POLITECNICO DI MILANO
Master Degree in Mathematical Engineering
Dipartimento di Ingegneria Matematica

TIME SERIES ANALYSIS TO DETECT CRISIS IN SYSTEMIC CAPILLARY LEAK SYNDROME

Relatore: Prof.ssa Alessandra Guglielmi
Correlatore: Dott.ssa Ilenia Epifani
Prof. Giancarlo Ripamonti

Tesi di Laurea di:
Marija Zdolsek, matricola 854414

Accademic Year 2017/18

A big thank you to my mother, my father and my sister, without their unconditionas support I would not be here today.
To all my friends from Campus Martinitt (especially Alfredo, Jelena, Dajana, Alex, Martina, Jalal, Bence and Kasra). Thank you for sharing the most important experiences with me in those three years, I love you.

Abstract

Systemic capillary leak syndrome (SCLS) is a rare disorder characterized by repeated flares of massive leakage of plasma from blood vessels into neighbouring body cavities and muscles, which results in a sharp decrease in blood pressure, that can lead to organ failure and death, and typically an increase in body weight.

The available data analysed in this thesis contains information about a single patient having this disease. Some of the variables contained in dataset are the daily body weight, diastolic and systolic pressure, heart rate of patient. Beside this, there are some information that are collected while the patient was hospitalized, including the type of crisis that the patient had as well as levels of hemoglobin, hematocrit and neutrophils in the patient's blood.

The aim of this study is to obtain a statistical description af the dataset, and appropriate statistical models in order to predict possible future attacks taking in consideration the most significant features, like the body weight of the patient.

Sommario

La sindrome sistemica da aumentata permeabilità capillare (SCLS) è una malattia rara caratterizzata da scoppi ripetuti di fuoriuscite massicce di plasma dai vasi sanguigni nelle vicine cavità del corpo e muscoli, che si traduce in una forte diminuzione della pressione sanguigna, che può portare a insufficienze d'organo e morte. Le crisi possono iniziare con un aumento del peso corporeo.

I dati disponibili analizzati in questa tesi contengono informazioni su un singolo paziente che ha questa malattia. Alcune delle variabili contenute nel set di dati sono il peso corporeo giornaliero, la pressione diastolica e sistolica e la frequenza cardiaca. Altre misurazioni sono state raccolte durante i ricoveri in ospedale del paziente e riguardano il grado di gravità della crisi avuta dal paziente e livelli di emoglobina, ematocrito e neutrofili.

Lo scopo di questo studio è stata un'esplorazione statistica del set di dati e un'analisi di alcuni modelli statistici idonei a prevedere possibili crisi future, sulla base delle caratteristiche più significative, come per esempio il peso corporeo.

Contents

Abstract **I**

1 Autoregressive Moving Average Models for time series **5**
1.1 Stationarity and Strict Stationarity 6
1.2 ARIMA models 8
1.2.1 Autoregressive processes 8
1.2.2 Moving average processes 9
1.3 Stationarity tests 11
1.4 Estimation and Elimination of the Trend and Seasonal Components 13
1.4.1 Elimination of a trend in the absence of seasonality 13
1.4.2 Elimination of both Trend and Seasonality 14
1.5 Model Diagnostic 15

2 Bayesian Statistic for Structural Time Series **18**
2.1 Structural Time Series 19
2.1.1 Local Level Trend 20
2.1.2 Local Linear Trend 20
2.1.3 AR(p) model with time varying coefficients 20
2.2 Spike and Slab Variable selection 21

3 Analysis of data set for person with Capillary Leak Syndrome **24**
3.1 Capillary Leak Syndrome 24
3.2 Analysis of weights 26
3.2.1 Fitting ARIMA model to the first phase of the wight time series 29
3.2.2 Fittig ARIMA Models to the second and third phases of the weight time series 34
3.2.3 Conclusion 42
3.3 Fitting ARIMA models to nine subseries of weight 43
3.4 Seasonality 46
3.5 Analysis of blood pressure and heart beat 48
3.5.1 ARIMA model for heart beat rate 48
3.5.2 ARIMA model for systolic blood pressure 51
3.5.3 ARIMA model for diastolic blood pressure 54

3.5.4 Desriptive analysis of blood pressure and heart beat . . . 56
3.5.5 Conclusion . . . 58
3.6 Analysis of data collected in hospital . . . 59
3.6.1 Descriptive statistics . . . 59
3.6.2 Fitting ARIMA model to full weight time series . . . 66
3.6.3 Conclusion . . . 72

4 Bayesian regression models for the body weight and for the indicator of crisis 75
4.1 Regression models for the body weight . . . 75
4.1.1 Model with only Local Level Trend for weight time series 76
4.1.2 Model with Local Level Trend and $AR(5)$ for weight time series . . . 80
4.1.3 Local Level Trend with $AR(5)$ and linear regression for weight time series . . . 83
4.2 Regression for crisis indicator . . . 86
4.3 Conclusion . . . 87

5 Conclusions and further developments 89

List of Figures

3.1 Time series of the daily body weight in the log scale: the two vertical lines split the time series into the three different subseries. 27
3.2 Time series of the daily body weight regards to time in days: blue lines represent missing parts of time series. 28
3.3 Dates for which values of time series of weights are missing. . . . 28
3.4 Red line represents the first subseries in the log scale and blue line is fit of $ARIMA(0,1,0)$ model. 30
3.5 ACF and PACF of residuals for model $ARIMA(0,1,0)$, for the time series of first phase. 31
3.6 Red line represents the first subseries in log scale and blue line is fit of $ARIMA(0,2,0)$ model 32
3.7 ACF and PACF of residuals for model $ARIMA(0,2,0)$. 33
3.8 Second phase of time series for the weights in log scale: blue dots are estimations of NA values using function *na.interpolation*. . . 35
3.9 Third phase of time series for the weights in log scale: blue dots are estimations of NA values using function *na.interpolation*. . . 35
3.10 Red line represents second subseries in log scale: blue line is the fit of $ARIMA(4,1,0)$ model. 36
3.11 ACF and PACF of residuals for model $ARIMA(4,1,0)$. 37
3.12 Red line represents the second phase data in log scale: blue line is fit of $ARIMA(5,2,0)$ model. 38
3.13 ACF and PACF of estimated residuals for model ARIMA(5,2,0). 38
3.14 Red line represents third part in log scale: blue line is the fit of $ARIMA(4,1,2)$ model. 39
3.15 ACF and PACF of residuals for model $ARIMA(4,1,2)$. 40
3.16 Red line represents the third subseries in log scale: blue line is fit of $ARIMA(5,2,0)$ model. 40
3.17 ACF and PACF of residuals for model ARIMA(5,2,0). 41
3.18 Nine subseries of time series of weight. 44
3.19 Box plots for five subseries in log scale. 46
3.20 Box plots for twenty subseries in log scale: verical line is median of data from 18.11.2013 until 29.08.2016. 47
3.21 Heart beat rate time series. 48
3.22 Red line represents heart rate time series: blue line is the fit of $ARIMA(2,1,4)$ model. 49

3.23 ACF and PACF of estimated residuals for $ARIMA(2,1,4)$ model. 49
3.24 Red line represents heart rate time series: blue line is the fit of $ARIMA(5,2,0)$ model. 50
3.25 ACF and PACF of estimated residuals for $ARIMA(5,2,0)$ model. 50
3.26 Systolic blood pressure time series. 51
3.27 Red line represents systolic blood pressure time series: blue line is the fit of $ARIMA(4,1,3)$ model. 51
3.28 ACF and PACF of estimated residuals for $ARIMA(4,1,3)$ model. 52
3.29 Red line represents systolic blood pressure time series: blue line is the fit of $ARIMA(1,2,2)$ model. 52
3.30 ACF and PACF of estimated residuals for $ARIMA(4,1,3)$ model. 53
3.31 Diastolic blood pressure time series. 54
3.32 ACF and PACF of estimated residuals for $ARIMA(1,1,2)$ model. 55
3.33 ACF and PACF of estimated residuals for $ARIMA(5,2,0)$ model. 55
3.34 Left figure: boxplots of systolic blood pressure; right figure: boxplots of diastolic blood pressure. 56
3.35 Boxplots of heart beat rate 56
3.36 Boxplot of hemoglobin levels during days before crisis, days after crisis and crisis days 60
3.37 Boxplot of hemtrocit levels during days before crisis, days after crisis and crisis days 62
3.38 Boxplot of neutrophils levels during days before crisis, days after crisis and crisis days 64
3.39 Boxplot of absolute count of neutrophils during days before crisis, days after crisis and crisis days 64
3.40 Boxplots of the body weight days before crisis, days after crisis and crisis days 65
3.41 Red line represents new weight time series in log scale: blue line is the fit of $ARIMA(4,1,5)$ model. 66
3.42 ACF and PACF of residuals for model $ARIMA(4,1,5)$. 67
3.43 Red line represents new weight time series in log scale: blue line is fit of $ARIMA(5,2,0)$ model. 68
3.44 ACF and PACF of residuals for model ARIMA(5,2,0). 69
3.45 Red line represents new weight time series in log scale: blue line is fit of $ARIMA(4,1,0)$ model. 70
3.46 ACF and PACF of residuals for model ARIMA(4,1,0). 70
3.47 Red line represents new weight time series in log scale: blue line is fit of $ARIMA(4,1,0)$ model and circles are missing values from Section 3.4.1 71
3.48 Boxplots of the body weight two days before crisis, days before crisis, crisis days and days after crisis. 72

4.1 Local Level Trend model for weight time series 78
4.2 Trend for weight time series 78
4.3 Trace plot for Local Level Trend model 79
4.4 Gelman plot for Local Level Trend model 79

4.5 Local Level Trend with AR(5) model for weight time series . . . 80
4.6 Components of model 81
4.7 Trace plot for Local Level Trend with $AR(5)$ model 81
4.8 Gelman plot for Local Level Trend with $AR(5)$ model 82
4.9 Gelman plot for coefficients of $AR(5)$ model 82
4.10 Geweke plots for $AR(5)$ coefficients 85

List of Tables

3.1 Table of results 42
3.2 Table of results 45
3.3 Table of results 58
3.4 Table of descriptive statistics for hemoglobin 59
3.5 Table of descriptive statistics for hematrocit 61
3.6 Table of descriptive statistics for percentage level of neutrophils . 63
3.7 Table of descriptive statistics for absolute count of neutrophils . 63
3.8 Table of descriptive statistics for the body weight of the patient . 65
3.9 Table of results from Section 3.7.2 72

4.1 Posterior estimates of the regression coefficients 83
4.2 Posterior estimates of the regression coefficients of reduced model 84
4.3 Posterior estimates of $AR(5)$ coefficients 84
4.4 Gewek diagnostic of $AR(5)$ coefficients 85
4.5 Posterior estimatets of regression coefficients in (4.9) 86
4.6 Posterior estimates of regression coefficients of reduced model . . 86
4.7 Descriptive statistics of coefficients for reduced model 87

Introduction

In this paper, we analyse the data on a single patient having Systemic capillary leak syndrome. Systemic capillary leak syndrome (SCLS) is a condition in which fluid and proteins leak out of tiny blood vessels, into surrounding tissues. This can result in dangerously low blood pressure and a decrease in plasma volume. Initial symptoms may include fatigue, nausea, abdominal pain, extreme thirst, and sudden increase in body weight. Episodes of SCLS vary in frequency, with some people having one episode in their lifetime, and others having several per year. The severity also varies, and the condition can be fatal.
The main question that should be answered is whether there exists a statistical relationship between weight and crisis indicator, i.e. can the future attack be predicted by daily measuring body weight?

There are two sources of data used in this work. The first one was collected daily by the patient himself, i.e. he was measuring body weight, diastolic and systolic blood pressure and heart rate, except on the days when the crisis happened and he was hospitalized. The other data are collected by physicians, while the patient was in hospital. Apart from the body weight, diastolic and systolic blood pressure and heart rate, this data contains also information about levels of hemoglobin, hematocrit and neutrophils in patient's blood. Both data collected by the patient and by physicians are for period from 16.10.2013.until 20.06.2017, but we focused on analysis of data from 18.11.2013 until 31.08.2016. In this period many attacks occured and during this period immunoglobulin therapy was applied.

We used two modelling approaches. The model developed for forecasting the body weight, blood pressure and heart rate of the patient is Autoregressive Integrated Moving Average (ARIMA) model, while for finding the relation between the body weight and indicator of crisis Bayesian regression models are used. The first approach was used to check autoregressive type in behaviour of the daily body weight, suggested by physicians. The second approach was adopted to answer the main question of the analysis using Bayesian approach through R packages.

ARIMA models were introduced in Time Series Analysis: Forecasting and Control by Box and Jenkins. The main reason of choosing ARIMA models in this

study for the forecasting the patient's body weight is because they are flexible, and represent stationary as well as non-stationary time series. They also allow for accurate prediction of a future outcome.

The first and the most important step in fitting an ARIMA model to the body weight time series is to determine if the series is stationary and if there is any significant seasonality that needs to be modelled. If data is non-stationary, we should identify the order of differencing needed for it to be stationary, d. Differencing can help in stabilizing the mean of a time series by removing changes in the level of a time series, and so eliminating trend and seasonality. In this study it was suggested to use $d = 1$ and $d = 2$. After a time series has been stationarized and once and seasonality has been addressed by differencing, the next step in fitting an ARIMA model is to determine Autoregressive (AR), p, and Moving average (MA), q, terms order.

Estimating the parameters for the ARIMA models is a quite complicated nonlinear estimation problem. There are many approaches to fitting Box-Jenkins models, but maximum likelihood estimation is the preferred technique. The likelihood equations for the full ARIMA model are complicated and are not included here. See Brockwell and Davis (1991) for the mathematical details.

Model diagnostics suppose checking if the error term is assumed to follow the assumptions for a stationary univariate process. The residuals should be independent and identically distributed from a fixed distribution with a constant mean and variance. If the ARIMA model is a good model for the data, the residuals should satisfy these assumptions. For checking residuals correlation, Box-Ljung test and Autocorrelation and Partial Autocorrelation function plots are used. If these assumptions are not satisfied, we need to fit a more appropriate model. That is, we go back to the model identification step and try to develop a better model.

Since the indicator of crisis is a binary variable, equal to 1 if the patient was hospitalized and 0 otherwise, there are many frequentist statistical methods that could be used for predicting it, as for example generalized linear regression, more over logistic regression. Our problem was that those approaches assume that the variables in model have autocorrelation function equal to zero, which is not the case here. We used Bayesian approach for fitting the regression models of the body weight and crisis indicator because it allows the use of additional information on the patient and because it already exists a R package that we know.

The major difference in the frequentist and Bayesian approaches is that in a frequentist approach, unknown parameters are often, but not always, treated as having fixed but unknown values that are not capable of being treated as random variates in any sense, and hence there is no way that probabilities can be associated with them. In contrast, a Bayesian approach does allow probabilities to be associated with unknown parameters. The Bayesian approach allows

these probabilities to have an interpretation as representing the scientist's belief that given values of the parameter are true. The Bayesian approach is based on Bayes' theorem.

In this work we propose a structural time series model. One of the most important models for time series is the basic structural model: this consists of a trend, a seasonal and an irregular component. Our computational results are centred on the model consisting of a level trend and time dependent components. The temporal effects are modelled with a fifth order autoregressive process.
Once the regression model is chosen, a prior probability distribution on the model parameters is specified, representing our knowledge about these parameters before any data is observed. Once the data has been observed, the likelihood function, that is the distribution of the observed data conditional on parameters, is computed.
The posterior distribution is the distribution of the parameters after taking into account the observed data. Multiplying the prior distribution with the likelihood function a posterior distribution is provided. For obtaining parameters of our Bayesian model Markov Chain Monte Carlo (MCMC) is used to sample from the posterior distribution.
Next and final step is model diagnostics. In this step we are checking if obtained MCMC chain is converging to the target distribution after infinite iterations. We used two approaches for checking convergence, the Geweke (1992), compares means calculated from distinct segments of Markov chain, Gelman and Rubin (1992), computes m independent Markov chains and compares variances between chains.

The work is organised as follows: in Chapter 1, after a brief introduction to the theory of time series, we present the main properties of the Autoregressive Integrated Moving Average models. In Chapter 2 Bayesian statistic for structural time series models are reviewed. In Chapter 3 we present inference on ARIMA models for the body weight, blood pressure and heart rate of the patient and we analysed the data collected by physicians. In Chapter 4 we present the inference for Bayesian structural models for the body weight and the indicator of crisis.

The open source statistical software 'R' and various statistical and time series packages as 'tseries', 'forecast' and 'bsts' were used for this study.

Chapter 1

Autoregressive Moving Average Models for time series

In this chapter we introduce some basic ideas of time series analysis, and an extremely important class of time series defined in terms of linear difference equations with constant coefficients, calles *Autoregressive moving average (ARMA)* models. Chapter 1 is based on the book by Brockwell and Davis (2002).

A time series $\{Y_t, t \in T_0\}$ is a set of random variables, each one being recorded at a specified time t. A discrete-time series is one in which the set T_0 of times at which observations are made is a discrete set, as it is the case for example when observations are made at fixed time intervals. Continuous-time series are obtained when observations are recorded continuously over some time interval, e.g. when $T_0 = \big[0, 1\big]$. Most commonly, a time series is a sequence taken at successive equally spaced points in time.

Let $T_0 = \{0, \pm 1, \pm 2, \dots\}$, then the time series $\{Y_t, t \in T_0\}$ is called discrete time series. A basic model for representing a discrete time series $\{Y_t, t \in T_0\}$ is the additive model:

$$Y_t = \mu_t + \gamma_t + \varepsilon_t, \qquad t \in T_0 \tag{1.1}$$

Model (1.1) is also known as Classical Decomposition, where
Y_t is the observation at time t,
μ_t is the trend at time t,
γ_t is the seasonal at time t,
ε_t is the noise disturbance at time t.

1.1 Stationarity and Strict Stationarity

Definition of autocovariance function. *If* $\boldsymbol{Y} = \{Y_t, \quad t \in T_0\}$ *is a random process such that* $Var(Y_t) < \infty$ *for each* $t \in T_0$*, then the autocovariance function* $\gamma_Y(\cdot, \cdot)$ *of* Y_t *is defined by*

$$\gamma_Y(r, s) = Cov(Y_r, Y_s) = E[(Y_r - E[Y_r])(Y_s - E[Y_s])], \qquad r, s \in T_0. \tag{1.2}$$

Definition of autocorrelation fuction. *The autocorrelation function of the* Y_t *process is given by:*

$$\rho_Y(r, s) = \frac{\gamma_Y(r, s)}{\sqrt{var(Y_r)var(Y_s)}}, \quad r, s \in T_0 \tag{1.3}$$

Definition of stationarity. *The time series* $\{Y_t, t \in T_0\}$ *is said to be stationary if*

1. $E[|Y_t|^2] < \infty$ for all $t \in T_0$,
2. $E[Y_t] = m$ for all $t \in T_0$,
3. $\gamma_Y(r, s) = \gamma_Y(r + t, s + t)$ for all $r, s, t \in T_0$,

Stationarity as just defined is frequently referred to as weak stationarity, covariance stationarity, stationarity in the wide sense or second-order stationarity. For us however the term stationarity will always refer to the properties specified by this definition.

If $\{Y_t, \quad t \in T_0\}$ is stationary then $\gamma_Y(r, s) = \gamma_Y(r - s, 0)$ for all $r < s \in T_0$. It is therefore convenient to redefine the autocovariance function of a stationary time series as the function of the gap h between times t and $t + h$, i.e.:

$$\gamma_Y(h) \equiv \gamma_Y(h, 0) = Cov(Y_{t+h}, Y_t) \quad \forall t, h \in T_0 \tag{1.4}$$

Example(White noise process) A process $\{X_t\}$ is called *white noise*, and is indicated by the acronym *WN*, if it has the following properties:

$$E(X_t) = 0 \quad \forall t$$
$$E(X_t X_{t-k}) = \begin{cases} 0 & k \neq 0 \\ \sigma_X^2 & k = 0 \end{cases}$$

The autocorrelation function of the WN process is:

$$\rho_X(r,s) = \rho_X(\tau = s - r) = \begin{cases} 0 & \tau \neq 0 \\ 1 & \tau = 0 \end{cases}$$

Example(Sinusoidal process) A process $\{C_t\}$ is called *sinusoidal process* if:

$$C_t = Acos(\lambda t) + Bsin(\lambda t) \quad \lambda \in [-\pi, \pi], \quad t = 0, \pm 1, \pm 2, \dots$$

with A and B uncorrelated random variables with zero mean and equal variances:

$$E(A) = E(B) = 0$$
$$Var(A) = Var(B) = \sigma^2$$

λ is called the angular frequency of the process, $T = 2\pi/\lambda$ is the period of the process and $1/T$ is frequency. The autocovariance function of the sinusoidal process C_t is:

$$\gamma_C(r,s) = \gamma_C(\tau = s - r) = \sigma^2 cos(\lambda\tau)$$

Definition of strict stationarity. *The time series $\{Y_t, t \in T_0\}$ is said to be strictly stationary if the joint distribution of $(Y_{t_1}, ..., Y_{t_k})\prime$ and $(Y_{t_1+h}, ..., Y_{t_k+h})\prime$ are the same for all positive integers k, $\forall t_1, ..., t_k, h \in T_0$.*

If $T_0 = \{0, \pm 1, \pm 2, \dots\}$, the previous definition is equivalent to the statement that $\{Y_1, Y_2, \dots, Y_k\}\prime$ and $\{Y_{1+h}, Y_{2+h}, \dots, Y_{k+h}\}\prime$ have the same joint distribution for all positive integers k and integers h.

1.2 ARIMA models

An autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series. ARIMA models are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity.

ARIMA models are generally denoted by $ARIMA(p, d, q)$ where parameters p, d and q are non-negative integers, p is the order of the autoregressive model, d is the degree of differencing, and q is the order of the moving-average model.

1.2.1 Autoregressive processes

Definition of AR(1) models. *The $AR(1)$ model is defined as:*

$$Y_t = a + bY_{t-1} + \varepsilon_t \tag{1.5}$$

where ε_t is white noise.

An $AR(1)$ process is stationary if and only if $|b| < 1$.
The moments of a process $AR(1)$ are

$$E(Y_t) = \frac{a}{1-b}$$
$$var(Y_t) = \frac{\sigma_\varepsilon^2}{1-a^2}$$
$$cov(X_t, X_{t+\tau}) = \sigma_\varepsilon^2 \frac{a^\tau}{1-a^2}$$

Definition of $AR(p)$ models. *The notation AR(p) indicates an autoregressive model of order $p = 2, 3, \ldots$. The $AR(p)$ model is defined as:*

$$Y_t = c + a_1 Y_{t-1} + ... + a_p Y_{t-p} + \varepsilon_t \tag{1.6}$$

where $a_1, a_2, \ldots, a_p$ are the parameters of the model, c is a constant, and ε_t is white noise.

Some parameter constraints are necessary for the model to remain wide-sense stationary. For example, processes in the AR(1) model with $a_1 \geq 1$ are not stationary. More generally, for an AR(p) model to be wide-sense stationary, the

roots of the polynomial $z^p - \sum_{i=1}^{p} a_i z^{p-i} = 0$ must lie inside of the unit circle, i.e., each (complex) root z_i must satisfy $z_i < 1$.

