3.3.5 Analysis of collinearity . . . 28
3.3.6 Box-Cox transformation . . . 29
3.3.7 Confidence intervals and prediction intervals . . . 30
3.4 Variable selection . . . 30

3.4.1 Steps in selecting the best regression equation 30
3.4.2 Specification of the maximum model 31
3.4.3 Selection criteria 31
3.4.4 Specification of a strategy for selecting variables 33
3.5 Dummy regression 38
3.6 Regression surfaces 38
3.7 Decision trees 40

4 Data preparation **45**
4.1 Received data 45
4.1.1 XML files for data of the ergometries 45
4.1.2 CSV files for the blood parameters, diagnosis and medication 45
4.2 Step 1: Transformation of the XML files in a CSV file 47
4.3 Step 2: Integration of the CSV files in MySQL database 48
4.4 Step 3: Integration of the blood parameters, diagnosis and medication in the table maindata with Java programs 48
4.4.1 Columns for the blood parameters 48
4.4.2 Column for diagnosis 49
4.4.3 Columns for medication 49
4.5 Step 4: Reading the table maindata into R 49
4.6 Step 5: Adaptation of the table maindata in R 49
4.7 Variables used in the statistical analysis 50

5 Exploratory data analysis **51**
5.1 Univariate analysis 51
5.1.1 Composition of the data regarding gender, medication and diagnosis 51
5.1.2 Statistical indicators of the variables 52
5.1.3 Analysis of the variables age, height, weight and waist circumference 53
5.1.4 Analysis of the additionally calculated variables body mass index and waist-to-height-ratio 56
5.1.5 Analysis of the blood pressure 57
5.1.6 Analysis of the heart rate 58
5.1.7 Analysis of the blood parameters 59
5.1.8 Additionally calculated variables TC to HDL ratio, LDL to HDL ratio and TG to HDL ratio 61
5.1.9 Analysis of the maximum workload and maximum workload-to-weight-ratio 62
5.2 Multivariate analysis 64
5.2.1 Correlation 65
5.2.2 Scatter plots 67
5.3 Quality control of the data 68
5.4 Discussion 69

6 Normal values for the maximum workload **71**
6.1 Correlation 71
6.2 Regression trees 72
6.3 Model P 73

6.4 Model 1 74
6.4.1 Analysis of the model 74
6.4.2 Diagnostics of the model 75
6.4.3 Confidence intervals and prediction intervals for Model 1 . 76
6.4.4 Regression surface of Model 1 77
6.5 Model 2 79
6.5.1 Variable selection 79
6.5.2 Analysis of the model 81
6.5.3 Diagnostics of the model 81
6.5.4 Confidence intervals and prediction intervals for Model 2 . 83
6.5.5 Regression surface of Model 2 84
6.6 Model 3 85
6.6.1 Variable selection 85
6.6.2 Analysis of the model 87
6.6.3 Diagnostics of the model 88
6.6.4 Confidence intervals and prediction intervals for Model 3 . 90
6.6.5 Regression surface of Model 3 91
6.7 Model 4 92
6.7.1 Variable selection 93
6.7.2 Analysis of the model 95
6.7.3 Diagnostics of the model 96
6.7.4 Confidence intervals and prediction intervals for Model 4 . 97
6.7.5 Regression surface of Model 4 98
6.8 Comparison 100
6.8.1 Summary of the models 100
6.8.2 Comparison of the sum of squares and the adjusted coefficient of determination 101
6.8.3 Comparison of the surfaces 102
6.9 Model D 103
6.9.1 Regression tree 103
6.9.2 Variable selection 103
6.9.3 Analysis of the model 105
6.9.4 Diagnostics of the model 105
6.10 Discussion 106

7 Cardiac out-patient rehabilitation **109**
7.1 Composition of the data 110
7.2 Normal values of the entry examination of stage II 111
7.2.1 Correlation 112
7.2.2 Regression trees 113
7.2.3 Variable selection 113
7.2.4 Analysis of the model 116
7.2.5 Diagnostics of the model 117
7.3 Contribution during stage II 118
7.3.1 Correlation 119
7.3.2 Regression trees 121
7.3.3 Regression model 121
7.4 Normal values of the entry examination of stage III 121

7.4.1 Correlation . . . 121
7.4.2 Regression trees . . . 123
7.4.3 Variable selection . . . 123
7.4.4 Analysis of the model . . . 126
7.4.5 Diagnostics of the model . . . 127
7.5 Contribution during stage III . . . 128
7.5.1 Correlation . . . 129
7.5.2 Regression trees . . . 131
7.5.3 Regression model . . . 132
7.6 Discussion . . . 132

8 Conclusion and outlook for further research 133
8.1 Summary . . . 133
8.2 Limitation and Recommendations . . . 134
8.3 Further research questions . . . 134

Bibliography 135

Appendices 139

A XSL stylesheet 141

B MySQL script 145

C Java program for diagnosis 149

D Java program for medication 157

E Java program for laboratory 165

F R script 173

List of figures

3.1 Box plots for the variable maximum workload 11
3.2 Number of ergometries seperated by gender 13
3.3 Histogram for the variable maximum workload 14
3.4 Density plot for the variable maximum workload 15
3.5 Plot of the empirical distribution function for the variable maximum workload 16
3.6 Q-Q plot for the variable maximum workload 18
3.7 Scatter plot for the variables age and maximum workload 19
3.8 Correlation matrix between the variables weight, height, age and maximum workload 20
3.9 Diagnostic plot of the example linear regression model 28
3.10 Box-Cox plot of the example linear regression model 29
3.11 Results for different selection criteria for the example linear regression 34
3.12 Information criteria for the example linear regression 35
3.13 Plots of the regression line and confidence bands of the regression surface 39
3.14 Plots of the regression surface of the maximum workload depending on weight and height 40
3.15 Regression tree for the example linear regression 41
3.16 Regression tree for the example linear regression with the package `party` 43

4.1 Workflow for the data preparation 46

5.1 Number of ergometries separated by gender, medication and diagnosis 52
5.2 Exploratory data analysis for the variable age 54
5.3 Exploratory data analysis for the variable height 54
5.4 Exploratory data analysis for the variable weight 55
5.5 Exploratory data analysis for the variable waist 55
5.6 Exploratory data analysis for the variable BMI 56
5.7 Exploratory data analysis for the variable WHR 56
5.8 Exploratory data analysis for the variable diastolic blood pressure 57
5.9 Exploratory data analysis for the variable systolic blood pressure . 57
5.10 Increase of the blood pressure during ergometry 58

5.11 Exploratory data analysis for the variable heart rate 59
5.12 Scatter plot for the resting heart rate and the increase of the heart rate during an ergometry 59
5.13 Box plots for the variables TC, LDL, HDL, TG 60
5.14 Box plots for Cr, FBS and UA 61
5.15 Box plots for the calculated variables TC:HDL, LDL:HDL and TG:HDL 62
5.16 Exploratory data analysis for the variable maximum workload . . 63
5.17 Exploratory data analysis for the variable maximum-workload-to-weight-ratio 64
5.18 Correlation between the variables of the basic information and the measured parameters during an ergometry, the WL and the WWR for all ergometries together 65
5.19 Correlation of the blood parameters, the WL and the WWR of all ergometries together 66
5.20 Scatter plots between the variables of the basic information and the maximum workload-to-weight-ratio 67
5.21 Scatter plots between the calculated varaibles of the basic information and the maximum workload-to-weight-ratio 68
5.22 Quality control for the maximum workload during time 69

6.1 Analysis of correlation with the variables weight, height, age, BMI, BSA, maximum workload and maximum workload-to-weight-ratio 72
6.2 Regression trees for males with the independent variables weight, height, age, body surface area and the dependent variables maximum workload and maximum workload-to-weight-ratio 72
6.3 Regression trees for females with the independent variables weight, height, age, BSA and the dependent variables maximum workload and maximum workload-to-weight-ratio 73
6.4 Box-Cox plots of Model 1 75
6.5 Diagnostic plots of Model 1 for males 76
6.6 Diagnostic plots of Model 1 for females 77
6.7 Plots of the regression surface of Model 1 78
6.8 Plots of the regression line and confidence bands of the BSA for different ages of Model 1 79
6.9 Information criteria for deriving Model 2 80
6.10 Box-Cox plots of Model 2 82
6.11 Diagnostic plots of Model 2 for males 82
6.12 Diagnostic plots of Model 2 for females 83
6.13 Plots of the regression surface for Model 2 85
6.14 Plots of the regression lines and confidence bands of the surface for different ages of Model 2 85
6.15 Plots of the regression lines and confidence bands of the age for different surfaces of Model 2 86
6.16 Information criteria for deriving Model 3 87
6.17 Box-Cox plot of Model 3 89
6.18 Diagnostic plots of Model 3 for males 89
6.19 Diagnostic plots of Model 3 for females 90

6.20 Plots of the regression surface for Model 3 92
6.21 Plots of the regression lines and confidence bands of the surface for different ages for Model 3 92
6.22 Plots of the regression lines and confidence bands of the age for different surfaces for Model 3 93
6.23 Information criteria for deriving Model 4 94
6.24 Box-Cox plots of Model 4 96
6.25 Diagnostic plots of Model 4 for males 96
6.26 Diagnostic plots of Model 4 for females 97
6.27 Plots of the regression surface for Model 4 100
6.28 Plots of the regression lines and confidence bands of the weight for different ages of Model 4 100
6.29 Plots of the regression lines and confidence bands of the age for different surfaces of Model 4 101
6.30 Surface of Model P and Model 1: orange surface belongs to Model P and the blue surface belongs to Model 1 102
6.31 Scatter plot of the maximum workloads for Model P and Model 1 102
6.32 Regression tree for the variables of Model D 103
6.33 Information criteria for deriving Model D 104
6.34 Box-Cox plots of Model D 106
6.35 Diagnostic plots of Model D 106

7.1 Number of ergometries during cardiac out-patient rehabilitation split by the reason for examination and gender 110
7.2 Number of ergometries during cardiac out-patient rehabilitation separated by medication or diagnosis 111
7.3 Box plots of the maximum workload during the different stages of cardiac out-patient rehabilitation split by gender 112
7.4 Analysis of correlation with the variables weight, height, body mass index, age, waist, waist-to-height-ratio, maximum workload and maximum workload-to-weight-ratio for the entry examination of stage II 113
7.5 Regression trees with the independent variables weight, height, age, body mass index, waist, waist-to-height-ratio and the dependent variable maximum workload-to-weight-ratio for the entry examination of stage II 114
7.6 Information criteria for deriving Model RII 115
7.7 Box-Cox plots of Model RII 117
7.8 Diagnostic plots of Model RII for males 117
7.9 Diagnostic plots of Model RII for females 118
7.10 Graphical illustration of the improvement of the maximum workload in stage II of cardiac out-patient rehabilitation 119
7.11 Correlation matrix for stage II of cardiac out-patient rehabilitation between the variables height, age, weight, waist, body mass index, waist-to-height-ratio and the improvement 120

7.12 Correlation matrix for stage II of cardiac out-patient rehabilitation between the variables maximum workload at entry examination of stage II, maximum workload-to-weight-ratio at entry examination of stage II, duration of stage II and the improvement 120
7.13 Regression trees for the improvement of stage II with the independent variables weight, height, age, waist-to-height-ratio and the dependent variable maximum workload-to-weight-ratio 122
7.14 Analysis of correlation with the variables weight, height, body mass index, age, waist, waist-to-height-ratio, maximum workload and maximum workload-to-weight-ratio for the entry examination of stage III 123
7.15 Regression trees with the independent variables weight, height, age, body mass index, waist, waist-to-height-ratio and the dependent variable maximum workload-to-weight-ratio for the entry examination of stage III 124
7.16 Information criteria for deriving Model RIII 125
7.17 Box-Cox plots of Model RIII 127
7.18 Diagnostic plots of Model RIII for males 127
7.19 Diagnostic plots of Model RIII for females 128
7.20 Graphical illustration of the improvement of the maximum workload in stage III of cardiac out-patient rehabilitation 129
7.21 Correlation matrix for stage III of cardiac out-patient rehabilitation between the variables height, age, weight, waist, body mass index, waist-to-height-ratio and the improvement 130
7.22 Correlation matrix for stage III of cardiac out-patient rehabilitation between the variables maximum workload at entry examination of stage III, maximum workload-to-weight-ratio at entry examination of stage II, duration of stage III and the improvement 130
7.23 Regression trees for the improvement of stage III with the independent variables weight, height, age, waist-to-height-ratio and the dependent variable maximum workload-to-weight-ratio 131

List of tables

4.1 Description of the variables in CSV file blood parameters 46
4.2 Description of the variables in Table diagnosis 47
4.3 Description of the variables in Table medication 47
4.4 Description of the variables in Table maindata 48
4.5 Description of the variables used in the statistical analysis 50

5.1 Descriptive statistics of all variables with the following indicators: minimum, 1-quartile, median, average, 3-quartile, maximum, standard deviation and the number of missing values 53

6.1 Results of Model P 74
6.2 Summary and ANOVA of Model 1 for males 74
6.3 Summary and ANOVA of Model 1 for females 75
6.4 Prediction values for the maximum workload [W] (first row), confidence intervals (second row) and prediction intervals (third row) of Model 1 for males 77
6.5 Prediction values for the maximum workload [W] (first row), confidence intervals (second row) and prediction intervals (third row) of Model 1 for females 78
6.6 Summary of the best models with 1 to 3 parameters for deriving Model 2 for males and females 79
6.7 Summary and ANOVA of Model 2 for males 81
6.8 Summary and ANOVA of Model 2 for females 81
6.9 Prediction values for the maximum workload [W] (first row), confidence intervals (second row) and prediction intervals (third row) of Model 2 for males 84
6.10 Prediction values for the maximum workload [W] (first row), confidence intervals (second row) and prediction intervals (third row) of Model 2 for females 84
6.11 Summary of the best models with 1 to 3 parameters for deriving Model 3 86
6.12 Summary and ANOVA of Model 3 for males 88
6.13 Summary and ANOVA of Model 3 for females 88
6.14 Prediction values for the maximum workload-to-weight-ratio [W/kg] (first row), confidence intervals (second row) and prediction intervals (third row) of Model 3 for males 91

6.15 Prediction values for the maximum workload-to-weight-ratio [W/kg] (first row), confidence intervals (second row) and prediction intervals (third row) of Model 3 for females 91
6.16 Summary of the best models with 1 to 7 parameters for deriving Model 4 for males 93
6.17 Summary of the best models with 1 to 7 parameters for deriving Model 4 for females 94
6.18 Summary and ANOVA of Model 4 for males 95
6.19 Summary and ANOVA of Model 4 for females 95
6.20 Prediction values for the maximum workload-to-weight-ratio [W/kg] (first row), confidence intervals (second row) and prediction intervals (third row) of Model 4 for males 98
6.21 Prediction values for the maximum workload-to-weight-ratio [W/kg] (first row), confidence intervals (second row) and prediction intervals (third row) of Model 4 for females 99
6.22 Comparison of the Models for males 101
6.23 Comparison of the Models for females 101
6.24 Summary of the best models with one to six parameters for deriving Model D 104
6.25 Summary and ANOVA of Model D 105

7.1 Summary of the best models with one to four parameters for deriving Model RII for males 113
7.2 Summary of the best models with one to four parameters for deriving Model RII for females 115
7.3 Summary of Model RII for males 116
7.4 Summary of Model RII for females 116
7.5 Descriptive statistics of the improvement during stage II: minimum, 1-quartile, median, average, 3-quartile, maximum, standard deviation and the number of missing values 119
7.6 Summary of the best models with 1 to 5 parameters for deriving Model RIII for males 123
7.7 Summary of the best models with 1 to 5 parameters for deriving Model RIII for females 125
7.8 Summary of Model RIII for males 126
7.9 Summary of Model RIII for females 126
7.10 Descriptive statistics of the improvement during stage III: minimum, 1-quartile, median, average, 3-quartile, maximum, standard deviation and the number of missing values 129

List of abbreviations

AIC Akaike's Information Criterion

AIC$_c$ Corrected Akaike's Information Criterion

ANOVA Analysis of Variance

BIC Bayesian Information Criterion

BMI body mass index

BSA body surface area

C$_p$ Mallow's C_p

CAD Coronary Artery Disease

CP complexity parameter

Cr creatinine

DBP diastolic blood pressure

DBP$_{max}$ diastolic blood pressure at maximum workload

DBP$_r$ resting diastolic blood pressure

DIAG diagnosis

EDA Explorative Data Analysis

FBS fasting blood sugar level

HDL high density lipoprotein

HR heart rate

HR$_{max}$ heart rate at maximum workload

HR$_r$ resting heart rate

LDL low density lipoprotein

MED medication

MSE mean square error

MSR mean square regression

MST mean square total

Q-Q plot Quantile-Quantile plot

RSE residual standard error

SBP$_{max}$ systolic blood pressure at maximum workload

SBP$_r$ resting systolic blood pressure

SSE sum of squares error

SSR sum of squares regression

SST sum of squares total

SBP systolic blood pressure

TC total cholesterol

TG triglyceride

UA urid acid

VLDL very low density lipoprotein

WHR waist to height ratio

WL maximum workload

WWR workload to weight ratio

ZARG Zentrum für ambulante Rehabilitation

Chapter 1

Introduction

Cardiovascular diseases are responsible for 45% of all deaths in Europe. In 2015, more than 85 million people across Europe were living with cardiovascular diseases, 48% of cases were in males and 52% in females. In that year, Austria had 399,912 females and 409,542 males suffering from cardiovascular diseases [Wilkins et al., 2017]. For more than six decades now, maximal cardiac exercise stress testing is used to evaluate the functional capacity of the heart non-invasively and is an efficient and a cost effective way. Moreover, it is useful in determining the risk and extent of coronary artery diseases (CAD), prognosis and prediction of cardiovascular events, and the effect of therapy [Fletcher et al., 2013]. The testing may involve ergometers such as stationary bicycles, treadmills, rowing machines, arm ergometer and monitoring of cardiac activities by electrocardiography and blood pressure monitoring (see e.g. [Pina et al., 1995]). Such tests involve exercising at several workloads or exercise intensities, during which cardiovascular and respiratory measurements are recorded. Depending on the rationale for testing, the type of test employed, and the health and fitness status of the individual taking part, exercise stress tests might terminate on volitional exhaustion or upon observation of a particular clinical symptom, or at some other predetermined outcome measure [Turner, 2013].

Cardiac rehabilitation is a multidisciplinary approach at physical, social and psychological levels to reduce the risk and secondary progression of cardiovascular diseases and reverse the disease progressions [Price et al., 2016]. Studies have shown the benefits of cardiac rehabilitation and it has been recommended by the European Society of Cardiology, American Heart Association and the American College of Cardiology for the treatment of cardiovascular diseases [Bjarnason-Wehrens et al., 2010]. Traditionally, cardiac rehabilitation is divided into three stages (I, II, III). Stage I is started when the patient is in the hospital whereas stage II (short-term stage) and stage III (long-term stage) are on outpatient or in-patient basis. Cardiac rehabilitation is provided through different models which differs across Europe depending upon local and national health guidelines [Bjarnason-Wehrens et al., 2010]. The assessment of physical activity using cardiac exercise stress test in clinical settings is an important task for the physicians involved in preventive cardiology [Piepoli et al., 2010]. According to Austrian outpatient rehabilitation guidelines, the cardiac exercise stress test (ergometry) is performed three times, initial (beginning of stage II), inter-

mediate (end of stage II or/and beginning of stage III) and final (end of stage III) [Niebauer et al., 2012].

One of the crucial measured parameters in ergometry is the maximum workload achieved. For the interpretation of the results of an ergometry, the normal values of maximum workload are used. These normal values are expected values, which a person should achieve. The literature for calculating the normal values of the maximum workload during an ergometry is scant. For Austria, [Wonisch et al., 2008] suggested to continue using the same normal values as in the old guideline [Niederberger et al., 1974]. These normal values are dated back to [Arstila, 1972]. Unfortunately, it was not possible to obtain further information on the population and number of subjects used to develop a model for calculating the normal values. These normal values are still used, although they have not been updated for over 40 years. In last few decades, new mathematical models were developed to update the normal values by [Jones et al., 1985] and [Mohammad et al., 2012], however the sample sizes used to calculate normal values were very small (100 and 34 respectively).

The data used in this thesis is provided by *Zentrum für ambulante Rehabilitation Graz* (ZARG). ZARG is a center for out-patient rehabilitation. Currently, ZARG uses the suggested normal values for the maximum workload from [Wonisch et al., 2008], but they are interested in new approaches for normal values. Note that ZARG performs an extra ergometry in addition to suggested guidelines at intermediate examination of stage III.

The aims of this thesis are 1) to evaluate the presently used model in Austria and to develop new models to calculate the normal values of maximum workload ($n = 28625$) and 2) to calculate the normal values of maximum workload of the entry examination for stage II ($n = 779$) and stage III ($n = 674$) of cardiac out-patient rehabilitation and the expected improvement in maximum workload during these stages (stage II: $n = 644$ and stage III: $n = 787$).

For answering these questions, statistical methods are used. For getting an overview about the data, an exploratory data analysis is utilized. For deriving normal values, regression trees as well as a regression analysis is done. For finding the best model, variable selection is used.

In Chapter 2 there is an overview about the medical background. Chapter 3 gives an introduction about the statistical methods used in this thesis. Afterwards Chapter 4 deals with the data preparation and shows the final data which are analyzed in the further thesis. Chapter 5 gives an exploratory data analysis both univariate and multivariate. The first research question to derive normal values for the maximum workload, is answered in Chapter 6. Chapter 7 analyzes the data of cardiac out-patient rehabilitation. Finally, Chapter 8 contains a conclusion and outlook for further research.

Chapter 2

Medical literature overview

In this chapter, a short overview of the relevant medical topics is given. This starts with cardiovascular diseases, risk factors for cardiovascular diseases, cardiovascular stress testing and medications used in treatment of cardiovascular diseases.

2.1 Cardiovascular diseases

The cohort used for the analysis in this thesis consists of patients with different diagnoses such as coronary artery diseases, arrhythmias and fibrillation, conduction disorders and valvular heart diseases. Some of these diseases are explained here in short:

Coronary Artery Disease (CAD) ([DeSilva, 2013, Chapter 3]): Coronary artery disease or ischemic heart diseases are caused due to blockage or narrowing of coronary arteries due to plaques resulting in reduction or complete obliteration of blood to heart muscles. The disease diagnosis, which falls under this category in our cohorts are CAD, myocardial infarction, atherosclerosis, stenosis etc.

Cardiac arrhythmias ([DeSilva, 2013, Chapter 6]): The group of disorders due to irregular rhythm and rate of heartbeat. The heart rate could be too fast or too slow, or the rhythm is irregular. The diagnosis from our cohort, which falls under this category are ventricular tachycardia, atrial fibrillation, fibrillation arrhythmias etc.

Conduction disorders ([DeSilva, 2013, Chapter 6]): The group of cardiac disorders caused due to improper progression of electrical impulses through the heart, or complete blockage thereof. A diagnosis such as disturbance of conduction left bundle branch block, sinoatrial block, atrioventricular blocks from our cohort, falls under this category.

Valvular heart diseases ([DeSilva, 2013, Chapter 4]): It is caused by malfunctioning or nonfunctioning of one or more heart valves. This includes regurgitation - leakage of blood back into the heart chambers and stenosis - due to thickening, stiffening or fusion of valve flaps resulting in failure in the opening of heart valves completely. The diagnoses, which fall under this category are aortic insufficiency, aortic stenosis, mitral insufficiency, mitral regurgitation, tricuspid insufficiency, tricuspid stenosis etc.

2.2 Risk factors for cardiovascular diseases

Conventionally, older age is associated with an increased risk of heart diseases, over 45 years in men and 55 years in women. In addition, family history of early heart diseases put an individual at additional risk. Other risk factors for heart diseases are obesity, high blood pressure, high cholesterol level etc. Obesity is associated with increased risk of vascular risks. Waist circumference, waist-to-high-ratio (WHR) and body mass index (BMI) are the measures of obesity. The surveys indicate that there is a coherence between the waist circumference and the risk of getting a heart attack (see e.g. [De Koning et al., 2007]). Increasing risks arise for the waist circumference of more than 80 cm for women and more than 94 cm for men. A better measurement than the waist circumference is the WHR, which is the quotient between the waist circumference and the height. If this value is above 0.95 for men and above 0.8 for women, they will have a high cardiovascular risk. BMI is calculated from body weight and height by the formula

$$BMI[kg/m^2] = \frac{weight[kg]}{size[m]^2}$$

and is an estimate of body fat. Values above 25 identify overweight and above 30 obesity [Han et al., 1995].

High blood pressure is defined as systolic blood pressure > 140 mmHg and diastolic blood pressure >90 mmHg. It increases the risk for a variety of cardiovascular diseases. High levels of blood cholesterol are associated with greater risk of heart diseases. Total cholesterol level is the sum of three components HDL cholesterol, LDL cholesterol and 20% of triglycerides. HDL cholesterol is, because of its higher level of proteins, considered as good cholesterol. LDL and VLDL are considered as bad cholesterol because of their high level of cholesterol and increased levels are associated with a substantial increase in the risk of heart diseases. Total cholesterol level > 240 mg/dL, LDL level >160 mg/dL and triglycerides >200 mg/dL are considered high whereas HDL levels <40 mg/dL are considered as a major risk factor for heart diseases [Nelson, 2014].

2.3 Maximal cardiac exercise stress testing

Maximal exercise stress testing (ergometry) is a noninvasive way used to assess the diagnostic and prognostic information of cardiovascular diseases, specifically coronary artery disease (CAD). It studies the physical work activity, including that performed by specific muscles or muscle groups. The testing may involve ergometers such as stationary bicycles, treadmills, rowing machines, arm ergometer and monitoring of cardiac activities by electrocardiography and blood pressure monitoring [Pina et al., 1995]. Such tests involve exercising at several workloads or exercise intensities/difficulties, during which cardiovascular and respiratory measurements are recorded. Depending on the rationale for testing, the type of test employed, and the health and fitness status of the individual taking part, exercise stress tests might terminate at volitional exhaustion, upon observation of a particular clinical symptom, or at some other predetermined outcome measure [Turner, 2013].

2.4 Normal values for the maximum workload of maximal cardiac exercise stress testing

The articles addressing the normal values of parameters such as resting oxygen consumption, the maximal oxygen consumption, the maximum workload and the maximal heart rate obtained during the exercise stress testing are abundant (see e.g. [Koch et al., 2009], [Luks et al., 203] and [Brown et al., 2005]). Many articles focus on the prediction of the maximal oxygen consumption, while only few deal with the prediction of the maximum workload. However, this thesis explicitly deals with the maximum workload during ergometry.

In the Austrian practice guidelines for ergometry [Wonisch et al., 2008], the maximum workload is calculated separately for males and females using the following model. First the body surface area is calculated with the Du Bois formula (see [Du Bois and Du Bois, 1916])

$$BSA[m^2] = 0.007148 workload[kg]^{0.425} height[cm]^{0.725}.$$

followed by actual calculation of the maximum workload.

Formula for the maximum workload for men

$$WL_{max}[W] = 6.773 + 136.141 BSA[m^2] - 0.916 BSA[m^2] age[years].$$

Formula for the maximum workload for women

$$WL_{max}[W] = 3.933 + 86.641 BSA[m^2] - 0.346 BSA[m^2] age[years].$$

In addition to the Du Bois formula, the body surface area can be calculated using other formulas; for the calculation of the body surface we refer to [Verbraecken et al., 2006].

There also exist other models to predict the maximum workload (see e.g. [Nogueira and Pompeu, 2006]), but most of them use just small sample sizes. In [Jones et al., 1985], the maximum workload is calculated using

$$WL_{max}[kpm/min] = 20.4 height[cm] - 8.74 age[years] - 288 gender^1 - 1.909 kpm/min.$$

However, these values are calculated using 50 males and 50 females constituting 100 ergometries. In [Mohammad et al., 2012] the maximum workload is predicted with a regression. They developed two formulas

$$WL_{max}[W] = -436 - 1.2 age[years] + 1.8 weight[kg]$$

and

$$WL_{max}[W] = -436 + 3.6 height[cm].$$

This study was a pilot study and involved only 34 participants from Iran.

[1] 0..man, 1..woman

2.5 Medications used in the treatment of cardiovascular diseases

There are many medications used in the treatment of cardiovascular diseases, however, in this thesis only two main drugs, beta-blockers and statins are considered for further evaluation of normal values of maximum workload.

Beta blockers are the group of drugs used to manage high blood pressure and other cardiac ailments. Cardioselective beta blockers are Carvedilol, Dilatrend, Hemangiol, Inderal, Sotacor, Sotahexal and Sotastad. Non cardioselective beta blockers are Atehexal, Atenolan, Atenolol, Beloc, Bilokord, Bisocor, Bisoprolol, Bisostad, Brevibloc, Concor, Esmolol, Hypoloc, Lanoc, Metohexal, Metoprolol, Metoprololsuccinat, Nebivolol, Nomexor, Rapibloc, Rivacor, Seloken and Tenormin.

Statins are the group of compounds used to reduce the lipid levels hence reducing the risk of heart diseases. The types of statins are Atorvadivid, Atorvalan, Atorvastatin, Atozet, Crestor, Fluvastatin, Gerosim, Inegy, Lovastatin, Nyzoc, Panchol, Pravastatin, Simvastad, Simvarcana, Simvastatin, Simvatin, Sortis and Zocord.

Chapter 3

Statistical literature overview

The aim of this chapter is to provide an overview of basic statistics used in analysis of the data in this thesis. Theoretical description of each statistical topic is followed by an example in the statistical programming language R. The remainder of this chapter is organized as follows: Section 3.1 contains an introduction of R. Section **??** deals with the exploratory data analysis. Section 3.3 gives an overview of multiple linear regression. Finally in Section 3.4 the variable selection for multiple linear regression is clarified and in Section 3.5 dummy regression is explained. The following notation is used in this chapter: The observations are notated by $x_1, x_2, \ldots, x_n$ and the ordered observations are notated by $x_{(1)} \leq \cdots \leq x_{(n)}$.

3.1 Statistical programming language R

R is a free language and environment for statistical computing and graphics. There exist different implementations of R for Linux, Windows and MacOS. R is a highly extendable environment used for linear and nonlinear modelling, time-series analysis, classification and clustering etc. Moreover, R provides a high flexibility to plot well-designed publication-quality graphics including mathematical symbols and formulas ([R Core Team, 2017]).

There are several statistical books which use R. E.g. [Wollschläger, 2010] contains a good introduction and the installation guides. [Faraway, 2009] explains use of R to generate linear models. [Muche et al., 2011] uses Microsoft Excel combined with the Add-In 'RExcel', with which R can be used within Excel. [Ramachandran and Tsokos, 2014] provides computer examples of many statistical programs including R for each statistical topic. Additional information about R is available on [Cornelissen, 2018] and for information about the package `ggplot2`, which is used in this thesis, see [Wickham, 2009].

3.2 Exploratory data analysis

Exploratory data analysis (EDA) is a use of tools of the descriptive statistics to summarize and to find structures and coherences in the data. With the help of these methods, patterns and irregularities in one- and multidimensional data sets can be found, in addition, the information about the dispersion, symmetry,

concentration, and outliers can be obtained (see for example [Akkerboom, 2008, Chapter 9], [Burkschat et al., 2012, Chapter 1] or [Fahrmeir et al., 2016, Chapter 2]).

The EDA can be classified in two different ways. The first approach is to classify the EDA in univariate and multivariate analysis. Often as a first step, an univariate analysis is performed in which, as the name suggests, each variable is investigated separately. In a second step the relationship between two or more variables can be analyzed by using multivariate analysis ([Fahrmeir et al., 2016, Chapter 2], [Cleff, 2014, Chapter 3]).

In the second approach, the EDA can be classified by graphical methods and methods involving the calculation of numerical summaries. Examples for graphical methods of the EDA are histograms, box plots, quantile-quantile plots and scatter plots. Examples for methods involving the calculation of numerical summary measures are means, standard deviations and correlation coefficients (see for example [Devore and Berk, 2012, Chapter 1]).

In the following subsections, EDA methods used in the thesis are explained. For this purpose, after explanation an example is given. For this example, the data of patients between the age of 30 and 60 years in year 2016 is analyzed

3.2.1 Numerical summary measures

Numerical summary measures can be used to characterize the data set. They can be characterized in measures of location and measures of variability. The following definitions for these measures are taken from [Devore and Berk, 2012, Chapter 1].

Mean

The sample mean $\bar{x}$ of observations $x_1, x_2, \ldots, x_n$ is given by $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. To calculate the mean with R, the command `mean` can be used.

Example 3.1. The calculation of the mean of the variable *workload_max* of table *exampledata* can be done in R in the following way:

```
mean(exampledata$workload_max)

## [1] 179.3922
```

Therefore, the mean of the maximum workload of the ergometries is 179 W.

Median

The sample median $\tilde{x}$ is indeed the middle value of the ordered observations, i.e. it divides the data set into two parts of equal size. If n is odd, then it is the single middle value $x_{(\frac{n+1}{2})}$, and if n is even, then it is the average of the two middle values, i.e. $\frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})$. To calculate the median with R, the command `median` can be used.

Example 3.2. The calculation of the median of the variable *workload_max* of table *exampledata* can be done in R in the following way:

```
median(exampledata$workload_max)

## [1] 180
```

Therefore, the median of the maximum workload of all ergometries is 180 W.

Quartiles

The quartiles divide the data set into four equal parts. The first quartile $Q_1 = q_{0.25}$ separates the lower quarter from the upper three-quarters, the second quartile $Q_2 = q_{0.5}$ is equivalent to the median and the third quartile $Q_3 = q_{0.75}$ constitutes the upper quarter of the data set. To calculate the quartiles with R the command `quantile` can be used.

Example 3.3. The calculation of the quartiles of the variable *workload_max* of table *exampledata* can be done in R in the following way:

```
quantile(exampledata$workload_max)

##   0%  25%  50%  75% 100%
##   30  142  180  210  400
```

Therefore the minimum of the maximum workload is 30 W, the maximum is 400 W and the median is 180 W (as calculated above). Moreover we know that 25% of the maximum workload is less than 142 W and 25% of the maximum workload is higher than 210 W.

Inter quartile range

The inter quartile range (IQR) is defined as $Q_3 - Q_1$. To calculate the IQR with R the command `IQR` can be used.

Example 3.4. The calculation of the IQR of the variable *workload_max* of table *exampledata* can be done in R in the following way:

```
IQR(exampledata$workload_max)

## [1] 68
```

Therefore the range of 50% of the data around the median is 68 W.

Quantiles

The quantiles are a generalization of the quartiles with the property that at least a relative share of p data are smaller or equal to q_p and maximal a relative share of $(1 - p)$ is bigger than q_p (see for example [Cleff, 2014, Chapter 3]). In R it can be done with the command `quantile`, but additional percentages for calculating the quantiles have to be defined.

Example 3.5. The calculation of the 5%-and 95%-quantiles of the above mentioned variable *workload_max* of table *exampledata* can be done in R in the following way:

```
quantile(exampledata$workload_max, probs=c(0.05, 0.95))

##  5% 95%
## 100 262
```

Therefore 5% of the ergometries have a maximum workload of less than 100 W and 5% of them have a maximum workload of more than 262 W.

Variance and standard deviation

The sample variance s^2 is given by $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. To calculate the variance with R the command `var` can be used.

The sample standard deviation s is the square root of the variance, i.e. $\sqrt{s^2}$. To calculate the standard deviation with R the command `sd` can be used.

Example 3.6. The calculation of the variance and standard deviation of the variable *workload_max* of table *exampledata* can be done in R in the following way:

```
var(exampledata$workload_max)

## [1] 2662.902

sd(exampledata$workload_max)

## [1] 51.60331
```

Therefore the variance of the maximum workload is 2662.9 W and the standard deviation is 51.6 W.

3.2.2 Box plot

A box plot is a pictorial summary used to describe most prominent features of data sets. These include the center, the spread, the extend and nature of any deviation from symmetry and the identification of outliers, which are observations that are located unusually far from the main body of the data (see for example [Devore and Berk, 2012, Chapter 1]).

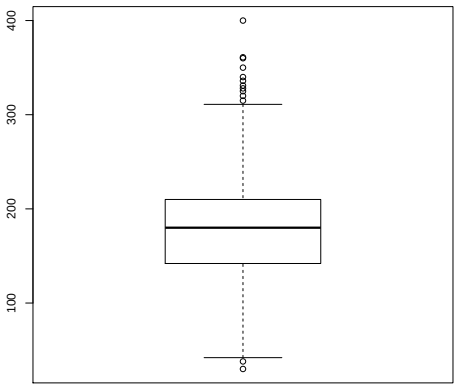
There exist several variations of box plots. The box plots which are created in this thesis provide the following information: The bottom and the top of the boxes are the first and third quartiles $q_{0.25}$ and $q_{0.75}$. The notches display a 95%-confidence interval around the median $\tilde{x}$, i.e. the true median is with a probability of 95% covered by this interval. The width of the boxes is proportional to the number of observations. These are also written in the middle of the box. The upper whisker is defined as $\min(\max(x), Q_3 + 1.5IQR)$ and the lower whisker is $\max(\min(x), Q_1 - 1.5IQR)$. The points which are less than $Q_1 - 1.5IQR$ or

greater than $Q_3 + 1.5IQR$ are classified as outliers and represented by points in the box plots.

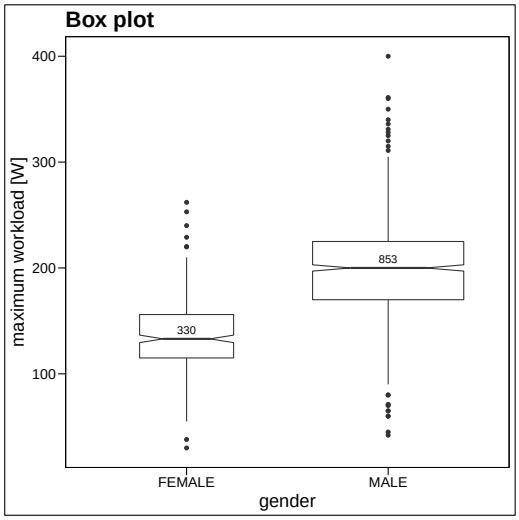
The creation of a box plot can be done in R in different ways. The easiest way is to use the command `boxplot`. With the package `ggplot2` and the command `geom_boxplot` more adjustments can be done.

Example 3.7. This example shows two possibilities to create a box plot of the variable *workload_max* of table *exampledata* in R.

```
#Variant 1:
boxplot(exampledata$workload_max)

#Variant 2:
#Selection of the data
ggplot(exampledata, aes(x=gender, y=workload_max)) +
  #Box plot with notches and variable width
  geom_boxplot(notch=TRUE, varwidth=TRUE)+
  #Addition of the number of observations above the median
  stat_summary(fun.data = function(workload_max)
    {return(c(y = median(workload_max)+8,
              label = length(workload_max)))}, geom = "text")+
  #Title and axis label
  ggtitle("Box plot")+
  labs(y="maximum workload [W]")
```



(a) Variant 1: Standard box plot in R



(b) Variant 2: Box plot with the package `ggplot2` and additional adjustments

Figure 3.1: Box plots for the variable maximum workload

Figure 3.1(a) shows the standard box plot in R. Without an additional package just limited adjustments can be done. With the help of package `ggplot2` many additional characteristics can be adjusted, as in Figure 3.1(b) shown. Here the box plots for the maximum workload are separated by gender by using

`aes(x=gender, y=workload_max)`. The width of the boxes are correlated to the number of observations with `varwidth=TRUE` and additionally the number of observations is given with the command `stat_summary`. There are additional notches, which represent 95%-confidence intervals around the median, done with `notch=TRUE`.

3.2.3 Bar plot

In a bar plot rectangular bars represent categorical variables. The length of these bars are proportional to the number of observations in the corresponding value of the categorical variable (see e.g. [Akkerboom, 2008, Chapter 7]).

Again like for the box plot, the creation of a bar plot can be done in R in different ways. The easiest way is to use the command `barplot`. With the package `ggplot2` and the command `geom_bar` more adjustments can be done.

Example 3.8. This example shows two possibilities to create a bar plot of the variable *gender* of table *exampledata* in R.

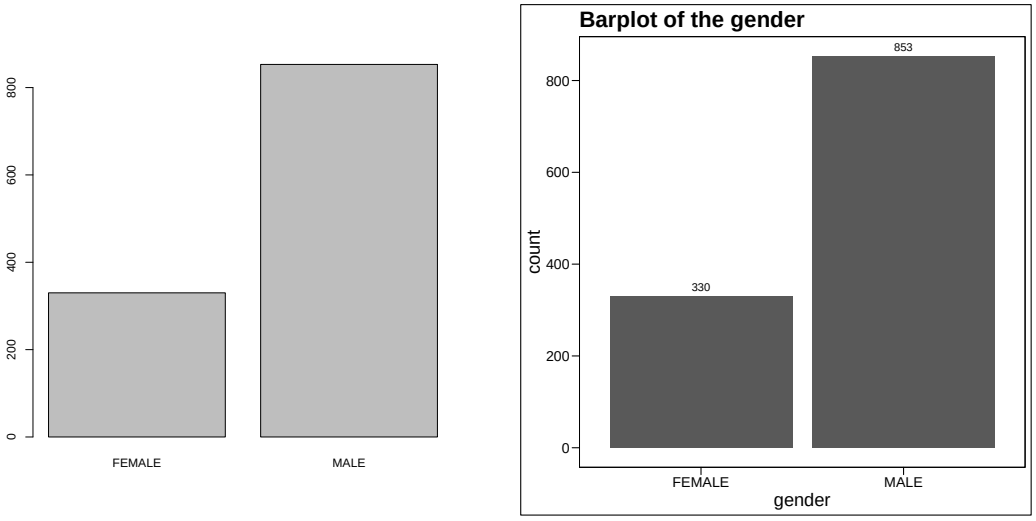
```
#Variant 1:
barplot(table(exampledata$gender))

#Variant 2:
#Selection of the data
ggplot(data=exampledata) +
  #Bar plot
  geom_bar(aes(factor(gender)),stat="count",
           position = position_dodge(width=0.9))+
  #Addition of the number of observations above the bars
  geom_text(aes(x=gender, label=..count..),stat="count",
            position = position_dodge(width = 0.8), vjust = -0.6)+
  #Title and axis label
  ggtitle('Barplot of the gender')+
  labs(x='gender')
```

Figure 3.2(a) shows the standard bar plot in R. The bar plot with the help of package `ggplot2` is shown in Figure 3.2(b). Here the number of observations is given with the command `geom_text`.

3.2.4 Histogram and density plot

In a histogram the n values $x_1, \ldots, x_n$ of a sample will be summarized in k uniform classes. For each class the frequency is determined and above each class interval a rectangle is drawn whose height is the corresponding frequency.

In order to create the histogram, the ordered data $x_{(1)} \leq \cdots \leq x_{(n)}$ are used to calculate the total area of the sample $[x_{(1)}, x_{(n)}]$. This area is divided into k uniform classes $[a_0 \leq x_{(1)}, a_1), [a_1, a_2), \ldots, [a_{(k-1)}, a_k \geq x_{(n)}]$, which all have the same width b and the height h_j with



(a) Variant 1: Standard bar plot in R

(b) Variant 2: Bar plot with the package `ggplot2` and additional adjustments

Figure 3.2: Number of ergometries seperated by gender

$$h(x) = \begin{cases} h_j, & \text{if } a_{j-1} \leq x < a_j, j = 1, \ldots, k; a_j - a_{j-1} = b \\ 0, & else. \end{cases}$$

The height h_j can be scaled through using the absolute frequency n_j, relative frequency $\frac{n_j}{n}$ or density $\frac{n_j}{nb}$. With defining the height of the rectangles as the density, classes with different widths can be used ([Cleff, 2014, Chapter 3]), [Devore and Berk, 2012, Chapter 1] and [Toutenburg et al., 2009, Chapter 2]).

Note, that the selection of suitable intervals is important, because different class widths have a big influence on the appearance of the histogram ([Cleff, 2014, Chapter 3]).

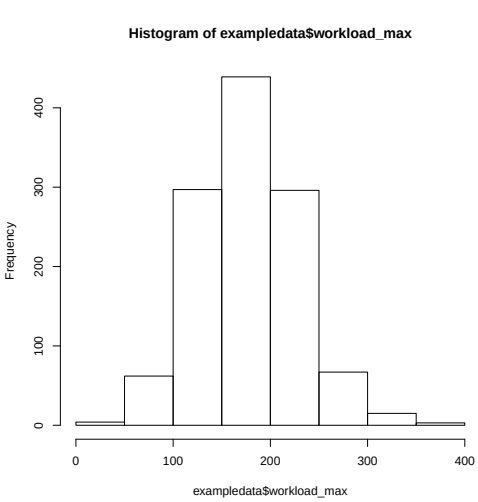
Additional information in a histogram can be obtained, if the mean is marked with a line.

Like before, the creation of a histogram can be done in R in different ways. The easiest way is to use the command `histo`. With the package `ggplot2` and the command `geom_histogram` more adjustments can be done.

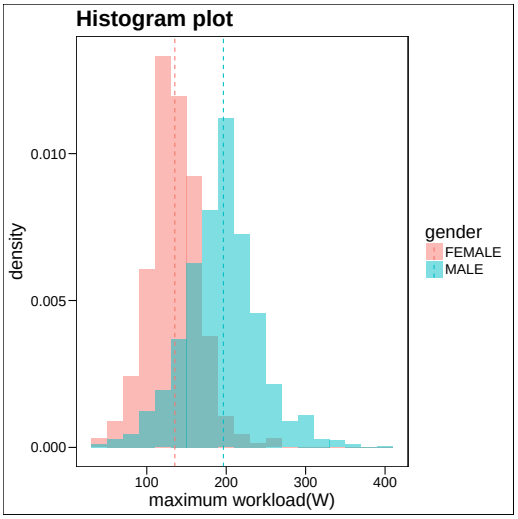
Example 3.9. This example shows two possibilities to create a histogram of the variable *workload_max* of table *exampledata* in R.

```
#Variant 1:
hist(exampledata$workload_max)

#Variant 2:
#Selection of the data
ggplot(exampledata, aes(x=workload_max, fill=gender)) +
  #Histogram
  geom_histogram(aes(y=..density..), position="identity",
                 alpha=0.5,binwidth=20)+
  #Additional line to show the mean
```

```
  geom_vline(data=ddply(exampledata, "gender", summarise,
                        grp.mean=mean(workload_max)),
             aes(xintercept=grp.mean, color=gender),
             linetype="dashed")+
  #Title and axis label
  ggtitle("Histogram plot")+
  labs(x="maximum workload(W)", y = "density")
```



(a) Variant 1: Standard histogram in R



(b) Variant 2: Histogram with the package ggplot2 and additional adjustments

Figure 3.3: Histogram for the variable maximum workload

Figure 3.3(a) shows the standard histogram in R. The histogram with the help of package ggplot2 is shown in Figure 3.3(b). Here the histograms are separated by the gender and the additional dashed lines mark the means.

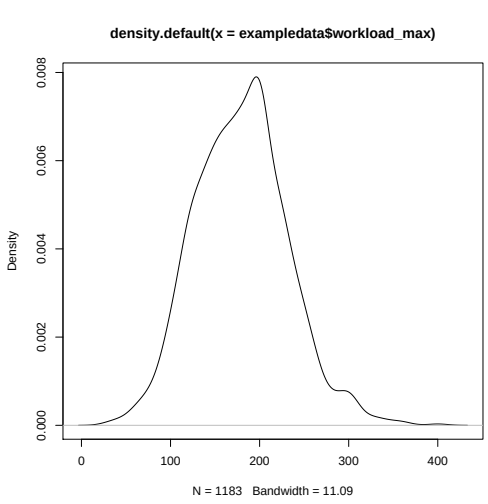
Histograms can be approximated with a smooth curve, the so-called density plot (see [Fahrmeir et al., 2016, Chapter 2]). The density plot visualizes the distribution of data over a continuous interval of a variable (see for example [Devore and Berk, 2012, Chapter 4]).

The creation of a density plot can be done in R in different ways. The easiest way is to use the command `density` and plot it with the command `plot`. With the package ggplot2 and the command `geom_density` more adjustments can be done.

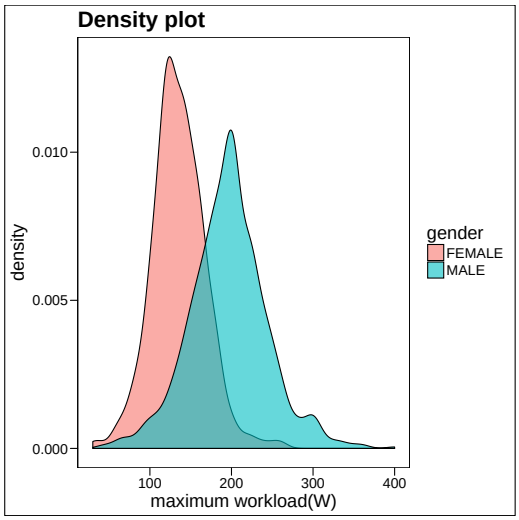
Example 3.10. This example shows two possibilities to create a density plot of the variable *workload_max* of table *exampledata* in R.

```
#Variant 1:
plot(density(exampledata$workload_max))
#Variant 2:
#Selection of the data
ggplot(exampledata, aes(x=workload_max, fill=gender))+
```

```
#Density plot
geom_density(alpha=0.6)+
#Title and axis label
ggtitle("Density plot")+
labs(x="maximum workload(W)", y = "density")
```



(a) Variant 1: Standard density plot in R



(b) Variant 2: Density plot with the package `ggplot2` and additional adjustments

Figure 3.4: Density plot for the variable maximum workload

Figure 3.4(a) shows the standard density plot in R. The density plot with the help of package `ggplot2` is shown in Figure 3.4(b), where the density plots are separated by the gender.

Note, histogram and density plot can be also combined in one plot, as shown in the Figures in Chapter 6.

3.2.5 Plot of the empirical distribution function

The empirical distribution function (see e.g. [Fahrmeir et al., 2016, Chapter 2]) of a random sample is defined as

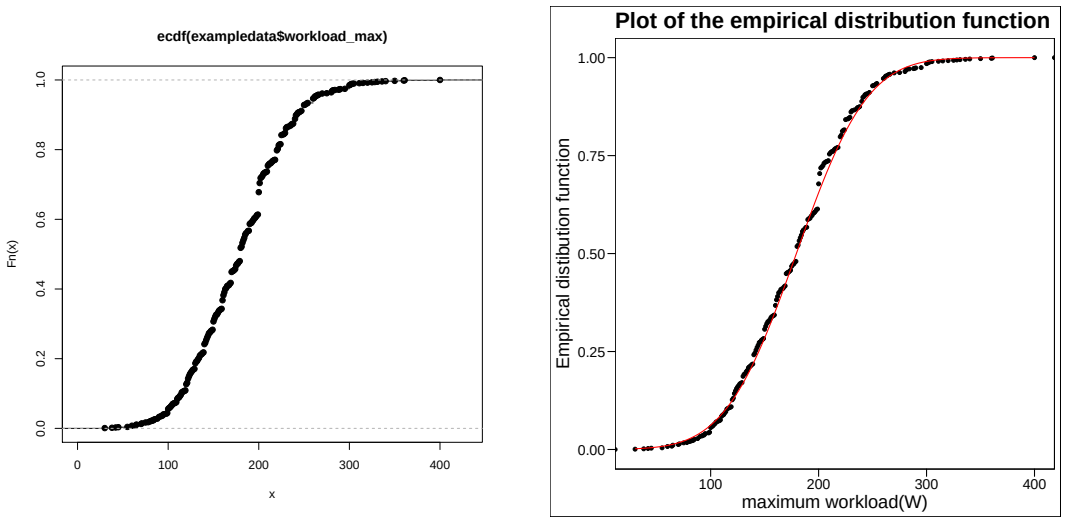
$$F_n(x) = \begin{cases} 0, & \text{if } x < x_{(i)} \\ \frac{i}{n}, & \text{if } x_{(i)} \leq x_{(i+1)}, i = 1, \ldots, n-1. \\ 1, & \text{if } x_{(n)} \leq x \end{cases}$$

In the corresponding plot the shifted step function $F_n^s(x) = F_n(x) - \frac{1}{2n}$ for $x_{(i)} \leq x < x_{(x+1)}$ is presented instead of plotting directly $F_n(x)$.

The easiest way to create the plot of the empirical distribution function in R is to use the command `ecdf` and plot it with the command `plot`. With the package `ggplot2` and the command `stat_ecdf` more adjustments can be done.

Example 3.11. This example shows two possibilities to plot the empirical distribution function of the variable *workload_max* of table *exampledata* in R.

```
#Variant 1
plot(ecdf(exampledata$workload_max))

#Variant 2:
#Selection of the data
ggplot(exampledata, aes(x=workload_max))+
  #Empicial distribution function
  stat_ecdf(geom = "point")+
  stat_function(fun = pnorm, n = 101, args =
                  list(mean = mean(exampledata$workload_max),
                       sd = sd(exampledata$workload_max)),
                colour='red')+
  #Title and axis label
  ggtitle("Plot of the empirical distribution function")+
  labs(x="maximum workload(W)", y="Empirical distibution function")
```



(a) Variant 1: Standard plot of the empirical distribution function in R

(b) Variant 2: Plot of the empirical distribution function with the package `ggplot2` and additional adjustments

Figure 3.5: Plot of the empirical distribution function for the variable maximum workload

Figure 3.5(a) shows the standard plot for the empirical distribution function in R. The plot for the empirical distribution function with the help of package `ggplot2` is shown in Figure 3.5(b). It contains also the distribution function of the corresponding normal distribution with mean $\bar{x}$ and standard deviation sd.

3.2.6 Quantile-quantile plot

In a Quantile-quantile plot (Q-Q plot), the distribution of two variables is compared by plotting their quantiles against each other (see [Kohn and Öztürk, 2013, Chapter 15]). Often this plot is used to compare the distribution of a sample with

the theoretical distribution, i.e. the ordered sample values $x_{(i)}$ against the theoretical quantile $F^{-1}(\frac{i}{n+1})$, with $i = 1, \dots n$. Therefore the deviation from the reference line, a line with a slope of 45°, can be evaluated.

The easiest way to create a Q-Q plot to compare the empirical quantiles with the quantiles of the standard normal distribution $N(0,1)$ is to use the command `qqnorm` to create the points and the command `qqline` to create the line. With the package `ggplot2` the command `stat_qq` can be used to create the dots and the command `geom_abline` can be used to create the line.

Example 3.12. This example shows two possibilities to make a Q-Q plot of the variable *workload_max* of table *exampledata* in R.

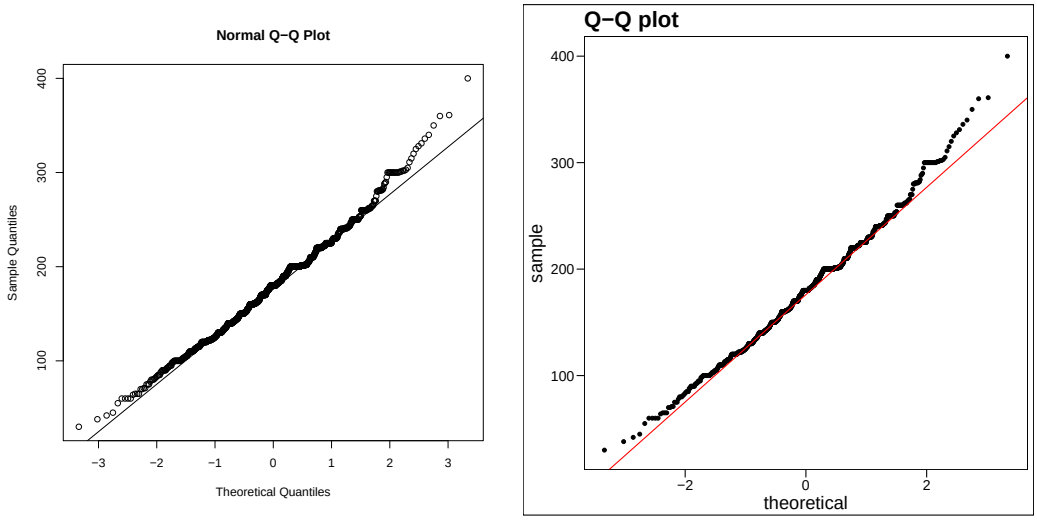
```
#Variant 1:
qqnorm(exampledata$workload_max)
qqline(exampledata$workload_max)

#Variant 2:
# Find the 1st and 3rd quartiles
y     <- quantile(exampledata$workload_max, c(0.25, 0.75), type=5)
# Find the matching normal values on the x-axis
x     <- qnorm( c(0.25, 0.75))
# Compute the line slope
slope <- diff(y) / diff(x)
# Compute the line intercept
int   <- y[1] - slope * x[1]
#Selection of the data
ggplot(exampledata, aes(sample=workload_max))+
  #QQ-plot
  stat_qq(geom = "point")+
  geom_abline(intercept=int, slope=slope,colour='red')+
  #Title
  ggtitle("Q-Q plot")
```

Figure 3.6(a) shows the standard Q-Q plot in R. The Q-Q plot with the help of package `ggplot2` is shown in Figure 3.6(b). Here the line has to be calculated separately. The plot indicates a remarkable deviation from the normal distribution.

