University of Economics, Prague

Faculty of Informatics and Statistics

MODEL BUILDING IN REGRESSION ANALYSIS

MASTER THESIS

Study programme: Quantitative Methods in Economics

Field of study: Quantitative Economic Analysis

Author: Bc. Kristina Lobanova

Supervisor: Mgr. Milan Bašta, Ph.D.

Prague, June 2019

Declaration

I hereby declare that I am the sole author of the thesis entitled “Model Building in Regression Analysis”. I duly marked out all quotations. The used literature and sources are stated in the attached list of references.

In Prague on ....

Bc. Kristina Lobanova

Acknowledgement

I hereby wish to express my deepest appreciation and gratitude to the supervisor of my thesis, Mgr. Milan Bašta, Ph.D., for his professional guidance, encouragement, insightful comments and recommendations that I received throughout my time as his student.

I would also like to extend my sincere gratitude to the Academic Guarantors of the major and minor field specializations, prof. Ing. Josef Arlt, CSc. and Ing. Pavel Zimmermann, Ph.D., for making an invaluable contribution to our educational process and providing us with the opportunity to study for this degree.

I gratefully acknowledge the effort of the coordinator of the QEA, MOS, and ISM programs, Mgr. Veronika Brunerová, who extended a great amount of assistance and actively supported us throughout these two years of studies.

My deep appreciation goes out to all professors who provided me with profound knowledge and shared their expertise in various fields of statistics.

I would also like to say a heartfelt thank you to my family for always being my mainstay and foremost support.

Abstract

Regression analysis is an increasingly used statistical technique for examining and modeling the relationship between various phenomena, which evolves formulation of a mathematical expression that characterizes the behavior of a particular random variable and its dependence on the set of external factors. The fundamental goal of the thesis is to illustrate the main steps of the model-building procedure, enhance understanding of the least squares estimation technique, and associated statistical methods. The emphasis of the theoretical part is placed on the discussion of the essential linear regression concepts and provision of tools necessary for utilizing a modeling approach for statistical analysis of the response variable. The practical part of the thesis aims at the illustration of the regression model-building process implemented using the actual data on the life expectancy at birth in various countries in order to investigate its dependence on the socio-economic development, demographic indicators, immunization coverage, nutritional status, and risk factors. The regression analysis is entirely conducted in the R statistical computing environment, which provides a broad spectrum of statistical and graphical techniques.

Keywords

Linear regression, model building, ordinary least squares, weighted least squares, life expectancy

Content

List of figures

List of tables

List of abbreviations

AIC	Akaike Information Criterion
AQI	Air Quality Index
BIC	Bayesian Information Criterion
BMI	Body Mass Index
CI	Confidence Interval
CLRM	Classical Linear Regression Model
CV	Cross-Validation
ECDF	Empirical Cumulative Distribution Function
FGLS	Feasible Generalized Least Squares
FWLS	Feasible Weighted Least Squares
GDP	Gross Domestic Product
GHO	Global Health Observatory
GLS	Generalized Least Squares
GNI	Gross National Income
LM	Lagrange Multiplier
LOESS	Locally Estimated Scatterplot Smoothing
MAE	Mean Absolute Error
OLS	Ordinary Least Squares
PM	Particulate Matter
PPP	Purchasing Power Parity
PRF	Population Regression Function
RESET	Regression Specification Error Test
RMSE	Root Mean Square Error
SRF	Sample Regression Function
SSE	Explained Sum of Squares
SSR	Residual Sum of Squares
SST	Total Sum of Squares
VIF	Variance Inflation Factor
WB	World Bank
WHO	World Health Organization
WLS	Weighted Least Squares

Introduction

Regression analysis is a statistical technique for examining and modeling the relationship between various phenomena, which is being used increasingly in different scientific areas. Regression analysis is attractive theoretically because of the elegant mathematics and well-designed statistical theory. Successful use of the regression methods demands a comprehension of both the theory and the practical problems that arise when the technique is applied to the real-world data (Montgomery et al., 2012).

Modeling refers to the formulation of mathematical expressions that, in some sense, characterize the behavior of a particular random variable. Such a variable of interest is called the dependent (response) variable and is denoted as y. Generally, the modeling aims at describing how the expected value of the dependent variable, $E(y)$, changes with varying conditions.

Other variables, incorporated into the regression model, which provide information on the behavior of the response, are known as independent (explanatory) variables. These variables are denoted by X_j and are assumed to be known constants. Additionally, all regression models include unknown constants, parameters, which define the behavior of the model. These parameters are identified by the Greek letters and need to be estimated from the data.

The degree of mathematical complexity of the model depends on the purpose of the modeling and knowledge about the process being analyzed (Rawlings et al., 1998).

- **Regression Model-Building Process**

The model-building process in the regression analysis is an iterative process, as depicted in figure 1. It starts with usage of the theoretical knowledge of the phenomenon under consideration and available data to formulate an initial regression model. Graphical visualization of the data may assist in the specification of the initial model. Then the parameters of the model are estimated, frequently employing the least squares method, to evaluate the quantitative effect of the regressors upon the variable of interest. Afterward, the researcher must assess the model adequacy by looking for potential functional form misspecification, unusual data, or failure to include important predictors. If the diagnostics suggest the inadequacy of the model, then the model should be altered and the parameters estimated again. This procedure may be repeated until a satisfactory model is obtained.

Finally, it is necessary to validate the model to ensure that it produces the results that are suitable in the final application (Montgomery et al., 2012).

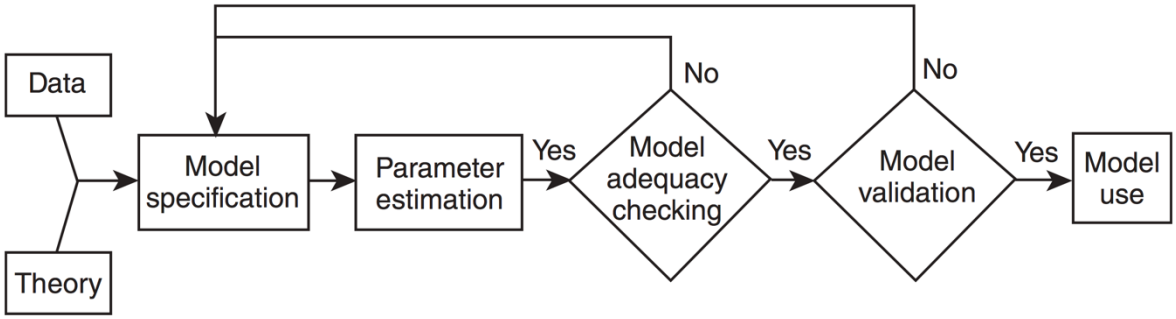

Figure 1: Regression model-building process (Montgomery et al., 2012)

- **Objective and Structure of Thesis**

The fundamental goal of the thesis is to illustrate the main steps of the model-building procedure, enhance understanding of the least squares estimation technique, and associated statistical methods. The emphasis is placed on the discussion of the essential linear regression concepts and provision of tools necessary for utilizing a modeling approach for statistical analysis of the response variable.

The first chapter provides an insight into the specification and assumptions of the linear regression model, the properties of the least squares estimators, measures of fit, and generalization of the Ordinary Least Squares method in the presence of heteroskedasticity. The second chapter discusses the classical hypothesis tests conducted in the regression analysis in order to assess the statistical significance of specific parameters and the model as a whole, as well as the methods for constructing individual and joint confidence intervals that serve for making inferential statements about the population. Chapter 3 reviews the techniques for diagnostics of a possible violation of the underlying assumptions on the error term in the regression model. Chapter 4 outlines methods for identification of the unusual observations which are, in some sense, remote from the rest of the data and may potentially affect the estimation and prediction results. The fifth chapter concludes the theoretical part by briefly covering several procedures for the features selection, which help to distinguish between the active and inactive predictors.

The practical part of the thesis aims at the illustration of the regression model-building process implemented on the actual data. For that purpose, the life expectancy at birth has been taken as the random variable whose behavior will be studied from the statistical point of view.

Life expectancy is one of the key indicators reflecting the population's health, which is broadly used by the researchers and policymakers to supplement economic measures of a nation's prosperity, such as GDP per capita. The data on the indicators, which may potentially be connected with the life expectancy, were retrieved from the official databases of international institutions: Global Health Observatory (GHO) - a World Health Organization's (WHO) data repository, and the World Bank's (WB) databank. All the features which act as explanatory variables involve economic, demographic factors, as well as indicators based on the nutritional status, immunization coverage, and factors which may put a person's life at risk.

The regression analysis is entirely conducted in the R statistical computing environment (R Core Team, 2018), which provides a broad spectrum of statistical and graphical techniques. Appendix A3 contains the complete reproducible R code with commented commands for better comprehension of the steps of the analysis.

Theoretical Part

1. Linear Regression Model

In a preliminary analysis of a particular phenomenon or in the case where predictions are the main objectives, the models usually belong to the group of models that are linear in the parameters. That is, the relationships are modeled as linear functions of predictors, and the parameters enter the model as simple coefficients. These models are referred to as linear regression models (Rawlings et al., 1998).

1.1. Theoretical Regression Model

The theoretical regression model is assumed to hold in the population of interest and is represented by the following equation

$$y_i = \eta_i + \varepsilon_i, \quad \text{for } i = 1, 2, \ldots, n, \tag{1.1}$$

where

n is the number of observations,

y_i is the value of the response variable y for the i^{th} observation,

η_i is the population (theoretical) regression function corresponding to the i^{th} observation,

ε_i is an additive error term such that

$$E(\varepsilon_i) = 0, \quad \text{for } i = 1, 2, \ldots, n. \tag{1.2}$$

A population regression function (PRF) η_i is a systematic component, represented by a linear function of the predictor variables and unknown constants, which hypothesizes a theoretical relationship between a dependent variable and a set of independent variables.

It is convenient to consider the regressors $X_1, \ldots, X_k$ as controlled by the researcher and measured with negligible error, while the response y is a random variable. That is, there is a conditional probability distribution for y at each possible value for $X_1, \ldots, X_k$.

For a simple linear regression model with a single regressor X, the regression function describing the relationship with a response y is a straight line, and in accordance with (1.2) the mean of the distribution is

$$\eta_i = E(y_i) = E(\beta_0 + \beta_1 x_i + \varepsilon_i) = \beta_0 + \beta_1 x_i + E(\varepsilon_i) = \beta_0 + \beta_1 x_i \,, \qquad (1.3)$$

where

x_i are the values of the explanatory variable X for the i^{th} observation,

β_0 is the intercept of the regression line (i.e., the expected value of y when $X = 0$),

β_1 is the slope of the regression line (i.e., the change in the mean of the distribution of y produced by a unit change in X).

If the range of X does not include zero, then β_0 has no practical interpretation.

Generally, the response y may be related to k explanatory variables. The regression function for a multiple regression, involving more than one predictor, is a hyperplane in a $(k+1)$-dimensional space and is given as

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}, \qquad (1.4)$$

where

k is the number of regressors,

$x_{i1}, \ldots, x_{ik}$ are the values of the explanatory variables $X_1, \ldots, X_k$ for the i^{th} observation,

β_0 is the intercept of the regression line (i.e., the expected value of y when $X_1, \ldots, X_k = 0$),

β_j, for $j = 1, 2, \ldots, k$ are partial regression coefficients, representing the expected change in y per unit change in X_j when all of the remaining regressor variables are held constant (Montgomery et al., 2012).

Consequently, the theoretical regression model is defined as

$$y_i = \eta_i + \varepsilon_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \qquad (1.5)$$

where ε_i is an error term or random disturbance, named so because it "disturbs" an otherwise stable relationship. The disturbance arises for several reasons, principally because it is merely possible to capture every impact on an economic variable in a model, no matter how elaborate (Greene, 2003). Thus, it is a proxy of all factors other than predictors under consideration that could possibly influence the dependent variable.

Under matrix notation, the equation (1.5) can be rewritten as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \qquad (1.6)$$

where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \tag{1.7}$$

and

$\boldsymbol{y}$ is the $(n \times 1)$ column vector of observations on the dependent variable y_i,

$\boldsymbol{X}$ is the $(n \times p)$ model matrix consisting of a column of ones allowing for estimation of the intercept, followed by the k column vectors of the observations on the independent variables,

$\boldsymbol{\beta}$ is the $(p \times 1)$ vector of parameters,

$\boldsymbol{\varepsilon}$ is the $(n \times 1)$ vector of the error terms.

Due to the presence of the intercept, the number of parameters in the model is equal to $(p = k + 1)$. The vectors $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ are stochastic vectors; elements of these vectors are random variables. The matrix $\boldsymbol{X}$ is regarded as a matrix of known constants. The vector $\boldsymbol{\beta}$ is a vector of fixed, but unknown, population parameters (Rawlings et al., 1998).

1.2. Empirical Regression Model

Multiple linear regression models are frequently applied as empirical models or approximating functions for the true underlying functional relationship between y and $X_1, \ldots, X_k$. This relationship is not known, but over certain sets of the predictor variables, the linear regression model may be a suitable approximation to the true unknown function (Montgomery et al., 2012). The fundamental purpose of the regression model is to estimate the population parameters β_j based on the data from a given sample.

The sample regression function (SRF) is the counterpart of the fixed, but unknown population regression function (PRF). Since the SRF, which is an estimation of the PRF, is obtained for a given sample drawn from the population, a new sample will produce different parameter estimates. The SRF is defined as

$$\hat{\eta}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik}, \tag{1.8}$$

where b_j are the estimators of the parameters β_j.

Consequently, the empirical regression model is expressed as

$$y_i = \hat{\eta}_i + e_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik} + e_i, \tag{1.9}$$

where

y_i is the observed value of the response variable y for the i^{th} observation,

e_i is the residual for i^{th} observation.

Using matrix notation, the equation (1.9) can be rewritten as

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{e}, \tag{1.10}$$

where

$\boldsymbol{b}$ is the $(p \times 1)$ vector of estimators of $\boldsymbol{\beta}$,

$\boldsymbol{e}$ is the $(n \times 1)$ vector of the residuals (i.e., estimators of $\boldsymbol{\varepsilon}$).

It follows that

$$\hat{y}_i = \hat{\eta}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik}, \tag{1.11}$$

where $\hat{y}_i$ is the fitted value of y for observation i, when $X_1 = x_{i1}, \ldots, X_k = x_{ik}$,

or equivalently

$$\hat{\boldsymbol{y}} = \boldsymbol{Xb}, \tag{1.12}$$

where $\hat{\boldsymbol{y}}$ is the $(n \times 1)$ vector of fitted values.

The residual is the difference between the observed value y_i and the corresponding fitted value $\hat{y}_i$, which provides a basis for the estimation of the realized value of the error term ε_i. Mathematically, the i^{th} residual is

$$e_i = y_i - \hat{y}_i, \tag{1.13}$$

or the vector of residuals

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}. \tag{1.14}$$

Since the residuals measure the discrepancy between the actual data and the fitted model, they play a significant role in examining model adequacy (Montgomery et al., 2012). The subsequent sections discuss the main underlying assumptions of the linear regression models, methods for detection of departures from these assumptions, and possible solutions to such problems.

1.3. Assumptions of the Classical Linear Regression Model

The linear regression is a parametric approach, which means that the model consists of a set of the underlying assumptions. Since the population regression function (PRF) is unobservable, one has to „guess“ it from the sample regression function (SRF) based on a particular sample drawn randomly from the entire population. The Classical Linear Regression Model (CLRM) provides a framework which assists in the achievement of the best possible guess (Gujarati, 2018), based on the assumptions discussed below. For successful regression analysis, proper estimation and inference procedures, it is crucial to evaluate whether these assumptions on the form of the model and relationships between its parts are satisfied.

A1. Linearity

The model (1.5) determines a linear relationship between y and $X_1, \ldots, X_k$. In such context, this assumption requires that the response variable is a linear combination of the explanatory variables and the error term. Nonetheless, by including non-linear independent variables, such as power transformations, it is possible to model curvilinear relationships.

A2. Full rank of the model matrix X

There cannot be perfect linear dependence (multicollinearity) among any of the independent variables in the model. Perfect multicollinearity suggests exact linear relationship, that is, knowing the value of one regressor allows to precisely predict the values of the other regressors. If this is not the case, the columns of the model matrix **X** are linearly independent, and the rank of the model matrix is equal to the number of its columns. The assumption of the full column rank of **X** is necessary for estimation of the parameters of the model.

A3. Exogeneity of the independent variables

The expected value of the error term for the i^{th} sample observation should not be a function of the values of the explanatory variables at any observation, including the i^{th} one. That is disturbance ε is assumed to have zero conditional mean

$$E[\varepsilon_i|\boldsymbol{X}] = 0, \quad \text{for all} \quad i = 1, 2, \ldots, n. \tag{1.15}$$

This assumption requires that the predictors do not contain any useful information for prediction of the random error ε_i.

A4. Homoskedasticity and nonautocorrelation of the error term

This assumption requires that the error terms have finite constant variance σ^2

$$D[\varepsilon_i|\boldsymbol{X}] = \sigma^2 < \infty, \quad \text{for all} \quad i = 1, 2, \ldots, n \tag{1.16}$$

and are not correlated across observations

$$C[\varepsilon_i, \varepsilon_j|\boldsymbol{X}] = 0, \quad \text{for all} \quad i \neq j. \tag{1.17}$$

The homoskedasticity (1.16) suggests an equal degree of variability of the disturbance across the range of the independent variables. The heteroskedasticity occurs when the variance of the error term changes across values of the predictors. In the presence of the heteroskedasticity, inferences about the population based on the Ordinary Least Squares estimation, discussed in chapter 2, may be generally incorrect.

Uncorrelatedness implies that observations of the error term should not predict each other. The assumption (1.17) requires that deviations of observations y_i and y_j from their expected values are uncorrelated.

A5. Data generation

It is customary to assume that elements of **X** are non-stochastic, whereby the researcher chooses the values of the regressors and then observes y_i. This assumption is a mathematical convenience, which allows simplifying the assumptions A3, A4, and A6 by considering the probability distribution of the error to be unconditional. That is, the distribution of ε_i does not involve any of the constants in **X**.

A6. Normality of the error term

In addition to the assumptions A3 and A4, the disturbances are supposed to follow normal distribution

$$\boldsymbol{\varepsilon}|\boldsymbol{X} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I}), \tag{1.18}$$

where **I** is the identity matrix with ones on the main diagonal and zeros elsewhere.

The violation of the normality assumption does not lead to biased or inefficient estimation of the regression parameters. Fulfillment of this assumption is essential for performing appropriate hypothesis testing and generating reliable confidence and prediction intervals. However, this is only a concern when the sample size is very small. When the sample size is sufficiently large, the Central Limit Theorem ensures that the distribution of the unobservables will be approximately normal (Greene, 2003).

1.4. Least Squares Estimation

There are various approaches to parameter estimation in the model. For many reasons, the method of least squares remains the benchmark technique, and in practice, the preferred method frequently results in a modification of the least squares (Greene, 2003). This section summarizes some of the features of the Ordinary Least Squares (OLS) method and its modification known as the Weighted Least Squares (WLS).

1.4.1. Ordinary Least Squares

The method of the Ordinary Least Squares (OLS) chooses the estimates to minimize the sum of squared residuals. In the multivariate case with *k* independent variables, that is, given *n* observations on *y*, X_1, ..., X_k, the least squares estimators of β_j are obtained by minimizing the following expression

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik})^2 \,. \tag{1.19}$$

This minimization problem consists of taking partial derivatives of the (1.19) with the respect to each β_j and setting them to 0, leading to *(k* + 1*)* linear equations in *(k* + 1*)* unknowns $b_0, b_1, \ldots, b_k$

$$\begin{gathered} n^{-1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_k x_{ik})^2 \\ n^{-1} \sum_{i=1}^{n} x_{i1} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_k x_{ik})^2 \\ \ldots \\ n^{-1} \sum_{i=1}^{n} x_{ik} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_k x_{ik})^2 \end{gathered} \tag{1.20}$$

These equations are often referred to as the OLS first order conditions, which can be computed by the method of moments under the exogeneity assumption A3 (Wooldridge, 2015).

Recall the equation (1.15) $E[\varepsilon_i | \boldsymbol{X}] = 0$, which can be written as $E[\varepsilon] = 0$. The probability theory implies that

$$C[X_j, \varepsilon] = E\left[X_j \varepsilon\right] - E\left[X_j\right] E[\varepsilon] = 0. \tag{1.20 a}$$

Given the mean value of the random element $E[\varepsilon] = 0$ by the assumption and independence of the error term form the j^{th} regressor, it follows that $E\left[X_j \varepsilon\right] = 0$.

Using these assumptions and $\varepsilon = y - \beta_0 - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_k X_k$, the population moment conditions can be expressed as

$$\begin{gathered} E(y - \beta_0 - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_k X_k) = 0 \\ E[X_1(y - \beta_0 - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_k X_k)] = 0 \\ \cdots \\ E[X_k(y - \beta_0 - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_k X_k)] = 0 \end{gathered} \tag{1.20 b}$$

The method of moments is used to estimate population moments by their sample counterpart. Therefore, the equations (1.20) are the sample analogs to the population restrictions (1.20 b).