1.2.2 Moving average processes

Definition of $MA(1)$ models. *Let $\{\varepsilon_t\}$ be a white noise process with variance $E(\varepsilon_t) = \sigma_\varepsilon^2$.* If

$$Y_t = \mu + \varepsilon_t + b\varepsilon_{t-1}$$

with μ, b deterministic constants, then $\{Y_t\}$ is called the moving average process of order $1(MA(1))$. An $MA(1)$ process has mean μ and variance

$$var(Y_t) = (1 + b^2)\sigma_\varepsilon^2$$

Definition of $MA(q)$ models. *The notation MA(q) refers to the moving average model of order $q = 2, 3, \ldots$:*

$$Y_t = \mu + \varepsilon_t + b_1\varepsilon_{t-1} + ... + b_q\varepsilon_{t-q} \tag{1.7}$$

where μ is the mean of each Y_t, $b_1, \ldots, b_q$ are the parameters of the model and $\varepsilon_t, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q}$ are white noise errors.

The $MA(q)$ process has mean and variance, respectively,

$$\begin{aligned} E(Y_t) &= \mu \\ Var(Y_t) &= \sigma_\varepsilon^2(1 + b_1^2 + \cdots + b_q^2) \end{aligned}$$

A moving-average model is conceptually a linear regression of the current value of the series against current and previous (unobserved) white noise error terms or random shocks. The random shocks at each point are assumed to be mutually independent and to come from the same distribution, typically a normal distribution, with location at zero and constant scale.

Definition of ARMA models. *Given a time series of data Y_t where t is an integer index and the Y_t's are real numbers, an ARMA(p,q) model is given by:*

$$Y_t - a_1Y_{t-1} - ... - a_pY_{t-p} = \varepsilon_t + b_1\varepsilon_{t-1} + ... + b_q\varepsilon_{t-q} \tag{1.8}$$

Or equivalently:

$$\Big(1 - \sum_{i=1}^{p} a_iB^i\Big)Y_t = \Big(1 + \sum_{i=1}^{q} b_iB^i\Big)\varepsilon_t \tag{1.9}$$

where B is the lag operator, meaning:

$$BY_t = Y_{t-1}, \quad B^kY_t = Y_{t-k} \tag{1.10}$$

The a_i are the parameters of the autoregressive part of the model, the b_i are the parameters of moving average part and the ε_t are the error terms. The error terms are generally assumed to be independent, identically distributed variables, with zero mean and equal variance σ^2.

Definition of ARIMA models. *An ARIMA(p, d, q) process is given by*

$$\Big(1 - \sum_{i=1}^{p} a_i B^i\Big)\Big(1 - B\Big)^d Y_t = \Big(1 + \sum_{i=1}^{q} b_i B^i\Big)\varepsilon_t \tag{1.11}$$

Note that, if $d = 0$, ARIMA$(p, 0, q)$ model is equivalent to ARMA(p, q) model. And if $d = 1$, we have that ARIMA$(p, 1, q)$ model applied on series $\{Y_t, t \in T_0\}$ is same as if we apply ARMA(p, q) model to series $\{Y'_t = Y_{t+1} - Y_t, t \in T_0\}$. In general, ARIMA$(p, q, d)$ for $\{Y_t, t \in T_0\}$ is equivalent to ARMA(p, q) on difference with lag d of $\{Y_t, t \in T_0\}$.

Some well-known special cases arise naturally or are mathematically equivalent to other popular forecasting models. For example:

- An ARIMA(0,1,0) model is given by $X_t = X_{t-1} + \varepsilon_t$ — which is simply a random walk.

- An ARIMA(0,1,0) with a constant, given by $X_t = c + X_{t-1} + \varepsilon_t$ — which is a random walk with drift.

- An ARIMA(0,0,0) model is a white noise model.

To determine the order of a non-seasonal ARIMA model, a useful criterion is the *Akaike information criterion* (AIC). It is written as

$$AIC = -2log(L) + 2(p + q + k + 1) \tag{1.12}$$

where L is the likelihood of the data, p is the order of the autoregressive part and q is the order of the moving average part. The parameter k in AIC formula is defined as the number of parameters in the model being fitted to the data. For AIC, if $k = 1$ then $c \neq 0$ and if $k = 0$ then $c = 0$.

The objective is to minimize the AIC value for a good model. The lower the value of AIC for a range of models being investigated, the better the model will suit the data.

1.3 Stationarity tests

Stationarity tests allow verifying whether a series is stationary or not. There are two different approaches: stationarity tests such as the KPSS test that consider null hypothesis H_0 that the series is stationary, and unit root tests, such as the Dickey-Fuller test and its augmented version, the Augmented Dickey-Fuller test (ADF), or the Phillips-Perron test (PP), for which the null hypothesis H_0 is on the contrary that the series possesses a unit root and hence is not stationary.

Unit root

A linear stochastic process has a unit root if 1 is a root of the process's characteristic equation. Such a process is non-stationary.

If the other roots of the characteristic equation lie inside the unit circle-that is, have a modulus (absolute value) less than one, then the first difference of the process will be stationary; otherwise, the process will need to be differenced multiple times to become stationary.

If a root of the process's characteristic equation is larger than 1, then it is called an *explosive process.*

Unit root processes may sometimes be confused with trend-stationary processes; it is possible for a time series to be non-stationary, yet have no unit root and be trend-stationary.

Consider a discrete-time stochastic process $\{Y_t, \quad t = 1, ..., \infty\}$, and suppose that it can be written as an autoregressive process of order p:

$$Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + ... + a_p Y_{t-p} + \varepsilon_t. \tag{1.13}$$

Here, $\{\varepsilon_t, \quad t = 1, ..., \infty\}$, is a serially uncorrelated, zero-mean stochastic process with constant variance σ^2. For convenience, assume $Y_0 = 0$. If $m = 1$ is a root of the characteristic equation:

$$m^p - a_1 m^{p-1} - a_2 m^{p-2} - ... - a_p = 0 \tag{1.14}$$

then the stochastic process has a unit root.

Augmented Dickey-Fuller test

The Augmented Dickey Fuller (ADF) test is a unit root test for stationarity.

The hypotheses for the test are as follows:

- the null hypothesis H_0 is that there is a unit root;

- the alternative hypothesis differs slightly according to which equation we are using. The basic alternative is that the time series is stationary (or trend-stationary).

Before running an ADF test, an inspection of data is needed to figure out an appropriate regression model. For example, a nonzero mean indicates the regression will have a constant term. The three basic regression models are:

- No constant, no trend: $Y_t = a_1 Y_{t-1} + \varepsilon_t$
- Constant, no trend: $Y_t = \alpha + a_1 Y_{t-1} + \varepsilon_t$
- Constant and trend: $Y_t = \alpha + a_1 Y_{t-1} + \lambda_t + \varepsilon_t$

The Augmented Dickey Fuller adds lagged differences to these models:

- No constant, no trend: $Y_t = \sum_{i=1}^{p} a_i Y_{t-i} + \varepsilon_t$
- Constant, no trend: $y_t = \alpha + \sum_{i=1}^{p} a_i Y_{t-i} + \varepsilon_t$
- Constant and trend: $Y_t = \alpha + \sum_{i=1}^{p} a_i Y_{t-i} + \lambda_t + \varepsilon_t$

A lag length needs to be chosen to run the test. The lag length should be chosen so that the residuals are not serially correlated. There are several options for choosing lags: Minimize Akaike's information criterion (AIC) or Bayesian information criterion (BIC).

KPSS test

The Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test figures out if a time series is stationary around either a deterministic trend or a level trend, or is non-stationary due to a unit root.

Let $\{Y_t, \quad t = 1, 2, ..., N\}$ be the observerd series for which we wish to test stationarity. Assume that we can decompose the series into the sum of a deterministic trend, a random walk and a stationary error with the following linear regression model:

$$Y_t = r_t + \beta t + \varepsilon_t \tag{1.15}$$

where r_t is a random walk, i.e., $r_t = r_{t-1} + u_t$ and u_t are independent and identically distributed (iid) $N\big(0, \sigma_u^2\big)$, βt is deterministic trend and ε_t is serially uncorrelated, zero-mean stochastic process of errors with constant variance σ_ε^2.

To test in this model if Y_t is a trend stationary process, namely, the time series is stationary around deterministic trend, the null hypothesis will be $\sigma_u^2 = 0$, against the alternative of a positive σ_u^2.
In another stationarity case, level stationarity, namely, the series is stationary around a fixed level, and the null hypothesis is $\beta = 0$.

1.4 Estimation and Elimination of the Trend and Seasonal Components

The first step in the analysis of any time series is to plot the data to see if there are apparent discontinuities in the series, such as a sudden change of level, or if there are outlying observations. Inspection of graph may also suggest the possibility of representing the data as a realization of the process (the "classical decomposition" model)

$$Y_t = \mu_t + \gamma_t + \varepsilon_t \tag{1.16}$$

where μ_t is a slowly changing function known as a trend component, γ_t is a function whit known period d and refers to as a seasonal component, and ε_t is a random noise component which is stationary.

Our aim is to estimate and extract the deterministic components μ_t and γ_t in the hope that the noise component ε_t will turn out to be a stationary random process. We can then use the theory of such processes to find a satisfactory probabilistic model for the process ε_t.

An approach, developed by Box and Jenkins in 1970, is to apply the difference operator repeatedly to the data Y_t until the differenced observations resemble a realization of some stationary process W_t.

1.4.1 Elimination of a trend in the absence of seasonality

In the absence of a seasonal component the model becomes

$$Y_t = \mu_t + \varepsilon_t, \quad t = 1, ..., n \tag{1.17}$$

where, without loss of generality, we can assume that $E[\varepsilon_t] = 0$.

Now we attempt to eliminate the trend term by differencing. We define the

first differencing operator ∇ by

$$\nabla Y_t = y_t - Y_{t-1} = (1 - B)Y_t, \tag{1.18}$$

where B is the backward shift operator,

$$BY_t = Y_{t-1}. \tag{1.19}$$

Powers of the operators B and ∇ are defined in the following way

$$B^j(Y_t) = Y_{t-j} \tag{1.20}$$

$$\nabla^j(Y_t) = \nabla(\nabla^{j-1}(Y_t)), j \geq 1 \tag{1.21}$$

$$\nabla^0(Y_t) = Y_t. \tag{1.22}$$

If the operator ∇ is applied to a linear trend function $\mu_t = at + b$, then we obtain constant function $\nabla\mu_t = a$. In the same way any polynomial trend of degree k can be reduced to a constant by application of the operator ∇^k.
Starting therefore with the model $Y_t = \mu_t + \varepsilon_t$ where $\mu_t = \sum_{j=0}^{k} a_j t^j$ and ε_t is stationary with zero mean, we obtain

$$\nabla^k Y_t = k!a_k + \nabla^k \varepsilon_t \tag{1.23}$$

that is a stationary process with mean $k!a_k$. These considerations suggest the possibility, given any sequence $\{y_t\}$ of data, of applying the operator ∇ repeatedly until we find a sequence $\{\nabla^k y_t\}$ which can be modelled as a realization of stationary process. It is often found in practice that the required order of differencing k is quite small, frequently one or two.

1.4.2 Elimination of both Trend and Seasonality

The method described for the removal of trend can be applied in a natural way to eliminate both trend and seasonality in the general model

$$Y_t = \mu_t + \gamma_t + \varepsilon_t \tag{1.24}$$

where $E[\varepsilon_t] = 0$, $\gamma_{t+d} = \gamma_t$ and $\sum_{j=1}^{d} \gamma_j = 0$.

The technique of differencing which we applied earlier to non-seasonal data can be adapted to deal with seasonality of period d by introducing the lag-d difference operator ∇_d defined by

$$\nabla_d y_t = y_t - y_{t-d} = (1 - B^d)y_t. \tag{1.25}$$

(This operator should not be confused with operator $\nabla^d = (1 - B)^d$ defined earlier.)

Applying the operator ∇_d to the model $Y_t = \mu_t + \gamma_t + \varepsilon_t$, where $\{\gamma_t\}$ has period d, we obtain

$$\nabla_d Y_t = \mu_t - \mu_{t-d} + \varepsilon_t - \varepsilon_{t-d} \tag{1.26}$$

which gives a decomposition of the difference $\nabla_d y_t$ into a trend component $(\mu_t - \mu_{t-d})$ and a noise term $(\varepsilon_t - \varepsilon_{t-d})$. The trend component, $\mu_t - \mu_{t-d}$ can be eliminated using the method already described.

1.5 Model Diagnostic

The residuals from a model are calculated as the difference between the actual values and the fitted values: $\hat{\varepsilon}_t = x_t - \hat{x}_t$. Each residual is the unpredictable component of the associated observation.

After fitting a model, it is necessary to check that the assumptions of the model's residuals have been satisfied. For checking the iid hypothesis on the errors, we can check the *Autocorrelation Function* and the *Partial Autocorrelation Function* or resort to *Box-Ljung test*.

Autocorrelation Function

The correlation between two variables Y_1 and Y_2 is defined as:

$$\rho = \frac{E[(Y_1 - \mu_1)(Y_2 - \mu_2)]}{\sigma_1 \sigma_2} = \frac{Cov(Y_1, Y_2)}{\sigma_1 \sigma_2} \tag{1.27}$$

where E is the expectation operator, μ_1 and μ_2 are the means respectively for Y_1 and Y_2 and σ_1, σ_2 are their standard deviations.

Upon the above definition, residuals sample autocorrelations of order $k = 0, 1, 2, ...$ can be obtained by computing the following expression with the series of estimated errors e_t:

$$\hat{\rho}_k = \frac{\sum_{t=1}^{n-k}(e_t - \bar{e})(e_{t+k} - \bar{e})}{\sum_{t=1}^{n}(e_t - \bar{e})^2} \tag{1.28}$$

where $\bar{e}$ is the sample mean of the estimated residuals.

Partial Autocorrelation Function

The partial autocorrelations measure the linear dependence of one variable after removing the effect of other variable(s) that may affect both variables. For example, the partial autocorrelation of order h measures the effect of x_{t-h} on x_t after removing the effect of $x_{t-h+1}, x_{t-h+2}, ...x_{t-1}$ on both x_{t-h} and x_t.

The 1st order partial autocorrelation of residuals will be defined equal to the 1st order autocorrelation.
The 2nd order (lag) partial autocorrelation is:

$$\frac{Cov(\varepsilon_{t-2}, \varepsilon_t|\varepsilon_{t-1})}{\sqrt{Var(\varepsilon_t|\varepsilon_{t-1})Var(\varepsilon_{t-2}|\varepsilon_{t-1})}} \tag{1.29}$$

The 3rd order (lag) partial autocorrelation is:

$$\frac{Cov(\varepsilon_{t-3}, \varepsilon_t|\varepsilon_{t-1}, \varepsilon_{t-2})}{\sqrt{Var(\varepsilon_t|\varepsilon_{t-1}, \varepsilon_{t-2})Var(\varepsilon_{t-3}|\varepsilon_{t-1}, \varepsilon_{t-2})}} \tag{1.30}$$

And, so on, for any lag.

Ljung-Box test

The autocorrelation function (ACF) and partial autocorrelation function (PACF) are useful qualitative tools to assess the presence of autocorrelation at individual lags. The Ljung-Box Q-test is a quantitative way to test for autocorrelation at multiple lags jointly. The null hypothesis for this test is that the first m autocorrelations between residuals are jointly zero.

The Ljung–Box test may be defined as:
H_0 : "the residuals are independently distributed"
H_1 : "the residuals are not independently distributed".

The test statistic is:

$$Q = n(n+2)\sum_{k=1}^{m}\frac{\hat{\rho}_k}{n-k} \tag{1.31}$$

where n is the sample size, $\hat{\rho}_k$ is the sample autocorrelation at lag k, and m is the number of lags being tested. Under H_0 the statistic Q follows a chi-squared distribution with h degrees of freedom (χ^2_h). For a significance level α, the critical region for rejecting the null hypothesis of randomness is:

$$Q > \chi^2_{1-\alpha,h} \tag{1.32}$$

where $\chi^2_{1-\alpha,h}$ is the (1-α)-quantile of the χ^2_{rh}. Because the test is applied to the estimated residuals, the degrees of freedom must account for the estimated model parameters so that $h = m$ - p - d - q, where p, d and q indicate the number of parameters from the $ARIMA(p, d, q)$ model fit to the data.

Chapter 2

Bayesian Statistic for Structural Time Series

In this chapter is provided an introduction to Bayesian statistical inference, in quite general terms, how Bayesian data analysis proceeds. At a high level of abstraction, Bayesian data analysis is extremely simple, following the same, basic recipe: via Bayes Rule, we use the data to update prior beliefs. More detailed theory about the Bayesian paradigm can be found in Jackman (2009).

In several applications there are enough (statistical) information on the most likely values on the parameters θ to be estimated even before making any experiment or any observation. The information is given by the corresponding probability density function (pdf) $\pi[\theta]$ that accounts for all the statistical properties of θ before any observation. Since $\pi[\theta]$ accounts for the statistical properties before any experiment, this is referred to as *a-priori pdf.* Bayesian methods make an efficient use of the a-priori pdf to yield the "best estimate" given both the observation $\{y_t\}$ and the a-priori knowledge $\pi[\theta]$.

Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ be random sample from probability distribution f, conditionally to some unknown parameters θ, then a Bayesian model is given by:

$$Y_1, Y_2, \ldots, Y_n | \boldsymbol{\theta} \sim f(\boldsymbol{y}, \boldsymbol{\theta}) \tag{2.1}$$

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}) \tag{2.2}$$

Bayesian inference relies on the Bayes theorem:

$$P(A_i|E) = \frac{P(E|A_i)P(A_i)}{P(E)} = \frac{P(E|A_i)P(A_i)}{\sum_{k=1}^{n} P(E|A_k)P(A_k)}$$

where $A_1, A_2, \ldots, A_n$ is finite or infinite partition of sample space $(\Omega, \boldsymbol{B})$ such that $P(A_j) > 0 \quad \forall j$ and $P(E) > 0$.

Given Equations (2.1) and (2.2), the goal is to calculate *a-posterior distribution* $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$ for the parameters $\boldsymbol{\theta}$ given data $\mathbf{Y}$, this represents an update of $\pi(\boldsymbol{\theta})$ after conditioning on the sample data. From Bayesian theorem we get:

$$\pi(\boldsymbol{\theta}|\boldsymbol{Y}) = \frac{f(\boldsymbol{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{Y})} \quad .$$

where $f(\boldsymbol{Y}|\boldsymbol{\theta})$ is the likelihood and $f(\boldsymbol{Y})$ is the marginal distribution of the data. Since $f(\boldsymbol{Y})$ is independent from the parameters $\boldsymbol{\theta}$, we can calculate the posterior distribution to a constant:

$$\pi(\boldsymbol{\theta}|\boldsymbol{Y}) \propto f(\boldsymbol{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \quad .$$

2.1 Structural Time Series

A structural time series model is defined by two equations. The *observation equation* relates the observed data y_t to a vector of latent variables α_t known as the state:

$$Y_t = Z_t'\theta_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2) \tag{2.3}$$

The *transition equation* describes how the latent state evolves through time.

$$\theta_{t+1} = B_t\theta_t + R_t\eta_t, \quad \eta_t \sim N(0, \tau^2) \tag{2.4}$$

The error terms ε_t and η_t are Gaussian and independent of everything else.The arrays Z_t , B_t and R_t are structural parameters. They may contain parameters in the statistical sense, but often they simply contain strategically placed 0's and 1's indicating which bits of θ_t are relevant for a particular computation.
A model that can be described by equations (2.3) and (2.4) is said to be in *state space form*.

For example, one useful model can be obtained by adding a regression component to the popular "basic structural model." This model can be written as:

$$\begin{cases} Y_t = \mu_t + +\tau_t + \beta' \boldsymbol{X_t} + \varepsilon_t \\ \mu_{t+1} = \mu_t + \delta_t + \eta_t \\ \delta_{t+1} = \delta_t + \xi_t \\ \tau_t = -\sum_{s=1}^{S-1} \tau_{t-s} + w_t \end{cases} \tag{2.5}$$

where $\varepsilon_t, \eta_t, \xi_t, w_t$ are independent components of Gaussian random noise with variances $\sigma_\varepsilon^2, \sigma_\eta^2, \sigma_\xi^2, \sigma_w^2$. The current level of the trend is μ_t, the current "slope" of the trend is δ_t and the seasonal component is τ_t.

2.1.1 Local Level Trend

The simplest useful model is the *local level model*, in which the vector θ_t is just a scalar μ_t. The local level model is a random walk.

$$y_t = \mu_t + \varepsilon_t \tag{2.6}$$
$$\mu_{t+1} = \mu_t + \eta_t \tag{2.7}$$

Here $\theta_t = \mu_t$, and Z_t, B_t, and R_t are all the scalar value 1. The probabilistic assumptions are $\varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$ and $\eta_t \overset{iid}{\sim} N(0, \tau^2)$.

If $\tau^2 = 0$ then μ_t is a constant, so the data are IID Gaussian noise. In that case the best estimator of y_{t+1} is the mean of $y_1, y_2, \ldots, y_t$. Conversely, if $\sigma^2 = 0$ then the data follow a random walk, in which case the best estimator of y_{t+1} is y_t. Notice that in one case the estimator depends on all past data (weighted equally) while in the other it depends only on the most recent data point, giving zero weight to the past data .

2.1.2 Local Linear Trend

The LLT model extends the LL model with a slope:

$$y_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \tag{2.8}$$
$$\mu_{t+1} = \mu_t + \delta_t + \eta_t, \quad \eta_t \sim N(0, \sigma_\eta^2) \tag{2.9}$$
$$\delta_{t+1} = \delta_t + \xi_t, \quad \xi_t \sim N(0, \sigma_\xi^2) \tag{2.10}$$

If $\sigma_\varepsilon^2 = 0$ the trend is a random walk with constant drift β_1; if $\beta_1 = 0$ the model reduces to a LL model. If additionally $\sigma_\eta^2 = 0$ the trend is a straight line with slope β_1 and intercept μ_1. If $\sigma_\xi^2 = 0$ but $\sigma_\eta^2 = 0$, then the trend is smooth curve or Integrated Random Walk.

2.1.3 AR(p) model with time varying coefficients

The $AR(p)$ model with time varying coefficients takes the form:

$$Y_t = \alpha_{0,t} + \alpha_{1,t} y_{t-1} + \cdots + \alpha_{p,t} y_{t-p} + \varepsilon_t \tag{2.11}$$
$$\alpha_{i,t+1} = \alpha_{i,t} + \eta_t \tag{2.12}$$

where we assume that the error terms are independent normals:

$$\varepsilon_t \sim N(0, \sigma_\varepsilon^2), \quad \eta_t \sim N(0, \sigma_\eta^2)$$

2.2 Spike and Slab Variable selection

Equation (2.5) contains a regression component that allows a set of external factors to contribute to the prediction. *Spike and slab variable selection* is a Bayesian variable selection technique that is particularly useful when the number of possible predictors is larger than the number of observations.

Our problem is now to establish whether Y_i is associated with X_k. Let $\gamma_k = 1$ if $\beta_k \neq 0$, and $\gamma_k = 0$ if $\beta_k = 0$. Let β_γ denote the subset of elements of β where $\beta_k \neq 0$. A spike-and-slab prior may be written

$$p(\beta, \gamma, \sigma_\varepsilon^2) = p_3(\beta_\gamma|\gamma, \sigma_\varepsilon^2)p_2(\sigma_\varepsilon^2|\gamma)p_1(\gamma) \tag{2.13}$$

The marginal distribution $p_1(\gamma)$ is the *"spike"* so named because it places positive probability mass at zero. In practice it is convenient to simply use an independent Bernoulli prior for $p_1(\gamma)$:

$$\gamma \sim \prod_{k=1}^{K} \pi_k^{\gamma_k}(1-\pi_k)^{\gamma_k} \tag{2.14}$$

Equation (2.17) is often further simplified by assuming all the π_k are the same value π. A natural way to elicit π is to ask the analyst for an "expected model size," so that if one expects p nonzero predictors then $\pi = p/K$, where K is the dimension of the design matrix $\boldsymbol{X_t}$.