3.2.7 Scatter plot

A scatter plot uses Cartesian coordinates to display values of two variables (see [Devore and Berk, 2012, Chapter 12]). Therefore it belongs to the multivariate analysis. The resulting point cloud gives information about the correlation between these variables (see e.g. [Pruscha, 2015, Chapter 1]).

The easiest way to create the scatter plot in R is to use the command `plot`. With the package `ggplot2` the command `geom_point` can be used to create the dots. To add an additional regression line the command `geom_smooth` can be used. For this function different methods like linear regression or generalized



(a) Variant 1: Standard Q-Q plot in R

(b) Variant 2: Q-Q plot with the package ggplot2 and additional adjustments

Figure 3.6: Q-Q plot for the variable maximum workload

additive model, which is the standard method for more than 1000 observations, can be applied.

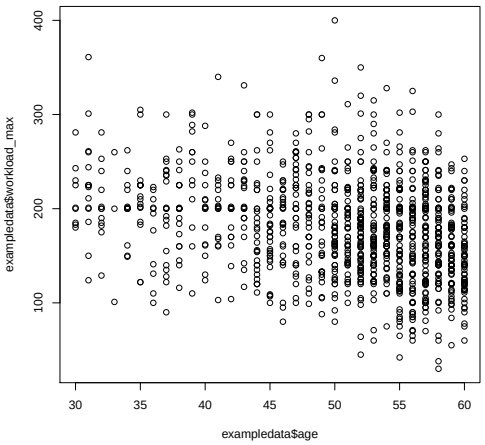
Example 3.13. This example shows two possibilities to make a scatter plot of the variables *age* and *workload_max* of table *exampledata* in R.

```
#Variant 1:
plot(exampledata$age,exampledata$workload_max)

#Variant 2:
#Selection of the data
ggplot(exampledata, aes(age,workload_max))+
  #Scatter plot
  geom_point()+
  #Regression line
  geom_smooth()+
  #Title
  ggtitle("Scatter plot")+
  labs(x="age", y="maximum workload(W)")
```

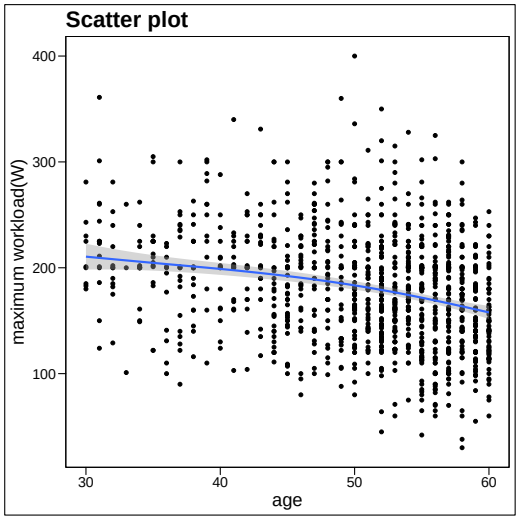
Figure 3.7(a) shows the standard scatter plot in R. The scatter plot created with the help of package ggplot2 is shown in Figure 3.7(b). Here additionally a regression line including a 95%-confidence interval around it is added.

3.2.8 Correlation coefficient

A measurement for the linear correlation between two samples x and y is the (Pearson) correlation coefficient (see e.g. [Devore and Berk, 2012, Chapter 5],



(a) Variant 1: Standard scatter plot in R



(b) Variant 2: Scatter plot with the package `ggplot2` and additional adjustments

Figure 3.7: Scatter plot for the variables age and maximum workload

[Cleff, 2014, Chapter 4]). The correlation coefficient is defined by

$$r_{x,y} = \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n-1}\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}.$$

s_{xy} is the co-variance between x and y, s_x is the standard deviation of x and s_y is the standard deviation of y.

For the Pearson correlation coefficient the following properties hold:

- r is normalized, i.e. $-1 \leq r \leq +1$.
- If $r > 0$ then there is a positive linear correlation, if $r < 0$ then there is a negative linear correlation and if $r = 0$ then there is no linear correlation.
- If x and y come from independent random variables X and Y, then $r \approx 0$, the inversion is not valid in general. Two variables can be uncorrelated, but dependent because of a nonlinear relationship.

In R the Pearson correlation coefficient can be calculated with the function `cor(x,y)`. Additionally it is possible to create correlation matrices for several variables. This can be done for example with the package `lattice` [Sarkar, 2008] and the command `corrgram`.

Example 3.14. This example shows how to calculate the correlation between two variables and how to create a correlation matrix between several variables in R.

```
cor(exampledata$workload_max,exampledata$age)

## [1] -0.2815862
```

The correlation between the maximum workload and age is -0.28, consequently they are negatively correlated.

```
corrgram(subset(exampledata, select = c("weight","height","age",
                                        "workload_max")),
         lower.panel = panel.ellipse, upper.panel = panel.pie)
```

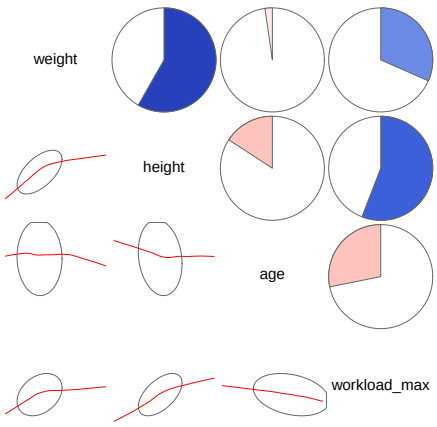


Figure 3.8: Correlation matrix between the variables weight, height, age and maximum workload

Figure 3.8 shows the correlation matrix with the help of package `lattice`. The variables are written on the diagonal line.

The lower triangle matrix shows both ellipses and loess lines[1]. Long, narrow ellipses represent high correlations while circular ellipses represent low correlations.

The value of the correlation can be deducted from the upper triangle matrix. For a positive correlation, blue color is used and the correlation is seen clockwise. A full clock represents a correlation of 1. On the contrary to this a clock with a negative correlation is colored red and counter clockwise.

3.3 Multiple linear regression

In general a regression analysis is a statistical method to model the relationship between two or more variables to gain information about one of them through knowing the values of the other(s). In the simple linear regression the depended or response variable is predicted by a single independent, explanatory or predictor variable. However, in most cases more than one predictor variable is useful. In this case we use multiple linear regression (see e.g. [Devore and Berk, 2012, Chapter 12], [Kleinbaum et al., 1998, Chapter 8], [Fahrmeir et al., 2008, Chapter 2], [Sheather, 2009, Chapter 5], [Gareth et al., 2014, Chapter 3]).

[1]These lines are constructed by local polynomial regression fitting. For the documentation of the algorithm see [Cleveland et al., 1992]

3.3.1 Multiple linear regression model

The standard model for multiple linear regression (see [Sheather, 2009, Chapter 5] or [Bingham and Fry, 2010, Chapter 3]) is the following:

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i, \quad i = 1, ..., n,$$

with

$Y_1, ..., Y_n \ldots$ observed metric variables

$x_{1j}, ..., x_{nj} \ldots$ deterministic values of the variables X_j

$\epsilon_1, ..., \epsilon_n \ldots$ unobserved random variables, which are independent and identically distributed with $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$ and $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$

$\beta_0, ..., \beta_p \ldots$ coefficients of the regression

The matrix notation of this is:

$$\begin{aligned} Y &= \beta_0 1 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon \\ &= X\beta + \epsilon, \end{aligned}$$

with

$Y = (Y_1, ..., Y_n)^T \ldots$ $n \times 1$-vector of dependent observations

$X \ldots$ $n \times k$-design matrix (with a full column rank)

$\epsilon \ldots$ $n \times 1$-vector of unobserved residuals

$\beta \ldots$ $k \times 1$-parameter vector

With given values of the regressors, the $Y_1, ..., Y_n$ are independent with the expected value and variance

$$\begin{aligned} E(Y_i) &= \mu_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}, \quad i = 1, ..., n \\ Var(Y_i) &= \sigma^2, \quad i = 1, ..., n. \end{aligned}$$

With the assumption of normal distribution for the error values

$$\epsilon_i \sim N(0, \sigma^2), \quad i = 1, .., n$$

follows the normal distribution for the target variable, therefore

$$Y_i \sim N(\mu_i, \sigma^2), \quad i = 1, ..., n.$$

Note, that the dependent variable is a linear combination of the parameters, but is not necessarily linear in the independent variables [Kleinbaum et al., 1998, Chapter 8]. For example the model with all possible first- and second order terms

$$Y = \beta_0 1 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon$$

can be rewritten as

$$Y = \beta_0 1 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon.$$

In R for the linear regression the command `lm` can be used. Note that in R for high order terms and other nonlinear relations the function `I(.)` and for interactions between variables the symbol `:` can be used.

Example 3.15. This example shows how to create a linear regression for the response variable *workload_max* and the predictor variables *age*, *weight* and *height* in R. Additionally the second order term age^2 and interaction between *weight* and *height* are used in the regression.

```
regression_model=lm(workload_max ~ age+weight+height+I(age^2)+
                      weight:height,data=exampledata)
```

3.3.2 Estimating the regression coefficients

The fitted regression model is given by

$$\hat{Y} = \hat{\beta}_0 1 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p.$$

The regression coefficients $\hat{\beta}_0, \ldots, \hat{\beta}_p$ can be determined with the least-squares method ([Kleinbaum et al., 1998, Chapter 8], [Devore and Berk, 2012, Chapter 12]). This method minimizes the sum of squares of the distances between the observed responses and those predicted by the fitted model, i.e.

$$SSE(\beta) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \to \min_{\beta_0,\ldots,\beta_p}.$$

This difference is called as residual sum of squares, error sum of squares or sum of squares about regression.

Another possibility to estimate the regression coefficients is the maximum likelihood estimation (see [Fahrmeir et al., 2008, Chapter 3]). The likelihood function, under the assumption of normal distributed residuals, is given by

$$L(y|\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp(-\frac{1}{2\sigma^2} SSE).$$

Therefore the Log-Likelihood function is

$$\log L(y|\beta, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} SSE.$$

One can see that maximizing the Log-Likelihood function leads to minimizing the SSE. This is because the constant and the factor in front of the SSE can be eliminated for the maximization and maximizing the negative SSE is equivalent to minimizing SSE. Therefore the least-squares method and the maximum likelihood estimation lead to the same result.

The maximum likelihood estimator of the variance σ^2 can be calculated by setting $\frac{\partial log L(y|\beta,\sigma^2)}{\partial \sigma}$ to zero and is

$$\hat{\sigma}^2 = \frac{SSE}{n}.$$

Because this is a biased estimator[2], the unbiased estimator

$$\tilde{\sigma}^2 = \frac{SSE}{n-p} = MSE$$

[2]The bias of an estimator is the difference between an estimator's expected value and the true value of the parameter.

is used.

For receiving the regression coefficients of a linear regression model in R just the name of the model has to be typed. With the command `summary` before the name of the model, more information about the coefficients can be achieved.

The summary consists of the following parts:

Call

The first part shows again the predictor variable and the response variables as seen before.

Residuals

The next part shows the residuals. These are the differences between the actual observed response values and the response values that the model predicted, i.e. $Y_i - \hat{Y}_i$. For these residuals, five summary measures are given, i.e. the minimum, the first quartile, the median, the third quartile and the maximum. The residuals should be symmetrical distributed around the mean value 0.

Coefficients

The next part of the summary gives information about the coefficients of the model. Therefore, first, the coefficient estimates are given. The first one is the intercept and the others refer to the predictor variables. Additionally the standard error of the coefficients are given. This measures the average amount that the coefficient estimate varies from the actual average value of the response variable. Therefore the standard error should be low compared to its coefficients. For the t-value the coefficients estimate is divided by its standard error, therefore it should be far away from zero as this would indicate to reject the null hypothesis, and there is a relationship between the coefficient and the dependent variable. With this t-value, the p-value (in the R output Pr(>|t|)) can be calculated. The p-value gives the probability that the null hypothesis is true. In our case the null hypothesis says that there is no relationship between the predictor and the response. Finally, the significance codes evaluate the p-values. For example, three stars refer to a highly significant p-value of less than 0.001.

Residual standard error

The residual standard error (RSE) is a measure of the quality of a linear regression fit. It is defined as the square root of the residual sum of squares SSE (see Subsection 3.3.3) and therefore should be small. The degree of freedom is the number of data points minus the number of parameters.

Multiple R-squared, Adjusted R-squared

The coefficient of determination R^2 as well as the adjusted coefficient of determination R^2_{adj} are measures of how well the model is fitting the actual data. More information about these coefficients can be found in Subsection 3.3.3.

F-Statistic

F-statistic is a good indicator of whether there is a relationship between the predictors and the response variables. The further the F-statistic is from 1 the better it is. However, how much larger the F-statistic needs to be, depends on both the number of data points and the number of predictors. Generally, when the number of data points is large, an F-statistic that is only a little bit larger than 1 is already sufficient to reject the null hypothesis, which says that there is no relationship between the predictor and the response variables. The reverse is true as if the number of data points is small, a large F-statistic is required to be able to ascertain that there may be a relationship between predictor and response variables. Additionally, the p-value for the test is given.

Example 3.16. In the last example the regression model was stored in the variable `regression_model`. To get the coefficients this variable has to be called:

```
regression_model

##
## Call:
## lm(formula = workload_max ~ age + weight + height + I(age^2) +
##     weight:height, data = exampledata)
##
## Coefficients:
##   (Intercept)            age         weight         height
##    -812.36511        3.66709        4.97133        5.40829
##      I(age^2)  weight:height
##      -0.05335       -0.02832
```

As a result we can see the previously defined formula of the regression model and the calculated coefficients. Therefore the regression model is

$$\begin{aligned} workload\_max = \quad & -812.37 + 3.67age + 4.97weight + 5.41height - 0.05age^2 \\ & -0.03weight : height. \end{aligned}$$

The summary of the previously model provides the following information:

```
summary(regression_model)

##
## Call:
## lm(formula = workload_max ~ age + weight + height + I(age^2) +
##     weight:height, data = exampledata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -144.019  -25.229   -0.348   23.652  185.211
##
## Coefficients:
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -812.36511  121.33092  -6.695 3.32e-11 ***
## age              3.66709    1.84674   1.986 0.047298 *
## weight           4.97133    1.39243   3.570 0.000371 ***
## height           5.40829    0.66814   8.095 1.42e-15 ***
## I(age^2)        -0.05335    0.01949  -2.736 0.006303 **
## weight:height   -0.02832    0.00792  -3.576 0.000363 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.29 on 1177 degrees of freedom
## Multiple R-squared:  0.3624,Adjusted R-squared:  0.3597
## F-statistic: 133.8 on 5 and 1177 DF,  p-value: < 2.2e-16
```

We see a considerable symmetrical distribution of the residuals. As mentioned above, the standard error should be small compared to the coefficients' estimate, which is the case in our example. The t-statistic values are relatively far away from zero and are considerably large relative to the standard error. The p-values are very close to zero and the significance codes show a high significance for all predictor variables. The residual standard error is 41.29 calculated with 1,177 degrees of freedom. We had 1,183 data points and 6 parameters. The coefficient of determination is 0.3624 and the adjusted coefficient of determination is 0.3597. The F-statistic is 133.8 which is relatively larger than 1, given the size of our data.

3.3.3 Analysis of variance and the coefficient of determination

The analysis of variance (ANOVA) provides an overall summary of a multiple regression analysis (see [Kleinbaum et al., 1998, Chapter 8]). The total sum of squares (SST) consists of the regression sum of squares (SSR) and the error sum of squares (SSE), i.e.

$$\begin{aligned} SST &= SSR + SSE \\ \sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2. \end{aligned}$$

The mean of squares are defined as the sum of squares divided by the degree of freedom:

$$\begin{aligned} MST &= \tfrac{SST}{n-1} \\ MSR &= \tfrac{MSR}{p-1} \\ MSE &= \tfrac{MSE}{n-p} \end{aligned}$$

The sum of squares as well as the mean of squares can be calculated in R with the command `anova`. It presents the degree of freedom (df), the sum of squares (Sum Sq), the mean of squares (Mean Sq), the F-value, the p-value (Pr(>F)) and the significance codes. It tests whether the model terms are significant.

Example 3.17. The ANOVA for the previous example in R is the following:

```
anova(regression_model)

## Analysis of Variance Table
##
## Response: workload_max
##                 Df  Sum Sq Mean Sq  F value    Pr(>F)
## age              1  249572  249572 146.3661 < 2.2e-16 ***
## weight           1  301878  301878 177.0419 < 2.2e-16 ***
## height           1  552022  552022 323.7438 < 2.2e-16 ***
## I(age^2)         1   15350   15350   9.0023 0.0027532 **
## weight:height    1   21803   21803  12.7870 0.0003632 ***
## Residuals     1177 2006926    1705
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this table the SSE is 46987470. The SSR is the sum of the first five sum of squares, i.e. age to weight:height. Finally, the SST is the sum of all of them.

The coefficient of determination, which is a quantitative measure of how well the fitted model predicts the dependent variable, is defined as

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}.$$

The coefficient of determination R^2 lies between 0 and 1. If it is equal to 1, then the model fits perfect (SSE=0) and if it is equal to 0, then there is no linear correlation (SSR=0).

Different models can just be compared under the following three assumptions (see [Fahrmeir et al., 2008, Chapter 3]):

1. In all models the same dependent variable y is used, i.e. a comparison of models with objective criterion y and $\log(y)$ is not possible.
2. In all models the number of regression coefficients is the same.
3. All models have a constant β_0.

Therefore, in general, different models cannot be compared with the coefficient of determination, because the coefficient of determination R^2 is dependent on the number of predictor variables. If the number of predictor variables gets height, then R^2 rises. It is therefore desirable to adjust the coefficient of determination. The adjusted coefficient of determination (see [Devore and Berk, 2012, Chapter 12]) is defined by

$$R^2_{adj} = 1 - \frac{MSE}{MST} = 1 - \frac{n-1}{n-p}\frac{SSE}{SST}.$$

To display the coefficient of determination and the adjusted coefficient of determination from the summary of a linear model, the commands `r.squared` and `adj.r.squared` can be used.

Example 3.18. The summary of the previous model provides the following information:

```
summary(regression_model)$r.squared

## [1] 0.3623848

summary(regression_model)$adj.r.squared

## [1] 0.3596762
```

3.3.4 Diagnostics of the model

As mentioned before, several assumptions had to be made to use the multiple linear regression. The diagnostics of the model can be separated into the following parts (see [Faraway, 2009, Chapter 4]):

1. The error values are independent and identically normally distributed $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \quad i = 1, .., n$
2. The structural part of the model $\hat{Y} = X\beta$ is correct.
3. Unusual observations, which do not fit the model and therefore might change the selection and fit of the model, should be identified.

Constant variance

To identify a constant variance, the residual vs. fitted plots can be used. This plot is expected to appear random.

Normality

For the normality check a Q-Q plot can be used. It will give a straight line if the errors are distributed normally.

Correlated errors

To identify correlated errors a Scale-Location plot can be used. This plot is expected to appear random with no discernable patterns.

Identification of unusual observations

To identify unusual observations the residual vs. leverage plot can be used. In this plot, the Cook's distance is marked to show which points have the greatest influence on the regression, i.e. are leverage points.

This diagnostics of the model can be done in R with the command `plot`.

Example 3.19. The diagnostics for the above example can be done in the following way:

```
plot(regression_model)
```

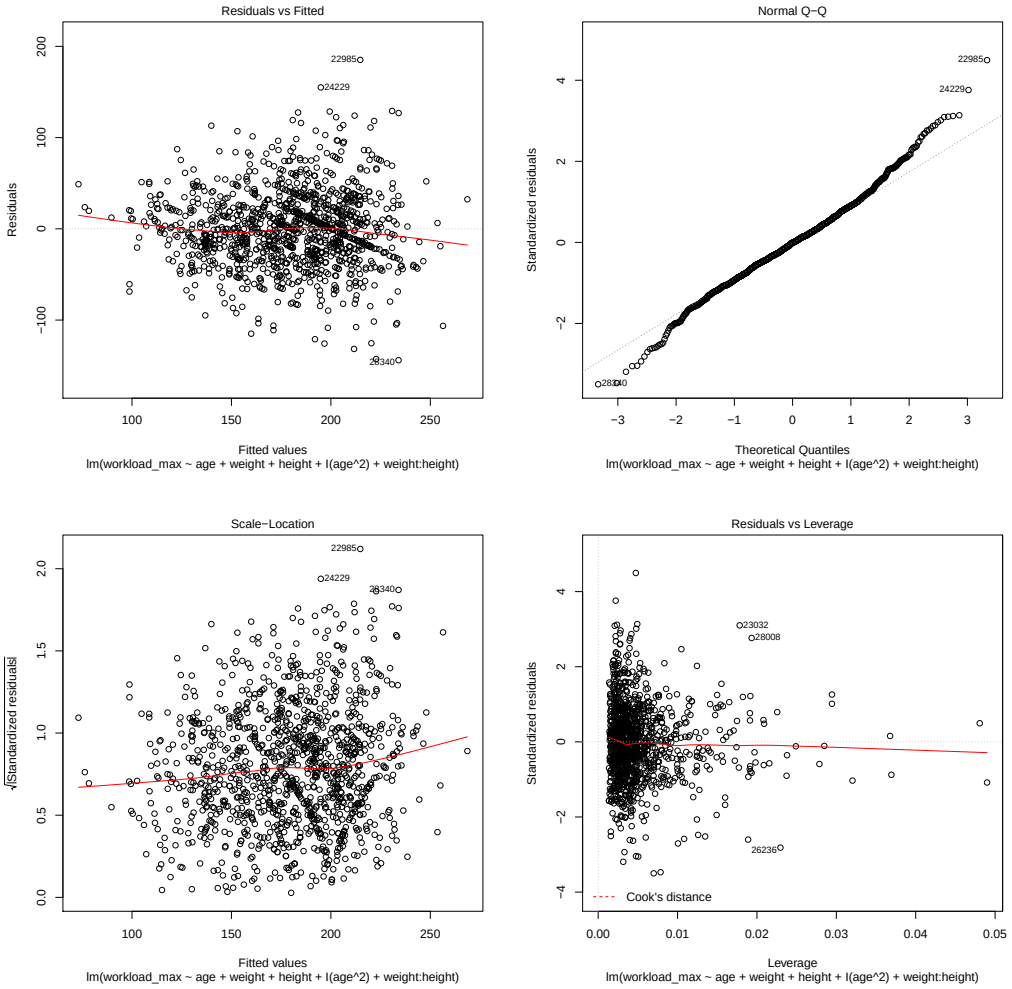


Figure 3.9: Diagnostic plot of the example linear regression model

In the first plot of Figure 3.9 the residuals are plotted against the predicted values. One can see that the variance is not constant. The second plot of Figure 3.9 shows a Q-Q plot of the standardized residuals. For small and big values there is a deviation from the line, i.e. there is a skinny positive and skinny negative tail. The third plot of Figure 3.9 should appear random with no patterns, as with the first plot. But this is again not the case, therefore the errors are correlated. Finally, the last plot of Figure 3.9 shows which points have the greatest influence on the regression (leverage points) which are observations 23032, 28008 and 26236.

3.3.5 Analysis of collinearity

In multiple linear regression it can happen that the predictor variables are correlated (see [Fahrmeir et al., 2008, Chapter 3]). The higher the linear correlation between the predictor variables, the higher the variance. In this case the estimations can be extremely inexact.

To remove strong collinearity, the relevant covariables can be omitted, from them a jointly (easy interpreted) variable can be built, a ridge-regression can be done or a main component regression can be used. More information about these methods can be found in [Fahrmeir et al., 2008, Chapter 3].

3.3.6 Box-Cox transformation

In Section 3.3.1 it has already been mentioned that predictor variables can be transformed for example by adding a polynomial term. Also the response variable can be transformed. A popular method to determine a transformation on the response variable is the Box-Cox transformation (see [Faraway, 2009, Chapter 7]). The goal of this method is to find a parameter λ for a variance stabilizing transformation, i.e. to make sure that the variance of the transformed response variable $y^*(\lambda)$ is independent of its mean. The transformation is given by

$$y^*(\lambda) = \begin{cases} \frac{y^\lambda}{\lambda} & \lambda \neq 0 \\ \log(y) & \lambda = 0. \end{cases}$$

An estimator of the parameter λ can be obtained by the profile log-likelihood function.

In R the command `boxcox` of the package `MASS`([Venables and Ripley, 2002]) calculates the estimator for λ and draws the profile of the likelihood function including a 95%-confidence interval for λ.

Example 3.20. For the above example the Box-Cox plot can be shown in the following way:

```
boxcox(regression_model)
```

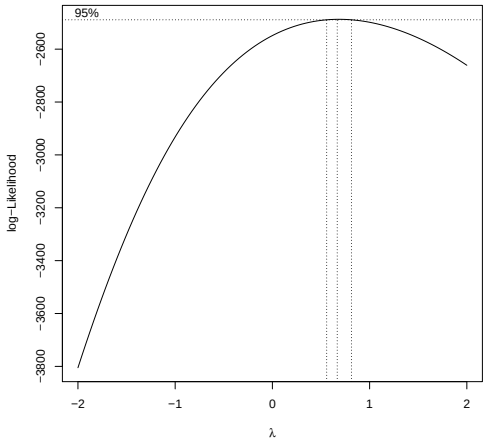


Figure 3.10: Box-Cox plot of the example linear regression model

One can see in Figure 3.10 that a transformation with $\lambda \approx 0.4$ is necessary.

3.3.7 Confidence intervals and prediction intervals

In the regression for a new given observation, the value of the response variables can be predicted. For this value, also confidence intervals and prediction intervals can be calculated (see [Faraway, 2009, Chapter 3]).

The confidence interval covers the expected value of the dependent variable $E(y_+)$ of a new observation x_+ with a probability of $(1-\alpha)$ and is defined as

$$\hat{y}_+ \pm t_{n-p,1-\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2 x_+^T (X^T X)^{-1} x_+}.$$

The prediction interval covers the actual new observation y_+ of a new observation x_+ with a probability of $(1-\alpha)$ and is defined as

$$\hat{y}_+ \pm t_{n-p,1-\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2 (1 + x_+^T (X^T X)^{-1} x_+)}.$$

Note, that by definition the prediction interval is always bigger than the confidence interval.

In R both intervals can be calculated with the command `predict`. For a confidence interval the additional option `interval="confidence"` has to be chosen and for a prediction interval the option `interval="prediction"` has to be added. In both cases the level can also be specified.

Example 3.21. For our example the intervals can be calculated for an observation with a height of 170 cm, a weight of 60 kg and an age of 50 years in the following way:

```
test <- data.frame(height=170,weight=60,age=50)
predict(regression_model, test, se.fit=TRUE, interval="confidence",
        level=0.90)$fit

##        fit      lwr      upr
## 1 166.4537 162.4279 170.4795

predict(regression_model, test, se.fit=TRUE, interval="prediction",
        level=0.90)$fit

##        fit      lwr      upr
## 1 166.4537 98.35998 234.5474
```

The new observation has y_+ 166.45 W. The confidence interval for this observation is $[162.43; 170.48]$, which means that the expected value of the new observation $E(y_+)$ is to 90% in this interval. The prediction interval is $[98.36; 234.55]$, which means that the actual new observation y_+ is to 90% in this interval.

3.4 Variable selection

3.4.1 Steps in selecting the best regression equation

For selecting the best regression model, the following steps have to be followed (see e.g. [Kleinbaum et al., 1998, Chapter 16]):

1. Specification of the maximum model
2. Specification of a criterion for selecting a model
3. Specification of a strategy for selecting variables
4. Conduction of the specified analysis
5. Evaluating the reliability of the chosen model

3.4.2 Specification of the maximum model

The maximum model should contain the following elements (see for more information [Kleinbaum et al., 1998, Chapter 16]):

1. All conceivable basic predictor variables
2. High-order powers of basic predictor variables
3. Other transformations of predictor variables
4. Interactions among predictor variables, including two-way and higher-order interactions
5. All possible 'control' variables, as well as their powers and interactions

Note that overfitting a model, i.e. including variables in the model with truly zero regression coefficients in the population, will not introduce bias on the result if usual regression assumptions are met, but it must be ensured that overfitting does not introduce harmful collinearity (see Subsection 3.3.5). Also underfitting the model, i.e. by leaving important predictor variables out, will introduce bias in the estimated regression coefficients.

3.4.3 Selection criteria

In general, larger models will fit better and have smaller error sum of squares SSE, but use more parameters. Thus for the choice of the model these two factors have to be balanced. Therefore, for evaluating potential subsets of predictor variables, selection criteria can be used. A selection criteria is an index, which can be calculated for each candidate model to compare the models. Thus, for a given selection criteria, the different models can be ordered from the best to the worst (see [Kleinbaum et al., 1998, Chapter 16]).

However, more than one selection criteria should be considered, because no single criterion works best. In practice, using a different selection criteria may lead to a different model choice (see [Kleinbaum et al., 1998, Chapter 16]).

The discussion of the following four criteria for evaluating subsets of predictor variables are taken from [Sheather, 2009, Chapter 7].

R^2 and adjusted R^2

The coefficient of determination, as already mentioned in Subsection 3.3.3, is defined as

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}.$$

Because adding a new variable to a model can only decrease the coefficient of determination R^2, always the biggest model would be chosen and the adjusted coefficient of determination is a better choice. The adjusted coefficient of determination, as already mentioned in Subsection 3.3.3, is defined by

$$R^2_{adj} = 1 - \frac{MSE}{MST} = 1 - \frac{n-1}{n-p}\frac{SSE}{SST},$$

where p is the number of predictor variables in the current model.

Using the adjusted R^2 as the selection criterion, the model with the highest R^2_{adj} should be selected.

Akaike's Information Criterion AIC

The AIC evaluates the

$$AIC = n \log\left(\frac{SSE}{n} + 2p\right).$$

The smaller the value of the AIC, the better the model.

Corrected Akaike's Information Criterion AIC_C

$$AIC_C = AIC + \frac{2(p+2)(p+3)}{n-p-1}.$$

The AIC_C is especially suitable for a small sample size compared to the number of predictor variables, i.e. $\frac{n}{p+1} \leq 40$.

Bayesian Information Criterion BIC

$$BIC = -2 \log L(\hat{\beta}, \hat{\sigma}^2, y) + (p+1)\log(n).$$

The smaller the value of the BIC, the better the model.

Mallow's C_p

Mallow's C_p (see [Kleinbaum et al., 1998, Chapter 16]) is defined as

$$C_p = \frac{SSE}{MSE} - (n - 2(p+1)).$$

A small value of C_p means that the model is relatively precise.

Mallows's C_p has been shown to be equivalent to the AIC in the special case of Gaussian linear regression (see [Boisbunon et al., 2014]).

Using the BIC results in simpler models, which are more robust. Therefore in Chapter 6 the BIC will be the chosen criterion.

3.4.4 Specification of a strategy for selecting variables

The third step in choosing the best model is to specify the strategy for selecting variables. [Sheather, 2009, Chapter 7] classifies these methods into two different approaches to choose the potential subsets of predictor variables, namely

1. all possible subsets, and
2. stepwise methods.

The stepwise methods include the forward and backward selection.

All subset selection

The all subset selection is to be preferred over all other variable selection strategies, if it is possible (see [Kleinbaum et al., 1998, Chapter 16]). This is the only method, where it is guaranteed that the model with the best value of the selected selection criterion is found.

In R all subsets can be calculated with the function `regsubsets`, which is in the package `leaps`. Useful options are `nbest` and `nvmax`. With `nbest=n` it can be specified, that n models of each size are kept in the result object. Without this option $n = 1$ is assumed. With `nvmax=n` the maximal size model, i.e. the number of variables to include are defined. Without this option $n = 8$ is assumed.

Example 3.22. The following example shows the summary of the two best regression subsets for one to five variables of the above model:

```
model_regsubsets<-regsubsets(workload_max ~ age+weight+height+
                               I(age^2)+weight:height,
                             data=exampledata,nbest=1,nvmax=5)
summary(model_regsubsets)$outmat

##          age weight height I(age^2) weight:height
## 1  ( 1 ) " " " "    "*"    " "       " "
## 2  ( 1 ) " " " "    "*"    "*"       " "
## 3  ( 1 ) "*" " "    "*"    "*"       " "
## 4  ( 1 ) " " "*"    "*"    "*"       "*"
## 5  ( 1 ) "*" "*"    "*"    "*"       "*"
```

To evaluate which model is the best, the selection criteria BIC (standard criterion), R^2, R^2_{adj} and C_p from Subsection 3.4.3 can be used. The first line, which has no option scale automatically uses BIC:

```
plot(model_regsubsets)
plot(model_regsubsets, scale="r2")
plot(model_regsubsets, scale="adjr2")
plot(model_regsubsets, scale="Cp")
```

Figure 3.11 shows the results for the different selection criteria.

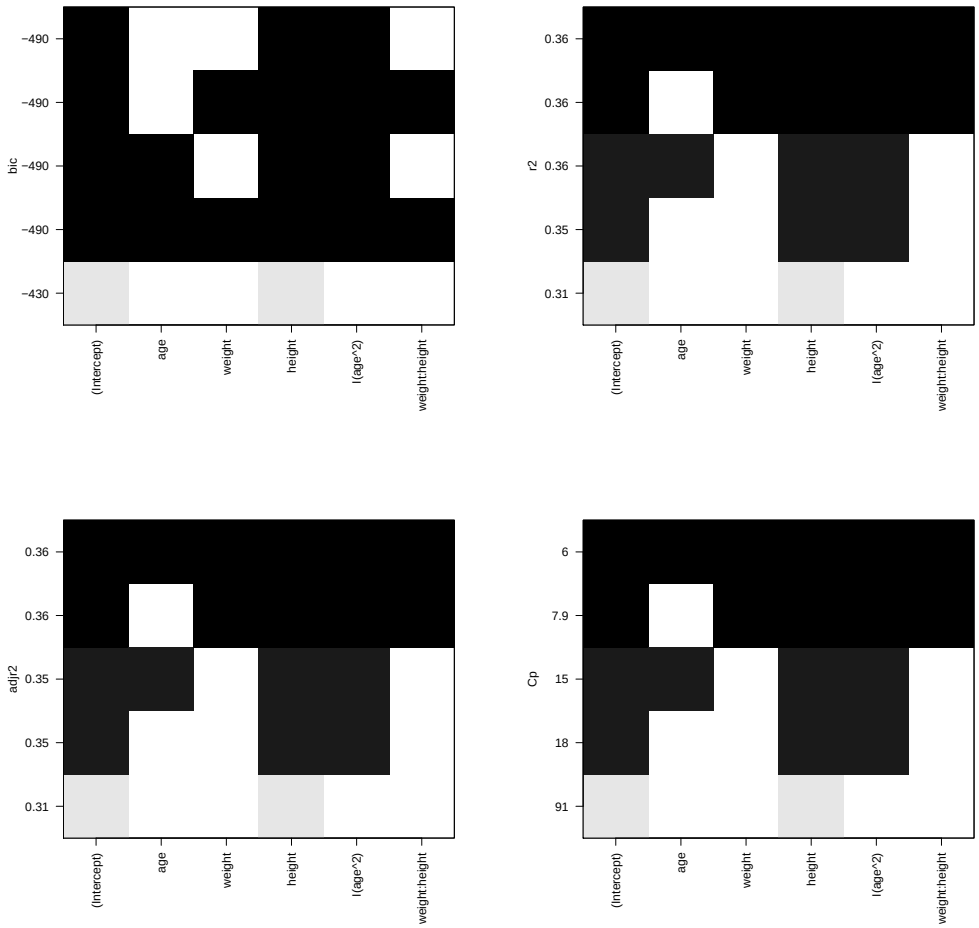


Figure 3.11: Results for different selection criteria for the example linear regression

```
plot(model_regsubsets$rss, xlab="Number of Variables", ylab="SSE",
     type="l")
plot(summary(model_regsubsets)$adjr2, xlab="Number of Variables",
     ylab=expression(R[adj]^2), type="l")
points(which.max(summary(model_regsubsets)$adjr2),
       summary(model_regsubsets)$adjr2[which.max(
         summary(model_regsubsets)$adjr2)]
       , col="red", cex =2, pch =20)
plot(summary(model_regsubsets)$cp, xlab="Number of Variables",
     ylab="CP", type="l")
points(which.min(summary(model_regsubsets)$cp),
       summary(model_regsubsets)$cp[which.min(
         summary(model_regsubsets)$cp)]
       , col="red", cex =2, pch =20)
plot(summary(model_regsubsets)$bic, xlab="Number of Variables",
     ylab="BIC", type="l")
points(which.min(summary(model_regsubsets)$bic),
```

```
        summary(model_regsubsets)$bic[which.min(
          summary(model_regsubsets)$bic)],
        col="red", cex =2, pch =20)
```

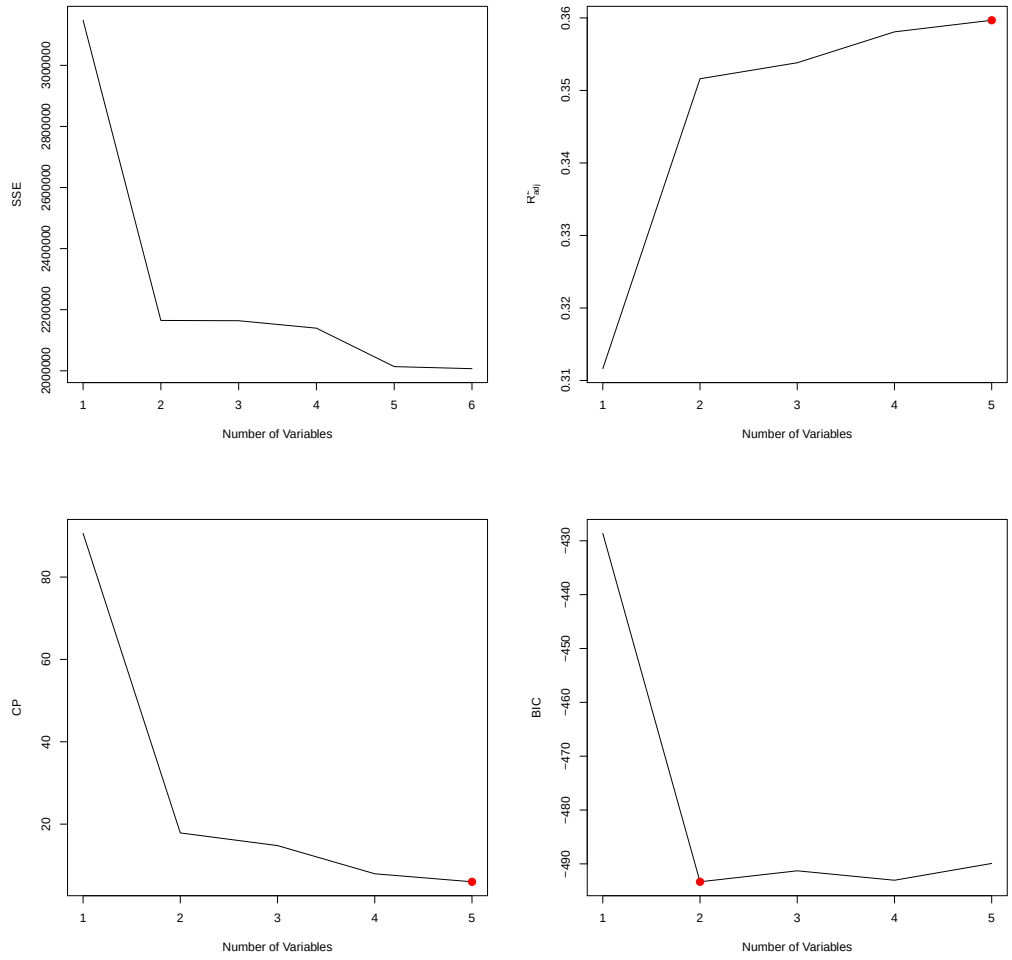


Figure 3.12: Information criteria for the example linear regression

The first plot in Figure 3.12 shows the SSE for the best models with one to five variables. The other plots show the values of the different selection criteria. As a result the model with five parameters is the best for all selection criteria.

Note, that also the function `leaps` exists in R, which is similar to the function `regsubsets`.

However, if the number of variables in the maximum model exceeds a certain size, it becomes impractical and instead it is preferable to use a stepwise method. The most popular stepwise methods are the forward selection and the backward selection.

Forward selection

The forward selection starts with a model, which just includes the intercept. In each iteration one predictor variable, taken from the set of the variables of

the maximum model, is included. The predictor variable is chosen such that the resulting model improves the selection criterion as much as possible. The algorithm ends, if all predictor variables are in the model or if no further improvement is possible.

Backward selection

The backward selection is the opposite of the forward selection, i.e. it starts with the maximum model with all predictor variables. In each step the predictor variable, which improves the selection criterion as much as possible, gets excluded. The algorithm ends, if all predictor variables got excluded or no further improvement is possible.

Stepwise selection

The stepwise selection combines the forward and the backward selection. In each iteration it will be decided, if a further variable comes into a model or if a variable gets excluded.

In R the variable selection can be done with the function `step`. The used selection criterion in this function is the AIC.

Example 3.23. The following is an example for the forward selection. For each step first the actual AIC and regression coefficients are given. For the coefficients, which can be included, the resulting sum of squares, the residual sum of squares RSS and the AIC are given:

```
null<-lm(workload_max ~ 1,data=exampledata)
full<-lm(workload_max ~ age+weight+height+I(age^2)+weight:height,
         data=exampledata)
modell_regression_step_fs<-step(null, scope=list(lower=null,
                                                 upper=full),
                                direction="forward")

## Start:  AIC=9331.52
## workload_max ~ 1
##
##             Df Sum of Sq     RSS    AIC
## + height     1    982714 2164836 8890.8
## + weight     1    314399 2833151 9209.0
## + I(age^2)   1    257059 2890491 9232.7
## + age        1    249572 2897978 9235.8
## <none>                   3147550 9331.5
##
## Step:  AIC=8890.75
## workload_max ~ height
##
##             Df Sum of Sq     RSS    AIC
## + I(age^2)   1    127448 2037388 8821.0
## + age        1    120669 2044167 8824.9
```

```
## <none>                  2164836 8890.8
## + weight    1       424 2164413 8892.5
##
## Step:  AIC=8820.97
## workload_max ~ height + I(age^2)
##
##          Df Sum of Sq     RSS    AIC
## + age     1    8657.5 2028730 8817.9
## <none>                 2037388 8821.0
## + weight  1      64.1 2037324 8822.9
##
## Step:  AIC=8817.93
## workload_max ~ height + I(age^2) + age
##
##          Df Sum of Sq     RSS    AIC
## <none>                 2028730 8817.9
## + weight  1    1.4482 2028729 8819.9
```

Afterwards a summary of the best model can be given by:

```
summary(modell_regression_step_fs)

##
## Call:
## lm(formula = workload_max ~ height + I(age^2) + age,
## data = exampledata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -139.584  -25.128   -0.323   23.625  188.828
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -420.55335   50.28850  -8.363   <2e-16 ***
## height         3.10155    0.13891  22.327   <2e-16 ***
## I(age^2)      -0.05838    0.01949  -2.995   0.0028 **
## age            4.13965    1.84554   2.243   0.0251 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.48 on 1179 degrees of freedom
## Multiple R-squared:  0.3555,Adjusted R-squared:  0.3538
## F-statistic: 216.7 on 3 and 1179 DF,  p-value: < 2.2e-16
```

The backward selection can be done in the following way:

```
modell_regression_step_bs<-step(full, direction="backward")
```

And finally the combination of both forward and backward selection can be done in the following way:

```
modell_regression_step_fbs<-step(null, scope = list(upper=full),
                                 data=exampledata, direction="both")
```

All directions give the same result.

3.5 Dummy regression

In Section 3.3 only continuous predictor variables were considered. To extend the predictor variables to categorical variables, dummy variables can be used. The dummy variable in a regression equation takes on a finite number of values, so that different categories of a nominal variable can be defined (see [Gareth et al., 2014, Chapter 3] and [Kleinbaum et al., 1998, Chapter 14]).

The term dummy is due to the fact that the values taken on by such variables do not indicate meaningful measurements but rather the categories of interest (see [Kleinbaum et al., 1998, Chapter 14]).

[Kleinbaum et al., 1998, Chapter 14] defines a rule for defining dummy variables for regression analysis to avoid collinearity: "If the nominal independent variable of interest has k categories, then exactly $k-1$ dummy variables must be defined to index these categories, provided that the regression model contains a constant term (i.e. an intercept β_0). If the regression model does not contain an intercept, then k dummy variables are needed to index the k categories of interest."

[Fahrmeir et al., 2008, Chapter 3] gives an explicit definition to model the categorical variable with c levels $x \in \{1, \ldots, c\}$ as $c-1$ dummy variables:

$$x = \begin{cases} 1 \\ 2 \\ \vdots \\ c \end{cases} \Rightarrow x_1 = \begin{cases} 1 & x = 1 \\ 0 & else \end{cases} \ldots x_{c-1} = \begin{cases} 1 & x = c-1 \\ 0 & else \end{cases}$$

The resulting dummy variables can be used in R in the same way as continuous variables.

3.6 Regression surfaces

Regression surfaces are used to examine the relationship between a response variable and the exploratory variables. The visualization of a regression surface can be done with the package `visreg` [Breheny and Burchett, 2017].

In this thesis, two functions from this package are used. The first is `visreg` with which the regression model can be visualized. The default plots contain

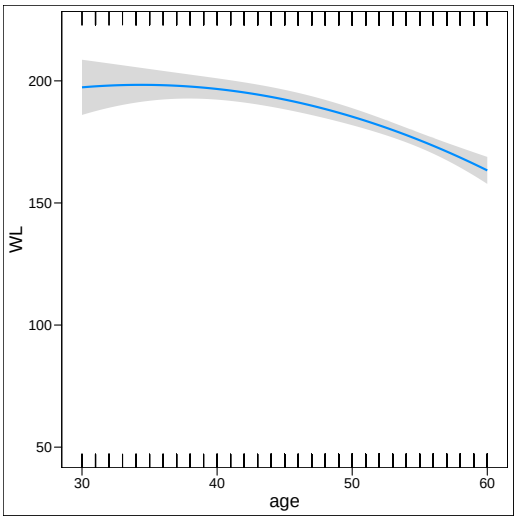
a confidence band, prediction line, and partial residuals. Also, options for factors, transformations, conditioning and interactions are supported. The second function is `visreg2d` and with this function it can be shown how two variables interact to affect the response in a regression model.

Example 3.24. This example shows regression surfaces for the above regression with the formula

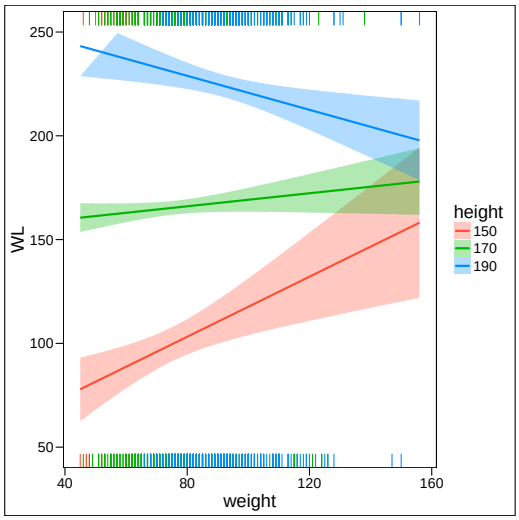
$$WL = age + weight + height + I(age^2) + weight : height.$$

The function `visreg` can be used in the following way:

```
visreg(lm(workload_max ~ age+weight+height+I(age^2)+weight:height,
          data=exampledata),
       'age',band=TRUE,gg=TRUE,overlay=TRUE, partial=FALSE,alpha=0.05,
       breaks=c(25,50,75),ylab='WL',print.cond=FALSE)+ylim(50, 220)
visreg(lm(workload_max ~ age+weight+height+I(age^2)+weight:height,
          data=exampledata),
       'weight',by='height',plot.type="persp",band=TRUE,gg=TRUE,
       overlay=TRUE,
       partial=FALSE,alpha=0.05,breaks=c(150,170,190),ylab='WL')+
  ylim(50,250)
```



(a) Regression line for the maximum workload denpendent on age



(b) Regression lines for the maximum workload depending on weight for three different breaks of height

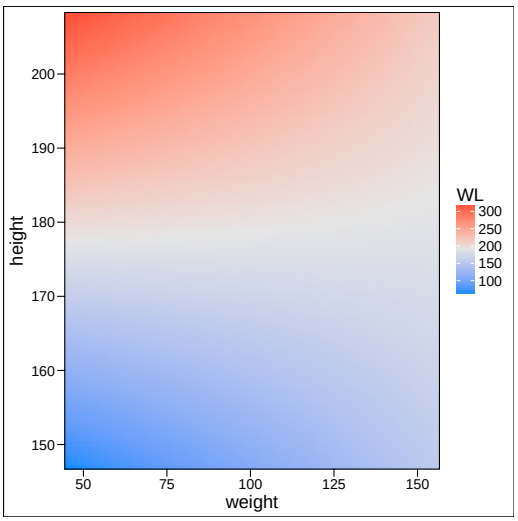
Figure 3.13: Plots of the regression line and confidence bands of the regression surface

Figure 3.14 shows these plots. Figure 3.14(a) shows the maximum workload depending on the age. To plot this, the other parameters have to be fixed. $I(age^2)$ is calculated by the age. The other parameters are set to the median values., i.e. weight is set to 78 and height is set to 173. The function `visreg` can also be used to see how the effect of weight differs depending on the height with the

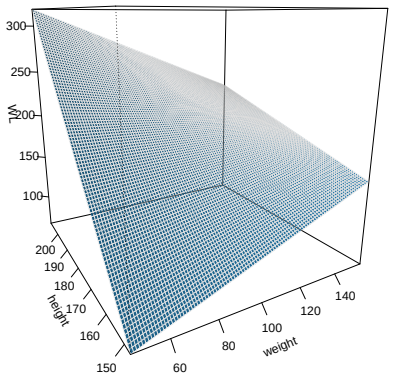
option `by`. Figure 3.14(b) shows this. Here the height is fixed at three different stages. The slope of the regression lines of the surface is different for each of these heights. In both plots additionally the 95%-confidence bands are plotted.

The function `visreg2d` can be used in the following way:

```
visreg2d(lm(workload_max ~ age+weight+height+I(age^2)+weight:height,
            data=exampledata),
         'weight','height',plot.type='gg',zlab='WL')
visreg2d(lm(workload_max ~ age+weight+height+I(age^2)+weight:height,
            data=exampledata),
         'weight','height',plot.type='persp',zlab='WL')
```



(a) Contour plot



(b) Perspective plot

Figure 3.14: Plots of the regression surface of the maximum workload depending on weight and height

The resulting plots are in Figure 3.14. The first plot, Figure 3.14(a), shows a contour plot (option `plot.type='gg'`) and Figure 3.14(b) shows a perspective plot (option `plot.type='persp'`). In both plots the interaction between height and weight can be observed.

3.7 Decision trees

Decision trees are a way to visualize regression models (e.g. [Falk et al., 2014, Chapter 9] or [Bankhofer and Vogel, 2008, Chapter 18]). There exist two variants of decision trees: classification trees and regression trees. The classification trees are used for nominal dependent variables and decision trees for quantitative dependent variables.

Decision trees are used for an automatic classification of data objects in order to solve decision problems. A decision tree consists of a root vertex, inner vertexes and leaves. Every vertex except the leaves represent one logical decision. The leaves represent an answer of the decision tree. This answer could be in the

form of the mean or box plots. Therefore, in the tree, the target variable for each observation can be predicted by following the path from the root node to the corresponding leaf.

To create a decision tree in R, the command `tree` from the package `tree` can be used.

Example 3.25. This example shows a regression tree for the maximum workload. As stated before, the maximum workload depends on age, height and weight. The right branch refers in this tree to `yes` and the left one to `no`.

```
plot(tree(workload_max ~ age+height+weight,data=exampledata))
text(tree(workload_max ~ age+height+weight,data=exampledata))
```

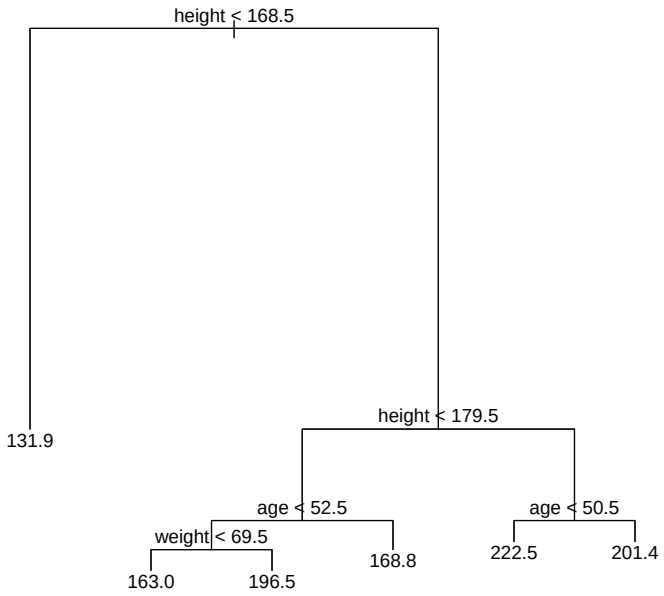


Figure 3.15: Regression tree for the example linear regression

We can see that the regression tree only contains the variables height and age. This corresponds to the result that the correlation between the maximum workload and the height and age is higher than with the weight.

If a new observation is 40 years old and has a height of 170 cm, then expected maximum workload is 165.1 W.

For the following variant of a regression tree the packages `rpart` and `party` are needed.

Example 3.26. This example shows another variant to get a regression tree. First the tree can be created with the command `rpart`.

```
tree<-rpart(workload_max ~ age+height+weight,data=exampledata)
tree

## n= 1183
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 1183 3147550.00 179.3922
##    2) height< 168.5 281  325207.10 131.8826 *
##    3) height>=168.5 902 1990486.00 194.1929
##      6) height< 179.5 497  886373.40 181.0966
##       12) age>=52.5 221  380658.10 168.7783 *
##       13) age< 52.5 276  445328.60 190.9601
##         26) weight< 69.5 46   66799.91 163.0435 *
##         27) weight>=69.5 230  335509.10 196.5435 *
##      7) height>=179.5 405  914264.70 210.2642
##       14) age>=50.5 235  437771.70 201.4340 *
##       15) age< 50.5 170  432840.40 222.4706 *
```

The `cptable` gives information about the complexity parameter (CP), which is used to control the size of the regression tree.

```
tree$cptable

##           CP nsplit rel error    xerror       xstd
## 1 0.26428697      0 1.0000000 1.0012694 0.04480537
## 2 0.06031623      1 0.7357130 0.7469441 0.03803352
## 3 0.01918529      2 0.6753968 0.6917521 0.03440656
## 4 0.01386877      3 0.6562115 0.6846830 0.03403352
## 5 0.01366764      4 0.6423427 0.6799076 0.03439675
## 6 0.01000000      5 0.6286751 0.6708352 0.03421516
```

To adjust the size of the tree the command `prune` can be used.

```
tree_size<-prune(tree, cp = tree$cptable[3,"CP"])
tree_size

## n= 1183
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
## 1) root 1183 3147550.0 179.3922
##   2) height< 168.5 281  325207.1 131.8826 *
##   3) height>=168.5 902 1990486.0 194.1929
##     6) height< 179.5 497  886373.4 181.0966 *
##     7) height>=179.5 405  914264.7 210.2642 *
```

Finally the regression tree can be plotted with the command `plot`.