In matrix terms, minimizing the sum of squared residuals requires to select a vector **b** such that the following function of $\boldsymbol{\beta}$ is as small as possible

$$\boldsymbol{\varepsilon^T \varepsilon} = (\boldsymbol{y - X\beta})^{\boldsymbol{T}} (\boldsymbol{y - X\beta}). \tag{1.21}$$

Taking partial derivatives of the expression with respect to $\boldsymbol{\beta}$ and setting them to null vector leads to the least squares normal equations for **b**

$$\boldsymbol{X^T X b = X^T y}. \tag{1.22}$$

If the square matrix $(\mathbf{X^T X})$ is non-singular, following from the full column rank assumption A2, the inverse of this matrix exists, and there is a unique solution to (1.22) obtained as

$$\boldsymbol{b = (X^T X)^{-1} X^T y}. \tag{1.23}$$

Hence, **b** is given by a linear transformation of the random vector **y** (Bašta, 2017).

1.4.2. Goodness of Fit

Once the parameter estimates have been obtained, it is necessary to assess how well the regression model fits the data at hand. Measures of goodness of fit summarize the disparity

between actual values of the dependent variable and the values expected under the model in consideration. Both with simple and multiple regression, it is reasonable to define the explained sum of squares (SSE), the residual sum of squares (SSR) and the total sum of squares (SST) as

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2, \tag{1.24}$$

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{1.25}$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2. \tag{1.26}$$

The explained sum of squares (SSE) is the sum of squared differences between the fitted values and the mean of the response variable, which describes how well the model fits the data. The residual sum of squares (SSR) is the sum of squared distances between observed and predicted values, which quantifies the remaining variability which was not captured by the model. The total sum of squares (SST) is the sum of squared differences between the observed response variable and its mean, which measures the dispersion of the response around its average value.

Thus, the total variation in y can be expressed as the sum of the explained and unexplained variation

$$SST = SSE + SSR \tag{1.27}$$

Considering that the total sum of squares, SST, being not equal to zero (which is true except the very rare case when all the y_i are equal to the same value) it is possible to derive the coefficient of determination, or R-squared, as

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}. \tag{1.28}$$

R^2 indicates the proportion of the sample variation in y that is explained by independent variables **X**. The value of R^2 is always between zero and one because SSE cannot exceed SST. A value of R^2that is nearly equal to zero is an evidence of a poor fit of the OLS model. On the contrary, the values close to 1 may signify that the OLS estimation provides a good fit to the data. For the purpose of interpretation, R^2 is usually multiplied by 100 to express the percentage of the variation in y explained by the model.

An important fact about the coefficient of determination, R^2, is that it never decreases, and moreover, usually increases when another regressor is added to the model. On the contrary, the adjusted R^2 imposes a penalty for the inclusion of an additional predictor to a model. The formula (1.29) for the adjusted R^2 shows that it depends explicitly on the number of independent variables k. Therefore, the adjusted R^2 can either increase or decrease, depending on the contribution of the new regressor to the fit of the regression (Wooldridge, 2015):

$$R^2_{adj} = 1 - (1 - R^2)\frac{(n-1)}{(n-k-1)}. \tag{1.29}$$

1.4.3. Properties of the OLS Estimators

Under the CLRM assumptions, discussed in section 1.3, the OLS estimators b_j are unbiased estimators of the population parameters β_j

$$E(b_j) = \beta_j, \text{ for all } j = 0, 1, \dots, k, \tag{1.30}$$

with the sampling variances

$$D(b_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}, \text{ for } j = 1, 2, \dots, k. \tag{1.31}$$

where

σ^2 is the error variance,

$SST_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ is the total sample variation in x_j and R^2 is the R-squared from regressing x_j on all other independent variables, and including an intercept (Wooldridge, 2015).

Under the matrix notation, the properties (1.30) and (1.31) are defined as

$$E(\boldsymbol{b}) = \boldsymbol{\beta}, \tag{1.32}$$

and

$$C(\boldsymbol{b}) = \sigma^2(\boldsymbol{X^T X})^{-1}. \tag{1.33}$$

The main-diagonal elements of the covariance matrix $\boldsymbol{C}(b)$ are variances of the least-squares estimators of individual regression parameters, and the off-diagonal elements are

covariances between the estimators. The matrix $\boldsymbol{C}(b)$ is entirely determined by the σ^2 and the model matrix **X**. Furthermore, such OLS estimators follow approximately the multivariate normal distribution:

$$\boldsymbol{b} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X^T X})^{-1}). \tag{1.34}$$

For construction of the confidence intervals and conducting hypothesis tests presented in chapter 2, it is necessary to estimate the standard deviation of b_j, which is the square root of the estimators variance

$$sd(b_j) = \frac{\sigma}{\sqrt{SST_j(1 - R_j^2)}}. \tag{1.35}$$

Since the theoretical error variance σ^2 is unknown in real life, it must be estimated from the available sample data. In the general multiple regression case, an unbiased estimator of σ^2 is the residual variance calculated as

$$s^2(e) = \frac{SSR}{n - k - 1}. \tag{1.36}$$

It follows that σ is replaced with its estimator, which gives the standard error of b_j

$$se(b_j) = \frac{s(e)}{\sqrt{SST_j(1 - R_j^2)}}. \tag{1.37}$$

Therefore, the unbiased estimator of the covariance matrix *C(**b**)* (Bašta, 2017) is defined as

$$S(\boldsymbol{b}) = s^2(e)(\boldsymbol{X^T X})^{-1}. \tag{1.38}$$

1.4.4. Weighted Least Squares

In response to the situation when the assumption of the constant error variance (A4) is violated, that is, in the presence of heteroskedasticity, a Weighted Least Squares (WLS) estimation may serve as an alternative to the Ordinary Least Squares. If the form of the heteroskedasticity as a function of explanatory variables is specified correctly, then the Weighted Least Squares approach is more efficient than the OLS and leads to the new *t* and *F* statistics that have *t* and *F* distributions (discussed in chapter 2).

Let **X** denote the model matrix containing all the information on the explanatory variables and assume that

$$D(\varepsilon|\boldsymbol{X}) = \sigma^2 w(\boldsymbol{X}), \tag{1.39}$$

where $w(\boldsymbol{X})$ is some function of the independent variables that determines the shape of the heteroskedasticity. Since variances must be positive, $w(\boldsymbol{X}) > 0$for all possible values of the explanatory variables. For a random drawing from the population, it can be written

$$\sigma_i^2 = D(\varepsilon_i|\boldsymbol{X_i}) = \sigma^2 w_i, \tag{1.40}$$

where $\boldsymbol{X_i}$ denotes all independent variables for observation *i*, and w_i changes with each observation because the independent variables change across observations.

To estimate the parameters β_j, the original equation (1.5) containing heteroskedastic errors

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \,,$$

is transformed into an equation that has homoskedastic errors and satisfies the other CLRM assumptions. Since w_i is just a function of X_i the following holds for the transformed error term, stemming from (1.40):

$$E\left(\frac{\varepsilon_i}{\sqrt{w_i}}|\boldsymbol{X_i}\right) = 0 \,, \tag{1.41}$$

$$D\left(\frac{\varepsilon_i}{\sqrt{w_i}}|\boldsymbol{X_i}\right) = \sigma^2 \,. \tag{1.42}$$

The equation (1.5) can be, therefore, divided by $\sqrt{w_i}$ to get

$$\frac{y_i}{\sqrt{w_i}} = \beta_0 \frac{1}{\sqrt{w_i}} + \beta_1 \frac{x_{i1}}{\sqrt{w_i}} + \beta_2 \frac{x_{i2}}{\sqrt{w_i}} + \cdots + \beta_k \frac{x_{ik}}{\sqrt{w_i}} + \frac{\varepsilon_i}{\sqrt{w_i}} \tag{1.43}$$

or equivalently

$$y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + \cdots + \beta_k x_{ik}^* + \varepsilon_i^* \tag{1.44}$$

where $x_{i0}^* = \frac{1}{\sqrt{w_i}}$.

The modified equation (1.44) satisfies the classical linear model assumptions (A1 through A6) if the initial model does so except for the homoskedasticity assumption. The parameter estimators b_j from this model will differ from the OLS estimators in the original equation and are the examples of Generalized Least Squares (GLS) estimators. In this particular case, the GLS estimators are used to correct for the heteroskedasticity in the errors and are termed the Weighted Least Squares (WLS) estimators. This name arises from the fact that the b_j minimize the weighted sum of squared residuals, where each squared residual is

weighted by $\frac{1}{w_i}$. The concept of the WLS is that less weight is given to the observations with a higher error variance, whereby the OLS assigns the same weight to each observation, assuming identical error variance for the whole population.

Mathematically, the WLS estimators are the values of the b_j that make the following expression as small as possible

$$\sum_{i=1}^{n} \frac{(y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik})^2}{w_i}. \tag{1.45}$$

In most situations, the exact form of heteroskedasticity is not apparent; hence, it is difficult to find the function $w(\boldsymbol{X})$. Nevertheless, it is convenient to model the function w_i and use the data to estimate the unknown parameters in this model. This results in an estimate of each w_i indicated as $\widehat{w}_i$. Using $\widehat{w}_i$ in place of w_i in the GLS transformation yields an estimator known as the Feasible Weighted Least Squares (FWLS) estimator (a special case of the Feasible Generalized Least Squares, FGLS, whereby the error terms are not correlated (Franzese and Kam, 2009).

There are many approaches to modeling heteroskedasticity, but one particular, reasonably flexible approach is considered in this section. Assume that

$$D(\varepsilon|\boldsymbol{X}) = \sigma^2 exp(\delta_0 + \delta_1 X_1 + \delta_2 X_2 + \cdots + \delta_k X_k)\,, \tag{1.46}$$

where

$X_1, \ldots, X_k$ are the independent variables appearing in the regression model equation (1.5) (for convenience, the subscripts *i* are omitted),

δ_j are unknown parameters.

The function $w(\boldsymbol{X})$ is then

$$w(\boldsymbol{X}) = exp(\delta_0 + \delta_1 X_1 + \delta_2 X_2 + \cdots + \delta_k X_k)\,. \tag{1.47}$$

The exponential function in (1.46) ensures that predicted values are positive since the estimated variances have to be positive in order to implement WLS. The parameters δ_j estimated from the sample data will serve for construction of the weights. Under the assumption (1.45), it can be written

$$\varepsilon^2 = \sigma^2 exp(\delta_0 + \delta_1 X_1 + \delta_2 X_2 + \cdots + \delta_k X_k)\nu. \tag{1.48}$$

where ν has a mean equal to unity, conditional on **X**. If ν is assumed to be independent of **X**, it is possible to write

$$\log(\varepsilon^2) = \alpha_0 + \delta_1 X_1 + \delta_2 X_2 + \cdots + \delta_k X_k + \nu'. \tag{1.49}$$

where ν' has a zero mean and does not depend on X. The intercept in this model differs from δ_0; however, it is not important in performing WLS. Since (1.49) satisfies the main assumptions, the unbiased estimators of δ_j can be obtained using OLS.

First, it is necessary to replace the unobserved ε with the OLS residuals *e*. Consequently, we run the regression of

$$\log(e^2) \; on \; X_1, X_2, \ldots, X_k \; . \tag{1.50}$$

After obtaining the fitted values from this regression, the estimates of $\widehat{w}_i$ can be simply derived through exponentiation

$$\widehat{w}_i = \exp\left(\widehat{\log(e_i^2)}\right) . \tag{1.51}$$

Now, the w_i are substituted with $\widehat{w}_i$ in the expression (1.45). It is necessary to remember that each squared residual is weighted by $\frac{1}{\widehat{w}_i}$. If all the variables are transformed in the first place and then the OLS is applied, each variable gets multiplied by $\frac{1}{\sqrt{\widehat{w}_i}}$ including the intercept.

Similarly to the OLS, the FGLS estimation measures the marginal impact each X_j has on *y*. However, if the heteroskedasticity problem arises, the FWLS estimators are usually more efficient, and associated test statistics have the usual *t* and *F* distributions, at least in large samples (Wooldridge, 2015).

In the matrix notation, the heteroskedastic regression model has the error covariance matrix

$$C(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{\Omega} = \sigma^2 \boldsymbol{W} \tag{1.52}$$

where **Ω** is a diagonal positive semidefinite matrix. The disturbances are still regarded as uncorrelated across observations, so the off-diagonal elements of the covariance matrix would be zeros

$$C(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \sigma^2\boldsymbol{W} = \sigma^2\begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} \tag{1.53}$$

where the variance of the disturbances depends on the predictor values of the respective observation *i*.

Thereby, the classical linear regression with homoskedastic error terms is a special case with $w_i = 1$ for all $i = 1,2,\ldots, n$ (Greene, 2003). The matrix **W** equals to the identity matrix **I**, and the resulting the covariance matrix is

$$C(\boldsymbol{\varepsilon}) = \boldsymbol{\sigma^2 I}\,. \tag{1.54}$$

It is possible to find an invertible matrix **P** such that

$$\boldsymbol{P^T P = W^{-1}}, \tag{1.55}$$

and

$$\boldsymbol{I = PWP^T}. \tag{1.56}$$

If both sides of the equation $\boldsymbol{y = X\beta + \varepsilon}$, are premultiplied by the matrix **P**, the modified regression model is defined as

$$\boldsymbol{Py = PX\beta + P\varepsilon}\,. \tag{1.57}$$

Defining $\mathbf{q} \equiv \mathbf{Py}$, $\mathbf{Q} \equiv \mathbf{PX}$ and $\mathbf{u} \equiv \mathbf{P}\boldsymbol{\varepsilon}$, equation (1.57) can be equivalently written as

$$\boldsymbol{q = Q\beta + u}\,. \tag{1.58}$$

It can be proved, that in this transformed equation, the expectation and the variance of the error term $\boldsymbol{u}$, conditioned on the model matrix **X** are

$$E(\boldsymbol{u}) = E(\boldsymbol{P\varepsilon}) = \mathbf{0}\,, \tag{1.59}$$

$$C(\boldsymbol{u}) = C(\boldsymbol{P\varepsilon}) = \sigma^2\boldsymbol{I}\,. \tag{1.60}$$

Therefore, the classical regression model applies to this transformed model. The vector of the error terms $\boldsymbol{u}$ in the equation (1.58) satisfied the assumption A4. Thus, OLS estimator of $\boldsymbol{\beta}$ becomes a GLS estimator, denoted as $\boldsymbol{b}_G$, which is obtained by minimizing the generalized sum of squares with respect to $\boldsymbol{\beta}$

$$\boldsymbol{u^T u = (q - Q\beta)^T (q - Q\beta)}\,, \tag{1.61}$$

or equivalently

$$(\boldsymbol{y} - \boldsymbol{X\beta})^{\boldsymbol{T}}\mathbf{W}^{-\mathbf{1}}(\boldsymbol{y} - \boldsymbol{X\beta}) , \tag{1.62}$$

and is given as

$$\boldsymbol{b}_G = (\boldsymbol{Q}^{\boldsymbol{T}}\boldsymbol{Q})^{-\mathbf{1}}\boldsymbol{Q}^{\boldsymbol{T}}\boldsymbol{q} . \tag{1.63}$$

Since **W** is a diagonal matrix such that

$$\boldsymbol{W} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix},$$

the diagonal elements of $\mathbf{W}^{-\mathbf{1}}$ are given as $\frac{1}{w_i}$

$$\boldsymbol{W}^{-\mathbf{1}} = \begin{bmatrix} 1/w_1 & 0 & \cdots & 0 \\ 0 & 1/w_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1/w_n \end{bmatrix}. \tag{1.64}$$

Consequently, the matrix **P** can be chosen such that its diagonal values are equal to $\frac{1}{\sqrt{w_i}}$ (Bašta, 2017):

$$\boldsymbol{P} = \begin{bmatrix} 1/\sqrt{w_1} & 0 & \cdots & 0 \\ 0 & 1/\sqrt{w_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1/\sqrt{w_n} \end{bmatrix}. \tag{1.65}$$

Since the matrix of weights is unknown in the real-life situation, the procedure described above is used to estimate the weights and to transform the original regression equation. Hence, finding the weighted least-squares estimators amounts to minimizing

$$\sum_{i=1}^{n} \frac{e_i^2}{w_i}. \tag{1.66}$$

All the results for the classical model, such as usual inference procedures, apply to the transformed model in (1.58).

However, there is no explicit counterpart to R^2 in the generalized regression model. As seen from the equation (1.43), the transformed regression (1.58) need not have a constant intercept, so the R^2 is not bounded by zero and one.

2. Statistical Inference

This chapter addresses the problem of testing the hypotheses about the parameters in the population regression model.

2.1. Hypothesis Testing

Once the parameters in the model (1.5) have been estimated, it is necessary to assess the overall adequacy of the model and the importance of specific regressors. Several hypothesis testing methods may serve for this purpose. To ensure that the formal tests provide reliable results, it is essential that the random disturbances follow approximately normal distribution with zero mean and constant variance.

For a full comprehension of hypothesis testing, it is necessary to remember that the β_j are unknown characteristics of the population, and they will never be known with certainty. Nevertheless, an analyst can hypothesize about the value of β_j and then conduct statistical inference to test the hypothesis of interest.

The null hypothesis, shortly H_0, is the hypothesis being tested. To perform the testing of H_0, one must calculate a test statistic, which is a random variable with a known distribution under the null hypothesis. When the null hypothesis is false, the test statistic has some other distribution (Davidson and MacKinnon, 2003).

The explicit rejection rule depends on the alternative hypothesis, against which H_0 is tested, and the chosen significance level of the test α, that is, the probability of rejecting H_0 when it is, in fact, true (Wooldridge, 2015).

2.1.1. Test for Overall Significance of a Regression: The *F*-Test

The test for significance of regression helps to see whether a linear relationship between the response y and any of the regressor variables $X_1, \ldots, X_k$ exists or not. This procedure often evaluates overall adequacy of the model. The tested null hypothesis is

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0\,. \tag{2.1}$$

This test is a joint test of the hypothesis that all the coefficients except the constant term are zero; thus, none of the explanatory variables has an impact on y. The alternative hypothesis is then

$$H_1: \beta_j \neq 0\,, \quad \text{for at least one } j, \tag{2.2}$$

which implies that at least one of the predictors $X_1, \ldots, X_k$ contributes significantly to the model.

The *F*-test is an example of a set of multiple restrictions since several restrictions are imposed on the regression parameters.

If $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ is not rejected, it indicates that all explanatory variables $X_1, \ldots, X_k$ have no effect on the response variable and might be excluded from the model.

In its general form, the *F*-statistic (or *F*-ratio) used for testing the null hypothesis is given as

$$F = \frac{(SSR_r - SSR_{ur})/J}{SSR_{ur}/(n-k-1)}\,, \tag{2.3}$$

where

J is the number of explicitly imposed restrictions on the parameters of the general linear hypothesis in the regression (J parameters are equal to 0),

SSR_r, SSR_{ur} are the sums of squared residuals from the restricted and unrestricted models, respectively.

For testing restrictions, it is often convenient to compute the *F*-statistic using the coefficients of determination, R^2, from the restricted and unrestricted models. Thus, the formula in (2.3) can be equivalently defined as

$$F = \frac{(R^2_{ur} - R^2_r)/J}{(1-R^2_{ur})/(n-k-1)}\,, \tag{2.4}$$

where R^2_r and R^2_{ur} are the R-squareds from the restricted and unrestricted models respectively.

Assuming the CLRM assumptions hold, it can be shown that under H_0, F is distributed as an F random variable with $(J, n-k-1)$ degrees of freedom

$$F \sim F_{J,\ n-k-1}\,. \tag{2.5}$$

When testing for the global significance of a regression model, $J = k$ meaning that there are k restrictions in (1.5), and when they are imposed, the restricted model takes the form

$$y_i = \beta_0 + \varepsilon_i\,. \tag{2.6}$$

That is, all independent variables have been dropped from the equation. Now, the R_r^2 from estimating (2.6) is zero: the model explains none of the variation in *y* because it does not contain explanatory variables. Therefore, the *F*-statistic for testing (2.1) is

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)}, \tag{2.7}$$

Where R^2 is just the usual R-squared from the regression of *y* on all independent variables, and the test statistic has the following distribution

$$F \sim F_{k,\ n-k-1}\,. \tag{2.8}$$

One will reject H_0 in favor of H_1 when *F* is sufficiently "large", exceeding the $(1-\alpha)\times 100\%$ percentile of an *F* distribution with *(k, n – k – 1)* degrees of freedom. The rejection region is defined as

$$W_\alpha = \{F > F_{1-\alpha,\ k,\ n-k-1}\}\,. \tag{2.9}$$

If H_0 is rejected, it can be stated that $X_1, \ldots, X_k$ are jointly statistically significant at the corresponding significance level. This test alone does not allow to determine, which of the variables have a partial effect on *y*: they may all have an impact on *y*, or maybe only one predictor affects *y*. If H_0 is not rejected, then the regressors are jointly insignificant, which often justifies dropping them from the model (Wooldridge, 2015).

2.1.2. Test on Individual Regression Coefficients: The *t*-Test

Once the *F*-test detected that at least one of the regressors is significant, the next step is to define which one. Adding a variable to a regression equation always causes the explained sum of squares (SSE) to increase. However, the inclusion of a regressor also increases the variance of the fitted value $\hat{y}$, so one must preferably include only those regressors that are useful for explaining the response (Montgomery, 2013).

The null hypotheses for testing the significance of any individual regression coefficient β_j, are

$$H_0: \beta_j = 0\,, \tag{2.10}$$

where *j* corresponds to any of the *k* independent variables. Since β_j reflects the partial effect of X_j on the expected value of *y* under ceteris paribus condition, (2.10) means that, once $X_1, X_2, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k$ have been controlled for, X_j has no influence on the

expectation of y. Thereby, this is a test of the contribution of X_j given the other variables in the model (Wooldridge, 2015).