Example: Let us consider general linear model and suppose that we have only one covariate:

$$Y_i|\theta_i, \eta \overset{iid}{\sim} f(y_i; \theta_i, \eta) \tag{2.15}$$

$$g(\theta_i) = \beta_0 + \beta_1 X_{i1} \tag{2.16}$$

$$\beta_0 \sim \pi_0(\beta_0) \quad \beta_1 \sim \pi_1(\beta_1) \quad \eta \sim \pi_2(\eta) \tag{2.17}$$

We can perform a statistical test to verify: $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$

Then the quasi spike-and-slab prior is:

$$\beta_1|y_1 \sim \gamma_1 N(0, c_1^2\tau_1^2) + (1-\gamma_1)N(0, \tau_1) \tag{2.18}$$

$$\gamma_1 \sim Bernoulli(\pi_1) \tag{2.19}$$

$$\pi_1 \sim Unif(0,1) \quad or \quad \pi_1 = 0.5 \tag{2.20}$$

τ_1^2 is sufficiently small to approximate δ_0 function and c_1^2 is sufficiently large. If:

$$\beta_1 \in [-\delta, \delta] \Rightarrow \beta_1 \simeq 0 \tag{2.21}$$

$$\delta = \pm\tau_1 \sqrt{\frac{2ln(\,c_1)\,c_1^2}{c_1^2 - 1}} \tag{2.22}$$

Chapter 3

Analysis of data set for person with Capillary Leak Syndrome

Here we will introduce the data set and the general goal of analysis, then we try to find the best models for time series of weight of patient.
Body weight, blood pressure and heart rate are simple measurements that can be done daily by the patient at home, while other measurements potentially more fit to explore the syndrome (such as measurements on blood samples) need professional assistance typically in a hospital. By using weight data and blood pressure and heart beat rate, it would be interesting to forecast a crisis with at least some hour anticipation, and even more useful to predict if such crisis will need hospitalization or can be handled at the patient's home. Missing data in the time series are due to the presence of crises, which do not allow the patient to stay still on the weight. Furthermore, the effectiveness of a therapy can be possibly assessed by looking at the changes in the model of the weight series before and after the beginning of the therapy. This can be of help to the physician, in absence of better indicators of the effectiveness of the therapy.

3.1 Capillary Leak Syndrome

In this section, we will describe Capillary Leak Syndrome, the information about this disease are collected from reports from Mayo Clinic and NORD (National Organization of Rare Disorders).

Systemic capillary leak syndrome (SCLS) is a rare disorder characterized by repeated flares of massive leakage of plasma from blood vessels into neighbouring body cavities and muscles. This results in a sharp drop in blood pressure

that, if not treated, can lead to organ failure and death. SCLS occurs most often in adults and the disease is very rare in children.

Also called Clarkson's disease, this condition can be mistaken for severe reactions to widespread infections (septic shock) or severe allergic reactions (anaphylactic shock). The frequency of attacks can range from several per year to a single instance in a lifetime.

Attacks may be triggered by an upper respiratory infection or intense physical exertion. Attacks are often preceded by one to two days of one or more nonspecific symptoms that may include:

- Irritability
- Fatigue
- Abdominal pain
- Nausea
- Muscle aches
- Increased thirst
- Sudden increase in body weight

As the fluid leaks out from the bloodstream, blood volume and blood pressure drop. This can starve tissues in the kidneys, brain and liver of the oxygen and nutrients they need for normal function.

More than one half of patients have a monoclonal or M protein detected in the blood. The level of M protein is usually low. The M protein is produced by what usually amount to small numbers of plasma cells in the marrow. The M protein itself does not appear to cause the attacks. Recently it has been suggested that capillary lining cells may be damaged by a factor in the blood, which is produced during the acute attack. SCLS has been recognized in a range of racial backgrounds and nationalities. There appears to be no genetic predisposition to the disease. The cause of SCLS is unknown, and there's no known cure. Treatment during episodes aims to stabilize symptoms and prevent severe complications. The use of intravenous fluids must be monitored carefully.

Once an attack is underway, treatment is directed toward controlling blood pressure to maintain blood flow to vital organs as well as preventing excessive swelling and fluid accumulation.

Treatment of a fully developed SCLS episode requires recognition that there are two phases of the acute attack. The first phase, which often lasts several days is called the resuscitation phase aimed at controlling the capillary leak and maintaining blood pressure. In that phase an albumin and fluid leak from the capillaries into the tissue spaces causes swelling. This loss of fluid has similar effects on the circulation as dehydration, slowing the flow of oxygen carrying blood to tissues. The blood pressure falls and the red cells concentrate. Intravenous fluid replacement is usually required. In most cases intravenous fluids must be administered immediately and in high-volume in order to prevent excessive drops in blood pressure. Constant check of the fluid loss is important because sustained low blood pressure can damage vital organs such as the kidneys.

The second phase of the treatment is sometimes called the recruitment phase as fluids and albumin are reabsorbed from the tissues. In this phase the capillary leak has abated and the main threat is fluid overload. Even though the blood pressure may still be low, it is important to avoid overly aggressive intravenous fluid administration causing massive swelling of the extremities requiring surgical decompression. In this procedure the skin of the arms or legs is incised to release the compressive pressure the retained fluid is having on blood flow to and from the extremities. Excessive intravenous fluids may also cause accumulation of fluid in the lungs and around other vital organs. Many of the deaths happen during this recruitment phase. The goal during the acute phase is NOT to attempt to maintain absolutely normal blood pressure or urine flow but to maintain the blood pressure at just sufficiently high enough levels to avoid permanent damage to vital organs yet spare the patient from the risks of excess fluid administration.

Monthly infusions of intravenous immunoglobulin (IVIG) can help prevent future episodes. Preventive treatment with certain oral medications originally designed to treat asthma also can be helpful, but these drugs may produce troublesome side effects, such as tremors.

3.2 Analysis of weights

The goal of this section is to analyse and fit some models to time series of the weight of a patient measured each day from 16.10.2013 until 20.06.2017.

The model developed for forecasting is an Autoregressive Integrated Moving Average (ARIMA) model. This model was introduced by Box and Jenkins in 1960 and hence this model is also known as Box-Jenkins model which is used to

forecast a single variable. The main reason of choosing ARIMA model in this study for the forecasting is because this model assumes and takes into account the non-zero autocorrelation between the successive values of the time series data.

For better analysis, data are first split into three phases:

1. From beginning until 17.11.2013: in this period two-drug therapy was applied (theophylline and terbutaline), which has been revealed to be irrelevant. This period can be taken as a reference for evolution in the absence of external therapeutic interventions.

2. From 18.11.2013 until 31.08.2016: immunoglobulin therapy (ig_vena). Clinically, it is observed that effectivness of therapy decrease, in the begining the therapy was very effective, but as time was passing, influenc of therapy on patient health was smaller and smaller. From a statistical point of view, it is interesting to evaluate the evolution of the parameters over time after each cycle of infusions, taking as zero time the day of the beginning of the cycle.

3. From 01.09.2016 until 20.06.2017: the therapy consists of plasma exchange cycles.

Figure 3.1 shows the daily body weight time series:

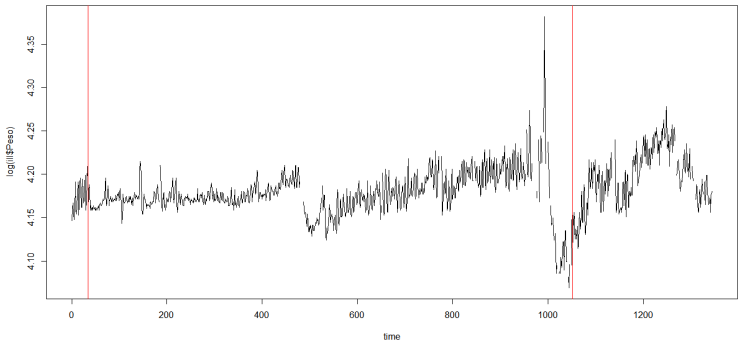

Figure 3.1: Time series of the daily body weight in the log scale: the two vertical lines split the time series into the three different subseries.

We notice difference in variation in these three phases, meaning that in last two phases we have bigger changes in the body weight, also we can notice some missing data, which are shown in Figure 3.2 through vertical lines.

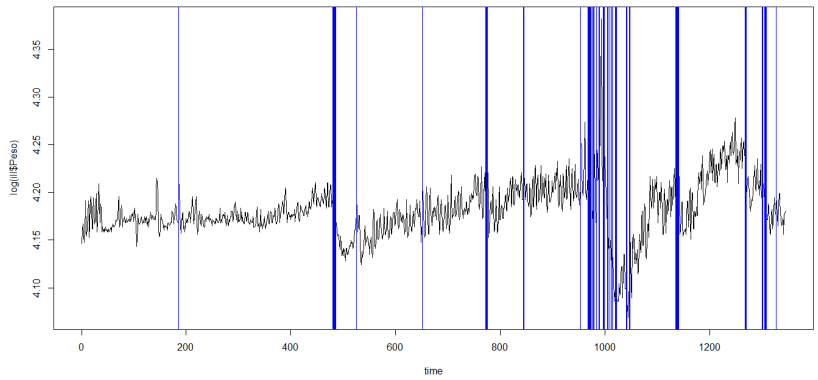

Figure 3.2: Time series of the daily body weight regards to time in days: blue lines represent missing parts of time series.

The dates with missing values are displayed in Figure 3.3:

```
"2014-04-18" "2015-02-07" "2015-02-08" "2015-02-09" "2015-02-10" "2015-02-11"
"2015-02-12" "2015-02-13" "2015-03-24" "2015-07-28" "2015-11-26" "2015-11-27"
"2015-11-28" "2015-11-29" "2016-02-06" "2016-02-07" "2016-05-25" "2016-06-08"
"2016-06-09" "2016-06-10" "2016-06-11" "2016-06-12" "2016-06-13" "2016-06-14"
"2016-06-16" "2016-06-19" "2016-06-20" "2016-06-24" "2016-06-29" "2016-06-30"
"2016-07-07" "2016-07-08" "2016-07-09" "2016-07-10" "2016-07-15" "2016-07-16"
"2016-07-20" "2016-07-24" "2016-07-25" "2016-07-30" "2016-07-31" "2016-08-01"
"2016-08-02" "2016-08-20" "2016-08-21" "2016-08-22" "2016-08-26" "2016-08-27"
"2016-11-23" "2016-11-24" "2016-11-25" "2016-11-26" "2016-11-27" "2016-11-28"
"2017-04-04" "2017-04-05" "2017-04-06" "2017-05-07" "2017-05-08" "2017-05-12"
"2017-05-13" "2017-05-14" "2017-05-15" "2017-06-02" "2017-06-03"
```

Figure 3.3: Dates for which values of time series of weights are missing.

The main idea of this project is to examine if each subseries of the weights time series can be fitted by some of $ARIMA(p, d, q)$ models (introduced in Chapter 2 Section 3). We can do that with the help of *auto.arima* function, contained in R package called *forecast* (more about this package can be found in article by Hyndman and Khandakar (2008.)). This function gives us the best model according to AIC value; the smaller AIC value is, the better the model is.

First stage of ARIMA model building is to identify whether the variable, which is being forecasted, is stationary in time series or not. By stationary we mean, the values of variable over time varies around a constant mean and variance. The time plot of the body weight data in Figure 3.1 above clearly shows that the data is not stationary. The ARIMA model cannot be built until we make this series stationary. We first have to difference the time series d times to obtain a stationary series in order to have an ARIMA(p,d,q) model with d as the order of differencing used. Package *forecast* contains also function *ndiffs* which helps to estimate the number of differences required to make a given time series stationary (i.e. parameter d in $ARIMA(p, d, q)$ *model*).

For checking residuals correlation, it will be used *acf, pacf* and *Box.test* functions in the *stats* package of `R`.

3.2.1 Fitting ARIMA model to the first phase of the wight time series

The first part of data is from 16.10.2013 until 17.11.2013.

We start with the estimation of parameter d using function *ndiffs* and we got that difference at lag 1 gives us stationary data. Therefore we will search for the best $ARIMA(p, 1, q)$ model, but also for the best model among all possibles with $d = 2$, which was suggested by the team of physicians in charge of the patient. At the end, as the best model for this subseries, we will choose the better model among then with respect to both their AIC values and fulfillness of residuals assumptions.

The best model among all possible ones with $d = 1$

As the result of *auto.arima* function we have that the best model in this case is $ARIMA(0, 1, 0)$ and its AIC value is -176.56, meaning:

$$X_t - X_{t-1} = \varepsilon_t \tag{3.1}$$

where $X_t, \quad t \in \{1, 2, \dots\}$ represents daily log weight time series and $\varepsilon_t, \quad t \in \{1, 2, \dots\}$ are iid variables with zero mean.

The plot for this fitted model is represented on Figure 3.4.

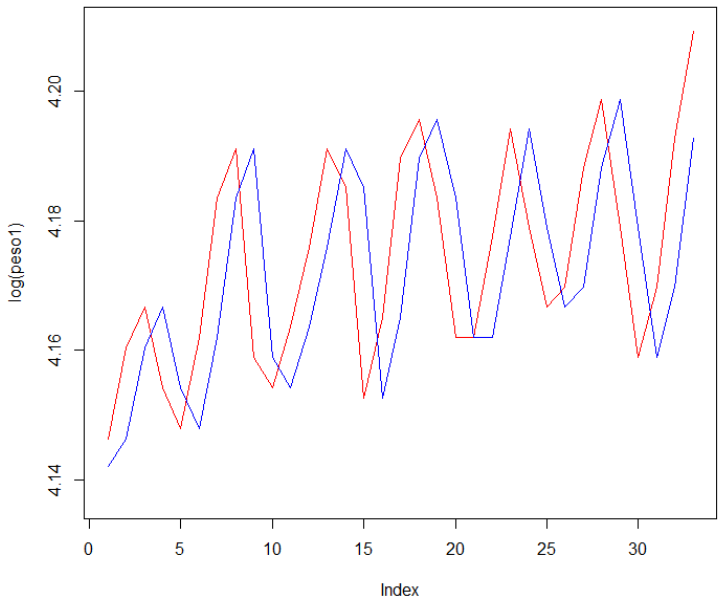

Figure 3.4: Red line represents the first subseries in the log scale and blue line is fit of $ARIMA(0,1,0)$ model.

From Figure 3.4 we can notice that $ARIMA(0,1,0)$ model estimation has a small delay with respect to the original data.

For residual assumption checking, Box-Ljung test was used and the p-value= 0.1527 is obtained If p-value > 0.05: there is no enough statistical evidence to reject the null hypothesis. So it can not be assumed that values are dependent. This could mean that values are dependent anyway or it can mean that they are independent. But it is not proven any specific possibility, what test actually said is that we can not assert the dependence of the values, neither can we assert the independence of the values. Regarding this, ACF and PACF plots are drawn and we can see them in Figure 3.5.

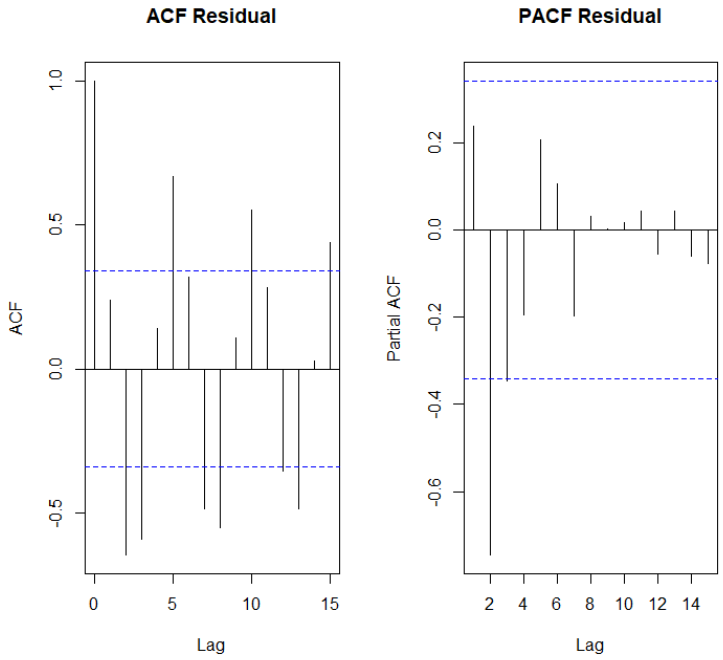

Figure 3.5: ACF and PACF of residuals for model $ARIMA(0,1,0)$, for the time series of first phase.

Autocorrelation plots are a commonly-used tool for checking randomness in a data set. This randomness is ascertained by computing autocorrelations for data values at varying time lags. If random, such autocorrelations should be near zero for any and all time-lag separations. If non-random, then one or more of the autocorrelations will be significantly non-zero.
The above plot also contains several horizontal reference lines. The middle line is at zero. The other two lines are 95% confidence bands.
The ACF plot shows that there is some autocorrelation remaining in the residuals. This means there is some information remaining in the residuals that can be exploited to obtain better forecasts.

The best model among all possibles with $d = 2$

The best model in this case is $ARIMA(0, 2, 0)$ and its AIC value is -153.34. Fitted plot for this model is represented on Figure 3.6.

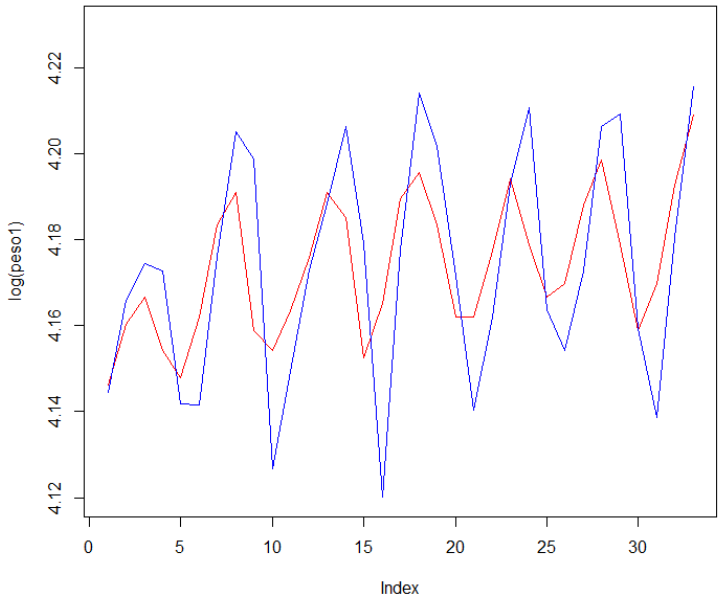

Figure 3.6: Red line represents the first subseries in log scale and blue line is fit of $ARIMA(0, 2, 0)$ model

From Figure 3.6 we can notice that $ARIMA(0, 2, 0)$ model is weak in fitting extremes values of the weight.

P-value $= 0.6254$ of Box-Ljung test for residual correlation is bigger than the significance level $\alpha = 0.05$, which tells us that it can not be assumed that estimated residuals are dependent.

Figure 3.7 represents ACF and PACF plots of residuals for $ARIMA(0,2,0)$ model.

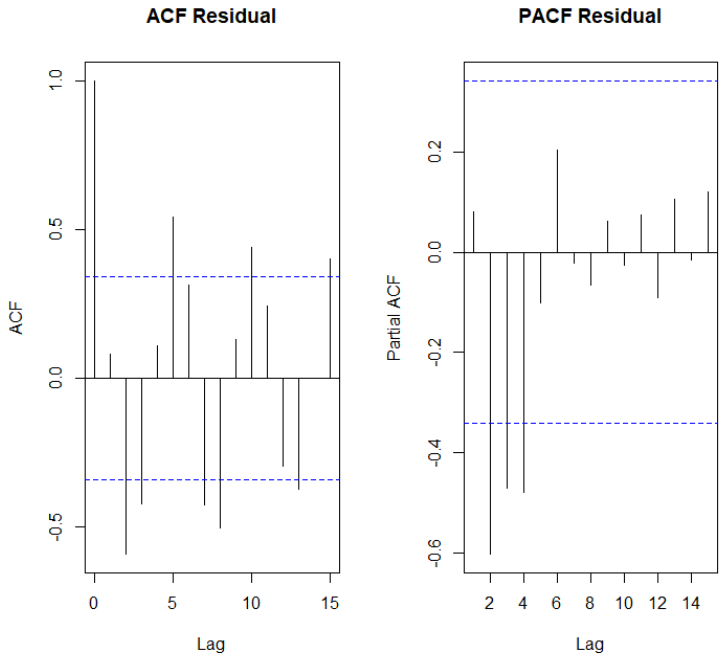

Figure 3.7: ACF and PACF of residuals for model $ARIMA(0,2,0)$.

According to ACF and PACF plots in Figure 3.7 residuals of estimated model $ARIMA(0,2,0)$ are dependent.

From the results obtained in this section, the best model that we got for the first subseries is $ARIMA(0,2,0)$, since for its residuals we have stronger evidence for stationarity and the difference between AIC value of $ARIMA(0,2,0)$ and $ARIMA(0,1,0)$ is small.
The above selected model ARIMA(0,2,0) means that we are fitting ARMA(0,0) model of second order difference to our time series, i.e second order differencing to our time series gives times series that is iid with zero mean. Therefore, this model can be expressed as:

$$Y_t - 2Y_{t-1} + Y_{t-2} = \varepsilon_t$$

3.2.2 Fittig ARIMA Models to the second and third phases of the weight time series

In second and third subseries missing values occur. They are shown on Figure 3.2. For dealing with these there are few methods:

1. Ignoring NA values - we ignore them and use *auto.arima* function on original subseries to fit the best models;

2. Predict NA values - we iteratively predict missing values in following steps:

   (a) If NA value occurs at time t but weight data are available at times $t-1$ and $t+1$ we do not have missing data, then the estimate of data x_t is the average of x_{t-1} and x_{t+1}, meaning $x_t = (x_{t-1} + x_{t+1})/2$.
   For example, on April 18th 2014, the weight is missing, but the day before, on April 17th 2014, and the day after, April 19th 2014, weight is measured. If we denote as x_{170414}, x_{180414} and x_{190414} weights on 17th April, 18th April and 19th April 2014 retrospectively, our estimated value for 18th April 2014 is given by $x_{180414} = (x_{170414} + x_{190414})/2$

   (b) If we have sequence of missing data at times $t, t+1, ..., t+k$, then *auto.arima* is used to fit the best model to series $x_1, x_2, ..., x_{t-1}$ and then data $x_t, x_{t+1}, ..., x_{t+k}$ are estimated with help of *predict* function, forecast from models fitted by auto.arima and as a result gives a time series of predictions, used on the best model that we got for data $x_1, x_2, ..., x_{t-1}$.
   For example, from February 7th 2015 until February 14th 2015 , the weight is missing (in total 7 missing values), but, after using method (a) for estimating $x_{1804214}$, there are no NA values before 7th February 2015, so first thing that is done is using auto.arima function on time series starting from 18.11.2013 until 06.02.2015 to fit the best model, and then prediction function is used to estimate new 7 data given the best model for data until 07.02.2015.

3. Interpolation of missing data. NA values are estimated using *na.interpolation* function from R package *imputeTS* which fills in missing values according to some polynomial function.