```
plot(as.party(tree_size))
```

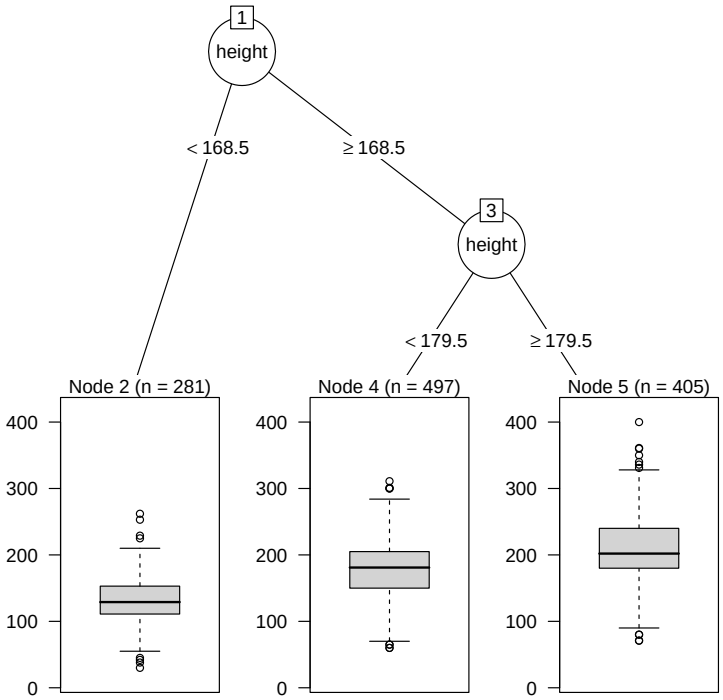


Figure 3.16: Regression tree for the example linear regression with the package `party`

Chapter 4

Data preparation

In this chapter, steps involving the preparation of raw data (received from ZARG) for statistical evaluation are explained in detail. The chapter starts with a description of the received raw data. Afterwards each step involved in the data preparation is described.

The data preparation required 5 steps:

Step 1: Transformation of the XML files in a CSV file

Step 2: Integration of the CSV files in a MySQL database

Step 3: Integration of the files with information on blood parameters, medication and diagnosis in the main table with a Java program

Step 4: Reading the MySQL database into R

Step 5: Adaptation of the table in R

Figure 4.1 shows these steps.

4.1 Received data

The received data consists of separate files with information on ergometries, blood parameters, diagnoses and medication prescribed.

4.1.1 XML files for data of the ergometries

The data from the ergometries have been provided in XML files. For each ergometry, there was one XML file. These XML files consist of variables such as age, gender, height, weight, maximum workload. In addition, files had information about the time series of the ergometry, start time of the step, the workload, systolic blood pressure, diastolic blood pressure, heart rate and cadence.

4.1.2 CSV files for the blood parameters, diagnosis and medication

The information on blood parameters, medication and diagnoses of the participants has been provided in CSV files.

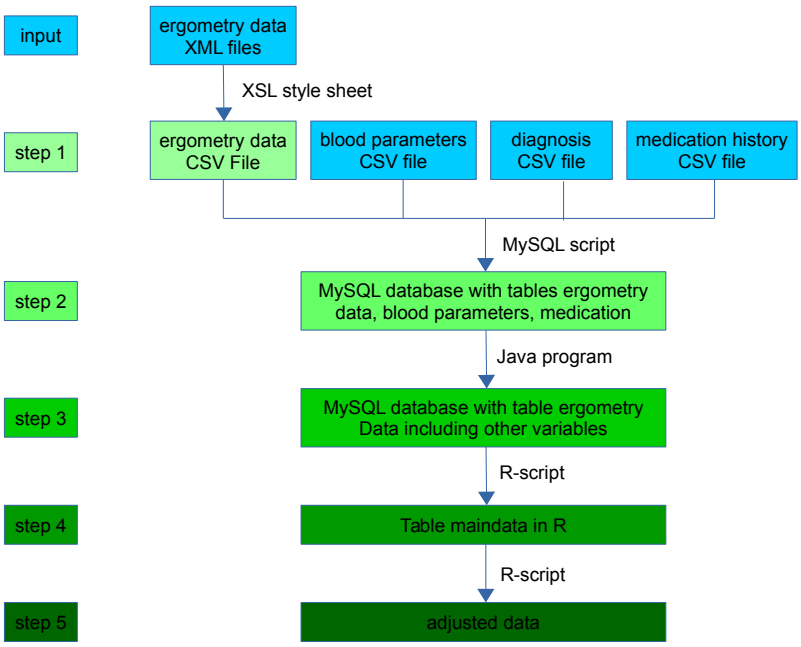


Figure 4.1: Workflow for the data preparation

CSV file: Blood parameters

This CSV file contains information about total cholesterol (TC), low-density lipoprotein (LDL), high-density lipoprotein (HDL), triglyceride (TG), creatinine (Cr), fasting blood sugar level (FBS) and uric acid (UA) levels in blood. The information on the variables present in this file are shown in Table 4.1. The first column contains the value of that parameter if it was measured within the range of two weeks before or after the relevant ergometry, and the second column states whether this measured value is pathological (T) or not (F).

Variables	Units	Description
patient_id	-	Patient ID
entry_id	-	identification number for the entry
lab_date	date	date when the laboratory was taken
parameter	-	parameter of the lab
lab_value	mg/dl	corresponding value of the parameter
evaluation	-	T for pathological and F for not pathological

Table 4.1: Description of the variables in CSV file blood parameters

CSV file: Diagnosis

The information about the diagnosis of participants who performed at least one ergometry is present in this CSV file. The information on the variables present

in this file are shown in Table 4.2.

Variables	Units	Description
patient_id	-	Patient ID
entry_id	-	identification number for the entry
date	date	date of the diagnosis
diagnosis	-	type of the diagnosis

Table 4.2: Description of the variables in Table diagnosis

CSV file: Medication

This file contains information about the types and doses of medication prescribed to participants. The information on the variables present in this file are shown in Table 4.3.

Variables	Units	Description
patient_id	-	Patient ID
entry_id	-	identification number for the entry
date	date	date on which the medication was prescribed
medication	-	type of the drug
doses	-	doses of the medication
medication_id	-	identification number of the medicine

Table 4.3: Description of the variables in Table medication

4.2 Step 1: Transformation of the XML files in a CSV file

As first step, the data from the XML files have been transformed into a CSV file using XSL stylesheet. The stylesheet is attached in Appendix A. The variables of the resulting CSV file are described in Table 4.4. This table contains the general information about the participants and some parameters of the corresponding ergometry. The general information includes patient ID, age, gender, height, weight and abdominal measurement. The ergometry information includes start time, end time, device, pace maker, resting heart rate, maximal heart rate, maximal power, resting systolic blood pressure, resting diastolic blood pressure, maximal systolic blood pressure, maximal diastolic blood pressure and the reason for termination.

In total, there were 29,805 records each containing information about one ergometry. After quality control, some of the records were deleted due to lack of correct information. Finally, 28,625 records were used for downstream analysis which passed the quality control step. These records were collected between 2004-01-27 and 2017-06-13 (14 years).

Variables	Units	Description
patient_id	-	Patient ID
start_time	date and time	date and time of the ergometry's start
end_time	date and time	date and time of the ergometry's end
age	years	patient's age
gender	-	patient's gender
height	*cm*	patient's height
weight	*kg*	patient's weight
waist	*cm*	patient's waist circumference
pacemaker	-	binary variable for existence of pacemaker
device	-	type of device, ergometer / treadmill
WL	*W*	maximum workload the patient could achieve
SBP_r	*mmHg*	resting systolic blood pressure
DBP_r	*mmHg*	resting diastolic blood pressure
SBP_{max}	*mmHg*	systolic blood pressure at maximum workload
DBP_{max}	*mmHg*	diastolic blood pressure at maximum workload
HR_r	*bpm*	resting heart rate
HR_{max}	*bpm*	maximal heart rate during the ergometry
termination	-	reason for termination of the ergometry
purpose	-	purpose for the ergometry

Table 4.4: Description of the variables in Table maindata

4.3 Step 2: Integration of the CSV files in MySQL database

Four CSV files `maindata`, `blood parameters`, `diagnosis` and `medication` are integrated in a MySQL database. The MySQL program for this is attached in Appendix B.

4.4 Step 3: Integration of the blood parameters, diagnosis and medication in the table maindata with Java programs

For the sake of convenience in table maindata additional columns for blood parameters, diagnosis and medication were generated with the help of Java programs. These programs are attached in Appendix C to Appendix E.

4.4.1 Columns for the blood parameters

For each ergometry, the information on each blood parameter is inserted in two new columns. These columns refer to the value of the corresponding blood parameter and its evaluation as mentioned in Table 4.1.

4.4.2 Column for diagnosis

The column for diagnosis contains information in the form of binary numbers. The participant with a diagnosis of cardiovascular diseases is assigned 1, otherwise 0.

4.4.3 Columns for medication

The information on beta blockers and statins is inserted in two separate columns. Here, also the binary numbers are used to indicate if the participant is taking medication or not. The participant taking beta blockers, from the list mentioned in Section 2.5, is assigned 1, otherwise 0. The convention is the same for statins too.

4.5 Step 4: Reading the table maindata into R

The finished table maindata is exported into R. The corresponding R script is attached in Appendix F.

4.6 Step 5: Adaptation of the table maindata in R

In the last step, additional variables calculated from existing variables are added. For this calculation, the following formulas are used:

Formula for the body mass index:

$$BMI[kg/m^2] = \frac{weight[kg]}{(\frac{height[cm]}{100})^2}$$

Formula for the waist-to-height-ratio:

$$WHR = \frac{waist[cm]}{height[cm]}$$

Formula for the workload-to-weight-ratio:

$$WWR[W/kg] = \frac{WL[W]}{weight[kg]}$$

$$TC : HDL = \frac{TC}{HDL}$$
$$LDL : HDL = \frac{LDL}{HDL}$$
$$TG : HDL = \frac{TG}{HDL}$$

The body surface area is calculated from weight and height using the Du Bois formula (see Chapter 2) and the maximum workload using the model from

[Wonisch et al., 2008]. Furthermore, the column showing the year in which erogometry was performed has been added.

Finally, two tables with information which correspond to stage II and stage III of cardiac out-patient rehabilitation are created. In these tables, above mentioned information available from these two stages for a single participant is presented in one row.

The R script for this is added in Appendix F.

4.7 Variables used in the statistical analysis

Table 4.5 shows a summary of all variables which will be used in the next chapters.

	Variables	Units	Description
Nominal	gender	-	patient's gender[1]
	medication	-	medication history[2]
	diagnosis	-	diagnosis[3]
Basic information	age	years	age
	height	cm	height
	weight	kg	weight
	waist	cm	waist circumference
	BMI	kg/m^2	body mass index
	WHR	-	waist to height ratio
	BSA	m^2	body surface area
Cardiac param.	SBP_r	$mmHg$	systolic blood pressure at the beginning of the ergometry
	DBP_r	$mmHg$	diastolic blood pressure at the beginning of the ergometry
	SBP_{max}	$mmHg$	maximal systolic blood pressure during the ergometry
	DBP_{max}	$mmHg$	maximal diastolic blood pressure during the ergometry
	HR_r	bpm	heart rate at the beginning of the ergometry
	HR_{max}	bpm	maximal heart rate during the ergometry
Blood param.	TC	mg/dl	total cholesterol
	LDL	mg/dl	low-density lipoprotein
	HDL	mg/dl	high-density lipoprotein
	TG	mg/dl	triglyceride
	Cr	mg/dl	creatinine
	FBS	mg/dl	fasting blood sugar level
	UA	mg/dl	uric acid
	TC:HDL	-	ratio between the TC and the HDL
	LDL:HDL	-	ratio between the LDL and the HDL
	TG:HDL	-	ratio between the TG and the HDL
Output	WL	W	maximum workload the patient could achieve
	WWR	W/kg	maximum workload to weight ratio

Table 4.5: Description of the variables used in the statistical analysis

[1] M... male, F... female
[2] 1... statin, 2... beta blocker, 3... both, 0... none
[3] 1... cardiac diagnosis, 0... non-cardiac or no diagnosis

Chapter 5

Exploratory data analysis

In this chapter, an exploratory data analysis (EDA) is used to analyze the data for patterns and relationships. A mathematical description of EDA and its methods are given in Section 3.2. In Section 5.1, a univariate analysis is done, where each variable is analyzed separately by a calculation of statistical indicators and presented as an illustration. In Section 5.2, a multivariate analysis, correlation matrices are used to provide an insight into the linear relationship between paired variables, also scatter plots are used to identify patterns between multiple variables. In Section 5.3, a gender specific comparison between the maximum workload obtained in ergometries from the first (2004 & 2005) and last two years (2016 & 2017) of study is done for the quality control of the data. Finally a short conclusion of this chapter is given in Section 5.4.

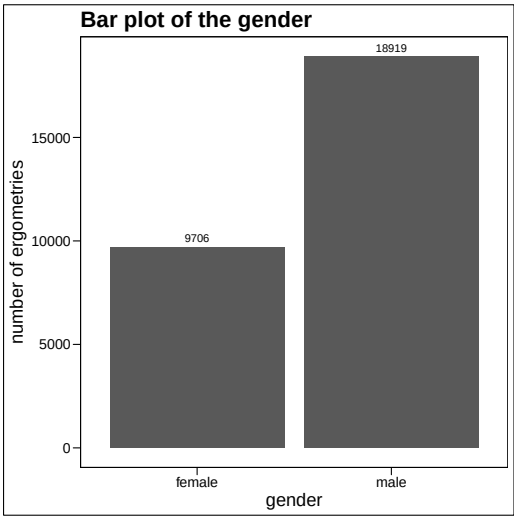
5.1 Univariate analysis

The EDA for each variable is performed to summarize their main characteristics and to find structures. The aim of this analysis is to verify the data and to find and eliminate errors in it.

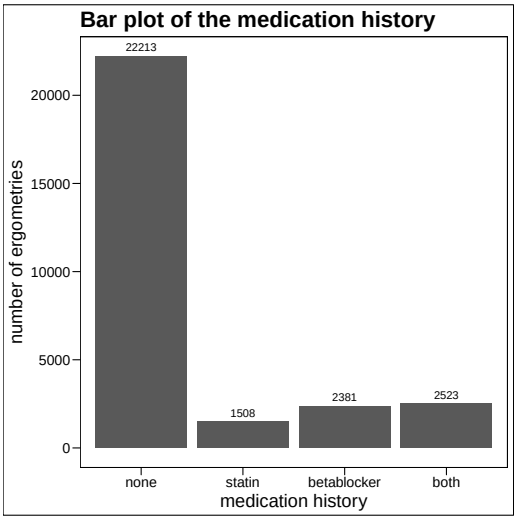
5.1.1 Composition of the data regarding gender, medication and diagnosis

Figure 5.1 shows the categorization of the ergometries by three nominal variables namely gender, medication and diagnosis. Figure 5.1(a) shows that 18,919 ergometries are done by males and 9,706 by females. Males performed twice as many ergometries as females. For the categorization using medications only two medication groups are taken into account as mentioned in Chapter 4. Figure 5.1(b) shows the number of ergometries in patients with no medication, with either one of statin or beta blocker and with both. In total, 4,031 ergometries are done by patients taking statin and 4,904 taking beta blocker. In most of the cases, the participants were not taking statins or beta blocker. Figure 5.1(c) shows the categorization of ergometries in two groups. The first group of 17,544 ergometries is a group with the diagnosis of non-cardiac diseases or no diagnosis and the second group with 11,081 ergometries is with the diagnosis of cardiac diseases. So the majority of data for the ergometries is from the participants without

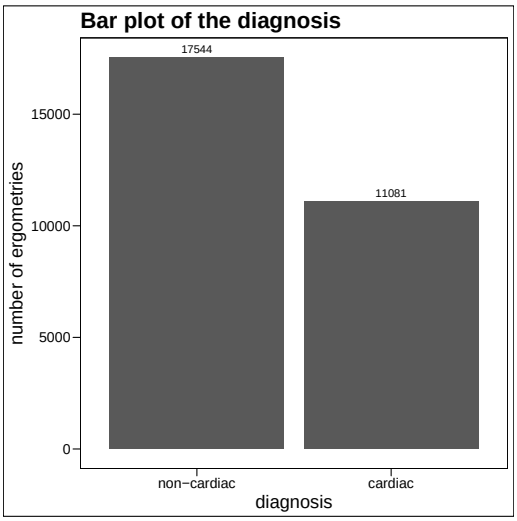
any cardiac disease diagnosis.



(a) Number of ergometries performed by females and males



(b) Number of ergometries divided by medication status



(c) Number of ergometries divided by diagnosis status

Figure 5.1: Number of ergometries separated by gender, medication and diagnosis

5.1.2 Statistical indicators of the variables

Detailed assessment of statistical indicators is given in Table 5.1 along with the missing values in the data. It shows that just 8 variables have no missing values. Unfortunately the waist circumference is measured during just 3,800 ergometries. Each variable is discussed thoroughly in the upcoming subsections.

variable	min	$q_{0.25}$	$q_{0.5}$	$\bar{x}$	$q_{0.75}$	max	s	NA's
age	10	49	60	57.92	69	94	14.84	0
height	139	167	173	172.60	179	208	8.98	0
weight	32	69	78	78.99	88	160	14.82	0
waist	62	92	99	99.55	107	158	12.04	24825
BMI	15.22	23.74	26.03	26.44	28.73	55.28	4.06	0
WHR	0.37	0.53	0.57	0.57	0.61	0.89	0.07	24825
BSA	1.15	1.77	1.91	1.91	2.05	2.76	0.020	0
DBP_r	36	71	79	79.06	87	223	19.30	1339
DBP_max	36	76	86	86.67	96	257	30.88	315
SBP_r	6	106	118	119.40	130	291	13.23	1390
SBP_max	73	164	184	185.10	204	347	18.54	31
HR_r	35	67	76	77.61	86	179	14.88	98
HR_max	55	127	148	146.20	166	324	26.44	0
TC	78	158	188	193.19	223	434	47.82	9767
LDL	8.6	82.2	105.8	112.21	137.4	345.0	40.63	9839
HDL	16	43	51	52.83	61	157	14.09	9872
TG	27	86	118	140.75	167	1204	88.31	9767
Cr	0.58	0.97	1.09	1.12	1.23	3.29	0.024	14219
FBS	42	96	104	110.36	115	378	25.50	13443
UA	1.50	5.09	5.92	5.98	6.81	12.86	1.38	14152
TC:HDL	1.51	3.04	3.61	3.80	4.38	10.62	1.02	9872
LDL:HDL	0.29	1.62	2.04	2.19	2.63	6.18	0.80	9912
TG:HDL	0.30	1.53	2.35	3.02	3.63	45.19	0.56	9872
WL	21	113	150	158	200	440	58.53	0
WWR	0.24	1.50	1.96	2.02	2.46	6.10	0.73	0

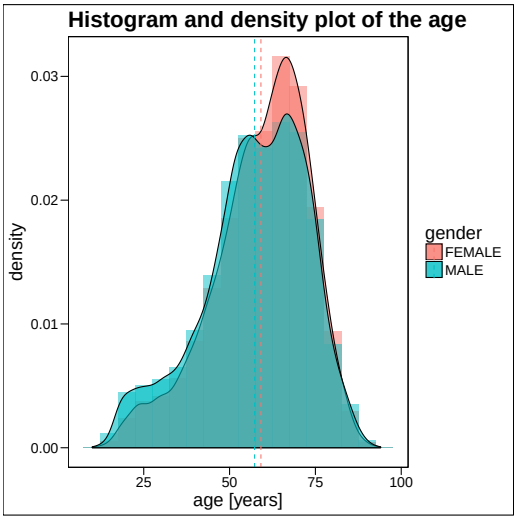
Table 5.1: Descriptive statistics of all variables with the following indicators: minimum, 1-quartile, median, average, 3-quartile, maximum, standard deviation and the number of missing values

5.1.3 Analysis of the variables age, height, weight and waist circumference

The following analysis shows box plots, histograms and density plots for the variables age, height, weight and waist circumference categorized by gender. These variables age, height and weight are measured for each ergometry. In contrast to that, the waist circumference is just available for 3,800 ergometries as it is measured only during out-patient rehabilitation.

Figure 5.2 shows a box plot, a histogram and a density plot for age. All of them are categorized by gender. The mean age of females is 59.13 with a standard deviation of 14.00, whereas for males mean age is 57.30 and the standard deviation of 15.22. Both plots show that the distribution of age is similar for both genders, although the median as well as the mean in females is slightly higher and the standard deviation is slightly smaller. The range for females and males is similar, the minimum is 10 and 11 and the maximum is 93 and 94, respectively.

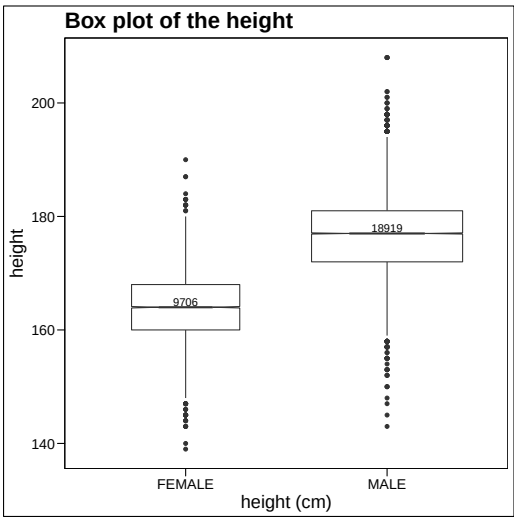
(a) Box plot of the age divided by gender



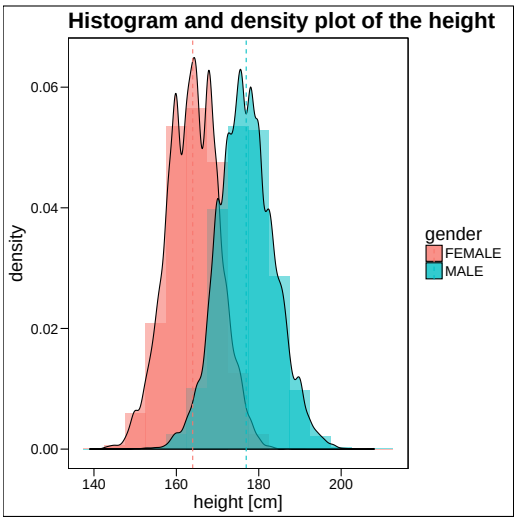
(b) Histogram (binwidth=5) and density plot of the age with marked mean divided by gender

Figure 5.2: Exploratory data analysis for the variable age

The mean height in females (164 cm) is 13 cm less than in males (177 cm). The standard deviation is similar in females and males. It is 6.16 and 6.74. The height of females spans from 139 cm to 190 cm while for male it is from 143 cm to 208 cm as seen in Figure 5.3.



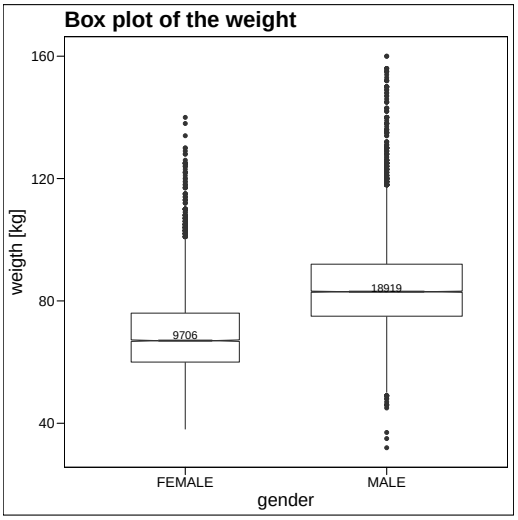
(a) Box plot of the height divided by gender



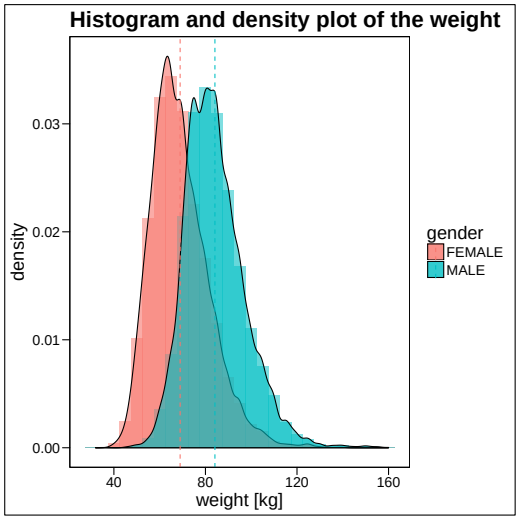
(b) Histogram (binwidth=5) and density plot of the height with marked mean divided by gender

Figure 5.3: Exploratory data analysis for the variable height

As that of variable height, similar trends, i.e. similar standard deviation and different mean, are present in weight as seen in Figure reffig:weight. The mean weight for females is 68.94 kg, whereas for men it is 84.14 kg. The maximum weight for male in the cohort is 160 kg and female 140 kg.



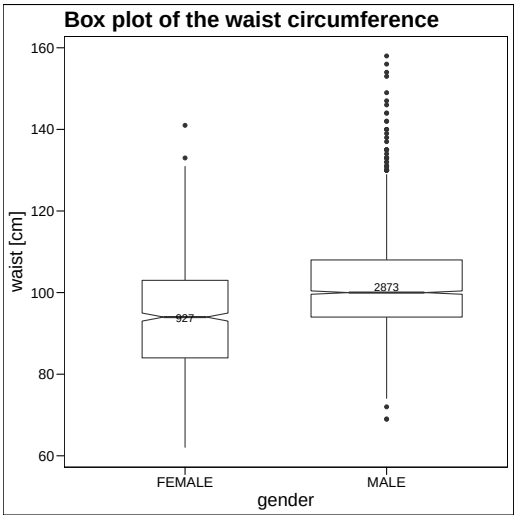
(a) Box plot of the weight divided by gender



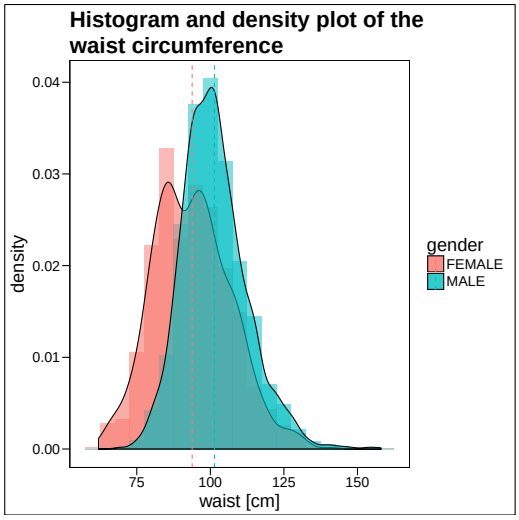
(b) Histogram (binwidth=5) and density plot of the weight with marked mean divided by gender

Figure 5.4: Exploratory data analysis for the variable weight

Figure 5.5 shows the plots for the waist circumference. Although the mean for female is 7.58 cm lower compared to male, the standard deviation is higher. Equivalent to the mean, the median is higher for male, and the length of the 95%-confidence interval around the median is much higher for both genders as compared to other variables due to availability of only 3,800 measurements.



(a) Box plot of the waist divided by gender

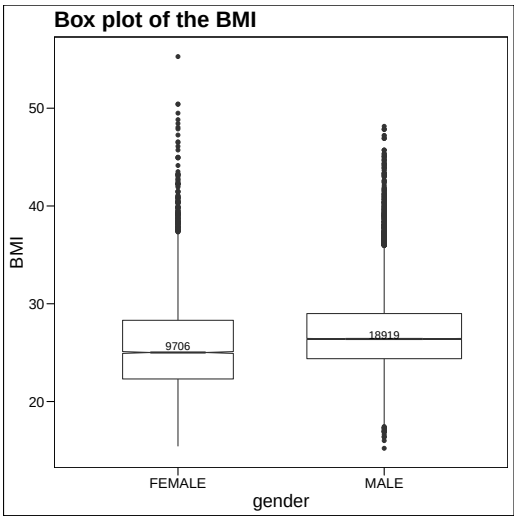


(b) Histogram (binwidth=5) and density plot of the waist with marked mean divided by gender

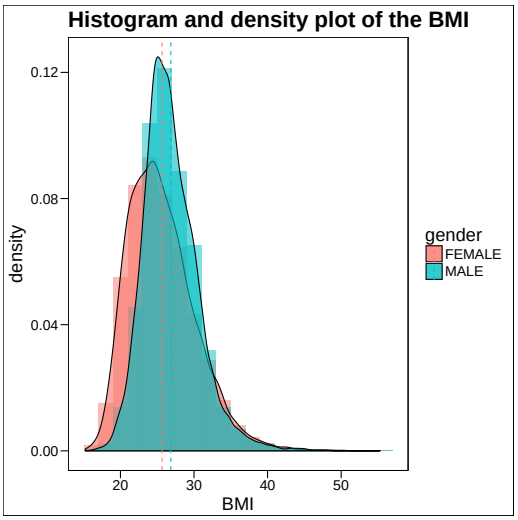
Figure 5.5: Exploratory data analysis for the variable waist

5.1.4 Analysis of the additionally calculated variables body mass index and waist-to-height-ratio

Both variables BMI and WHR are calculated from variables in the section above. The relevance of these variables for the heart attack risk is given in Chapter 2. Although the weight and the height of males and females are different, the BMI is similar for both genders as shown in Figure 5.6.



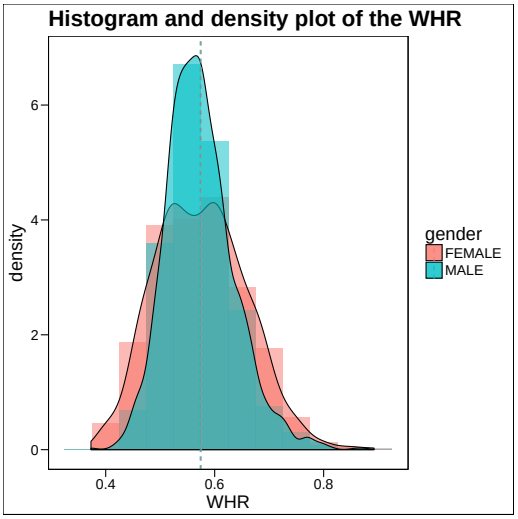
(a) Box plot of the BMI divided by gender



(b) Histogram (binwidth=2) and density plot of the BMI with marked mean divided by gender

Figure 5.6: Exploratory data analysis for the variable BMI

(a) Box plot of the of the WHR divided by gender



(b) Histogram (binwidth=0.05) and density plot of the of the WHR with marked mean divided by gender

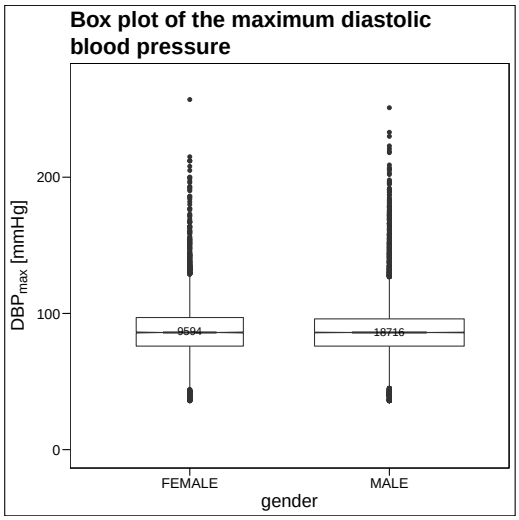
Figure 5.7: Exploratory data analysis for the variable WHR

Similar to the BMI, the analysis of variable WHR yields similar results for males and females. Although the mean of WHR is similar, the standard deviation is higher for females as seen in Figure 5.7.

5.1.5 Analysis of the blood pressure

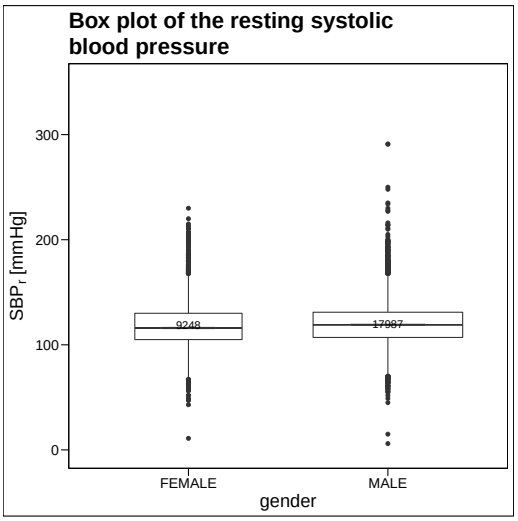
Figure 5.8 and Figure 5.9 show box plots of the resting blood pressure and blood pressure at maximum workload, respectively, separated by the gender. There is only a big difference in the systolic blood pressure at maximum workload among males and females but other values are similarly distributed in both genders.

(a) Box plot of the resting diastolic blood pressure divided by gender



(b) Box plot of the diastolic blood pressure at maximum workload divided by gender

Figure 5.8: Exploratory data analysis for the variable diastolic blood pressure

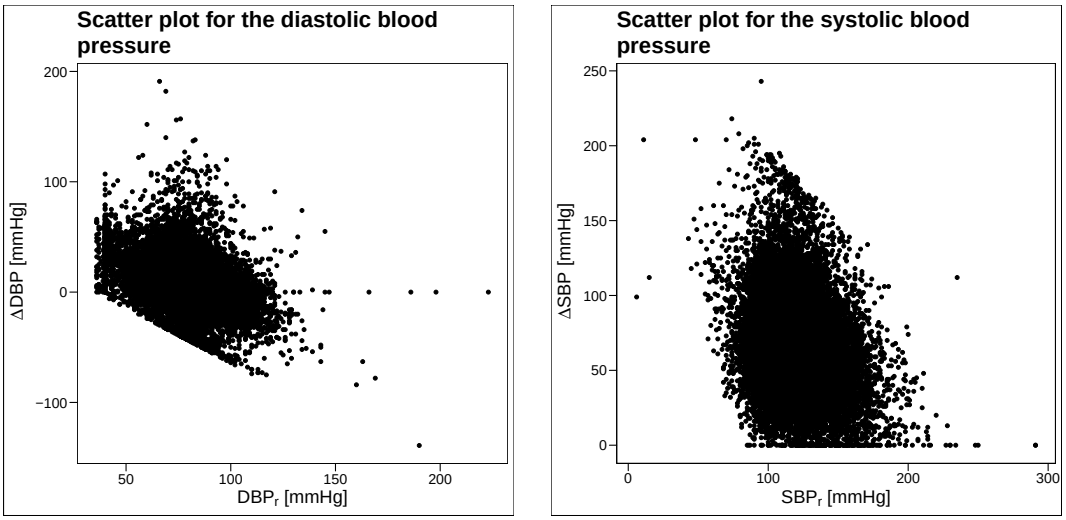


(a) Box plot of the resting systolic blood pressure divided by gender

(b) Box plot of the systolic blood pressure at maximum workload divided by gender

Figure 5.9: Exploratory data analysis for the variable systolic blood pressure

To analyze the correlation between the resting blood pressure and the increase in the blood pressure during an ergometry, the difference between blood pressure at the maximum workload and the resting blood pressure is calculated. Figure 5.10 shows the scatter plot of the resting blood pressure compared with the increase in blood pressure. It is interesting to note that the increase in the diastolic blood pressure is also negative. The reason for this could be the difficulty in measuring the diastolic blood pressure during the exercise, which results in lower diastolic blood pressure values than actual. For both systolic and diastolic blood pressure the increase can get higher if the resting blood pressure is low.



(a) Scatter plot for the resting diastolic blood pressure and the increase of the blood pressure during the ergometry

(b) Scatter plot for the resting systolic blood pressure and the increase of the blood pressure during the ergometry

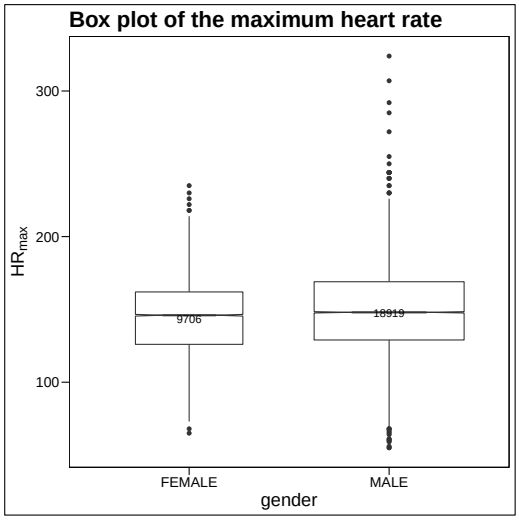
Figure 5.10: Increase of the blood pressure during ergometry

5.1.6 Analysis of the heart rate

Figure 5.11 shows gender specific box plots for the resting and the maximum heart rate. It shows that the heart rate is very similar between the genders, but the median of the resting heart rate is slightly higher for females while the median of the maximum heart rate is higher for males.

Figure 5.12 shows the relationship between resting heart rate and increase of the heart rate during an ergometry. The increase of the heart rate is a difference between the resting heart rate and the heart rate at the end of that ergometry. No negative values indicate that during all ergometries, heart rate was increased.

(a) Box plot of the resting heart rate separated by gender



(b) Box plot of the maximum heart rate separated by gender

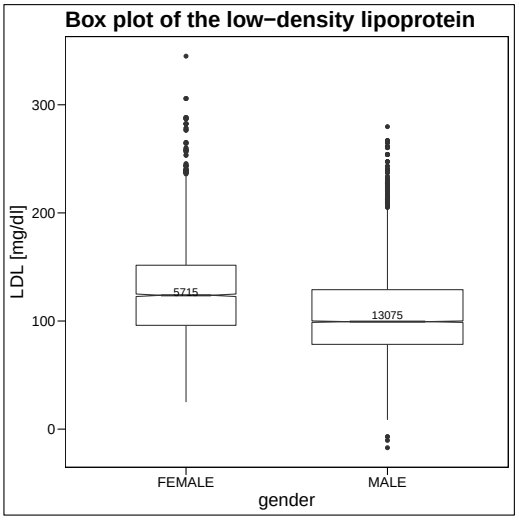
Figure 5.11: Exploratory data analysis for the variable heart rate

Figure 5.12: Scatter plot for the resting heart rate and the increase of the heart rate during an ergometry

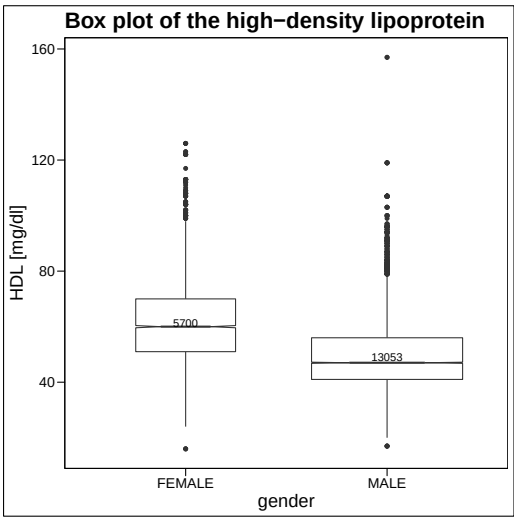
5.1.7 Analysis of the blood parameters

For the ergometries 7 different blood parameters were measured. Figure 5.13 shows the box plots of the variables TC, LDL, HDL, TG. The median of the variables TC, LDL and HDL are higher for females. In contrary, the median for the variable TG is similar, but some outliers are far away from the median.

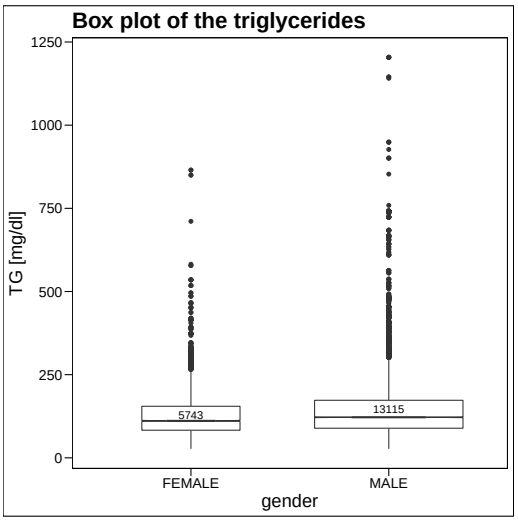
(a) Box plot of the total cholesterol separated by gender



(b) Box plot of the low-density lipoprotein separated by gender



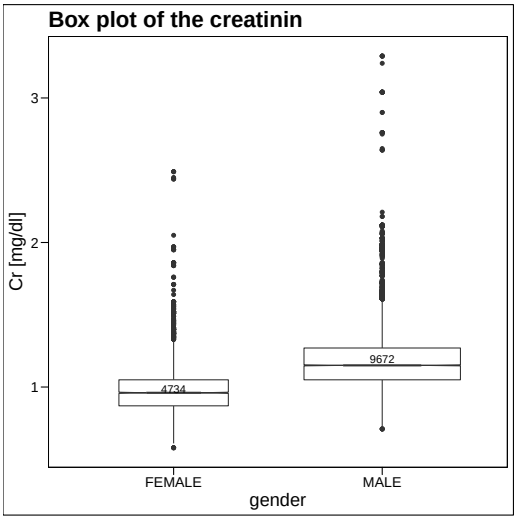
(c) Box plot of the high-density lipoprotein separated by gender



(d) Box plot of the triglycerides separated by gender

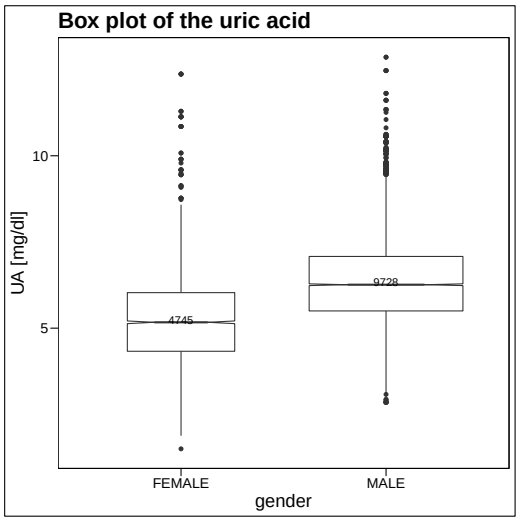
Figure 5.13: Box plots for the variables TC, LDL, HDL, TG

The analysis of the variables Cr, FBS and UA, Figure 5.14, shows higher medians for male than for female.



(a) Box plot of the creatinin separated by gender

(b) Box plot of the fasting blood sugar level separated by gender

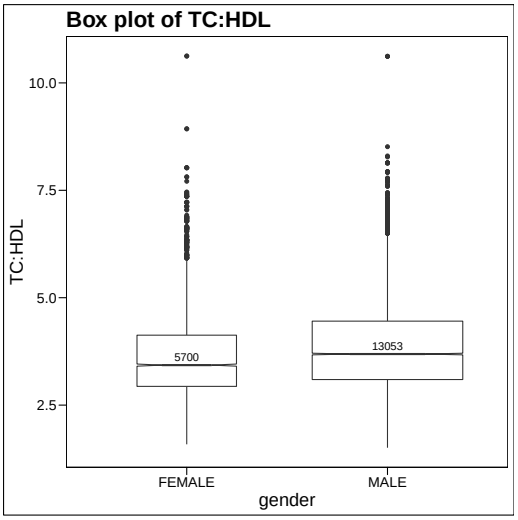


(c) Box plot of the uric acid separated by gender

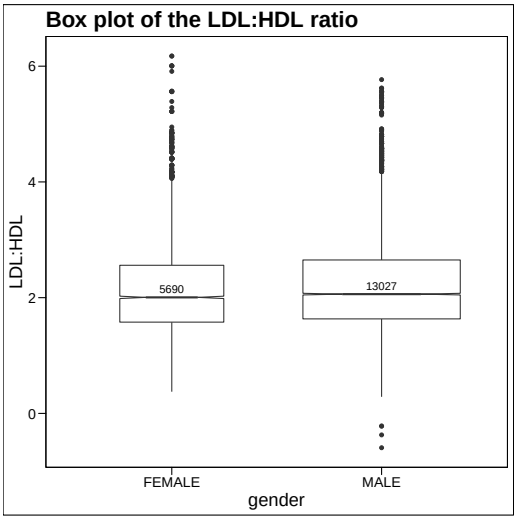
Figure 5.14: Box plots for Cr, FBS and UA

5.1.8 Additionally calculated variables TC to HDL ratio, LDL to HDL ratio and TG to HDL ratio

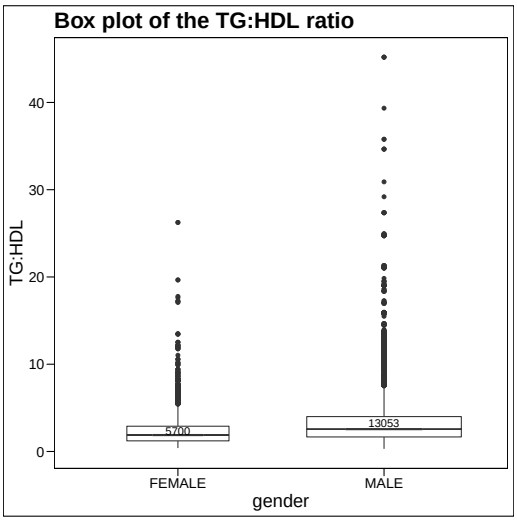
For analyzing the effect of blood cholesterol level on cardiac health, the TC:HDL ratio, the LDL:HDL ratio and the TG:HDL ratio are calculated. These parameters are better indicators of cardiac health. Figure 5.15 shows that there is no remarkable difference between the genders.



(a) Box plot of TC-to-HDL-ratio separated by gender



(b) Box plot of LDL-to-HDL-ratio separated by gender

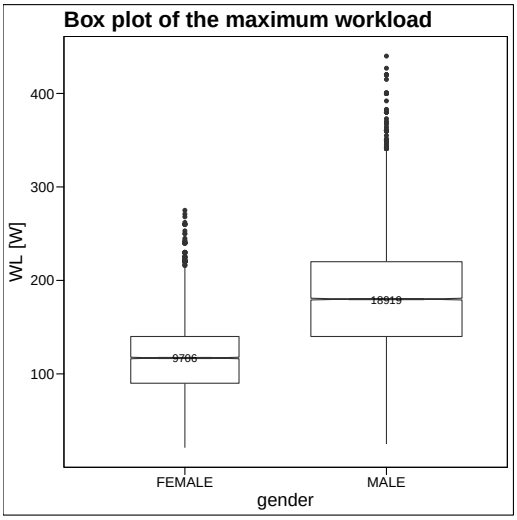


(c) Box plot of TG-to-HDL-ratio separated by gender

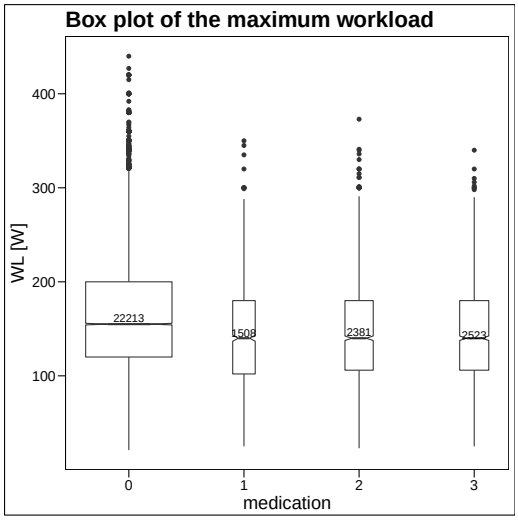
Figure 5.15: Box plots for the calculated variables TC:HDL, LDL:HDL and TG:HDL

5.1.9 Analysis of the maximum workload and maximum workload-to-weight-ratio

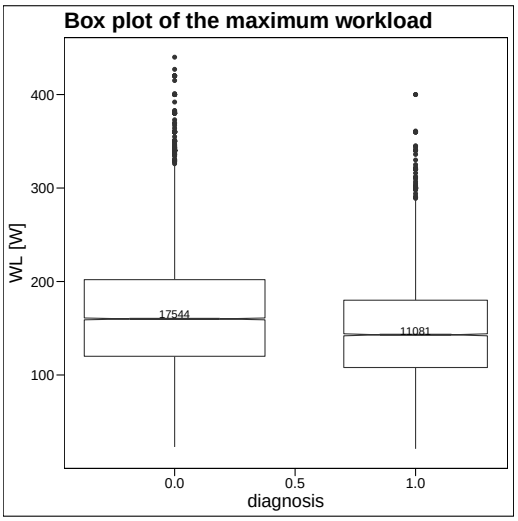
Finally the variables WL and WWR are analyzed. Figure 5.16 shows box plots categorized by gender, medication and diagnosis, respectively, and a histogram and density plot of the maximum workload. The mean of the variable maximum workload is much higher for men. If participants of the ergometry have an cardiac diagnosis or take statin or beta blocker, then the maximum workload is decreasing.



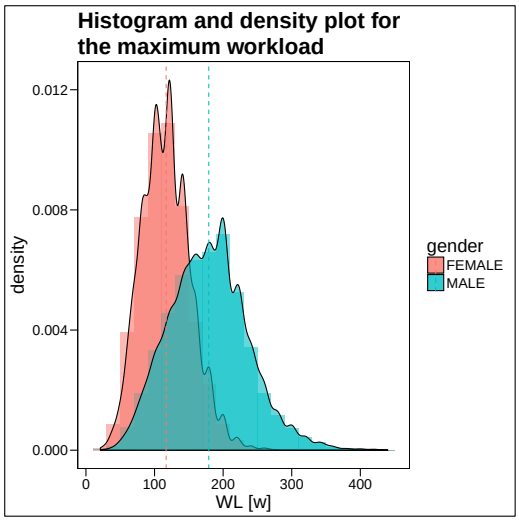
(a) Box plot of the maximum workload divided by gender



(b) Box plot of the maximum workload divided by medication history



(c) Box plot of the maximum workload divided by diagnosis

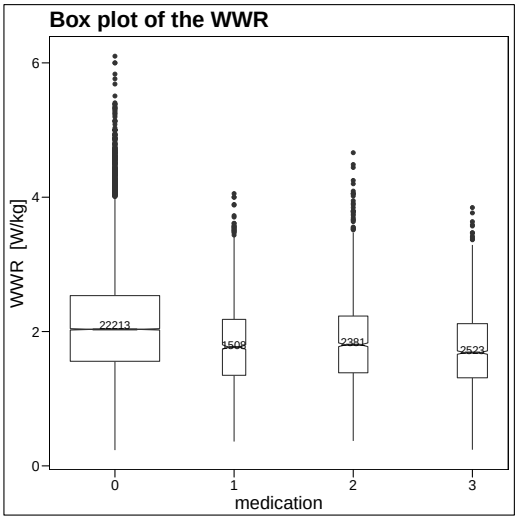


(d) Histogram and density plot for the maximum workload separated by gender

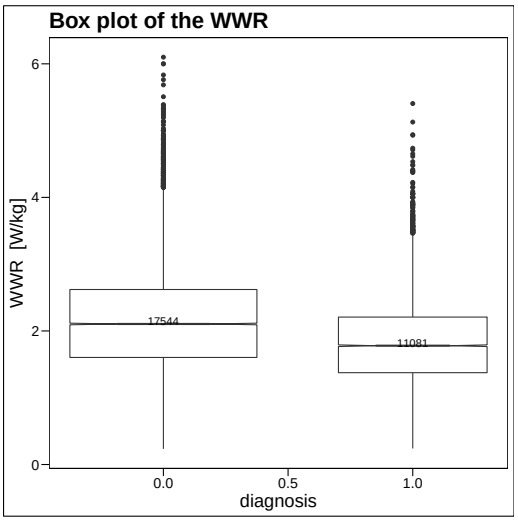
Figure 5.16: Exploratory data analysis for the variable maximum workload

Figure 5.17 shows box plots categorized by gender, medication and diagnosis, respectively, and a histogram of the maximum workload-to-weight-ratio. The difference between the maximum workload-to-weight-ratio between males and females is less than in the maximum workload.

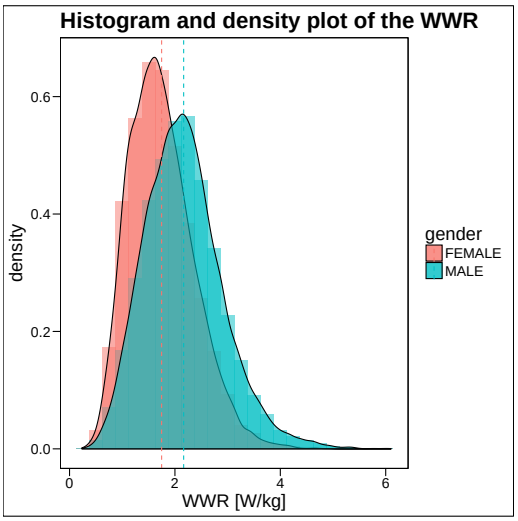
(a) Box plot of the maximum workload-to-weight-ratio divided by gender



(b) Box plot of the maximum workload-to-weight-ratio divided by medication



(c) Box plot of the maximum workload-to-weight-ratio divided by diagnosis



(d) Histogram and density plot of the maximum workload-to-weight-ratio divided by gender

Figure 5.17: Exploratory data analysis for the variable maximum-workload-to-weight-ratio

5.2 Multivariate analysis

The multivariate analysis is used to investigate the correlation between more than one parameters.

5.2.1 Correlation

The correlation between the variables is explored in a correlation matrix, which explores the linear relationship. Figure 5.18 and Figure 5.19 show the correlation between the variables. One can see that the correlation to the WL is in average less than to the WWR. The WWR has a correlation of more than 0.5 with the age and the waist-to-height-ratio. Further, four variables, height, waist, BMI and FBS have a correlation of more than 0.25 to the WWR.

Due to the low correlation between the cardiac parameters and the blood parameters with the WWR, in the regression analyses of the next chapter these variables will not be taken into account.

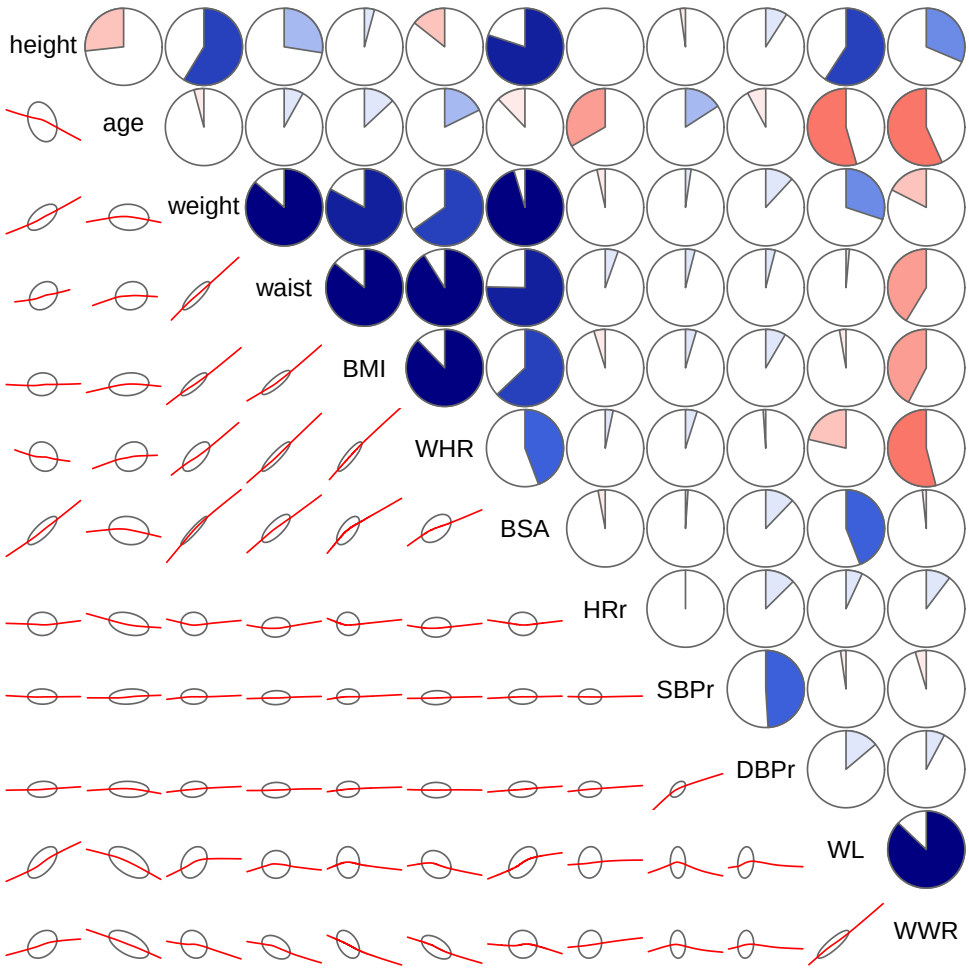


Figure 5.18: Correlation between the variables of the basic information and the measured parameters during an ergometry, the WL and the WWR for all ergometries together

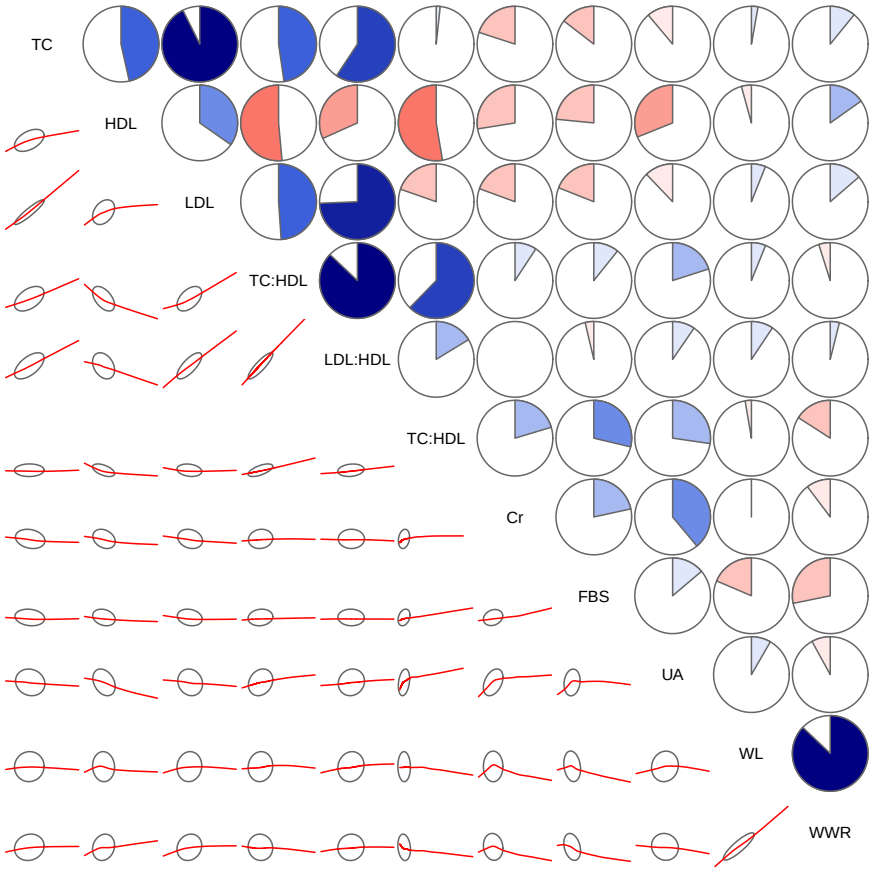
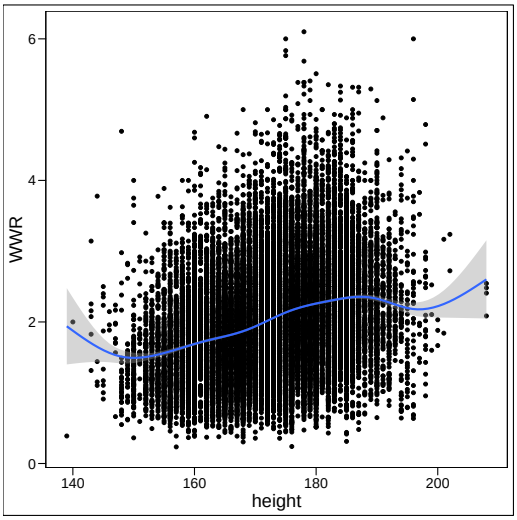


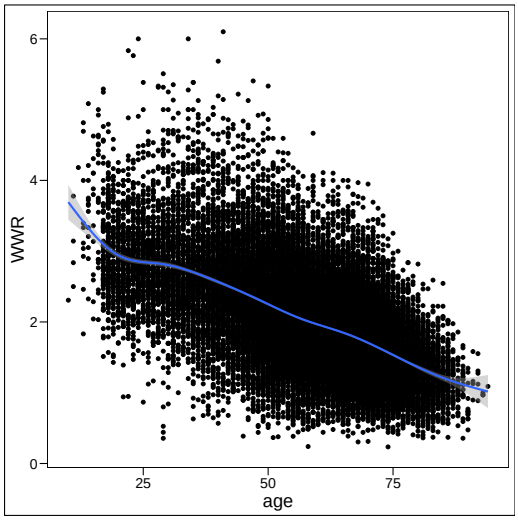
Figure 5.19: Correlation of the blood parameters, the WL and the WWR of all ergometries together

5.2.2 Scatter plots

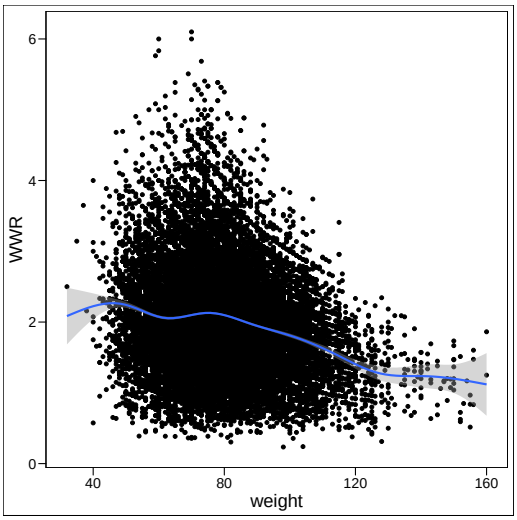
For the variables of the basic information scatter plots with the WWR are shown. Figure 5.20 shows scatter plots for the variables height, age, weight and waist and the WWR and Figure 5.21 shows the plots for the calculated variables BMI, WHR and BSA and the WWR. In both Figures also regression lines are drawn in.



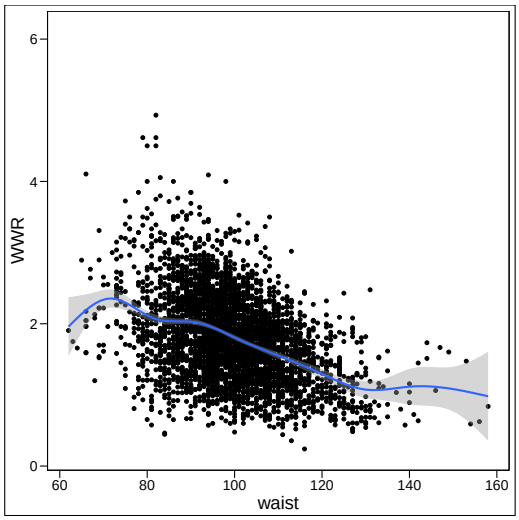
(a) Scatter plot between height and maximum workload-to-weight-ratio



(b) Scatter plot between age and maximum workload-to-weight-ratio

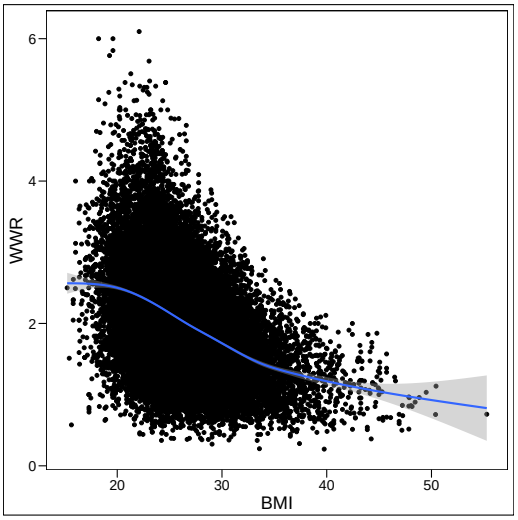


(c) Scatter plot between weight and maximum workload-to-weight-ratio

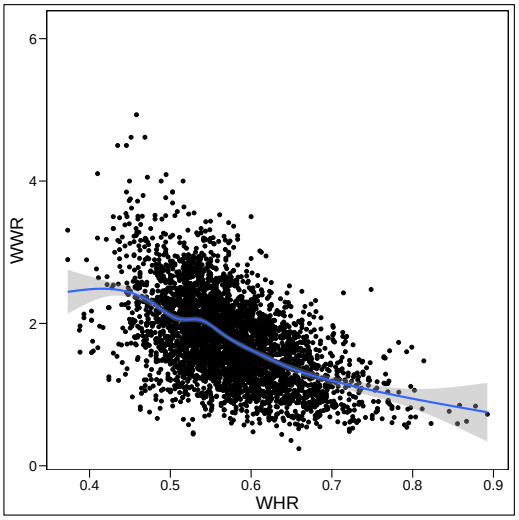


(d) Scatter plot between waist and maximum workload-to-weight-ratio

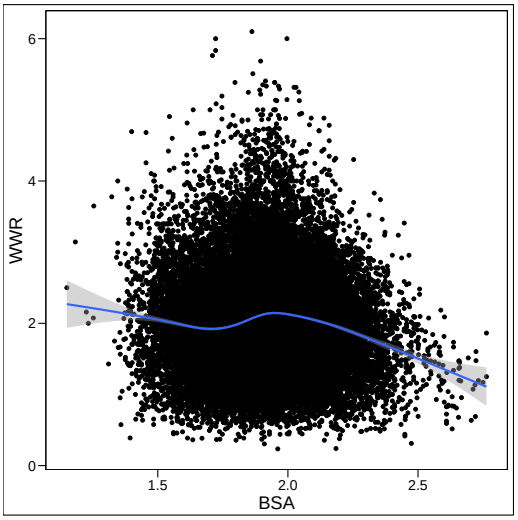
Figure 5.20: Scatter plots between the variables of the basic information and the maximum workload-to-weight-ratio



(a) Scatter plot between body mass index and maximum workload-to-weight-ratio



(b) Scatter plot between waist-to-height-ratio and maximum workload-to-weight-ratio



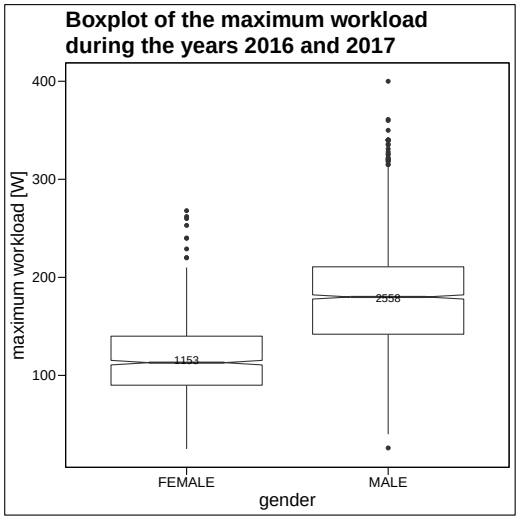
(c) Scatter plot between body surface area and maximum workload-to-weight-ratio

Figure 5.21: Scatter plots between the calculated varaibles of the basic information and the maximum workload-to-weight-ratio

5.3 Quality control of the data

Figure 5.22 asserts that the maximum workload is distributed similarly at the beginning (2004 & 2005) and the end of the study (2016 & 2017). A deviation of the maximum workload in the beginning and the end of the study would point to an error in the measurement.

(a) Boxplot of the maximum workload during the years 2004 and 2005



(b) Boxplot of the maximum workload during the years 2016 and 2017

Figure 5.22: Quality control for the maximum workload during time

5.4 Discussion

In summary, in this chapter, the data is analyzed using both univariate and multivariate EDA. The first aim of this analysis was to analyze the data to find incorrect entries and eliminate them. Without checking the data for quality, incorrect entries would have an influence on the results of the regression ultimately yielding wrong results. The second objective was to get a general overview of the data and find its structures. These results can be further used in the next chapters, in which normal values for the maximum workload and the maximum workload-to-weight-ratio are derived.

The first observation of this analysis is that the variables weight, height and waist differ a lot in gender, whereas the BMI and WHR are similar for both genders. This observation can be used for choosing the variables for the regression analysis. Another observation is that the variable WWR has a higher correlation to the other variables than the maximum workload. Consequently, it is preferable to use the WWR as a prediction variable in the regression analysis instead of the maximum workload.

The cardiac parameters and the blood parameters have a low correlation to the maximum workload and the maximum workload-to-weight-ratio. Whereas, the variables of basic information have quite a good correlation to them. For this reason, the variables for cardiac parameters and blood parameters are omitted from further analysis. The EDA showed that there is approximately a linear relationship between the variables of basic information and the maximum workload and maximum workload-to-weight-ratio. Therefore a linear regression can be used for the calculation of normal values.