In case of a one-tailed t-test, we are testing for the possibility of the relationship in one direction and absolutely ignoring the possibility of a relationship in the other direction:

$$H_1: \beta_j > 0 \quad \text{or} \quad H_1: \beta_j < 0\,. \tag{2.11}$$

That is the expected population value of β_j is significantly greater (or less) than 0, and the corresponding predictor X_j has a positive (or negative) effect on the expected value of the outcome.

Classical statistical inference procedures presume that the null and alternative hypothesis are stated before analyzing the data. Using a two-sided test prevents the researched from looking at the estimated model and then formulating hypotheses on the population coefficients. Hence, in many applications, it is common to test the null hypothesis (2.10) against a two-sided alternative

$$H_1: \beta_j \neq 0. \tag{2.12}$$

Under this alternative hypothesis, X_j has either positive or negative ceteris paribus effect on y. The two-tailed test is a suitable alternative when the sign of β_j is not entirely determined by the theory or common sense.

If we fail to reject $H_0: \beta_j = 0$, the regressor X_j might be omitted from the model. Nonetheless, the results of the statistical hypothesis test should be viewed as a guideline, and deleting the variables from the model depends on the research question and knowledge about the process.

The statistic used to test (2.7) against any alternative is called the t-statistic or the t-ratio of β_j and is defined as

$$t_{b_j} = \frac{(b_j - \beta_j)}{se(b_j)} = \frac{b_j}{se(b_j)}, \tag{2.13}$$

where

β_j is the hypothesized value of the population parameter being tested (zero in this particular case),

$se(b_j)$ is the standard error of b_j from the equation (1.37).

Under the CLRM assumptions A1 through A6, the *t*-statistic has *t* distribution with *(n – k – 1)* degrees of freedom

$$t_{b_j} \sim t_{n-k-1}\,. \tag{2.14}$$

When the test is two-tailed, the absolute value of the *t*-statistic is taken. The rejection rule for H_0 against H_1 is

$$W_\alpha = \{|t_{b_j}| \geq t_{1-\frac{\alpha}{2},\ n-k-1}\}\,, \tag{2.15}$$

where the critical value is chosen to make the area in each tail of the *t* distribution equal to $\left(\frac{\alpha}{2}\right) \times 100\%$.

In case of rejection of H_0 in favor of H_1 at the $\alpha \times 100\%$ significance level, X_j is considered to be statistically different from zero. Otherwise, the explanatory variable X_j is thought of as statistically insignificant (Wooldridge, 2015).

2.2. Univariate and Joint Confidence Regions on Regression Coefficients

Point estimation returns single values as an estimation of the unknown population parameters. The point estimators are considerably useful; however, they do not carry as much information on the parameters of interest as the interval estimators.

Since there is a degree of uncertainty whether the estimated value is close to the true parameter value or not, the interval estimation solves this issue by construction an interval around the point estimate. Each interval built with regard to a prespecified confidence level $(1-\alpha) \times 100\%$ is called confidence interval, and it is supposed to contain the real value of population parameter with $(1-\alpha) \times 100\%$ probability.

2.2.1. Univariate Confidence Intervals

Accounting for the fact, that t_{b_j} has a *e* distribution with *(n – k – 1)* degrees of freedom and assuming that the interval of interest would be symmetric around b_j, a $(1-\alpha) \times 100\%$ confidence interval for parameter β_j is derived as

$$b_j \pm t_{1-\frac{\alpha}{2},\ n-k-1} \times se(b_j)\,. \tag{2.16}$$

More precisely, the lower and upper limits of the confidence interval are given by

$$\underline{b_j} \equiv b_j - t_{1-\frac{\alpha}{2},\ n-k-1} \times se(b_j) \tag{2.17}$$

and

$$\overline{b_j} \equiv b_j + t_{1-\frac{\alpha}{2},\ n-k-1} \times se(b_j) \tag{2.18}$$

respectively. This means, that if random samples were drawn from the population over and over again, with $\underline{b_j}$ and $\overline{b_j}$ calculated each time, then the unknown population value β_j would lie in the interval $[\underline{b_j}; \overline{b_j}]$ for $(1-\alpha) \times 100\%$ of the samples (Wooldridge, 2015).

2.2.2. Simultaneous Confidence Intervals

For the classical univariate case, the confidence coefficient $(1-\alpha)$ applies to each confidence interval. Nevertheless, in the multiple linear regression, the degree of confidence connected to the statement that all $(p = k + 1)$ intervals simultaneously comprise their respective parameters is much lower. One of the relatively simple procedures that retains the joint confidence coefficient for several simultaneous statements near a preselected level $(1-\alpha)$ is called the Bonferroni method.

The confidence intervals are constructed as given in (2.16), but using $\alpha^* = \frac{\alpha}{p}$, where p is the number of simultaneous intervals or statements. That is, in the expression (2.16), $t_{1-\frac{\alpha}{2},\ n-k-1}$ is substituted by $t_{1-\frac{\alpha}{2p},\ n-k-1}$. This approach ensures that the true joint confidence coefficient for the p simultaneous statements is at least $(1-\alpha)$.

The Bonferroni simultaneous confidence intervals for the p parameters β_j are given by

$$b_j \pm t_{1-\frac{\alpha}{2p},\ n-k-1} \times se(b_j)\,. \tag{2.19}$$

Therefore, Bonferroni confidence intervals look somewhat like the regular one-at-a-time confidence intervals based on the t distribution, except that each Bonferroni interval has a confidence coefficient $\left(1-\frac{\alpha}{p}\right)$ instead of $(1-\alpha)$.

The simultaneous confidence intervals obtained using Bonferroni approach provides confidence intervals for each individual parameter β_j in such a way that the p-dimensional region produced by the intersection of the p simultaneous confidence intervals gives at

least a $(1-\alpha)\times 100\%$ joint confidence interval for all parameters. The shape of this simultaneous confidence interval is rectangular (for $p = 2$) or cubic (for $p = 3$).

When the number of parameters is small, the Bonferroni simultaneous confidence intervals are not very wide. However, if p is large, the Bonferroni-adjusted intervals tend to be very broad, and the simultaneous coverage may be much larger than the specified confidence level $(1-\alpha)$ (Rawlings et al., 1998).

2.2.3. Joint Confidence Regions

In case of the confidence intervals, the inferences about the population parameters are implicitly based on the marginal distributions of their estimates. However, if those estimates are not independent, it is preferable to construct a confidence region that captures their joint distribution.

The confidence intervals described in the previous sections are all derived by inverting the *t*-tests. Each *t*-statistic depends explicitly on individual parameters and their standard errors. In order to construct the confidence regions, the joint tests for several parameters should be inverted. These are generally the tests based on the statistics which follow the *F* distribution, as they depend on the vectors of estimates b_j and their covariance matrix ***C(b)*** (Davidson and MacKinnon, 2003).

Mathematically, the joint confidence region for all p parameters in $\boldsymbol{\beta}$ results from the inequality

$$(\boldsymbol{\beta}-\boldsymbol{b})^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})(\boldsymbol{\beta}-\boldsymbol{b}) \leq p\, s^2(e)\, F_{1-\alpha,\ p,\ n-p} \qquad (2.20)$$

where

$F_{1-\alpha,\ p,\ n-p}$ is the value of the *F* distribution with *(p, n – p)* degrees of freedom that leaves probability α in the upper tail,

$s^2(e)$ is the residual variance, defined in the equation (1.36).

The left-hand side of the inequality (2.17) expresses a quadratic form in $\boldsymbol{\beta}$, because **b**, $\mathbf{X^TX}$, and the right-hand side are known quantities calculated from the data. Solving this quadratic form for the inequality boundary establishes a p – dimensional ellipsoid, which represents the $(1-\alpha)\times 100\%$ joint confidence region for all the parameters in the

model. The slopes of the axes and eccentricity of the ellipsoid display the direction and strength, respectively, of the correlation between the parameter estimates.

A disadvantage of this statistical technique is that the ellipsoidal confidence regions with more than two or three dimensions do not have straightforward interpretation. A possible approach to using the p-dimensional joint confidence region for all regression parameters is to construct confidence regions for two parameters at a time ignoring the other $(p - 2)$ parameters. The confidence level $(1 - \alpha) \times 100\%$ then applies to the joint statement with regards to two parameters being analyzed at the time. This procedure focuses on the joint distribution of two parameter estimates but disregards the values of the other parameters. For this reason, the joint confidence region approach suffers from the same conceptual issue as the univariate confidence intervals (Rawlings et al., 1998).

3. Residual Diagnostics

3.1. Assessment of Regression Function Specification: RESET test

Misspecification of the functional form in a multiple regression model occurs when it does not correctly capture the relationship between the response and the observed explanatory variables (Wooldridge, 2015).

If in the original model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{3.1}$$

the assumption $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ does not hold, then $\boldsymbol{y} \neq \boldsymbol{X\beta}$, and the regression function is regarded as misspecified. Improperly specifying the functional form can have serious consequences if the aim of the analysis is statistical inference about the population. On the contrary, the assumption of a correct regression function specification is not required for constructing models for prediction purposes (Bašta, 2017).

A useful way to conduct specification tests is as if the original model (3.1) is the null hypothesis, and the alternative is some unstated generalization of that model. Ramsey's (1969) Regression Specification Error Test (RESET) is one such test which attempts to reveal the nonlinearities in the functional form. A straightforward approach would be to include squares, cubes, and interactions of the regressors to the equation and view H_0 as a restriction on the extended model. Thus, one way to formulate the hypothesis is

$$H_0: \ y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon \tag{3.2}$$

$$H_1: y = \beta_0 + \beta_1 X_1 + \cdots + higher\ order\ powers\ of\ X_k\ and\ other\ terms + \cdots + \beta_k X_k + \varepsilon. \tag{3.3}$$

The complication is that with a large number of variables in **X**, the model could become cumbersome.

As an alternative solution, Ramsey proposed to add powers of the OLS fitted values $\hat{y}$ - typically, the square and, possibly, the cube - to detect general forms of the model misspecification. This approach requires a two-step estimation procedure since the coefficients are needed in order to obtain $\hat{y}^2$ and $\hat{y}^3$. The suggestion is to fit the null model first, applying the least squares. Afterward, the second step is to add the squares and cubes of the predicted values from the first-step estimation to the equation and refit it with these additional variables (Greene, 2003). Hence, the alternative hypothesis is stated as

$$H_1: \ y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + \varepsilon \, . \tag{3.4}$$

Functions of the fitted values from the initial regression now act as explanatory variables; however, the estimated parameters from (3.4) are not of the primary interest. This equation serves for testing whether some significant nonlinearities have not been accounted for.

Under the null hypothesis, the regression equation (3.2) is correctly specified, and the parameters corresponding to the powers of the fitted values are statistically insignificant. The RESET is the *F*-statistic for testing

$$H_0: \ \delta_1 = \delta_2 = 0 \tag{3.5}$$

in the expanded model (3.4). A significant *F*-statistic implies a rejection of the null hypothesis and assumes some functional form problem. In large samples, under H_0 and the CLRM assumptions, the *F*-statistic has approximately *F* distribution with $(2, n - k - 3)$ degrees of freedom (Wooldridge, 2015):

$$F \sim F_{2, \ n-k-3} \, . \tag{3.6}$$

The prominent advantage of such a test is that it ensures much greater generality than a simple test of restrictions such as whether a coefficient or a set of coefficients are equal to zero. Still, a shortcoming of the RESET test is that it does not suggest any direction in which the researcher should proceed if the null model is rejected. This is a common feature of the specification tests, whereby the rejection of the null model does not presume any particular alternative (Greene, 2003).

3.2. Assessment of Homoskedasticity of Errors

A variety of tests for heteroskedasticity of the disturbances have been applied over the years. Some of them, while being able to detect heteroskedasticity, do not test the assumption that the error variance is independent from the explanatory variables. This section describes the two modern tests that detect the nature of heteroskedasticity, under which the usual OLS statistics become invalid. If heteroskedasticity is discovered by one of those tests, the Weighted Least Squares (WLS) described in the section (1.4.4.) is a suitable alternative to the OLS.

3.2.1. The Breusch-Pagan Test for Heteroskedasticity

For the linear model

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon\,, \tag{3.7}$$

the condition that $E(\varepsilon|X) = 0$ is assumed to hold so that OLS estimators are unbiased and consistent. The null hypothesis states that the assumption of homoskedasticity is true:

$$H_0\colon\ D(\varepsilon|\boldsymbol{X}) = \sigma^2. \tag{3.8}$$

Due to the assumption of a zero conditional expectation of the error, the variance can be written as $D(\varepsilon|X) = E(\varepsilon^2|X)$, and so the null hypothesis of homoskedasticity is equivalent to

$$H_0\colon\ E(\varepsilon^2|\boldsymbol{X}) = E(\varepsilon^2) = \sigma^2. \tag{3.9}$$

Thus, the test for violation of the homoskedasticity assumption constitutes of testing whether the expected value of ε^2 is associated with one or more of the explanatory variables. If H_0 is rejected, the expectation of ε^2, given the predictors, can be practically any function of the X_j. One of the approaches is to assume a linear function:

$$\varepsilon^2 \;=\; \delta_0 + \delta_1 X_1 + \delta_2 X_2 + \cdots + \delta_k X_k + \nu\,, \tag{3.10}$$

where ν is an error term with mean zero given the X_j.

The null hypothesis of homoskedasticity is formulated as

$$H_0\colon\ \delta_1 = \delta_2 = \cdots = \delta_k = 0\,. \tag{3.11}$$

The actual errors in the population model are never known but have to be estimated by the OLS residuals, e_i. Thus, after estimation of the equation

$$e^2 \;=\; \delta_0 + \delta_1 X_1 + \delta_2 X_2 + \cdots + \delta_k X_k + error\,, \tag{3.12}$$

it is possible to compute the *F* or *LM* (Lagrange multiplier) statistics for the joint significance of $X_1, \ldots, X_k$. The *F* and *LM* statistics both depend on the R-squared from regression (3.12) denoted as $R^2_{e^2}$.

The *LM* version of the test is referred to as the Breusch-Pagan test for heteroskedasticity (BP test), and the *LM* statistic for amounts to the product of the sample size and the R-squared from (3.12):

$$LM = n\, R^2_{e^2}\,. \tag{3.13}$$

The *LM* statistic has Chi-Square distribution with *k* degrees of freedom

$$LM \sim \chi_k^2 . \tag{3.14}$$

A sufficiently small *p*-value, that is, below the chosen significance levelα, leads to the rejection of the null hypothesis of homoskedasticity (Breusch and Pagan, 1980).

3.2.2. The White Test for Heteroskedasticity

According to Wooldridge (2015), the assumption $D(\varepsilon|\boldsymbol{X}) = \sigma^2$ can be replaced with a weaker assumption that the squared disturbance, ε^2, is uncorrelated with all the explanatory variables X_j, their squares X_j^2, and all the interaction terms $X_j X_h$ for $j \neq h$, This feature motivated White (1980) to introduce a test for heteroskedasticity that incorporates the squares and cross products of all the predictors to the original model equation (3.7). The test explicitly aims to test for nature of heteroskedasticity that renders the OLS standard errors and test statistics invalid.

However, the pure form of the White's test is weakened by the abundance of regressors: it consumes many degrees of freedom for models even with a small number of explanatory variables. For instance, if only three independent variables appear in the initial equation, the transformed model will have nine independent variables (three original predictors, three squared predictors, and three interaction terms).

The idea of the White test can be retained while saving the degrees of freedom by utilizing the OLS fitted values to test for heteroskedasticity. The squared fitted values represent a specific function of all the squares and cross products of the explanatory variables. Similarly to the RESET test, the modified version of the White test suggests regressing the original OLS residuals on the fitted values $\hat{y}$ and their squares $\hat{y}^2$

$$e^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + error . \tag{3.15}$$

The *F* or *LM* statistics can be used to test the null hypothesis

$$H_0: \ \delta_1 = \delta_2 = 0 \tag{3.16}$$

in the equation (3.15). The null hypothesis now results in testing two restrictions which support homoskedasticity assumption, regardless of the number of predictors in the original regression model.

The *LM* statistic for this test is calculated in the same way as for the Breusch-Pagan test discussed above. However, the test statistic has Chi-Square distribution with only 2 degrees of freedom

$$LM \sim \chi_2^2 \,. \tag{3.17}$$

3.3. Assessment of Normality of Errors

Deviation from the normality assumption may result in the suboptimal estimation, invalid inferential procedures, and erroneous conclusions, highlighting the necessity of the assumption assessment (Jarque and Bera, 1987).

Typically, evaluation of the relevant residual plots is sufficient to diagnose departures from normality. However, more formal quantification of normality should be used together with the graphical diagnostics. Therefore, the researcher can conduct hypothesis tests stating the null hypothesis that the errors are normally distributed. For each test reviewed below, the formal hypotheses are written as:

$$H_0 \colon \ \varepsilon \sim N(0, \sigma^2) \tag{3.18}$$

$$H_1 \colon \ non - H_0 \,. \tag{3.19}$$

A large *p*-value suggests that it is plausible to assume that the disturbances follow a normal distribution. This section describes some common non-parametric testing methods for normality.

3.3.1. The Shapiro-Wilk Test

Shapiro and Wilk proposed a test for normality of a sample data, which is now one of the most frequently used tests. The *W*-test is based on the *W* test statistic

$$W = \frac{(\sum_{i=1}^{n} a_i e_{(i)})^2}{\sum_{i=1}^{n} (e_i - \bar{e})^2} \tag{3.20}$$

where

i is the rank of each value of e_i,

$e_{(i)} \colon e_{(1)} \leq e_{(2)} \leq \cdots \leq e_{(n)}$ are the ordered values of a sample $e_1, e_2, \ldots, e_n$,

a_i are constants generated from the means, variances and covariances of the order statistics of a sample of size *n* from a normal distribution (for the derivation of a_i see Shapiro and Wilk, 1965).

Obtained test statistic *W* is compared against tabulated values of its distribution. Its maximum achievable value is 1, and small values of *W* are evidence of departure from normality and lead to rejection of the null hypothesis.

3.3.2. The Lilliefors Test

The Lilliefors Test is a modification of the Kolmogorov-Smirnov test that compares the empirical cumulative distribution function (ECDF) of the sample with the distribution anticipated if the data were normal. If the difference between these distributions is significantly large, the test rejects the null hypothesis of the population normality. The test statistic is given by:

$$D = max_e\left|F_n\left(e_{(i)}\right) - F\left(e_{(i)}\right)\right| \ , \tag{3.21}$$

where

$F\left(e_{(i)}\right)$ is the cumulative normal distribution function (CDF) with parameters given as the residual mean $\mu = \bar{e}$ and variance $\sigma^2 = s^2(e)$,

$F_n\left(e_{(i)}\right)$ is the empirical distribution function for *n* ordered variables $e_{(1)} \leq e_{(2)} \leq \cdots \leq e_{(n)}$.

The rejection of the null hypothesis requires that the maximum discrepancy is large enough opposed to the tabulated critical values to be statistically significant (Lilliefors, 1967).

3.3.3. The Cramér-von Mises Test

The Cramér-von Mises test is another approach for evaluating the goodness of fit, which tests whether the sample follows a specified continuous distribution. It uses the squared differences between observed and theoretical cumulative distribution functions as the test statistic. In case of the one-sample test, the Cramér-von Mises criterion is defined as

$$\omega^2 = \int_{-\infty}^{\infty}\left[F_n\left(e_{(i)}\right) - F\left(e_{(i)}\right)\right]^2 dF\left(e_{(i)}\right) \ , \tag{3.22}$$

where $F(e_{(i)})$ is the specified normal cumulative distribution function equivalent to that in the case of the Lilliefors test. If values $e_{(i)}$ have been standardized in advance, $F(e_{(i)}) = \Phi(e_{(i)})$ representing the standard normal distribution.

The test statistic T (Anderson, 1962) is then

$$T = n\,\omega^2\,, \tag{3.23}$$

or equivalently (Stephens, 1974)

$$T = \frac{1}{12n} + \sum_{i=1}^{n}\left[\frac{2i-1}{2n} - F(e_{(i)})\right]^2\,. \tag{3.24}$$

If the computed value of T exceeds the tabulated critical value, then the null hypothesis that the disturbances follow the normal distribution can be rejected.

3.3.4. The Anderson-Darling Test

Similarly to the Cramér-von Mises test, the Anderson-Darling statistic belongs to the family of quadratic EDF statistics.

Compared to the Lilliefors and Cramér-von Mises tests, the Anderson-Darling test puts more weight to the tails of the distribution. It is commonly viewed as one of the most powerful tests of normality, even when applied to small samples.

Mathematically, the test statistic is given by:

$$A^2 = n\int_{-\infty}^{\infty}[F_n(e_{(i)}) - F(e_{(i)})]^2\;\psi(e_{(i)})\;dF(e_{(i)})\,, \tag{3.25}$$

where $\psi(e_{(i)})$ is a non-negative weighting function obtained as

$\psi(e_{(i)}) = \left[F(e_{(i)})\left(1 - F(e_{(i)})\right)\right]^{-1}$ (Anderson and Darling, 1954).

Equivalently to the previous procedures, the Anderson-Darling test is a one-sided test, which rejects the hypothesis of the error normality if the test statistic, A^2, is greater than the critical value.

4. Outliers and Influential Observations

In some practical applications, mainly, but not only, with small data sets, the OLS parameter estimates are susceptible to the presence of observations which are significantly remote from the majority of the data points. Outliers can occur due to the inaccurate data entry or when the sample is drawn from a small population. In the latter case, one or several members of the population may greatly differ in some certain features from the rest of the population (Wooldridge, 2015). Typically, practical applications distinguish between two types of observations that substantially differ from all other ones: outliers in the response variable *y*, known simply as „outliers“, and outliers with respect to the explanatory variables, called „leverage points“ (Blatná, 2006).