The most accurate method from the three methods mentioned above is the one that use *na.interpolation* function. For the first method, the disadvantage is that NA values occur randomly, in not equally spaced times. As for second method, it happens that we have NA values at times $t, t+1, ..., t+k$, but the best model for data $x_1, x_2, ..., x_{t-1}$ is such that observation at time m depends

on l previous data with $l < k$, meaning, that for estimation of missing data at times $x_{t+l+1}, ..., x_{t+k}$ we use only estimated values and not original data. Estimation of missing values using *na.interpolation* function for second and third subseries are shown on Figure 3.8 and Figure 3.9:

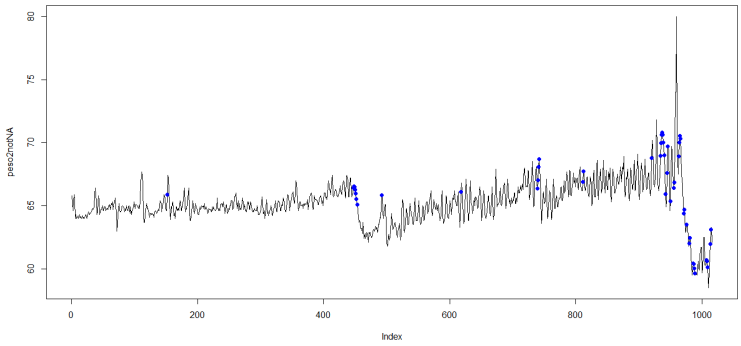

Figure 3.8: Second phase of time series for the weights in log scale: blue dots are estimations of NA values using function *na.interpolation*.

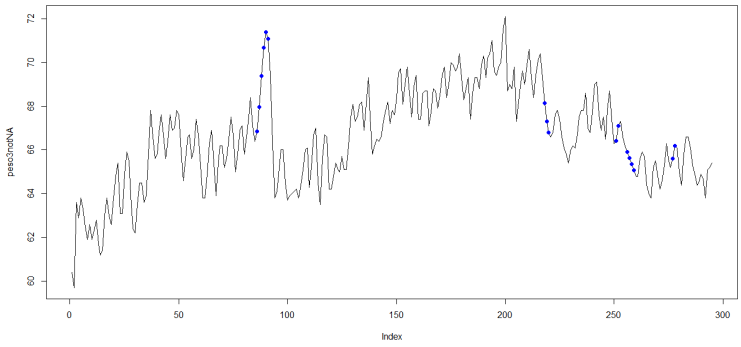

Figure 3.9: Third phase of time series for the weights in log scale: blue dots are estimations of NA values using function *na.interpolation*.

We estimate parameter d for the second and the third subseries, as we did it for the first phase, and we got that difference at lag 1 gives us stationary data. Therefore we will search for the best $ARIMA(p, 1, q)$ model, but also for the best model among all possibles with $d = 2$.

Using *pp.test* (that represents Phillips-Perron Unit Root Test in R studio) and *adf.test* (that represents Augmented Dickey-Fuller Test in R studio) stationarity of difference at lag 1 of the second and the third part of data are checked and the results are:

- For the second phase:
  both tests give p-value=0.01, which allows us to reject null hypothesis and accept that data are stationary.

- For the third phase:
  the results of tests are the same as for the second subseries, p-value=0.01, data are stationary.

ARIMA Models for the second phase

The second part of data is from 18.11.2013 until 31.08.2016.

The best model among all possible with $d = 1$ is $ARIMA(4, 1, 0)$ and its AIC value is -6380.66; the fitted plot for this model is represented in Figure 3.10.

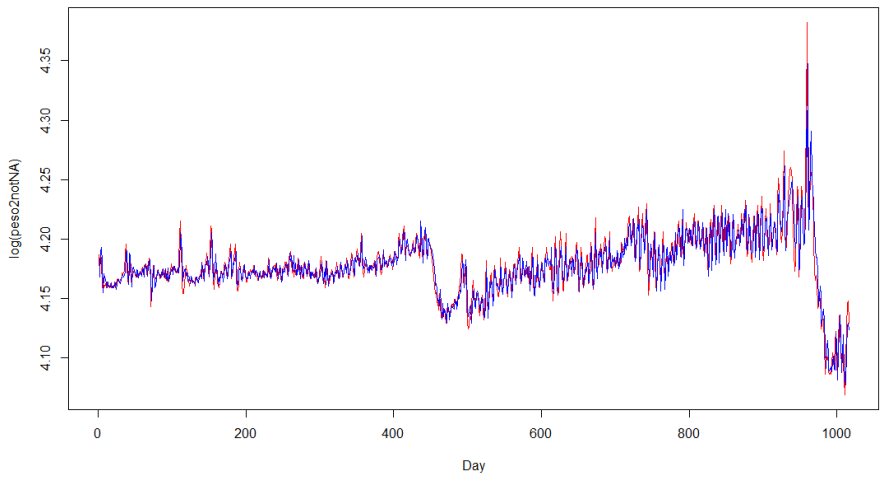

Figure 3.10: Red line represents second subseries in log scale: blue line is the fit of $ARIMA(4, 1, 0)$ model.

Next, we will investigate the forecast errors of our ARIMA(4,1,0) model, whether there are any correlations between successive forecast errors.
Ljung-Box test gave as result p-value=0.796, large p-value in the test is suggesting us to accept the null hypothesis that all of the autocorrelation functions are zero. In other words, we can conclude that there is no evidence for non-zero autocorrelations in the forecast errors in our fitted model.

ACF and PACF plots are shown in Figure 3.11:

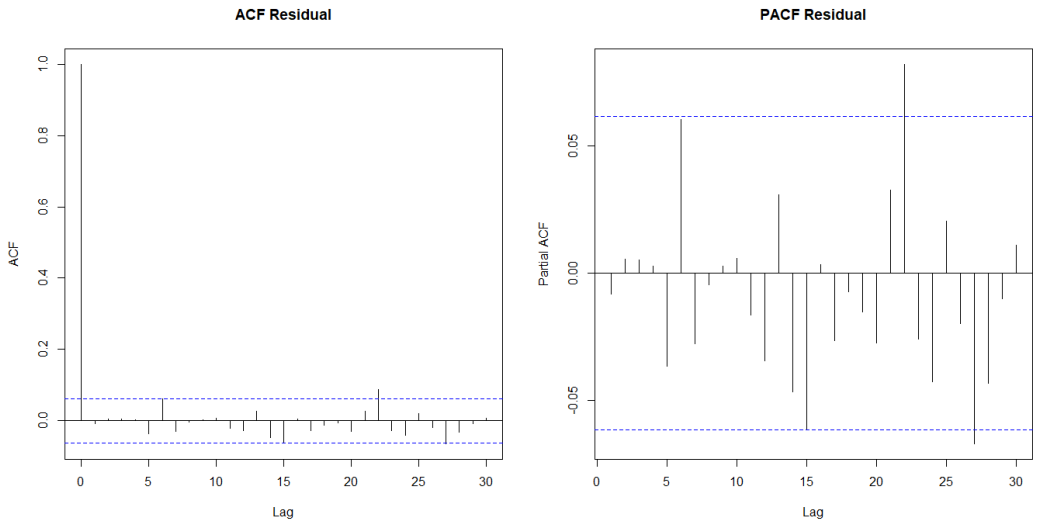

Figure 3.11: ACF and PACF of residuals for model $ARIMA(4,1,0)$.

On the other hand, from Figure 3.11 we can conclude that there is no significant autocorrelation between residuals for model $ARIMA(4,1,0)$. It is clearly evident from the ACF and PACF plots in Figure 3.11 above that none of the autocorrelation coefficients between lag 1 and 30 are breaching the significant limits i.e. all the ACF values, except two, are well within the significant bounds. This means ACF and PACF concluded that there is no autocorrelations in the forecast residuals (or standard errors) at lag 1 to 30 in the fitted ARIMA(4,1,0) model.

The best model among all possible with $d = 2$ is $ARIMA(5, 2, 0)$ and its AIC value is -6118.28; the fitted plot for this model is represented in Figure 3.12:

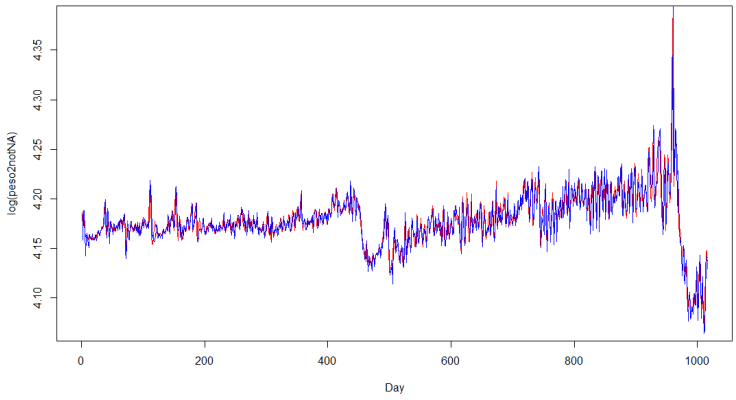

Figure 3.12: Red line represents the second phase data in log scale: blue line is fit of $ARIMA(5, 2, 0)$ model.

As far as the analysis of residuals correlation:
Ljung-Box test gave as result p-value=0.01975, small p-value tells us that it can be assumed that estimated residuals from $ARIMA(5, 2, 0)$ model are dependent. To investigate further whether there are any correlation between sucessive forecast errors, we will plot the ACF and PACF of the forecast errors. Figure 3.13 represents ACF and PACF plots of estimated residuals for $ARIMA(5, 2, 0)$ model.

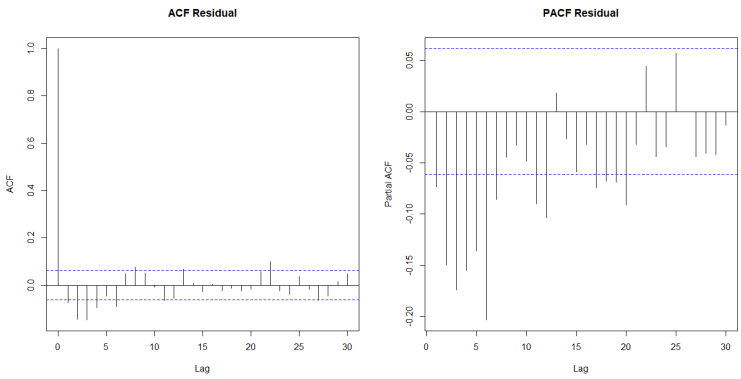

Figure 3.13: ACF and PACF of estimated residuals for model ARIMA(5,2,0).

By looking at Figure 3.13 seams that the residuals are correlated.

ARIMA Models for the third phase

The third phase of data is from 01.09.2016 until 20.06.2017.

The best model among all possible with $d = 1$ is $ARIMA(4, 1, 2)$ and its AIC value is -1714.59; the fitted plot for this model is represented in Figure 3.14:

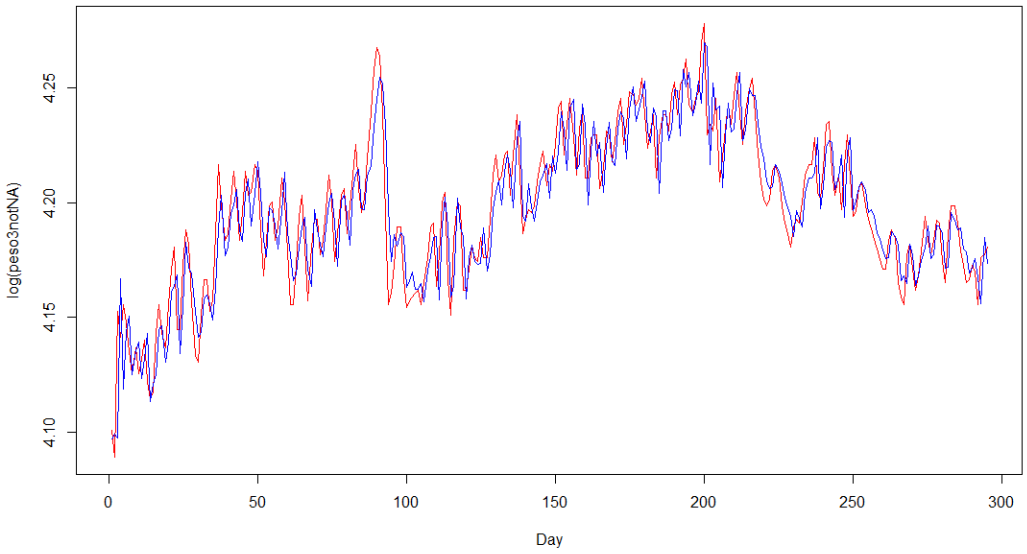

Figure 3.14: Red line represents third part in log scale: blue line is the fit of $ARIMA(4, 1, 2)$ model.

As far as analysis of residuals correlation:
Ljung-Box test gave as result p-value=0.9435, which tells us that it can not be assumed that estimated residuals for $ARIMA(4, 1, 2)$ model are dependent. This is confirmed by the ACF and PACF plots of residuals autocorrelation between lag 1 and lag 25. ACF and PACF plots of residuals for $ARIMA(4, 1, 2)$ model are shown in Figure 3.15.

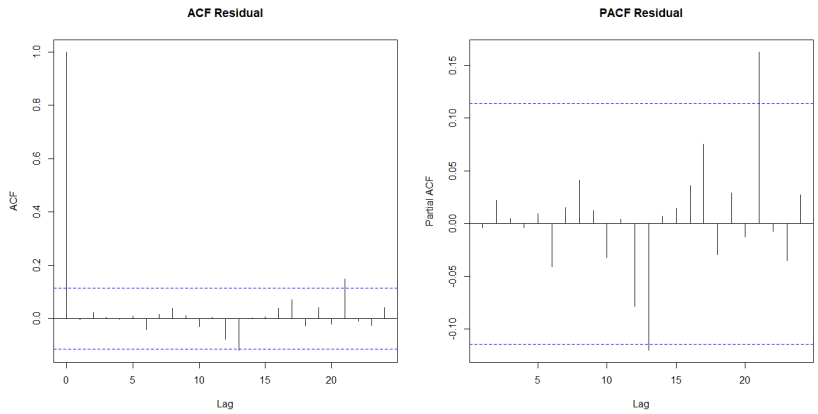

Figure 3.15: ACF and PACF of residuals for model $ARIMA(4,1,2)$.

From Figure 3.15 we can conclude that there is no significant autocorrelation between residuals for model $ARIMA(4,1,2)$.

The best model among all possible with $d = 2$ is $ARIMA(5,2,0)$ and its AIC value is -1635.28; the fitted plot for this model is represented in Figure 3.16:

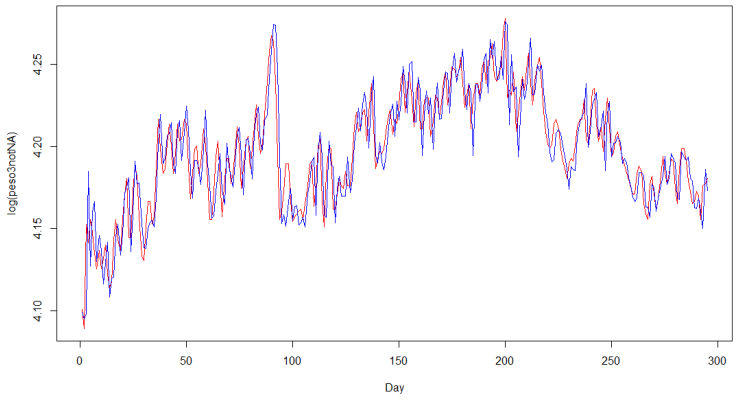

Figure 3.16: Red line represents the third subseries in log scale: blue line is fit of $ARIMA(5,2,0)$ model.

Moving to analysis of residuals correlation:
Ljung-Box test gave as result p-value=0.3645, which tells us that it can not be assumed that estimated residuals for $ARIMA(5,2,0)$ model are dependent.
Figure 3.17 represents ACF and PACF plots of estimated residuals for $ARIMA(5,2,0)$ model.

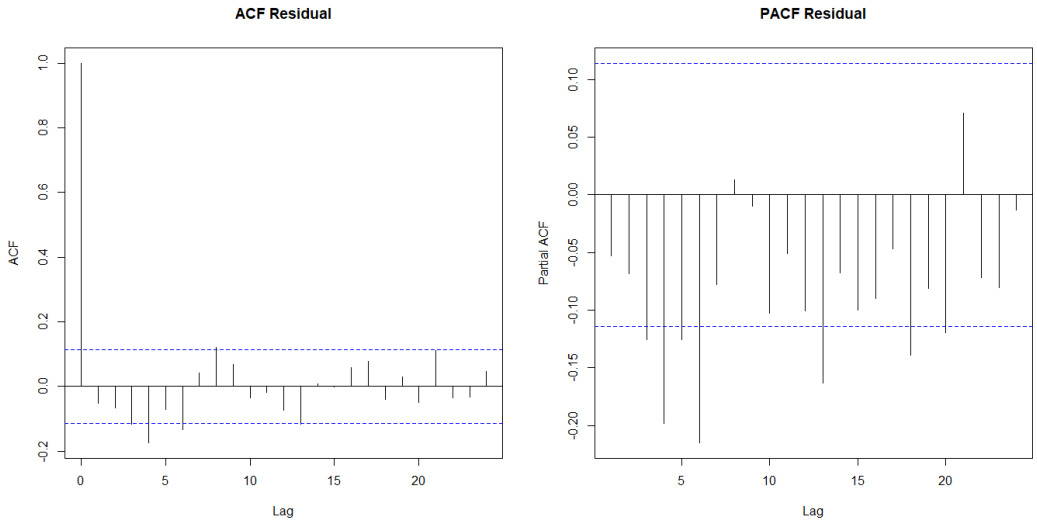

Figure 3.17: ACF and PACF of residuals for model ARIMA(5,2,0).

Furthermore, from Figure 3.17 it can be observed that the estimated residuals are uncorrelated.

3.2.3 Conclusion

Finally, Table 3.1 summarises the result obtained in this section for each of the fitted ARIMA model in our three subseries.

Phase	ARIMA	AIC	Ljung-Box
First	ARIMA$(0,1,0)$	-176.56	0.1527
	ARIMA$(0,2,0)$	-153.34	0.6254
Second	ARIMA$(4,1,0)$	-6380.66	0.976
	ARIMA$(5,2,0)$	-6118.28	0.01975
Third	ARIMA$(4,1,2)$	-1714.59	0.9435
	ARIMA$(5,2,0)$	-1635.28	0.3645

Table 3.1: Table of results

The best fit model is selected based on Akaike Information Criterion (AIC) value. The ideq is to choose a model with minimum AIC value. In Table 3.1 are displayed AIC values for models, as well as p-values of Ljung-Box test for correlation of estimater residuals for each model.

In addition, the best model that we get for the first subseries is $ARIMA(0,2,0)$, since for its residuals we have stronger evidence for stationarity and the difference between AIC value of $ARIMA(0,2,0)$ and $ARIMA(0,1,0)$ is small.
As for second and third subesries, we can cleraly observe in the table above that the lowest AIC values are for $ARIMA(4,1,0)$ for second part of the body weight time series and for $ARIMA(4,1,2)$ for third phase. Another thing we can validate by looking at p-values of Ljung-box test that the forecast errors of this two models are not correlated and hence this two models can be the best predictive models for making forecasts of the second and the third subseries.

3.3 Fitting ARIMA models to nine subseries of weight

Our goal in this section is to inspect if NA values that occur in weight time series affect in fitting ARIMA models to data. In order to check results we obtained in Section 3.3 and Section 3.4 (they are shown in Table 3.1), the second and third time subseries of weight are divided into totally 9 parts, with respect to missing data (sequences of NA values are skipped, if only one day with NA value occurs, its value is estimated with the average of previous and following day).

New subseries are:

1. from 18.11.2013 until 06.02.2015,
2. form 14.02.2015 until 21.11.2015,
3. from 30.11.2015 until 05.02.2016,
4. from 08.02.2016 until 07.06.2016,
5. from 30.08.2016 until 22.11.2016,
6. from 29.11.2016 until 03.04.2017,
7. from 07.04.2017 until 06.05.2017,
8. from 16.05.2017 until 01.06.2017,
9. from 04.06.2017 until 20.06.2017,

On Figure 3.18 are displayed plots for all nine subseries:

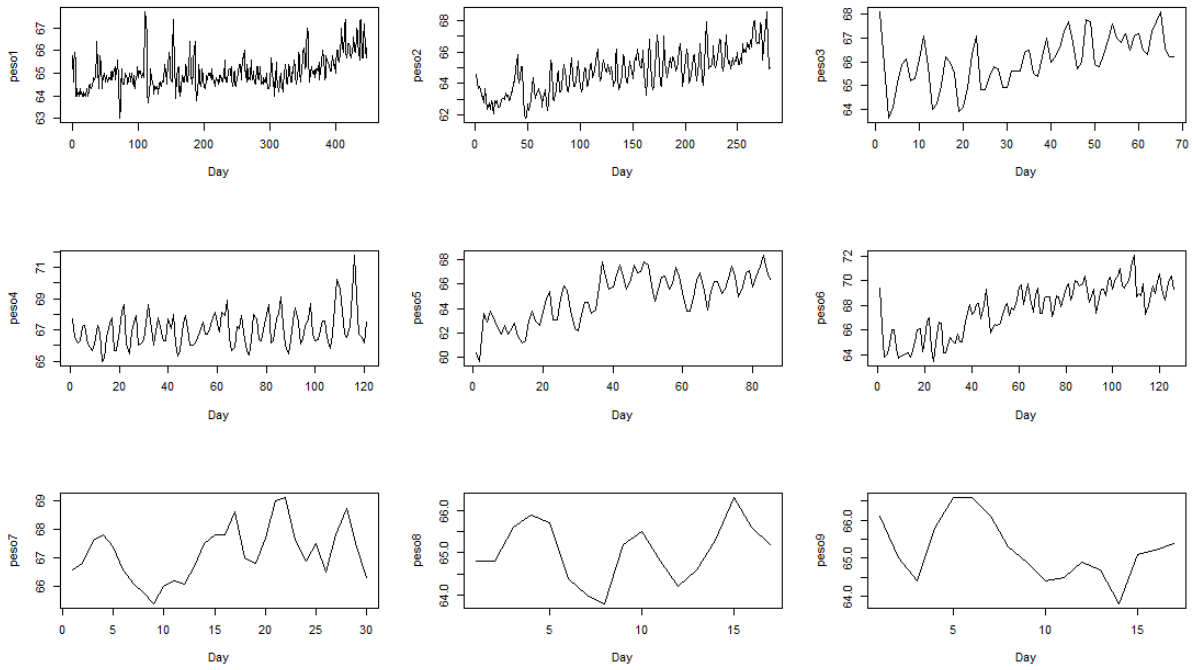

Figure 3.18: Nine subseries of time series of weight.

First thing we do is to find the best model among all $ARIMA(p, d, q)$ models with $d \in \{0, 1\}$ and the best model among all $ARIMA(p, d, q)$ models with $d = 2$, in both cases *auto.arima* function is used. Next thing that is going to be done is to compare results obtained for nine subsreries with the results obtained for the second and the third phase in Section 3.4.

Period	ARIMA	AIC	Box-Ljung
18.11.2013-06.02.2015	ARIMA$(3, 1, 3)$	–3258.03	0.705
	ARIMA$(5, 2, 0)$	-3101.23	0.07866
14.02.2015-21.11.2015	ARIMA$(3, 1, 4)$	-1872.7	0.9023
	ARIMA$(5, 2, 0)$	-1872.3	0.05334
30.11.2015-05.02.2016	ARIMA$(2, 1, 2)$	-425.67	0.6565
	ARIMA$(2, 2, 2)$	-411.46	0.8581
08.02.2016-07.06.2016	ARIMA$(0, 1, 0)$	-647.54	0.08204
	ARIMA$(5, 2, 0)$	-681.69	0.3742
30.08.2016-22.11.2016	ARIMA$(3, 1, 0)$	-475.53	0.7711
	ARIMA$(5, 2, 0)$	-446.31	0.8435
16.05.2017-01.06.2017	ARIMA$(0, 0, 2)$	-109.48	0.9104
	ARIMA$(0, 2, 0)$	-83.86	0.524
04.06.2017-20.06.2017	ARIMA$(0, 1, 0)$	-96.85	0.7667
	ARIMA$(0, 2, 0)$	-82.28	0.2204

Table 3.2: Table of results

From Table 3.2 it can be noticed that data from 07.04.2017 until 06.05.2017, data from 16.05.2017 until 01.06.2017 and data from 04.06.2017 until 20.06.2017 are different from other 6 subseries of weight time series, which may be considered as due to sort time interval of this three subseries.