Chapter 6

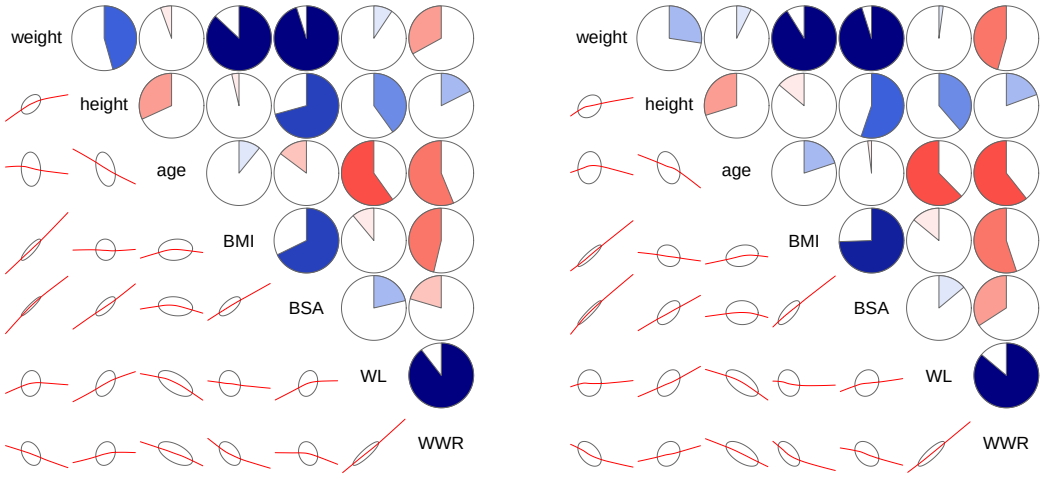
Normal values for the maximum workload of exercise testing

The aim of this chapter is to model normal values for the maximum workload of ergometries. The chosen method for this is a linear regression. The currently used normal values for ergometries in Austria are suggested in the Austrian practice guidelines for ergometries ([Wonisch et al., 2008]). However, the information about the applied data to develop this model is not available. For this reason, in this Chapter, the currently used normal values are analyzed and compared with new approaches which use similar independent variables. As a starting point, a correlation analysis is done in Section 1 followed by regression trees in Section 2. In Section 3 to Section 7 different models are introduced. Section 3 explains the model used. First, a linear regression model with the same effects is derived in Section 4. Afterwards, the best model with the same independent variables but with all main effects and all intersections is built in Section 5. Next, the model is extended by the variables weight and height in Section 6. As the last step in Section 7, instead of the maximum workload, the maximum workload-to-weight-ratio is used as the dependent variable. In Section 8, all models are compared along with a conclusion. Finally, in Section 9 a dummy regression model is introduced.

6.1 Correlation

Figure 6.1 shows the correlation between weight, height, age, body surface area (calculated with the formula of Du Bois), the outputs of an ergometry maximum workload and maximum workload-to-weight-ratio. In this chapter, there are regression models for the maximum workload as well as for the maximum workload-to-weight-ratio. One result of Chapter 5 is that the maximum workload-to-weight-ratio is preferable over the maximum workload.

Since the body surface area is calculated with the formula of Du Bois, the correlation in Figure 6.1 between it and the weight and height is very high. The correlation between the other independent variables is quite low.

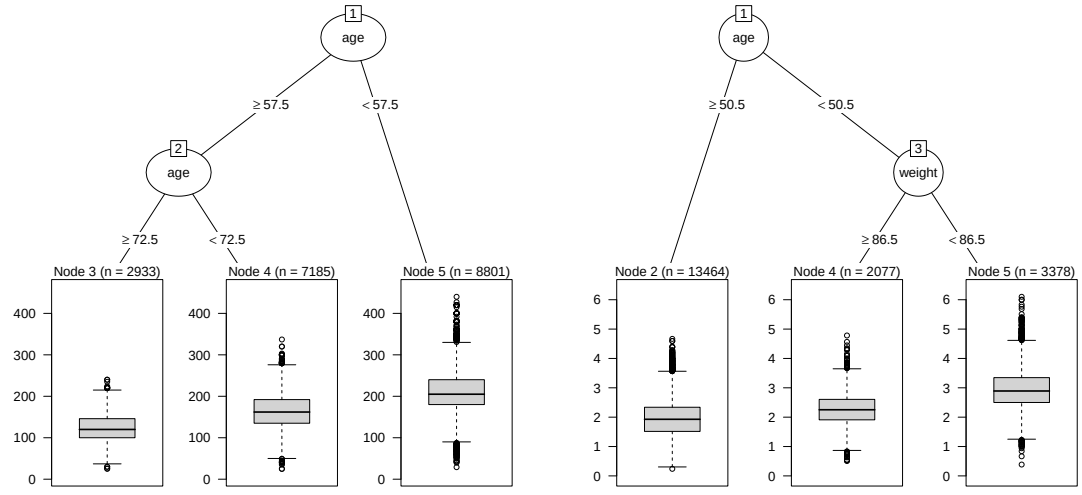


(a) Male (b) Female

Figure 6.1: Analysis of correlation with the variables weight, height, age, BMI, BSA, maximum workload and maximum workload-to-weight-ratio

6.2 Regression trees

This section shows regression trees for the maximum workload and the maximum workload-to-weight-ratio, respectively, in Figure 6.2 and Figure 6.3.

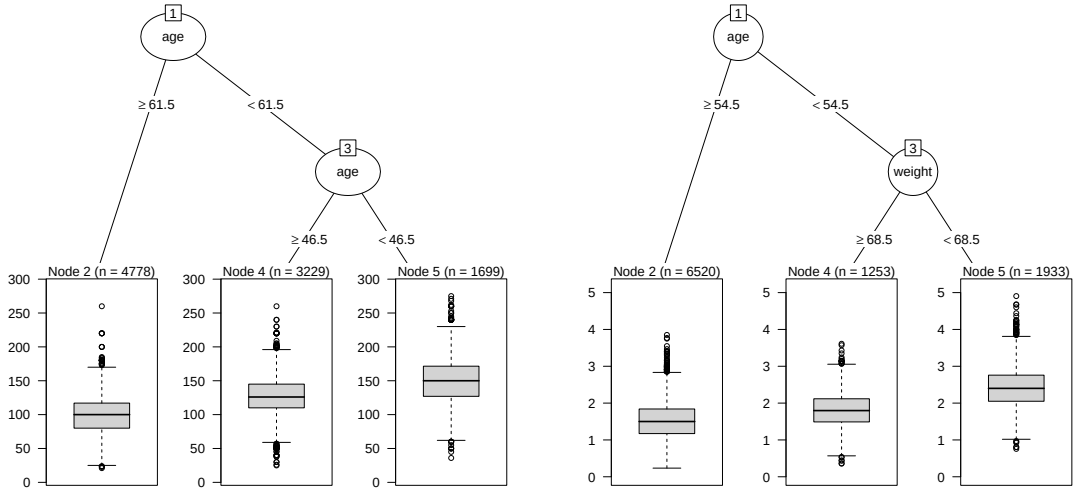


(a) Regression tree for the maximum workload (b) Regression tree for the maximum workload-to-weight-ratio

Figure 6.2: Regression trees for males with the independent variables weight, height, age, body surface area and the dependent variables maximum workload and maximum workload-to-weight-ratio

The structure of the trees is the same for males and females, i.e. in the regression trees for the maximum workload in both decision vertices the age

plays a role and in the regression trees for the maximum workload-to-weight-ratio, the first vertex splits the age and the second one the weight.



(a) Regression tree for the maximum workload (b) Regression tree for the maximum workload-to-weight-ratio

Figure 6.3: Regression trees for females with the independent variables weight, height, age, BSA and the dependent variables maximum workload and maximum workload-to-weight-ratio

6.3 Model P: Linear regression model from the Austrian practice guidelines for ergometries

The Austrian practice guidelines for ergometries ([Wonisch et al., 2008]) contains a formula for the physical performance. This formula traced back to the publication from [Arstila, 1972]. With this formula, the normal values of the maximum workload can be calculated by knowing the body surface area and the age. Because males and females reach different values, there exist two different formulas separated by gender:

$$\begin{aligned} \text{male: } WL_{max}[W] &= 6.773 + 136.141 BSA[m^2] - 0.916 BSA[m^2] age[y] \\ \text{female: } WL_{max}[W] &= 3.933 + 86.641 BSA[m^2] - 0.346 BSA[m^2] age[y]. \end{aligned}$$

For determining the body surface area the formula of Du Bois is used in the formulas above (see [Du Bois and Du Bois, 1916]):

$$BSA[m^2] = 0.007148 weight[kg]^{0.425} height[cm]^{0.725}$$

Applying this formula on the data from ZARG provides the results in Table 6.1.

response variable	Model P	
	male	female
SSR	20430888	1745237
SSE	39733086	9513851
SST	60964358	12347738
R^2_{adj}	0.3483	0.2295

Table 6.1: Results of Model P

6.4 Model 1: Linear regression model for the maximum workload depending on the surface and the interaction of the BSA and the age

Model 1 has the same main effect and interactions as the currently used model, i.e. the surface and the interaction between the BSA and the age:

$$\hat{WL}[W] = \hat{\beta}_0 + \hat{\beta}_1 BSA[m^2] + \hat{\beta}_2 BSA[m^2] age[y].$$

6.4.1 Analysis of the model

Table 6.2 and Table 6.3 summarize the results of Model 1 for males and females. In the models for both genders, all coefficients are highly significant. The adjusted coefficient of determination R^2_{adj} is for male 0.3744 and for female it is 0.4055.

$WL \sim BSA + BSA : age$						
Coefficients:						
	Estimate	Std. Error	t value	Pr(>|t|)		
(Intercept)	93.19284	4.05615	22.98	<2e-16	***	
BSA	105.67996	2.02190	52.27	<2e-16	***	
BSA:age	-1.09894	0.01102	-99.74	<2e-16	***	
RSE	44.9 on 18916 degrees of freedom					
R^2	0.3745					
R^2_{adj}	0.3744					
F-statistic	5663 on 2 and 18916 DF					
p-value:	< 2.2e-16					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
BSA	1	2777209	2777209	1377.6	< 2.2e-16	***
BSA:age	1	20054085	20054085	9947.9	< 2.2e-16	***
Residuals	18916	38133065	2016			
Signif. codes	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

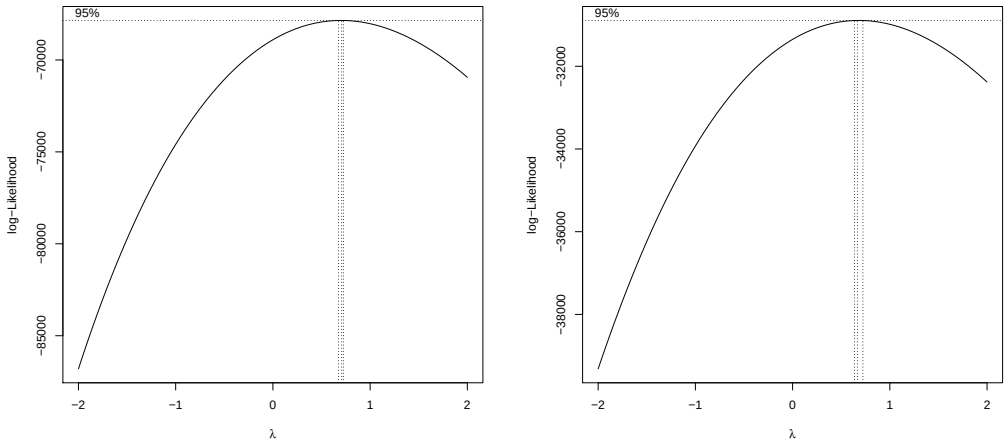
Table 6.2: Summary and ANOVA of Model 1 for males

$WL \sim BSA + BSA : age$						
Coefficients:						
	Estimate	Std. Error	t value	Pr(>|t|)		
(Intercept)	67.2302	3.1903	21.07	<2e-16	***	
BSA	83.0788	1.9365	42.90	<2e-16	***	
BSA:age	-0.9214	0.0116	-79.41	<2e-16	***	
RSE	27.5 on 9703 degrees of freedom					
R^2	0.4057					
R^2_{adj}	0.4055					
F-statistic	3311 on 2 and 9703 DF					
p-value:	< 2.2e-16					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
BSA	1	240237	240237	317.64	< 2.2e-16	***
BSA:age	1	4768876	4768876	6305.32	< 2.2e-16	***
Residuals	9703	7338625	756			
Signif. codes	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 6.3: Summary and ANOVA of Model 1 for females

6.4.2 Diagnostics of the model

The Box-Cox plots in Figure 6.4 show that a transformation of the maximum workload is necessary for both genders. Due to better comparability with the formula of [Wonisch et al., 2008] this is not done in Model 1.



(a) Male (b) Female

Figure 6.4: Box-Cox plots of Model 1

Following the regression, the requirements of the regression model have to be checked. Figure 6.5 and Figure 6.6 show the results of the diagnostics for male and female, respectively.

The first plot of Figure 6.5 and Figure 6.6, respectively, shows a scatter plot between residuals and predicted values. For both genders, a small pattern is

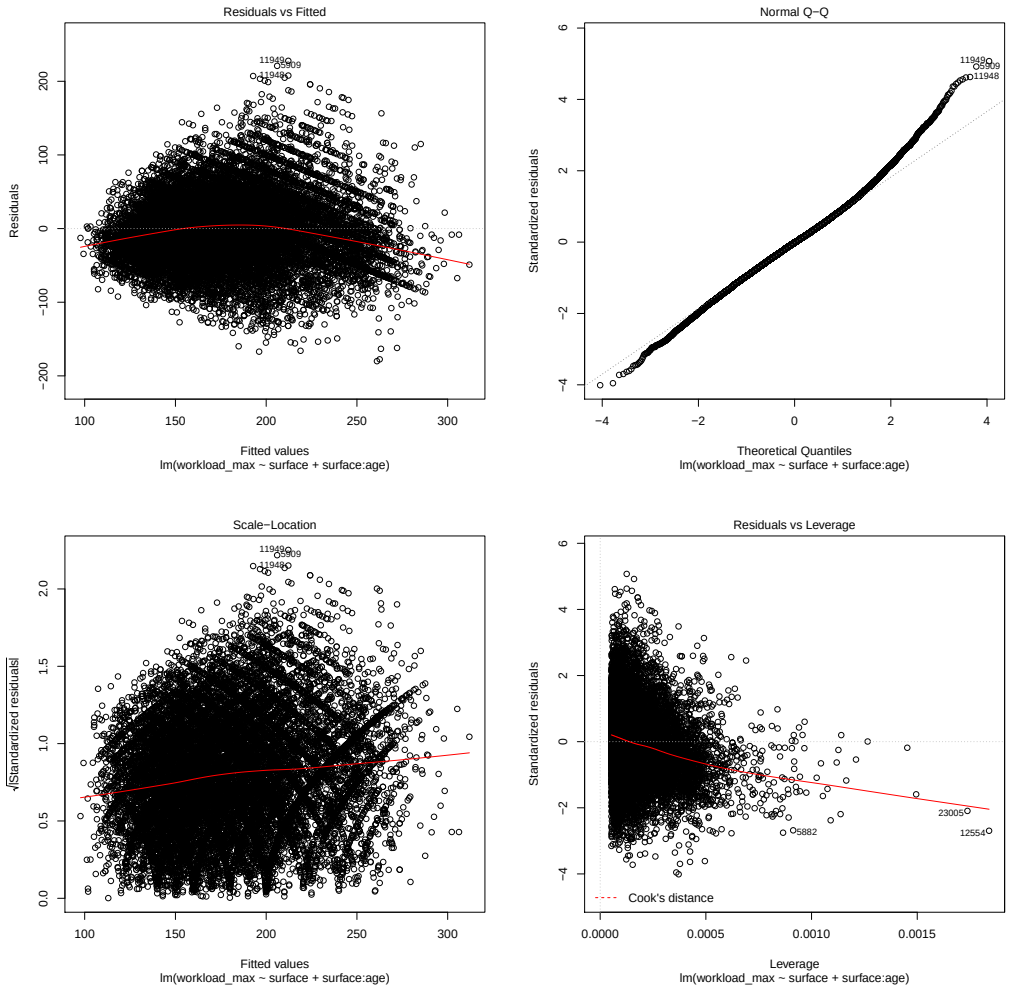


Figure 6.5: Diagnostic plots of Model 1 for males

recognizable, although it should look random. The second plot of Figure 6.5 and Figure 6.6 shows a Q-Q plot of the residuals. It will give a straight line if the errors are normally distributed. Both plots for males and females show that for very small and big values there is a deviation from this line, i.e. there is a skinny positive and skinny negative tail. The third plot of Figure 6.5 and Figure 6.6 should look random with no patterns, which is the case in both plots. The last plot of Figure 6.5 and Figure 6.6 identifies the points which have the greatest influence on the regression (leverage points).

6.4.3 Confidence intervals and prediction intervals for Model 1

In this subsection, the confidence intervals and prediction intervals are created in Table 6.4 for males and in Table 6.5 for females. In this table, for the age and BSA, three different values are chosen. One can see that the confidence intervals are much tighter than the prediction intervals. Note, that the intervals are not reasonable for Model 1, because in Model 1 there was no transformation done although it was necessary.

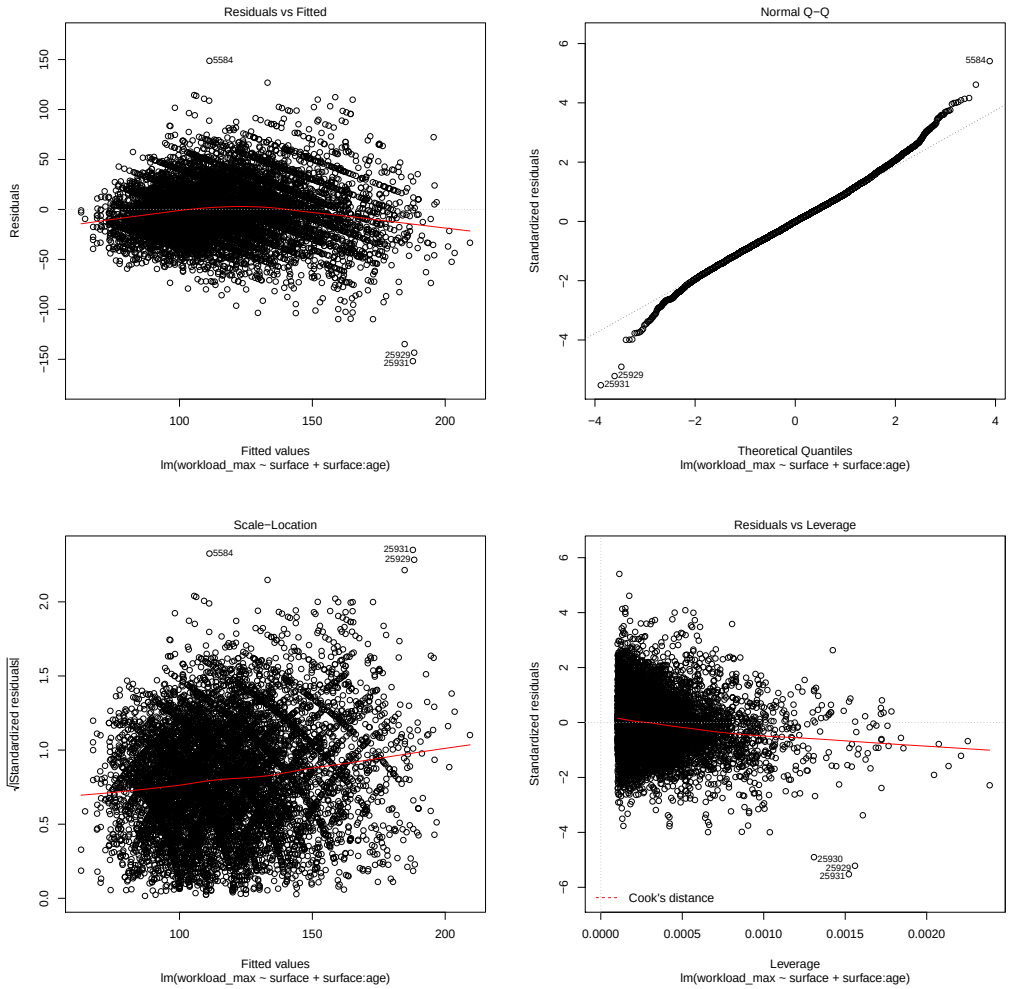


Figure 6.6: Diagnostic plots of Model 1 for females

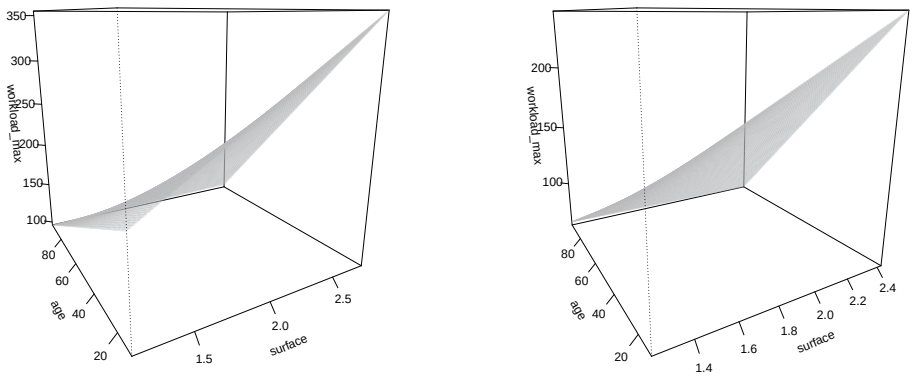
Men			
age/BSA	1.5	2	2.5
25	210.5 [208.04,212.96] [122.46,298.54]	249.61 [248.08,251.13] [161.59,337.63]	288.71 [286.2,291.21] [200.67,376.75]
50	169.29 [167.16,171.42] [81.26,257.32]	194.66 [193.95,195.37] [106.65,282.67]	220.03 [217.97,222.09] [132,308.06]
75	128.08 [126,130.16] [40.05,216.11]	139.71 [138.71,140.71] [51.7,227.72]	151.34 [148.92,153.76] [63.3,239.38]

Table 6.4: Prediction values for the maximum workload [W] (first row), confidence intervals (second row) and prediction intervals (third row) of Model 1 for males

6.4.4 Regression surface of Model 1

Women			
age/BSA	1.5	2	2.5
25	157.3 [155.74,158.85] [103.36,211.23]	187.32 [185.44,189.19] [133.38,241.26]	217.34 [213.98,220.7] [163.33,271.35]
50	122.74 [121.68,123.81] [68.82,176.66]	141.25 [140.09,142.41] [87.33,195.17]	159.75 [156.93,162.58] [105.77,213.73]
75	88.19 [87.06,89.33] [34.27,142.11]	95.18 [93.85,96.5] [41.25,149.1]	102.16 [99.21,105.12] [48.18,156.15]

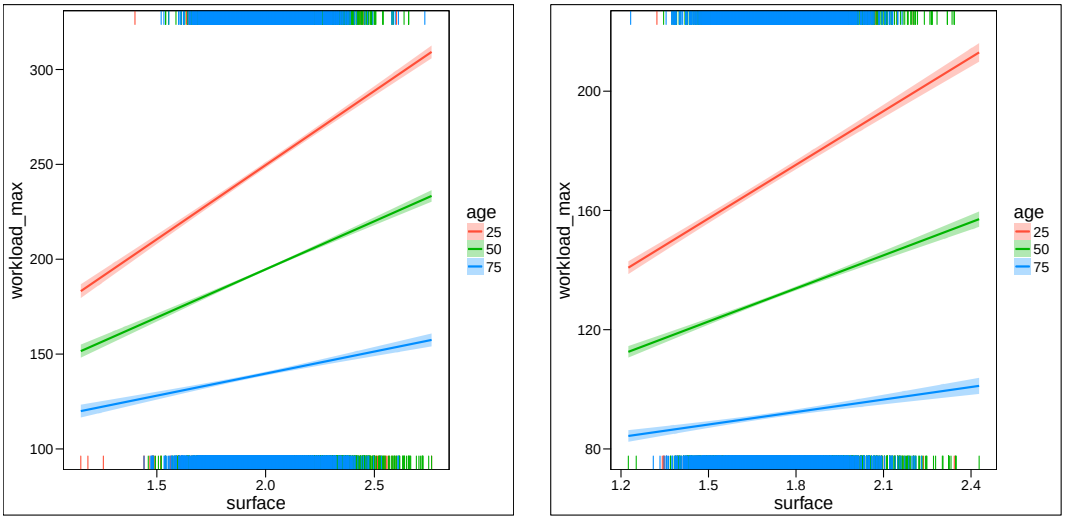
Table 6.5: Prediction values for the maximum workload [W] (first row), confidence intervals (second row) and prediction intervals (third row) of Model 1 for females



(a) Male (b) Female

Figure 6.7: Plots of the regression surface of Model 1

Figure 6.7 shows the regression surface of Model 1. One can see that there is a main effect and also interaction. This interaction can be seen more precisely in Figure 6.8. Here the age is fixed at the three different stages from the tables above. The slope of the regression lines of the surface is different for each of these ages. Additionally, the 95%-confidence bands are plotted.



(a) Male (b) Female

Figure 6.8: Plots of the regression line and confidence bands of the BSA for different ages of Model 1

6.5 Model 2: Linear regression model for the maximum workload depending on the body surface area and the age and its interaction

Based on the same independent variables, body surface area and age, the best model is chosen. Both main effects and the interaction can be considered, but to reduce the number of variables, a variable selection is done. Therefore, all possible subsets of the pool of the explanatory variables are calculated, and the model that fits the data best according to the chosen criterion is found.

6.5.1 Variable selection

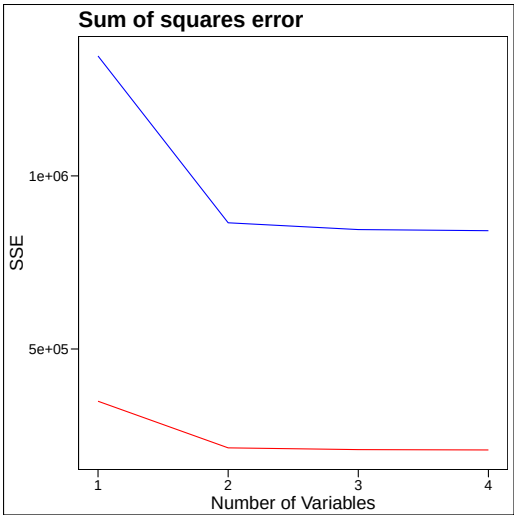
Table 6.6 shows the best model for different numbers of parameters. In the best model with one parameter, just the age is chosen and in the best model with two parameters the BSA, and the age are chosen. This is in contrast to the model which is currently used, where the BSA and the interaction are used. This will hold for both males and females.

no. of parameters	BSA	age	BSA:age
1		*	
2	*	*	
3	*	*	*

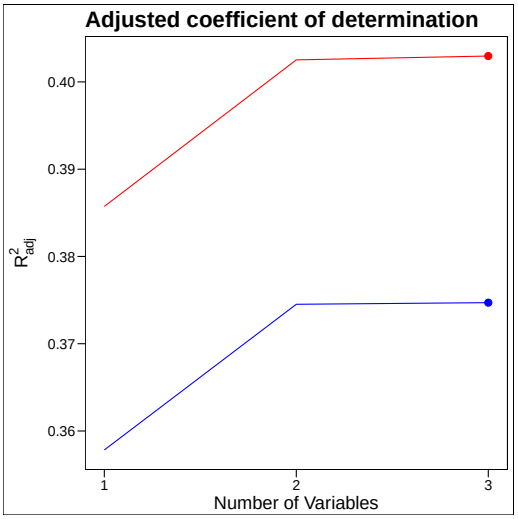
Table 6.6: Summary of the best models with 1 to 3 parameters for deriving Model 2 for males and females

In a next step, the best model among these models in Table 6.6 is chosen. For this selection, Figure 6.9 shows plots for the different selection criteria. Fig-

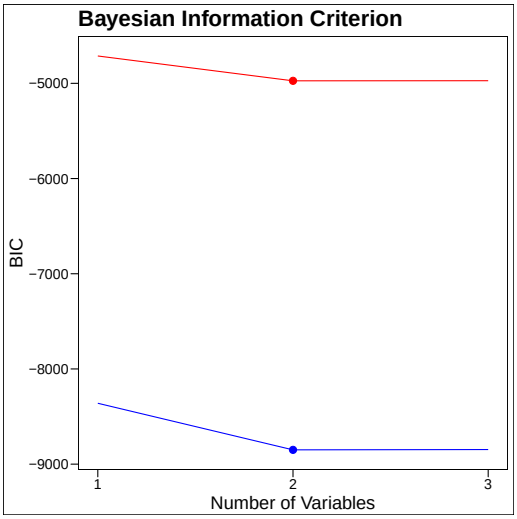
ure 6.9(a) shows the error sum of squares SSE, which should be small. The adjusted coefficient of determination R^2_{adj} is shown in Figure 6.9(b). The resulting value of the R^2_{adj} should be as high as possible. The chosen selection criterion for our models is the Bayesian Information Criterion because this criterion chooses quite small models.



(a) Graphical description of the error sum of squares for the best models with 1 to 4 parameters (blue lines refer to male, red lines refer to female)



(b) Graphical description of the R^2_{adj} for the best models with 1 to 3 parameters (blue lines refer to male, red lines refer to female)



(c) Graphical description of the BIC for the best models with 1 to 3 parameters

Figure 6.9: Information criteria for deriving Model 2

As a result in the final version of Model 2 for both males and females, the two variables BSA and age are chosen. Therefore, the following regression model is used:

$$\hat{WL}^{0.7}[W] = \hat{\beta}_0 + \hat{\beta}_1 BSA[m^2] + \hat{\beta}_2 age[y].$$

6.5.2 Analysis of the model

Table 6.7 and Table 6.8 show a summary of the chosen regressions for males and females. All coefficients are highly significant.

$WL^{0.7} \sim BSA + age$						
Coefficients:						
	Estimate	Std. Error	t value	Pr(>|t|)		
(Intercept)	42.240422	0.655197	64.47	<2e-16	***	
BSA	6.747887	0.300029	22.49	<2e-16	***	
age	-0.320909	0.003223	-99.58	<2e-16	***	
RSE	6.672 on 18916 degrees of freedom					
R^2	0.3746					
R^2_{adj}	0.3745					
F-statistic	5665 on 2 and 18916 DF					
p-value:	< 2.2e-16					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
BSAe	1	62933	62933	1413.8	< 2.2e-16	***
age	1	441375	441375	9915.7	< 2.2e-16	***
Residuals	18916	842000	45			
Signif. codes	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

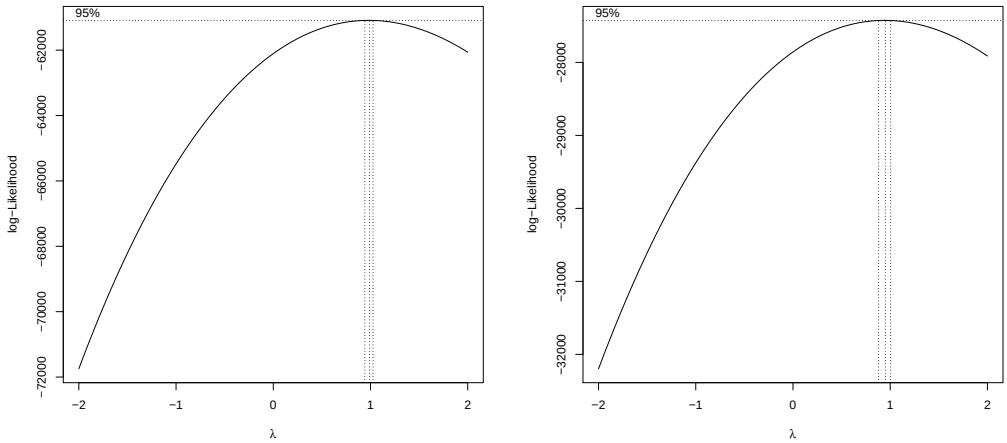
Table 6.7: Summary and ANOVA of Model 2 for males

$WL^{0.7} \sim BSA + age$						
Coefficients:						
	Estimate	Std. Error	t value	Pr(>|t|)		
(Intercept)	34.562048	0.576806	59.92	<2e-16	***	
BSA	5.102416	0.308410	16.54	<2e-16	***	
age	-0.265126	0.003363	-78.83	<2e-16	***	
RSE	4.636 on 9703 degrees of freedom					
R^2	0.4026					
R^2_{adj}	0.4025					
F-statistic	3270 on 2 and 9703 DF					
p-value:	< 2.2e-16					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
BSA	1	6998	6998	325.6	< 2.2e-16	***
age	1	133566	133566	6214.7	< 2.2e-16	***
Residuals	9703	208535	21			
Signif. codes	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 6.8: Summary and ANOVA of Model 2 for females

6.5.3 Diagnostics of the model

The Box-Cox plots in Figure 6.10 show that the already performed transformation of the maximum workload was necessary.



(a) Male (b) Female

Figure 6.10: Box-Cox plots of Model 2

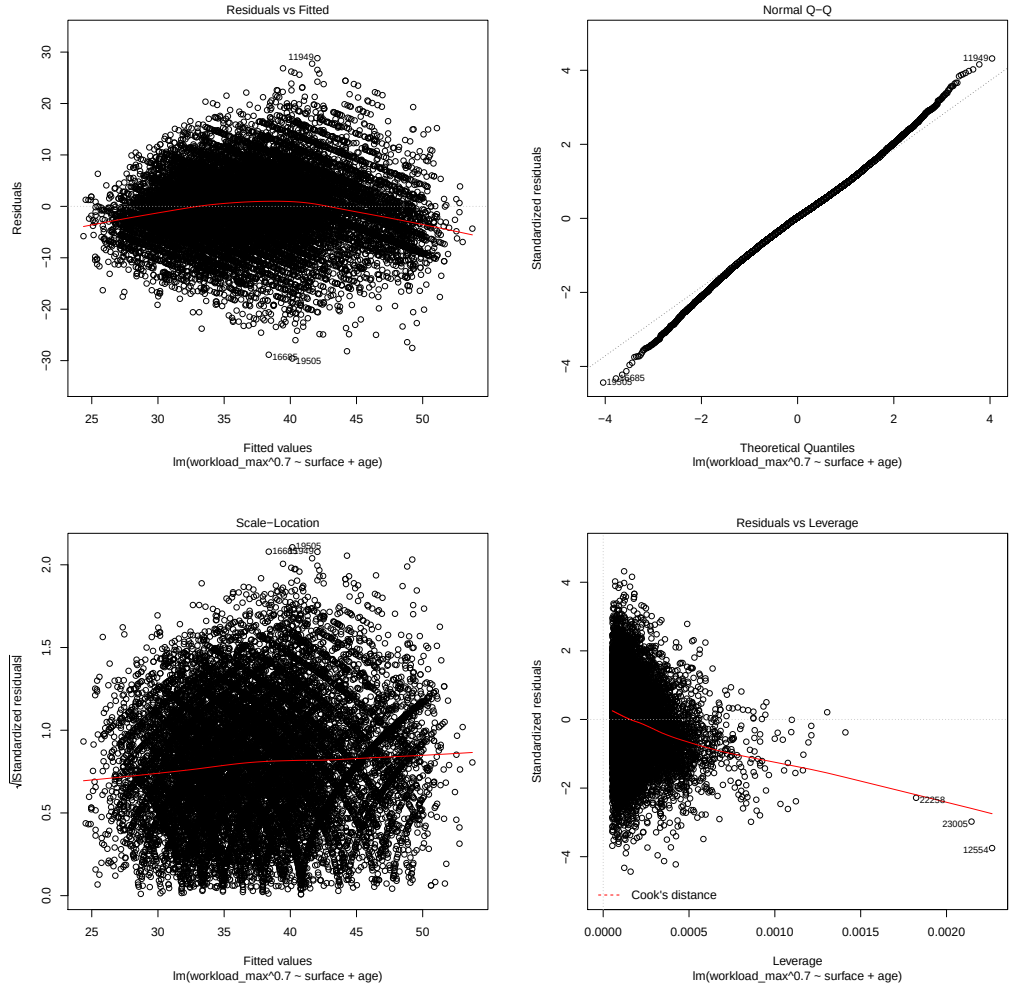


Figure 6.11: Diagnostic plots of Model 2 for males

Figure 6.11 and Figure 6.12 analyze Model 2 to check the requirements of the regression model for both male and female.

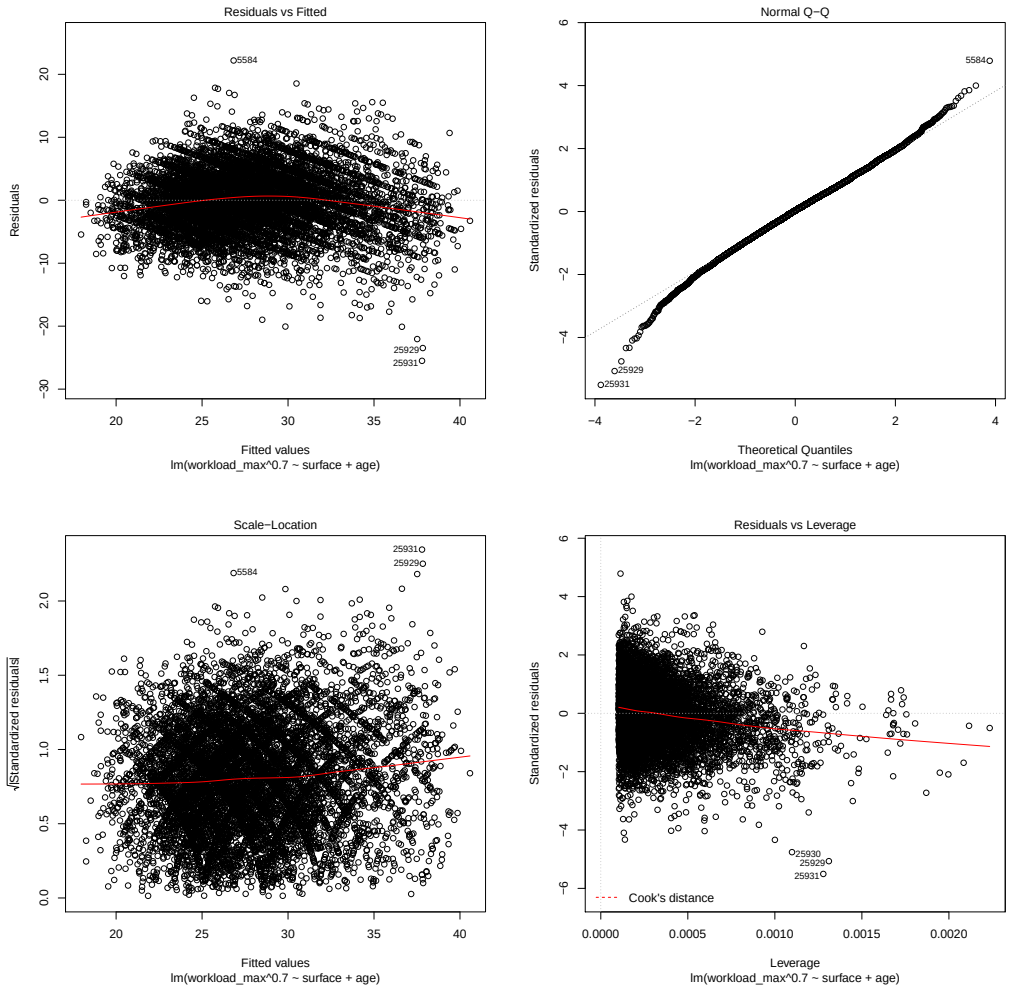


Figure 6.12: Diagnostic plots of Model 2 for females

6.5.4 Confidence intervals and prediction intervals for Model 2

In this subsection the confidence intervals and prediction intervals are created in Table 6.9 and in Table 6.10. In this table, for the age and BSA, the same values as for Model 1 are chosen. Note that the values of this tables are transformed back to the maximum workload itself.

Men			
age/BSA	1.5	2	2.5
25	225.2 [222.35,228.06] [136.66,325.82]	250.07 [248.38,251.76] [158.23,353.48]	275.71 [273.05,278.37] [180.69,381.85]
50	169.33 [167.21,171.45] [89.46,262.78]	192.24 [191.51,192.97] [108.6,288.76]	216 [213.81,218.2] [128.78,315.52]
75	118.54 [116.66,120.42] [48.85,204]	139.23 [138.31,140.16] [65.05,228.15]	160.9 [158.65,163.15] [82.53,253.15]

Table 6.9: Prediction values for the maximum workload [W] (first row), confidence intervals (second row) and prediction intervals (third row) of Model 2 for males

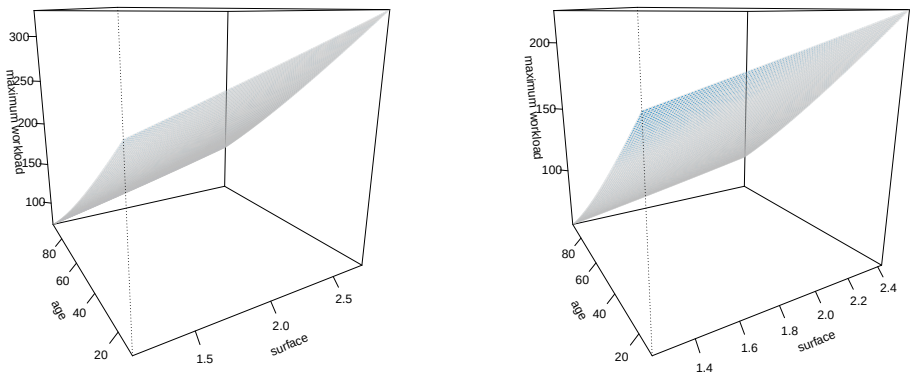
Women			
age/BSA	1.5	2	2.5
25	164.49 [162.62,166.37] [107.92,227.66]	181.59 [179.63,183.56] [123.07,246.46]	199.19 [195.58,202.82] [138.73,265.78]
50	122.54 [121.44,123.64] [71.54,180.98]	138.25 [137.04,139.45] [85.02,198.56]	154.51 [151.45,157.58] [99.11,216.7]
75	84.53 [83.46,85.61] [40.07,137.68]	98.66 [97.45,99.87] [51.53,153.93]	113.42 [110.57,116.29] [63.77,170.77]

Table 6.10: Prediction values for the maximum workload [W] (first row), confidence intervals (second row) and prediction intervals (third row) of Model 2 for females

6.5.5 Regression surface of Model 2

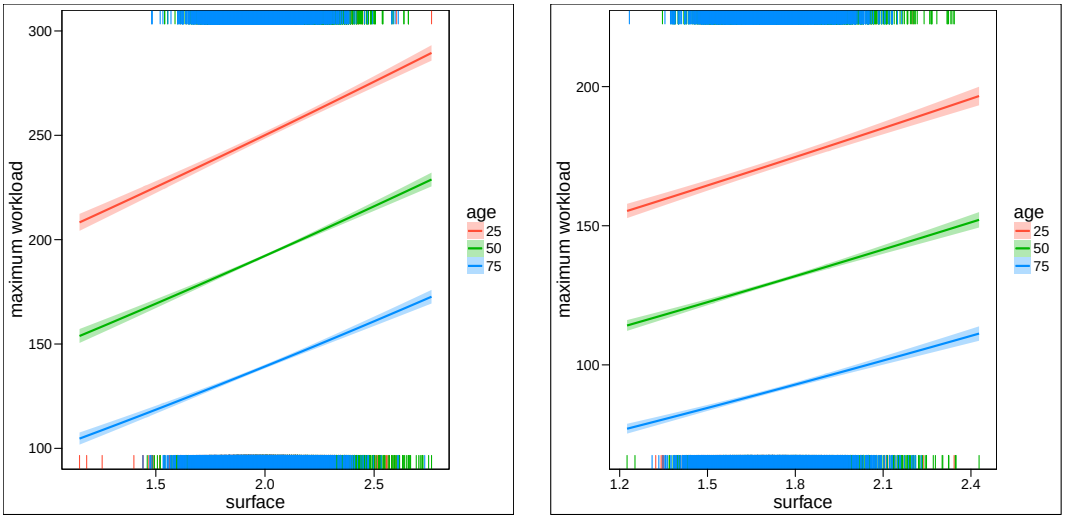
Figure 6.13 shows the regression surface of the regression for males and females.

In Figure 6.14 as well as in Figure 6.15 the age is fixed in three different stages from the tables above. In both figures, the slope of the regression lines is always the same, because there is no interaction. Additionally, the 95%-confidence bands are plotted.



(a) Male (b) Female

Figure 6.13: Plots of the regression surface for Model 2



(a) Male (b) Female

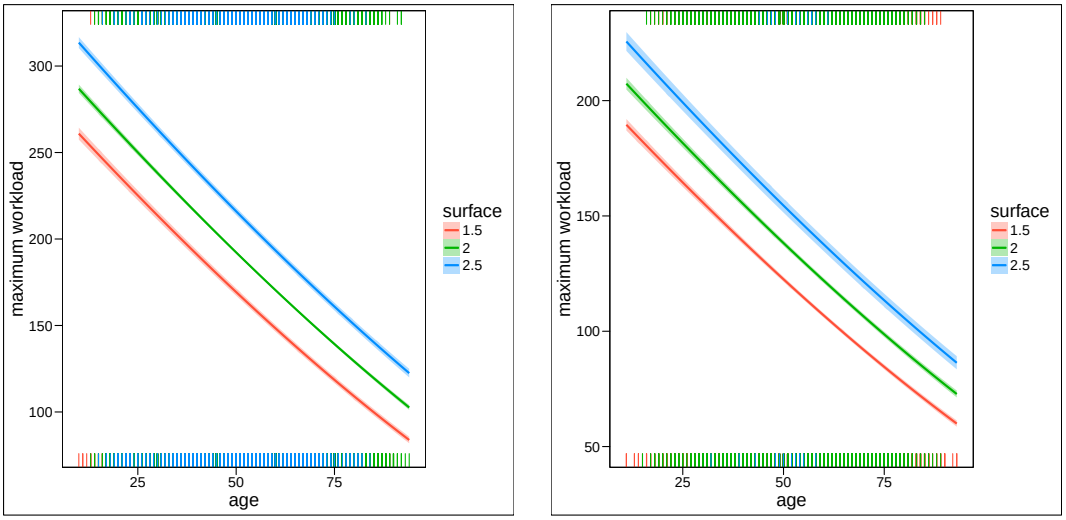
Figure 6.14: Plots of the regression lines and confidence bands of the surface for different ages of Model 2

6.6 Model 3: Linear regression model for the maximum workload to weight ratio depending on BSA and age and its interaction

In the next modification, the maximum-workload-to-weight-ratio is the depended variable instead of the maximum workload.

6.6.1 Variable selection

The best models with one to three variables for males and females are shown in Table 6.11.



(a) Male (b) Female

Figure 6.15: Plots of the regression lines and confidence bands of the age for different surfaces of Model 2

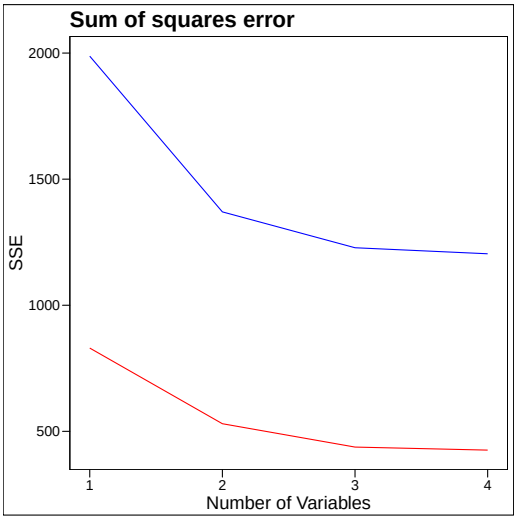
no.	BSA	age	BSA:age
1			*
2	*	*	
3	*	*	*

Table 6.11: Summary of the best models with 1 to 3 parameters for deriving Model 3

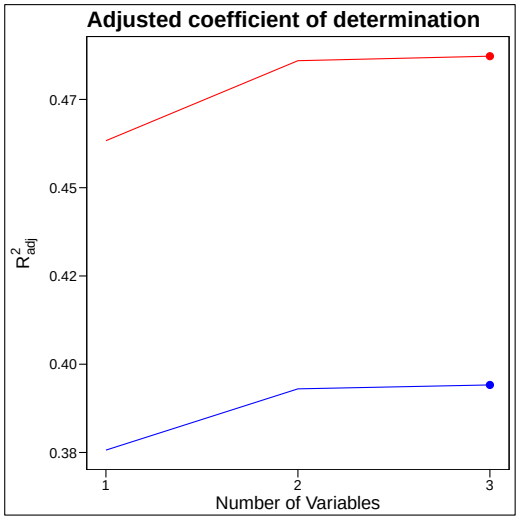
For choosing the best model out of these, Figure 6.16 shows plots for the information criteria.

As a result, 3 parameters can be chosen. Therefore, the following regression model will be used:

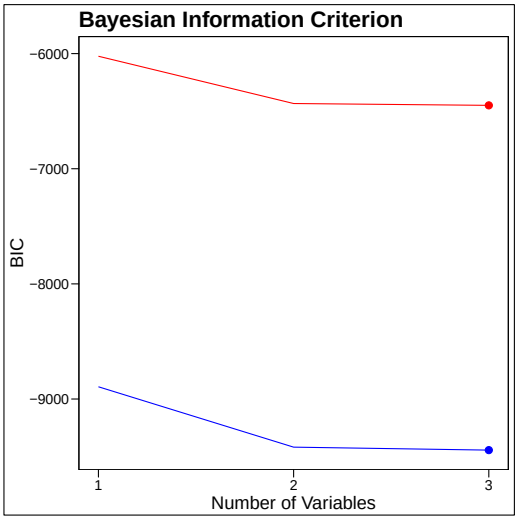
$$W\hat{W}R^{0.6}[W] = \hat{\beta}_0 + \hat{\beta}_1 BSA[m^2] + \hat{\beta}_2 age[y] + \hat{\beta}_3 BSA[m^2] age[y].$$



(a) Graphical description of the error sum of squares for the best models with 1 to 3 variables



(b) Graphical description of the R^2_{adj} for the best models with 1 to 3 variables



(c) Graphical description of the BIC for the best models with 1 to 3 variables

Figure 6.16: Information criteria for deriving Model 3

6.6.2 Analysis of the model

In Table 6.12 and Table 6.13 summary and ANOVA are presented for males and females.

$WWR^{0.6} \sim BSA * age$					
Coefficients:					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.9290491	0.0837277	46.927	< 2e-16	***
BSA	-0.8184991	0.0422183	-19.387	< 2e-16	***
age	-0.0213401	0.0014378	-14.842	< 2e-16	***
BSA:age	0.0043623	0.0007305	5.972	2.39e-09	***
RSE	0.2523 on 18915 degrees of freedom				
R^2	0.3942				
R^2_{adj}	0.3941				
F-statistic	4103 on 3 and 18915 DF				
p-value:	< 2.2e-16				
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
BSA	1	80.87	80.87	1270.325	< 2.2e-16	***
age	1	700.48	700.48	11003.647	< 2.2e-16	***
BSA:age	1	2.27	2.27	35.664	2.387e-09	***
Residuals	18915	1204.12	0.06			
Signif. codes	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

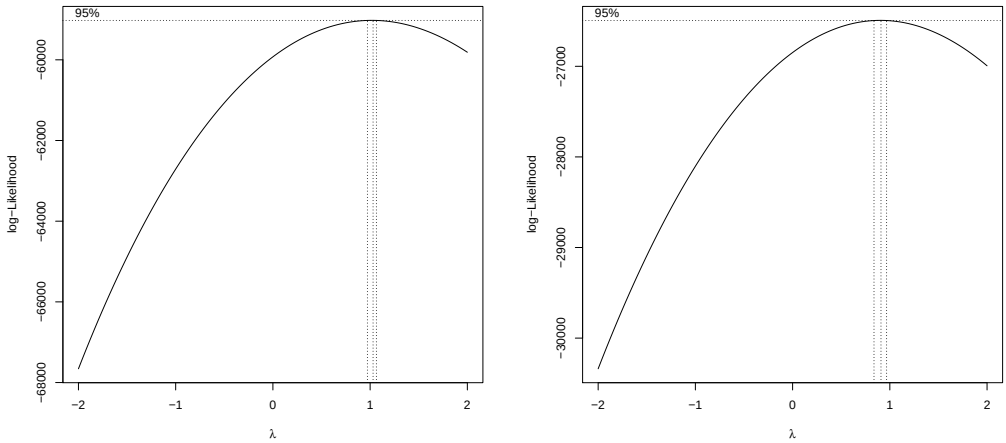
Table 6.12: Summary and ANOVA of Model 3 for males

$WWR^{0.6} \sim BSA * age$					
Coefficients:					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.786664	0.099688	37.98	< 2e-16	***
BSA	-0.958777	0.057894	-16.56	< 2e-16	***
age	-0.021056	0.001673	-12.59	< 2e-16	***
BSA:age	0.004878	0.000973	5.01	5.4e-07	***
RSE	0.21 on 9702 degrees of freedom				
R^2	0.487				
R^2_{adj}	0.487				
F-statistic	3.08e+03 on 3 and 9702 DF				
p-value:	< 2.2e-16				
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
BSA	1	97	96.9	2209.8	< 2e-16	***
age	1	307	306.6	6990.3	< 2e-16	***
BSA:age	1	1	1.1	25.1	5.4e-07	***
Residuals	9702	426	0.0			
Signif. codes	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 6.13: Summary and ANOVA of Model 3 for females

6.6.3 Diagnostics of the model

The Box-Cox plots in Figure 6.17 show that the already performed transformation of the maximum workload was necessary.



(a) Male (b) Female

Figure 6.17: Box-Cox plot of Model 3

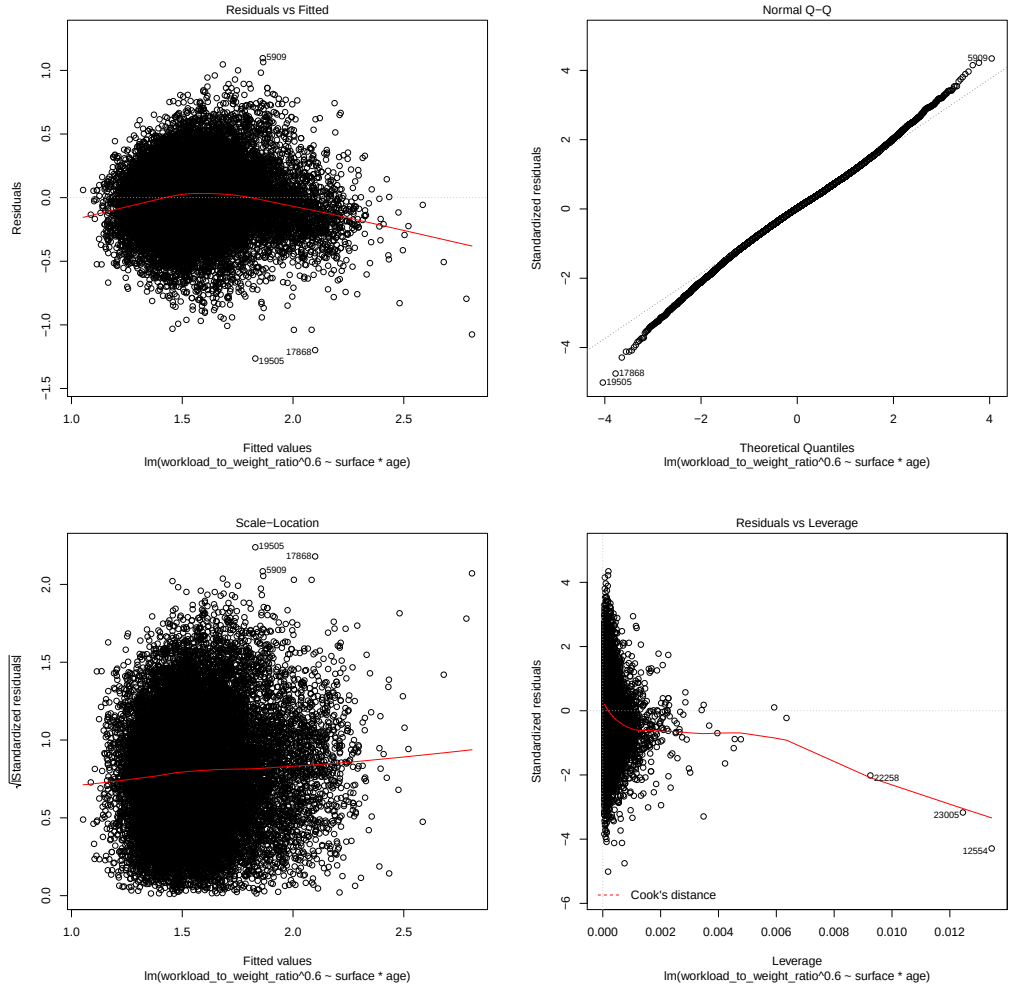


Figure 6.18: Diagnostic plots of Model 3 for males

Figure 6.18 and Figure 6.19 analyze Model 3 to check the requirements.

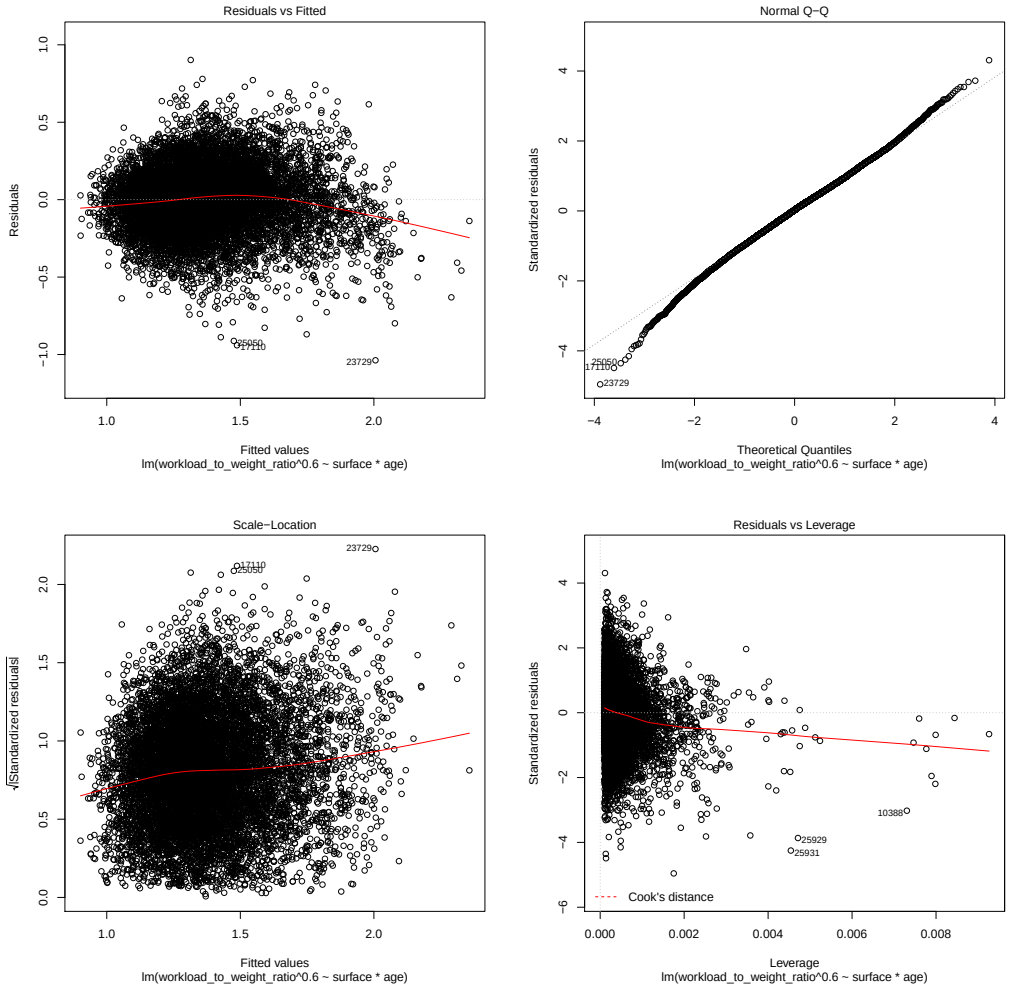


Figure 6.19: Diagnostic plots of Model 3 for females

6.6.4 Confidence intervals and prediction intervals for Model 3

In this subsection, the confidence intervals and prediction intervals are created in Table 6.14 and in Table 6.15. In these tables for the age and BSA, again the same values are chosen. Note that the values of this tables refer to the maximum workload to weight ratio instead of the maximum workload.

Men			
age/BSA	1.5	2	2.5
25	4.1 [4.02,4.17] [2.75,5.65]	3.11 [3.09,3.14] [1.93,4.52]	2.24 [2.18,2.3] [1.22,3.49]
50	3.07 [3.04,3.11] [1.89,4.47]	2.33 [2.32,2.34] [1.29,3.6]	1.67 [1.65,1.7] [0.79,2.8]
75	2.17 [2.13,2.21] [1.17,3.41]	1.64 [1.63,1.65] [0.76,2.76]	1.17 [1.14,1.21] [0.43,2.18]

Table 6.14: Prediction values for the maximum workload-to-weight-ratio [W/kg] (first row), confidence intervals (second row) and prediction intervals (third row) of Model 3 for males

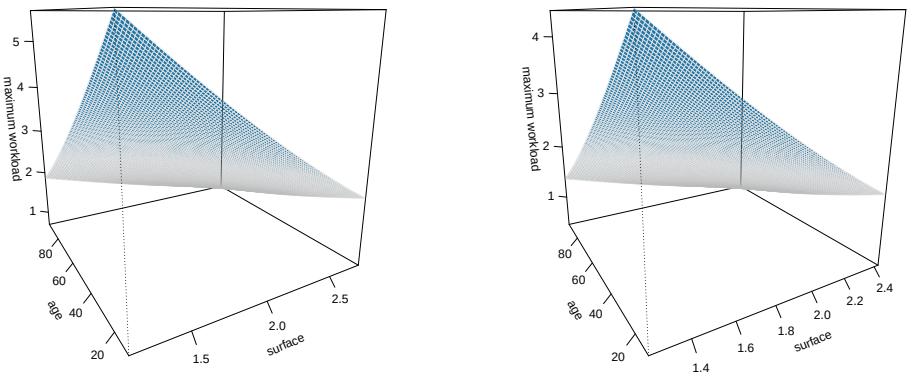
Women			
age/BSA	1.5	2	2.5
25	3.19 [3.14,3.24] [2.18,4.35]	2.16 [2.11,2.21] [1.31,3.17]	1.3 [1.2,1.4] [0.62,2.15]
50	2.33 [2.31,2.35] [1.45,3.37]	1.56 [1.54,1.58] [0.83,2.46]	0.91 [0.87,0.95] [0.35,1.67]
75	1.58 [1.56,1.61] [0.85,2.49]	1.04 [1.01,1.06] [0.44,1.82]	0.59 [0.54,0.63] [0.14,1.24]

Table 6.15: Prediction values for the maximum workload-to-weight-ratio [W/kg] (first row), confidence intervals (second row) and prediction intervals (third row) of Model 3 for females

6.6.5 Regression surface of Model 3

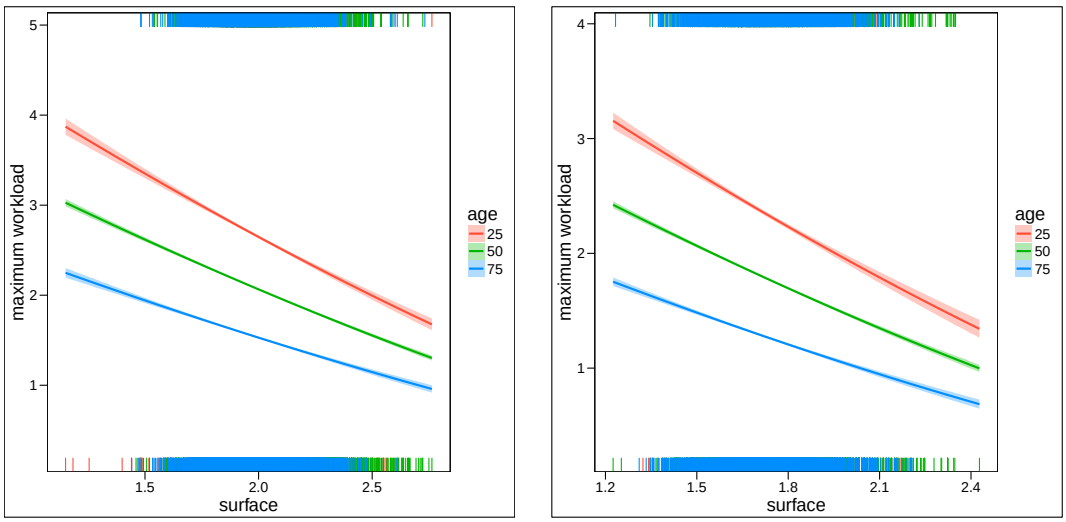
Figure 6.20 shows the regression surface of the regression for males and females.

In Figure 6.21 as well as in Figure 6.22, the age is fixed in three different stages from the tables above. In both figures, the slope is changing because of the interaction. Additionally, the 95%-confidence bands are plotted.



(a) Male (b) Female

Figure 6.20: Plots of the regression surface for Model 3

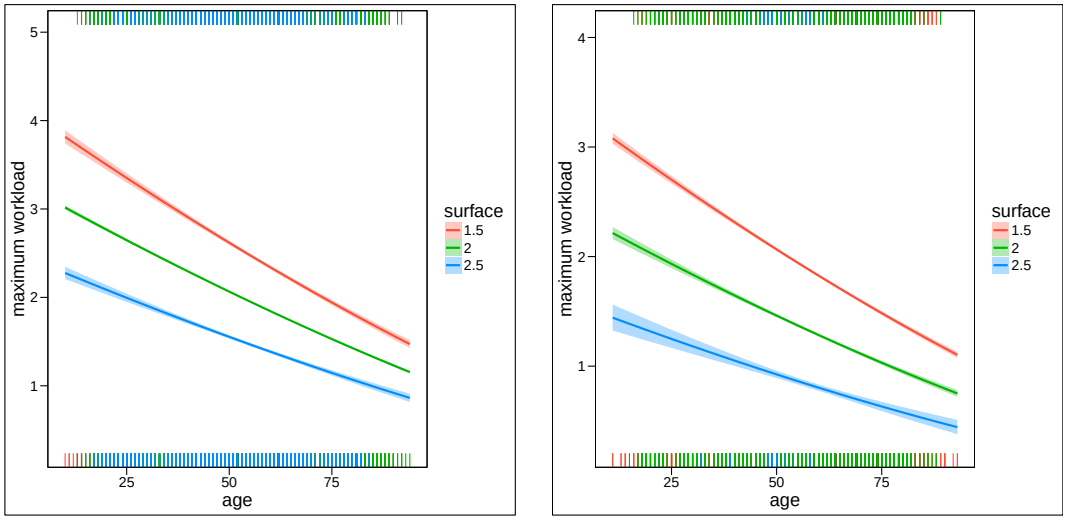


(a) Male (b) Female

Figure 6.21: Plots of the regression lines and confidence bands of the surface for different ages for Model 3

6.7 Model 4: Linear regression model for the maximum workload to weight ratio depending on height, weight and age

To reduce the number of variables a variable selection is made. Therefore, all possible subsets of the pool of explanatory variables are calculated, and the model that best fits the data according to the BIC is developed.



(a) Male (b) Female

Figure 6.22: Plots of the regression lines and confidence bands of the age for different surfaces for Model 3

6.7.1 Variable selection

Table 6.16 and Table 6.17 show the best model for different numbers of parameters for males and females, respectively. For example, in the best model for males with four parameters the height, age and the interaction between height and weight and between height and age are chosen. Whereas, in the best model for females with four parameters the weight, age and the interaction between weight and age and between height, weight and age are chosen.

no.	height	weight	age	height:weight	age:height	weight:age	height:weight:age
1						*	
2						*	*
3		*	*		*		
4	*		*	*	*		
5	*		*	*	*		*
6	*	*	*	*		*	*
7	*	*	*	*	*	*	*

Table 6.16: Summary of the best models with 1 to 7 parameters for deriving Model 4 for males

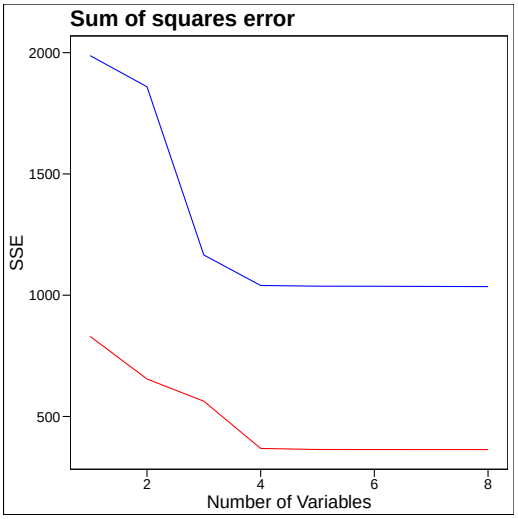
In the next step, the best model among these models in Table 6.16 is chosen. As in the model above, the BIC is the chosen criterion. Figure 6.23 shows plots for the information criteria.

For convenience, for both males and females, three parameters are chosen. This has the advantage that both regressions have the same parameters

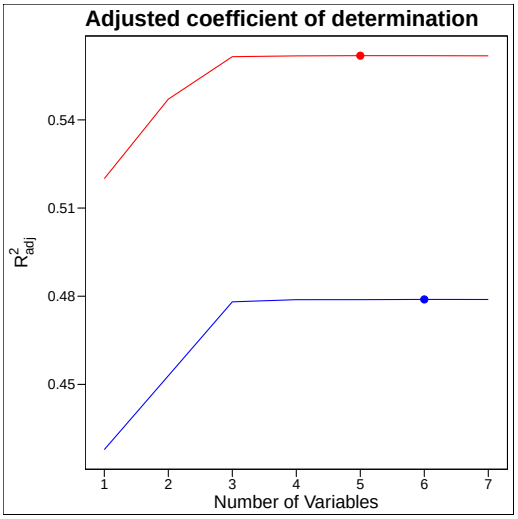
$$\hat{WWR}^{0.6}[W] = \hat{\beta}_0 + \hat{\beta}_1 weight[kg] + \hat{\beta}_2 age[y] + \hat{\beta}_3 height[cm] age[y].$$

no.	height	weight	age	height:weight	age:height	weight:age	height:weight:age
1						*	
2						*	*
3		*	*		*		
4		*	*			*	*
5		*	*		*	*	*
6	*	*	*	*	*		*
7	*	*	*	*	*	*	*

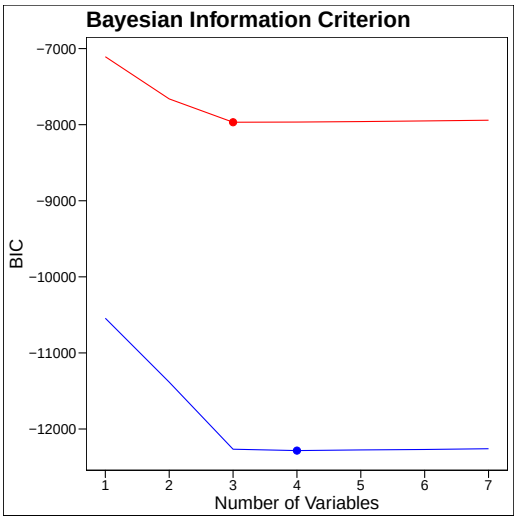
Table 6.17: Summary of the best models with 1 to 7 parameters for deriving Model 4 for females



(a) Graphical description of the error sum of squares for the best models with 1 to 7 variables



(b) Graphical description of the R^2_{adj} for the best models with 1 to 7 variables



(c) Graphical description of the BIC for the best models with 1 to 7 variables

Figure 6.23: Information criteria for deriving Model 4

6.7.2 Analysis of the model

In Table 6.18 and in Table 6.19 a summary and ANOVA are presented.

$WWR^{0.6} \sim weight + age + age : height$						
Coefficients:						
	Estimate	Std. Error	t value	Pr(>|t|)		
(Intercept)	3.160e+00	1.362e-02	232.07	<2e-16	***	
weight	-1.140e-02	1.463e-04	-77.91	<2e-16	***	
age	-4.413e-02	8.675e-04	-50.87	<2e-16	***	
age:height	1.875e-04	5.067e-06	37.01	<2e-16	***	
RSE	0.2342 on 18915 degrees of freedom					
R^2	0.4782					
R^2_{adj}	0.4781					
F-statistic	5777 on 3 and 18915 DF					
p-value:	< 2.2e-16					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
weight	1	214.85	214.85	3918.0	< 2.2e-16	***
age	1	660.53	660.53	12045.1	< 2.2e-16	***
age:height	1	75.09	75.09	1369.4	< 2.2e-16	***
Residuals	18915	1037.26	0.05			
Signif. codes	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 6.18: Summary and ANOVA of Model 4 for males

$WWR^{0.6} \sim weight + age + age : height$						
Coefficients:						
	Estimate	Std. Error	t value	Pr(>|t|)		
(Intercept)	2.775e+00	1.330e-02	208.68	<2e-16	***	
weight	-1.089e-02	1.639e-04	-66.45	<2e-16	***	
age	-3.374e-02	9.083e-04	-37.15	<2e-16	***	
age:height	1.394e-04	5.732e-06	24.32	<2e-16	***	
RSE	0.1937 on 9702 degrees of freedom					
R^2	0.5616					
R^2_{adj}	0.5615					
F-statistic	4143 on 3 and 9702 DF					
p-value:	< 2.2e-16					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
weight	1	175.68	175.680	4683.15	< 2.2e-16	***
age	1	268.44	268.435	7155.75	< 2.2e-16	***
age:height	1	22.19	22.187	591.46	< 2.2e-16	***
Residuals	9702	363.95	0.038			
Signif. codes	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 6.19: Summary and ANOVA of Model 4 for females

6.7.3 Diagnostics of the model

The Box-Cox plots in Figure 6.24 show that the already performed transformation of the maximum workload was necessary.

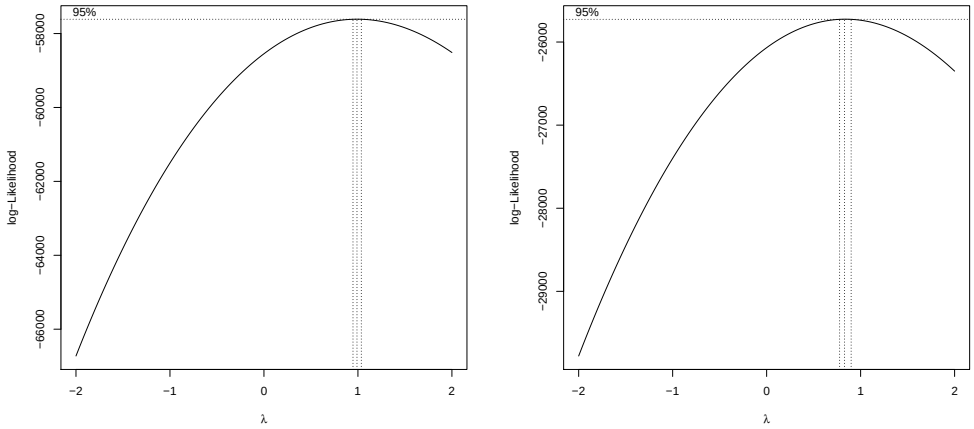


Figure 6.24: Box-Cox plots of Model 4

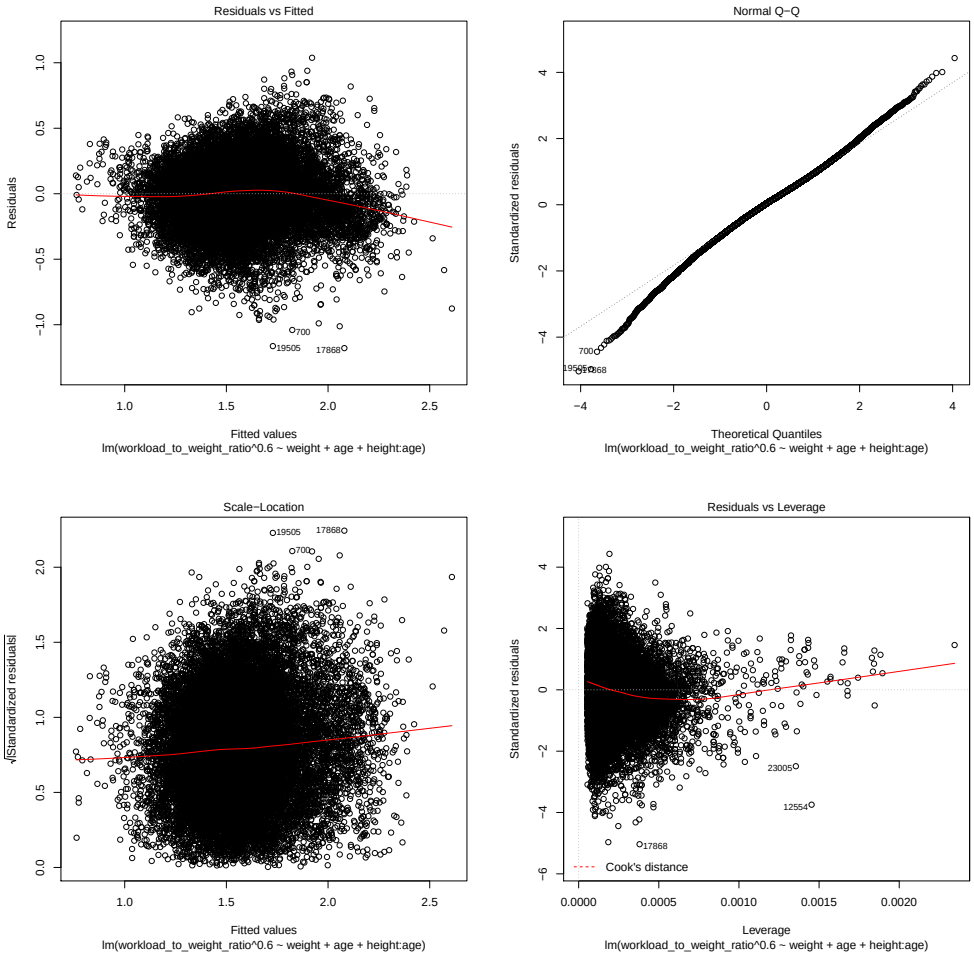


Figure 6.25: Diagnostic plots of Model 4 for males

Figure 6.25 and Figure 6.26 analyze Model 4 to check the requirements of the regression model for males and females, respectively.

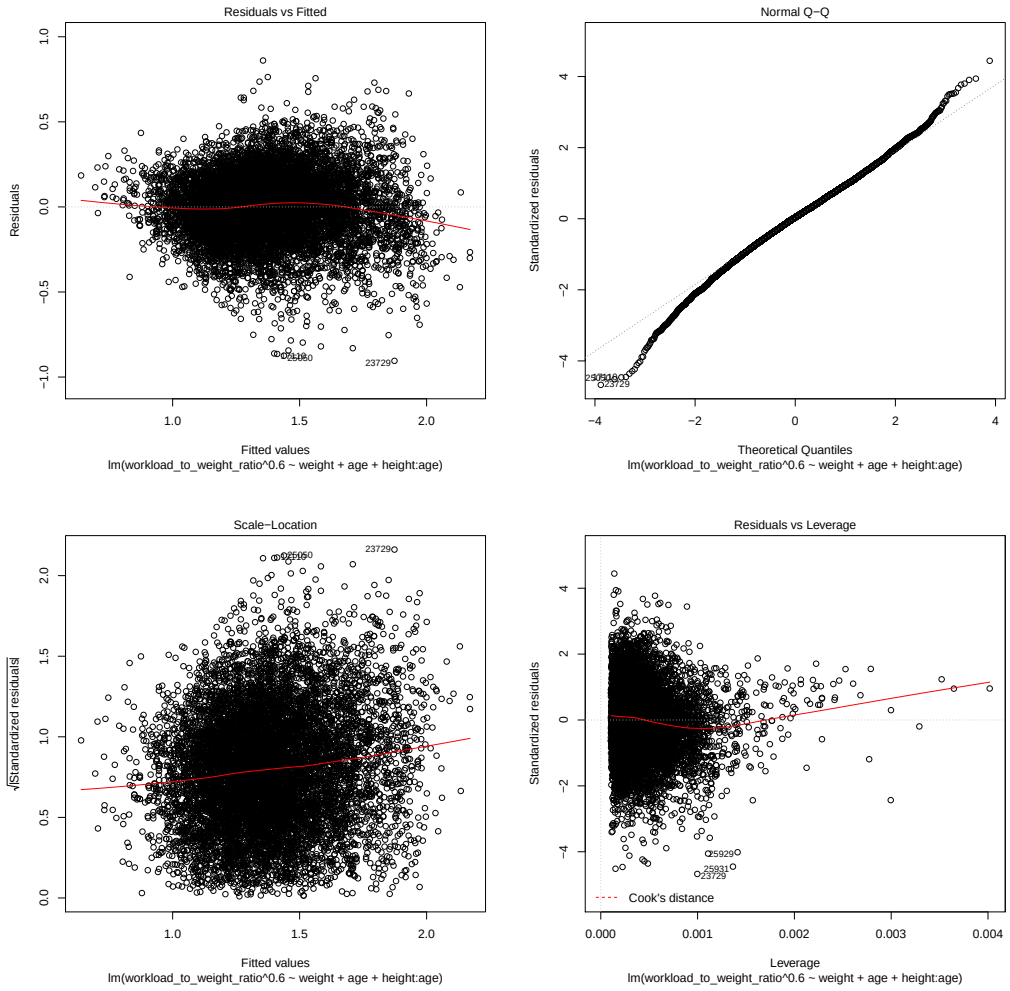


Figure 6.26: Diagnostic plots of Model 4 for females

6.7.4 Confidence intervals and prediction intervals for Model 4

In this subsection, the confidence intervals and prediction intervals are created in Table 6.20 for men and in Table 6.21 for women. In this tables, the variables age, weight and height are necessary.

Men, weight=70			
age/height	170	180	190
25	3.32 [3.3,3.35] [2.18,4.65]	3.45 [3.43,3.48] [2.29,4.8]	3.58 [3.55,3.61] [2.4,4.94]
50	2.54 [2.53,2.55] [1.53,3.75]	2.77 [2.76,2.79] [1.72,4.01]	3.01 [2.99,3.04] [1.92,4.29]
75	1.84 [1.83,1.86] [0.97,2.92]	2.15 [2.13,2.17] [1.22,3.29]	2.48 [2.45,2.51] [1.48,3.68]
Men, weight=90			
age/height	170	180	190
25	2.73 [2.71,2.76] [1.69,3.97]	2.85 [2.83,2.87] [1.79,4.11]	2.97 [2.95,2.99] [1.88,4.25]
50	2.01 [2,2.03] [1.11,3.12]	2.22 [2.21,2.23] [1.27,3.37]	2.44 [2.43,2.46] [1.45,3.63]
75	1.38 [1.37,1.4] [0.63,2.36]	1.66 [1.65,1.67] [0.83,2.7]	1.96 [1.93,1.98] [1.06,3.06]
Men, weight=110			
age/height	170	180	190
25	2.19 [2.16,2.22] [1.25,3.33]	2.3 [2.27,2.32] [1.33,3.46]	2.41 [2.38,2.43] [1.42,3.59]
50	1.54 [1.52,1.56] [0.74,2.55]	1.73 [1.71,1.74] [0.88,2.78]	1.92 [1.91,1.94] [1.04,3.02]
75	0.98 [0.96,1] [0.35,1.85]	1.22 [1.21,1.24] [0.51,2.16]	1.49 [1.46,1.51] [0.7,2.49]

Table 6.20: Prediction values for the maximum workload-to-weight-ratio [W/kg] (first row), confidence intervals (second row) and prediction intervals (third row) of Model 4 for males

6.7.5 Regression surface of Model 4

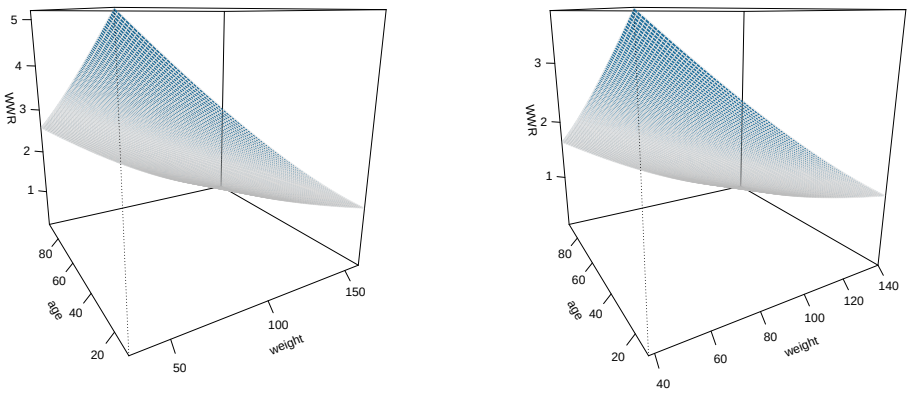
Figure 6.27 shows the regression surface of the regression for males and females.

Figure 6.28 and Figure 6.29 visualize the regression function. In both plots, the age is fixed in three different stages from the tables above. Figure 6.28 shows that there is no interaction between weight and height. In contrary to that, the interaction between age and height can be observed in Figure 6.29. In both plots, the parameter which is not used is set to the median value and additionally, the

Women, weight=50			
age/height	150	165	180
25	2.94 [2.91,2.97] [2.03,3.98]	3.07 [3.05,3.1] [2.15,4.13]	3.21 [3.18,3.24] [2.27,4.28]
50	2.16 [2.14,2.19] [1.37,3.09]	2.41 [2.39,2.42] [1.58,3.37]	2.66 [2.63,2.69] [1.79,3.66]
75	1.49 [1.46,1.51] [0.82,2.3]	1.8 [1.78,1.83] [1.08,2.68]	2.15 [2.11,2.19] [1.36,3.08]
Women, weight=70			
age/height	150	165	180
25	2.4 [2.37,2.43] [1.57,3.37]	2.53 [2.5,2.55] [1.68,3.51]	2.65 [2.63,2.68] [1.79,3.65]
50	1.69 [1.67,1.71] [0.99,2.54]	1.91 [1.9,1.92] [1.17,2.8]	2.14 [2.12,2.17] [1.36,3.07]
75	1.09 [1.07,1.11] [0.51,1.82]	1.37 [1.36,1.38] [0.73,2.16]	1.68 [1.65,1.71] [0.97,2.53]
Women, weight=90			