The researchers are mostly interested in the “regression outliers”, which correspond to the observations whose values of both the response and explanatory variables deviate from the regression relationship followed by the majority of the data. The OLS is sensitive to such outlying data points because it minimizes the sum of squared residuals: large residuals (either positive or negative) receive high weights in the least squares minimization procedure (Wooldridge, 2015).

The observations are said to be influential if their inclusion or exclusion from the estimation procedure leads to significant changes in the fitted model – regression coefficients and fitted values (Blatná, 2006). One should be concerned if the estimates change by a noticeable large amount when the sample is slightly modified.

4.1. Leverage: Hat-Values

Observations that are comparatively distant from the center of the regressor space **X**, accounting for the correlational pattern among the regressors, have a potentially greater impact on the OLS estimates of the regression coefficients. Such points are assumed to have high leverage. The most well-known measures of leverage are the hat-values, which come from the relationship between the observed vector of the dependent variable and the vector of fitted values (Fox and Weisberg, 2011). The fitted values are given by

$$\hat{\boldsymbol{y}} = \boldsymbol{Xb} = \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} = \boldsymbol{Hy} \ , \tag{4.1}$$

where **H** is the hat (projection) matrix

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}^{\boldsymbol{T}} . \tag{4.2}$$

The hat matrix (or the projection matrix) maps the vector of the response values to the vector of fitted values and plays an important role in identifying influential observations.

The hat matrix **H** defines the variance-covariance matrices of $\hat{\boldsymbol{y}}$ and **e**, such that

$$C(\hat{\boldsymbol{y}}) = \boldsymbol{\sigma}^{\mathbf{2}}\boldsymbol{H} \tag{4.3}$$

and

$$C(\boldsymbol{e}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H}). \tag{4.4}$$

The off-diagonal elements h_{ij} of the matrix **H** may be viewed as the measure of leverage exerted by the i^{th} observation y_i on the j^{th} fitted value $\hat{y}_j$.

However, the attention is usually focused on the diagonal elements h_{ii} of the hat matrix **H** - leverage values - written as

$$h_{ii} = \boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{x}_{\boldsymbol{i}} \ , \tag{4.5}$$

where $\boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{T}}$ is the i^{th} row of the **X** matrix.

The hat matrix diagonal measures the distance between the i^{th} observation and the centroid of the **X** space. The hat-values are bounded by 0 and 1, and those values that are close to 1 indicate observations that are likely to be influential because they are far from the rest of the sample in the **X** space.

Since the trace of **H**, given as the sum of diagonal values, is equal to the number of parameters *p*

$$tr(\boldsymbol{H}) = \sum_{i=1}^{n} h_{ii} = p \ , \tag{4.6}$$

it follows that the average size of a hat diagonal is

$$\overline{h} = \frac{p}{n} \ . \tag{4.7}$$

Traditionally, the rule of thumb is that any observation for which the hat-value is greater than twice the average $2\frac{p}{n}$ is remote enough from the remaining data to be treated as a leverage point.

Not all leverage points necessarily influence the regression coefficients. Some data points that have a large value on the hat diagonal and are definitely a leverage point might have

almost no effect on the regression coefficients if they lie virtually on the line passing through the rest observations. Since the hat diagonals explore only the location of observations in the regressor space, it is useful to examine the studentized residuals in connection with the h_{ii}. Observations with both large hat-values and large residuals are most likely to be influential (Montgomery et al., 2012).

4.2. Regression Outliers: Externally Studentized Residuals

The studentized residuals serve as a helpful criterion for regression outliers identification, that is the observations whose values of the response variable y_i conditional on the combination of the regressors $x_{i1}, x_{i2}, \ldots, x_{ik}$ considerably differ from the linear relationship which holds for the major part of the data.

The basic concept of the externally studentized residuals is to remove the observations one at a time, each time refitting the regression model using the remaining *(n* – 1*)* observations. Then, the observed response values are compared to their fitted values based on the models with the ith observation omitted. This technique produces deleted (predicted) residuals which, after standardization, are known as the studentized residuals.

Computationally, a studentized deleted (or externally studentized) residual is obtained as

$$e_{Ji} = \frac{e_i}{s_{(-i)}(e)\sqrt{1 - h_{ii}}}\ , \tag{4.8}$$

where

e_i is the ordinary residual from the model estimated with the complete number of observartions,

$s_{(-i)}(e)$ is the standard error of the estimated model with the ith observation deleted (Bašta, 2017).

The logic behind the equation (4.8) is that if the i^{th} observation y_i is, in fact, unusual, the regression model based on the complete set of observations may be overly affected by this particular observation. This estimation could produce a fitted value $\hat{y}_i$ that is quite similar to the observed value y_i, and as a result, the ordinary residual e_i will be small, leading to difficulty of detecting the outlier. However, if the i^{th} observation is deleted, it will not influence the newly fitted value $\hat{y}_{(-i)}$, so the resulting residual should be likely to indicate the presence of the regression outlier (Montgomery et al., 2012).

Under the classical regression assumptions, the externally studentized residual e_{Ji} follows the *t* distribution with *(n – p – 1)* degrees of freedom

$$e_{Ji} \sim t_{n-p-1} \ . \tag{4.9}$$

An outlier test for studentized residuals is performed by comparing the absolute value of the studentized residual with the threshold value |2|. Points with the corresponding studentized residuals that exceed the threshold value are reported as significant differences, that is regression outliers in this case. Special attention should be paid to the observations whose residuals exceed |3| (Blatná, 2006).

4.3. Influence Measures

As mentioned earlier, a data point that is both outlying and has high leverage exerts influence on the regression coefficients, in the sense that removal of this observation leads to a considerable change in the coefficients (Fox and Weisberg, 2011).

4.3.1. Cook's Distance

Cook's distance (Cook's *D*) is developed to measure the shift in **b** when a particular observation is excluded from the estimation. It is a measure of the contribution of that observation on all regression coefficients. To assess the degree of influence the i^{th} observation has on the coefficient estimate **b** in a linear model, a logical first step would be to obtain the least squares estimate of $\boldsymbol{\beta}$ with the i^{th} point deleted.

Let $\mathbf{b}_{(-i)}$ define the estimated value of the vector $\boldsymbol{\beta}$, with with the observation *i* omitted. Then the difference $(\mathbf{b}_{(-i)} - \mathbf{b})$ directly quantifies the influence of the i^{th} observation on the estimate of $\boldsymbol{\beta}$. If this difference is small, then the i^{th} observation does not significantly affect the estimates.

Formally, Cook's *D* (Cook, 2000) is found as

$$D_i = \frac{(\boldsymbol{b}_{(-i)} - \boldsymbol{b})^T (\boldsymbol{X}^T \boldsymbol{X})(\boldsymbol{b}_{(-i)} - \boldsymbol{b})}{p s^2(e)} \ , \tag{4.10}$$

where $s^2(e)$ is the residual variance obtained according to the equation (1.36).

Computationally, D_i is more easily found from the diagnostic statistics as

$$D_i = \frac{{e_{Si}}^2}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right), \tag{4.11}$$

where

h_{ii} is the hat-value for observation *i* from the equation (4.5),

${e_{Si}}^2$ is the squared standardized (internally studentised) residual, obtained from

$$e_{Si} = \frac{e_i}{s(e)\sqrt{1-h_{ii}}}. \tag{4.12}$$

Hence, in the formula (4.12), the first part may be regarded as a measure of remoteness, and the second as a measure of leverage corresponding to the point *i* (Fox and Weisberg, 2011).

D_i is large when the standardized residual is large, and the data point is located far from the centroid of the **X** space — that is, in case of large values of h_{ii} (Rawlings et al., 1998). The most commonly quoted criterion declares the i^{th} point as influential if D_i exceeds the median of the *F* distribution with *(p, n – p)* degrees of freedom, where *p* is the number of regression coefficients including the intercept (McDonald, 2002)

$$D_i > F_{0.5,\ p,\ n-p}. \tag{4.13}$$

If any noteworthy D_i is evident, it is reasonable to remove the respective case temporarily from the data, refit the regression model, and observe how the results change (Fox and Weisberg, 2011).

4.3.2. DFFITS

The deletion effect of i^{th} observation on the predicted (or fitted) value $\hat{y}_i$ can be investigated using diagnostic *DFFITS* (Belsley et al., 1980). It provides a measure of the shift in $\hat{\boldsymbol{y}}$ when the i^{th} observation is not used in the estimation of the population parameters $\boldsymbol{\beta}$.

Let $\hat{y}_i$ and $\hat{y}_{(-i)}$ be the predictions for the i^{th} observation with and without point *i* participating in the estimation of $\boldsymbol{\beta}$. Then the DFFITS for the i^{th} point is formulated as

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(-i)}}{s_{(-i)}(e)\sqrt{1 - h_{ii}}} \,, \tag{4.14}$$

where $s_{(-i)}(e)$ is the estimate of σ obtained without the i^{th} observation (Rawlings et al., 1998).

Alternatively, equation (4.14) can be written as

$$DFFITS_i = e_{Ji}\sqrt{\frac{h_{ii}}{1 - h_{ii}}} \,, \tag{4.15}$$

where e_{Ji} the externally studentized residual from the equation (4.8).

The $DFFITS_i$ can be large in any of the cases when e_{Ji} is large is the magnitude (the observation is an outlier) or when the h_{ii} is close to unity (the data point has high leverage). Therefore, $DFFITS_i$ is affected by both prediction error and leverage. The common suggestion is that any observation for which the following inequality holds

$$|DFFITS_i| > 2\sqrt{\frac{p}{n - p}} \tag{4.16}$$

can be used to indicate influential observations (Harrell, 2001).

Although the values resulting from solution of the equations (4.11) and (4.15) are different, Cook's distance and DFFITS are conceptually identical, and a closed-form formula for conversion of one value to the other can be analytically derived.

5. Variable Selection Procedures

In linear and generalized linear regression models, the "variable selection" (or "feature selection") means selecting which variables to include in the model. As such, it is a special case of model selection. Having a response variable y and a set of explanatory variables $\mathbf{X}$, the aim is to divide $\mathbf{X}$ into two groups – active and inactive predictors. Thereby, all essential information about y is contained in the active predictors, and the redundant predictor is eliminated.

The variable selection procedures follow two main objectives. First, a model should be as complete and realistic as possible, such that every regressor that somehow influences the dependent variable is included. Second, it is preferable to engage as few variables as possible since each irrelevant regressor weakens the precision of the estimated coefficients and fitted values. Also, the more variables are present in the model, more complex, costly, and time-consuming the data collection process becomes. Thus, it is necessary to achieve a balance between simplicity (as few regressors as possible) and fit (as many regressors as needed).

General model selection can concern not just finding the active predictors, but also building the model itself, including defining regressors for the predictors (i.e., transformations).

There are various types of variable selection procedures, such as, for instance, the best subset regression. This method compares all potential models that can be created based on a set of predictors and selects the subset of predictors that do the best at meeting some objective criterion, such as having the largest R^2 value or the smallest values of MSE (Mean Squared Error) and information criteria. The procedure fits $2k$ regression models, where k is the total number of predictors. It may be suitable in case of a low-dimensional data, while with a large number of explanatory variables the algorithm becomes time-consuming, effort-demanding, and at the end, it does not provide useful results.

A better alternative to the best subset regression is stepwise (or stagewise) variable selection – a family of methods based on adding or removing variables from a model sequentially. It includes three algorithms, which work in three directions: backward, forward, and forward-backward.

Two main differences from the best subset regression are that at each step:

- The procedure is not considering every single possible model that contains *k* predictors, but just the models that contain the *(k – 1)* predictors, which have already been chosen in the previous step.
- The goal is to select the variable that gives the most significant improvement to the model.

5.1. Backward Elimination

The starting point of the backward elimination is the full model estimated by the OLS containing all *k* predictors. Then iteratively the least useful predictor is removed, one-at-a-time, based on the threshold significance level α_{remove} (e.g., 5%). The algorithm is defined as follows:

Step 1: Start with a full model with all variables

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \ . \tag{5.1}$$

Step2: Remove the variable with the largest *p*-value of the corresponding two-sided *t*-test (that is, the least statistically significant variable), provided it exceed the threshold α_{remove}.

Step3: Re-estimate the model with *(k – 1)* predictors and return to Step 2.

The procedure continues until all explanatory variables remaining in the model have *p*-values less or equal to α_{remove}. Once the regressor has been excluded from the model, it cannot be entered back.

5.2. Forward Selection

Forward selection is the opposite of the backward elimination. It begins with the least squares model without any predictors, that if only with intercept β_0 (the mean over *y*), and then iteratively adds the most useful predictor, one-at-a-time, based on the threshold significance level α_{enter} (e.g., 5%). The procedure is then the following:

Step1: Start with a null model with no predictors

$$y_i = \beta_0 + \varepsilon_i \ . \tag{5.2}$$

Step2: Fit *k* simple linear regression models, each with one of the variables in and the intercept. All the single-variable models are considered to select the best one, that is, the one that results in the lowest *p*-value of the respective two-sided *t*-test. This variable is permanently fixed in the model.

Step3: Search through the remaining *(k – 1)* regression models with two variables (the first being fixed and the second being successively added) and determine the variable which should be added to the current model based on the lowest *p*-value.

The process continues until the lowest *p*-value exceeds the threshold α_{enter} (Bašta, 2018). Once the variable has been defined as significant and included in the model, it cannot be deleted.

5.3. Stepwise Regression

The stepwise (bidirectional) regression is a modification of the forward selection. After each step, whereby a variable was added, all candidate predictors in the model are investigated to test if their significance has been reduced below the specified tolerance level. If an insignificant variable is detected, it is excluded from the model. This procedure requires two significance levels: one for adding variables α_{enter}, and one for removing variables α_{remove}.

All three methods may be based not only on the significance level α, but also on some other criteria such as AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion), which reflect the relative quality of each model. Both criteria evaluate the likelihood of the model, understood as the probability of obtaining the data that we have, given the model being tested. Mathematically, the AIC and BIC can be derived as

$$AIC = -2\, l(\boldsymbol{b}) + 2k \ , \tag{5.3}$$

$$BIC = -2\, l(\boldsymbol{b}) + log(n)k \ , \tag{5.4}$$

where

$l(\boldsymbol{b})$ measures the improvement of the model fit when variables are added,

$2k$, $log(n)k$ are the penalizations for the number of parameters in the model.

The preference is then given to the regression model with the lowest criterion value.

The alternative ways to calculate AIC and BIC is to replace the $l(\boldsymbol{b})$ by $\log\left(\frac{SSR}{n}\right)$, where *SSR* is the residual sum of squares from the equation (1.25):

$$AIC = -2\log\left(\frac{SSR}{n}\right), +2k\,, \tag{5.5}$$

$$BIC = -2\log\left(\frac{SSR}{n}\right), +log(n)k\,. \tag{5.6}$$

The stepwise selection approach based on the information criteria lightly alters Step 2 of each procedure: the terms are dropped or added to the model until removal or inclusion of another term makes the criterion of interest worse, that is increasing it.

Automatic variable selection methods do not guarantee optimal results and should be used as a guide only. Moreover, these procedures do not take into account the researchers knowledge about the problem: it may add redundant predictors to the model or delete the important ones. Therefore, it may be necessary to force the algorithm to include the predictors of interest. It is essential to assess the research problem carefully, costs of the data collection, consistency, and meaningfulness of the results when determining the variables to be included in the regression model.

Practical Part

6. Data

The dataset used for the implementation of the regression model building process is presented in the appendix A1. It contains information on 183 countries and nine attributes for the year 2016. Not all of the currently existing countries have been considered due to the unavailability of the data for the majority of features. Such areas are primarily those with a low development level and small land area, which often fail to provide reliable information on the indicators. All explanatory variables can be assigned to one of the five topics: Economy, Demographics, Immunization, Nutrition, and Risk Factors.

6.1. Definition of Response and Explanatory Variables

Dependent variable

- *Life expectancy at birth (both genders)*: the average number of years newborns could expect to live, provided they follow current age- and gender-specific mortality conditions prevailing in a specific geographic area.

Explanatory variables

Economy

- *Income level:* income groups based on the gross national income (GNI) per capita.
- *GDP per capita:* a country's total economic output in the current 2016 US dollars divided by the total mid-year population.
- *Current health expenditure per capita:* average expenditure on health per capita converted to a US dollar and adjusted for the purchasing power of the national currencies using the economy-wide PPP (Purchasing Power Parity).

Demographics

- *Adult mortality rate:* the probability of dying in the interval from 15 to 60 years per hundred of population per year, not separated by gender.

Immunization

- *Hepatitis B (HepB3) immunization coverage:* the percentage of one-year-old children who have received three doses of a HepB3 vaccine in 2016.

Nutrition

- *Mean Body Mass Index (BMI):* a ratio of a person's weight in kilograms to the square of height in meters $BMI = \frac{Weight}{Height^2} = \frac{kg}{m^2}$.

Risk factors

- *Alcohol:* total per capita (15+ years) consumption in liters of pure alcohol.
- *Concentration of PM2.5:* the mean annual concentration of particulate matter of less than 2.5 microns of diameter ($PM_{2.5}$) in urban areas.

The following table summarizes the definition of the variables in the model and units of their measurement:

Variable	Unit of measurement
life_exp	Years
inc_level	Ordered factor with levels: Low < Lower-middle < Upper-middle < High
adult_m	Permille
HepB	Percent
BMI	kg/m^2 (kilogram per square meters)
alcohol	Litres
pm2.5	$\mu g/m^3$ (micrograms per cubic meter)
GDP	US dollars (per capita)
hlth_exp	US dollars (per capita)

Table 1: Definition of response and explanatory variables

6.2. Expected Influence on Response Variable

Economy

Currently, the World Bank (WB) divides all economies into four income groups based on the gross national income (GNI) per capita converted using Atlas method: low-income, lower-middle-income, upper-middle-income, and high-income countries. For the 2016 calendar year, the income level ranges were defined as follows:

Analytical classification	GNI per capita in USD
Low income	Below 1,005
Lower-middle income	1,006 - 3,955
Upper-middle income	3,956 - 12,235
High income	Above 12,235

Table 2: Country classification by income group (World Bank, n.d.)

The distributional characteristics of the life expectancy at birth by income group are presented in table 3:

	Income group			
Statistic	Low	Lower-middle	Upper-middle	High
Minimum	53.0	52.9	59.5	71.8
Lower quartile	59.8	65.4	71.9	77.2
Mean	61.4	69.5	75.1	81.2
Median	61.8	68.6	73.6	79.8
Upper quartile	64.1	72.6	76.4	82.4
Maximum	71.9	76.3	79.6	89.5
Standard deviation	4.4618	5.6104	4.2824	3.4327
Skewness	0.0335	-0.9313	-1.3561	-0.1632
Kurtosis	3.0565	3.5829	4.5022	2.9772

Table 3: Descriptive statistics of life expectancy by income level

The life expectancy at birth for both genders is normally distributed for the low-income countries, among which the Central African Republic has the lowest value of 53 years, whereas people in North Korea expect to live on average 71.9 years (see figure 2). The distribution of the average number of years to be lived of by the lower-middle and upper-middle income countries is negatively skewed, that is, there are several countries, particularly from the African region, with low life expectancy compared to others. From the countries with high disposable income per capita, the Southern American Trinidad and Tobago represents the minimum of 71.8 years, whereas an absolute record holder is Monaco, one of the wealthiest countries in the world, where Monacoians expect to live 89.5 on average.

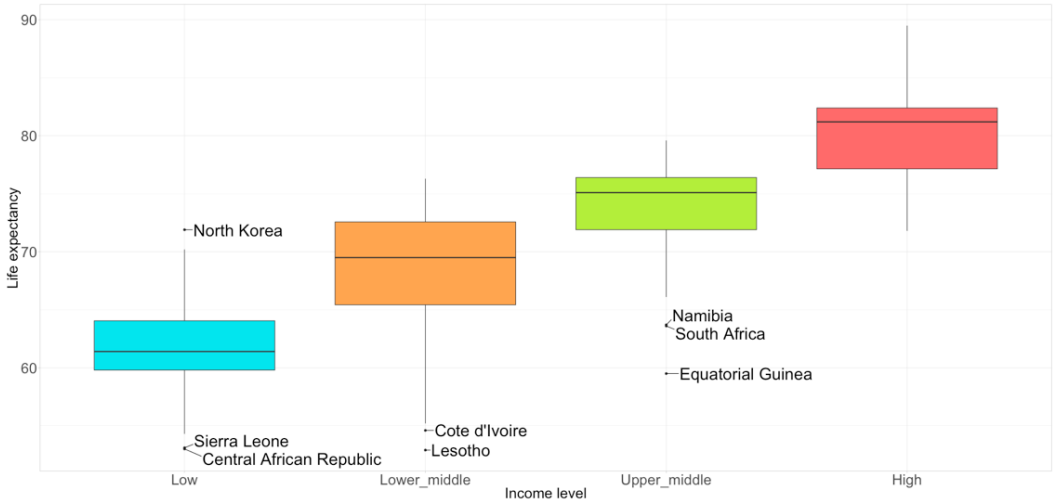

Figure 2: Distribution of life expectancy at birth by income level

GDP per capita allows taking an insight into the inhabitants' prosperity. There is a non-linear relationship between the GDP per capita and life expectancy since people fulfill both their needs and wants. In order to survive, people need to satisfy their basic needs, such as food, clothing, shelter, healthcare, education. Once those needs are met, people can spend the remaining money on the non-necessities.