Comparing results in Table 3.1 and Table 3.2 it can be also observed that there is no need for dividing data into 9 subseries, that division in three phases is good enough for analysis. Also it can be perceived that the best models for data from 18.11.2013 until 31.08.2016 and data from 01.09.2016 until 20.06.2017 in both cases are actually $ARIMA(5, 2, 0)$, since for all six subseries (not including last three), that were analysed in this section, have $ARIMA(5, 2, 0)$ as one of the two options for best model.

3.4 Seasonality

In time series data, seasonality is the presence of variations that occur at specific regular intervals called period. If seasonality is present, it must be incorporated into the time series model. Multiple boxplots can be used as a tool to detect seasonality.The boxplot shows the seasonal difference between group patterns quite well, but it does not show within group patterns. If there is significant seasonality, the boxplots between groups shoud differ.

For time series for weight from 18.11.2013 until 31.08.2016, it will be checked two seasonalities, one with period $d = 5$ and another with period d $d = 20$, owing to the fact that the period between two consecutive immunoglobuline(IG) therapies is in avegare 21 days, the first subseries contains all first days of IG therapy, the second subseries contains all second days of IG therapy, and so on until the twentieth subseries which contains all Twentieth days ot therapy.

Seasonality with period $d = 5$

In order to check this seasonality, data from 18.11.2013 until 29.08.2016 is divided into five subseries in following way:
If $x_1, x_2, x_3, ..., x_n$ represent the weight from 18.11.2013 until 31.08.2016, than the first subseries contains all x_i such that $i \mod 5 = 1$, i.e. $x_1, x_6, x_{11}, ...$, the second subseries contains all x_i such that $i \mod 5 = 2$, i.e. $x_2, x_7, x_{12}, ...$, and so on until the fifth subseries which contains all x_i such that $i \mod 5 = 0$, i.e. $x_5, x_{10}, x_{15}, ...$

Figure 3.19 shows the boxplots for five subseries:

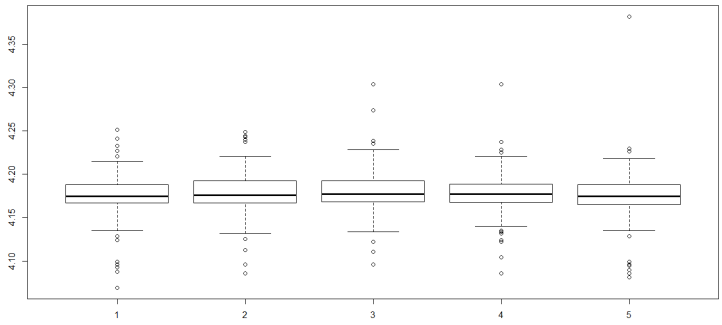

Figure 3.19: Box plots for five subseries in log scale.

From the Figure 3.19 it can be observed that there is no seasonality with period $d = 5$, since all five boxplots have almost identical shape.

Seasonality with period $d = 20$

Since the period between two consecutive immunoglobuline therapies is 20 days, goal of examining this seasonality is to see if therapy influences weight of patient.

In order to check this seasonality, data from 18.11.2013 until 29.08.2016 is divided into twenty subseries with respect to first day of immunoglobulin therapy, meaning that first subseries contains all first days of therapy, second subseries contains all second days and so on.

On Figure 3.20 are shown box plots for twenty subseries:

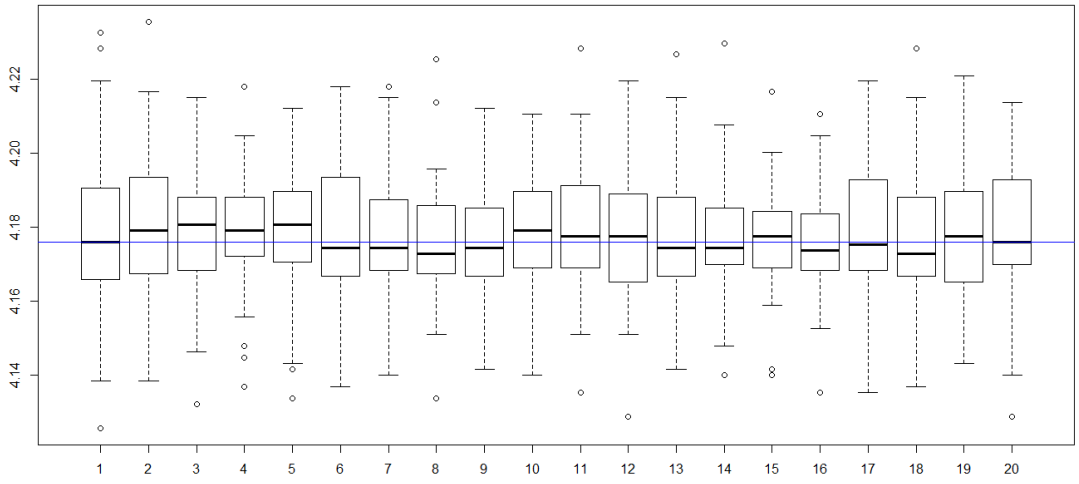

Figure 3.20: Box plots for twenty subseries in log scale: verical line is median of data from 18.11.2013 until 29.08.2016.

From the Figure 3.20 we can conclude that there is no also seasonality with period $d = 20$.

3.5 Analysis of blood pressure and heart beat

In this section we fit Autoregressive Integrated Moving Average (ARIMA) models to the time series for the patient's diastolic and systolic blood pressure and for patient's heart beat. Analysis in this section follows the same steps as analysis in Section 3.2.

3.5.1 ARIMA model for heart beat rate

Time series for heart rate (HR) of the patient is shown in Figure 3.21.

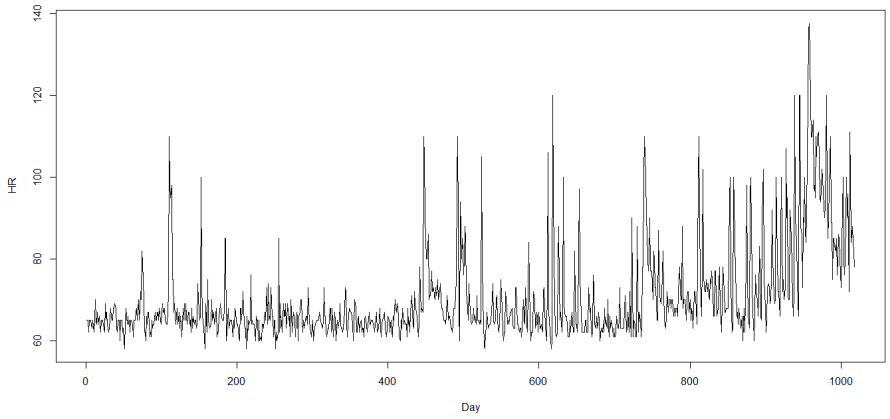

Figure 3.21: Heart beat rate time series.

The best model among all possible with $d = 1$ is $ARIMA(2, 1, 4)$and its AIC value is 6970.23; the fitted plot for this model is represented in Figure 3.22:

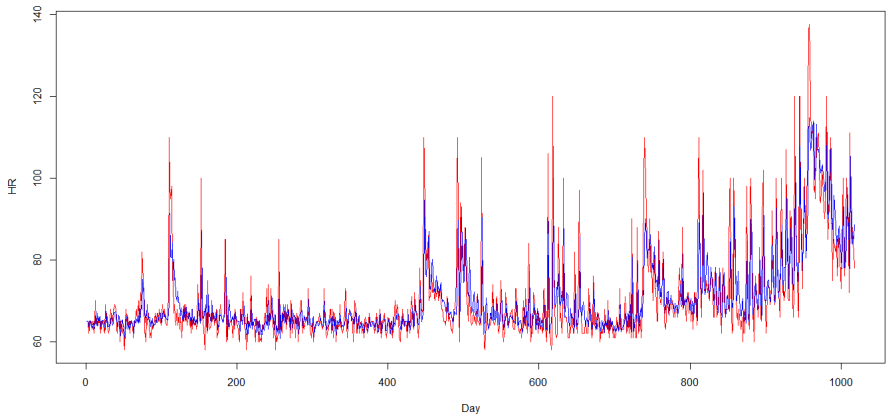

Figure 3.22: Red line represents heart rate time series: blue line is the fit of $ARIMA(2,1,4)$ model.

As far as the analysis of residual correlation:
Ljung-Box test gave as result p-value=0.9536, large p-value in the test suggests that all autocorrelation functions are zero, meaning that we can conclude that there is no evidence for non-zero autocorrelation in the estimated residuals for $ARIMA(2,1,4)$ model.

ACF and PACF plots are shown in Figure 3.23:

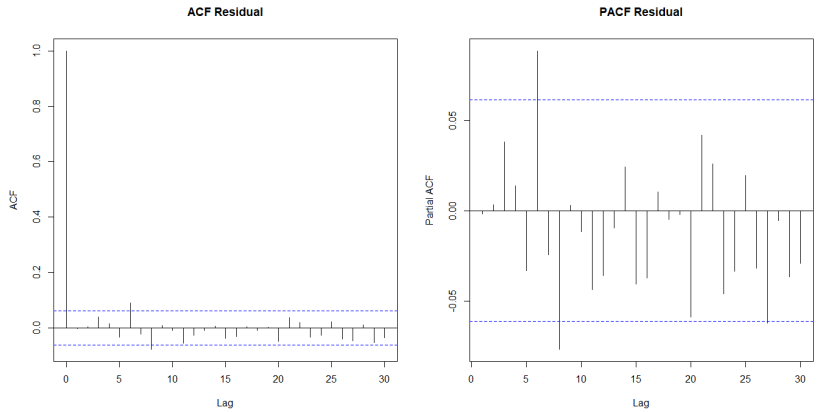

Figure 3.23: ACF and PACF of estimated residuals for $ARIMA(2,1,4)$ model.

From the ACF and PACF plots in Figure 3.23 it is evident that there is no autocorrelation in the forecast residuals in the fitted $ARIMA(2,1,4)$ model.

The best model among all possible with $d = 2$ is $ARIMA(5, 2, 0)$ and its AIC value is 7319.01; the fitted plot for this model is represented in Figure 3.24:

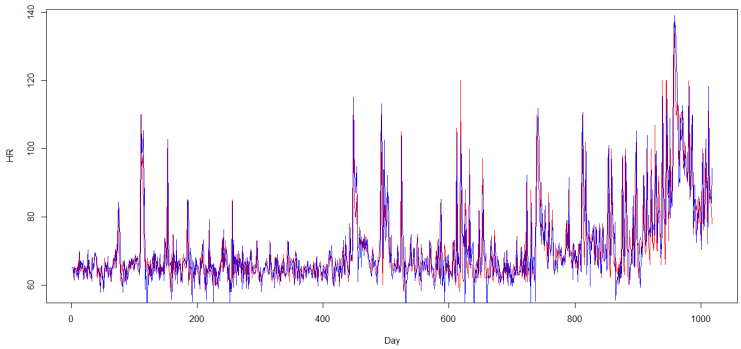

Figure 3.24: Red line represents heart rate time series: blue line is the fit of $ARIMA(5, 2, 0)$ model.

As far as the analysis of residual correlation:
Ljung-Box test gave as result small p-value, which tells us that it can be assumed that the estimred error terms from $ARIMA(5, 2, 0)$ model are dependent. To be sure about this, we will plot ACF and PACF of residuals. Figure 3.25 represents ACF and PACF plots of estimated residuals for $ARIMA(5, 2, 0)$ model.

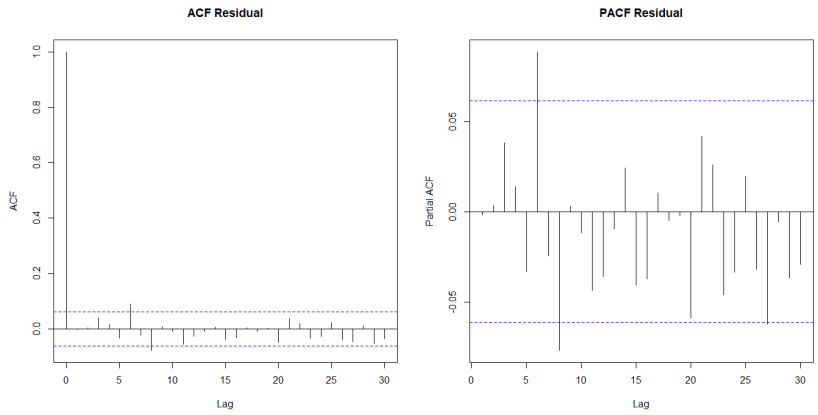

Figure 3.25: ACF and PACF of estimated residuals for $ARIMA(5, 2, 0)$ model.

By looking at Figure 3.25 seems that the residuals are correlated.

3.5.2 ARIMA model for systolic blood pressure

Time series for systolic blood pressure (SBP) of the patient is shown in Figure 3.26.

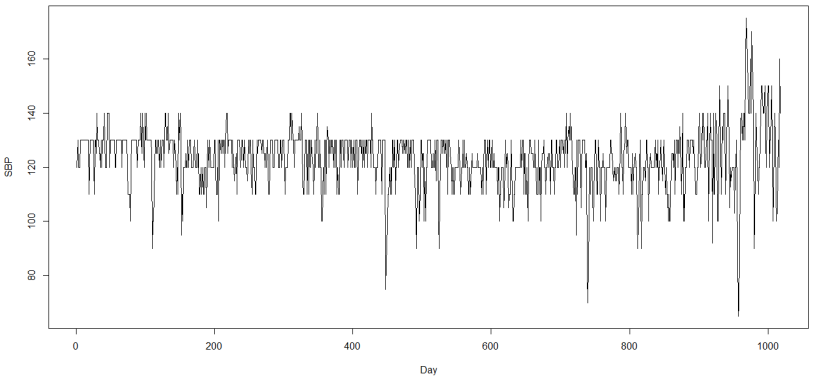

Figure 3.26: Systolic blood pressure time series.

The best model among all possible with $d = 1$ is $ARIMA(4, 1, 3)$ and its AIC value is 7446.87; the fitted plot for this model is represented in Figure 3.27:

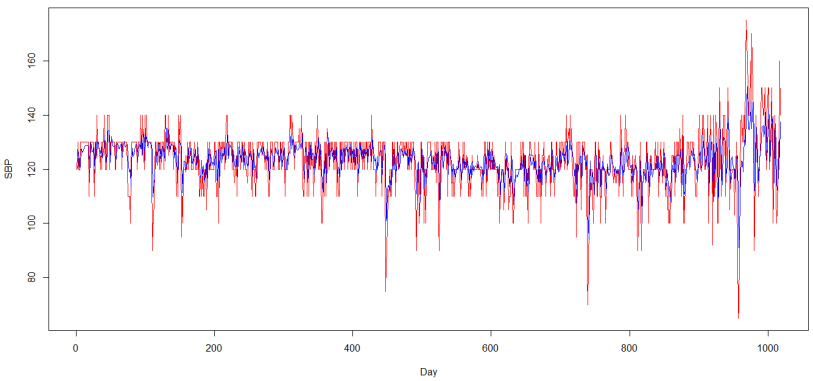

Figure 3.27: Red line represents systolic blood pressure time series: blue line is the fit of $ARIMA(4, 1, 3)$ model.

As far as analysis of estimated errors:
Ljung-Box test gave as the result p-value=0.9438, which tells us that it can not be assumed that estimated residuals for $ARIMA(4,1,3)$ model are dependent. This is confirmed by the ACF and PACF plots of residuals autocorrelation. ACF and PACF plots of residuals for $ARIMA(4,1,3)$ model are shown in Figure 3.28.

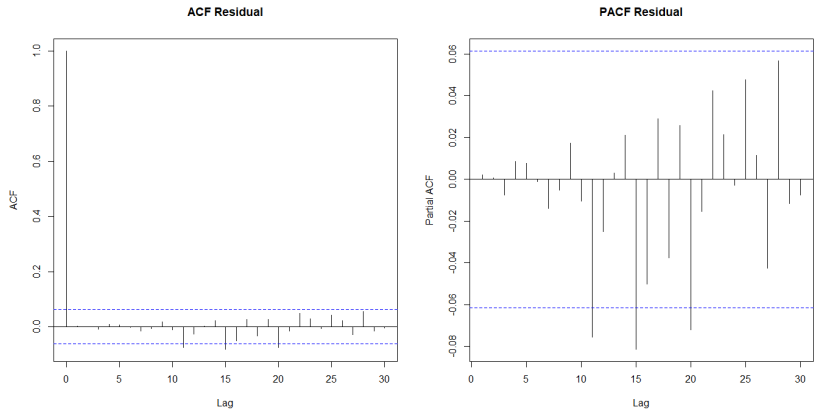

Figure 3.28: ACF and PACF of estimated residuals for $ARIMA(4,1,3)$ model.

The best model among all possible with $d = 2$ is $ARIMA(1,2,2)$ and its AIC value is 7479.16; the fitted plot for this model is represented in Figure 3.29:

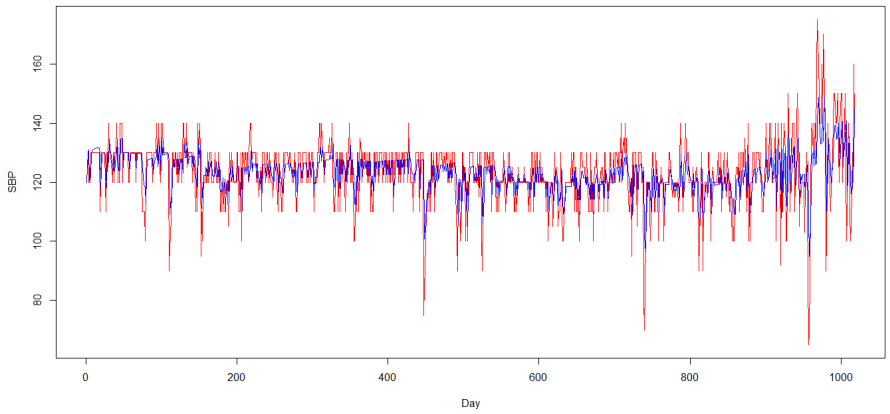

Figure 3.29: Red line represents systolic blood pressure time series: blue line is the fit of $ARIMA(1,2,2)$ model.

Moving to analysis of residuals correlation: Ljung-Box test gave as result p-value=0.661, which tells that it can not be assumed that estimated residuals for $ARIMA(1,2,2)$ model are dependent.
Figure 3.30 represents ACF and PACF plots of estimated residuals for $ARIMA(1,2,2)$ model.

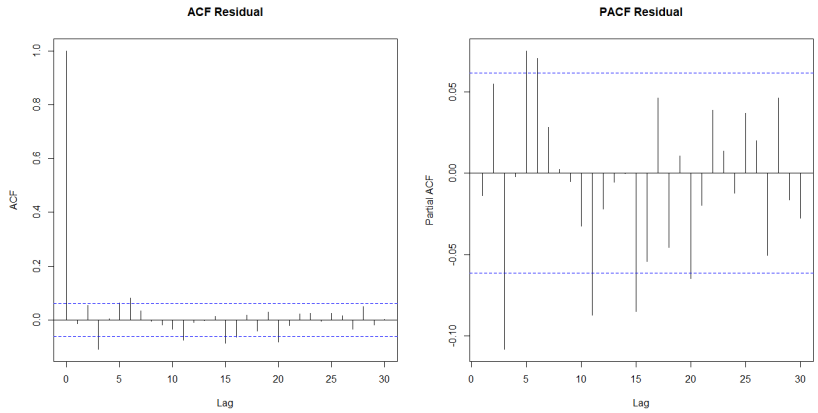

Figure 3.30: ACF and PACF of estimated residuals for $ARIMA(4,1,3)$ model.

From Figure 3.30 we conclude that the forecast errors are uncorrelated.

3.5.3 ARIMA model for diastolic blood pressure

Time series for diastolic blood pressure of the patient is shown in Figure 3.31.

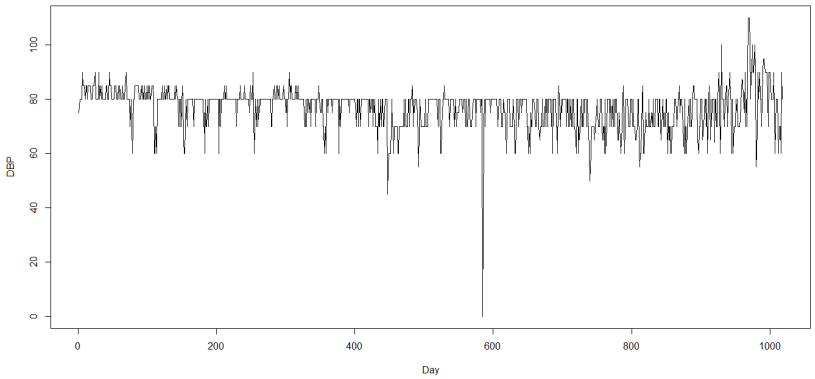

Figure 3.31: Diastolic blood pressure time series.

The best model among all possible with $d = 1$ is $ARIMA(1, 1, 2)$ and its AIC value is 6765.2.

As far as the residuals correlation: Ljung-Box test gave as result p-value=0.9776, which tell us that there is no evidence for non-zero autocorrelation in the estimated residuals for $ARIMA(1, 1, 2)$ model. From ACF and PACF plots of errors, shown in Figure 3.32, we can confirm that residuals are not dependent.

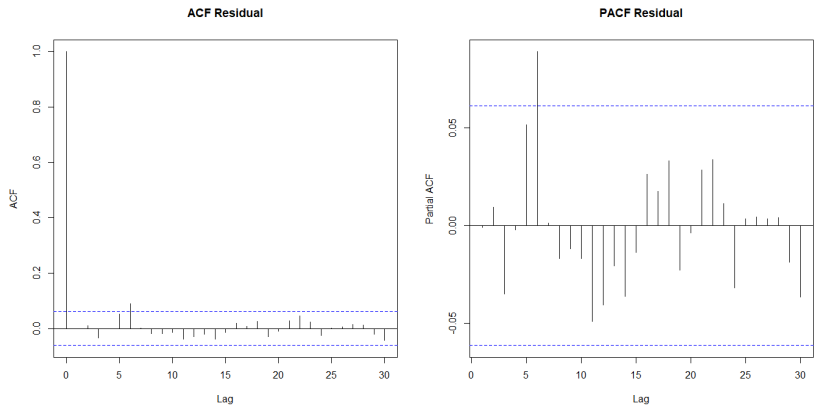

Figure 3.32: ACF and PACF of estimated residuals for $ARIMA(1, 1, 2)$ model.

The best model among all possible with $d = 1$ is $ARIMA(5, 2, 0)$ and its AIC value is 7174.61.

Moving to the residuals correlation: Ljung-Box test gave as result p-value=0.009552, small p-value suggests that it can be assumed thet the estimated residuals of $ARIMA(5, 2, 0)$ model are correlated.