age/height	150	165	180
25	1.91 [1.88,1.94] [1.16,2.8]	2.02 [1.99,2.05] [1.26,2.93]	2.14 [2.11,2.17] [1.35,3.07]
50	1.27 [1.24,1.29] [0.65,2.04]	1.47 [1.45,1.48] [0.81,2.28]	1.67 [1.65,1.7] [0.97,2.52]
75	0.74 [0.72,0.76] [0.27,1.38]	0.98 [0.97,1] [0.44,1.69]	1.26 [1.23,1.28] [0.64,2.02]

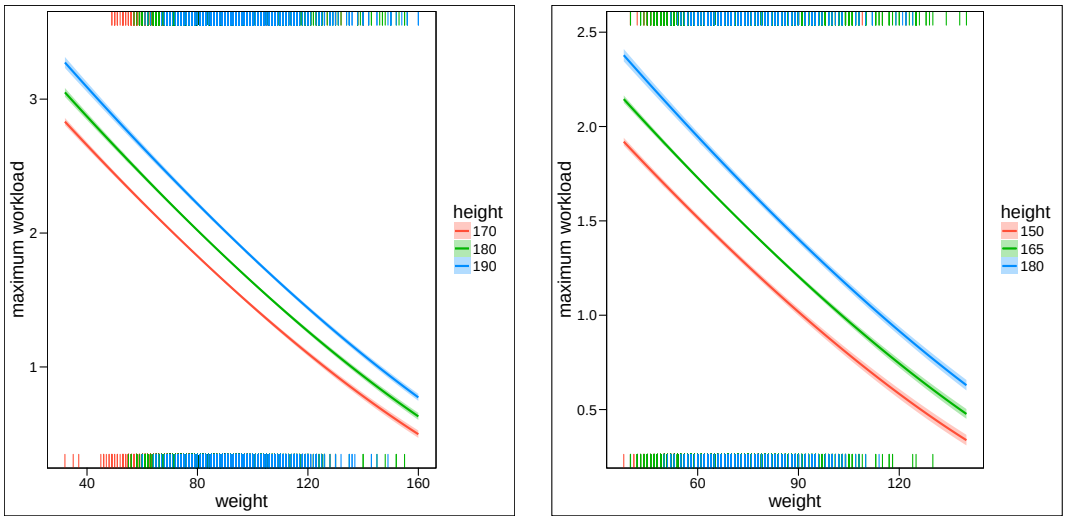
Table 6.21: Prediction values for the maximum workload-to-weight-ratio [W/kg] (first row), confidence intervals (second row) and prediction intervals (third row) of Model 4 for females

95%-confidence bands are plotted.



(a) Men (b) Women

Figure 6.27: Plots of the regression surface for Model 4



(a) Male (b) Female

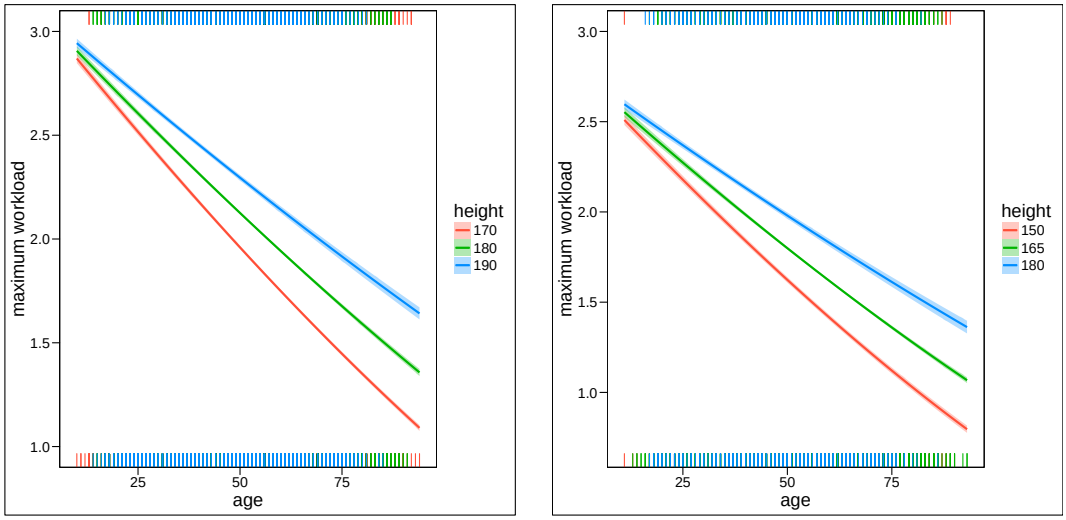
Figure 6.28: Plots of the regression lines and confidence bands of the weight for different ages of Model 4

6.8 Comparison of the models

In this section the models are compared.

6.8.1 Summary of the models

Model 1: $\hat{WL}[W] = \hat{\beta}_0 + \hat{\beta}_1 BSA[m^2] + \hat{\beta}_2 BSA[m^2] age[y]$
Model 2: $\hat{WL}^{0.7}[W] = \hat{\beta}_0 + \hat{\beta}_1 BSA[m^2] + \hat{\beta}_2 age[y]$
Model 3: $\hat{WWR}^{0.6}[W/kg] = \hat{\beta}_0 + \hat{\beta}_1 BSA[m^2] + \hat{\beta}_2 age[y] + \hat{\beta}_3 BSA[m^2] age[y]$
Model 4: $\hat{WWR}^{0.6}[W/kg] = \hat{\beta}_0 + \hat{\beta}_1 weight[kg] + \hat{\beta}_2 age[y] + \hat{\beta}_3 height[cm] age[kg]$



(a) Male (b) Female

Figure 6.29: Plots of the regression lines and confidence bands of the age for different surfaces of Model 4

6.8.2 Comparison of the sum of squares and the adjusted coefficient of determination

To compare these models the coefficient of determination R^2 and the correlation are compared for males in Table 6.22 and for females in Table 6.23.

response variable	W		$W^{0.7}$	$WWR^{0.6}$	
	Model P	Model 1	Model 2	Model 3	Model 4
SSR	20430888	22831293	504308	783.6	950.5
SSE	39733086	38133065	842000	1204.12	1037.26
SST	60964358	60964358	1346308	1988	1988
R^2_{adj}	0.3483	0.3744	0.3745	0.3941	0.4781
corr	0.9740719				

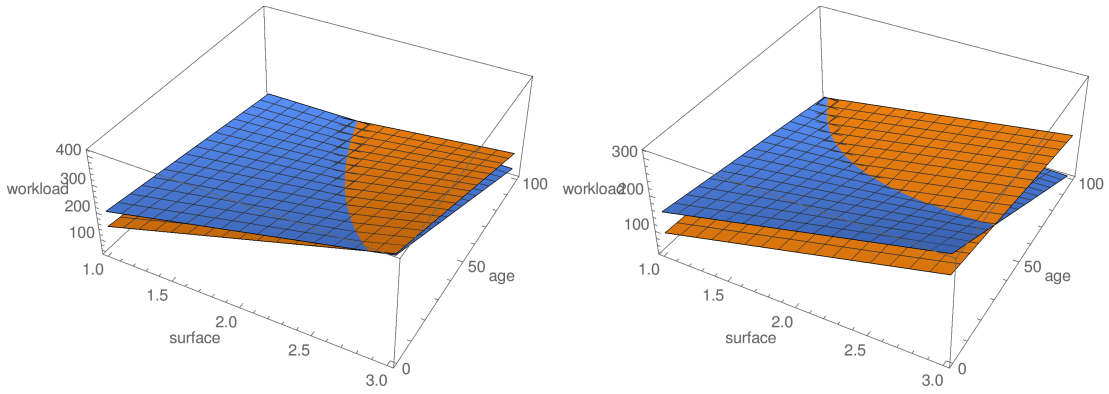
Table 6.22: Comparison of the Models for males

response variable	W		$W^{0.7}$	$WWR^{0.6}$	
	Model P	Model 1	Model 2	Model 3	Model 4
SSR	1745237	5009113	140564	405	466.3
SSE	9513851	7338625	208535	426	363.95
SST	12347738	12347738	349099	831	830.2
R^2_{adj}	0.2295	0.4055	0.4025	0.4870	0.5615
corr	0.7827556				

Table 6.23: Comparison of the Models for females

6.8.3 Comparison of the surfaces

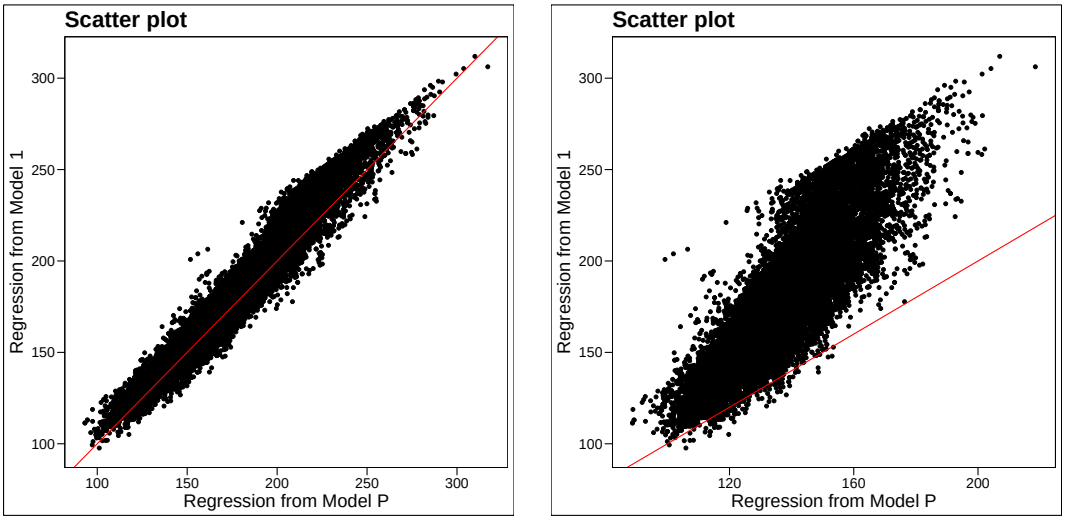
The three-dimensional regression surfaces are compared to show the difference between Model 1 and Model P. In both models, the same parameters are taken. Unfortunately, for the model from [Wonisch et al., 2008] no additional information, e.g. the distribution of age, is available. Figure 6.30 shows the difference between the surfaces of Model P and Model 1 for both males and females.



(a) Male (b) Female

Figure 6.30: Surface of Model P and Model 1: orange surface belongs to Model P and the blue surface belongs to Model 1

Figure 6.31 shows scatter plots between Model P and Model 1. In Figure 6.31(a), which shows the situation for males, one can see that the points are close to the red line. In Figure 6.31(b), which shows the situation for females, one can see that the point cloud differs from the red line. This effect can be also seen in the discrepancy of the correlation.



(a) Male (b) Female

Figure 6.31: Scatter plot of the maximum workloads for Model P and Model 1

6.9 Model D: Dummy linear regression model for the maximum workload to weight ratio depending on the height, weight and age and the dummy variables gender, diagnosis and medication

In addition to the maximum workload to weight ratio, the diagnosis and medication are included. To reduce the number of variables a variable selection is done. Therefore, all possible subsets of the pool of explanatory variables are calculated and the model that best fits the data according to the BIC is developed.

6.9.1 Regression tree

Figure 6.32 shows a regression tree for the variable WWR. The independent variables are BSA, age, height, weight, gender, diagnosis and medication.

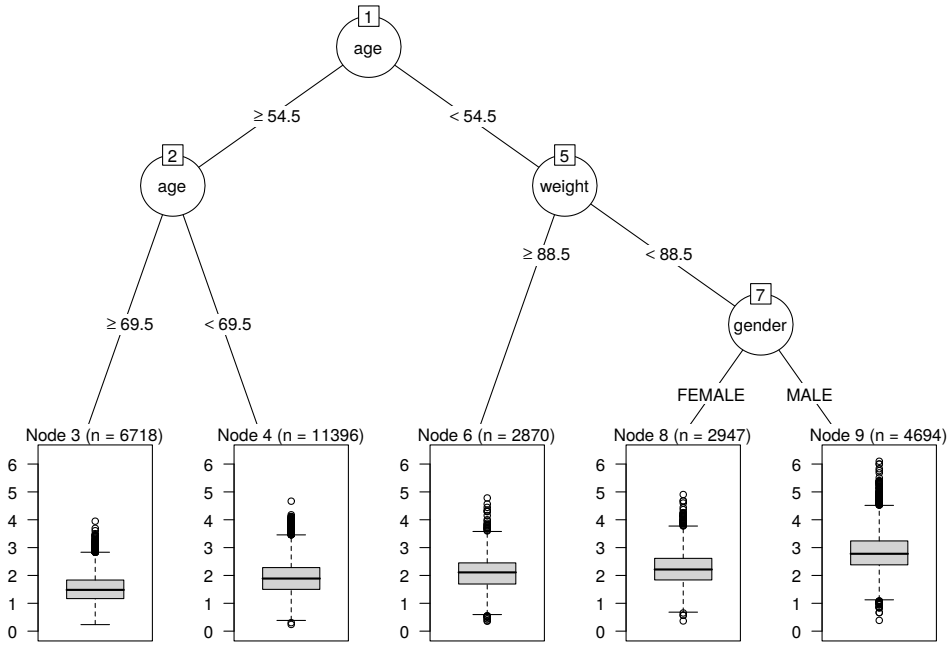


Figure 6.32: Regression tree for the variables of Model D

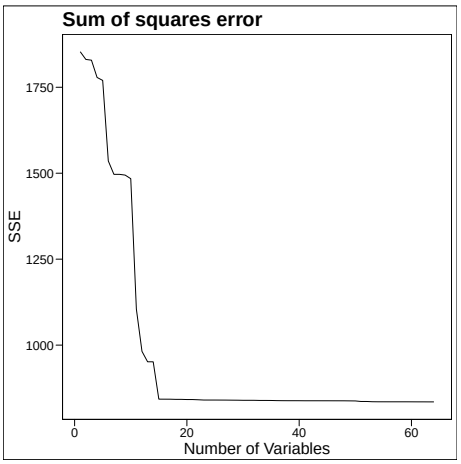
6.9.2 Variable selection

Table 6.24 shows the best model for different numbers of parameters. For example in the best model with four parameters height, weight, age and the interaction between height and gender is chosen.

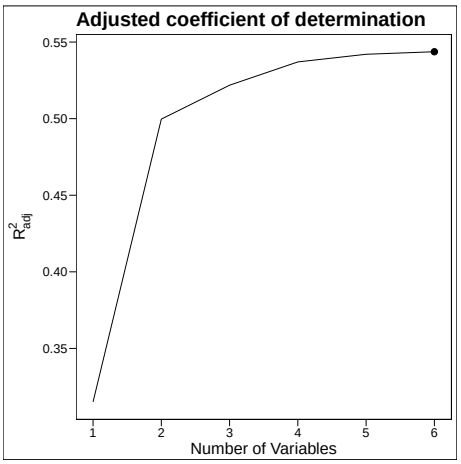
In a next step, the best model among models in Table 6.24 is chosen. As in the above model the BIC is the chosen criterion. Figure 6.33 shows plots for the information criteria.

no.	height	weight	age	weight:age	height:age	height:MALE	age:MALE	height:MED	MALE:MED	height:age:MALE
1			*							
2				*		*				
3	*			*						*
4	*	*	*			*				
5		*	*		*	*			*	
6		*	*	*		*	*	*		

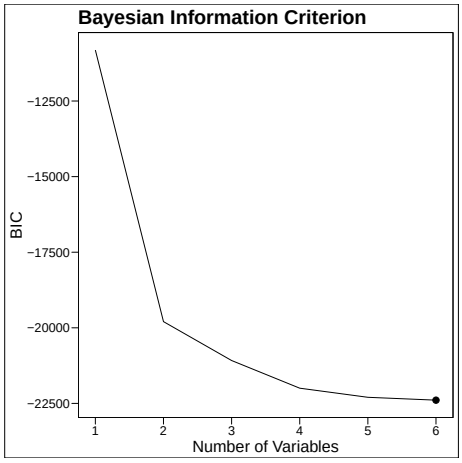
Table 6.24: Summary of the best models with one to six parameters for deriving Model D



(a) Graphical description of the error sum of squares for the best models with 1 to 64 variables



(b) Graphical description of the R^2_{adj} for the best models with 1 to 6 variables



(c) Graphical description of the BIC for the best models with 1 to 6 variables

Figure 6.33: Information criteria for deriving Model D

For Model D the best model with six parameters is chosen:

$$\begin{aligned} W\hat{W}R^{0.5}[W] = \quad & \hat{\beta}_0 + \hat{\beta}_1 weight[kg] + \hat{\beta}_2 age[y] + \hat{\beta}_3 height[cm]age[y] \\ & + \hat{\beta}_5 height[cm]MALE + \hat{\beta}_6 age[y]MALE + \hat{\beta}_4 height[cm]MED. \end{aligned}$$

6.9.3 Analysis of the model

In Table 6.25 a summary and ANOVA is presented.

$WWR^{0.6} \sim weight + age + height : age$					
Coefficients:					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.777e+00	1.158e-02	239.82	<2e-16	***
weight	-1.100e-02	1.114e-04	-98.79	<2e-16	***
age	-3.463e-02	6.441e-04	-53.76	<2e-16	***
age:height	1.465e-04	3.846e-06	38.08	<2e-16	***
height:MALE	1.989e-03	6.320e-05	31.46	<2e-16	***
age:MALE	-2.053e-03	1.804e-04	-11.38	<2e-16	***
height:MED	-3.225e-04	1.843e-05	-17.50	<2e-16	***
RSE	0.2201 on 28618 degrees of freedom				
R^2	0.5455				
R^2_{adj}	0.5454				
F-statistic	5726 on 6 and 28618				
p-value:	< 2.2e-16				
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
weight	1	94.58	94.58	1952.41	< 2.2e-16	***
age	1	995.48	995.48	20550.35	< 2.2e-16	***
age:height	1	388.27	388.27	8015.30	< 2.2e-16	***
height:MALE	1	164.06	164.06	3386.81	< 2.2e-16	***
age:MALE	1	6.89	6.89	142.15	< 2.2e-16	***
height:MED	1	14.83	14.83	306.21	< 2.2e-16	***
Residuals	28618	1386.29	0.05			
Signif. codes	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 6.25: Summary and ANOVA of Model D

6.9.4 Diagnostics of the model

The Box-Cox plots in Figure 6.34 show that the already performed transformation of the maximum workload was necessary.

Figure 6.35 analyzes Model D to check the requirements of the regression model.

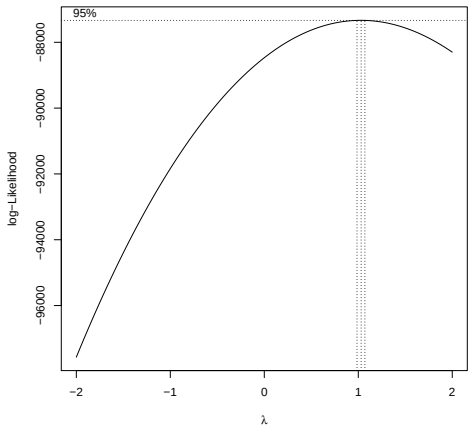


Figure 6.34: Box-Cox plots of Model D

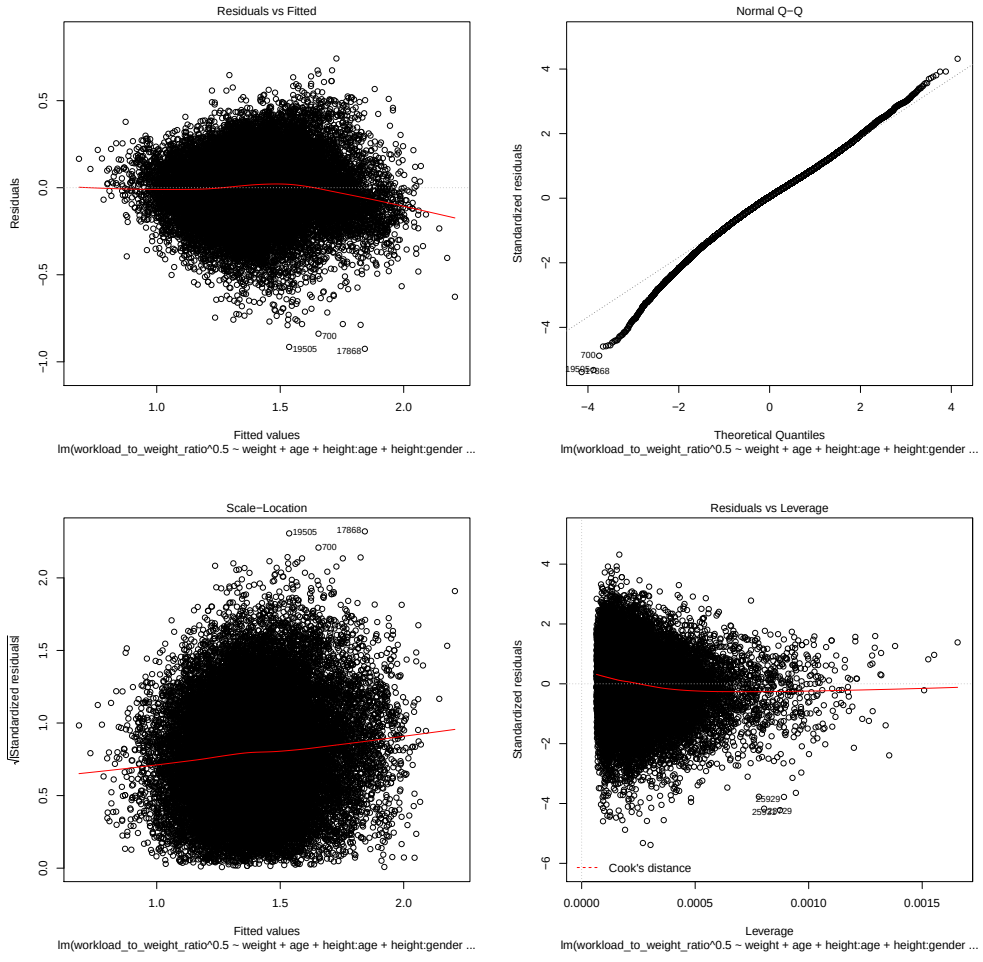


Figure 6.35: Diagnostic plots of Model D

6.10 Discussion

In summary, in this chapter, different models for obtaining normal values for the maximum workload and the maximum workload-to-weight-ratio are ana-

lyzed. In Austria, Model P is currently used to calculate the normal values of ergometries. Based on this model, new models are derived. New models show that an improvement of the currently used model is possible. In Model P, the maximum workload is calculated based on BSA and the interaction between age and BSA. When compared to the newly developed Model 1, it can be observed that Model P has a poor goodness of fit for females. This could be due to the different structure of underlying data while deriving Model P. Secondly, Model 1 shows that a transformation of the prediction variable is necessary. The next model, Model 2, already includes the transformation of the prediction variable and uses a maximum model which includes both main effects and the interaction. Comparison of Model 2 shows that this model has no particular effect on the adjusted coefficient of determination, though it produces correct prediction and confidence intervals. Use of the maximum workload-to-weigth-ratio as a prediction variable instead of the maximum workload results in Model 3. Comparison with prior models shows that an increase of the adjusted coefficient of determination is possible. Model 4 is using weight and height instead of BSA and yields the best results among all these models.

Finally, a dummy regression model is introduced at the end of this chapter. This model D uses the information about the diagnosis and medication (statin and beta blocker) along with age, gender, weight, and height. In contrary to the other models, Model D used just one regression for males and females.

Chapter 7

Normal values for the maximum workload of exercise testing during cardiac out-patient rehabilitation

This chapter deals with the results of ergometries during cardiac out-patient rehabilitation. The data of the cardiac out-patient rehabilitation consists of the ergometries during stage II and stage III (see Chapter 2). For stage II there exists one ergometry at the entry examination and one at the final examination. During stage III there exists an ergometry additionally at the intermediate examination.

One of the reasons to execute these ergometries is to discover the physical constitution of the patients. This can be measured inter alia by the maximum workload which is achieved at the ergometries. To evaluate these results normal values can be used. Generally speaking the goal of this chapter is to find normal values of the maximum workload for these ergometries. Therefore, the following research questions are answered in this chapter:

1. What are the normal values of the maximum workload at the entry examination of stage II in cardiac out-patient rehabilitation?

2. How much should be contributed in stage II of cardiac out-patient rehabilitation?

3. What are the normal values of the maximum workload at the entry examination of stage III in cardiac out-patient rehabilitation?

4. How much should be contributed in stage III of cardiac out-patient rehabilitation?

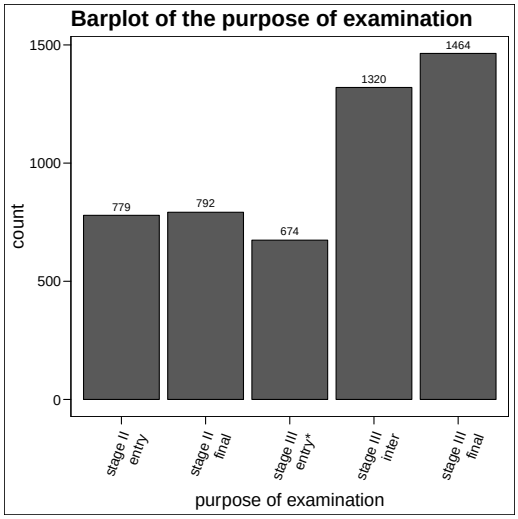
The remainder of this chapter is the following: Section 7.1 shows the composition of the data and an analysis of the maximum workload. The next four sections are answering the research questions. Section 7.2 deals with the normal values for the entry examination of stage II. Afterwards the contribution during stage II is covered in Section 7.3. Section 7.4 deals in the same manner with the normal values of stage III and Section 7.5 deals with the contribution during stage III. Finally Section 7.6 gives a short conclusion of the results.

7.1 Composition of the data

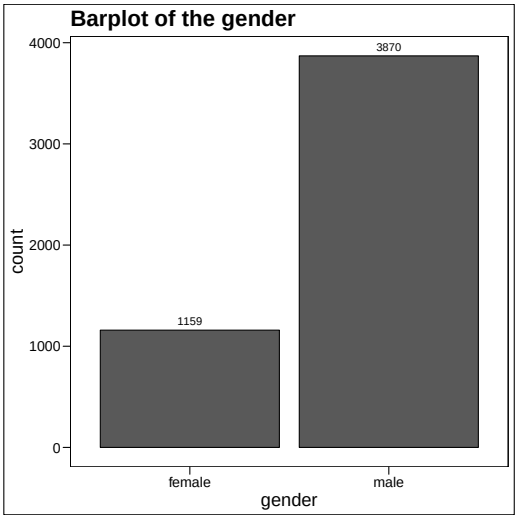
There is data of 5,029 ergometries during cardiac out-patient rehabilitation available in total. As a first step the composition of the data is analyzed in this section. For this purpose the data is split into four categories: reason for examination, gender, medication and diagnosis.

Starting with the purpose of examination, Figure 7.1(a) shows the number of ergometries during the different stages. The available data for stage II consist of 779 ergometries at the entry examination and 792 ergometries at the final examination. For stage III the data of 674 ergometries at the entry examination, 1,320 ergometries at the intermediate examination and 1,464 ergometries at the final examination are available. Note that there is no entry examination of stage III if stage III starts directly after stage II. In this case the final examination of stage II is the entry examination of stage III.

Figure 7.1(b) shows the split of ergometries by gender. In total the data consist of 3,870 ergometries done by men and 1,159 ergometries done by women. Concluding that the number of men in the cardiac out-patient rehabilitation is more than three times higher than the number of women.



(a) Number of ergometries during the different stages of cardiac out-patient rehabilitation (*Number without the occurences in which the final examination of stage II is also the entry examination of stage III)

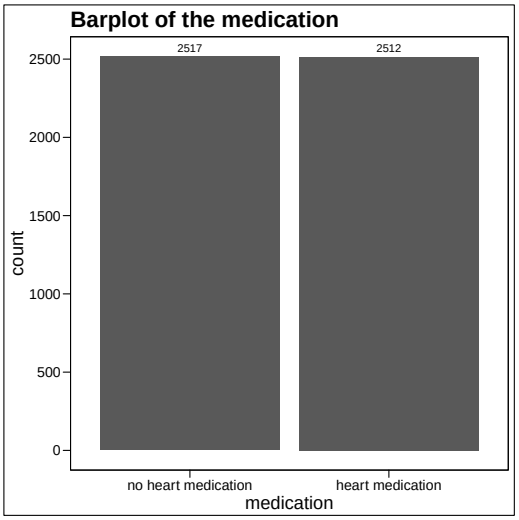


(b) Number of ergometries split by gender

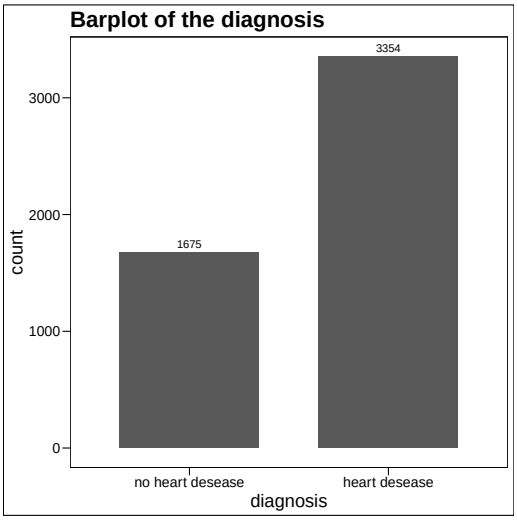
Figure 7.1: Number of ergometries during cardiac out-patient rehabilitation split by the reason for examination and gender

Another way of splitting the data is the medication and diagnosis. For the medication the drug groups statin and beta blocker are used and are noted in the following by heart medication. For more information about these medications see Section 2.5. Figure 7.2(a) shows split of ergometries in these medications. In the data 2,517 ergometries were done without statin and beta blocker and 2,512 ergometries were done under a heart medication. This means that around the half of the people did not take any heart medication.

Figure 7.2(b) shows the split of ergometries by diagnosis. The diagnosis is split into people which have a heart disease, e.g. infarct or coronal heart disease. For more information about the actual used diagnoses see Chapter 2. 4,062 ergometries were done with a heart disease and just 947 patients of cardiac out-patient rehabilitation does not have any heart disease. This was unexpected because typically just people with a heart disease make a cardiac rehabilitation.



(a) Number of ergometries separated by medication (nothing means neather statin nor betablocker)



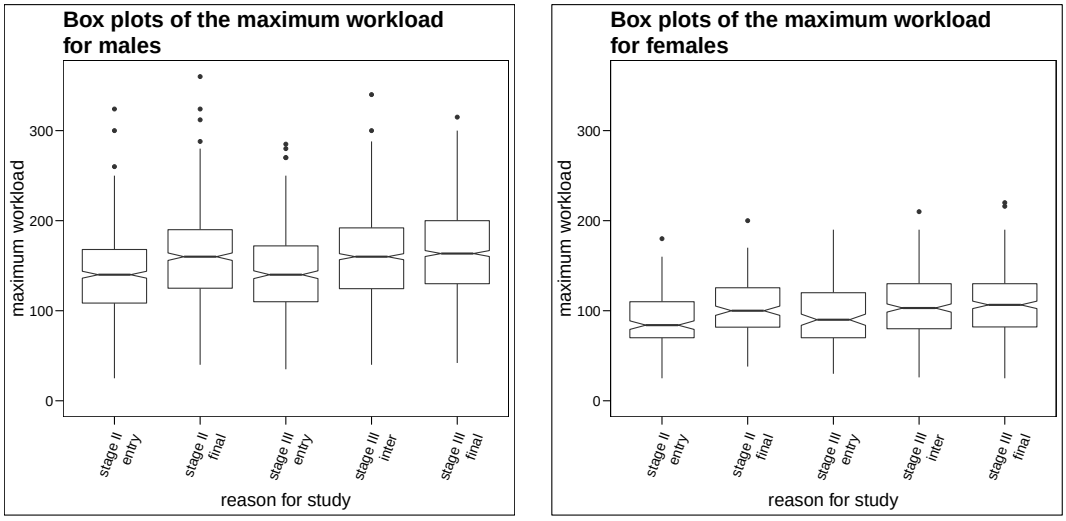
(b) Number of ergometries separated by diagnosis (nothing means neather infarkt nor khk nor another kardio disease)

Figure 7.2: Number of ergometries during cardiac out-patient rehabilitation separated by medication or diagnosis

Finally the analysis of the maximum workload which is achieved in the ergometries during cardiac out-patient rehabilitation is shown in Figure 7.3. Figure 7.3(a) shows the maximum workload for males and Figure 7.3(b) shows the maximum workload for females. The first fact which can be observed is that the maximum workload is much less for females. In each sub-figure the data are further split into the different stages during cardiac outpatient-rehabilitation. For both males and females an increase of the maximum workload during stage II as well as during stage III can be observed. Another interesting result is that the maximum workload decreases between the two stages. One reason for this could be that the patients do not do any exercises during the break between stage II and stage III. A more precise analysis of the maximum workload during stage II and stage III is done in the following sections.

7.2 Normal values for the maximum workload of the entry examination of stage II during cardiac out-patient rehabilitation

In this section a linear regression is done for deriving normal values for the maximum workload of the entry examination of stage II during cardiac out-



(a) Male (b) Female

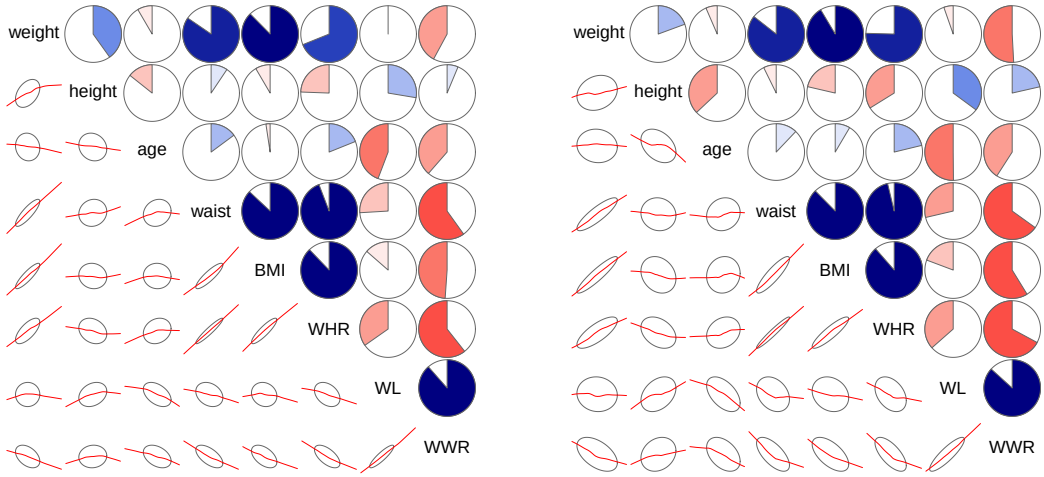
Figure 7.3: Box plots of the maximum workload during the different stages of cardiac out-patient rehabilitation split by gender

patient rehabilitation. First the relevant variables are identified by a correlation analysis. Afterwards a regression tree is built to identify the main variables. Finally this section includes a regression model for deriving the normal values. Therefore also a variable selection is done. The resulting model is notated by Model RII (rehabilitation stage II).

For deriving normal values the waist circumference is a crucial variable. Unfortunately, not for all data the waist circumference is available. For this reason just the data which include values for the waist circumference are included. This constraint reduces the data from 779 to 430 ergometries.

7.2.1 Correlation

Figure 7.4 shows the correlation between all variables for male and female. In both correlation matrices the maximum workload-to-weight-ratio has a higher correlation to the other variables than the maximum workload. For this reason the normal values refer to the maximum workload-to-weight-ratio. Note that also further independent variables were considered at the beginning. The first group of variables are the variables which are measured in the beginning of the ergometry, i.e. resting heart rate, resting systolic blood pressure and resting diastolic blood pressure. The second group were the results of the laboratory, i.e. cholesterol, low-density lipoprotein, high-density lipoprotein, triglyceride, creatinine, fasting blood sugar level and uric acid. Due to the complexity in practice and the low correlation to the maximum workload-to-weight-ratio these variables are not used for the regression. The remaining independent variables weight, height, age, waist and waist-to-height-ratio form the pool of variables for the regression.



(a) Male (b) Female

Figure 7.4: Analysis of correlation with the variables weight, height, body mass index, age, waist, waist-to-height-ratio, maximum workload and maximum workload-to-weight-ratio for the entry examination of stage II

7.2.2 Regression trees

In a next step regression trees are built for the maximum workload-to-weight-ratio. For the regression tree the independent variables waist and BMI are removed because of the high correlation to other independent variables. The resulting variables are weight, height, age, and waist-to-height-ratio. Figure 7.5 shows the regression trees. For both males and females the variables waist-to-height-ratio and age are relevant. For females also the variable weight is crucial. To sum it up, the structure of the regression trees varies and the resulting box plots of the leaf vertices are crucially different for both genders.

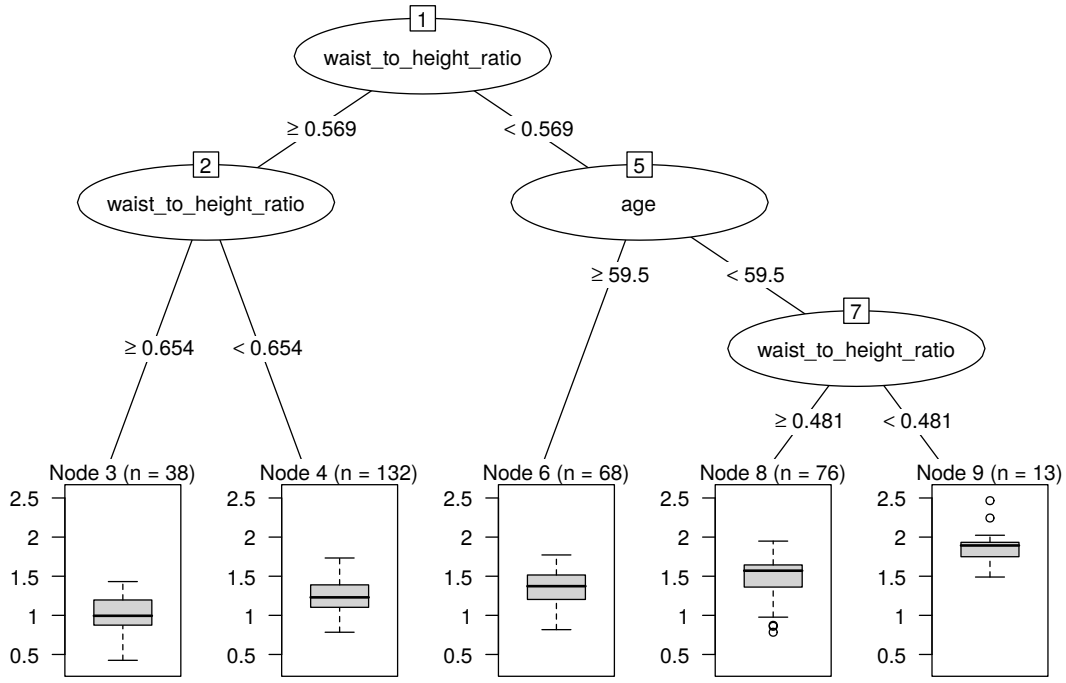
7.2.3 Variable selection

For deriving the regression model a variable selection is done first to find out the best variables. Table 7.1 and Table 7.2 show the best models for different numbers of parameters.

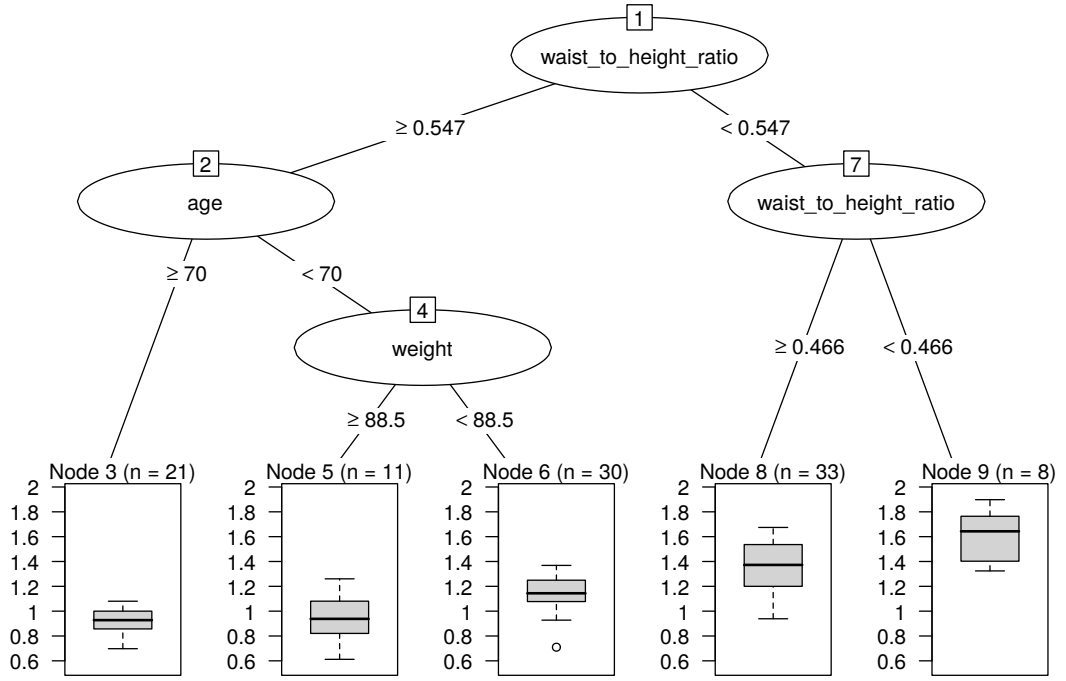
no.	height	age	WHR	height:weight	height:age	weight:age	age:WHR	weight:age:WHR
1								*
2			*		*			
3				*		*	*	
4	*	*	*		*			

Table 7.1: Summary of the best models with one to four parameters for deriving Model RII for males

In a next step the best model in Table 7.1 and in Table 7.2 has to be chosen. The variables in the best models are different for males and females, hence there



(a) Male

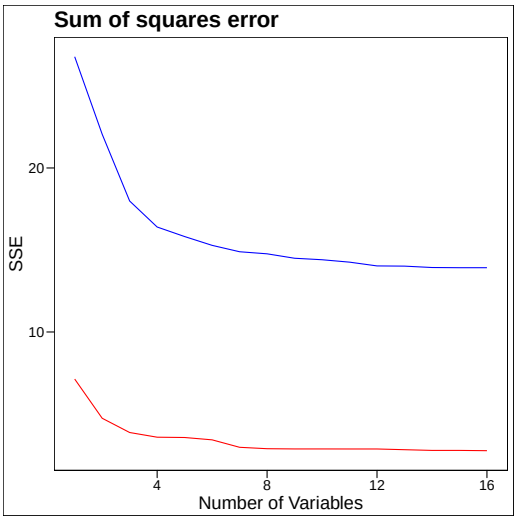


(b) Female

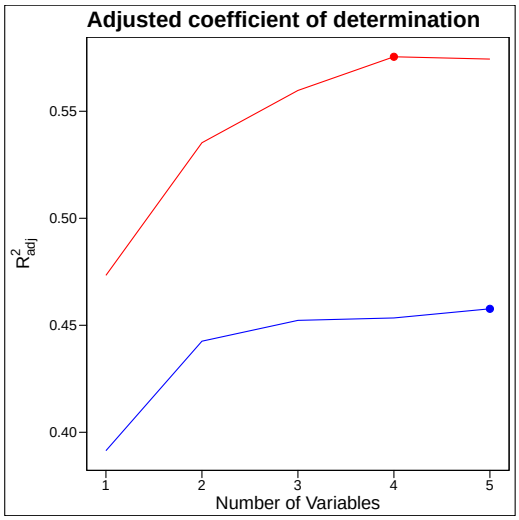
Figure 7.5: Regression trees with the independent variables weight, height, age, body mass index, waist, waist-to-height-ratio and the dependent variable maximum workload-to-weight-ratio for the entry examination of stage II

no.	height:weight	weight:age	height:WHR	age:WHR	height:weight: age	height:weight: WHR	weight:age: WHR
1							*
2			*	*			
3	*	*		*			
4	*			*	*	*	

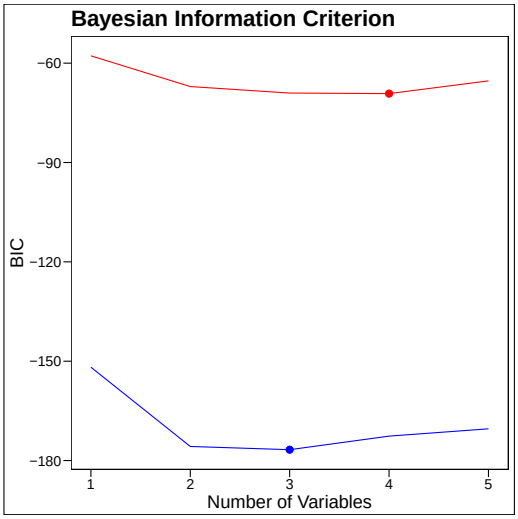
Table 7.2: Summary of the best models with one to four parameters for deriving Model RII for females



(a) Graphical description of the error sum of squares for the best models with 1 to 16 variables



(b) Graphical description of the R^2_{adj} for the best models with 1 to 5 variables



(c) Graphical description of the BIC for the best models with 1 to 5 variables

Figure 7.6: Information criteria for deriving Model RII

exist different parameters for both models. This is consistent to the observation that the regression trees have a different structure. Figure 7.6 shows the plots for the error sum of squares, the adjusted coefficient of determination and the BIC.

For the resulting regression models for males 3 parameters are chosen and

for females 4 parameters are chosen:

Model RII for males:

$$\hat{WWR}^{0.6}[W] = \hat{\beta}_0 + \hat{\beta}_1 height[cm]weight[kg] + \hat{\beta}_2 weight[kg]age[y] + \hat{\beta}_3 age[y]WHR$$

Model RII for females:

$$\hat{WWR}^{0.6}[W] = \hat{\beta}_0 + \hat{\beta}_1 height[cm]weight[kg] + \hat{\beta}_2 age[y]WHR + \hat{\beta}_3 height[cm]weight[kg]age[y] + \hat{\beta}_4 height[cm] : weight[kg] : WHR$$

7.2.4 Analysis of the model

$WWR^{0.6} \sim height : weight + weight : age + age : WHR$					
Coefficients:					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.428e+00	8.497e-02	28.578	< 2e-16	***
height:weight	-5.521e-05	6.803e-06	-8.116	1.02e-14	***
weight:age	1.782e-04	3.159e-05	5.642	3.67e-08	***
age:WHR	-3.594e-02	3.979e-03	-9.032	< 2e-16	***
RSE	0.1697 on 323 degrees of freedom				
R^2	0.4539				
R^2_{adj}	0.4489				
F-statistic	89.5 on 3 and 323 DF				
p-value:	< 2.2e-16				
Signif. codes	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 7.3: Summary of Model RII for males

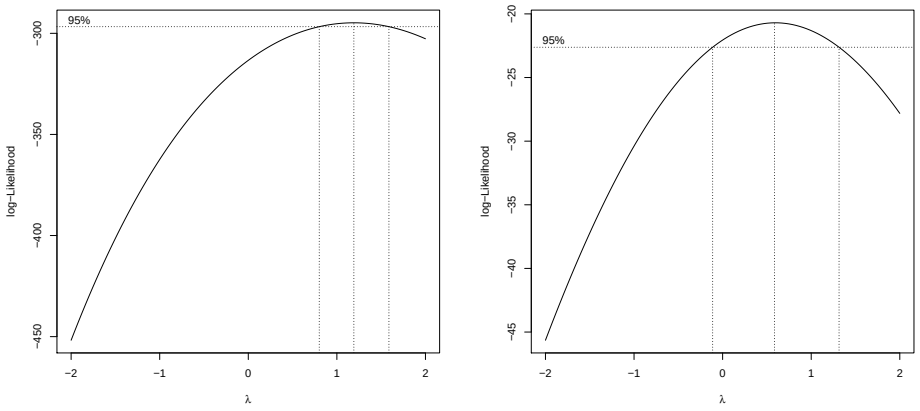
$WWR^{0.6} \sim height : weight + age : WHR + height : weight : age$ $+height : weight : WHR$					
Coefficients:					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.847e+00	3.339e-01	8.528	1.87e-13	***
height:weight	-1.754e-04	5.432e-05	-3.230	0.001687	**
age:WHR	-4.145e-02	8.710e-03	-4.759	6.72e-06	***
height:weight:age	1.485e-06	4.100e-07	3.622	0.000465	***
height:weight:WHR	1.116e-04	4.916e-05	2.271	0.025328	*
RSE	0.1393 on 98 degrees of freedom				
R^2	0.5929				
R^2_{adj}	0.5763				
F-statistic	35.68 on 4 and 98 DF				
p-value:	< 2.2e-16				
Signif. codes	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 7.4: Summary of Model RII for females

Table 7.3 and Table 7.4 show the summaries for the models. The adjusted coefficient of determination is 0.45 for males and 0.58 for females.

7.2.5 Diagnostics of the model

The Box-Cox plots in Figure 7.7 show that the already performed transformation of the maximum workload was necessary.



(a) Male (b) Female

Figure 7.7: Box-Cox plots of Model RII

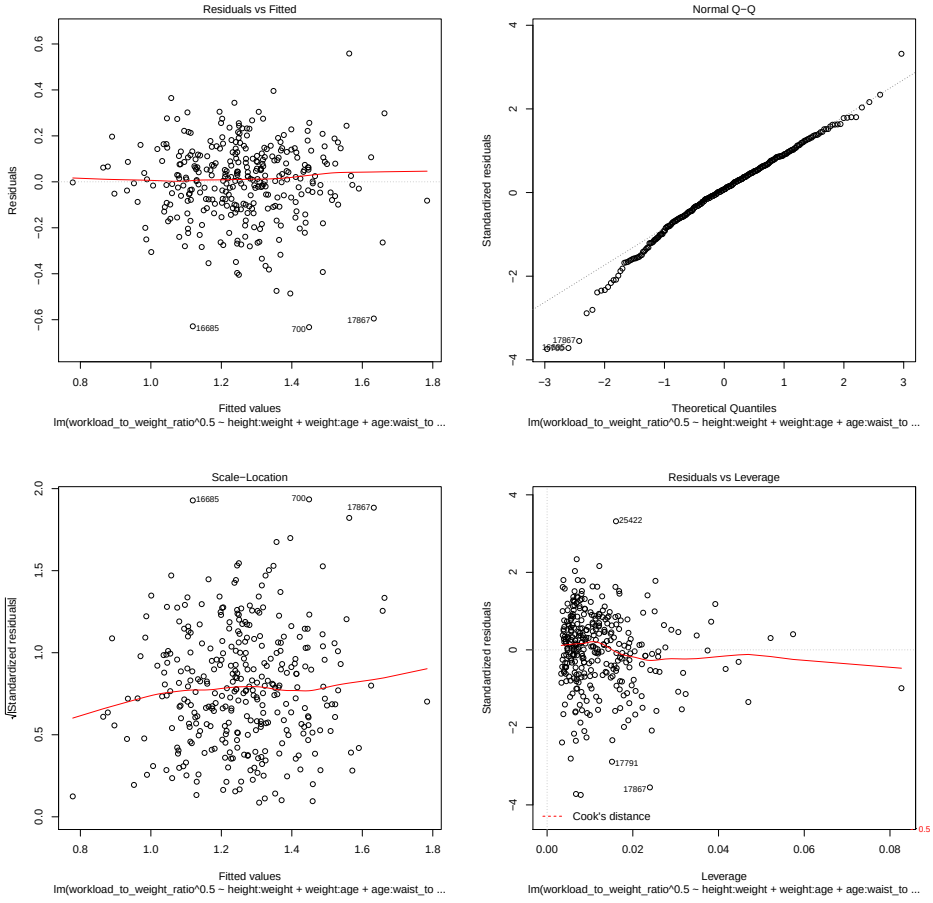


Figure 7.8: Diagnostic plots of Model RII for males

Figure 7.8 and Figure 7.9 analyze Model RII to check the requirements of the regression model for males and females, respectively.

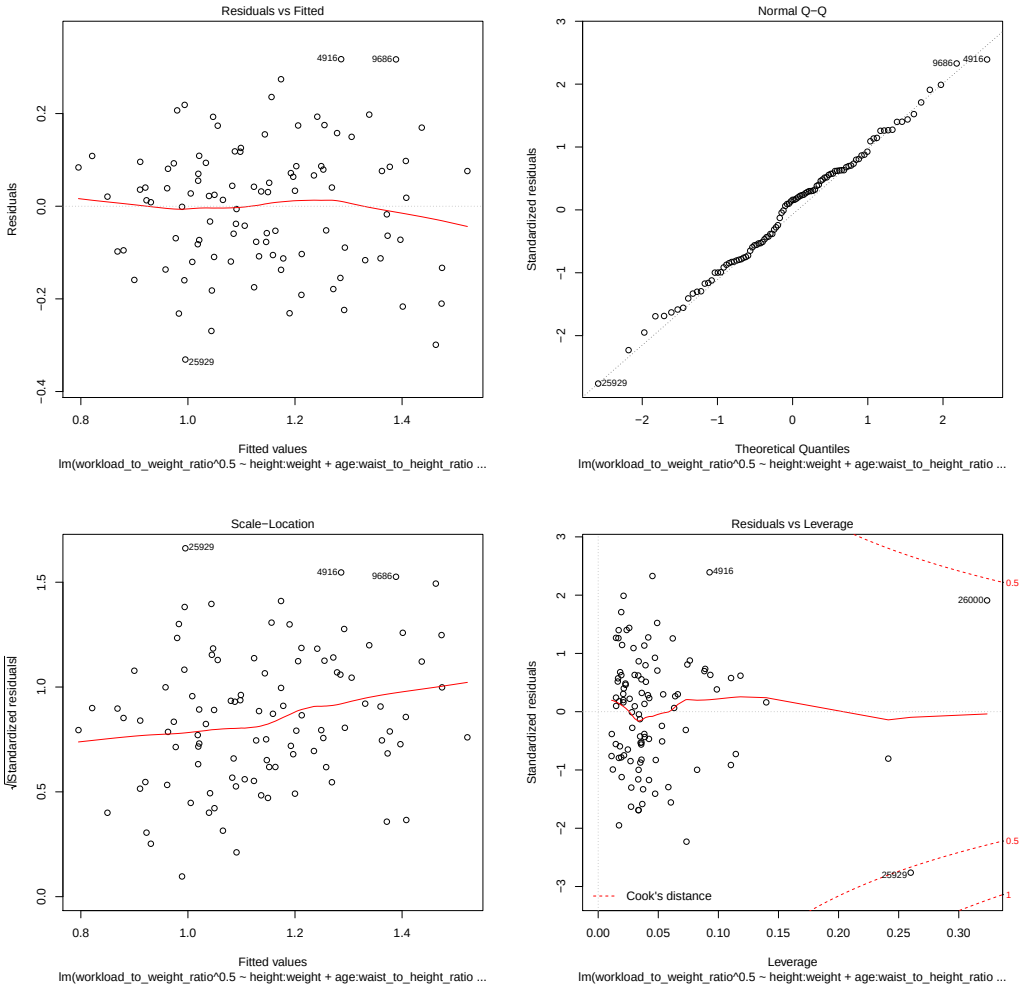


Figure 7.9: Diagnostic plots of Model RII for females

7.3 Normal values for the contribution of the maximum workload in stage II during cardiac outpatient rehabilitation

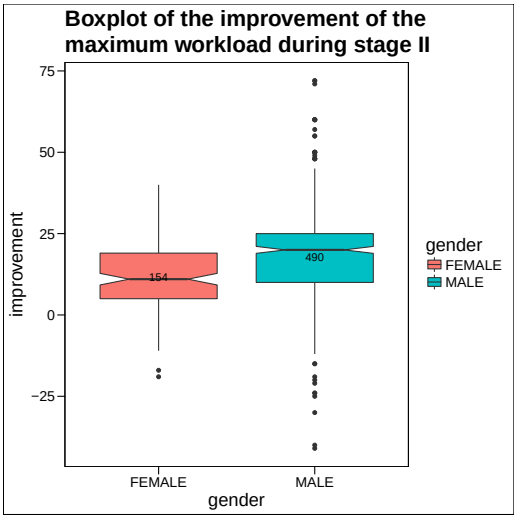
In this section the contribution of the maximum workload in stage II is analyzed. Therefore just the data of patients which attend both the entry examination and the final examination of stage II are considered. In total there exist 848 patients who did at least one of the ergometries during stage II. 644 patients completed both ergometries. For the following analysis just the patients, who completed both ergometries are relevant.

As a starting point the maximum workload in stage II is presented graphically. Figure 7.10(a) compares the maximum workload of stage II in the entry examination with the final examination. The red line highlights where both

workloads are equal. All the points above the red line have an improvement, all the points under the red line have a deterioration.

Figure 7.10(b) shows box plots of the improvement of the maximum workload itself. In this figure the data are additionally separated by gender. For both males and females the boxes are in the positive area. The median of the improvement for males is higher than the median of the improvement for females.

(a) Comparision of the maximum workload from the entry examination with the final examination of stage II



(b) Box plot of the improvement of the maximum workload during stage II split by gender

Figure 7.10: Graphical illustration of the improvement of the maximum workload in stage II of cardiac out-patient rehabilitation

Table 7.5 contains statistical indicators for the improvement. During stage II the mean of the improvement for males is about 17.9 W and for females it is about 11.8 W.

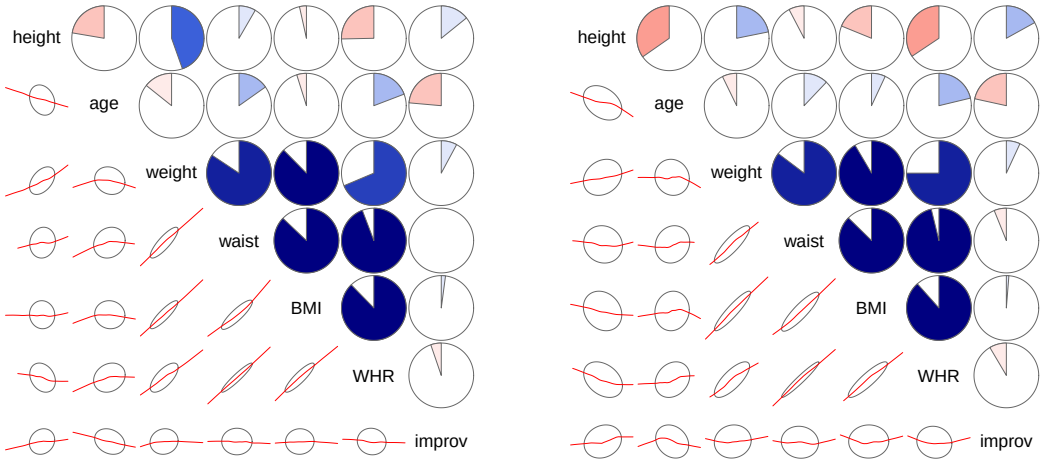
	min	$q_{0.25}$	$q_{0.5}$	$\bar{x}$	$q_{0.75}$	max	s^2	NA's
male	-41	10	20	17.86	25	72	234.24	165
female	-19	5	11	11.76	19	40	106.86	39

Table 7.5: Descriptive statistics of the improvement during stage II: minimum, 1-quartile, median, average, 3-quartile, maximum, standard deviation and the number of missing values

7.3.1 Correlation

Figure 7.11 shows the correlation between the independent variables height, age, weight, waist, body mass index, waist-to-height-ratio and the dependent variable improvement. It shows that the correlation between all independent variables and the improvement is quite low for males as well as for females.

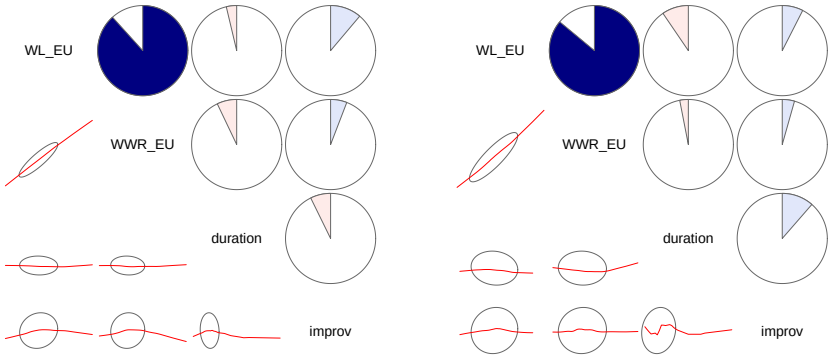
In addition to the variables specified above the normal values for the improvement can also be derived from the maximum workload, the maximum



(a) Male (b) Female

Figure 7.11: Correlation matrix for stage II of cardiac out-patient rehabilitation between the variables height, age, weight, waist, body mass index, waist-to-height-ratio and the improvement

workload-to-weight-ratio and the duration of stage II. The maximum workload and the maximum workload-to-weight-ratio only refer to the entry examination of stage II. Figure 7.12 shows the correlation matrix for these variables and the improvement. It shows that the correlation is very low.



(a) Male (b) Female

Figure 7.12: Correlation matrix for stage II of cardiac out-patient rehabilitation between the variables maximum workload at entry examination of stage II, maximum workload-to-weight-ratio at entry examination of stage II, duration of stage II and the improvement

7.3.2 Regression trees

This section shows regression trees for the improvement in Figure 7.13. The structure of the trees is different for males and females. The box plots in the leaf vertices are quite similar.

7.3.3 Regression model

Executing a linear regression with variable selection leads to a regression model with 5 parameters for males and 4 parameters for females. Both models have very low adjusted coefficients of determination: under 0.15 and about 0.1. For this reason the regression model is not shown in this chapter.

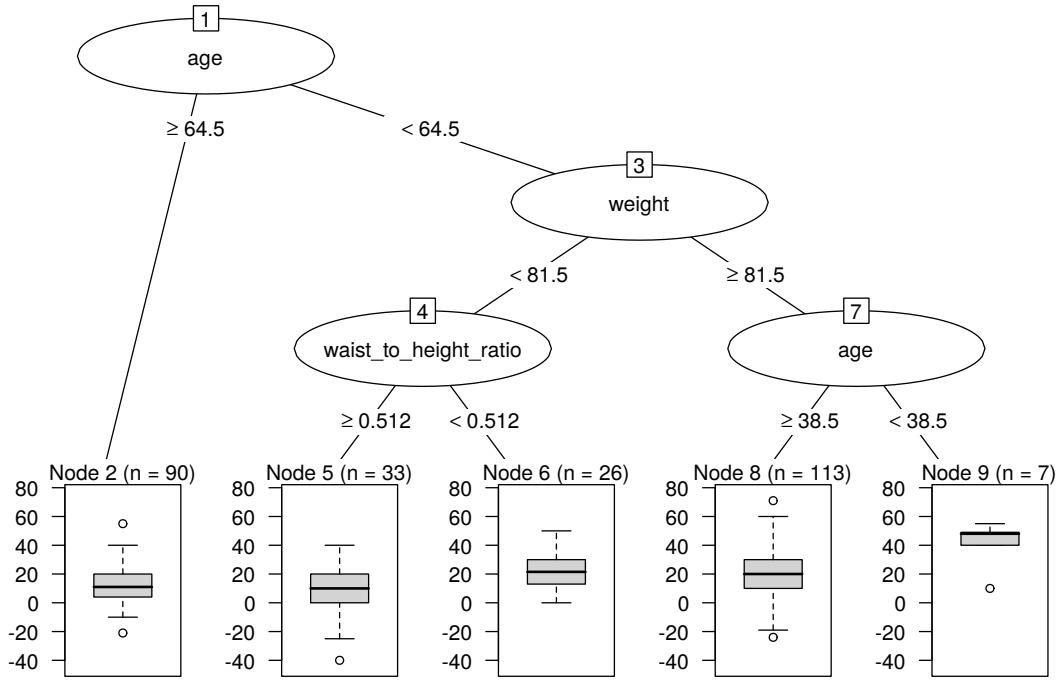
7.4 Normal values for the maximum workload of the entry examination of stage III during cardiac out-patient rehabilitation

In this section a linear regression is done for deriving normal values for the maximum workload of the entry examination of stage II during cardiac out-patient rehabilitation. First the relevant variables are identified by a correlation analysis. Afterwards to illustrate the main variables a regression tree is built. Finally this section includes a regression model for deriving the normal values. Therefore also a variable selection is done. The resulting model is notated by Model RIII (rehabilitation stage III).

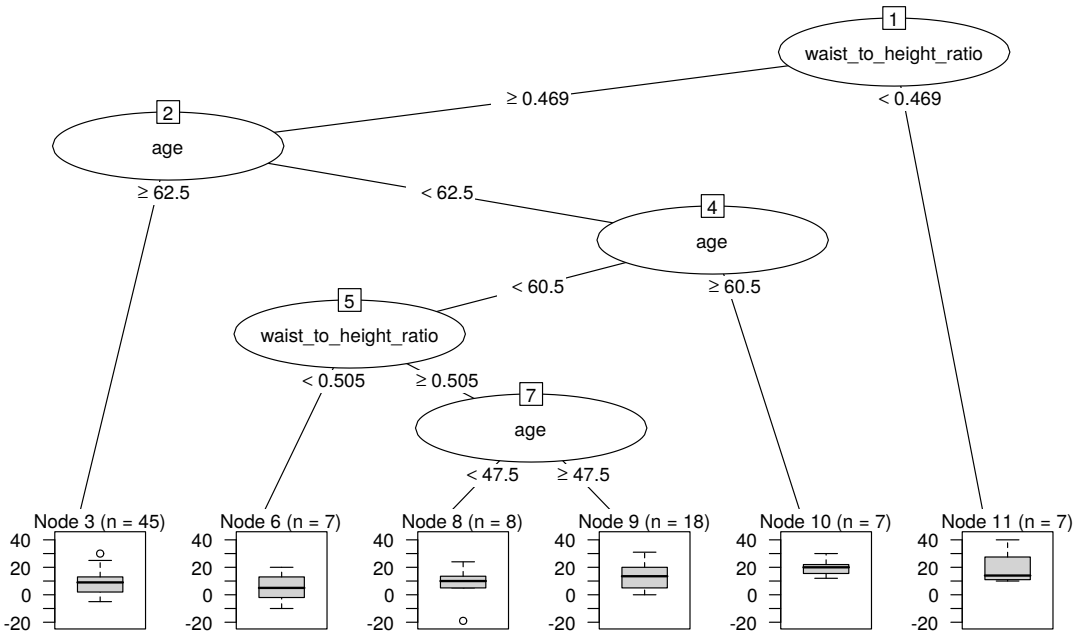
For deriving normal values the waist circumference is a crucial variable. Unfortunately, not for all data during rehabilitation the waist circumference is available. For this reason just the data which include values for the waist circumference are included. This constraint reduces the data from 674 to 340 ergometries.

7.4.1 Correlation

Figure 7.14 shows the correlation between all variables for males and females. In both correlation matrices the maximum workload-to-weight-ratio has a higher correlation to the other variables than the maximum workload. For this reason the normal values refer to the maximum workload-to-weight-ratio. Note that also further independent variables were considered in the beginning. The first group of variables are the variables which are measured at the beginning of the ergometry, i.e. resting heart rate, resting systolic blood pressure and resting diastolic blood pressure. The second group were the results of the laboratory, i.e. cholesterol, low-density lipoprotein, high-density lipoprotein, triglyceride, creatinine, fasting blood sugar level and uric acid. Due to the complexity in practice and the low correlation to the maximum workload-to-weight-ratio this variables are not used for the regression. The remaining independent variables weight, height, age, waist and waist-to-height-ratio form the pool of variables for the regression.

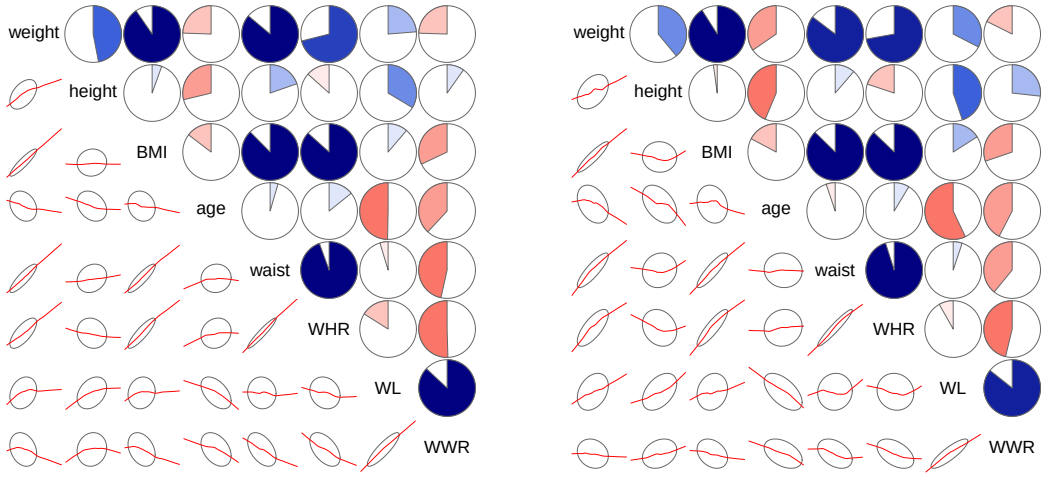


(a) Males



(b) Females

Figure 7.13: Regression trees for the improvement of stage II with the independent variables weight, height, age, waist-to-height-ratio and the dependent variable maximum workload-to-weight-ratio



(a) Male (b) Female

Figure 7.14: Analysis of correlation with the variables weight, height, body mass index, age, waist, waist-to-height-ratio, maximum workload and maximum workload-to-weight-ratio for the entry examination of stage III

7.4.2 Regression trees

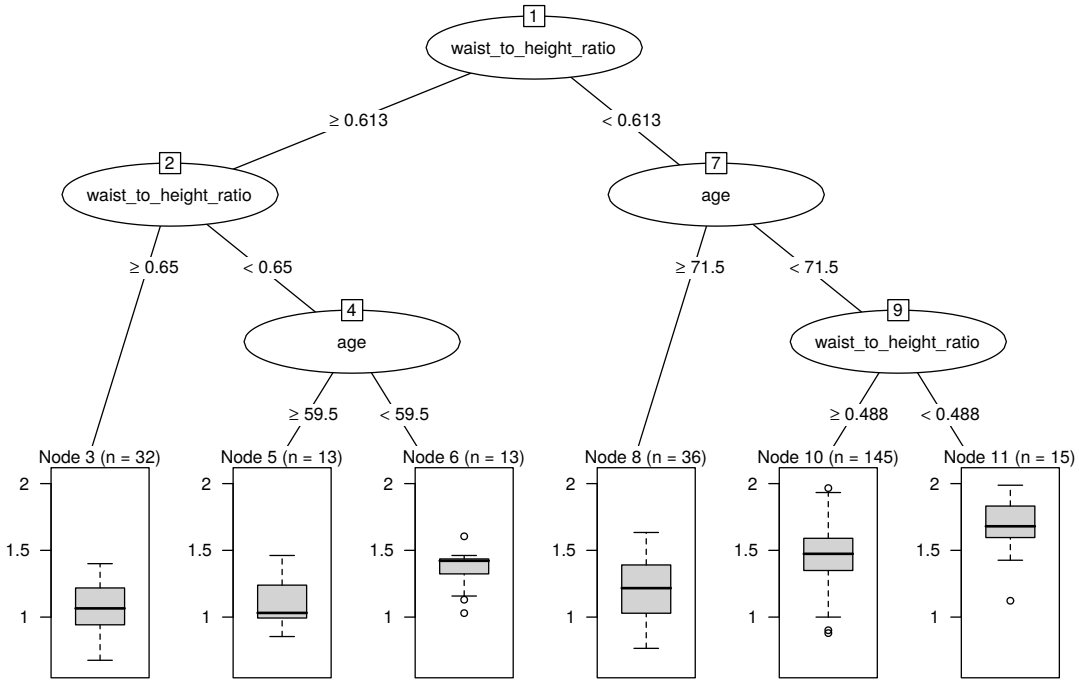
In a next step regression trees are built for the maximum workload-to-weight-ratio. For the regression tree the independent variables waist and BMI are removed because of the high correlation to other independent variables. The resulting variables are weight, height, age, and waist-to-height-ratio. Figure 7.15 shows the received the regression trees. For both males and females the variables waist-to-height-ratio and age are relevant. For females also the variable weight is crucial. To sum it up, the structure of the regression trees varies and the resulting box plots of the leaf vertices are crucially different for both genders.