Higher income suggests better food supply, access to housing, medical and educational facilities, and other factors that improve the quality of life, reduce mortality rates and, consequently, increase overall life expectancy. However, at a certain point, the relationship between life expectancy and income starts to weaken: if everyone satisfied the needs, any further growth in the GDP per capita would hardly affect the lifespan. The developed high-income countries serve as evidence of this phenomenon since higher disposable income goes in strong association with an unhealthy lifestyle, such as alcohol and tobacco consumption, which in turn negatively affects people's health.

Current health expenditures cover private sources spending on personal health care (curative, rehabilitative and long-term care, medical goods and ancillary services) and public sources spending on collective services (public health services, prevention programs, and health administration) (OECD, 2017). In line with the GDP per capita, the rising expenditures on healthcare in the population increase the expected lifespan of the population up to a certain point. Thus, it is reasonable to assume that the number of years one could expect to live non-linearly depends on the health services spendings (see figure 3).

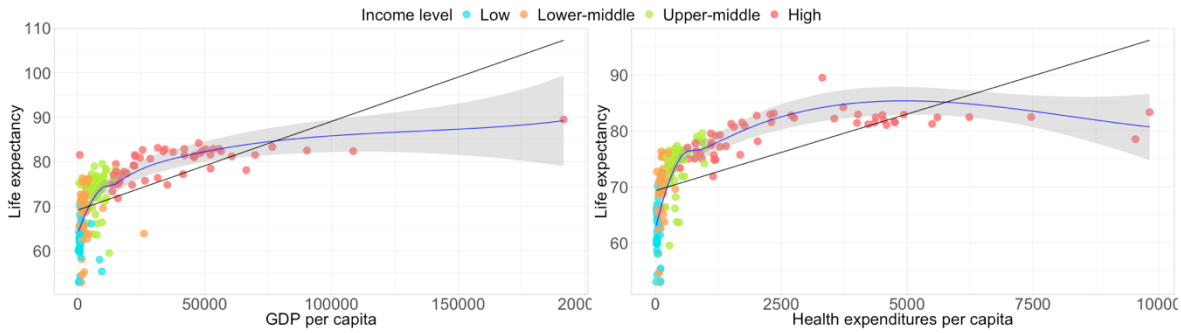

Figure 3: Scatterplots of life expectancy by GDP per capita (left) and by health expenditures per capita (right)

Demographics

The level of adult mortality is essential for evaluation of the population mortality pattern. Due to health transitions and population aging the incidence of the non-communicable

diseases among adults from 15 to 60 years old, the most productive group, is rapidly rising in developing countries. The adult mortality depicts the socio-economic, environmental, and health conditions in which the population lives; thus, it helps to identify population vulnerability. It is reasonable to assume that there is a strong negative linear dependence of the life expectancy on the adult mortality rate.

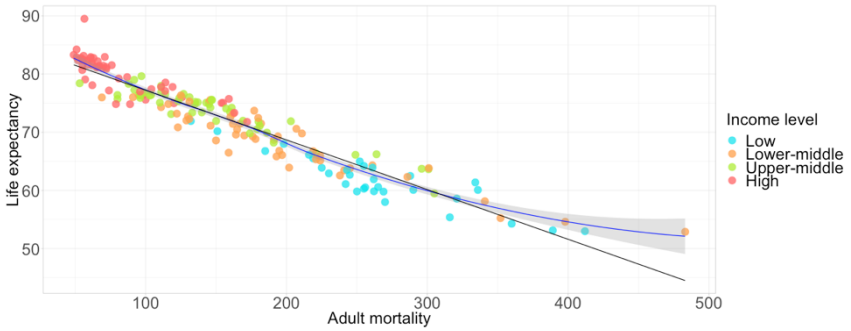

Figure 4: Scatterplot of life expectancy by adult mortality rate

Immunization

Immunization is a proven and powerful tool for controlling and eradicating life-threatening infectious diseases, being vital for reducing population mortality. Vaccination, primarily provided to the children in their first year of life, stimulate the body resistance that lasts for at least 20 years and may be lifelong, which prevents infection and a potential threat to a person's life such as the development of chronic disease, cirrhosis, and liver cancer due to hepatitis B. Hence, broader health service coverage implies rising life expectancy of the population.

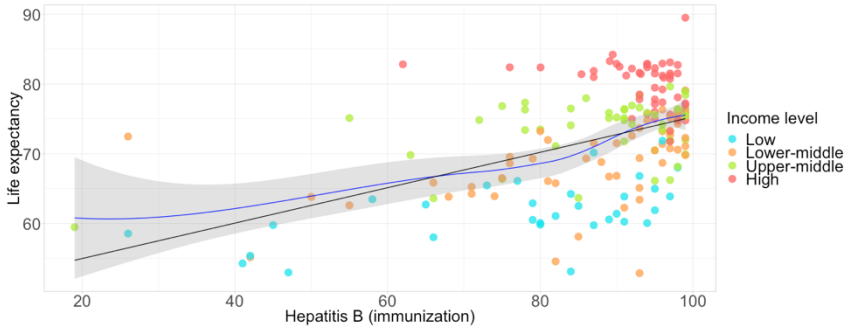

Figure 5: Scatterplot of life expectancy by hepatitis B immunization coverage

Nutrition

Body Mass Index (BMI) reflects the nutritional status of an adult and is a frequently used tool to associate the weight with the risk of various health problems. The BMI could serve as a good proxy for the investigation of the weight-related problems at the population level. The WHO classifies the adults over 20 years into the following categories:

Nutritional status	BMI
Underweight	Below 18.5
Normal weight	18.5–24.9
Pre-obesity	25.0–29.9
Obesity classes I, II, III	Above 30.0

Table 4: Classification of nutritional status in adults by BMI, WHO (2019)

A BMI falling below 18.5 signals insufficient weight, which may be a result of starvation, malnutrition, eating dysfunctions, or illnesses. On the contrary, a BMI greater than 25 and 30 is regarded as overweight and obese, respectively, to which some common conditions may relate, such as high blood pressure, cardiovascular diseases, diabetes and other. This knowledge, supported by the scatterplot in figure 6 allows concluding that there is some curvilinear (quadratic or cubic) form of dependence on the average BMI.

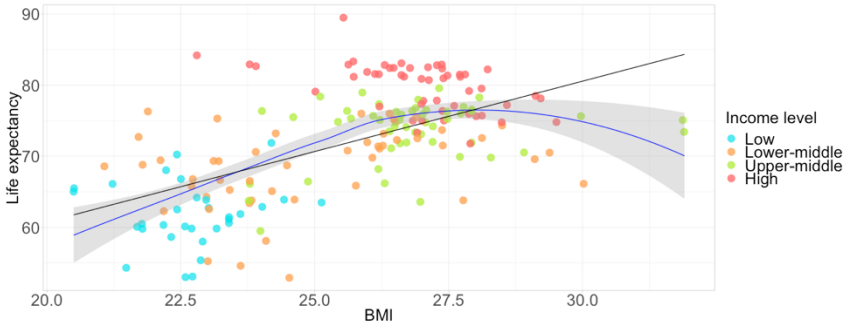

Figure 6: Scatterplot of life expectancy by BMI

Risk factors

Alcohol is a psychoactive beverage with addiction-producing features that many cultures widely used for centuries. Harmful alcohol consumption is associated with significant disease and socio-economic strain in societies. Drinking alcohol impacts health condition resulting in the incidence of alcohol intoxication, illnesses, behavioral and mental disorders. Figure 7 shows that the high-income developed countries tend to consume more alcoholic beverages, mainly due to affordability or cultural specifics.

A standard measure of air pollution is the concentration of the fine atmospheric particles $PM_{2.5}$ with a size of less than 2.5 micrometers (2.5×10^{-6} meters) contained in the industrial and automobile emissions, waste incineration, coal burning, and inorganic aerosols. Due to their small size, $PM_{2.5}$ penetrate deeply into the respiratory tract, and hence are dangerous for health and increase age-specific mortality for respiratory and cardiovascular diseases. According to the Air Quality Index (AQI), the 24-hour average concentration of the 2.5-type particles in the air should not exceed 12.0 µg/m3 (see table 5).

Air Quality Index (AQI)	$PM_{2.5}$ in µg/m3 (based on a 24-hour average)
Good	0 – 12.0
Moderate	12.1 – 35.4
Unhealthy for sensitive groups	35.5 – 55.4
Unhealthy	55.5 – 150.4
Very unhealthy	150.5 – 250.4
Hazardous	Above 250.5

Table 5: Air Quality Index based on 24-hour average concentration of fine particulate matter ($PM_{2.5}$) in the air, EPA (2013)

In developing countries, the air pollution is responsible for a high proportion of the burden of the diseases, however, in some high-income countries such as Qatar and Saudi Arabia, the concentrations of $PM_{2.5}$ are close to 100 µg/m3. The LOESS (locally estimated scatterplot smoothing) in figure 7 suggests that life expectancy relates to the amount of $PM_{2.5}$ in the air in a non-linear fashion due to the presence of several heavily polluted areas with high life expectancy, which are remote from the rest data points in the two-dimensional plane.

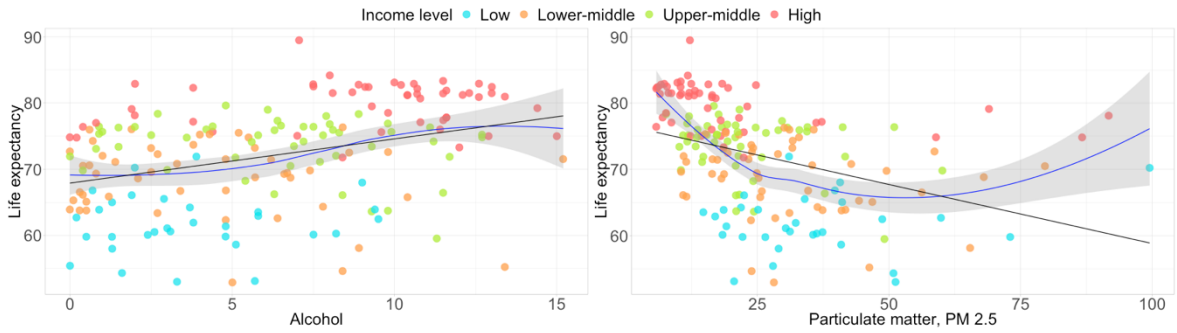

Figure 7: Scatterplots of life expectancy by alcohol consumption (left) and concentration of particulate matter $PM_{2.5}$ (right)

6.3. Missing Data

- **Missing data pattern**

A missing data pattern defines the configuration of observed and missing values in a data set. It depicts the location of the gaps in the data; however, it does not explain why the data are missing. In the investigated dataset, the missing data follows a general pattern, that is, unavailable values are scattered throughout the data matrix in a haphazard fashion. The missing values represent only 1.5% of the total data points (that is, out of $183 \times 9 = 1647$).

From the standpoint of observations, 90% of instances are fully observed, and only 10% of observations contain missing values in one, or in a combination of up to three variables. *HepB* is the variable that has the highest proportion of missingness, that is 5% of the values for that variable, which corresponds to 9 unavailable values. The rest five variables (*GDP, hlth_exp, BMI, alcohol, and adult_m*) contain from 1 to 7 missing values.

- **Multiple Imputation**

Imputation is one of the key procedures that researchers apply to fill in missing data in a dataset. There are different calculation techniques whereby the missing values are replaced with the most probable estimate allowing to conduct more accurate and reliable analyses. To ensure the correctness of the imputation procedure, one needs to asses the mechanism of missingness as a primary step. The missing data mechanisms describe potential associations between measured variables and the probability of missing data.

For this particular set of variables, the data are missing not at random (MNAR), since some countries tend to provide insufficient information. However, there is only 1% of data values unavailable, so we will neglect the nature of missingness and impute the values using Miss Forest approach to be able to run the regression analysis on the full data.

The Miss Forest approach, developed by Stekhoven and Buhlmann (2013) and specified in the {missForest} package, is frequently used to impute missing values, especially in the case of mixed-type data. It is a machine learning algorithm, which fits a random forest model using the available non-missing data to predict the missing data until convergence is attained iteratively. After each iteration, the difference between the preceding and the new imputed data matrix is evaluated for both the continuous and categorical parts. The process stops as soon as both differences go up for the first time.

Comparison of the descriptive statistics before and after the imputation justifies that full distribution of the variables with missing and imputed values do not significantly differ.

The most notable discrepancy is observed for the variable *GDP,* which stems from the high variability of the income across countries.

Statistic	Min	Lower quartile	Median	Upper quartile	Max	SD
adult_m	49	96.0	146.5	218.3	483	85.706
HepB	19	82.0	93.0	97	99	15.354
BMI	20.5	23.8	26.3	27.3	31.9	2.277
alcohol	0	2.4	6.1	9.3	15.2	4.108
pm2.5	5.8	14.7	21.5	33.0	99.5	16.911
GDP	219.2	1,716.3	5,787.3	15,265.9	191,586.6	23,157.240
hlth_exp	16.6	86.3	301.2	958.5	9,818.0	1,693.895

Table 6: Descriptive statistics of data containig missing values

Statistic	Min	Lower quartile	Median	Upper quartile	Max	SD
adult_m	49.0	96.0	146.0	217.5	483.0	85.825
HepB	19.0	82.0	93.0	97.0	99.0	15.022
BMI	20.5	23.8	26.2	27.3	32.0	2.278
alcohol	0	2.5	6.1	9.3	15.2	4.087
pm2.5	5.8	14.7	21.5	33.0	99.5	16.911
GDP	219.2	1,697.4	5,382.8	14,977.9	191,586.6	22,801.460
hlth_exp	16.6	91.0	295.4	936.5	9,818.0	1,679.517

Table 7: Descriptive statistics of data after multiple imputation

Although the missing data requires a more detailed and sophisticated approach, for this moment we will consider the result as acceptable.

7. Least Squares Estimation

7.1. Model Specification

According to the theoretical considerations described in section 6.2, the regression model which will be investigated in the subsequent chapters can be specified as

$$\begin{aligned} life\_exp = {} & \beta_0 + \beta_1\, Low + \beta_2\, Lower\text{-}middle + \beta_3\, Upper\text{-}middle \\ & + \beta_4\, adult\_m + \beta_5\, HepB + \beta_6\, BMI + \beta_7\, BMI^2 \\ & + \beta_8\, alcohol + \beta_9\, alcohol^2 + \beta_{10}\, log(pm2.5) \\ & + \beta_{11}\, GDP + \beta_{12}\, GDP^2 \\ & + \beta_{13}\, hlth\_exp + \beta_{14}\, hlth\_exp^2 + \varepsilon, \end{aligned} \tag{7.1}$$

where

Low, Lower-middle, Upper-middle are the levels corresponding to the *inc_level* indicator variable with *High* being the reference category,

$\beta_1, \beta_2, \beta_3$ express the expected difference in the life expectancy between countries with low, lower-middle or upper-middle income level and the high-income countries.

Four variables (*BMI, alcohol, GDP,* and *hlth_exp*) enter this equation as the second-degree polynomials, due to the anticipated non-linear impact they have on the dependent variable *life_exp*. For better interpretability, the variable *pm2.5* is included in the model as a logarithmic transformation. As a matter of convenience, the variables *GDP* and *hlth_exp* are intentionally divided by 1000 prior to the analysis, such that the resulting coefficients reflect a 1000 US dollars change in the income and health expenditures per capita.

Despite the fact that the response variable *life_exp* is negatively skewed (coefficient of skewness = - 0.473) and slightly platykurtic (coefficient of kurtosis = 2.56) compared to the normal distribution, it enters the model without any transformation. The reason is that taking the logarithm of *life_exp*, while having a practical interpretation, makes the distribution even more skewed (see figure 8). Some other power transformations may be applied, such as Box-Cox or Ordered Quantile normalizing in order to bring the distribution of the response closer to normal. However, such transformations complicate the interpretation of the regression coefficients and prediction of the response, since it is uncertain how to transform the predicted values back to the original units.

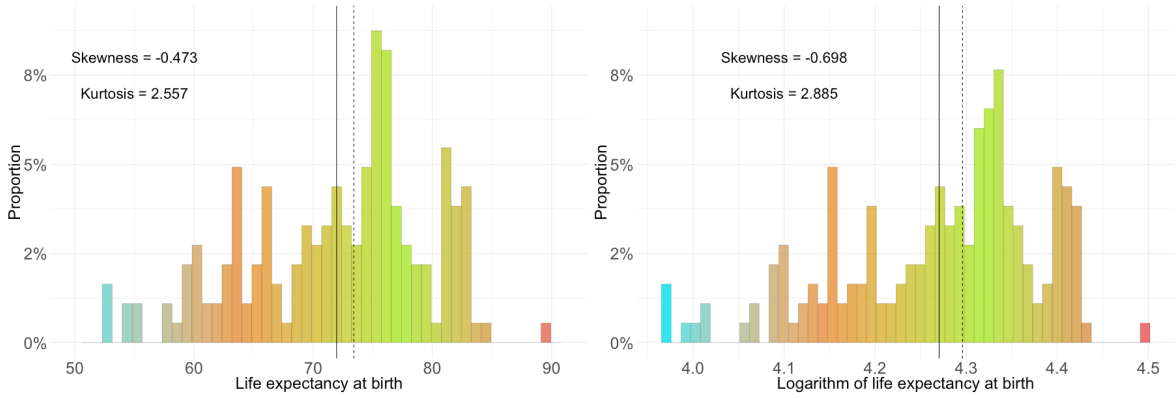

Figure 8: Histograms of life expectancy (left) and logarithm of life expectancy (right)

7.2. Ordinary Least Squares Estimation

Assuming the relationships between the life expectancy at birth and other factors discussed in section 6.1, the equation (7.1) was estimated using the OLS, whereby the variables *BMI, alcohol, GDP,* and *hlth_exp* were represented by the second-order raw (ordinary) polynomials. That is, each of these variables enters the regression model as the first and the second power. Intuitively, since X^2 is a vector whose elements are the squares of the respective elements of the vector X, they become almost perfectly linearly dependent. This may lead to a multicollinearity problem in a multiple regression, which in turn reduces the accuracy of the estimation. In the case of perfect multicollinearity, that is when an exact linear relationship exists among several variables, the matrix $(\mathbf{X}^{\mathbf{T}}\mathbf{X})$ is computationally singular, and the parameters $\boldsymbol{\beta}$ cannot be estimated by the OLS. To address this problem, the orthogonal polynomials may be used in place of the ordinary polynomials.

Orthogonalization is an attractive technique for fitting curve to the data by the method of the least squares because it provides coefficients which are statistically independent.

Figure 9 depicts the correlation between particular variables and their squares. Application of the raw polynomial fitting (correlogram to the left) leads to a strong positive linear relationship within each polynomial term with the Pearson correlation coefficients being higher than 0.86. However, implementation of the orthogonal polynomials (correlogram to the right) reduced the linear dependence to zero in each case.

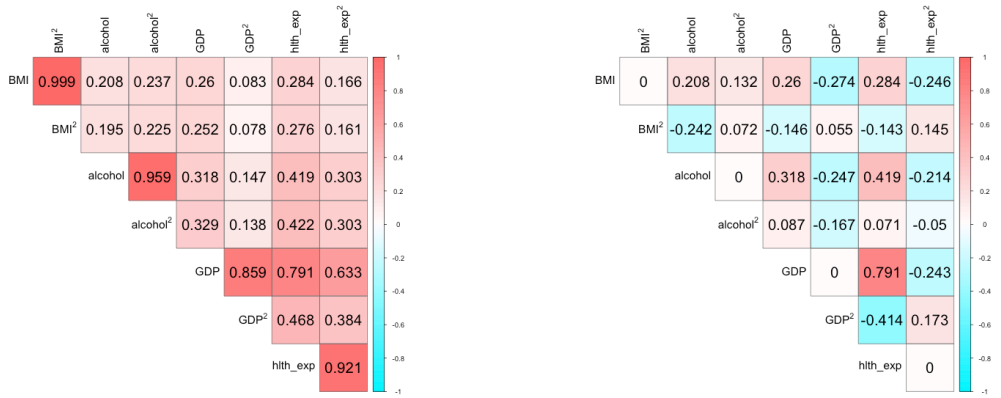

Figure 9: Pairwise Pearson correlation coefficients of ordinary (left) and orthogonal (right) polynomial regressors

The Variance Inflation Factor (VIF) is requently used to measure how much variance of the regression coefficient estimate is inflated in the presence of multicollinearity. Mathematically, the Variance Inflation Factor for the j^{th} predictor it is expressed as

$$VIF_j = \frac{1}{1 - R_j^2}\ , \tag{7.2}$$

where R_j^2 is the coefficient of determination obtained by regressing the j^{th} predictor on all the other explanatory variables. Typically, if VIF exceeds the value of 10, the multicollinearity is considered to be high.

Table 8 represents the values of VIF for the models fitted using raw and orthogonal polynomials. In the first case, the VIF for the *BMI* was 440.35 ($\sqrt{440.35} \approx 20.98$), which means that the standard error for its coefficient was approximately 21 times as large as it would be in case these variables were not correlated with other predictors. Orthogonalization considerably reduced the value of VIF from 440.35 to 2.29 for the linear term of BMI. The largest VIF corresponds to the regressor *hlth_exp* is equal to 10.79, meaning that the standard error for the coefficient b_{hlth_exp} is 3.3 times higher than in case of no collinearity. However, since this value does not dramatically exceed the threshold, the orthogonalized model will be considered as acceptable for the implementation of further analysis. A drawback of any polynomial regression is that the coefficients from such models are difficult to interpret since the ceteris paribus condition does not hold any longer.