ACF and PACF plot are displayed in Figure 3.33:

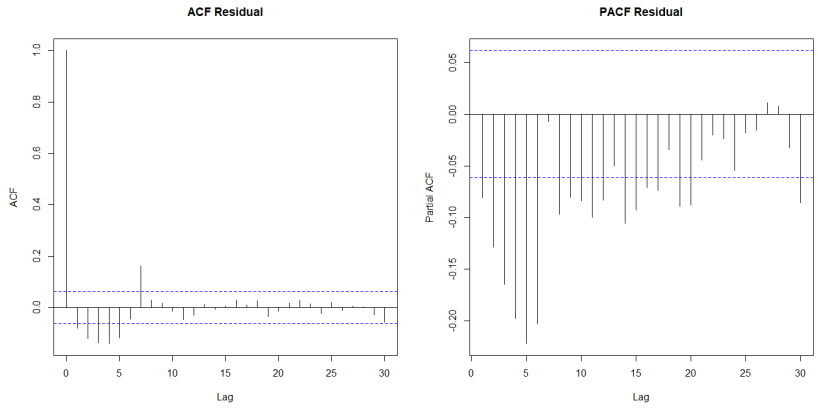

Figure 3.33: ACF and PACF of estimated residuals for $ARIMA(5, 2, 0)$ model.

Looking at Figure 3.33 it can be concluded that residuals for $ARIMA(5, 2, 0)$ model are dependent.

3.5.4 Desriptive analysis of blood pressure and heart beat

The goal of analysis in this section is to see if there are big changes in blood pressure and heart beat rate on the day of crisis comparing to the day before and day after the crisis.

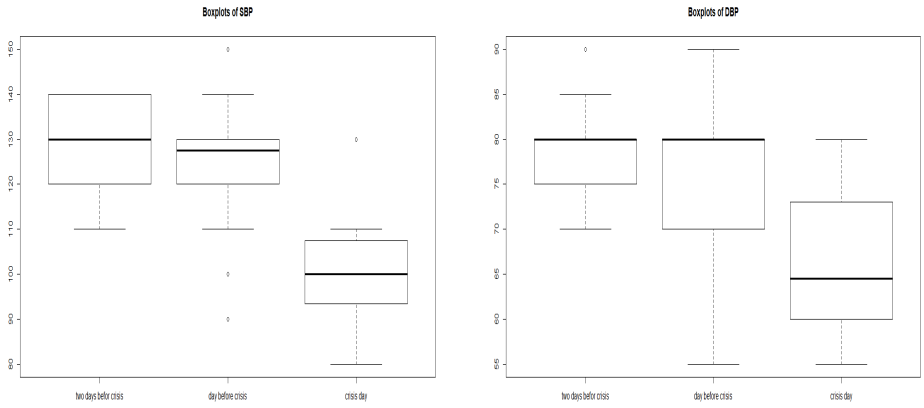

Figure 3.34: Left figure: boxplots of systolic blood pressure; right figure: boxplots of diastolic blood pressure.

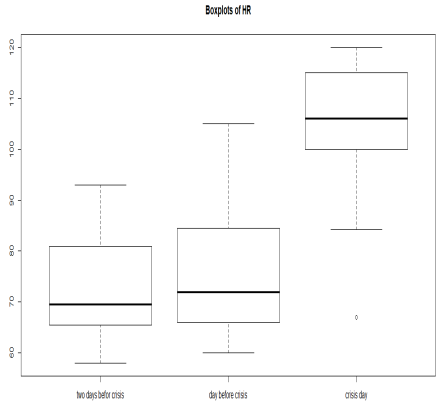

Figure 3.35: Boxplots of heart beat rate

From the boxplots in Figure 3.34 and Figure 3.35 we can see that the systolic and diastolic pressure are decreasing during crisis days, while heart beat rate is increasing.
Average decrease of the systolic pressure from day before crisis to day of crisis is equal to 23, while from two days before crisis to crisis day SBP decreases in average for 28.25.
Average decrease of the diastolic pressure from day before crisis to day of crisis is equal to 10, while from two days before crisis to crisis day SBP decreases in average for 12.
Average increase of heart beat from day before crisis to day of crisis is equal to 26, while from two days before crisis to crisis day HR increases in average for 29.5.

3.5.5 Conclusion

Table 3.3 summarises the result obtained in this section for each of the fitted ARIMA model in our three time series.

	ARIMA	AIC	Ljung-Box
HR	ARIMA$(2,1,4)$	6970.23	0.9536
	ARIMA$(5,2,0)$	7319.01	6.396×10^{-5}
SBP	ARIMA$(4,1,3)$	7446.87	0.9438
	ARIMA$(1,2,2)$	7479.16	0.661
DBP	ARIMA$(1,1,2)$	6765.	0.9776
	ARIMA$(5,2,0)$	7174.61	0.009552

Table 3.3: Table of results

The best fit model is selected based on Akaike Information Criterion (AIC) value. The ideq is to choose a model with minimum AIC value. In Table 3.3 are displayed AIC values for models, as well as p-values of Ljung-Box test for correlation of estimated residuals for each model.

In addition, the best model that we get for the heart rate is $ARIMA(2,1,4)$, since for its residuals we have stronger evidence for stationarity and the smaller AIC value.
As for blood pressure series, we can clearly observe in the table above that the lowest AIC values are for $ARIMA(4,1,3)$ for systolic blood pressure time series and for $ARIMA(1,1,2)$ for diastolic blood pressure. Another thing we can validate by looking at p-values of Ljung-box test that the forecast errors of this two models are not correlated and hence this two models can be the best predictive models for making forecasts of the systolic and diastolic blood pressure time series.

3.6 Analysis of data collected in hospital

In this section we will use data set with information about the patient collected in hospital during the crisis periods. This data set contains levels of hemoglobin, hematocrit and neutrophils in patient's blood in times when he was hospitalized, but also weight during these days, and other information.

Another thing that is done in this part is the descriptive analysis of hemoglobin, hematocrit and neutrophils. The goal is to see if there are big changes in their level in blood on the day of crisis comparing to the other days, but mostly to the day before and day after the crisis.

3.6.1 Descriptive statistics

Medical information about normal ranges of hemoglobin, hematocrit and neutrophils used in this subsection are taken from MedicineNet.

Hemoglobin

Hemoglobin is the protein molecule in red blood cells that carries oxygen from the lungs to the body's tissues and returns carbon dioxide from the tissues back to the lungs.

The hemoglobin level is expressed as the amount of hemoglobin in grams (gm) per decilitre (dL) of whole blood, a decilitre being 100 milliliters.

The normal range for hemoglobin depend on the age and the gender of the person. The normal ranges for men after middle age: 12.4 to 14.9 gm/dL.

Table 3.4 reports minimum, maximum, average and standard deviation of the level of hemoglobin in the blood of our patient on the days before crisis, crisis day and days after crisis.

Period	MIN	MAX	AVERAGE	STAN DEV
Day before crisis	8.8	14.5	12.12	1.98
Crisis day	15.2	22.4	18.8	2.16
Day after crisis	10.3	20.9	15.24	3.75

Table 3.4: Table of descriptive statistics for hemoglobin

Looking at minimum and maximum for both crisis days and not crisis days, we can observe that hemoglobin level of our patient differs from normal level of hemoglobin, it goes below but also above its normality. This conclusion is expected, since the state of our patient implies both losing or getting plasma. Another thing that we can notice from Table 3.4 is that during the crisis days level of hemoglobin is higher. This statement is tautological, given the explanation of how a crisis is defined, given before. To be sure of this observation we will check the boxplots of days before crisis, days after crisis and crisis days that are shown in Figure 3.36.

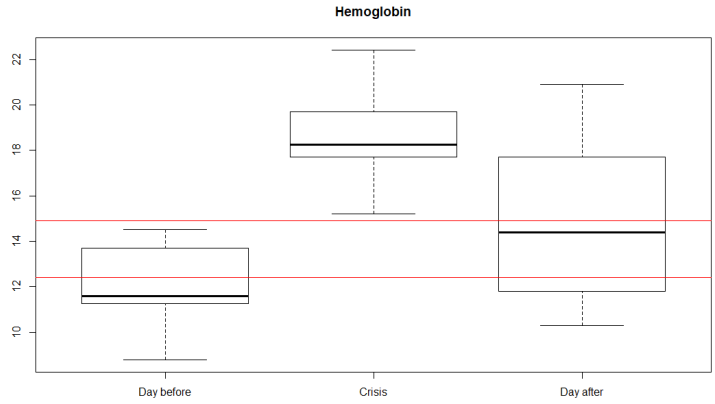

Figure 3.36: Boxplot of hemoglobin levels during days before crisis, days after crisis and crisis days

From the boxplots in Figure 3.36 we can see that the level of hemoglobin of the patient the day before crisis is inside the hemoglobin normal range and that during the days after crisis it can go above the normal range. As for the crisis days, the data confirm that there is a relevant increase of the hemoglobin level during attacks.

Hematocrit

The hematocrit blood test determines the percentage of red blood cells (RBC's) in the blood. Blood is composed mainly of red blood cells and white blood cells suspended in an almost clear fluid called serum. The hematocrit test indicates the percentage of blood by volume that is composed of red blood cells. Since hemoglobin is the protein molecule in red blood cells, hematocrit and hemoglobin levels are strictly correlated and so we expect same observations as above.

Normal values for the hematocrit test vary according to age, sex, pregnancy, altitude where people live, and even vary slightly between various testing methods. The following are reported ranges of normal hematocrit levels for males: $42\% - 54\%$.

Table 3.5 reports minimum, maximum, average and standard deviation of level of hematrocit in blood of our patient crisis on the days before crisis, crisis day and days after crisis.

Period	**MIN**	**MAX**	**AVERAGE**	**STAN DEV**
Day before crisis	25.8	44.3	37.34	6.88
Crisis day	44.5	65.9	55.4	6.23
Day after crisis	32.4	61.1	46.34	10.21

Table 3.5: Table of descriptive statistics for hematrocit

As for hemoglobin, if we look at minimum and maximum for both crisis days and not crisis days, we can observe that hematrocit level of our patient differs from the normal level of healthy person, it goes below but also above its normality.

From Table 3.5 we find that during crisis days level of hematrocit is higher than general. Again, to be sure of this observation we will check boxplots of days before crisis, days after crisis and crisis days which are represented in Figure 3.37.

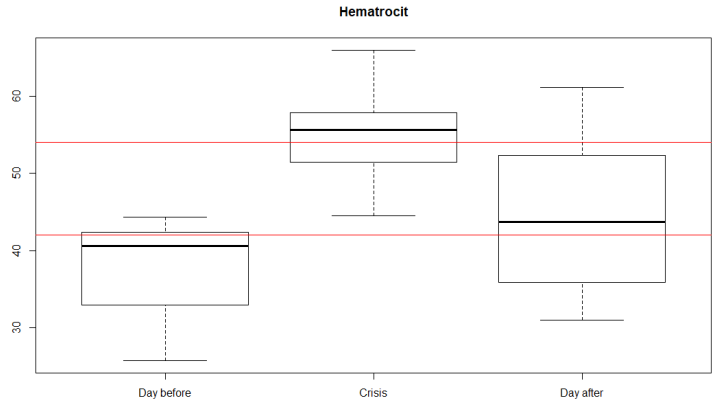

Figure 3.37: Boxplot of hemtrocit levels during days before crisis, days after crisis and crisis days

From Figure 3.37 we see that during days before crisis level of hematrocit is lower than normal range of hematocrit, also it can be confirmed that hematrocit level in blood is higher for crisis days than during periods without crisis.

Neutrophils

Neutrophils are a type of white blood cell. In fact, most of the white blood cells that lead the immune system's response are neutrophils. There are four other types of white blood cells. Neutrophils are the most plentiful type.

White blood cells produce chemicals that fight antigens by going to the source of the infection or inflammation.

Neutrophils are important because, unlike some of the other white blood cells, they are not limited to a specific area of circulation. They can move freely through the walls of veins and into the body tissues to immediately attack all antigens.

Normal neutrophils percentage level is $55\% - 70\%$ of white blood cells, while for the absolute neutrophil count the reference range in adults varies by study, but 1500 to 8000 cells per microliter is typical. We study both percentage level and absolute count of neutrophils of our patient.

Table 3.6 and Table 3.7 show minimum, maximum, average and standard deviation of percentage level and absolute count of neutrophils in blood of our patient during the days before crisis, crisis day and days after crisis.

Period	**MIN**	**MAX**	**AVERAGE**	**STAN DEV**
Day before crisis	49.5606	77.2627	67.2790	10.2063
Crisis day	57.1010	89.0444	76.4884	9.9923
Day after crisis	47.4255	87.4812	73.0185	13.0615

Table 3.6: Table of descriptive statistics for percentage level of neutrophils

Period	**MIN**	**MAX**	**AVERAGE**	**STAN DEV**
Day before crisis	2.8	13.3	6.4	3.925
Crisis day	3.9	26.09	14.961	7.462
Day after crisis	3.5	31.83	15.656	9.097

Table 3.7: Table of descriptive statistics for absolute count of neutrophils

Also for neutrophils, if we look at minimums and maximums in Table 3.6 and Table 3.7 for both crisis days and not crisis days, we can observe that neutrophil level of our patient differs from normal level of healthy person, it goes much more above its normality.

From Table 3.6 it seems that during crisis days percentage level of neutrophils is higher than the days before crisis. While as from Table 3.7 for absolute count it can be observed that it is much lower in the day before crisis than in days of crisis and days after crisis. Again, to be sure of this observation we will check box plots of days before crisis, days after crisis and crisis days.

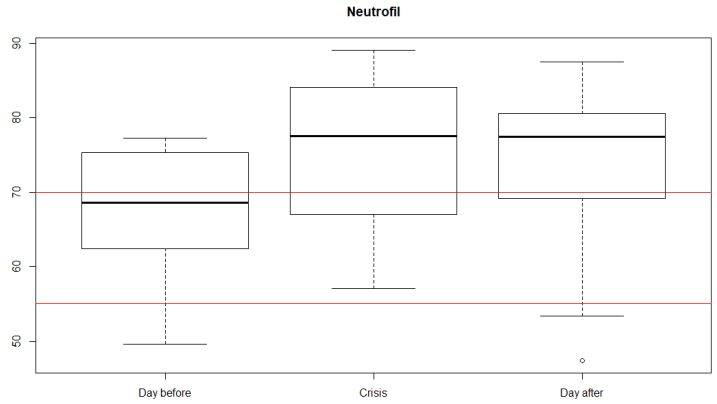

Figure 3.38: Boxplot of neutrophils levels during days before crisis, days after crisis and crisis days

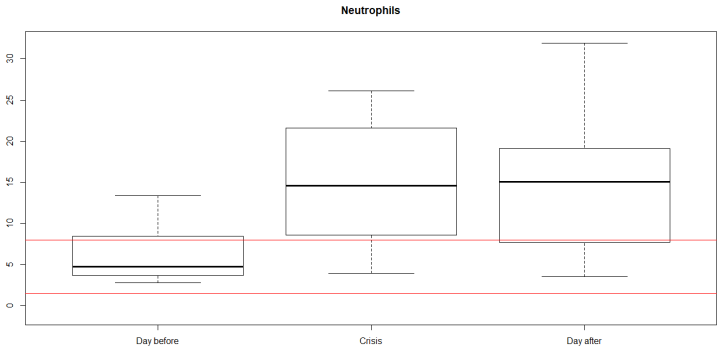

Figure 3.39: Boxplot of absolute count of neutrophils during days before crisis, days after crisis and crisis days

Looking the boxplots in Figure 3.38 and Figure 3.39 we can confirm that there is gtowth between neutrophils percentage levels during crisis days and during days before crisis in both percentage and absolute level of neutrophils. Another thing we can notice is that during crisis days and days after crisis, both levels of neutrophils are above their normal level range.

Weight

As for levels of hemoglobin, hematocrit and neutrophils in patient's blood, we will also compute descriptive statistics for the weight of the patient during the days before crisis, days after crisis and crisis days in order to check if crisis effects patient's body weight.

In Table 3.8 are presented minimum, maximum, average and standard deviation of level of hemoglobin in blood of our patient.

Period	MIN	MAX	AVERAGE	STAN DEV
Crisis day	60.5	69.4	65.7	2.37
Day before crisis	58.5	67.2	63.7	3.3
Day after crisis	61.2	71.8	66.4	3.2

Table 3.8: Table of descriptive statistics for the body weight of the patient

It can be observed from Table 3.8 that weight of patient grows during crisis days comparing to days before crisis. To be sure about this observation we will check the boxplots of days before crisis, days after crisis and crisis days which are shown in Figure 3.40.

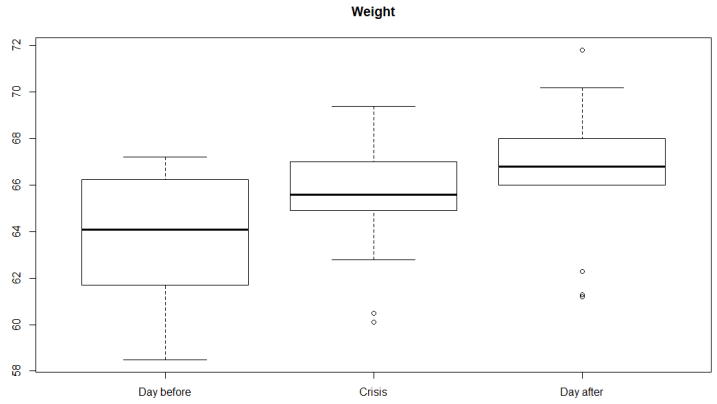

Figure 3.40: Boxplots of the body weight days before crisis, days after crisis and crisis days

From boxplots in Figure 3.40 we can notice difference between weight of patient during days before crisis and during other days (both crisis days and days after crisis), but most important during crisis days. Measurements confirm that during crisis days patient gains weight.

If we compare body weights between crisis days and days after crisis, it can be said that weight continuous to grows up to plus 1kg.

3.6.2 Fitting ARIMA model to full weight time series

We recall that in the analysis obtained in Section 3.4.1 there were 48 missing observation on body weight because patient was admitted to the hospital. Now we join the body weight data from hospital to the time series analuyed in section 3.4.1 (from 18.11.2013 until 31.08.2016). Now there are no missing data in the body weight.

Our goal now is to find the best ARIMA model for newly obtained data and compare it with results we got in Section 3.4.1 in order to check if NA values influence analysis.

The best model among all possible with $d = 1$ is $ARIMA(4, 1, 5)$ and its AIC value is -6328.24, the fitted plot for this model is represented in Figure 3.41:

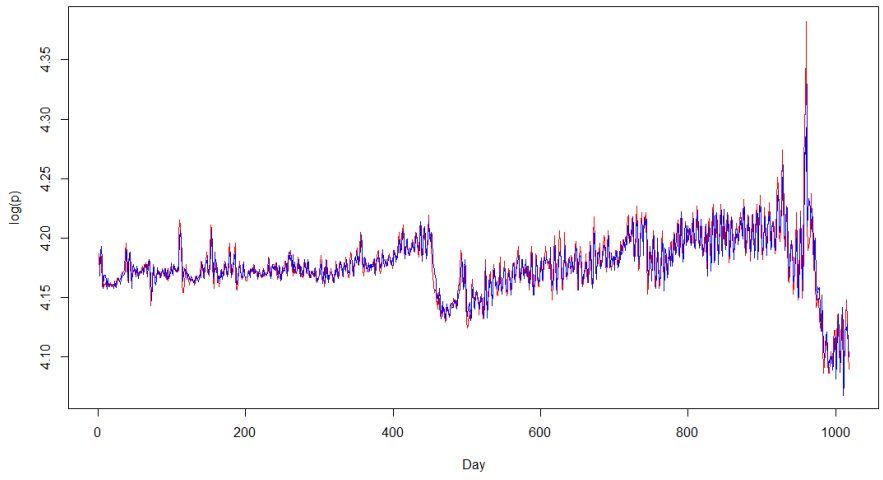

Figure 3.41: Red line represents new weight time series in log scale: blue line is the fit of $ARIMA(4, 1, 5)$ model.

As far as analysis of residuals correlation:
Ljung-Box test gave as result p-value=0.9709, which tells us that it can not be assumed that estimated residuals are dependent.

ACF and PACF plots of estimated residual for $ARIMA(4,1,5)$ model are shown in Figure 3.42:

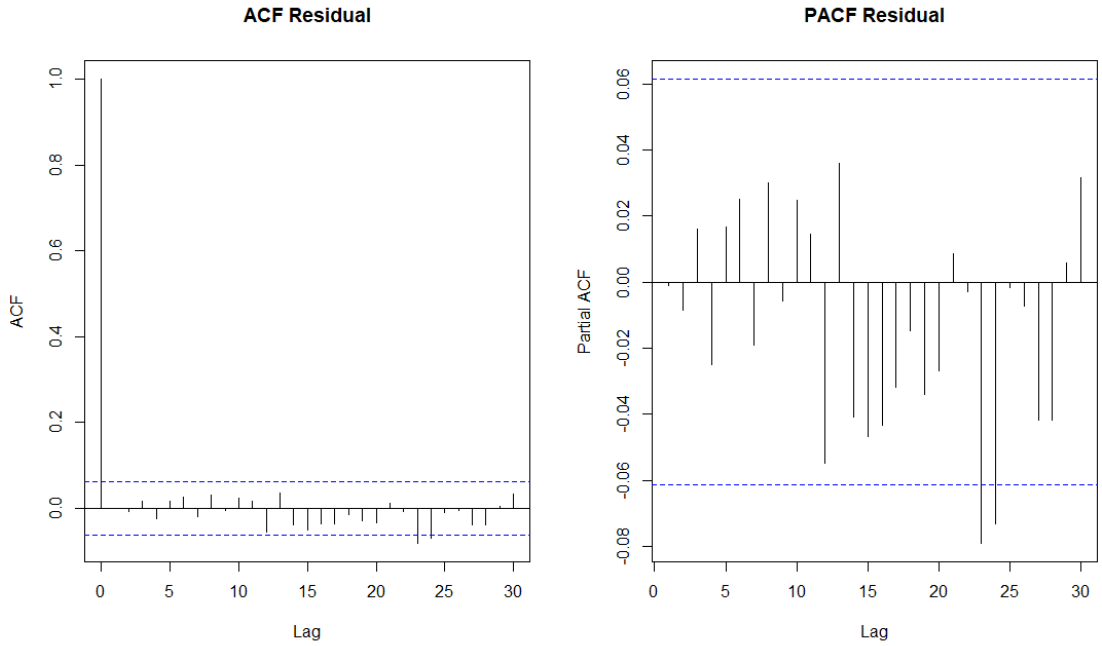

Figure 3.42: ACF and PACF of residuals for model $ARIMA(4,1,5)$.

From Figure 3.42 we can conclude that there is no significant autocorrelation between residuals for model $ARIMA(4,1,5)$.

The best model among all possible with $d = 2$ is $ARIMA(5, 2, 0)$ and its AIC value is -6035.49, the fitted plot for this model is represented in Figure 3.43:

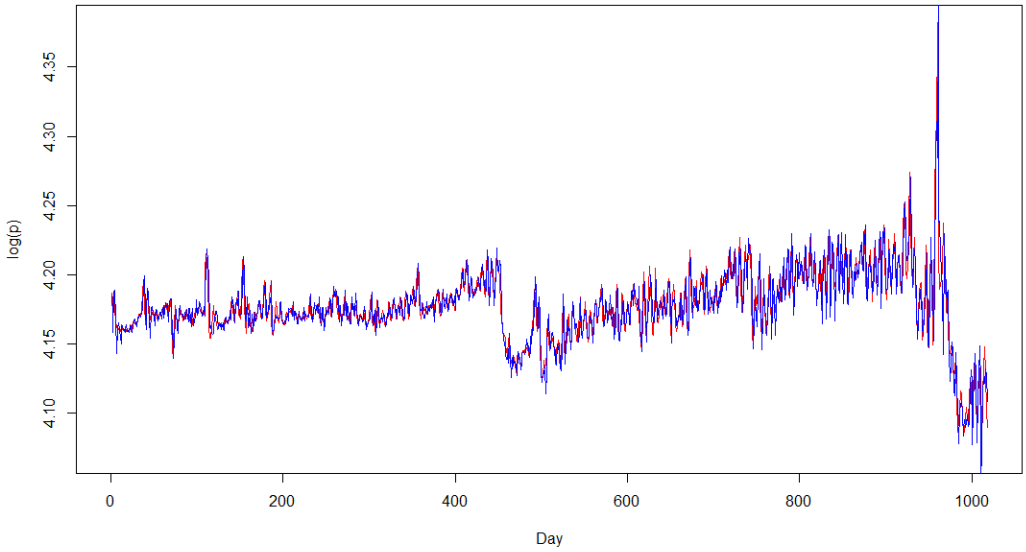

Figure 3.43: Red line represents new weight time series in log scale: blue line is fit of $ARIMA(5, 2, 0)$ model.