7.4.3 Variable selection

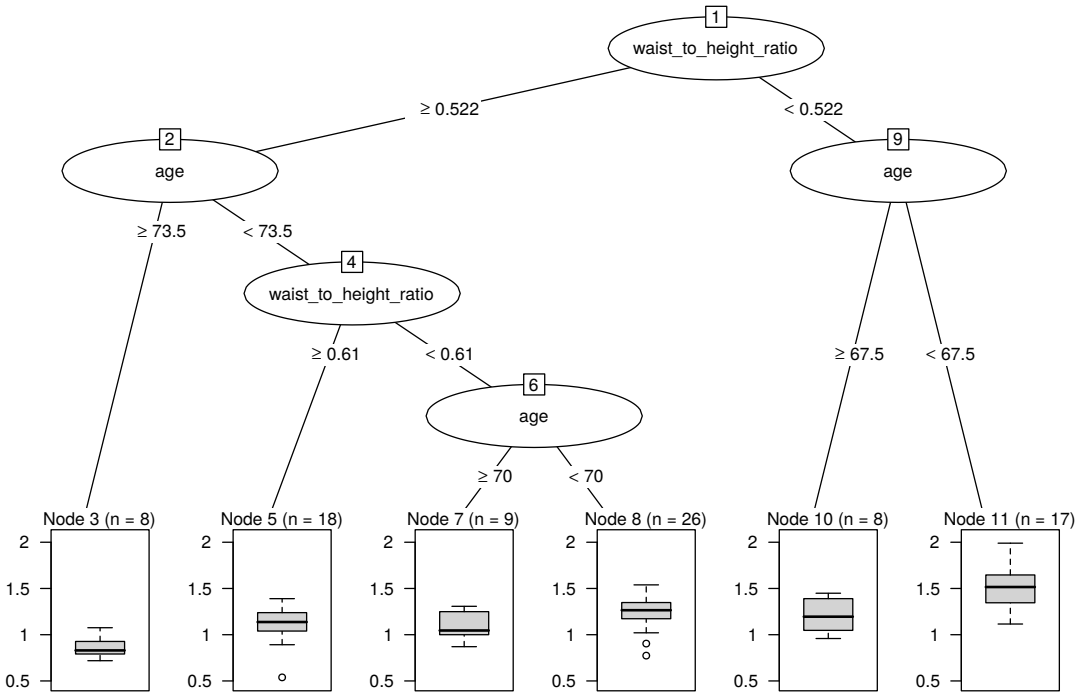
For deriving the regression model a variable selection is done first to find out the best variables. Table 7.6 and Table 7.7 show the best models for different numbers of parameters.

no.	height:age	weight:age	age:WHR	height:weight: age	height:age: WHR	weight:age: WHR	height:weight:age: WHR
1						*	
2	*				*		
3		*			*		*
4		*	*	*		*	
5		*	*	*	*	*	

Table 7.6: Summary of the best models with 1 to 5 parameters for deriving Model RIII for males



(a) Male

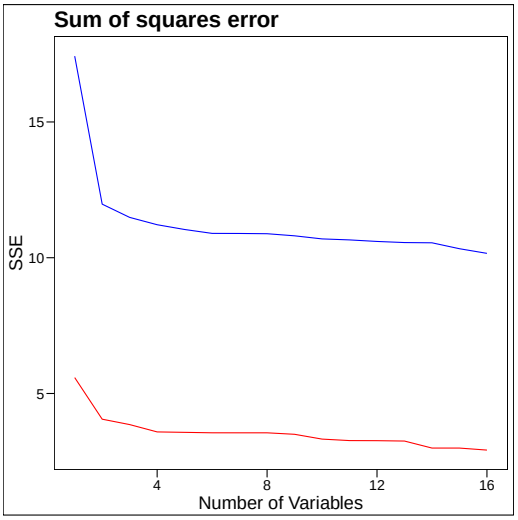


(b) Female

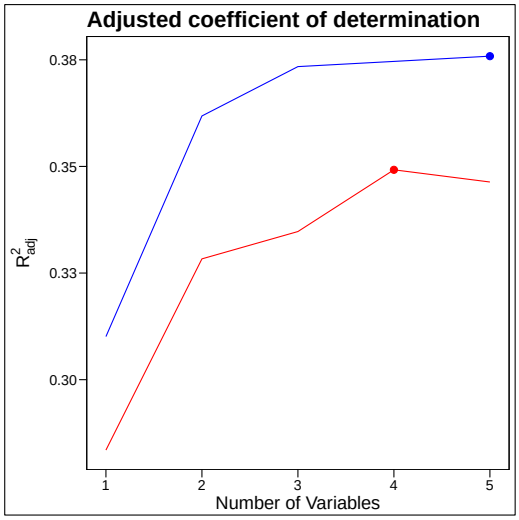
Figure 7.15: Regression trees with the independent variables weight, height, age, body mass index, waist, waist-to-height-ratio and the dependent variable maximum workload-to-weight-ratio for the entry examination of stage III

no.	height	age	WHR	height:WHR	age:WHR	height:weight: WHR	height:age: WHR	weight:age: WHR
1					*			
2		*	*					
3					*	*		*
4	*		*	*	*			
5	*		*	*	*		*	

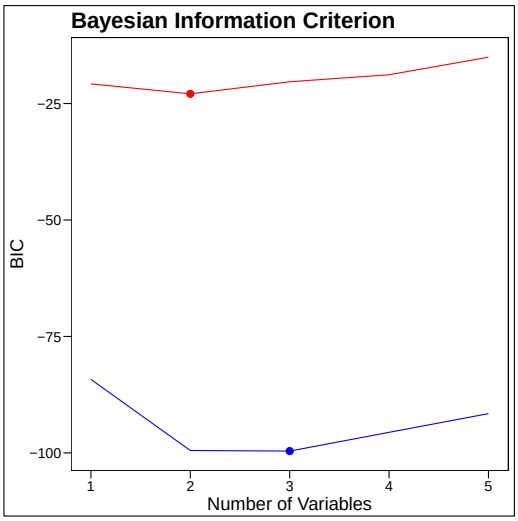
Table 7.7: Summary of the best models with 1 to 5 parameters for deriving Model RIII for females



(a) Graphical description of the error sum of squares for the best models with one to 16 variables



(b) Graphical description of the R^2_{adj} for the best models with one to 5 variables



(c) Graphical description of the BIC for the best models with one to 5 variables

Figure 7.16: Information criteria for deriving Model RIII

In a next step the best of the models in Table 7.6 and in Table 7.7 has to be chosen. Unfortunately the variables in the best models are different for males and females. This is consistent to the observation that the regression trees have

a different structure. Figure 7.16 shows the plots for the error sum of squares, the adjusted coefficient of determination and the Bayesian information criterion.

Using the BIC for men 3 parameters and for women 2 parameters are chosen.
male:

$$\begin{aligned} \hat{WWR}^{0.5}[W] &= \hat{\beta}_0 + \hat{\beta}_1 weight:age + \hat{\beta}_2 height:age:WHR \\ &\quad + \hat{\beta}_3 height:weight:age:WHR \end{aligned}$$

female:

$$\hat{WWR}^{0.5}[W] = \hat{\beta}_0 + \hat{\beta}_1 age + \hat{\beta}_2 WHR$$

7.4.4 Analysis of the model

Table 7.8 and Table 7.9 summarize both models.

$WWR^{0.6} \sim weight:age + age:height:WHR + weight:age:height:WHR$					
Coefficients:					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.597e+00	6.005e-02	26.601	< 2e-16	***
weight:age	2.131e-04	4.598e-05	4.636	5.73e-06	***
age:height:WHR	-1.197e-04	1.964e-05	-6.094	4.15e-09	***
weight:age:height:WHR	-1.283e-06	2.292e-07	-5.599	5.66e-08	***
Residual standard error	0.1653 on 250degrees of freedom				
R^2	0.3817				
R^2_{adj}	0.3743				
F-statistic	51.45 on 3 and 250 DF				
p-value:	< 2.2e-16				
Signif. codes	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

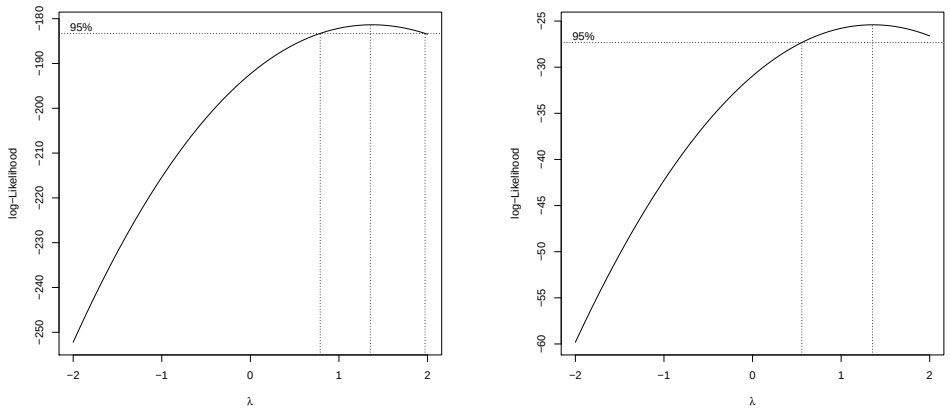
Table 7.8: Summary of Model RIII for males

$WWR^{0.6} \sim age + WHR$					
Coefficients:					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.227838	0.169508	13.143	< 2e-16	***
age	-0.005608	0.001355	-4.138	8.36e-05	***
WHR	-1.273880	0.274596	-4.639	1.29e-05	***
Residual standard error	0.1708 on 83 degrees of freedom				
R^2	0.3375				
R^2_{adj}	0.3215				
F-statistic	21.14 on 2 and 83 DF				
p-value:	3.805e-08				
Signif. codes	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 7.9: Summary of Model RIII for females

7.4.5 Diagnostics of the model

The Box-Cox plots in Figure 7.17 show that the already performed transformation of the maximum workload was necessary.



(a) Male (b) Female

Figure 7.17: Box-Cox plots of Model RIII

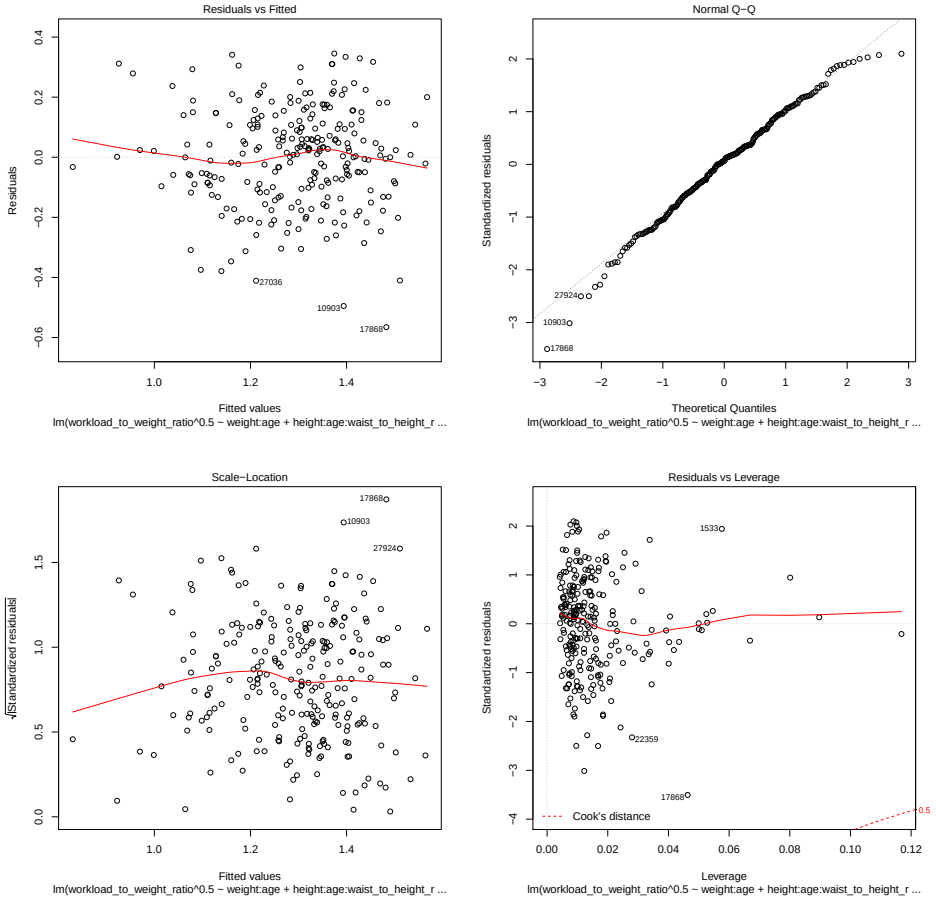


Figure 7.18: Diagnostic plots of Model RIII for males

Figure 7.18 and Figure 7.19 analyze Model RIII to check the requirements of the regression model for men and women,respectively.

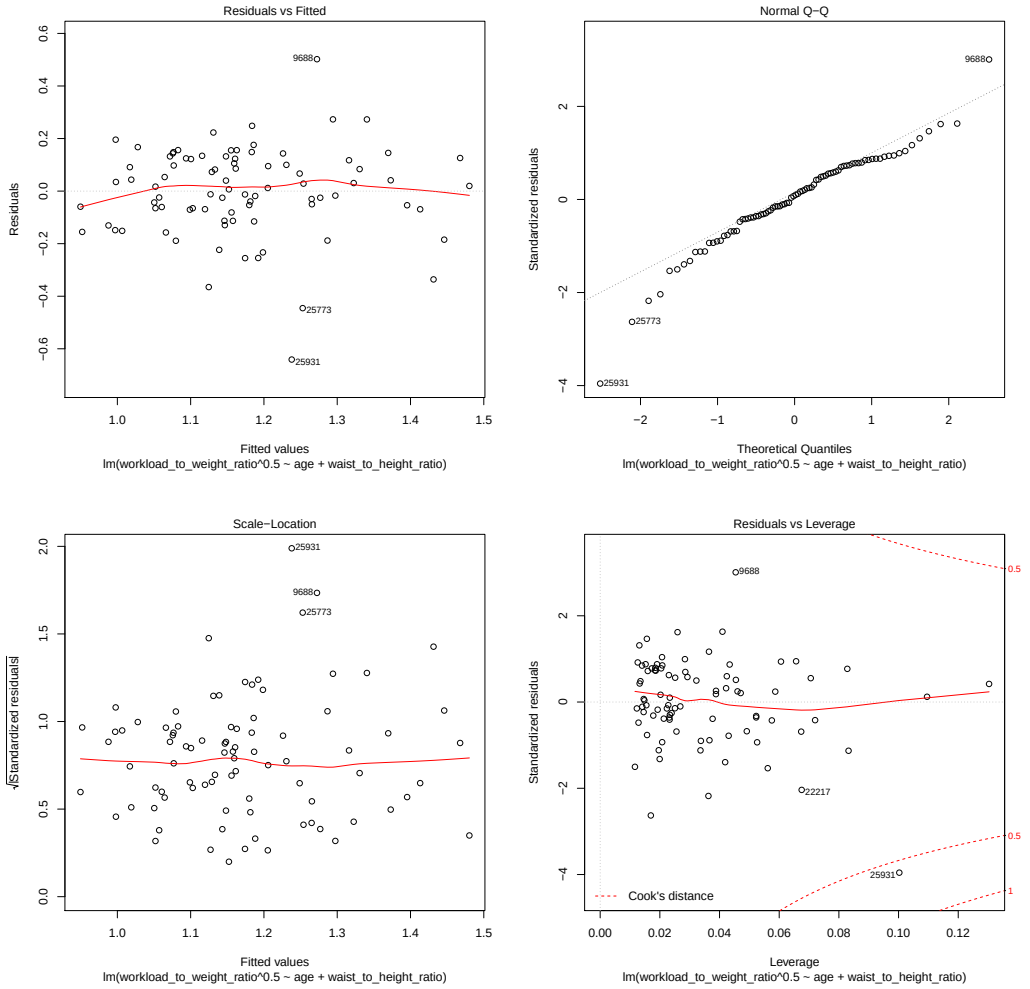


Figure 7.19: Diagnostic plots of Model RIII for females

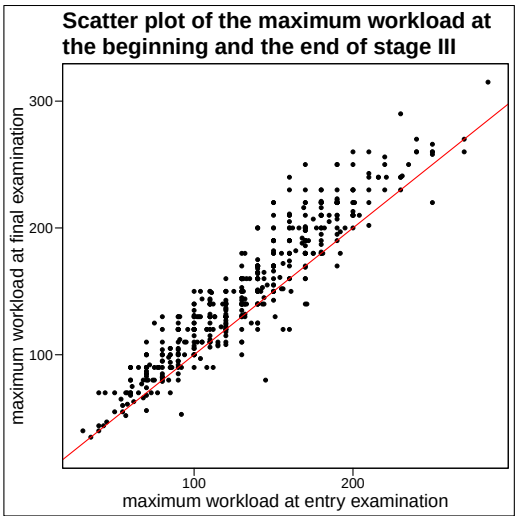
7.5 Normal values for the contribution of the maximum workload in stage III during cardiac outpatient rehabilitation

In this section the contribution of the maximum workload in stage III is analyzed. Therefore just the data of patients which attend both the entry examination and the final examination of stage III are considered. In total there exist 1476 patients who did at least one of the ergometries during stage III. 787 patients completed entry examination and the final examination. Note that for the missing entry examinations of stage III, the final examinations of stage II are used, which are missing in Figure 7.1.

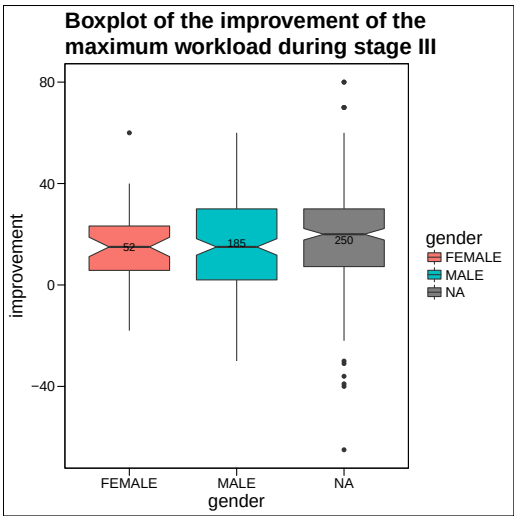
As a starting point the maximum workload in stage III is presented graphically. Figure 7.20(a) compares the maximum workload of stage III in the entry

examination with the final examination. The red line highlights where both workloads are equal. All the points above the red line have an improvement, all the points under the red line have a deterioration.

Figure 7.20(b) shows box plots of the improvement of the maximum workload itself. In this figure the data are additionally separated by gender. For both males and females the boxes are in the positive area. The median of the improvement for males is higher than the median of the improvement for females.



(a) Comparision of the maximum workload from the entry examination with the final examination of stage III



(b) Box plot of the improvement of the maximum workload during stage III split by gender

Figure 7.20: Graphical illustration of the improvement of the maximum workload in stage III of cardiac out-patient rehabilitation

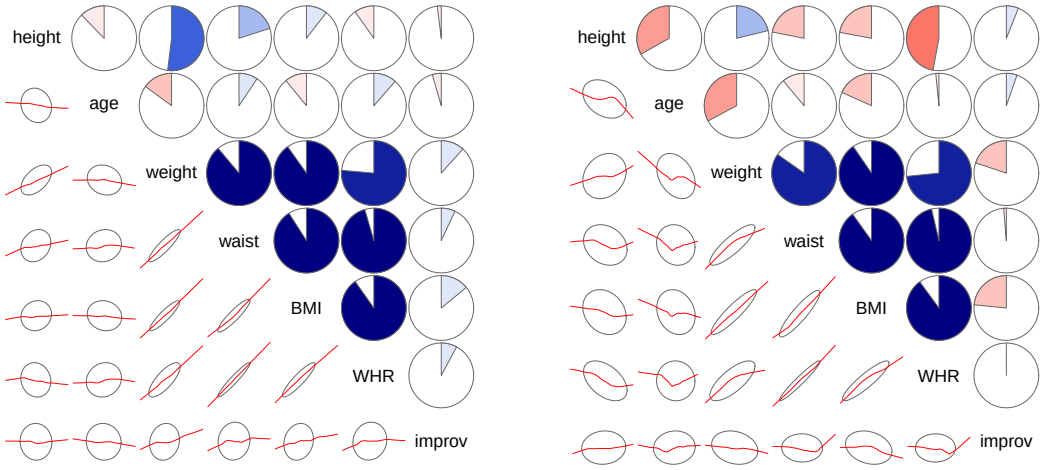
Table 7.10 contains statistical indicators for the improvement. During stage III the mean of the improvement for males is about 14.2 W and for females it is about 9.7 W.

	min	$q_{0.25}$	$q_{0.5}$	$\bar{x}$	$q_{0.75}$	max	s^2	NA's
male	-80	0	12	14.22	30	80	388.75	508
female	-60	0	10	9.65	20	60	234.27	143

Table 7.10: Descriptive statistics of the improvement during stage III: minimum, 1-quartile, median, average, 3-quartile, maximum, standard deviation and the number of missing values

7.5.1 Correlation

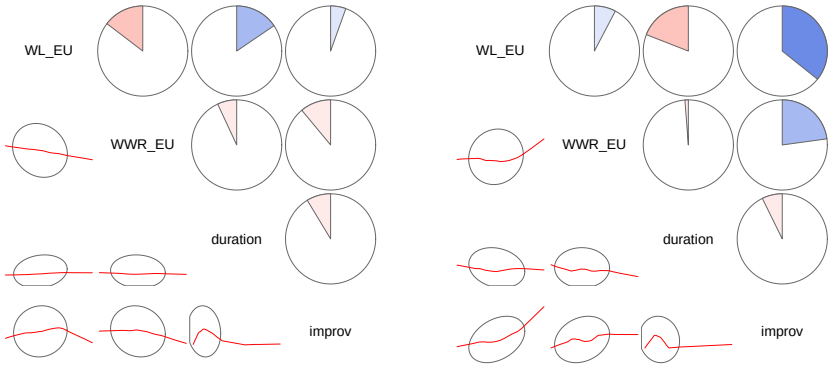
Figure 7.21 shows the correlation between the independent variables height, age, weight, waist, body mass index, waist-to-height-ratio and the dependent variable improvement. It shows that the correlation between all independent variables and the improvement is quite low for males as well as for females.



(a) Male (b) Female

Figure 7.21: Correlation matrix for stage III of cardiac out-patient rehabilitation between the variables height, age, weight, waist, body mass index, waist-to-height-ratio and the improvement

For deriving normal values for the improvement the maximum workload and the maximum workload-to-weight-ratio, respectively, during the entry examination and the duration of stage III can be used. Figure 7.22 shows the correlation matrix for these variables and the improvement. It shows that the correlation is very low.

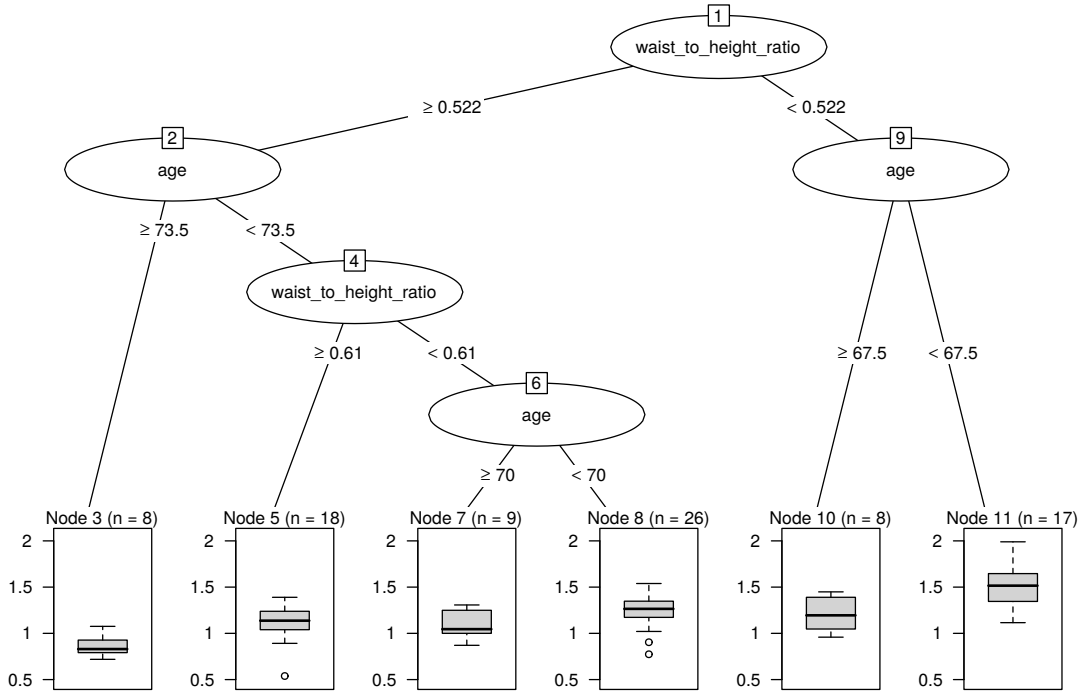


(a) Male (b) Female

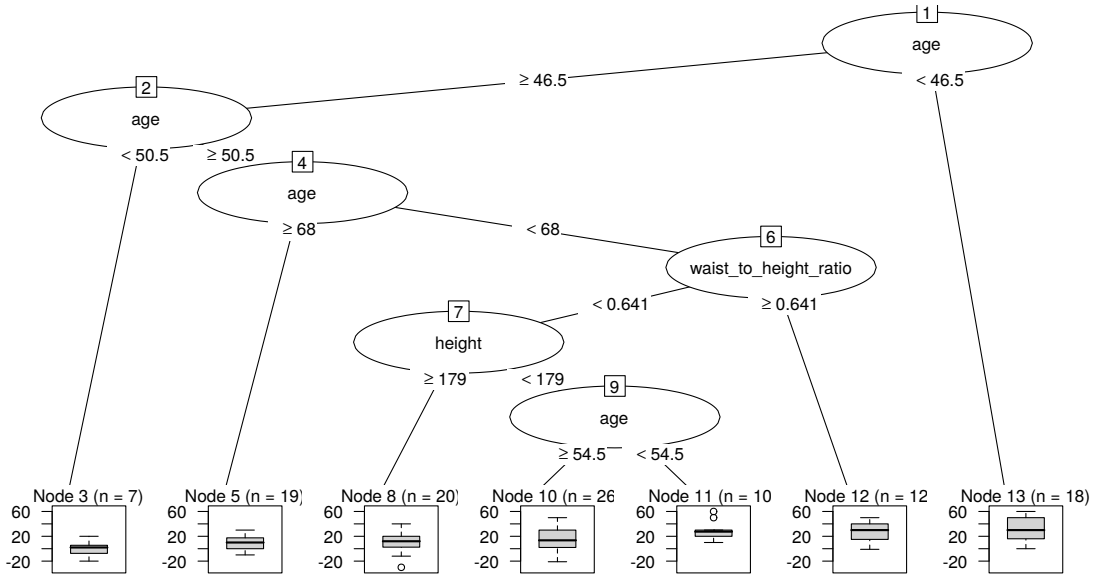
Figure 7.22: Correlation matrix for stage III of cardiac out-patient rehabilitation between the variables maximum workload at entry examination of stage III, maximum workload-to-weight-ratio at entry examination of stage II, duration of stage III and the improvement

7.5.2 Regression trees

This section shows regression trees for the improvement in Figure 7.23. The structure of the trees is different for males and females. The box plots in the leaf vertices are quite similar.



(a) Male



(b) Female

Figure 7.23: Regression trees for the improvement of stage III with the independent variables weight, height, age, waist-to-height-ratio and the dependent variable maximum workload-to-weight-ratio

7.5.3 Regression model

As for the normal values of the improvement of stage II the result of a linear regression has a very low adjusted coefficient of determination. For this reason the regression model is not shown in this chapter.

7.6 Discussion

At the moment, to my knowledge, there exist no normal values of the maximum workload in ergometries specific to the stages of rehabilitation. The same normal values are used for interpretation of ergometries from healthy people as well as people in rehabilitation. Nevertheless, these normal values should be adjusted to get rehabilitation specific normal values. In this chapter, normal values for the maximum workload for the entry examination stage II as well as for stage III are derived. The newly derived normal values have an adjusted coefficient of determination of 0.45 for males and 0.58 for females in stage II and 0.37 for males and 0.42 for females in stage III. Note, that the output of the regression models is the maximum workload-to-weight-ratio, which has to be multiplied with the weight to gain the maximum workload.

Furthermore, the improvement in the maximum workload during stage II and stage III is analyzed. For the improvement, the mean is calculated and also regression trees are created. Due to the low correlation between the independent variables and the maximum workload-to-weight-ratio, a regression analysis is not done.

A planned further research question is to find the optimal duration of stage III during cardiac out-patient rehabilitation. Due to only small changes in the duration of stage III and slight differences in the results, it was not possible to answer this by the underlying data.

Chapter 8

Conclusion and outlook for further research

8.1 Summary

In Austria, currently used reference values for the maximum workload reached during an ergometry are based on a model from 1972. ZARG uses the same values for the evaluation of ergometries performed at their center, but they doubt that the presently used reference values fit well.

For that reason, the evaluation of the currently used Model P with data from 28,625 ergometries against newly developed models was done to calculate the normal values of maximum workload. The analysis showed that the adjusted coefficient of determination for females in Model P is 0.23 as compared to 0.41 in Model 1, but there was not as much improvement seen in the adjusted coefficient of determination for males (0.35 vs 0.37). Therefore, it could be concluded that Model P does not fit well to calculate the normal values of maximum workload in females. This could be due to a different population used to calculate Model P. Furthermore, the use of the predicted variable maximum workload-to-weight-ratio and its box cox transformation leads to the development of another model (Model 3) which has an adjusted coefficient of determination of 0.39 for males and 0.49 for females, respectively. Hence, use of the variable maximum workload-to-weight-ratio instead of maximum workload leads to a better model. The use of the predicted variable maximum workload-to-weight-ratio and independent variables height and weight instead of the conventional variable body surface area resulted into Model 4. This model has an improved adjusted coefficient of determination of 0.48 for males and of 0.57 for females. Therefore, higher values of the adjusted coefficient of determination from Model 4 suggest this model is more precise in calculating the normal values of the maximum workload. Finally, development of a gender independent model using the history of medication as an independent dummy variable along with the previously used variable in Model 4 resulted in an adjusted coefficient of determination of 0.55. Therefore, we may conclude that the use of Model D eliminates the need for using gender specific models without compromising the value of the adjusted coefficient of determination.

In the current scenario, there is only one set of normal values used to interpret the maximum workload in ergometries irrespective of the cardiac fitness of

that individual. Even if the person is recovering from a heart attack or is entirely fit, the normal values of the maximum workload are the same. Therefore, the new normal values for the entry examination of stage II as well as for stage III of cardiac rehabilitation were calculated using the ergometry data of the people in cardiac rehabilitation. The newly derived normal values have an adjusted coefficient of determination of 0.45 for males and 0.58 for females in stage II, and of 0.37 for males and 0.42 for females in stage III. Henceforth, these cardiac rehabilitation stage-specific normal values could be used for the evaluation of ergometries during cardiac rehabilitation. Finally, the average improvement in the maximum workload during stage II of cardiac rehabilitation was 17.86 W in males and 11.76 W in females, whereas for stage III, the average improvement was 14.22 W for males and 9.65 W for females. Hence, these values could be used as the normal values to evaluate the improvement of a person in cardiac rehabilitation in the respective stages.

8.2 Limitation and Recommendations

The results of this thesis show that the waist circumference is a crucial parameter for deriving normal values of the maximum workload. However, in the cohort of 28,625 ergometries it was measured only during 3,800 ergometries. Therefore, it is suggested that for each ergometry the waist circumference should be measured. With more data of the waist circumference, the model would also contain information about the waist, to include the WHR in the model which should lead to an even better fit.

Because of non uniformity in the written diagnosis of the heart diseases in this cohort, the variable ‚diagnosis' was not precisely measured to be useful for the regression. Therefore, adaptation of the international statistical classification of diseases (ICD) codes for disease diagnosis will help to have more precise information for regression to calculate the normal values.

8.3 Further research questions

- Does statin therapy influence the maximum workload in cardiac exercise stress testing?
- What are the benefits of an out-patient cardiac rehabilitation over an in-patient cardiac rehabilitation?
- What can be concluded from a curve of the heart rate during a maximal cardiac exercise stress testing?

The analysis to answer the questions above could not be conducted due to the time constraint and unavailable data of in-patient cardiac rehabilitation.

To sum up the added value of this thesis, new improved normal values of the maximal workload for ergometries as well as specific normal values for out-patient rehabilitation can be directly applied in practice.

Bibliography

[Akkerboom, 2008] Akkerboom, H. (2008). *Wirtschaftsstatistik im Bachelor*. Gabler, Wiesbaden.

[Arstila, 1972] Arstila, M. (1972). Pulse-conducted triangular exercise-ecg test. *Acta Medica Austriaca*, 529(9):1–110.

[Bankhofer and Vogel, 2008] Bankhofer, U. and Vogel, J. (2008). *Datenanalyse und Statistik. Eine Einführung für Ökonomen im Bachelor*. Gabler, Wiesbaden.

[Bingham and Fry, 2010] Bingham, N. and Fry, J. M. (2010). *Regression - Linear Models in Statistics*. Springer, London.

[Bjarnason-Wehrens et al., 2010] Bjarnason-Wehrens, B., McGee, H., Zwisler, A. D., Piepoli, M. F., Benzer, W., Schmid, J. P., Dendale, P., Pogosova, N. G. V., Zdrenghea, D., Niebauer, J., and Mendes, M. (2010). Cardiac rehabilitation in Europe: Results from the European Cardiac Rehabilitation Inventory Survey. *European Journal of Preventive Cardiology*, 17(4):410–418.

[Boisbunon et al., 2014] Boisbunon, A., Canu, S., Fourdrinier, D., Strawderman, W., and Wells, M. T. (2014). AIC, Cp and estimators of loss for elliptically symmetric distributions. *International Statistical Review*, 82(3):422–439.

[Breheny and Burchett, 2017] Breheny, P. and Burchett, W. (2017). Visualization of Regression Models Using visreg. *The R Journal*, 9(2):56–71.

[Brown et al., 2005] Brown, S. P., Miller, W. C., and Eason, J. M. (2005). Clinical Exercise Testing. *European Respiratory Monograph*, 12(40):146–165.

[Burkschat et al., 2012] Burkschat, M., Cramer, E., and Kamps, U. (2012). *Beschreibende Statistik: Grundlegende Methoden der Datenanalyse*. Springer, Berlin.

[Cleff, 2014] Cleff, T. (2014). *Exploratory Data Analysis in Business and Economics*. Springer, Wiesbaden.

[Cleveland et al., 1992] Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992). *Local regression models*. Wadsworth & Brooks/Cole, California.

[Cornelissen, 2018] Cornelissen, J. (2018). RDocumentation: Integrate R with 'RDocumentation'. `https://cran.r-project.org/package=RDocumentation`.

[De Koning et al., 2007] De Koning, L., Merchant, A. T., Pogue, J., and Anand, S. S. (2007). Waist circumference and waist-to-hip ratio as predictors of cardiovascular events: Meta-regression analysis of prospective studies. *European Heart Journal*, 28(7):850–856.

[DeSilva, 2013] DeSilva, R. (2013). *Heart Disease*. Greenwood, Santa Barbara.

[Devore and Berk, 2012] Devore, J. L. and Berk, K. N. (2012). *Modern Mathematical Statistics with Applications*. Springer, New York.

[Du Bois and Du Bois, 1916] Du Bois, D. and Du Bois, E. (1916). A formula to estimate the approximate surface area if height and weight be known. *Arch Intern Med*, 17:863–871.

[Fahrmeir et al., 2016] Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I., and Tutz, G. (2016). *Statistik. Der Weg zur Datenanalyse*. Springer, Berlin.

[Fahrmeir et al., 2008] Fahrmeir, L., Kneib, T., and Lang, S. (2008). *Statistik und ihre Anwendungen- Regression- Modelle, Methoden und Anwendungen*. Springer, Berlin.

[Falk et al., 2014] Falk, M., Hain, J., Marohn, F., Fischer, H., and Michel, R. (2014). *Statistik in Theorie und Praxis*. Springer, Berlin.

[Faraway, 2009] Faraway, J. J. (2009). *Linear Models with R*. Taylor & Francis, Boca Raton.

[Fletcher et al., 2013] Fletcher, G. F., Ades, P. A., Kligfield, P., Arena, R., Balady, G. J., Bittner, V. A., Coke, L. A., Fleg, J. L., Forman, D. E., Gerber, T. C., Gulati, M., Madan, K., Rhodes, J., Thompson, P. D., and Williams, M. A. (2013). Exercise standards for testing and training: A scientific statement from the American heart association. *Circulation*, 128(8):873–934.

[Gareth et al., 2014] Gareth, J., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning*. Springer, New York.

[Han et al., 1995] Han, T. S., van Leer, E. M., Seidell, J. C., and Lean, M. E. (1995). Waist circumference action levels in the identification of cardiovascular risk factors: prevalence study in a random sample. *BMJ (Clinical research ed.)*, 311(7017):1401–1405.

[Jones et al., 1985] Jones, N. L., Makrides, L., Hitchcock, C., Chypchar, T., and McCartney, N. (1985). Normal standards for an incremental progressive cycle ergometer test. *The American review of respiratory disease*, 131(5):700–708.

[Kleinbaum et al., 1998] Kleinbaum, D. G., Kupper, L. L., Muller, K. E., and Nizam, A. (1998). *Applied Regression Analysis and Other Multivariable Methods*. Duxbury Press, Boston.

[Koch et al., 2009] Koch, B., Schäper, C., Ittermann, T., Spielhagen, T., Dörr, M., Völzke, H., Opitz, C. F., Ewert, R., and Gläser, S. (2009). Reference values for cardiopulmonary exercise testing in healthy volunteers: The SHIP study. *European Respiratory Journal*, 33(2):389–397.

[Kohn and Öztürk, 2013] Kohn, W. and Öztürk, R. (2013). *Statistik für Ökonomen - Datenanalyse mit R und SPSS*. Springer, Berlin.

[Luks et al., 203] Luks, A. M., Glenny, R. W., and Robertson, H. T. (203). Interpreting the Results of the Cardiopulmonary Exercise Test. In *Introduction to Cardiopulmonary Exercise Testing*. Springer, New York.

[Mohammad et al., 2012] Mohammad, M. M., Dadashpour, S., and Adimi, P. (2012). Predicted values of cardiopulmonary exercise testing in healthy individuals (A pilot study). *Tanaffos*, 11(1):18–25.

[Muche et al., 2011] Muche, R., Lanzinger, S., and Rau, M. (2011). *Medizinische Statistik mit R und Excel*. Springer, Heidelberg.

[Nelson, 2014] Nelson, R. H. (2014). Hyperlipidemia as a Risk Factor for Cardiovascular Disease Robert. *Primary Care*, 40(1):195–211.

[Niebauer et al., 2012] Niebauer, J., Mayr, K., Tschentscher, M., Pokan, R., and Benzer, W. (2012). Outpatient cardiac rehabilitation: The Austrian model. *European Journal of Preventive Cardiology*, 20(3):468–479.

[Niederberger et al., 1974] Niederberger, M., Kubicek, F., and Reiterer, W. (1974). Leitlinien für die Ergometrie. *Acta Medica Austriaca*, 2(2):33–48.

[Nogueira and Pompeu, 2006] Nogueira, F. S. and Pompeu, F. A. (2006). Maximal workload prediction models in the clinical cardio-pulmonary effort test. *Arq Bras Cardiol*, 87(2):137–145.

[Piepoli et al., 2010] Piepoli, M. F., Corra, U., Benzer, W., Bjarnason-Wehrens, B., Dendale, P., Gaita, D., McGee, H., Mendes, M., Niebauer, J., Zwisler, A. D. O., and Schmid, J. P. (2010). Secondary prevention through cardiac rehabilitation: Physical activity counselling and exercise training. *European Heart Journal*, 31(16):1967–1976.

[Pina et al., 1995] Pina, I. L., Balady, G. J., Hanson, P., Labovitz, A. J., Madonna, D. W., and Myers, J. (1995). Guidelines for Clinical Exercise Testing Laboratories. *Circulation*, 91:912–921.

[Price et al., 2016] Price, K. J., Gordon, B. A., Bird, S. R., and Benson, A. C. (2016). A review of guidelines for cardiac rehabilitation exercise programmes: Is there an international consensus? *European Journal of Preventive Cardiology*, 23(16):1715–1733.

[Pruscha, 2015] Pruscha, H. (2015). *Statistisches Methodenbuch*. Springer, Heidelberg.

[R Core Team, 2017] R Core Team (2017). R: A Language and Environment for Statistical Computing. `https://www.r-project.org/`.

[Ramachandran and Tsokos, 2014] Ramachandran, K. M. and Tsokos, C. P. (2014). *Mathematical Statistics with Applications in R*. Elsevier, London.

[Sarkar, 2008] Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. Springer, New York.

[Sheather, 2009] Sheather, S. J. (2009). *A modern approach to Regression with R*. Springer, New York.

[Toutenburg et al., 2009] Toutenburg, H., Schomaker, M., Wißmann, M., and Heumann, C. (2009). *Arbeitsbuch zur deskriptiven und induktiven Statistik*. Springer, Berlin.

[Turner, 2013] Turner, J. (2013). *Maximal Exercise Stress Test*, pages 1200–1201. Springer, New York.

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

[Verbraecken et al., 2006] Verbraecken, J., Van De Heyning, P., De Backer, W., and Van Gaal, L. (2006). Body surface area in normal-weight, overweight, and obese adults. A comparison study. *Metabolism: Clinical and Experimental*, 55(4):515–524.

[Wickham, 2009] Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis.

[Wilkins et al., 2017] Wilkins, E., Wilson, L., Wickramasinghe, K., Bhatnagar, P., Leal, J., Luengo-Fernandez, R., Burns, R., Rayner, M., and Townsend, N. (2017). *European Cardiovascular Disease Statistics*. European Heart Network, Brussels.

[Wollschläger, 2010] Wollschläger, D. (2010). *Grundlagen der Datenanalyse mit R: eine anwendungsorientierte Einführung*. Springer, Heidelberg.

[Wonisch et al., 2008] Wonisch, M., Berent, R., Klicpera, M., Laimer, H., Marko, C., Schwann, H., and Schmid, P. (2008). Praxisleitlinien Ergometrie. 15(Suppl A):3–17.

Appendices

Appendix A

XSL stylesheet for converting the XML files in a CSV file

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/
    Transform">
    <xsl:output method="text" />
    <xsl:template match="/">
    <xsl:choose>
        <xsl:when test="/CardiologyXML/PatientInfo/PID ! = ''">
            <xsl:value-of select="/CardiologyXML/PatientInfo/PID"/>
        </xsl:when>
        <xsl:otherwise>
            <xsl:text>0</xsl:text>
        </xsl:otherwise>
    </xsl:choose>
    <xsl:text>  </xsl:text>
    <xsl:value-of select="/CardiologyXML/ObservationDateTime/Year"/>
    <xsl:text>-</xsl:text>
    <xsl:value-of select="/CardiologyXML/ObservationDateTime/Month"/>
    <xsl:text>-</xsl:text>
    <xsl:value-of select="/CardiologyXML/ObservationDateTime/Day"/>
    <xsl:text> </xsl:text>
    <xsl:value-of select="/CardiologyXML/ObservationDateTime/Hour"/>
    <xsl:text>:</xsl:text>
    <xsl:value-of select="/CardiologyXML/ObservationDateTime/Minute"/>
    <xsl:text>:</xsl:text>
    <xsl:value-of select="/CardiologyXML/ObservationDateTime/Second"/>
    <xsl:text>  </xsl:text>
    <xsl:value-of select="/CardiologyXML/ObservationEndDateTime/Year"/>
    <xsl:text>-</xsl:text>
    <xsl:value-of select="/CardiologyXML/ObservationEndDateTime/Month"/
        >
    <xsl:text>-</xsl:text>
    <xsl:value-of select="/CardiologyXML/ObservationEndDateTime/Day"/>
    <xsl:text> </xsl:text>
    <xsl:value-of select="/CardiologyXML/ObservationEndDateTime/Hour"/>
    <xsl:text>:</xsl:text>
    <xsl:value-of select="/CardiologyXML/ObservationEndDateTime/Minute"
        />
    <xsl:text>:</xsl:text>
    <xsl:value-of select="/CardiologyXML/ObservationEndDateTime/Second"
        />
```

```
37      <xsl:text> </xsl:text>
38      <xsl:choose>
39          <xsl:when test="/CardiologyXML/PatientInfo/Age != ''">
40              <xsl:value-of select="/CardiologyXML/PatientInfo/Age"/>
41          </xsl:when>
42          <xsl:otherwise>
43              <xsl:text>-1</xsl:text>
44          </xsl:otherwise>
45      </xsl:choose>
46      <xsl:text> </xsl:text>
47      <xsl:value-of select="/CardiologyXML/PatientInfo/Gender"/>
48      <xsl:text> </xsl:text>
49      <xsl:choose>
50          <xsl:when test="/CardiologyXML/PatientInfo/Height != ''">
51              <xsl:value-of select="/CardiologyXML/PatientInfo/Height"/>
52          </xsl:when>
53          <xsl:otherwise>
54              <xsl:text>-1</xsl:text>
55          </xsl:otherwise>
56      </xsl:choose>
57      <xsl:text> </xsl:text>
58      <xsl:choose>
59          <xsl:when test="/CardiologyXML/PatientInfo/Weight != ''">
60              <xsl:value-of select="/CardiologyXML/PatientInfo/Weight"/>
61          </xsl:when>
62          <xsl:otherwise>
63              <xsl:text>-1</xsl:text>
64          </xsl:otherwise>
65      </xsl:choose>
66      <xsl:text> </xsl:text>
67      <xsl:value-of select="/CardiologyXML/Interpretation/
            CustomSummaryStatement"/>
68      <xsl:text> </xsl:text>
69      <xsl:value-of select="/CardiologyXML/PatientInfo/PaceMaker"/>
70      <xsl:text> </xsl:text>
71      <xsl:value-of select="/CardiologyXML/Protocol/Device"/>
72      <xsl:text> </xsl:text>
73      <xsl:value-of select="/CardiologyXML/ExerciseMeasurements/
            MaxWorkload"/>
74      <xsl:text> </xsl:text>
75      <xsl:choose>
76          <xsl:when test="/CardiologyXML/TrendData/TrendEntry[PhaseName='
                VORBEL. ' and SystolicBP][1]/SystolicBP != ''">
77            <xsl:value-of select="/CardiologyXML/TrendData/TrendEntry[
                  PhaseName='VORBEL. ' and SystolicBP][1]/SystolicBP"/>
78          </xsl:when>
79          <xsl:otherwise>
80              <xsl:text>-1</xsl:text>
81          </xsl:otherwise>
82      </xsl:choose>
83    <xsl:text> </xsl:text>
84      <xsl:choose>
85          <xsl:when test="/CardiologyXML/TrendData/TrendEntry[PhaseName='
                VORBEL. ' and DiastolicBP][1]/DiastolicBP != ''">
86            <xsl:value-of select="/CardiologyXML/TrendData/TrendEntry[
                  PhaseName='VORBEL. ' and DiastolicBP][1]/DiastolicBP"/>
87          </xsl:when>
88          <xsl:otherwise>
```

```
89              <xsl:text>-1</xsl:text>
90          </xsl:otherwise>
91      </xsl:choose>
92    <xsl:text>  </xsl:text>
93      <xsl:choose>
94          <xsl:when test="/CardiologyXML/ExerciseMeasurements/
                MaxSystolicBP != ''">
95              <xsl:value-of select="/CardiologyXML/ExerciseMeasurements/
                    MaxSystolicBP"/>
96          </xsl:when>
97          <xsl:otherwise>
98              <xsl:text>-1</xsl:text>
99          </xsl:otherwise>
100     </xsl:choose>
101     <xsl:text>  </xsl:text>
102     <xsl:choose>
103         <xsl:when test="/CardiologyXML/ExerciseMeasurements/
                MaxDiastolicBP != ''">
104             <xsl:value-of select="/CardiologyXML/ExerciseMeasurements/
                    MaxDiastolicBP"/>
105         </xsl:when>
106         <xsl:otherwise>
107             <xsl:text>-1</xsl:text>
108         </xsl:otherwise>
109     </xsl:choose>
110     <xsl:text>  </xsl:text>
111     <xsl:choose>
112         <xsl:when test="/CardiologyXML/ExerciseMeasurements/
                RestingStats/RestHR != ''">
113             <xsl:value-of select="/CardiologyXML/ExerciseMeasurements/
                    RestingStats/RestHR"/>
114         </xsl:when>
115         <xsl:otherwise>
116             <xsl:text>-1</xsl:text>
117         </xsl:otherwise>
118     </xsl:choose>
119     <xsl:text>  </xsl:text>
120     <xsl:choose>
121         <xsl:when test="/CardiologyXML/ExerciseMeasurements/
                MaxHeartRate != ''">
122             <xsl:value-of select="/CardiologyXML/ExerciseMeasurements/
                    MaxHeartRate"/>
123         </xsl:when>
124         <xsl:otherwise>
125             <xsl:text>-1</xsl:text>
126         </xsl:otherwise>
127     </xsl:choose>
128     <xsl:text>  </xsl:text>
129     <xsl:value-of select="/CardiologyXML/Interpretation/
            ReasonForTermination"/>
130     <xsl:text>  </xsl:text>
131     <xsl:value-of select="/CardiologyXML/ClinicalInfo/ReasonForStudy"/>
132     <xsl:text>
133     </xsl:text>
134 </xsl:template>
135 </xsl:stylesheet>
```