	Ordinary	Orthogonal
Low	6.6967	6.6967
Lower-middle	6.184	6.184
Upper-middle	4.2116	4.2116
adult_m	2.9816	2.9816
HepB	1.446	1.446
BMI	440.3532	2.2928
BMI^2	424.8769	1.2181
alcohol	15.2029	1.6519
$alcohol^2$	14.5139	1.1787
log(pm2.5)	1.784	1.784
GDP	33.4243	8.7242
GDP^2	13.5236	3.536
hlth_exp	41.5075	10.7903
$hlth\_exp^2$	16.5968	2.5234

Table 8: Variance Inflation Factors (VIF) for ordinary and orthogonal polynomial regressors

Unfortunately, many studies applying an orthogonal polynomial regression fitting procedure do not offer sufficient guidance for interpreting the coefficients estimates. One of the most straightforward approaches is to follow the signs of the higher order terms in order to determine convexity or concavity of the function and compare them with the expectations on the functional relationship.

The following equation depicts the estimated regression function obtained by the OLS method using orthogonalized polynomials:

$$\begin{aligned} life\_exp = {} & 83.32 \\ & - 3.06\, Low - 1.43\, Lower\text{-}middle - 0.47\, Upper\text{-}middle \\ & - 0.07\, adult\_m + 0.02\, HepB + 3.15\, BMI - 1.15\, BMI^2 \\ & + 5.68\, alcohol - 4.05\, alcohol^2 - 0.59\, log(pm2.5) \\ & - 4.68\, GDP + 10.54\, GDP^2 \\ & + 17.03\, hlth\_exp - 8.03\, hlth\_exp^2 + e. \end{aligned} \tag{7.3}$$

Each of the OLS slope coefficients except has the anticipated sign except for the GDP. In the countries with low- and middle-income level, the expected length of life is lower than in the areas with the GNI per capita above 12,235 USD. That is, transition to higher classes due to socio-economic advancement should potentially increase the number of years a child born in a particular territory could expect to live. Adult mortality and air pollution both have an adverse effect on longevity. The negative signs corresponding to the

quadratic terms of the variables *BMI, alcohol,* and *hlth_exp* reflect the anticipated concave form of relationship with the dependent variable. In consistency with the expectations made in chapter 6, the increasing body mass is associated with the rising life expectancy up to some point, at which the direction of the relationship changes. Similarly, positive relationship between life expectancy and health expenditures per capita gradually weakens, and additional spendings do not contribute to large changes in the response.

The analysis of variance from table 9 indicated that eight predictors and their transformations explained 96.3% of the variation in the *life_exp* ($R^2_{adj} = 0.9629$). The squared differences between the OLS fitted values and the mean life expectancy lead to a considerably high *F*-statistic for testing the joint significance of the regression coefficients. However, prior to the interpretation of the coefficients estimates and making inferential statements regarding the individual coefficients, the residual diagnostics need to be conducted in order to detect possible assumptions violation.

Source	Sum of Squares	Degrees of Freedom	Mean Square	*F*-statistic	*p*-value
Regression	10347.52	14	739.1086	338.6833	< 2.2e-16
Residual	366.63	168	2.1823	-	-
Total	10714.15	183	-	-	-
Residual standard error:			1.477		
Multiple R-squared:			0.9658		
Adjusted R-squared:			0.9629		

Table 9: Analysis of variance (model estimated by OLS)

- **RESET Test**

The RESET statistic for the equation (7.3) turns out to be 4.41, this is the value of an $F_{2, 166}$ random variable ($n = 183$, $k = 14$), and the associated *p*-value is 0.014 (table 10). Thus, we do not reject the null hypothesis and assume that at 1% significance level, the functional form is correctly specified. However, at 5% and 10% significance level, the null hypothesis can be rejected in favor of the misspecification.

F-test			
Degrees of freedom			2, 166
Test statistic			4.4133
Critical value	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
	4.7353	3.0505	2.3348
p-value			0.0136

Table 10: RESET test (model estimated by OLS)

- **Normality Tests**

All the normality tests described in section 3.3 advocate the normal distribution of the error term in the regression model being considered, since the respective *p*-values noticeably exceed any reasonable significance level.

Test	Statistic	*p*-value
Shapiro-Wilk	0.9919	0.4001
Lilliefors	0.0323	0.9134
Cramér-von Mises	0.0330	0.8018
Anderson-Darling	0.2660	0.6877

Table 11: Normality tests (model estimated by OLS)

The normal quantile-quantile plot in figure 10 confirms the result, as the distribution of the residuals is almost symmetrical around zero and the points in the lower and upper quartiles do not suggest either skewness or heavy tails.

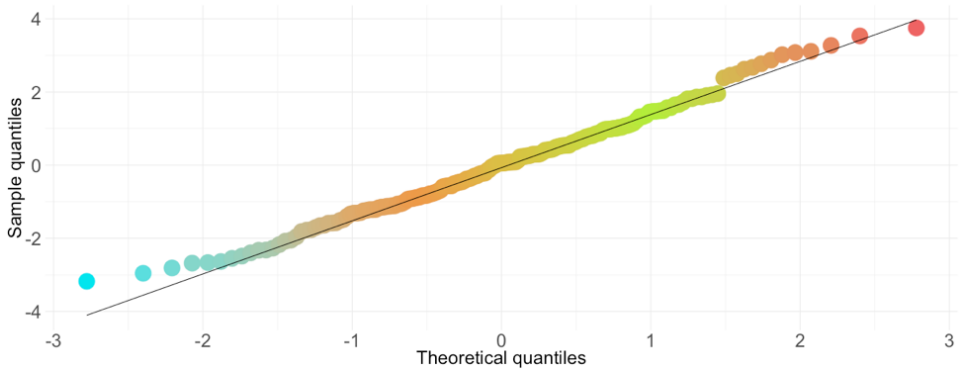

Figure 10: Quantile-comparison plot of ordinary residuals

- **Heteroskedasticity Tests**

Since the assumption of normality is satisfied, it is crucial to evaluate whether the variance of the error term is constant. A plot of the residuals versus the fitted values (figure 11) indicates possible nonconstant error variance. Low and lower-middle income countries (mostly African region) contribute to the higher spread of the residuals at lower values of the life expectancy, whereas countries from the upper-middle income group are more concentrated around zero.

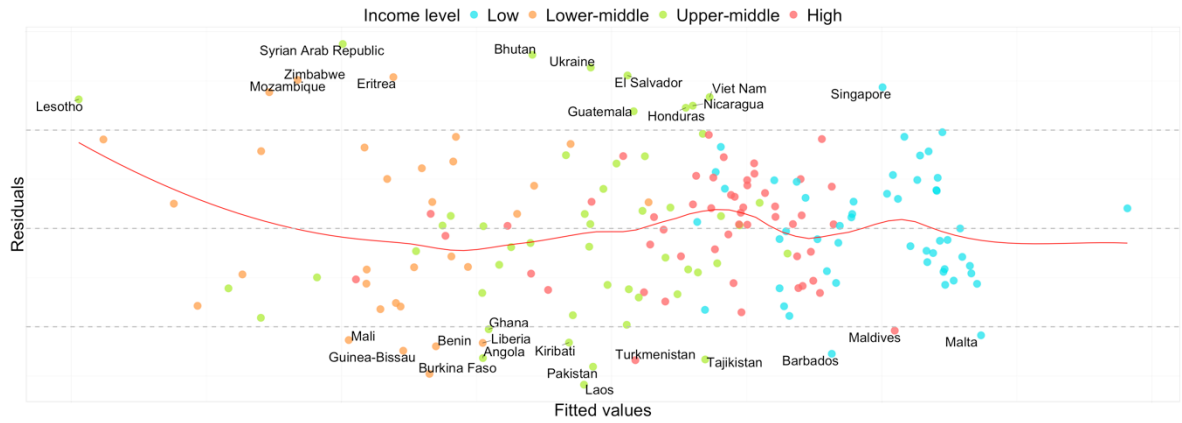

Figure 11: Scatterplot of ordinary residuals against fitted values

The Breusch-Pagan test rejects the null hypothesis of homoskedasticity at 10% and 5% significance levels, $\chi^2 = 28.443$, $p = 0.0124$, but fails to do so at 1% significance level. The special case of the White test, consisting of regressing residuals e^2 on the fitted values and $\widehat{life\_exp}$ and $\widehat{life\_exp}^2$, produces $R^2_{e^2} = 0.0852$; thus, $LM = 183 \times 0.0852 \approx 15.5871$, and the corresponding p-value < 0.001. This is another evidence of heteroskedasticity, which requires implementation of some techniques, such as the Weighted Least Squares.

	Breusch-Pagan	Modified White	
	LM test	*F*-test	*LM* test
DF	14	2, 180	2
Test statistic	28.4430	8.3795	15.5871
Critical value $\alpha = 0.05$	23.6850	3.0462	5.9915
p-value	0.0124	0.00033	0.00041

Table 12: Heteroskedasticity tests (model estimated by OLS)

7.3. Feasible Weighted Least Squares Estimation

In the OLS, all observations contribute equally to the parameter estimation, and each individual data point has a weight of one. Since there is strong evidence of heteroskedasticity, we need to estimate the model using the FWLS procedure. The logged squared residuals $\log(e_i^2)$, obtained from the estimation of the model (7.1) by OLS, were regressed on the fourteen predictors from that initial model. The estimates of the diagonal values $\widehat{w}_i$ of the matrix **W** were obtained through exponentiation of the resulting fitted values $\widehat{\log(e_i^2)}$, and the corresponding diagonal values of the weighting matrix **P** were found as $\frac{1}{\sqrt{\widehat{w}_i}}$.

The weights for each of 183 observations (diagonal elements of the weighting matrix **P** from equation (1.65)) have been estimated based on the equations (1.50-1.51) and (1.55). The full list of weights can be found in the appendix A2. Figure 12 depicts the distribution of weights assigned to the observations and grouped by the level of disposable income. In accordance with figure 12, the low-income countries have been given the lowest weights, all less than one, with an average of 0.7673. Since these underdeveloped areas are the main cause of high residual variance, their impact on the parameter estimates will be reduced after the weighting procedure. On the contrary, more weight is given to the upper-middle income areas, which may be thought to be more reliable source of information. The mean and median weights corresponding to the high-income group are less than that of the upper-middle income areas. The maximum weight of 3.374 is assigned to the USA, making its impact on the FWLS estimation procedure by 3.4 times larger than it had in the OLS.

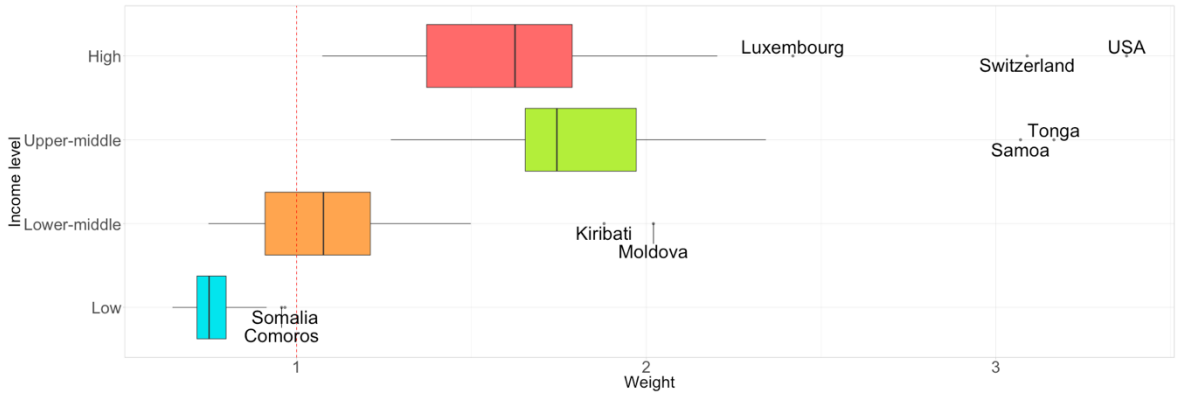

Figure 12: Distribution of estimated weights of observations by income group

	Mean weight	Median weight
High	1.6637	1.6246
Upper-middle	1.8275	1.7444
Lower-middle	1.1008	1.0767
Low	0.7673	0.7496

Table 13: Mean and median weights of observations by income group

After the data has been transformed using the computed weights, the regression coefficients were re-estimated by the OLS. The estimation outcome for the model adjusted for heteroskedasticity is as follows:

	Estimate	Std. Error	*t*-statistic	*p*-value
Non-constant	84.3125	1.1201	75.2700	< 2.2e-16
Low	-3.1896	0.7337	-4.3471	0.00002
Lower-middle	-1.6379	0.4861	-3.3695	0.00093
Upper-middle	-0.6706	0.3742	-1.7920	0.0749
adult_m	-0.0674	0.0022	-31.099	< 2.2e-16
HepB	0.0157	0.008	1.9695	0.0505
BMI	3.6234	2.3512	1.5411	0.1252
BMI^2	-2.3160	1.3044	-1.7755	0.0776
alcohol	7.6404	1.5120	5.0532	<0.0005
$alcohol^2$	-4.2545	1.2624	-3.3701	0.00093
log(pm2.5)	-0.5393	0.2043	-2.6396	0.0091
GDP	-4.4266	3.3612	-1.3170	0.1896
GDP^2	10.7095	2.1725	4.9295	<0.0005
hlth_exp	13.8027	3.7606	3.6704	0.00033
$hlth\_exp^2$	-5.2708	1.5581	-3.3829	0.00089

Table 14: Summary of regression model estimated by FWLS

Life expectancy at birth in the upper-middle-income countries is expected to be by 0.67 years lower compared to the high-income countries, according to the Feasible Weighted Least Squares. The lower-middle and low-income categories membership has an even larger negative effect on the number of years a newborn can expect to live on average, reducing the life expectancy by 1.64 and 3.19 years, respectively. The increase in adult mortality rate by ten permille (10 deaths per thousand of the population), holding other factors fixed, reduces the life expectancy by 0.67 years (eight months) on average. A 10% increase in the Hepatitis B immunization coverage among population prolongs the expected lifespan by 0.16 years on average. For a 10% increase in the $PM_{2.5}$ concentration in the air, the expected partial effect is a reduction of the life length by 0.05 years $(-0.5393 \times \log\left(\frac{110}{100}\right) = -0.0514)$.

- **RESET Test**

The RESET statistic for the model (7.1) estimated by the FWLS is equal to 0.811, which is sufficiently smaller than critical values for 1%, 5%, and 10% significance levels. Thus, the test statistic does not fall into the rejection region, and with the associated p-value of 0.446, we fail to reject the null hypothesis of a correct functional form specification.

F-test			
Degrees of freedom			2, 166
Test statistic			0.8113
Critical value	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
	4.7353	3.0505	2.3348
p-value			0.4460

Table 15: RESET test (model estimated by FWLS)

- **Normality Tests**

Test	Statistic	*p*-value
Shapiro-Wilk	0.9954	0.8497
Lilliefors	0.0352	0.8386
Cramér-von Mises	0.0248	0.9111
Anderson-Darling	0.1647	0.9409

Table 16: Normality tests (model estimated by FWLS)

All normality tests strongly advocate normality of the residuals with the *p*-values being even higher than in case of the OLS, almost approaching the value of one.

- **Heteroskedasticity Tests**

	Breusch-Pagan	Modified White	
	LM test	*F*-test	*LM* test
DF	14	2, 180	2
Test statistic	13.381	0.9507	1.9129
Critical value $\alpha = 0.05$	23.6850	3.0462	5.9915
p-value	0.4968	0.3884	0.3843

Table 17: Heteroskedasticity tests (model estimated by FWLS)

Both the Breusch-Pagan and modified White heteroskedasticity tests fail to reject the null hypothesis of a constant error variance at any significance level. For the modified White test, the R-squared produced by regressing the FWLS residuals u^2 on the fitted values and $l\imath\widehat{fe\_exp}^*$ and $l\imath\widehat{fe\_exp}^{2*}$ is $R^2_{u^2} = 0.0105$. That is a function of explanatory variables can explain only 1% of the residual variation. Failure to reject the homoskedasticity hypothesis signify that estimation of the weights and applying OLS on the weighted data solved the problem of the non-constant error variance. Thus, the usual inferential procedures can be conducted based on the FWLS estimates. Consequently, the final model which will serve as a basis for further analysis is expressed as

$$
\begin{aligned}
life\_exp^{*} = {} & 84.31\ Non\text{-}constant^{*} \\
& - 3.19\ Low^{*} - 1.64\ Lower\text{-}middle^{*} - 0.67\ Upper\text{-}middle^{*} \\
& - 0.07\ adult\_m^{*} + 0.02\ HepB^{*} + 3.62\ BMI^{*} - 2.32\ BMI^{2*} \\
& + 7.64\ alcohol^{*} - 4.25\ alcohol^{2*} - 0.54\ log(pm2.5)^{*} \\
& - 4.43\ GDP^{*} + 10.71\ GDP^{2*} \\
& + 13.80\ hlth\_exp^{*} - 5.27\ hlth\_exp^{2*} + e^{*} ,
\end{aligned}
\tag{7.4}
$$

where *Non-constant* term equals to the observation-specific values $\frac{1}{\sqrt{\widehat{w}_i}}$.

As a matter of convenience, the asterisk sign (*) corresponding to the transformed variables will be omitted from notation of the variables hereinafter.

It is necessary to remember, that according to the equation (1.43) the model estimated by the FWLS does not contain a constant term any longer since the weights of individual observations replace the column of ones in the model matrix **Q**. Therefore, the explained sum of squares (SSE) is represented as the sum of squared fitted values

$$SSE_{FWLS} = \sum_{i=1}^{n} \hat{q}_i^2 \ , \tag{7.5}$$

and the total sum of squares (SST) is no longer the sum of squared deviations from the mean, by merely the squared observed values of the transformed response q

$$SST_{WLS} = \sum_{i=1}^{n} q_i^2 \ . \tag{7.6}$$

Redefining the SSE and SST results in a highly inflated F-statistic for the FWLS model ($F_{15,\,168}$ = 46,827.35), which suggests rejection of the null hypothesis

$$H_0 : \beta_0 = \cdots = \beta_{14} = 0 \ . \tag{7.7}$$

in favor of the statement that at least one of the regression coefficients is non-zero (table 18). Thus, the whole model is assumed to be statistically significant at any level α, and individual t-statistics might be studied in order to judge the statistical relevance of the regression coefficients.

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-statistic	p-value
Regression	2,307,571.00	15	153,838.10	46,827.35	< 2.2e-16
Residual	551.92	168	3.29	-	-
Total	2,308,122.92	183	-	-	-
Residual standard error:			1.8125		
Multiple R-squared:			0.99976		
Adjusted R-squared:			0.99974		

Table 18: Analysis of variance (model estimated by FWLS)

According to the summary table 14, almost all explanatory variables have ceteris paribus statistically significant impact on the response variable. However, for certain regressors, the rejection of the null hypothesis depends on the significance level. Thereby, regressors *HepB* (t_{HepB}=1.9695, p = 0.0505) and BMI^2 (t_{BMI^2}= −1.7755 , p = 0.0776) are considered to be statistically significant at 10% level, but at 1% and at 5% significance levels we fail to reject the null hypothesis of the corresponding population coefficients being equal to zero. The *t*-test for the upper-middle income category suggests statistical significance at the level of 10% or higher ($t_{Upper-middle}$= −1.7920, p = 0.0749). Similarly, the logarithm of the $PM_{2.5}$ concentration ($t_{log(pm2.5)}$ = −2.6396, p = 0.0091) is thought to be insignificant at 1% significance level, but at 5% level it is assumed to be different from zero. Despite the fact that the *t*-test defined *BMI* and *GDP* as statistically indistinguishable from zero, they remain in the model since corresponding higher order (quadratic) terms exhibit statistical significance.

7.4. Confidence Intervals

Due to the computational specifics of the SSE and SST in case of the Weighted Least Squares, the R^2 does not provide a reliable basis for comparing the OLS and FWLS estimated models. For that purpose, better measures appear to be the standard errors and confidence intervals of the OLS and FWLS regressions.

The OLS and FWLS estimates are not expected to be identical, though, the difference between them is not large. Almost all the FWLS coefficients have smaller standard errors than the OLS. Therefore, the confidence intervals derived by the FWLS are mostly narrower than that of the OLS, except for the linear term *BMI*.