Moving to analysis of residuals correlation:
Ljung-Box test gave as result p-value=0.005792, which tells us that it can be assumed that estimated residuals of $ARIMA(5, 2, 0)$ are dependent. To investigate further wether there are correlation between successive estimated errors, we will observe ACF and PACF plots. ACF and PACF plots of estimated residuals are shown in Figure 3.44.

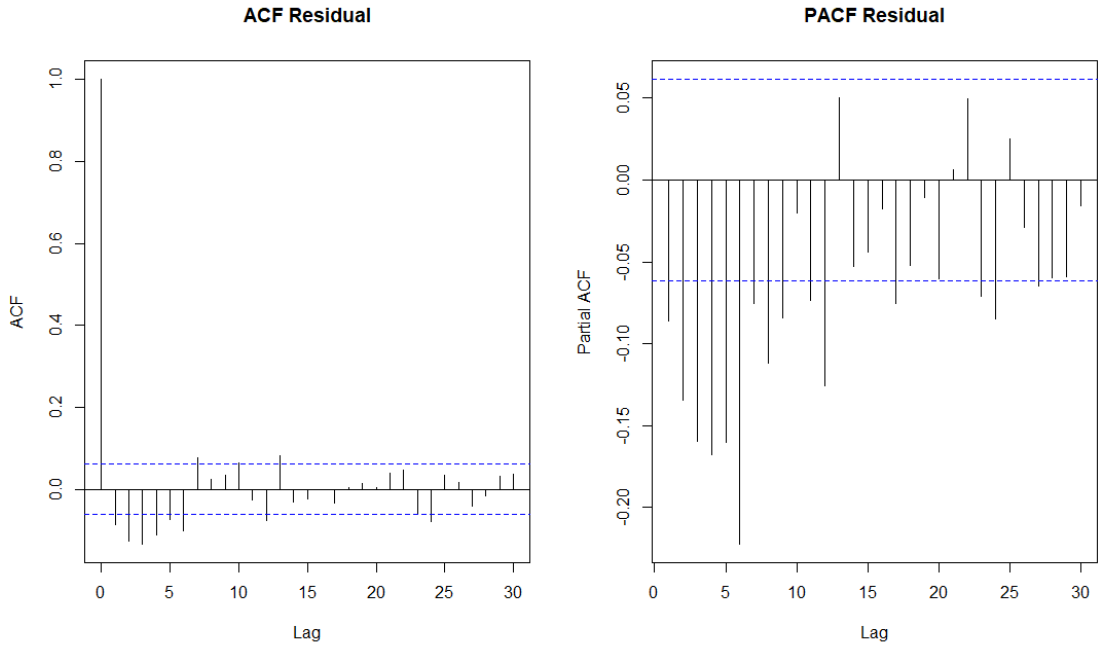

Figure 3.44: ACF and PACF of residuals for model ARIMA(5,2,0).

From Figure 3.44 can be observed that the residuals are correlated.

If we compare AIC values and p-values of $ARIMA(4,1,5)$ model and $ARIMA(5,2,0)$ model we get that the best model for weight time series in period from 18.11.2013 until 31.08.2016 is $ARIMA(4,1,5)$ since its AIC value is smaller. In Section 3.4.7 we have that the best model for the weight time series in the same period is $ARIMA(4,1,0)$, considering this, we will fit also this model to the complete weight time series. The AIC value of this model is -6310.6, its fitted plot is represented in the Figure 3.45.

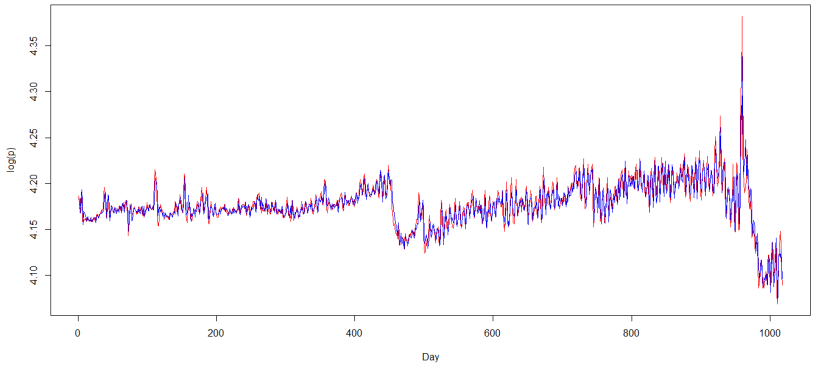

Figure 3.45: Red line represents new weight time series in log scale: blue line is fit of $ARIMA(4,1,0)$ model.

As far as analysis of residuals correlation:
P-value of Ljung-Box test is 0.4784, which tells that it can not be assumed that estimated residuals are correlated.

ACF and PACF plots of residuals of $ARIMA(4,1,0)$ model are shown in Figure 3.46:

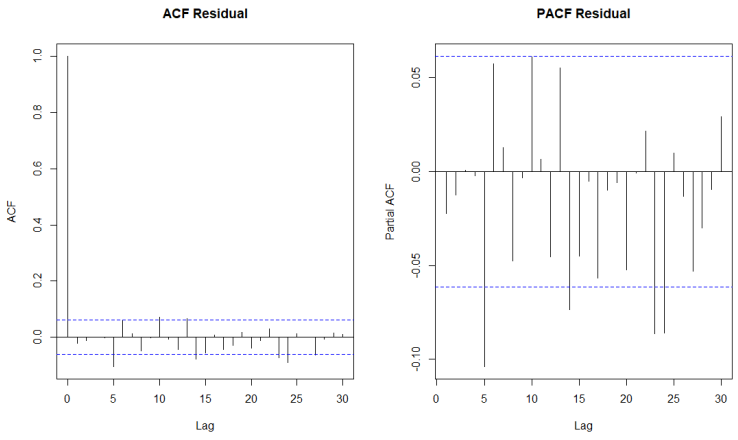

Figure 3.46: ACF and PACF of residuals for model ARIMA(4,1,0).

From Figure 3.46 we can conclude that there is no significant autocorrelation between residuals for model $ARIMA(4,1,0)$.

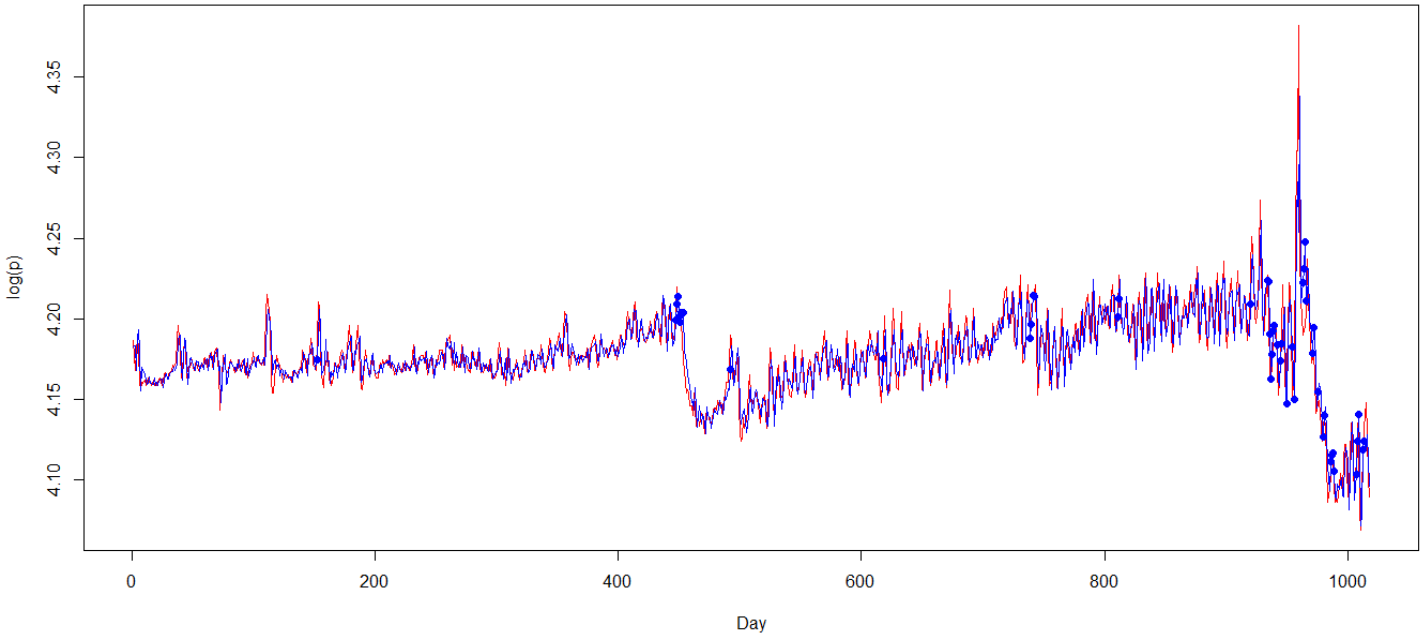

Figure 3.47: Red line represents new weight time series in log scale: blue line is fit of $ARIMA(4, 1, 0)$ model and circles are missing values from Section 3.4.1

3.6.3 Conclusion

From analysis obtained in Section 3.7.1 we see that by measuring every day the level of hemoglobin, neutrophils and hematrocit in patient's blood, it is possible to predict crisis; if increasing in their level is noticed, it can be assumed that the attack is coming to happen. Aapart from the hemoglobin, neutrophils and hematrocit levels, there are other features in the patient's data, as pressure and heart rate, that allow us to assess that there is an attack.
As for weight concerns, it can be confirmed that weight grows during crisis days, but for better understanding relationship between body weight and crisis, we should examine the change in weight at least two days before crisis, not only one. In Figure 3.48 are shown boxplots of the body weight of two days before crisis, days before crisis, crisis days and days after crisis.

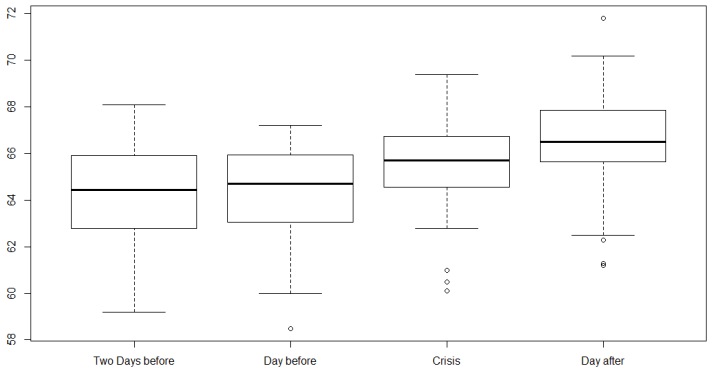

Figure 3.48: Boxplots of the body weight two days before crisis, days before crisis, crisis days and days after crisis.

From Figure 3.48 we notice that boxplots of the body weight two days before crisis and day before crisis have almost the same shape and we can confirm already observed, that during crisis days patients's weight increases.

Finally, Table 3.9 summarises the result obtained in Section 3.7.2.

ARIMA	**AIC**	**Ljun-Box**
ARIMA(4,1,5)	-6328.24	0.9709
ARIMA(5,2,0)	-6035.49	0.005792
ARIMA(4,1,0)	-6310.6	0.4784

Table 3.9: Table of results from Section 3.7.2

In this study, the $ARIMA(4,1,0)$ was the best candidate model selected for forecasting the body weight time series from 18.11.2013 until 31.8.2016 since the difference between its AIC value and AIC value of $ARIMA(4,1,0)$ is small and the sucessive residuals in $ARIMA(4,1,0)$ are not correlated and since $ARIMA(4,1,0)$ model has less parameter to estimate.
Hence we can coclude that the selected $ARIMA(4,1,0)$ model seem to provide an adequate predictive model for the body weight time series from 18.11.2013 until 31.08.2016, meaning that weight at time t can be estimated by weights measured in previous five days, i.e. weights at times $t-1, t-2, t-3, t-4, t-5$.

Chapter 4

Bayesian regression models for the body weight and for the indicator of crisis

In this chapter the goal is to find regression models for the time series of the body weight and the indicator of a crisis. As covariates, we us diastolic blood pressure (DBP) and systolic blood pressure (SBP), heart rate (HR) and the indicator of crisis (= 1 if at day t crisis occured or = 0 otherwise). To do this, it will be used Bayesian statistics, more precisely we will use *bsts* R package and its *bsts* function. The first thing to do when fitting a *bsts* model is to specify the contents of the latent state vector α_t. The bsts package offers a library of state models, which are included by adding them to a state specification (which is just a list with a particular format). The state specification is passed as an argument to bsts, along with the data and the desired number of MCMC iterations. The model is fitted using an MCMC algorithm. The returned object is a list (with class attribute "bsts").

4.1 Regression models for the body weight

Data included also a variable denoting increasing levels of crisis, from 0 to 4, where

0 stands for no crisis,

1 stands for crisis without resorting in hospital,

2 stands for crisis with hospitalization and infusion of liquids but not of vasoactive amines,

3 stands for crisis with hospitalization and infusion of both liquids and vasoactive amines

4 stands for crisis with death danger.

As we seek an algorithm to predict if next crisis would necessitate hospitalization or not, in our dataset five increasing levels of crisis from 0 to 4 have been simplified. Here we decided to distinguish only two levels "0" which represents no crisis and crisis without hospitalization (0 and 1 from our dataset), and "1" which represents other three levels, when patient was hospitalized.

4.1.1 Model with only Local Level Trend for weight time series

Let us fit a *bsts* model with just the linear trend. The first thing to do when fitting a *bsts* model is to specify the contents of the latent state vector μ_t. The *bsts* package offers a library of state models, which are included by adding them to a state specification (which is just a list with a particular format). The state specification is passed as an argument to *bsts*, along with the data and the desired number of MCMC iterations. We recall that the local level trend model is:

$$y_t = \mu_t + \varepsilon_t, \quad t = 1, 2, \dots \tag{4.1}$$

$$\mu_{t+1} = \mu_t + \eta_t \tag{4.2}$$

$$\mu_1 \sim N(\mu, \sigma_1^2) \tag{4.3}$$

$$\varepsilon_t \overset{iid}{\sim} N(0, \sigma^2), \quad \eta_t \overset{iid}{\sim} N(0, \tau^2) \quad t = 1, 2, \dots \tag{4.4}$$

Function *AddLocalLevel* adds a local level trend state component to an empty state specification. This function contains arguments:

- *Sigma.prior*: An object created by *SdPrior* function describing the prior distribution for the standard deviation of η_t.

- *Initial.state.value*: An object created using *NormalPrior* function, describing the prior distribution of the the μ_1.

SdPrior function specifies an Inverse Gamma prior for a variance parameter, but inputs are defined in terms of a standard deviation. It contains the following arguments:

- *Sigma.guess*: A prior guess at the value of standard deviation.

- *Sample.size*: The weight given to sigma.guess.

This puts a Gamma(α, β) prior on $1/\sigma^2$:

- Shape(α)= $sigma.guess^2 \times sample.size/2$

- Scale(β)= $sample.size/2$

NormalPrior function specifies a Normal(μ, σ^2) prior on μ_1. Its arguments are:

- *Mu*(μ): the mean of prior distribution

- *Sigma*(σ_1): the standard deviation of prior distribution

- *Initial.value*: the initial value of parameter being modeled in the MCMC algorithm.

There are several plot methods available. The default plot method plots the credible interval of model state, of μ_t against time t. Figure 4.1 and 4.2 show the estimated local level trend, where blue circles in Figure 4.1 represent original data of the body weight time series. Other plot methods can be accessed by passing a string to the plot function. For example, to see the contributions of the individual state components, pass the string "components" as a second argument.

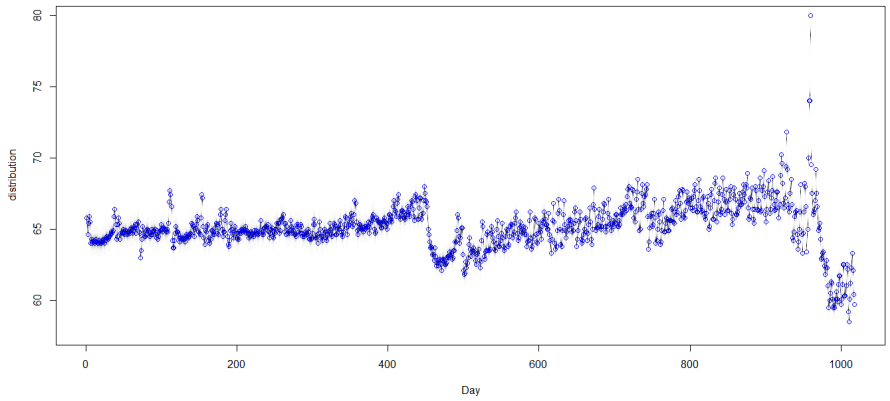

Figure 4.1: Local Level Trend model for weight time series

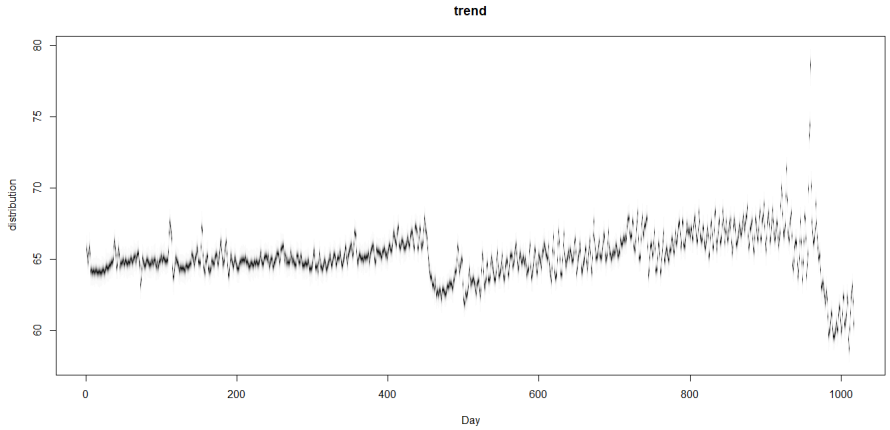

Figure 4.2: Trend for weight time series

Now, as far as the convergence, since *bsts* function returns a *bsts* object that contains several MCMC and an MCMC creates a sample from the posterior distribution, we usually want to know whether this sample is sufficiently close to the posterior to be used for analysis. There are several standard ways to check this, but I used the traceplot and the Gelman-Rubin diagnostic that evaluates MCMC convergence by analysing the difference between multiple Markov chains. The trace plot shows the sampled values of a parameter over time. This plot helps us to judge how quickly the MCMC procedure converges in distribution, that is, how quickly it forgets its starting values. Gelman and Rubin

propose a convergence test based on 2 or more parallel chains. Their method is based on a comparison of the within and between chain variances for each variable. Large differences between these variances indicate nonconvergence. See Gelman and Rubin [1992] for the detailed description of the method.

Gelman diagnostics is contained in *coda* R package, the function name is *gelman.plot*. The trace plots of the variance of ε_t and of η_t are shown in Figure 4.3, whereas in Figure 4.4 are their Gelman plots.

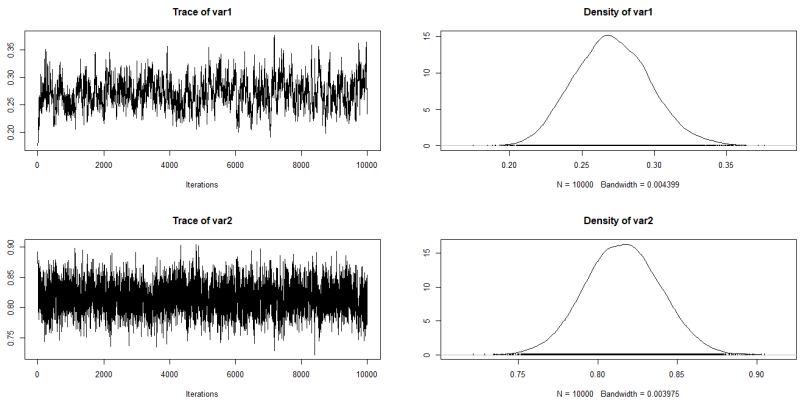

Figure 4.3: Trace plot for Local Level Trend model

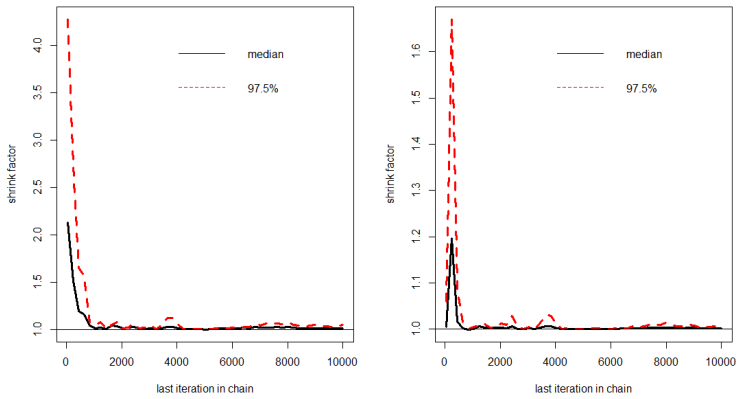

Figure 4.4: Gelman plot for Local Level Trend model

Figure 4.3 and Figure 4.4 suggest that the MCMC chains have reached convergence.

4.1.2 Model with Local Level Trend and $AR(5)$ for weight time series

The analysis in Section 3.7.2 suggests that the best model for the body weight time series in period from 18.11.2013 until 31.08.2016 is $ARIMA(4,1,0)$, hence $AR(5)$ model is added to the local level trend:

$$y_t = \alpha_{0,t} + \alpha_{1,t} y_{t-1} + \cdots + \alpha_{5,t} y_{t-5} + \varepsilon_t, \quad t = 1, 2, \ldots \tag{4.5}$$

$$\alpha_{i,t+1} = \alpha_{i,t} + \eta_t \tag{4.6}$$

$$\varepsilon_t \sim N(0, \sigma_\varepsilon^2), \quad \eta_t \sim N(0, \sigma_\eta^2) \tag{4.7}$$

Figures 4.5 and 4.6 show the estimated local level trend with AR(5), where the traceplots of variance of ε_t and variance of η_t in Figure 4.7 and their Gelman plots in Figure 4.8 suggest that the MCMC chains are converging.