Appendix B

MySQL script for integrating the CSV files in a MySQL data base

```
use ergometry;

drop table if exists timeseries;
drop table if exists maindata;
drop table if exists diagnosis;
drop table if exists medication;
drop table if exists lab;

SET SQL_SAFE_UPDATES = 0;
SET default_storage_engine=InnoDB;

CREATE TABLE maindata (
  patient_id int not null auto_increment,
  start_time datetime not null default current_timestamp,
  end_time datetime,
  age int(11) null,
  gender text,
  height int(11) null,
  weight int(11) null,
  waist int(11) null,
  pacemaker text null,
  device text null,
  workload_max decimal null,
  systolicbp_rest int(11) null,
  diastolicbp_rest int(11) null,
  systolicbp_max int(11) null,
  diastolicbp_max int(11) null,
  hr_rest int(11) null,
  hr_max int(11) null,
  termination_reason text,
  reason_for_study text,
  special_workload text,
  primary key (patient_id, start_time))
  ENGINE=InnoDB;

CREATE TABLE timeseries (
  patient_id int not null,
  start_time datetime not null default current_timestamp,
  entry_id int not null,
  start_entry time null,
```

```
41    workload decimal null,
42    diastolicbp int(11) null,
43    systolicbp int(11) null,
44    hr int(11) null,
45    cadence int(11) null,
46    primary key (patient_id, start_time, entry_id))
47    ENGINE=InnoDB;
48
49  -- add the foreign key constraint:
50  ALTER TABLE timeseries
51      ADD CONSTRAINT fk_timeseries_maindata
52          foreign key (patient_id, start_time)
53          references maindata (patient_id, start_time)
54          ON DELETE CASCADE;
55
56  -- deleting foreign key constraint
57  -- alter table timeseries drop foreign key fk_timeseries_maindata;
58
59  CREATE TABLE diagnosis (
60    patient_id int not null,
61    entry_id int not null,
62    diagnosis_date date not null,
63    diagnosis text,
64    primary key (patient_id, entry_id))
65    ENGINE=InnoDB;
66
67  CREATE TABLE medication (
68    patient_id int not null,
69    entry_id int not null,
70    medication_date date null,
71    medication text,
72    dosis text,
73    medication_id int(11) null,
74    primary key (patient_id, entry_id))
75    ENGINE=InnoDB;
76
77  CREATE TABLE lab (
78    patient_id int not null,
79    entry_id int not null,
80    lab_date date not null,
81    parameter text,
82    lab_value numeric(6,2) null,
83    evaluation text null,
84    primary key (patient_id, entry_id))
85    ENGINE=InnoDB;
86
87
88  -- add data to maindata:
89  load data infile '/var/lib/mysql/maindata.csv' into table maindata
        CHARACTER SET UTF8 fields terminated by '\t';
90
91  -- add data to timeseries:
92  load data infile '/var/lib/mysql/timeseries.csv' into table timeseries
        fields terminated by '\t';
93
94  -- add data to diagnosis:
95  load data infile '/var/lib/mysql/diagnosis.csv' into table diagnosis
        fields terminated by '\t';
```

```
96
97  -- add data to medication:
98  load data infile '/var/lib/mysql/medication.csv' into table medication
        fields terminated by '\t';
99
100 -- add data to lab:
101 load data infile '/var/lib/mysql/lab.csv' into table lab fields
        terminated by '\t';
102
103 -- test which data sets are in the timeseries but not in the maindata
104 -- select * from timeseries left join maindata on maindata.patient_id=
        timeseries.patient_id and maindata.start_time=timeseries.start_time
         where maindata.patient_id is NULL;
105
106
107 -- setting empty cells (-1) to NULL
108 UPDATE maindata SET patient_id = patient_id, start_time = start_time,
        age = NULL WHERE maindata.age =-1;
109 UPDATE maindata SET patient_id = patient_id, start_time = start_time,
        height = NULL WHERE maindata.height =-1;
110 UPDATE maindata SET patient_id = patient_id, start_time = start_time,
        weight = NULL WHERE maindata.weight =-1;
111 UPDATE maindata SET patient_id = patient_id, start_time = start_time,
        workload_max= NULL WHERE maindata.workload_max =-1;
112 UPDATE maindata SET patient_id = patient_id, start_time = start_time,
        workload_max= NULL WHERE maindata.workload_max =0;
113 UPDATE maindata SET patient_id = patient_id, start_time = start_time,
        systolicbp_rest= NULL WHERE maindata.systolicbp_rest=-1;
114 UPDATE maindata SET patient_id = patient_id, start_time = start_time,
        diastolicbp_rest = NULL WHERE maindata.diastolicbp_rest =-1;
115 UPDATE maindata SET patient_id = patient_id, start_time = start_time,
        systolicbp_max = NULL WHERE maindata.systolicbp_max =-1;
116 UPDATE maindata SET patient_id = patient_id, start_time = start_time,
        diastolicbp_max = NULL WHERE maindata.diastolicbp_max=-1;
117 UPDATE maindata SET patient_id = patient_id, start_time = start_time,
        hr_rest= NULL WHERE maindata.hr_rest=-1;
118 UPDATE maindata SET patient_id = patient_id, start_time = start_time,
        hr_max= NULL WHERE maindata.hr_max=-1;
119 UPDATE maindata SET patient_id = patient_id, start_time = start_time,
        waist= NULL WHERE maindata.waist=-1;
120
121 UPDATE timeseries SET patient_id = patient_id, start_time = start_time,
        workload= NULL WHERE timeseries.workload =-1;
122 UPDATE timeseries SET patient_id = patient_id, start_time = start_time,
         systolicbp= NULL WHERE timeseries.systolicbp=-1;
123 UPDATE timeseries SET patient_id = patient_id, start_time = start_time,
        diastolicbp = NULL WHERE timeseries.diastolicbp =-1;
124 UPDATE timeseries SET patient_id = patient_id, start_time = start_time,
        hr = NULL WHERE timeseries.hr =-1;
125 UPDATE timeseries SET patient_id = patient_id, start_time = start_time,
        cadence = NULL WHERE timeseries.cadence =-1;
126
127 -- deleting data sets
128 delete from maindata where device = 'Treadmill';
129 delete from maindata where workload_max is NULL;
130 delete from maindata where patient_id=3276; -- test ID
131 delete from maindata where workload_max<=20;
132 delete from maindata where termination_reason LIKE '';
```

```
133 delete from maindata where termination_reason='Kreislaufkollaps';
134
135 -- changing decimal to int (ergometer does not have decimal values)
136 ALTER TABLE maindata CHANGE workload_max workload_max INT(11) null;
137 ALTER TABLE timeseries CHANGE workload workload INT(11) null;
138
139 -- deleting the entries of the other tables whichs dates after after
        the last ergometry
140 delete from lab where lab_date > '2017-07-01';
141 delete from medication where medication_date > '2017-07-01';
142 delete from diagnosis where diagnosis_date > '2017-07-01';
143
144 -- adding lab
145 alter table maindata add CHOL int;
146 alter table maindata add e_CHOL Text;
147 alter table maindata add CREA decimal(7,2);
148 alter table maindata add e_CREA Text;
149 alter table maindata add HDL int;
150 alter table maindata add e_HDL Text;
151 alter table maindata add HSRE decimal(7,2);
152 alter table maindata add e_HSRE Text;
153 alter table maindata add LDL decimal(7,2);
154 alter table maindata add e_LDL Text;
155 alter table maindata add NBZ int;
156 alter table maindata add e_NBZ Text;
157 alter table maindata add TRIGL int;
158 alter table maindata add e_TRIGL Text;
```