The following table provides the 90% univariate and simultaneous (Bonferroni corrected) confidence intervals for the regression parameters:

	Univariate 90% CI				Simultaneous 90% CI			
	OLS		FWLS		OLS		FWLS	
	5%	95%	5%	95%	0.33%	99.67%	0.33%	99.67%
Non-constant	81.172	85.472	82.46	86.165	79.751	86.893	81.235	87.39
Low	-4.308	-1.815	-4.403	-1.976	-5.131	-0.992	-5.205	-1.174
Lower-middle	-2.434	-0.419	-2.442	-0.834	-3.101	0.248	-2.973	-0.302
Upper-middle	-1.292	0.361	-1.290	-0.052	-1.839	0.908	-1.699	0.357
adult_m	-0.069	-0.062	-0.071	-0.064	-0.071	-0.059	-0.073	-0.061
HepB	0.009	0.038	0.003	0.029	0.000	0.048	-0.006	0.038
BMI	-0.555	6.845	-0.265	7.512	-3.000	9.290	-2.836	10.083
BMI^2	-3.842	1.551	-4.473	-0.158	-5.625	3.333	-5.899	1.268
alcohol	2.538	8.818	5.140	10.141	0.462	10.894	3.487	11.794
$alcohol^2$	-6.707	-1.402	-6.343	-2.167	-8.46	0.352	-7.723	-0.786
log(pm2.5)	-0.987	-0.183	-0.877	-0.201	-1.253	0.083	-1.101	0.022
GDP	-11.902	2.532	-9.986	1.133	-16.672	7.302	-13.661	4.807
GDP^2	5.947	15.136	7.116	14.303	2.910	18.173	4.741	16.678
hlth_exp	9.000	25.053	7.583	20.023	3.695	30.358	3.472	24.134
$hlth\_exp^2$	-11.912	-4.15	-7.848	-2.694	-14.478	-1.584	-9.551	-0.991

Table 19: Univariate and Bonferroni simultaneous 90% confidence intervals for parameters estimated by OLS and FWLS

Since the FWLS estimator **b** is assumed to be normally distributed with the mean $\boldsymbol{\beta}$ and variance $\sigma^2(\boldsymbol{Q^T Q})^{-1}$, the marginal distribution of each regression coefficient b_j is normal with the mean β_j and variance $\sigma^2 C_{jj}$, where C_{jj} is the j^{th} diagonal element of the $(\boldsymbol{Q^T Q})^{-1}$ matrix. Consequently, the standard error of each regression coefficient can be expressed as

$$se(b_j) = \sqrt{s^2(e)C_{jj}}\,, \tag{7.8}$$

where $s^2(e)$ is the estimate of the error variance obtained from equation (1.36). It follows that the $(1-\alpha)\times 100\%$ confidence interval for the coefficient β_j can be obtained as

$$b_j \pm t_{1-\alpha/2,\ n-k-1} \times \sqrt{s^2(e)C_{jj}}\,. \tag{7.9}$$

For instance, a point estimate of the parameter β_4 associated with the variable *adult_m* obtained by the FWLS approach is $b_4 = -0.0674$, the diagonal element of $(\boldsymbol{Q^T Q})^{-1}$ corresponding to this parameter is $C_{44} = 0.00000143$, and the standard error of the model amounts to $s(e) = 1.8125$. Using expression (7.9), we find that

$$b_4 \pm t_{0.95,\ 168} \times \sqrt{s^2(e)C_{44}}\,, \tag{7.10}$$

and the resulting univariate 90% confidence for the parameter β_4 is [-0.071; -0.064]. The Bonferroni simultaneous confidence interval for $\alpha = 0.1$ is obtained as

$$b_4 \pm t_{0.9967,\ 168} \times \sqrt{s^2(e)C_{44}}\ , \tag{7.11}$$

resulting in a slightly wider confidence interval [-0.073; -0.061] with the confidence coefficient being $(1 - \frac{0.05}{15})$ instead of $(1 - 0.05)$. Thereby, there is at least 90% probability that the whole set of the confidence intervals contains the true values of the population parameters. Thus, one can be 99.67% certain that the population parameter corresponding to the adult mortality rate will be between -0.074 and -0.061.

Figure 13 provides a visualization of the 90% simultaneous confidence intervals for the regression parameters β_0 through β_{15} of the regression model (7.1)

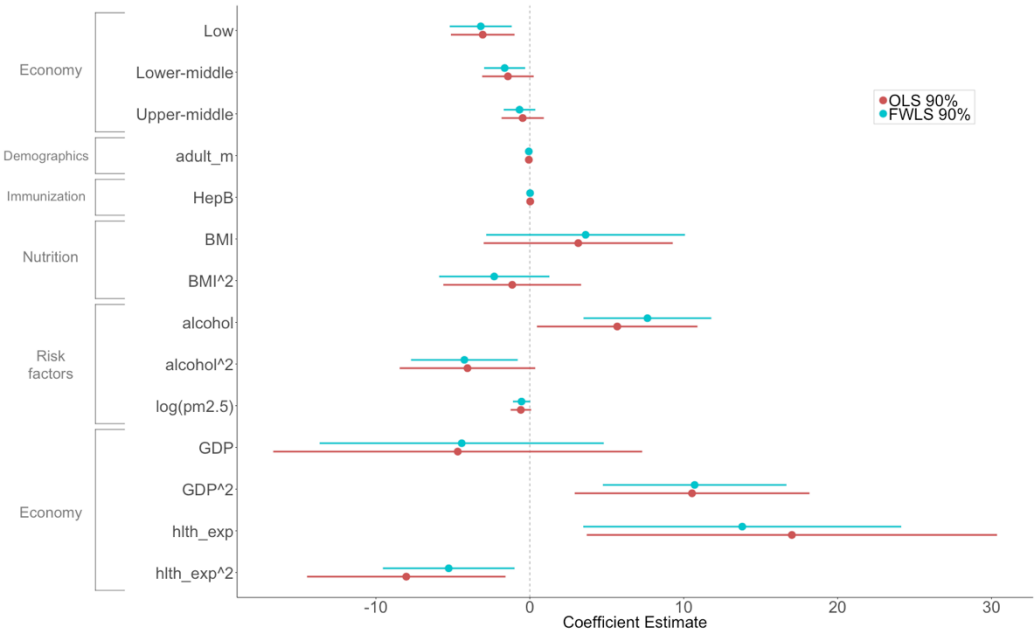

Figure 13: 90% Bonferroni simultaneous confidence intervals for parameters (models estimated by OLS and FWLS)

The 90% OLS confidence interval with Bonferroni correction suggested that the regressor *alcohol*^{2} has no effect on the dependent variable, whereas in the FWLS estimation, the effect is statistically significant. The body mass index is assumed to be insignificant at 10% level since its respective intervals cover zero, so the null hypotheses of no effect cannot be rejected. Nonetheless, assuming continuous scale, the interval contains a number of other values which cannot be rejected either. Moreover, statistical insignificance of a parameter does not necessarily imply an absence of a practical impact. Thus, one should analyze the process from both statistical significance and area-related knowledge. As

mentioned in section 6.2, people whose weight is much lower or much higher than is optimally healthy are more prone to various illnesses which may affect the life expectancy; thus, it was decided to keep the polynomial variable *BMI* in the model.

Furthermore, it is possible to measure the uncertainty related to any statistical estimator using the bootstrap resampling technique. The bootstrap approach involves repeatedly drawing a sample of *n* observations from the original data set of size *n* (with replacement), and estimation the model on each bootstrap sample. The coefficients were estimated on 500 bootstrap samples, each of size *n = 183,* the average of the standard errors after 500 estimations is taken, and the results provide an insight into the overall variance of the model performance.

	b_j	$se(b_j)$	$se_B(b_j)$
Non-constant	84.312	1.120	1.266
Low	-3.190	0.734	0.762
Lower-middle	-1.638	0.486	0.573
Upper-middle	-0.671	0.374	0.447
adult_m	-0.067	0.002	0.003
HepB	0.016	0.008	0.011
BMI	3.623	2.351	3.069
BMI^2	-2.316	1.304	1.715
alcohol	7.640	1.512	1.787
$alcohol^2$	-4.255	1.262	1.292
log(pm2.5)	-0.539	0.204	0.216
GDP	-4.427	3.361	5.884
GDP^2	10.709	2.173	6.308
hlth_exp	13.803	3.761	4.797
$hlth\_exp^2$	-5.271	1.558	4.365

Table 20: FWLS and bootstrap standard errors

The bootstrap standard errors are larger than the usual estimates of the regression coefficient standard errors. For instance, the standard error for the coefficient estimate corresponding to the lower-middle income group obtained using the formulas from section 1.4.3, is $se(b_{Lower-middle})$ = 0.486 as opposed to the bootstrap standard error estimate $se_B(b_{Lower-middle})$ = 0.573. The univariate 95% confidence interval for $b_{Lower-middle}$ is narrower in the first case, that is [- 2.610, - 0.667]. If the bootstrap estimate of $se_B(b_{Lower-middle})$ is used, the resulting confidence interval [- 2.784; -0.492] is slightly wider.

The traditional formula for the estimate standard error(1.35) involves the unknown parameter σ^2, which must be estimated and depends on the correctness of the linear model. Moreover, it is assumed that the x_{ik} are fixed and all the variability stems from the error term variation. Since the bootstrap approach does not rely upon these assumptions, it is likely to produce more accurate estimates of the standard errors of the b_j. Therefore, confidence intervals calculated on the basis of the bootstrap standard errors will be more appropriate for statistical inference (James et al., 2013).

7.5. Confidence Regions

Since there are 15 parameters in the regression model under consideration, it is impossible to examine the 15-dimensional joint confidence region directly. Still, as mentioned in section 2.2.3, two-dimensional confidence regions might be constructed ignoring the rest 13 parameters. The quadratic form for the joint confidence region for the parameters corresponding to the linear terms for the GDP and alcohol consumption is constructed from the equation (2.20) by:

1. replacing $(\boldsymbol{\beta} - \boldsymbol{b})$ with the (2×1) vector comprising only two parameters under consideration;

2. replacing $(\boldsymbol{X^T X})$ with the inverse of the (2×2) estimated variance-covariance matrix $S(\boldsymbol{b})$ for these two parameters;

3. replacing $p\ s^2(e)\ F_{1-\alpha,\ p,\ n-p}$ with $2\ F_{1-\alpha,\ 2,\ n-p}$.

The term $s^2(e)$ does not appear in the right-hand side of the inequality since it has been counted for in the variance-covariance matrix according to the formula (1.38).

The estimates of parameters $\beta_{alcohol}$ and β_{GDP} are 7.6404 and -4.4266, respectively. The elements of $S(\boldsymbol{b})$ corresponding to the $b_{alcohol}$ and b_{GDP} are as follows:

$$S(\boldsymbol{b}) = \begin{bmatrix} 2.2861 & 0.5922 \\ 0.5922 & 11.2978 \end{bmatrix}.$$

The main diagonal of the matrix contains the variances of parameter estimates, and the off-diagonal elements represent the joint variability of these estimates. After taking the inverse of the matrix, the 95% joint confidence region is obtained from

$$\begin{bmatrix} \beta_{alcohol} - 7.6404 \\ \beta_{GDP} - (-4.4266) \end{bmatrix}^{\boldsymbol{T}} \begin{bmatrix} 0.4434 & -0.0232 \\ -0.0232 & 0.0897 \end{bmatrix} \begin{bmatrix} \beta_{alcohol} - 7.6404 \\ \beta_{GDP} - (-4.4266) \end{bmatrix} \leq 2\ F_{0.95,\ 2,\ 168}.$$

Solving this inequality with respect to $\beta_{alcohol}$ and β_{GDP} for the boundary $F_{0.95,\ 2,\ 168}$= 3.0498 establishes a 95% confidence region for the two parameters in question. The outer ellipse in figure 14 displays the resulting 95% joint confidence region for $\beta_{alcohol}$ and β_{GDP}, meaning that there is 95% chance that these two population parameters will lie inside of the ellipse simultaneously. The ellipse is centered around the point whose coordinates are given by the estimates $b_{alcohol}$ and b_{GDP}. The upward slope of the ellipse stems from the positive covariance between the parameter estimates $S(b_{alcohol}, b_{GDP})$ =0.5922 suggesting the same direction of the errors in the point estimates.

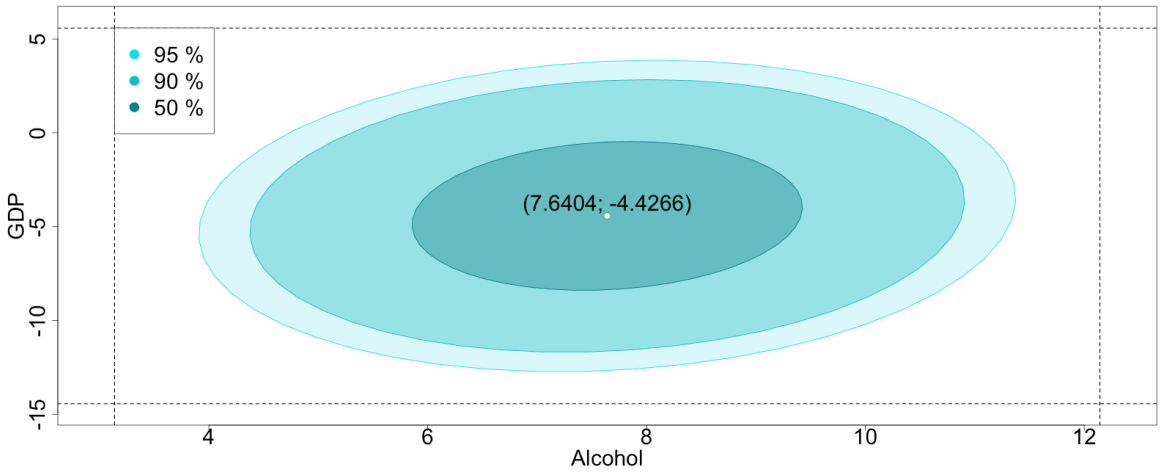

Figure 14: 50%, 90%, and 95% confidence ellipses for parameters $\beta_{alcohol}$ and β_{GDP}. Corners of rectangle formed by dashed lines represent the intersection of the 95% Bonferroni univariate confidence intervals.

8. Outliers and Influential Observations

8.1. Leverage: Hat-Values

Data points with high leverage, that is with large hat-values, have the potential to affect the fitted model considerably. In our case with number of parameters p = 15 and sample size n = 183, the average hat-value is $\overline{h}$ = 15/183 = 0.082. Recall that an observation can be regarded as extreme with respect to other data points if its hat-value is greater than two or three times the mean leverage $\overline{h}$.

Table 21 contains the diagonal values of the hat-matrix, which exceed the cut-off values $2\overline{h}$ = 0.1639 and $3\overline{h}$ = 0.2459. Out of 183 observations, 14 countries exhibit high leverage, indicating that their predictor values are unusual relative to other countries. In accordance with the equation (4.6), the trace of the **H** matrix expressed as the sum of its diagonal values with is exactly equal to the number of parameters $tr(\boldsymbol{H}) = \sum_{i=1}^{183} h_{ii} = 15$.

	Hat-value
Monaco	0.9259
USA	0.5954
Switzerland	0.5697
Samoa	0.4713
Tonga	0.4381
Equatorial Guinea	0.2829
Brunei Darussalam	0.2686
Moldova	0.2664
Luxembourg	0.2597
Malta	0.2406
Lithuania	0.2161
Qatar	0.1856
Iraq	0.1773
Ukraine	0.1655

Table 21: Hat-values exceeding thresholds $2\overline{h}$ (below dashed line) and $3\overline{h}$ (above dashed line)

As seen from figure 15, the more distant observations in terms of their positions in the **Q**-space (transformed **X**-space) are Samoa, Tonga, Switzerland, and the USA. Their leverage values are more than 5 times further from the mean hat-value (> 0.4098), partially due to the fact that they obtained the highest weights in the FWLS estimation (for each of these countries the weighting factors $1/\sqrt{w_i}$ are greater than 3). Furthermore, according to the original unweighted data, residents of Switzerland and the USA spend the most on the

healthcare services. At the same time, in 2016 Samoa and Tonga, island states located in the South Pacific Ocean, had the largest average body mass indices in the world, 31.9 for both, which corresponds to the obesity of the first type.

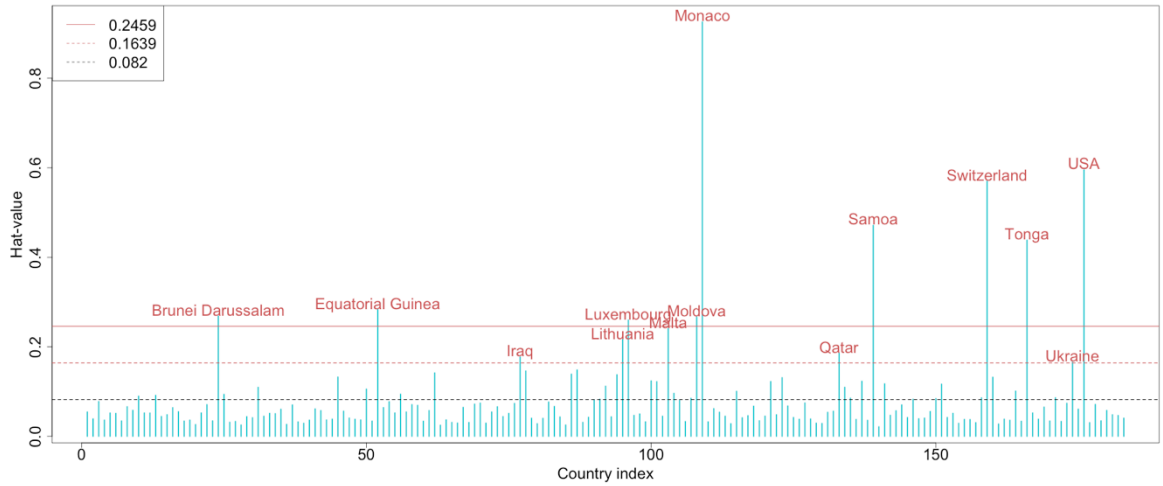

Figure 15: Hat-values

The most isolated data point is Monaco, with the hat value h_{Monaco} = 0.9259, close to the maximum possible of one. Even though the weighting factor for Monaco is almost twice smaller than that of the USA, its response and predictor values are still weighted heavier than for 75% of the countries in the FWLS estimation procedure. Thereby, the extremely large hat-value of Monaco may stem from the outstandingly high GDP per capita (191,586.6 USD), which is approximately 36 times higher than the median GDP per capita in 2016 (5,382.8 USD).

Since the leverage depends only on the regressor values, additional statistics should be calculated to determine the real influence a particular data point has on the estimated regression function.

8.2. Regression Outliers: Externally Studentized Residuals

Generally, the attention is drawn to the data points with the largest externally studentized residuals exceeding the threshold of |2|, and even more concern arises if the residuals exceed |3| (Blatná, 2006). Figure 16 provides evidence of the presence of outlying observations since nine points fall behind the limits [-2; 2].

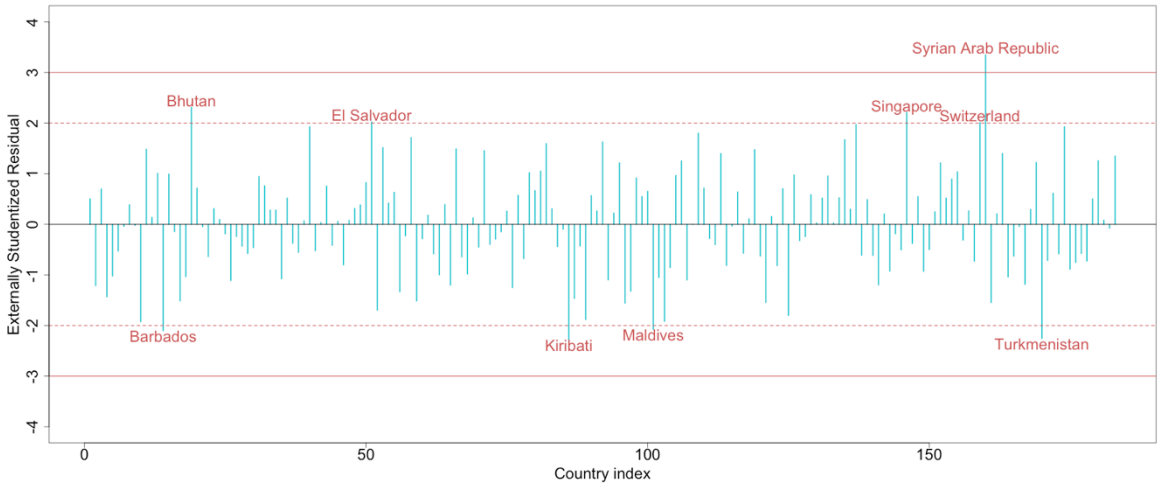

Figure 16: Externally studentized residuals

The mean-shift outlier model may serve as a useful way of regression outliers detection. Recall a modified model without an intercept from the equation (1.44)

$$y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \cdots + \beta_k x_{ik}^* + \gamma d_i + \varepsilon_i^*,$$

where d_i is a binary variable coded 1 for the observation i and 0 for all other observations. If parameter $\gamma \neq 0$, the conditional expectation of the i^{th} observation depend on the regressors $X_0^*, X_1^*, \ldots, X_k^*$ in the same manner as the other data points, but its intercept is shifted from 0 to γ. The t-statistic for testing the null hypothesis H_0: $\gamma = 0$ that the i^{th} observation is not a mean-shift outlier against the two-sided alternative is identical to the externally studentized residual e_{Ji}, that has t distribution with $(n - k - 2)$ degrees of freedom.

	Externally studentized residual
Syrian Arab Republic	3.3466
Bhutan	2.3103
Kiribati	2.2838
Turkmenistan	2.2534
Singapore	2.2085
Barbados	2.1006
Maldives	2.0741
El Salvador	2.0219
Switzerland	2.0077

Table 22: Absolute values of externally studentized residuals exceeding thresholds |2| (below dashed line) and |3| (above dashed line)

The largest e_{Ji} corresponds to the Syrian Arab Republic: $e_{J,Syria}$= 3.3466, and it is the only value which exceeds the cut-off |3| according to the table 22.

The p-value of the two-tailed t-test on parameter γ associated with the dummy variable d_{Syria} amounted to 0.001011. However, since the studentized residuals are not independent, the correct approach is to evaluate the Bonferroni-corrected t-test by multiplying the usual two-sided p-value by the sample size n = 183. The Bonferroni upper bound for the p-value = 0.18495 is not significant; thus, we fail to reject the null hypothesis that Syria is not a mean-shift outlier. On the contrary, Switzerland has one of the highest hat-values (0.5697) and the absolute value of residual value exceeding the cut-off value 2. Thus, it should be further examined for being an influential observation.