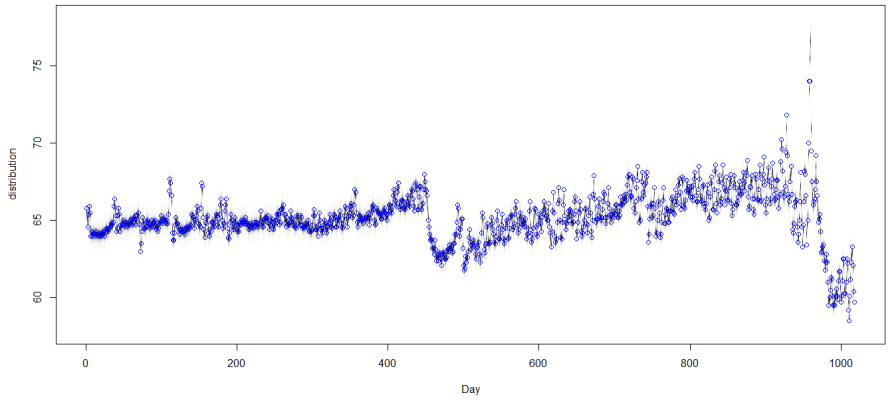

Figure 4.5: Local Level Trend with AR(5) model for weight time series

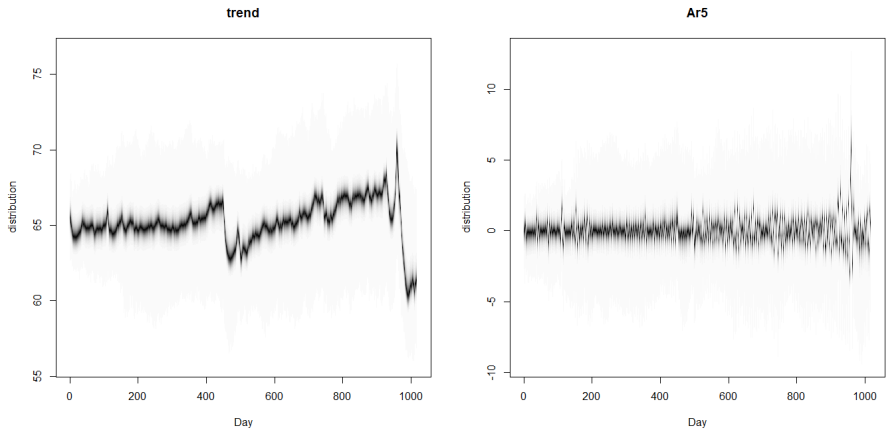

Figure 4.6: Components of model

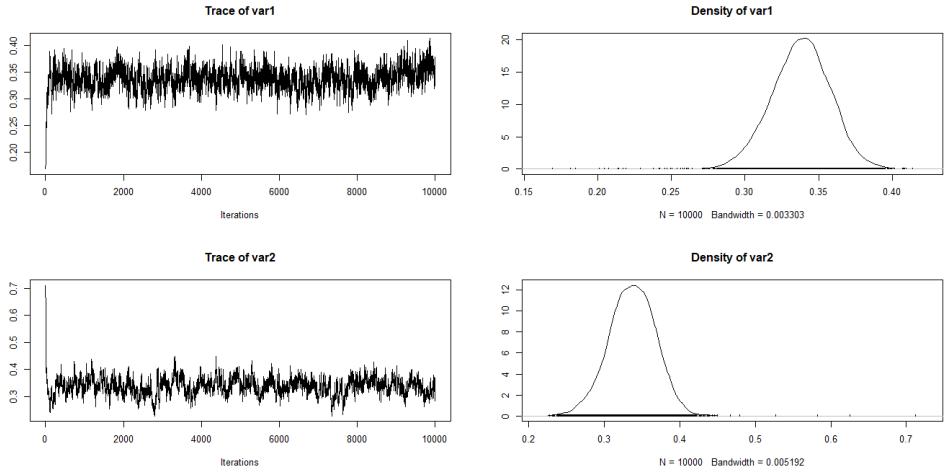

Figure 4.7: Trace plot for Local Level Trend with $AR(5)$ model

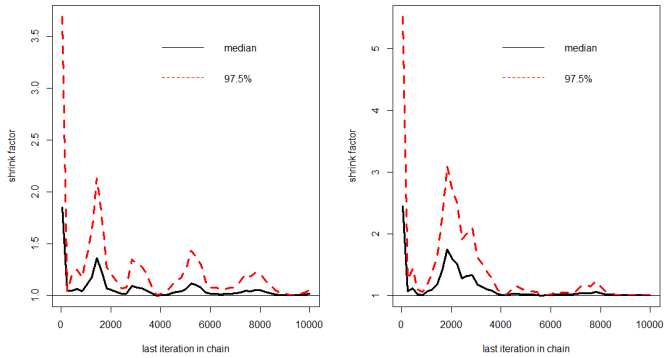

Figure 4.8: Gelman plot for Local Level Trend with $AR(5)$ model

Gelman plots of $AR(5)$ coefficients in Figure 4.9 are indicating that also coefficients are converging.

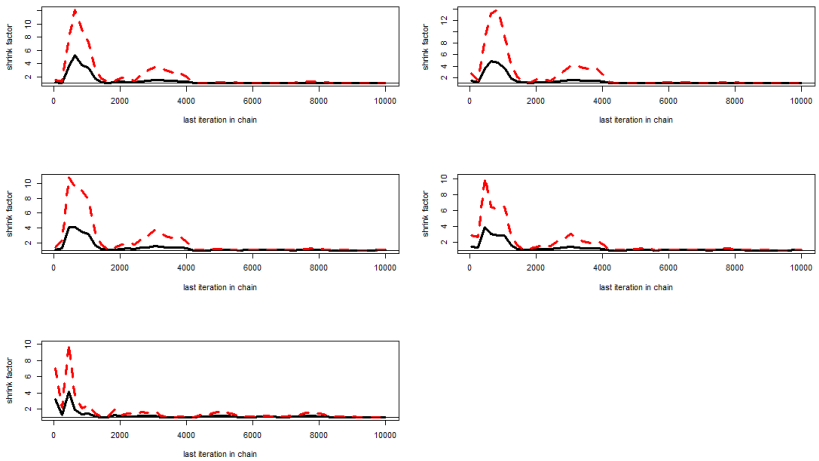

Figure 4.9: Gelman plot for coefficients of $AR(5)$ model

4.1.3 Local Level Trend with $AR(5)$ and linear regression for weight time series

Our main task is to find regression model for weight time series depending on SDP and DBP, heart beat and on whether day t was the crisis day, the day before crisis, the day after crisis or otherwise. Model representation is the following:

$$\begin{cases} y_t = \alpha_{0,t} + \sum_{i=1}^{5} \alpha_{i,t} y_{t-i} + aSBP_t + bDBP_t + cHR_t + dDayInd_t + \varepsilon_t, \\ \alpha_{i,t+1} = \alpha_{i,t} + \eta_t \qquad\qquad t = 1, 2, \dots \\ \varepsilon_t \sim N(0, \sigma_\varepsilon^2), \quad \eta_t \sim N(0, \sigma_\eta^2) \end{cases} \tag{4.8}$$

where DBP_t represents diastolic pressure at time t, SBP_t is systolic pressure at time t, HR_t is heart rate at time t and $DayInd_t$ is categorical time series equal to "c" if day t was crisis day, equal to "b" if day t was day before crisis, equal to "a" if day t was day after crisis and equal to "o" otherwise.

Summary of the estimates of the parameters in (4.7) is in Table 4.1.

	mean	sd	mean.inc	sd.inc	inc.prob
HR	0.01	0.00	0.01	0.00	1.00
DBP	-0.01	0.00	-0.01	0.00	0.97
SBP	-0.00	0.00	-0.01	0.00	0.10
DayIndc	-0.01	0.05	-0.31	0.14	0.02
DayIndo	-0.00	0.02	-0.15	0.15	0.01
DayIndb	-0.00	0.00	-0.08	0.11	0.00
(Intercept)	0.00	0.00	0.00	0.00	0.00

Table 4.1: Posterior estimates of the regression coefficients

The last column in Table 4.1, *inc.prob*, is the probability that shows how much important is variable for regression; the smaller it is the less significant the variable is. With respect to this we can observe that $DayInd$ can be removed. Also maximum pressure has a small probability with respect to minimum pressure and heart beat. Since minimum and maximum pressure are correlated($cor(mass, min) = 0.641311$) it is expected that only one of this two features is relevant for regression model for the body weight time series.

Considering previous results, we will fit model with minimum blood pressure and heart beat as only covariates. Summary of reduced model is shown in Table 4.2.

	mean	sd	mean.inc	sd.inc	inc.prob
HR	0.01	0.00	0.01	0.00	1.00
DBP	-0.01	0.00	-0.01	0.00	1.00
(Intercept)	0.00	0.00	0.00	0.00	0.00

Table 4.2: Posterior estimates of the regression coefficients of reduced model

From Table 4.2 we see that for both covariates $inc.prob = 1$, meaning that they are important for predicting body weight.

As fas as $AR(5)$ coefficients in (4.7) , Table 4.3 reports their posterior estimates.

	Mean	SD
α_1	1.13	0.21
α_2	-0.70	0.28
α_3	-0.17	0.27
α_4	0.29	0.25
α_5	-0.06	0.20

Table 4.3: Posterior estimates of $AR(5)$ coefficients

For checking convergence of $AR(5)$ coefficients it is used Geweke method. Geweke proposes a convergence diagnostic based on standard time series methods. The test is appropriate for use with single chains when convergence of the mean of the sampled variables is of interest. For each variable, the chain is divided into 2 "windows" containing the first 10% and the last 50% of the iterates. If the whole chain is stationary, the means of the values early and late in the sequence should be similar.
Its convergence diagnostic Z is the difference between these 2 means divided by the asymptotic standard error of their difference. As the chain length goes to infinity, the sampling distribution of $Z \sim N(0, 1)$ if the chain has converged. Hence values of Z which fall in the extreme of tails of a standard normal distribution suggest that the chain was not fully converging early on (i.e. during the 1st window).

From the Geweke diagnostic of $AR(5)$ coefficients, that is shown in Table 4.4, it can be concluded that the chains are converging.

var1	var2	var3	var4	var5
0.9962343	-1.2693594	1.9515477	-1.4228130	0.3131044

Table 4.4: Gewek diagnostic of $AR(5)$ coefficients

Geweke plot shows what happens to Geweke's Z-score when successively larger numbers of iterations are discarded from the beginning of the chain. The first half of the Markov chain is divided into $nbins - 1$ segments, then Geweke's Z-score is repeatedly calculated. The first Z-score is calculated with all iterations in the chain, the second after discarding the first segment, the third after discarding the first two segments, and so on. The last Z-score is calculated using only the samples in the second half of the chain.
Geweke plots for $AR(5)$ coefficients can be seen in Figure 4.10.

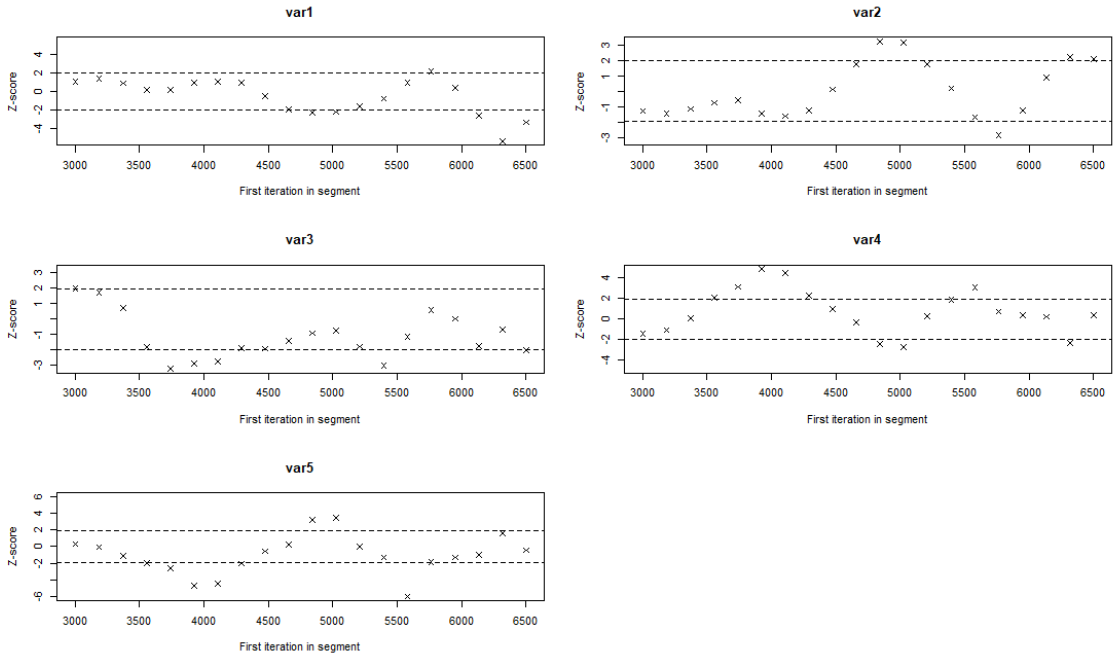

Figure 4.10: Geweke plots for $AR(5)$ coefficients

4.2 Regression for crisis indicator

Now we look for a Bayesian regression model where response is "crisis", the binary time series we have introduced in Section 4.1, i.e. $crisis_t = 1$ if patient was hospitalized at time t and 0 otherwise. Our goal now is to estimate regression model of type:

$$\begin{cases} crisis_t = ay_t + bSBP_t + cDBP_t + dHR_t + intercept + \varepsilon_t \\ \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \end{cases} \tag{4.9}$$

where y_t is weight time series, SBP_t represents systolic pressure, DBP_t is diastolic pressure, HR_t is heart rate and $intercept$ is a constant.

Posterior estimates are shown in Table 4.5.

	mean	sd	mean.inc	sd.inc	inc.prob
DBP	-0.00	0.00	-0.00	0.00	0.19
y	-0.00	0.00	-0.00	0.00	0.09
HR	0.00	0.00	0.00	0.00	0.02
SBP	-0.00	0.00	-0.00	0.00	0.02
(Intercept)	0.00	0.00	0.00	0.00	0.00

Table 4.5: Posterior estimatets of regression coefficients in (4.9)

From the last column in Table 4.5 we conclude that maximum pressure and heart beat can be removed. Summary of Model without maximum pressure and heart beat is shown in Table 4.6.

	mean	sd	mean.inc	sd.inc	inc.prob
y	-0.00	0.00	-0.00	0.00	0.23
DBP	0.00	0.00	0.00	0.00	0.21
(Intercept)	0.00	0.00	0.00	0.00	0.00

Table 4.6: Posterior estimates of regression coefficients of reduced model

Looking at Table 4.6 we can say that minimum pressure and weight influence on crisis. Descriptive statistics of coefficients for model with weight and minimum pressure as covariates are represented in Table 4.7.

	(Intercept)	y	min
1	Min. :0	Min. :-0.003931	Min. :-0.0006352
2	1st Qu.:0	1st Qu.: 0.000000	1st Qu.: 0.0000000
3	Median :0	Median : 0.000000	Median : 0.0000000
4	Mean :0	Mean :-0.000100	Mean : 0.0000913
5	3rd Qu.:0	3rd Qu.: 0.000000	3rd Qu.: 0.0000000
6	Max. :0	Max. : 0.002411	Max. : 0.0019196

Table 4.7: Descriptive statistics of coefficients for reduced model

4.3 Conclusion

In this study, the Bayesian statistic was used in order to find the best regression models for the body weight time series and for crisis indicator time series. The reason for using it is its possibility to combine AR models and linear regression in one model and its capabilities to fit regression models where output is binary time series.

Since in Section 3.7.2 we got that $ARIMA(4,1,0)$ can be the best model for the the body weight time series in period from 18.11.2013 until 31.08.2106, in this section for this time series we fitted $AR(5)$ model adding some regression models to it. As the result we got that the best model for the body weight in this period is:

$$y_t = 1.13 \times y_{t-1} - 0.7 \times y_{t-2} - 0.17 \times y_{t-3} + 0.29 \times y_{t-4} - 0.06 y_{t-5} - 0.01 \times DBP_t + 0.01 \times HR_t + \varepsilon_t$$

As fas as crisis indicator, the study showed that diastolic pressure and body weight have influence on it.

Chapter 5

Conclusions and further developments

In Section 3 we fitted ARIMA models to the time series of the body weight in three different time periods, the first is from 16.10.2013 until 17.11.2013, the second is from 18.11.2013 until 31.08.2016 and the third is from 01.09.2016 until 20.06.2017. For the first subseries the best model shows to be $ARIMA(0, 2, 0)$, which can be expressed as:

$$Y_t - 2Y_{t-1} + Y_{t-2} = \varepsilon_t$$

As for second and third subesries, we can observe that the best models are $ARIMA(4, 1, 0)$ and $ARIMA(4, 1, 2)$, retrospectively. The models are given by:

- ARIMA$(4, 1, 0)$:

$$(1 + 0.0203B^1 + 0.2871B^2 + 0.3769B^3 + 0.2643B^4)(Y_t - Y_{t-1}) = \varepsilon_t$$

- ARIMA$(4, 1, 2)$:

$$(1-0.3821B^1-0.1047B^2+0.2064B^3-0.3647B^4)(Y_t-Y_{t-1}) = \varepsilon_t-0.3164\varepsilon_{t-1}-0.5454\varepsilon_{t-2}$$

where Y_t is the body weight of the patient at time t, ε_t are independent, identically distributed variables, with zero mean and equal variance and B is lag operator, meaning:

$$BY_t = Y_{t-1}, \quad B^kY_t = Y_{t-k}$$

From this three models, we can observe that there is difference between the body weight in first period and the second two. The model for the first period suggests that the body weight at time t depends on the body weights in previous two days, i.e. at time $t-1$ and $t-2$, while the models for the second and third period suggest that the body weight in day t depends on previous five days; days $t-1$, $t-2$, $t-3$, $t-4$ and $t-5$. We can assume that the reason for this change in the body weight is immune globulin (IG) therapy since in the second period the patient was having monthly infusions of immune globulin. According to this models and analysis done in Section 3.4 we can conclude that there are no differences between the second and third phase, i.e. the patient's body weight does not change after ending with immune globulin therapy.
Pineton de Chambrun M., Gousseff M., Mauhin W. found study published in 2017 that preventive treatment with IG was the strongest factor associated with survival in people with SCLS. As far as our patient, looking at the available data we notice that in third period the attacks do not occur, we can suppose that this therapy had strong effect also on our patient, but we can not assume that attacks totally stopped since we do not have information from 20.06.2017 until now.

From analysis in Section 3.5 we obtained the best ARIMA models for the time series for the patient's diastolic and systolic blood pressure and for patient's heart beat.
The best model that we got for the heart rate is ARIMA$(2,1,4)$. As for blood pressure series, we observed that the best models are ARIMA$(4,1,3)$ for systolic blood pressure time series and ARIMA$(1,1,2)$ for diastolic blood pressure. In this section, from descriptive analysis, we also noticed that the systolic and diastolic pressure decrease from day before crisis to day of crisis, while heart rate increases.
Average decrease of the systolic pressure from day before crisis to day of crisis is equal to 23 mmHg, while from two days before crisis to crisis day SBP decreases in average for 28.25 mmHg.
Average decrease of the diastolic pressure from day before crisis to day of crisis is equal to 10 mmHg, while from two days before crisis to crisis day DBP decreases in average for 12 mmHg.
Average increase of heart rate from day before crisis to day of crisis is equal to 26 bpm, while from two days before crisis to crisis day HR increases in average for 29.5 bpm.

From analysis obtained in Section 3.7.1 we see that by measuring every day the level of hemoglobin, neutrophils and hematrocit in patient's blood, it is possible to predict crisis. If increase in their level is noticed, it can be assumed that the attack is coming to happen. Moreover, if we notice the following changes:

- increase in hemoglobin level bigger or equal to 5 gm per dL,

- increase in hematrocit level bigger or equal to 15%,

- increase in absolute count of neutrophils bigger or equal to 7 cells per liter,
- increase in percentage level of neutrophils bigger or equal to 9%

take a caution on possible attack and patient should be hospitalized.
Apart from the hemoglobin, neutrophils and hematrocit levels, there are other features in the patient's data, as pressure and heart rate, that allow us to assess that there is an attack. As for weight concerns, it can be confirmed that weight grows during crisis days. In this study, the ARIMA$(4, 1, 0)$ was the best candidate model selected for forecasting the body weight time series from 18.11.2013 until 31.8.2016, and its formula is:

$$(1 + 0.0354B^1 + 0.2779B^2 + 0.3817B^3 + 0.2700B^4)(Y_t - Y_{t-1}) = \varepsilon_t \qquad (5.1)$$

where Y_t is the body weight at time t.

As the result of Bayesian analysis in Section 5 we got that the best model for the body weight is:

$$y_t = 1.13 \times y_{t-1} - 0.7 \times y_{t-2} - 0.17 \times y_{t-3} + 0.29 \times y_{t-4} - 0.06 y_{t-5} - 0.01 \times DBP_t + 0.01 \times HR_t + \varepsilon_t$$

As fas as crisis indicator, the study showed that diastolic pressure and body weight have influence on it.

In the end, to predict future attack, after all analysis in this study it is suggested to continue with measuring not only the body weight but also blood pressure and heart rate every day. If significant changes in their values occur we should assume that crisis is coming. More specific, if we notice the following changes:

- increase in the body weight bigger or equal to 0.8kg,

- decrease of the systolic pressure bigger or equal to 20 mmHg,

- decrease of the diastolic pressure bigger or equal to 10 mmHg,

- increase of the heart rate bigger or equal to 25 bpm

we should pay attention and patient should be hospitalized.

In order to keep the patient's health stable , since by looking at available data we notice that the length of immune globulin therapy was from three to five days and since this therapy is said to be the most effective, we propose to use ARIMA models that we fitted to time series for heart rate, systolic and diastolic blood pressure and body weight in order to predicting their values for three to five days in advance, and if we observe variations in estimated values as mentioned above we can suppose attack in the following days and that patient should be hospitalized and he should start receive IG therapy .

Bibliography

[1] George Box and Gwilym Jenkins (1970). *Time Series Analysis: Forecasting and Control.* San Francisco: Holden-Day.

[2] Peter J. Brockwell and Richard A. Davis (2002). *Introduction to Time Series and Forecasting.* Springer, New York.

[3] Paul S. P. Cowpertwait and Andrew V. Metcalfe(2009). *Introductory Time Series with R.* Springer, New York.

[4] Simon Jackman (2009). *Bayesian Analysis for the Social Sciences.* Wiley.

[5] Kirk M. Druey and Philip R. Greipp (2010). *Narrative Review: Clarkson Disease-Systemic Capillary Leak Syndrome.* National Institutes of Health.

[6] Rob J. Hyndman and Yeasmin Khandakar (2008). *Automatic Time Series Forecasting: the forecast Package for R.* Journal of Statistical Software (https://www.jstatsoft.org/), Volume 27, Issue 3.

[7] Charles Patrick Davis and William C. Shiel Jr. *Hemoglobin (Low and High Range Causes).* MedicineNet.com

[8] William C. Shiel Jr. and Melissa Conrad Stöppler. *Hematocrit.* MedicineNet.com

[9] Steven L. Scott (2014). *Predicting the Present with Bayesian Structural Time Series.* International Journal of Mathematical Modelling and Numerical Optimisation (IJMMNO), Vol. 5 No.1/2, pages 4-23.

[10] Mary Kathryn Cowles, Nicky Best, Karen Vines and Martyn Plummer (1996). *CODA: Convergence Diagnosis and Output Analysis Software for Gibbs sampling output.* Institute of Public Health, Cambridge.

[11] A. Gelman and D. B. Rubin (1992). *Inference from Iterative Simulation Using Multiple Sequences.* Statist. Sci., Volume 7, Number 4, 457-472.

[12] Kristoffer Sahlin. *Estimating convergence of Markov chain Monte Carlo simulations.* Master Thesis in Mathematical Statistics, Stockholm University.