Appendix C

Java program to integrate the diagnosis in Table maindata

```
1  import java.sql.Connection;
2  import java.sql.Date;
3  import java.sql.DriverManager;
4  import java.sql.PreparedStatement;
5  import java.sql.ResultSet;
6  import java.sql.SQLException;
7  import java.sql.Statement;
8  import java.sql.Timestamp;
9  import java.util.ArrayList;
10 import java.util.HashMap;
11 import java.util.TreeMap;
12 import java.util.regex.Pattern;
13 import java.util.regex.Matcher;
14
15 // Program RefactorDiagnosisDB.java
16 // The colums "infarkt", "khk" and "kardio" will be appended in table
       "maindata".
17 // For each ergometry in table "maindata" the table "diagnosis" will be
        scanned for the strings "infarkt", "khk" oder "kardio" with the
       same patient_id as in table "maindata".
18 // In case of a hit, the particular columne will be marked
19
20 public class RefactorDiagnosisDB
21 {
22   Connection db;
23   boolean test = false;
24
25   private ArrayList<HashMap<String, Object>> maindata;
26   private ArrayList<HashMap<String, Object>> medication;
27   private ArrayList<HashMap<String, Object>> diagnosis;
28
29   public RefactorDiagnosisDB() throws SQLException
30   {
31     try {
32           // The newInstance() call is a work around for some broken
                Java implementations
33
34           Class.forName("com.mysql.jdbc.Driver").newInstance();
35       }
36       catch (Exception ex)
```

```
37          {
38              // handle the error
39          }
40
41      String username = "sqluser";
42      String password = "sqluser";
43      db = DriverManager.getConnection(
44          "jdbc:mysql://localhost:3306/ergometry?autoReconnect=true&
                useSSL=false",
45          username, password);
46
47      maindata = loadAllRows("maindata");
48      medication = loadAllRows("medication");
49      diagnosis = loadAllRows("diagnosis");
50
51  } // end constructor
52
53  ArrayList<HashMap<String, Object>> loadAllRows(String table) throws
        SQLException
54  {
55      PreparedStatement ps = db.prepareStatement("SELECT * FROM " + table
            );
56      ResultSet rs = ps.executeQuery();
57
58      java.sql.ResultSetMetaData md = rs.getMetaData();
59      int columns = md.getColumnCount();
60      ArrayList<HashMap<String, Object>> ret = new ArrayList<>();
61
62      while (rs.next())
63      {
64        HashMap<String, Object> row = new HashMap();
65        ret.add(row);
66        for (int i = 1; i <= columns; i++)
67          {
68              row.put(md.getColumnName(i), rs.getObject(i));
69          }
70      }
71      return ret;
72  } // end ArrayList
73
74
75  public void runMethod() throws SQLException
76  {
77      String[] diagnostic = new String[53] ;
78      diagnostic[0]  = "aorta";
79      diagnostic[1]  = "aorten";
80      diagnostic[2]  = "aorto";
81      diagnostic[3]  = "arrhythmie";
82      diagnostic[4]  = "asystolie";
83      diagnostic[5]  = "av-block";
84      diagnostic[6]  = "blutdruckentgleisung";
85      diagnostic[7]  = "bradykardie";
86      diagnostic[8]  = "cabg";
87      diagnostic[9]  = "cardial";
88      diagnostic[10]  = "cardio";
89      diagnostic[11]  = "cordis";
90      diagnostic[12]  = "coronar";
91      diagnostic[13]  = "diastolisch";
```

```
92      diagnostic[14]  = "endopericardiale";
93      diagnostic[15]  = "erregungsleitungsstörung";
94      diagnostic[16]  = "extrasystol";
95      diagnostic[17]  = "flimmer";
96      diagnostic[18]  = "herz";
97      diagnostic[19]  = "hhk";
98      diagnostic[20]  = "hinterwand";
99      diagnostic[21]  = "hypertonus";
100     diagnostic[22]  = "hypertropie";
101     diagnostic[23]  = "insuffienz";
102     diagnostic[24]  = "kardiovasculär";
103     diagnostic[25]  = "khk";
104     diagnostic[26]  = "klappe";
105     diagnostic[27]  = "koronar";
106     diagnostic[28]  = "ldh";
107     diagnostic[29]  = "linksschenkelblock";
108     diagnostic[30]  = "mins";
109     diagnostic[31]  = "mitral";
110     diagnostic[32]  = "myocardiopathie";
111     diagnostic[33]  = "nstemi";
112     diagnostic[34]  = "pericard";
113     diagnostic[35]  = "pfo";
114     diagnostic[36]  = "ptca";
115     diagnostic[37]  = "rca";
116     diagnostic[38]  = "rhythmopathia";
117     diagnostic[39]  = "rv-reiz";
118     diagnostic[40]  = "sa-block";
119     diagnostic[41]  = "schenkelblock";
120     diagnostic[42]  = "schrittmacher";
121     diagnostic[43]  = "stenose";
122     diagnostic[44]  = "stent";
123     diagnostic[45]  = "sves";
124     diagnostic[46]  = "svt";
125     diagnostic[47]  = "synkope";
126     diagnostic[48]  = "tachycardie";
127     diagnostic[49]  = "vetrik";
128     diagnostic[50]  = "ves";
129     diagnostic[51]  = "vorhof";
130     diagnostic[52]  = "zavk";
131
132     if(!maindata.iterator().next().containsKey("kardio"))
133     {
134       Statement stmt = db.createStatement();
135       stmt.execute("ALTER TABLE maindata ADD COLUMN kardio INT(1)");
136     }
137
138     TreeMap<Integer, ArrayList<HashMap<String, Object >>>
          idxPIDDiagnosis;
139     idxPIDDiagnosis = new TreeMap<>();
140
141     for(HashMap<String, Object> row : diagnosis)
142     {
143       Integer patient_id = (Integer) row.get("patient_id");
144       ArrayList<HashMap<String, Object >> list = idxPIDDiagnosis.get(
            patient_id);
145       if(list == null)
146       {
147         list = new ArrayList<>();
```

```
148         idxPIDDiagnosis.put(patient_id, list);
149       }
150       list.add(row);
151     }
152
153     PreparedStatement cardio = db.prepareStatement("UPDATE maindata SET
          kardio = ? WHERE patient_id = ? AND start_time = ?");
154
155     int cardioCount = 0, no_diag_count = 0;
156     if (test) System.out.println("");
157
158     for(HashMap<String, Object> mainrow : maindata)
159     {
160       Integer patient_id = (Integer) mainrow.get("patient_id");
161       Timestamp start_time = (Timestamp) mainrow.get("start_time");
162
163       ArrayList<HashMap<String, Object>> diaglist = idxPIDDiagnosis.get
            (patient_id);
164
165       if(diaglist == null)
166       {
167         System.err.println(String.format(" Patient %d has no diagnosis"
              , patient_id));
168         no_diag_count++;
169       }
170       else for(HashMap<String, Object> diagnosis : diaglist)
171       {
172         Date diagnosis_date = (Date) diagnosis.get("diagnosis_date");
173         String diags =  (String) diagnosis.get("diagnosis");
174
175         if(diagnosis_date == null || diags == null) continue;
176
177         if (test)
178         {
179           System.out.println("\n Patient       =" + patient_id);
180           System.out.println(" diagnosis     =" + diags);
181           System.out.println(" diagnosis-date=" + diagnosis_date);
182           System.out.println(" Start_time    =" + start_time);
183         }
184
185         if( diagnosis_date.before(start_time))
186         {
187           if( diags.toLowerCase().contains(diagnostic[0].toLowerCase())
                ||
188               diags.toLowerCase().contains(diagnostic[1].toLowerCase())
                    ||
189               diags.toLowerCase().contains(diagnostic[2].toLowerCase())
                    ||
190               diags.toLowerCase().contains(diagnostic[3].toLowerCase())
                    ||
191               diags.toLowerCase().contains(diagnostic[4].toLowerCase())
                    ||
192               diags.toLowerCase().contains(diagnostic[5].toLowerCase())
                    ||
193               diags.toLowerCase().contains(diagnostic[6].toLowerCase())
                    ||
194               diags.toLowerCase().contains(diagnostic[7].toLowerCase())
                    ||
```

```
195                 diags.toLowerCase().contains(diagnostic[8].toLowerCase())
                        ||
196                 diags.toLowerCase().contains(diagnostic[9].toLowerCase())
                        ||
197                 diags.toLowerCase().contains(diagnostic[10].toLowerCase()
                    ) ||
198                 diags.toLowerCase().contains(diagnostic[11].toLowerCase()
                    ) ||
199                 diags.toLowerCase().contains(diagnostic[12].toLowerCase()
                    ) ||
200                 diags.toLowerCase().contains(diagnostic[13].toLowerCase()
                    ) ||
201                 diags.toLowerCase().contains(diagnostic[14].toLowerCase()
                    ) ||
202                 diags.toLowerCase().contains(diagnostic[15].toLowerCase()
                    ) ||
203                 diags.toLowerCase().contains(diagnostic[16].toLowerCase()
                    ) ||
204                 diags.toLowerCase().contains(diagnostic[17].toLowerCase()
                    ) ||
205                 diags.toLowerCase().contains(diagnostic[18].toLowerCase()
                    ) ||
206                 diags.toLowerCase().contains(diagnostic[19].toLowerCase()
                    ) ||
207                 diags.toLowerCase().contains(diagnostic[20].toLowerCase()
                    ) ||
208                 diags.toLowerCase().contains(diagnostic[21].toLowerCase()
                    ) ||
209                 diags.toLowerCase().contains(diagnostic[22].toLowerCase()
                    ) ||
210                 diags.toLowerCase().contains(diagnostic[23].toLowerCase()
                    ) ||
211                 diags.toLowerCase().contains(diagnostic[24].toLowerCase()
                    ) ||
212                 diags.toLowerCase().contains(diagnostic[25].toLowerCase()
                    ) ||
213                 diags.toLowerCase().contains(diagnostic[26].toLowerCase()
                    ) ||
214                 diags.toLowerCase().contains(diagnostic[27].toLowerCase()
                    ) ||
215                 diags.toLowerCase().contains(diagnostic[28].toLowerCase()
                    ) ||
216                 diags.toLowerCase().contains(diagnostic[29].toLowerCase()
                    ) ||
217                 diags.toLowerCase().contains(diagnostic[30].toLowerCase()
                    ) ||
218                 diags.toLowerCase().contains(diagnostic[31].toLowerCase()
                    ) ||
219                 diags.toLowerCase().contains(diagnostic[32].toLowerCase()
                    ) ||
220                 diags.toLowerCase().contains(diagnostic[33].toLowerCase()
                    ) ||
221                 diags.toLowerCase().contains(diagnostic[34].toLowerCase()
                    ) ||
222                 diags.toLowerCase().contains(diagnostic[35].toLowerCase()
                    ) ||
223                 diags.toLowerCase().contains(diagnostic[36].toLowerCase()
                    ) ||
```

```
224             diags.toLowerCase().contains(diagnostic[37].toLowerCase()
                    ) ||
225             diags.toLowerCase().contains(diagnostic[38].toLowerCase()
                    ) ||
226             diags.toLowerCase().contains(diagnostic[39].toLowerCase()
                    ) ||
227             diags.toLowerCase().contains(diagnostic[40].toLowerCase()
                    ) ||
228             diags.toLowerCase().contains(diagnostic[41].toLowerCase()
                    ) ||
229             diags.toLowerCase().contains(diagnostic[42].toLowerCase()
                    ) ||
230             diags.toLowerCase().contains(diagnostic[43].toLowerCase()
                    ) ||
231             diags.toLowerCase().contains(diagnostic[44].toLowerCase()
                    ) ||
232             diags.toLowerCase().contains(diagnostic[45].toLowerCase()
                    ) ||
233             diags.toLowerCase().contains(diagnostic[46].toLowerCase()
                    ) ||
234             diags.toLowerCase().contains(diagnostic[47].toLowerCase()
                    ) ||
235             diags.toLowerCase().contains(diagnostic[48].toLowerCase()
                    ) ||
236             diags.toLowerCase().contains(diagnostic[49].toLowerCase()
                    ) ||
237             diags.toLowerCase().contains(diagnostic[50].toLowerCase()
                    ) ||
238             diags.toLowerCase().contains(diagnostic[51].toLowerCase()
                    ) ||
239             diags.toLowerCase().contains(diagnostic[52].toLowerCase()
                    ) )
240           {
241
242             int updateStatus=-1;
243
244             if (test)
245             {
246               System.out.println("␣patient_id=" + patient_id);
247               System.out.println("␣start_time=" + start_time);
248             }
249
250             cardio.setInt(1, 1);
251             cardio.setInt(2, patient_id);
252             cardio.setTimestamp(3, start_time);
253             updateStatus=cardio.executeUpdate();
254             cardioCount ++;
255
256             if (test) System.out.println("␣Update_Status=" +
                    updateStatus);
257             break;
258           } // end if diags.
259         } // end if diagnostics_date before start_time
260         else
261         {
262           System.out.println(String.format("␣Patient␣%d␣took␣after␣
                  ergo:␣%tF", patient_id, diagnosis_date));
263         } // end if diagnostics_date before start_time
```

```
264         } // end else for HashMap ...
265       } // end for HashMap
266
267       System.out.println(String.format("␣%d␣Patients␣had␣no␣diagnosis",
              no_diag_count ));
268       System.out.println(String.format("␣%d␣Patients␣had␣heart-Problems",
               cardioCount ));
269     } // end Method
270
271     public static void main(String[]  args) throws Exception
272     {
273       RefactorDiagnosisDB dbc = new RefactorDiagnosisDB();
274       dbc.runMethod();
275     }
276   }
```

Appendix D

Java program to integrate the medication in Table maindata

```
1  import java.sql.Connection;
2  import java.sql.Date;
3  import java.sql.DriverManager;
4  import java.sql.PreparedStatement;
5  import java.sql.ResultSet;
6  import java.sql.SQLException;
7  import java.sql.Statement;
8  import java.sql.Timestamp;
9  import java.util.ArrayList;
10 import java.util.HashMap;
11 import java.util.TreeMap;
12 import java.util.regex.Pattern;
13
14 // Program RefactorMedDB.java
15 // The colums "statin" and "betablocker" will be appended in table
       "maindata".
16 // For each ergometry in the table "maindata" the table "medication"
       will be scanned for the substrings "statin" and "betablocker".
17 // The drugs to be considered are stored in the string-arrays "statine"
        and betablocker.
18 // Is the date of the intake of the drug bofore the date of beginning
       of the ergometry, will this be marked in the particular columne.
19
20 public class RefactorMedDB {
21
22   Connection db;
23   boolean test = false; // for testprints
24
25   private ArrayList<HashMap<String, Object>> maindata;
26   private ArrayList<HashMap<String, Object>> medication;
27   private ArrayList<HashMap<String, Object>> diagnosis;
28
29
30 //-----------------------------------------------------------------------
31 // Constructor refactorDB. in main wird mit new das Object dbc erzeugt.
32 // runMethod ist eine Methode des Objectes dbc.
33
34   public RefactorMedDB() throws SQLException
35   {
36
```

```
37      try
38      {
39          // The newInstance() call is a work around for some broken Java
                implementations
40
41         Class.forName("com.mysql.jdbc.Driver").newInstance();
42      } catch (Exception ex)
43      {
44            // handle the error
45      }
46
47      String username = "sqluser";
48      String password = "sqluser";
49      db = DriverManager.getConnection(
50          "jdbc:mysql://localhost:3306/ergometry?autoReconnect=true&
                useSSL=false",
51          username, password);
52
53      maindata = loadAllRows("maindata");
54      medication = loadAllRows("medication");
55      diagnosis = loadAllRows("diagnosis");
56    } // end RefactorMedDB
57
58    ArrayList<HashMap<String, Object>> loadAllRows(String table) throws
         SQLException
59    {
60      PreparedStatement ps = db.prepareStatement("SELECT␣*␣FROM␣" + table
           );
61      ResultSet rs = ps.executeQuery();
62
63      java.sql.ResultSetMetaData md = rs.getMetaData();
64      int columns = md.getColumnCount();
65      ArrayList<HashMap<String, Object>> ret = new ArrayList<>();
66      while (rs.next())
67      {
68        HashMap<String, Object> row = new HashMap();
69        ret.add(row);
70        for (int i = 1; i <= columns; i++)
71        {
72            row.put(md.getColumnName(i), rs.getObject(i));
73        }
74      }
75      return ret;
76    } // end ArrayList
77
78  //-----------------------------------------------------------------
79  // Method runMethod
80
81    public void runMethod() throws SQLException
82    {
83
84  // 18 statines has been found in the Austrian Codex.
85    String[] statine = new String[18];
86    statine[0] = "Atorvadivid";
87    statine[1] = "Atorvalan";
88    statine[2] = "Atorvastatin";
89    statine[3] = "Atozet";
90    statine[4] = "Crestor";
```

```
91    statine[5] = "Fluvastatin";
92    statine[6] = "Gerosim";
93    statine[7] = "Inegy";
94    statine[8] = "Lovastatin";
95    statine[9] = "Nyzoc";
96    statine[10] = "Panchol";
97    statine[11] = "Pravastatin";
98    statine[12] = "Simvastad";
99    statine[13] = "Simvarcana";
100   statine[14] = "Simvastatin";
101   statine[15] = "Simvatin";
102   statine[16] = "Sortis";
103   statine[17] = "Zocord";
104
105   String[] betablocker = new String[29];
106 // Beta blocker without cardioselectivity
107   betablocker[0] = "Carvedilol";
108   betablocker[1] = "Dilatrend";
109   betablocker[2] = "Hemangiol";
110   betablocker[3] = "Inderal";
111   betablocker[4] = "Sotacor";
112   betablocker[5] = "Sotahexal";
113   betablocker[6] = "Sotastad";
114 // Beta blocker with relative cardioselectivity
115   betablocker[7] = "Atehexal";
116   betablocker[8] = "Atenolan";
117   betablocker[9] = "Atenolol";
118   betablocker[10] = "Beloc";
119   betablocker[11] = "Bilokord";
120   betablocker[12] = "Bisocor";
121   betablocker[13] = "Bisoprolol";
122   betablocker[14] = "Bisostad";
123   betablocker[15] = "Brevibloc";
124   betablocker[16] = "Concor";
125   betablocker[17] = "Esmolol";
126   betablocker[18] = "Hypoloc";
127   betablocker[19] = "Lanoc";
128   betablocker[20] = "Metohexal";
129   betablocker[21] = "Metoprolol";
130   betablocker[22] = "Metoprololsuccinat";
131   betablocker[23] = "Nebivolol";
132   betablocker[24] = "Nomexor";
133   betablocker[25] = "Rapibloc";
134   betablocker[26] = "Rivacor";
135   betablocker[27] = "Seloken";
136   betablocker[28] = "Tenormin";
137
138
139   String meds;
140
141     if(!maindata.iterator().next().containsKey("statin"))
142     {
143       Statement stmt = db.createStatement();
144       stmt.execute("ALTER TABLE maindata ADD COLUMN statin INT(1)");
145     }
146
147     if(!maindata.iterator().next().containsKey("betablocker"))
148     {
```

```
149         Statement stmt = db.createStatement();
150         stmt.execute("ALTER␣TABLE␣maindata␣ADD␣COLUMN␣betablocker␣INT(1)"
                );
151       }
152
153       TreeMap<Integer, ArrayList<HashMap<String, Object >>>
              idxPIDMedication;
154       idxPIDMedication = new TreeMap<>();
155
156       for(HashMap<String, Object> row : medication)
157       {
158         Integer patient_id = (Integer) row.get("patient_id");
159         ArrayList<HashMap<String, Object >> list = idxPIDMedication.get(
              patient_id);
160
161         if(list == null)
162         {
163           list = new ArrayList<>();
164           idxPIDMedication.put(patient_id, list);
165         }
166         list.add(row);
167       } // for HashMap ... medication
168
169   // Question marks are placeholder in the SQL-Statement. They will be
      replaced with actual values with the method statim.setInt(ind,
      value)
170
171       PreparedStatement statin = db.prepareStatement("UPDATE␣maindata␣SET
              ␣statin␣=␣?␣WHERE␣patient_id␣=␣?␣AND␣start_time␣=␣?");
172
173       PreparedStatement betablock = db.prepareStatement("UPDATE␣maindata␣
              SET␣betablocker␣=␣?␣WHERE␣patient_id␣=␣?␣AND␣start_time␣=␣?");
174
175       int statinCount = 0, betablockerCount = 0, no_med_count = 0;
176
177       for(HashMap<String, Object> mainrow : maindata)
178       {
179         Integer patient_id = (Integer) mainrow.get("patient_id");
180         Timestamp start_time = (Timestamp) mainrow.get("start_time");
181
182         boolean hasStatin = false, hasBetablocker = false;
183
184         ArrayList<HashMap<String, Object>> medlist = idxPIDMedication.get
              (patient_id);
185
186         if(medlist == null)
187         {
188           no_med_count++;
189         }
190         else for(HashMap<String, Object> medication : medlist)
191         {
192           Date medication_date = (Date) medication.get("medication_date")
                  ;
193           Integer medication_id = (Integer) medication.get("medication_id
                  ");
194           String smedication =  (String) medication.get("medication");
195
196           if ( smedication.indexOf("␣") < 0)
```

```
197         {
198           meds = smedication;
199         }
200         else
201         {
202           meds = smedication.substring(0, smedication.indexOf("␣"));
203         }
204 
205         if (test)
206         System.out.println( "index␣=" + smedication.indexOf("␣")+ "␣" +
                smedication);
207 
208         if(medication_date == null || medication_id == null) continue;
209 
210 // check for statines
211 
212         if( meds.equalsIgnoreCase(statine[0]) ||
213             meds.equalsIgnoreCase(statine[1]) ||
214             meds.equalsIgnoreCase(statine[2]) ||
215             meds.equalsIgnoreCase(statine[3]) ||
216             meds.equalsIgnoreCase(statine[4]) ||
217             meds.equalsIgnoreCase(statine[5]) ||
218             meds.equalsIgnoreCase(statine[6]) ||
219             meds.equalsIgnoreCase(statine[7]) ||
220             meds.equalsIgnoreCase(statine[8]) ||
221             meds.equalsIgnoreCase(statine[9]) ||
222             meds.equalsIgnoreCase(statine[10]) ||
223             meds.equalsIgnoreCase(statine[11]) ||
224             meds.equalsIgnoreCase(statine[12]) ||
225             meds.equalsIgnoreCase(statine[13]) ||
226             meds.equalsIgnoreCase(statine[14]) ||
227             meds.equalsIgnoreCase(statine[15]) ||
228             meds.equalsIgnoreCase(statine[16]) ||
229             meds.equalsIgnoreCase(statine[17]) )
230         {
231 
232           if( medication_date.before(start_time))
233           {
234               hasStatin = true;
235           } else
236           {
237             System.out.println(String.format("Patient␣%d␣took␣after␣
                  ergo:␣%tF", patient_id, medication_date));
238           }
239         } // meds.equalsIgnoreCase(statine ...
240 
241   // check for beta blocker
242 
243         if( meds.equalsIgnoreCase(betablocker[0]) ||
244             meds.equalsIgnoreCase(betablocker[1]) ||
245             meds.equalsIgnoreCase(betablocker[2]) ||
246             meds.equalsIgnoreCase(betablocker[3]) ||
247             meds.equalsIgnoreCase(betablocker[4]) ||
248             meds.equalsIgnoreCase(betablocker[5]) ||
249             meds.equalsIgnoreCase(betablocker[6]) ||
250             meds.equalsIgnoreCase(betablocker[7]) ||
251             meds.equalsIgnoreCase(betablocker[8]) ||
252             meds.equalsIgnoreCase(betablocker[9]) ||
```

```
253             meds.equalsIgnoreCase(betablocker[10]) ||
254             meds.equalsIgnoreCase(betablocker[11]) ||
255             meds.equalsIgnoreCase(betablocker[12]) ||
256             meds.equalsIgnoreCase(betablocker[13]) ||
257             meds.equalsIgnoreCase(betablocker[14]) ||
258             meds.equalsIgnoreCase(betablocker[15]) ||
259             meds.equalsIgnoreCase(betablocker[16]) ||
260             meds.equalsIgnoreCase(betablocker[17]) ||
261             meds.equalsIgnoreCase(betablocker[18]) ||
262             meds.equalsIgnoreCase(betablocker[19]) ||
263             meds.equalsIgnoreCase(betablocker[20]) ||
264             meds.equalsIgnoreCase(betablocker[21]) ||
265             meds.equalsIgnoreCase(betablocker[22]) ||
266             meds.equalsIgnoreCase(betablocker[23]) ||
267             meds.equalsIgnoreCase(betablocker[24]) ||
268             meds.equalsIgnoreCase(betablocker[25]) ||
269             meds.equalsIgnoreCase(betablocker[26]) ||
270             meds.equalsIgnoreCase(betablocker[27]) ||
271             meds.equalsIgnoreCase(betablocker[28]) )
272         {
273
274           if( medication_date.before(start_time))
275           {
276               hasBetablocker = true;
277           } else
278           {
279             System.out.println(String.format("Patient␣%d␣took␣after␣
                  ergo:␣%tF", patient_id, medication_date));
280           }
281         } // meds.equalsIgnoreCase(betablocker ...
282
283       } // else for HashMap ... medlist
284
285       if(hasStatin)
286       {
287         statin.setInt(1, 1);
288         statin.setInt(2, patient_id);
289         statin.setTimestamp(3, start_time);
290         statin.executeUpdate();
291         statinCount ++;
292       }
293       if(hasBetablocker)
294       {
295         betablock.setInt(1, 1);
296         betablock.setInt(2, patient_id);
297         betablock.setTimestamp(3, start_time);
298         betablock.executeUpdate();
299         betablockerCount ++;
300       }
301
302     } // end for hashMap ... maindata
303
304     System.out.println(String.format("%d␣Patients␣took␣nothing",
          no_med_count ));
305     System.out.println(String.format("%d␣Patients␣took␣statin",
          statinCount ));
306     System.out.println(String.format("%d␣Patients␣took␣betablocker",
          betablockerCount ));
```

```
307
308     } // end runMethod
309
310     public static void main(String[] args) throws Exception
311     {
312       RefactorMedDB dbc = new RefactorMedDB();
313       dbc.runMethod();
314     } // end main
315 }   // end class RefactorMedDB
```

Appendix E

Java program to integrate the laboratory in Table maindata

```
1
2  import java.sql.Connection;
3  import java.sql.DriverManager;
4  import java.sql.SQLException;
5  import java.sql.PreparedStatement;
6  import java.sql.ResultSet;
7  import java.sql.Statement;
8  import java.lang.*;
9  import java.io.*;
10 import java.util.*;
11 import java.text.*;
12
13 // Program RefactorLabDB.java
14 // This Programm searches for each ergometry in the table "maindata"
       for labvalues in the table "lab", which have been taken in the
       range +/- 14 days of the start of the ergometry.
15 // Within this range that value will be taken, which is nearest to the
       start of the ergometry.
16 // The scaned value and the corresponding flag will be inserted into
       the table "maindata".
17
18 public class RefactorLabDB
19 {
20   public static void main(String[] args)
21   {
22     Database db = new Database();
23     try
24     {
25       boolean test = true;
26       db.connect();      // open database
27       System.out.println("");
28       if (test) {System.out.println("␣␣Die␣Datenbank␣ist␣verbunden␣!")
             ;}
29
30       db.read();       // evaluate data
31
32     } catch (Exception e)
33     {
34       e.printStackTrace();
35       return;
```

```
36      }
37
38        db.close();
39     }     // end main
40  }    // end class
41
42  //--------------------------------------------------------------------
43
44  class Database
45    {
46      private Connection con = null;
47      private Statement stateMaindata = null;
48      private Statement stateLab = null;
49      private ResultSet reslab = null;
50      private ResultSet resmain = null;
51      private String string, device, sex;
52      private String datum;
53
54      private boolean update = false;
55      private boolean test = false;    // if true, activate testprints
56
57      public void connect() throws Exception
58      {
59         if(con != null) return;
60          try
61          {
62              Class.forName("com.mysql.jdbc.Driver");
63          } catch (ClassNotFoundException e)
64          {
65              throw new Exception("Kein␣MySQL-Driver␣installiert.␣Mit\n␣"
                    +
66              "$␣sudo␣apt-get␣install␣libmysql-java\n" +
67              "den␣Driver␣installieren␣und␣die␣Environment-Variable␣" +
68              "CLASSPATH␣setzen␣mit:\n" +
69              "export␣CLASSPATH=\n" +
70              "/Directory/von/TestDB:/usr/share/java/mysql-connector-java
                    .jar");
71          }
72
73          String connectionURL =
74          "jdbc:mysql://localhost/ergometry?autoReconnect=true&useSSL=
                false";
75
76          con = DriverManager.getConnection(connectionURL, "sqluser", "
                sqluser");
77          return;
78      }    // end connect
79
80  //-------------------------------------------------------------------
81  //
82      public void read() throws Exception
83      {
84        Calendar dateMin, dateMax = Calendar.getInstance();
85        Calendar calMain = Calendar.getInstance();
86        Calendar calLab  = Calendar.getInstance();
87        Calendar calMin  = Calendar.getInstance();
88        Calendar calMax  = Calendar.getInstance();
89        DateFormat format1, format2;
```

```
90        Date dateMain=null, dateLab=null;
91
92        int[] array_CHOL = new int[20];
93        float[] array_CREA = new float[20];
94        int[] array_HDL = new int[20];
95        float[] array_HSRE = new float[20];
96        float[] array_LDL = new float[20];
97        int[] array_NBZ = new int[20];
98        int[] array_TRIGL = new int[20];
99        String[] labWerte = new String[7];
100       String[] flags = new String[7];
101
102       long[] AtimeDiff = new long[20];
103       long[] AtimeDiffMin = new long[20];
104       int[] AindMin = new int[20];
105       int ind=0;
106       long timeDiff=0;
107       long[] timeDiffMin = new long[7];
108       int CHOL=0, CREA=1, HDL=2, HSRE=3, LDL=4, NBZ=5, TRIGL=6;
109       int id=0, pid=0, record_nr=0, noLabData=0;
110
111       format1 = new SimpleDateFormat("yyyy-MM-dd␣HH:mm:ss");
112       format2 = new SimpleDateFormat("yyyy-MM-dd");
113
114 // timeDiffMin must be initialized with a time-value of 14 days (in ms)
115
116       for (int i=0; i<7; i++)
117       {
118         timeDiffMin[i] = 1210000000;
119         flags[i] = null;
120       }
121
122       try
123       {
124         stateMaindata = con.createStatement();
125         stateLab = con.createStatement();
126
127         if (stateMaindata != null)
128         {
129           System.out.println("␣␣Das␣SQL-Object␣wurde␣instantiiert\n");
130         }
131
132         resmain = stateMaindata.executeQuery("SELECT␣*␣FROM␣maindata;")
              ;
133
134 // With object.next() the focus will be set to the next line.
135 // Within a record the data will be read by her columne-name possible
    methods are:
136 // obj.getString("name"), obj.GetInt("name")), obj.getFloat("name")
137
138         record_nr = 0;
139         while (resmain.next())
140         {
141         record_nr = record_nr + 1;
142         if (test) {System.out.println("␣␣Patientenrecord␣" + record_nr)
              ;}
143
144         int pidm = resmain.getInt("patient_id");
```

```
145         String datumMain = resmain.getString("start_time");
146
147 // To be able to calculate with a date, it must be existent in the
    Calendar-Format.
148 // The conversion of a string into the Calendar-format has to be done
    via the Date-format: String -> Date-format -> Calendar-format.
149 // With the method getTime() the type Calendar will be converted into
    the type Date.
150 // Example for Calendar to Date: Date dateM = calMax.getTime(string);
151 // Example for Date to Calendar: Calendar calMax.setTime(dateMain);
152 // Only the type Date can be converted into a String. This string is
    necessary in the SQL-SELECT-Statement for comparison of time .
153
154         dateMain = format1.parse(datumMain); // string to Date-format
155         calMax.setTime(dateMain);       // Date+format to Calendar-
              format
156         calMin.setTime(dateMain);       // Date+format to Calendar-
              format
157
158         calMax.add(Calendar.DAY_OF_MONTH, 14);
159         calMin.add(Calendar.DAY_OF_MONTH, -14);
160
161         String sdateMin = format2.format(calMin.getTime());
162         String sdateMax = format2.format(calMax.getTime());
163
164 // The variables for the lab-data must be initialized for each
    ergometry
165         if (test) {System.out.println("  Werte werden initialisiert");}
166
167         for (int i=0; i<7; i++)
168         {
169           timeDiffMin[i] = 1210000000;
170           flags[i] = null;
171           labWerte[i] = null;
172         }
173
174         if (test)
175         {
176           System.out.println("  Suchbereich  " + sdateMin + " bis " +
              sdateMax);
177           System.out.println("");
178           System.out.println("  patient_id           = " + pidm );
179           System.out.println("  Zeit des Pat.Records= " + datumMain);
180           System.out.println("");
181           System.out.println("  Tabelle lab mit: where patient_id=pid")
              ;
182         }
183
184 // The time-strings in the SELECT-Statement must be enclosed by single
    quotes ('yyyy-MM-dd')
185
186         String sel1 = "SELECT * FROM lab WHERE patient_id=" + pidm;
187         String sel2 = " and lab_date > " + "'" + sdateMin + "'";
188         String sel3 = " and lab_date < " + "'" + sdateMax + "'";
189
190         String selectstring = sel1  +  sel2 + sel3 ;
191
192         if (test) {System.out.println("  Select-String=\n" +
```

```
            selectstring);}
193
194 // reslab is an objekt of Typ ResultSet, in which the results will be
        stored.
195
196         reslab = stateLab.executeQuery(selectstring);
197
198         ind=0;
199         update=false;
200
201         while (reslab.next())
202         {
203           ind=ind+1;
204           if (test) {System.out.println("\n  nächster Record, ind=" +
                ind);}
205           String datumlab = reslab.getString("lab_date");
206           dateLab = (Date)format2.parse(datumlab);
207           calLab.setTime(dateLab);
208
209           timeDiff = Math.abs(dateMain.getTime() - dateLab.getTime() );
210
211           id = reslab.getInt("patient_id");
212           pid = reslab.getInt("entry_id");
213           String parameter = reslab.getString("parameter");
214           String wert = reslab.getString("lab_value");
215           String flag = reslab.getString("evaluation");
216
217           if (test)
218           {
219             System.out.println("  Zeitoffset in ms=" + timeDiff);
220             System.out.println("  " + pid+ "\t" + id +
221               "\t" + datumlab + "\t" + parameter + "\t" + wert + "\t" +
                      flag);
222           }
223
224             switch (parameter)
225             {
226               case "CHOL":
227               if (timeDiff < timeDiffMin[CHOL])
228               {
229                 labWerte[CHOL] = wert;
230                 flags[CHOL]="'" + flag + "'";
231                 timeDiffMin[CHOL] = timeDiff;
232                 update = true;
233               }
234                 break;
235               case "CREA":
236               if (timeDiff < timeDiffMin[CREA])
237               {
238                 labWerte[CREA] = wert;
239                 flags[CREA]="'" + flag + "'";
240                 timeDiffMin[CREA] = timeDiff;
241                 update = true;
242               }
243                 break;
244               case "HDL":
245               if (timeDiff < timeDiffMin[HDL])
246               {
```

```
247                 labWerte[HDL] = wert;
248                 flags[HDL]="'" + flag + "'";
249                 timeDiffMin[HDL] = timeDiff;
250                 update = true;
251               }
252                 break;
253               case "HSRE":
254               if (timeDiff < timeDiffMin[HSRE])
255               {
256                 labWerte[HSRE] = wert;
257                 flags[HSRE]="'" + flag + "'";
258                 timeDiffMin[HSRE] = timeDiff;
259                 update = true;
260               }
261                 break;
262               case "LDL":
263               if (timeDiff < timeDiffMin[LDL])
264               {
265                 labWerte[LDL] = wert;
266                 flags[LDL]="'" + flag + "'";
267                 timeDiffMin[LDL] = timeDiff;
268                 update = true;
269               }
270                 break;
271               case "NBZ":
272               if (timeDiff < timeDiffMin[NBZ])
273               {
274                 labWerte[NBZ] = wert;
275                 flags[NBZ]="'" + flag + "'";
276                 timeDiffMin[NBZ] = timeDiff;
277                 update = true;
278               }
279                 break;
280               case "TRIGL":
281               if (timeDiff < timeDiffMin[TRIGL])
282               {
283                 labWerte[TRIGL] = wert;
284                 flags[TRIGL]="'" + flag + "'";
285                 timeDiffMin[TRIGL] = timeDiff;
286                 update = true;
287               }
288                     break;
289             }   // end switch
290           }   // end while lab
291
292           if (test)
293           {
294             System.out.println("\n  Aktuellste Laborwerte");
295             System.out.println("\n  Wert       :" + labWerte[0]);
296             System.out.println("  Evaluation " + flags[0]);
297             System.out.println("  Tagesoffset:" + timeDiffMin
                    [0]/(24*3600*1000));
298
299             System.out.println("\n  Wert       :" + labWerte[1]);
300             System.out.println("  Evaluation :" + flags[1]);
301             System.out.println("  Tagesoffset:" + timeDiffMin
                    [1]/(24*3600*1000));
302
```

```
303             System.out.println("\n  Wert       :" + labWerte[2]);
304             System.out.println("  Evaluation :" + flags[2]);
305             System.out.println("  Tagesoffset:" + timeDiffMin
                    [2]/(24*3600*1000));
306
307             System.out.println("\n  Wert       :" + labWerte[3]);
308             System.out.println("  Evaluation :" + flags[3]);
309             System.out.println("  Tagesoffset:" + timeDiffMin
                    [3]/(24*3600*1000));
310
311             System.out.println("\n  Wert       :" + labWerte[4]);
312             System.out.println("  Evaluation :" + flags[4]);
313             System.out.println("  Tagesoffset:" + timeDiffMin
                    [4]/(24*3600*1000));
314
315             System.out.println("\n  Wert       :" + labWerte[5]);
316             System.out.println("  Evaluation :" + flags[5]);
317             System.out.println("  Tagesoffset:" + timeDiffMin
                    [5]/(24*3600*1000));
318
319             System.out.println("\n  Wert       :" + labWerte[6]);
320             System.out.println("  Evaluation :" + flags[6]);
321             System.out.println("  Tagesoffset:" + timeDiffMin
                    [6]/(24*3600*1000));
322           }
323
324         String selectstringu1 = "UPDATE maindata SET";
325         String selStringu11= " CHOL = "     + labWerte[0] + ", e_CHOL =
                ";
326         String selStringu21= ", CREA = "   + labWerte[1] + ", e_CREA = "
                ;
327         String selStringu31= ", HDL = "    + labWerte[2] + ", e_HDL = ";
328         String selStringu41= ", HSRE = "   + labWerte[3] + ", e_HSRE = "
                ;
329         String selStringu51= ", LDL = "    + labWerte[4] + ", e_LDL = ";
330         String selStringu61= ", NBZ = "    + labWerte[5] + ", e_NBZ = ";
331         String selStringu71= ", TRIGL = "  + labWerte[6] + ", e_TRIGL =
                ";
332         String selectstringu8 = " where patient_id = " + id + ";";
333
334         for (int i=0; i<7; i++)
335         {
336           System.out.println("flags(i)=" + flags[i]);
337           if (flags[i] == "null") {flags[i]="NULL";};
338         }
339
340         String selectstringUpdate = selectstringu1 +
341                                     selStringu11 + flags[0] +
342                                     selStringu21 + flags[1] +
343                                     selStringu31 + flags[2] +
344                                     selStringu41 + flags[3] +
345                                     selStringu51 + flags[4] +
346                                     selStringu61 + flags[5] +
347                                     selStringu71 + flags[6] +
348                                     selectstringu8;
349
350         if (test)
351         {System.out.println("\n  Selectstring für das Update der
```

```
            Laborwerte=\n" +              selectstringUpdate);
352         }
353
354         if (update)
355         {
356           System.out.println("      Labordaten bei patient_id= " + pidm
                );
357           int status = stateLab.executeUpdate(selectstringUpdate);
358         } else
359         {
360           noLabData++;
361           System.out.println("Keine Labordaten bei patient_id= " + pidm
                );
362         }
363
364         }      // end while maindata
365       }   // end try
366
367       catch (Exception e)
368       {
369         throw e;
370       }
371
372       System.out.println("\nAnzahl Ergometrie-Datensätze         =" +
            record_nr);
373       System.out.println("Anz. Ergometrie-Datensätze ohne Labor=" +
            noLabData);
374       System.out.println("\n");
375       return;
376     } // end function read
377
378 //-------------------------------------------------------------------
379
380     public void close()
381     {
382       if(con != null)
383       {
384         try
385         {
386             con.close();
387         } catch (SQLException e)
388         {
389             e.printStackTrace();
390         }
391       }
392     }   // end close
393   }   // end class Database
```

Appendix F

R script for adjusting the tables

```
#reading the data form the data base
library(DBI)
mydb = dbConnect(RMySQL::MySQL(), user='sqluser', password='sqluser',
    dbname='ergometry', host='127.0.0.1')
maindata <- dbReadTable(conn = mydb, name = 'maindata')

#calculating additional variables
maindata$bmi<-maindata$weight/(maindata$height/100)^2
maindata$waist_to_height_ratio<-maindata$waist/maindata$height
maindata$workload_to_weight_ratio<-maindata$workload/maindata$weight
maindata$TCtoHDL<-maindata$CHOL/maindata$HDL
maindata$LDLtoHDL<-maindata$LDL/maindata$HDL
maindata$TGtoHDL<-maindata$TRIGL/maindata$HDL
maindata$date <- as.Date(maindata$start_time,'%Y-%m-%d')
maindata$year <- format(as.Date(maindata$start_time, format="%Y-%m-%d")
    ,"%Y")
maindata$surface<-0.007148*maindata$weight^0.425*maindata$height^0.725
maindata$reg_paper_men<-6.773+136.141*maindata$surface -0.916*maindata$
    surface*maindata$age
maindata$reg_paper_women<-3.933+86.641*maindata$surface -0.346*maindata$
    surface*maindata$age

#creating values for regression from paper
for (i in 1:dim(maindata)[1]){
  maindata$reg_mod1[i]<-NA
}

for (i in 1:dim(maindata)[1]){
  if (maindata$gender[i]=='MALE'){
    maindata$reg_mod1[i]<-93.19284+105.67996*maindata$surface[i
        ]-1.09894*maindata$surface[i]*maindata$age[i]
  }
  if (maindata$gender[i]=='FEMALE'){
    maindata$reg_mod1[i]<-67.2302+83.0788*maindata$surface[i]-0.9214*
        maindata$surface[i]*maindata$age[i]
  }
}

#creating column for medication
for (i in 1:dim(maindata)[1]){
    maindata$medication[i]<-NA
}
```

```
37
38 for (i in 1:dim(maindata)[1]){
39   if (is.na(maindata$statin[i])&&is.na(maindata$betablocker[i])){
40     maindata$medication[i]<-0
41   }
42   if (!is.na(maindata$statin[i])){
43     maindata$medication[i]<-1
44   }
45   if (!is.na(maindata$betablocker[i])){
46     maindata$medication[i]<-2
47   }
48   if (!is.na(maindata$statin[i])&&!is.na(maindata$betablocker[i])){
49     maindata$medication[i]<-3
50   }
51 }
52
53 #creating binary column for medication
54 for (i in 1:dim(maindata)[1]){
55   maindata$medication_binary[i]<-NA
56 }
57
58 for (i in 1:dim(maindata)[1]){
59   if (maindata$medication[i]==0){
60     maindata$medication_binary[i]<-0
61   }
62   else {
63     maindata$medication_binary[i]<-1
64   }
65 }
66
67 #creating column for diagnosis
68 for (i in 1:dim(maindata)[1]){
69   maindata$diagnosis_binary[i]<-NA
70 }
71
72 for (i in 1:dim(maindata)[1]){
73   if (is.na(maindata$kardio[i])){
74     maindata$diagnosis_binary[i]<-0
75   }
76   else{
77     maindata$diagnosis_binary[i]<-1
78   }
79 }
80
81 #data for rehabilitation
82 data_rehab <- maindata[ which((maindata$reason_for_study=='Phase␣II␣EU'
     )|
83                                 (maindata$reason_for_study=='Phase␣II␣
                                     AU')|
84                                 (maindata$reason_for_study=='Phase␣III␣
                                     EU')|
85                                 (maindata$reason_for_study=='Phase␣III␣
                                     ZU' )|
86                                 (maindata$reason_for_study=='Phase␣III␣
                                     AU')),]
87
88 #creating data for stage 2
89 data_stage2 <- subset(maindata,(maindata$reason_for_study=='Phase␣II␣EU
```

```
    '|
90                                          maindata$reason_for_study=='Phase␣II␣
                                              AU'))
91 dim(data_stage2)[1]
92 data_stage2_duration <-  data.frame(patient_id = unique(data_stage2$
       patient_id))
93 for (i in 1:dim(data_stage2_duration)[1]){
94   data_stage2_duration$date_EU[i]<-NA
95   data_stage2_duration$date_AU[i]<-NA
96   data_stage2_duration$duration[i]<-NA
97   data_stage2_duration$workload_max_EU[i]<-NA
98   data_stage2_duration$gender[i]<-NA
99   data_stage2_duration$age[i]<-NA
100  data_stage2_duration$weight[i]<-NA
101  data_stage2_duration$height[i]<-NA
102  data_stage2_duration$waist[i]<-NA
103  data_stage2_duration$bmi[i]<-NA
104  data_stage2_duration$waist_to_height_ratio[i]<-NA
105  data_stage2_duration$workload_to_weight_ratio[i]<-NA
106  data_stage2_duration$systolicbp_rest[i]<-NA
107  data_stage2_duration$diastolicbp_rest[i]<-NA
108  data_stage2_duration$hr_rest[i]<-NA
109  data_stage2_duration$workload_max_AU[i]<-NA
110  data_stage2_duration$CHOL[i]<-NA
111  data_stage2_duration$e_CHOL[i]<-NA
112  data_stage2_duration$CREA[i]<-NA
113  data_stage2_duration$e_CREA[i]<-NA
114  data_stage2_duration$HDL[i]<-NA
115  data_stage2_duration$e_HDL[i]<-NA
116  data_stage2_duration$LDL[i]<-NA
117  data_stage2_duration$e_LDL[i]<-NA
118  data_stage2_duration$HSRE[i]<-NA
119  data_stage2_duration$e_HSRE[i]<-NA
120  data_stage2_duration$NBZ[i]<-NA
121  data_stage2_duration$e_NBZ[i]<-NA
122  data_stage2_duration$TRIGL[i]<-NA
123  data_stage2_duration$e_TRIGL[i]<-NA
124
125  for (j in 1:dim(data_stage2)[1]){
126    if (data_stage2$patient_id[j]==data_stage2_duration$patient_id[i] &
         & data_stage2$reason_for_study[j]=='Phase␣II␣EU' && is.na(data_
         stage2_duration$date_EU[j])){
127      data_stage2_duration$date_EU[i]<-data_stage2$date[j]
128      data_stage2_duration$workload_max_EU[i]<-data_stage2$workload_max
           [j]
129      data_stage2_duration$gender[i]<-data_stage2$gender[j]
130      data_stage2_duration$age[i]<-data_stage2$age[j]
131      data_stage2_duration$weight[i]<-data_stage2$weight[j]
132      data_stage2_duration$height[i]<-data_stage2$height[j]
133      data_stage2_duration$waist[i]<-data_stage2$waist[j]
134      data_stage2_duration$bmi[i]<-data_stage2$bmi[j]
135      data_stage2_duration$waist_to_height_ratio[i]<-data_stage2$waist_
           to_height_ratio[j]
136      data_stage2_duration$workload_to_weight_ratio[i]<-data_stage2$
           workload_to_weight_ratio[j]
137      data_stage2_duration$systolicbp_rest[i]<-data_stage2$systolicbp_
           rest[j]
138      data_stage2_duration$diastolicbp_rest[i]<-data_stage2$diastolicbp
```

```
          _rest[j]
139       data_stage2_duration$hr_rest[i]<-data_stage2$hr_rest[j]
140       data_stage2_duration$CHOL[i]<-data_stage2$CHOL[j]
141       data_stage2_duration$e_CHOL[i]<-data_stage2$e_CHOL[j]
142       data_stage2_duration$CREA[i]<-data_stage2$CREA[j]
143       data_stage2_duration$e_CREA[i]<-data_stage2$e_CREA[j]
144       data_stage2_duration$HDL[i]<-data_stage2$HDL[j]
145       data_stage2_duration$e_HDL[i]<-data_stage2$e_HDL[j]
146       data_stage2_duration$LDL[i]<-data_stage2$LDL[j]
147       data_stage2_duration$e_LDL[i]<-data_stage2$e_LDL[j]
148       data_stage2_duration$HSRE[i]<-data_stage2$HSRE[j]
149       data_stage2_duration$e_HSRE[i]<-data_stage2$e_HSRE[j]
150       data_stage2_duration$NBZ[i]<-data_stage2$NBZ[j]
151       data_stage2_duration$e_NBZ[i]<-data_stage2$e_NBZ[j]
152       data_stage2_duration$TRIGL[i]<-data_stage2$TRIGL[j]
153       data_stage2_duration$e_TRIGL[i]<-data_stage2$e_TRIGL[j]
154     }
155   }
156   for (j in 1:dim(data_stage2)[1]){
157     if (data_stage2$patient_id[j]==data_stage2_duration$patient_id[i] &
          & data_stage2$reason_for_study[j]=='Phase␣II␣AU'&& is.na(data_
          stage2_duration$date_AU[j])){
158       data_stage2_duration$date_AU[i]<-data_stage2$date[j]
159       data_stage2_duration$workload_max_AU[i]<-data_stage2$workload_max
            [j]
160       if(is.na(data_stage2_duration$gender[i])){
161         data_stage2_duration$gender[i]<-data_stage2$gender[j]
162       }
163     }
164   }
165   data_stage2_duration$duration[i]<-data_stage2_duration$date_AU[i]-
        data_stage2_duration$date_EU[i]
166   data_stage2_duration$improvement[i]<-data_stage2_duration$workload_
        max_AU[i]-data_stage2_duration$workload_max_EU[i]
167 }
168
169 # creating data for stage 3
170 data_stage3 <- subset(maindata,(maindata$reason_for_study=='Phase␣III␣
      EU'| maindata$reason_for_study=='Phase␣III␣ZU'| maindata$reason_for
      _study=='Phase␣III␣AU'))
171   dim(data_stage3)[1]
172   data_stage3_duration <-
173       data.frame(patient_id = unique(data_stage3$patient_id))
174   dim(data_stage3_duration)[1]
175   for (i in 1:dim(data_stage3_duration)[1]){
176     data_stage3_duration$date_EU[i]<-NA
177     data_stage3_duration$date_ZU[i]<-NA
178     data_stage3_duration$date_AU[i]<-NA
179     data_stage3_duration$duration[i]<-NA
180   data_stage3_duration$duration1[i]<-NA
181   data_stage3_duration$duration2[i]<-NA
182   data_stage3_duration$workload_max_EU[i]<-NA
183   data_stage3_duration$workload_max_ZU[i]<-NA
184   data_stage3_duration$workload_max_AU[i]<-NA
185   data_stage3_duration$gender[i]<-NA
186   data_stage3_duration$age[i]<-NA
187   data_stage3_duration$weight[i]<-NA
188   data_stage3_duration$height[i]<-NA
```

```
189   data_stage3_duration$waist[i]<-NA
190   data_stage3_duration$bmi[i]<-NA
191   data_stage3_duration$waist_to_height_ratio[i]<-NA
192   data_stage3_duration$workload_to_weight_ratio[i]<-NA
193   data_stage3_duration$systolicbp_rest[i]<-NA
194   data_stage3_duration$diastolicbp_rest[i]<-NA
195   data_stage3_duration$hr_rest[i]<-NA
196   data_stage3_duration$workload_max_AU[i]<-NA
197   data_stage3_duration$CHOL[i]<-NA
198   data_stage3_duration$e_CHOL[i]<-NA
199   data_stage3_duration$CREA[i]<-NA
200   data_stage3_duration$e_CREA[i]<-NA
201   data_stage3_duration$HDL[i]<-NA
202   data_stage3_duration$e_HDL[i]<-NA
203   data_stage3_duration$LDL[i]<-NA
204   data_stage3_duration$e_LDL[i]<-NA
205   data_stage3_duration$HSRE[i]<-NA
206   data_stage3_duration$e_HSRE[i]<-NA
207   data_stage3_duration$NBZ[i]<-NA
208   data_stage3_duration$e_NBZ[i]<-NA
209   data_stage3_duration$TRIGL[i]<-NA
210   data_stage3_duration$e_TRIGL[i]<-NA
211
212   for (j in 1:dim(data_stage3)[1]){
213     if (data_stage3$patient_id[j]==data_stage3_duration$patient_id[i] &
          & data_stage3$reason_for_study[j]=='Phase␣III␣EU'){
214      data_stage3_duration$date_EU[i]<-data_stage3$date[j]
215 data_stage3_duration$workload_max_EU[i]<-data_stage3$workload_max[j]
216 data_stage3_duration$gender[i]<-data_stage3$gender[j]
217 data_stage3_duration$age[i]<-data_stage3$age[j]
218 data_stage3_duration$weight[i]<-data_stage3$weight[j]
219 data_stage3_duration$height[i]<-data_stage3$height[j]
220 data_stage3_duration$waist[i]<-data_stage3$waist[j]
221 data_stage3_duration$bmi[i]<-data_stage3$bmi[j]
222 data_stage3_duration$waist_to_height_ratio[i]<-data_stage3$waist_to_
      height_ratio[j]
223 data_stage3_duration$workload_to_weight_ratio[i]<-data_stage3$workload
      _to_weight_ratio[j]
224 data_stage3_duration$systolicbp_rest[i]<-data_stage3$systolicbp_rest[j
      ]
225 data_stage3_duration$diastolicbp_rest[i]<-data_stage3$diastolicbp_rest
      [j]
226 data_stage3_duration$hr_rest[i]<-data_stage3$hr_rest[j]
227 data_stage3_duration$CHOL[i]<-data_stage3$CHOL[j]
228 data_stage3_duration$e_CHOL[i]<-data_stage3$e_CHOL[j]
229 data_stage3_duration$CREA[i]<-data_stage3$CREA[j]
230 data_stage3_duration$e_CREA[i]<-data_stage3$e_CREA[j]
231 data_stage3_duration$HDL[i]<-data_stage3$HDL[j]
232 data_stage3_duration$e_HDL[i]<-data_stage3$e_HDL[j]
233 data_stage3_duration$LDL[i]<-data_stage3$LDL[j]
234 data_stage3_duration$e_LDL[i]<-data_stage3$e_LDL[j]
235 data_stage3_duration$HSRE[i]<-data_stage3$HSRE[j]
236 data_stage3_duration$e_HSRE[i]<-data_stage3$e_HSRE[j]
237 data_stage3_duration$NBZ[i]<-data_stage3$NBZ[j]
238 data_stage3_duration$e_NBZ[i]<-data_stage3$e_NBZ[j]
239 data_stage3_duration$TRIGL[i]<-data_stage3$TRIGL[j]
240 data_stage3_duration$e_TRIGL[i]<-data_stage3$e_TRIGL[j]
241     }
```

```
242   }
243   for (j in 1:dim(data_stage2)[1]){
244     if (data_stage2$patient_id[j]==data_stage3_duration$patient_id[i] &
          & (is.na(data_stage3_duration$date_EU[j])&&data_stage2$reason_
          for_study[j]=='Phase␣II␣AU')){
245       data_stage3_duration$date_EU[i]<-data_stage2$date[j]
246       data_stage3_duration$workload_max_EU[i]<-data_stage2$workload_max
            [j]
247       data_stage3_duration$gender[i]<-data_stage2$gender[j]
248       data_stage3_duration$age[i]<-data_stage2$age[j]
249       data_stage3_duration$weight[i]<-data_stage2$weight[j]
250       data_stage3_duration$height[i]<-data_stage2$height[j]
251       data_stage3_duration$waist[i]<-data_stage2$waist[j]
252       data_stage3_duration$bmi[i]<-data_stage2$bmi[j]
253       data_stage3_duration$waist_to_height_ratio[i]<-data_stage2$waist_
            to_height_ratio[j]
254       data_stage3_duration$workload_to_weight_ratio[i]<-data_stage2$
            workload_to_weight_ratio[j]
255       data_stage3_duration$systolicbp_rest[i]<-data_stage2$systolicbp_
            rest[j]
256       data_stage3_duration$diastolicbp_rest[i]<-data_stage2$diastolicbp
            _rest[j]
257       data_stage3_duration$hr_rest[i]<-data_stage2$hr_rest[j]
258       data_stage3_duration$CHOL[i]<-data_stage2$CHOL[j]
259       data_stage3_duration$e_CHOL[i]<-data_stage2$e_CHOL[j]
260       data_stage3_duration$CREA[i]<-data_stage2$CREA[j]
261       data_stage3_duration$e_CREA[i]<-data_stage2$e_CREA[j]
262       data_stage3_duration$HDL[i]<-data_stage2$HDL[j]
263       data_stage3_duration$e_HDL[i]<-data_stage2$e_HDL[j]
264       data_stage3_duration$LDL[i]<-data_stage2$LDL[j]
265       data_stage3_duration$e_LDL[i]<-data_stage2$e_LDL[j]
266       data_stage3_duration$HSRE[i]<-data_stage2$HSRE[j]
267       data_stage3_duration$e_HSRE[i]<-data_stage2$e_HSRE[j]
268       data_stage3_duration$NBZ[i]<-data_stage2$NBZ[j]
269       data_stage3_duration$e_NBZ[i]<-data_stage2$e_NBZ[j]
270       data_stage3_duration$TRIGL[i]<-data_stage2$TRIGL[j]
271       data_stage3_duration$e_TRIGL[i]<-data_stage2$e_TRIGL[j]
272     }
273   }
274   for (j in 1:dim(data_stage3)[1]){
275     if (data_stage3$patient_id[j]==data_stage3_duration$patient_id[i] &
          & data_stage3$reason_for_study[j]=='Phase␣III␣ZU'){
276       data_stage3_duration$date_ZU[i]<-data_stage3$date[j]
277       data_stage3_duration$workload_max_ZU[i]<-data_stage3$workload_max
            [j]
278
279     }
280   }
281   for (j in 1:dim(data_stage3)[1]){
282     if (data_stage3$patient_id[j]==data_stage3_duration$patient_id[i] &
          & data_stage3$reason_for_study[j]=='Phase␣III␣AU'){
283      data_stage3_duration$date_AU[i]<-data_stage3$date[j]
284   data_stage3_duration$workload_max_AU[i]<-data_stage3$workload_max[j]
285   if (is.na(data_stage3_duration$gender[i])){
286     data_stage3_duration$gender[i]<-data_stage3$gender[j]
287   }
288     }
289   }
```

```
290     data_stage3_duration$duration[i]<-data_stage3_duration$date_AU[i]-
            data_stage3_duration$date_EU[i]
291     data_stage3_duration$duration1[i]<-data_stage3_duration$date_ZU[i]-
            data_stage3_duration$date_EU[i]
292     data_stage3_duration$duration2[i]<-data_stage3_duration$date_AU[i]-
            data_stage3_duration$date_ZU[i]
293     data_stage3_duration$improvement[i]<-data_stage3_duration$workload_
            max_AU[i]-data_stage3_duration$workload_max_EU[i]
294     data_stage3_duration$improvement1[i]<-data_stage3_duration$workload_
            max_ZU[i]-data_stage3_duration$workload_max_EU[i]
295     data_stage3_duration$improvement2[i]<-data_stage3_duration$workload_
            max_AU[i]-data_stage3_duration$workload_max_ZU[i]
296 }
```