8.3. Influence Measures

8.3.1. Cook's Distance

Figure 17 illustrates the hat-values, studentized residuals, and Cook's distance simultaneously. The radius of each circle is proportional to the square root of the Cook's D_i, and hence the area is proportional to the Cook's D of the i^{th} observation. It is clearly visible that Monaco dramatically differs from the rest of the countries, and thus it most likely has considerable influence on the regression estimation.

Figure 17: Plot of hat-values, externally studentized residuals and Cook's distances. Size of circles is proportional to Cook's D_i

As mentioned in the inequality (4.13), the observation i might be declared as influential if the D_i is greater than the median of an F distribution with (15, 168) degrees of freedom, which is $F_{0.5,\,15,\,168}$= 0.9597 in this particular case. Since the standardized residual for Monaco $e_{S,Monaco}$= 1.7884 exceeds the 95th percentile of its distribution, and the leverage is close to unity h_{Monaco} = 0.9259, the corresponding Cook's distance computed as

$D_{Monaco} = \left[\frac{1.7884^2}{15}\right] \times \left[\frac{0.9259}{1\text{-}0.9259}\right]$ = 2.6652 stands out substantially from the rest of the data points. The extreme remoteness of the statistic from the critical value implies that exclusion of this observation from the estimation procedure can substantially change the coefficients estimates.

Another widely used cut-off for detection of the influential observations based on the Cook's distance is $D_i > \frac{4}{n-p}$, which amounts to 0.0238 in the case of 183 observations and 15 estimated parameters. Based on this criterion, fourteen more countries may be viewed as influential according (see table 21). The second largest Cook's distance after Monaco corresponds to Switzerland, $D_{Switzerland}$ = 0.3495, which is 7.6 times less than that of Monaco.

Therefore, it is reasonable to exclude these two observations from the estimation and evaluate the change in the regression coefficients. Table 23 provides the results of dropping the data points for Monaco and Switzerland one-at-a-time and both simultaneously.

	All in	Monaco out	Switzerland out	Both out
Non-constant	84.3125	83.9802	83.8028	83.4653
Low	-3.1896	-2.7611	-3.1288	-2.6969
Lower-middle	-1.6379	-1.2556	-1.5729	-1.1876
Upper-middle	-0.6706	-0.3745	-0.6428	-0.3444
adult_m	-0.0674	-0.0675	-0.0661	-0.0662
HepB	0.0157	0.0165	0.0182	0.019
BMI	3.6234	3.5436	4.4215	4.3457
BMI^2	-2.316	-2.2147	-2.1729	-2.0701
alcohol	7.6404	7.7642	7.3341	7.457
$alcohol^2$	-4.2545	-3.9942	-4.4114	-4.1501
log(pm2.5)	-0.5393	-0.5465	-0.528	-0.5351
GDP	-4.4266	-5.8545	-5.7937	-7.2396
GDP^2	10.7095	4.291	11.3766	4.9162
hlth_exp	13.8027	13.148	15.5008	14.8512
$hlth\_exp^2$	-5.2708	-4.8779	-7.3279	-6.944

Table 23: Regression coefficients estimated with and without Monaco and Switzerland

Omitting Monaco decreased the coefficient for *GDP* by about 32%, GDP^2 by 60% and increased the coefficient for the upper-middle income group by 44%. Removing Switzerland and retaining Monaco leads to 12% and 39% increase in the coefficient estimates for *hlth_exp* and $hlth\_exp^2$ respectively, which stems from the highest spending

on the healthcare in that country. Exclusion of both, Monaco and Switzerland, results in the coefficient estimates for *GDP* and GDP^2 being reduced by 63% and 54%, respectively, as well as the estimated parameter for the upper-middle-income class being increased by 49% as compared to the model based on the complete dataset. Apparently, these two countries extensively affect the model estimation even after the weighting procedure.

8.3.2. DFFITS

The following table summarizes the Cook's distances, sorted by the absolute values of the $DFFITS_i$ exceeding the threshold $2\sqrt{\frac{15}{183\text{-}15}} = 0.5976$. Based on this cut-off, fifteen countries have a feasible influence on the fitted values of the life expectancy. However, the critical values for identifying the observations as influential are just the basic guidelines, and it is more important to pay attention to the data points which substantially stand out from the rest of the sample.

	Cook's distance	$\lvert DFFITS_i \rvert$
Monaco	2.6652	6.3649
Switzerland	0.3495	2.3101
Syrian Arab Republic	0.1071	1.3051
Malta	0.0764	1.0791
Equatorial Guinea	0.0749	1.0655
Luxembourg	0.0563	0.9228
Kiribati	0.0547	0.9169
USA	0.0555	0.9112
Ukraine	0.0483	0.8584
Maldives	0.0391	0.7733
Saint Lucia	0.0357	0.7377
Singapore	0.0286	0.6624
Lithuania	0.0270	0.6369
Kuwait	0.0247	0.6105
Azerbaijan	0.0239	0.6036

Table 24: Cook's distances exceeding thresholds [4/(183-15)] (below dashed line) and $F_{0.5,\ 15,\ 168}$ (above dashed line); $DFFITS_i$ exceeding threshold [2√(183-15)]

Since both the Cook's D_i and $DFFITS_i$ depend directly on the magnitude of the hat-values, they provide similar results. It is evident that Monaco and Switzerland have the highest scores for both measures, which distinguish them from the rest of the countries and proves

the fact that these they exert a significant influence on the parameter estimates and the fitted values.

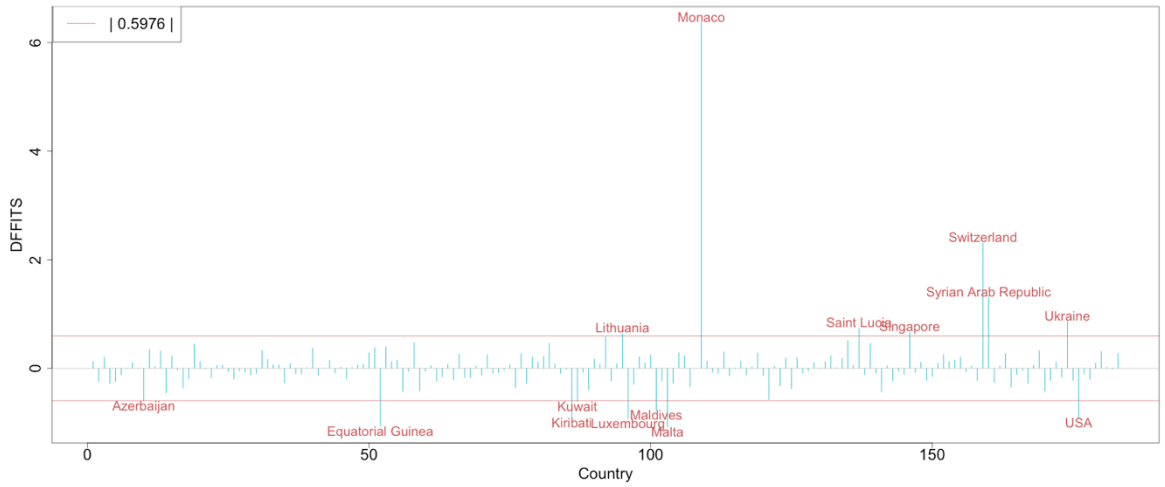

Figure 18: *DFFITS$_i$*

It is worth mentioning that Monaco is one of the most problematic observations in terms of data availability. Most of the official international databases do not provide information on many socio-economic indicators and other attributes for Monaco. Therefore, the necessity for imputation of the missing values contributes to additional risks when Monaco participates in the estimating and predicting procedures.

9. Variable Selection Procedures

The FWLS model from the equation (7.4) has been defined as the full starting model, with regard to which several variable selection procedures were applied. Recall this model:

$$\begin{aligned} life\_exp = {} & 84.31\ Non\text{-}constant \\ & - 3.19\ Low - 1.64\ Lower\text{-}middle - 0.67\ Upper\text{-}middle \\ & - 0.07\ adult\_m + 0.02\ HepB + 3.62\ BMI - 2.32\ BMI^2 \\ & + 7.64\ alcohol - 4.25\ alcohol^2 - 0.54\ log(pm2.5) \\ & - 4.43\ GDP + 10.71\ GDP^2 \\ & + 13.80\ hlth\_exp - 5.27\ hlth\_exp^2 + e\ . \end{aligned} \tag{9.1}$$

The values of the information criteria for the initial Model 1, computed on the basis of the formulas (5.5) and (5.6), are $AIC_{full} = 183 \times \log(\frac{551.92}{183}) + 2 \times 15 = 232.02$ and $BIC_{full} = 183 \times \log(\frac{551.92}{183}) + \log(183) \times 15 = 280.16$, which provide a benchmark for selection of the subset models relative to the full model. The null model, that is, the smallest model we are willing to entertain is specified as

$$\begin{aligned} life\_exp = {} & 80.11\ Non\text{-}constant \\ & - 18.44\ Low\ - 11.13\ Lower\text{-}middle - 6.30\ Upper\text{-}middle\ , \end{aligned} \tag{9.2}$$

where the first coefficient corresponds to the column of the weighting factors for the individual observations (diagonal elements of the transformation matrix **P**). Intentionally, R has been forced to include the *inc_level* variable, so that the algorithm does not omit individual levels, but involves all categories of the dummy variable in the model estimation. The information criteria for the null model $AIC_{null} = 652.25$ and $BIC_{null} = 665.09$ exceed the corresponding values for the full model by more than two times, implying a worse fit provided by the restricted model.

The backward elimination method based on the information criteria omitted the second order polynomial in BMI, which decreased the AIC from 232.02 to 231.68 and BIC from 280.16 to 273.4. Further variables removal increases the criteria, so the algorithm stoped after the regressors *BMI* and BMI^2 were dropped from the model. The forward selection procedure provided identical results: starting with the null model, the variables have been added sequentially according to the largest potential reduction of the information criteria. All variables except for *BMI* and BMI^2 have been included in the model, which reduced the

values of AIC from 652.25 to 232.02 and BIC from 665.09 to 231.68. Absolutely the same result was obtained using the stepwise regression.

The *p*-values for these two terms in the full model amounted to 0.1252 for *BMI* and 0.0776 for BMI^2, after exclusion of the squared term the *p*-value for the *BMI* coefficient increased to 0.6257. Thus, when the procedures were repeated using the stopping significance levels α_{remove} and α_{enter} being equal to 0.01, 0.05 and 0.1, the body mass index was equivalently considered to be insignificant in the prediction of the life expectancy, and the resulting model is expressed as

$$\begin{aligned} life\_exp = {} & 84.12\ Non\text{-}constant \\ & - 3.65\ Low - 1.72\ Lower\text{-}middle - 0.64\ Upper\text{-}middle \\ & - 0.07\ adult\_m + 0.02\ HepB \\ & + 7.97\ alcohol - 4.13\ alcohol^2 - 0.54\ log(pm2.5) \\ & - 4.49\ GDP + 10.61\ GDP^2 \\ & + 14.05\ hlth\_exp - 5.47\ hlth\_exp^2 + e\ . \end{aligned} \tag{9.3}$$

However, when the α_{remove} and α_{enter} were both set to 0.001, the three algorithms identified the variables *HepB* and *log(pm2.5)* as inactive on a par with *BMI* and BMI^2, resulting in the estimated model

$$\begin{aligned} life\_exp = {} & 84.71\ Non\text{-}constant \\ & - 3.94\ Low - 1.92\ Lower\text{-}middle - 0.80\ Upper\text{-}middle \\ & - 0.07\ adult\_m \\ & + 9.10\ alcohol - 4.59\ alcohol^2 \\ & - 6.50\ GDP + 11.90\ GDP^2 \\ & + 16.24\ hlth\_exp - 5.96\ hlth\_exp^2 + e\ . \end{aligned} \tag{9.4}$$

The next chapter aims at the investigation of the predictive abilities of the initial model (9.1), denoted as Model 1, and two models selected by the automated procedures (9.3) and (9.4), denoted as Model 2 and Model 3, respectively.

10. Cross-Validation

The regression models are useful not only in terms of estimation of the mean response for a particular set of observed values of the independent variables but also for prediction of the response values corresponding to the new observations. Cross-validation is one of the most widely applied resampling methods for assessment of the predictive performance of the model. It consists of repeatedly drawing samples from the available data and fitting the model under consideration on each sample in order to obtain information on the predictive abilities of a given model (James et al., 2013).

The family of cross-validation methods includes different approaches such as the validation set (hold-out set), leave-one-out cross-validation (LOOCV), *k*-fold, and repeated *k*-fold cross-validation. The underlying idea of these techniques consists of dividing the data into two subsets: the training set used to build the regression model, and the testing set used to validate the model by evaluating the error of prediction.

To compare the models by their predictive performance on the testing data, different metrics can be computed. The Mean Absolute Error (MAE) measures the average absolute difference between the actual response values and the predicted values:

$$MAE = \frac{\sum_{i=1}^{n} |q_i - \hat{q}_i|}{n}, \tag{10.1}$$

where $\hat{y}_i$ is a point predictor for the observation *i.* It quantifies the average magnitude of the prediction errors.

An alternative to MAE is the Root Mean Squared Error (RMSE), obtained as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(q_i - \hat{y}_i)^2}{n}}. \tag{10.2}$$

The advantage of RMSE over MAE is usefulness in cases when large prediction errors are undesirable since large errors are given higher weights after squaring.

- **Validation Set Approach**

The validation set approach is a simple strategy, which involves randomly splitting the available data into two parts: a training set and a hold-out (validation) set. The size of each of the subsets is arbitrary and depends on the sample size. For a sample of 183 countries, 80% of observations were randomly assigned to the training set (n_{Train} = 147) and 20% to

the validation set (n_{Test} = 36). The regression Models 1, 2, and 3 are estimated based on the training data, and then used to predict the values of the life expectancy for the data not used in the estimation.

- **Leave-One-Out Cross-Validation**

In the LOOCV approach, one data point is excluded from the estimation procedure, the model is built on *(n* – 1 = 183 – 1*)* observations and then tested against the left-out point. The process is repeated *n* = 183 times, and the LOOCV estimates are then the averages of the test error estimates:

$$CV_{(n)}^{MAE} = \frac{1}{n}\sum_{i=1}^{n} MAE_i \ , \tag{10.3}$$

$$CV_{(n)}^{RMSE} = \frac{1}{n}\sum_{i=1}^{n} RMSE_i \ . \tag{10.4}$$

The LOOCV decreases potential bias but may lead to a higher variation of the prediction error if some observations are outliers.

- ***k*-Fold Cross-Validation**

An alternative to the LOOCV is the *k*-fold cross-validation, which partitions the original sample into *k* approximately equal subsamples. The process consists of *k* iterations, each time a model is built on *(k* – 1*)* folds and validated using the remaining group. In most practical applications number of subsets is usually set to *k* = 5 or *k* = 10, as these values result in the test error estimates that do not suffer either from the high variability or from high bias (James et al., 2013). The *k*-fold CV estimates are obtained by averaging the MAE and RMSE across the *k* groups

$$CV_{(k)}^{MAE} = \frac{1}{k}\sum_{i=1}^{k} MAE_i, \tag{10.5}$$

$$CV_{(k)}^{RMSE} = \frac{1}{k}\sum_{i=1}^{k} RMSE_i. \tag{10.6}$$

- **Repeated *k*-Fold Cross-Validation**

The repeated k-Fold cross-validation is a modification of the previous approach, whereby the cross-validation is conducted multiple times, and each time the data is split into *k* folds

in a different way. The overall model error is taken as the average error from the number of repeats.

The four methods have been applied to the full dataset, and a dataset with Monaco and Switzerland excluded in order to evaluate how the main model and the models chosen by the automated selection procedures perform in the prediction of the life expectancy. It is necessary to remember that the procedures are based on the transformed data. Since the weights for 131 counties are greater than 1, the response values $\boldsymbol{q}$ may be out of the range of possible life length: for example, in 2016, life expectancy at birth in the Czech Republic was equal to 79.2 for both genders. After estimation of the matrix **P**, Czechia was assigned the weight of 2.11. The resulting transformed value of the response was obtained by multiplying these two values amounted to approximately 167.13 years, which is unrealistic. Therefore, the prediction errors calculated on the basis of the modified values of response and explanatory variables may yield higher values, than if the original observed values were used for modeling.

The following tables illustrate the resulting cross-validation estimates for the test errors:

Model	Model 1		Model 2		Model 3	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Train/Test Split	2.0661	1.7079	2.0683	1.6992	1.9294	1.6057
LOOCV	8.5171	6.0077	14.3160	8.9504	18.9640	12.8167
10-fold	7.9617	5.9289	15.4308	9.8870	17.6759	12.9732
Repeated 10-fold (3 repeats)	8.0383	5.9637	13.6831	9.3078	18.1521	12.9196

Table 25: Cross-validation RMSE and MAE of three models, obtained using full dataset

Model	Model 1		Model 2		Model 3	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Train/Test Split	1.7235	1.3374	1.74852	1.4039	1.8692	1.4517
LOOCV	7.3903	5.6202	12.0443	8.4055	16.3366	12.1641
10-fold	7.4114	5.6829	11.3270	8.4378	15.4737	11.9916
Repeated 10-fold (3 repeats)	7.2822	5.6463	11.5166	8.5418	16.4884	12.3920

Table 26: Cross-validation RMSE and MAE of three models, obtained using dataset with Monaco and Switzerland deleted

The resulting RMSE and MAE for the validation set approach are considerably less than the other estimates for the test errors. This may stem from the fact that random splitting of the dataset into training and validation sets depends on the seed specified prior to the division, and the validation set may contain observations that are relatively easy to predict. Deletion of Monaco and Switzerland from the sample data resulted in the reduction of the test error estimates, thus, advocating the improved accuracy of the predictions made by all three models. The lowest values of the RMSE statistics correspond to the initial model (9.1), suggesting that omitting of the variables *BMI* and BMI^2 increases the standard deviation of the predicted values from the actual ones by more than 50% in the second model (from 7.4 to 11.3 according to the 10-fold cross-validation). The third model, that is, with the predictors *BMI,* BMI^2*, HepB* and *log(pm2.5)* not participating in the procedures provides the most inaccurate prediction results, which is reflected in the highest validation error estimates.

Consequently, in all four validation methods, there is a clear superiority of the model with all variables over the stepwise selection procedures, based purely on results.

Still, the smallest average absolute difference between the predicted and observed values of approximately 5.6 years (table 24) is still considerably high and might suggest the inclusion of additional predictors to the model which contain relevant information necessary for prediction of the life expectancy.

Conclusion

The steps of the regression model-building process were presented in the context of the thesis introducing the principles of the multiple linear regression model and its basic properties. Based on a sample drawn randomly from a population, the ordinary least squares approach is employed to estimate the unknown intercept and slopes parameters in the population regression model. The multiple regression models allow examining the partial effect of a particular independent variable on the response while holding other factors fixed. The algebra of the OLS estimation was demonstrated, including computation of parameter estimates which reflect the expected changes in the dependent variable *life expectancy* for given changes in the predictors. Although the models are linear in parameters, they may be conveniently used to model non-linear relationships by selecting appropriate forms of the dependent and independent variables, such as power or logarithmic transformations.

The specification and data issues that frequently arise in the experimental cross-sectional analyses have been addressed. Incorrect specification of the functional form makes the estimated model difficult to interpret and can be detected by the RESET test, without additional data collection.

Section 1.3 described the assumptions of the classical linear regression model, under which the OLS estimators are unbiased, meaning that the expectations on the parameter estimates are the corresponding population parameters themselves. When the assumption of the constant error variance is added, we obtain simple formulas for the sampling variances of the OLS estimators. It can be seen from the equation (1.31), that the slope estimators' variance increases with rising error variance, while it decreases when there is more variation in the regressors.

Unfortunately, in many social science practical applications, the crucial assumption of zero mean of the error term ε conditional on the predictor variables $\mathbf{X}$ does not hold, since the omitted features in ε are often correlated with $\mathbf{X}$. The analysis of the life expectancy, although intentionally assumed exogeneity of the explanatory variables, still requires inclusion of important regressors which were not controlled for. However, adding more variables leads to increasing number of parameters to be estimated which results in the unreliable *t*-ratios and instability of parameters. Thus, in the context of this thesis only

several features have been selected to be able to fulfill its main goal of the model building process illustration.

Dealing with the omitted variables problem, that is when one or more relevant variables are left out, is more complicated, and one of the possible ways to address this issue is based on incorporating a proxy variable for the omitted variable. Under some rational assumptions, involving the proxy variable into the OLS regression may eliminate, or at least, mitigate the bias (Wooldridge, 2015).

Several common diagnostics used to evaluate whether the CLRM assumptions hold for a particular linear regression model were discussed in chapter 3 and applied to the real data in chapter 7. We addressed two techniques to test for heteroskedasticity: the Breusch-Pagan test and a modified version of the White test. Both of these test statistics engage regressing the squares of the OLS residuals on either the explanatory variables (Breusch-Pagan test) or the fitted and squared fitted values (a special case of the White test). The conclusions may rely upon either the F-tests or the Lagrange multiplier analogs of the tests. We revealed that countries with low level of development contribute to the higher spread of the residuals at lower values of the life expectancy, whereas countries with higher level of development provide a more reliable basis for the coefficients' estimation.

Heteroskedastic errors do not imply a bias in the OLS estimators, but the usual standard errors are no longer valid, and t- and F-test statistics do not have the expected t or F distribution, respectively. In the existence of heteroskedasticity OLS is not the best linear unbiased estimator any longer, that is the estimator with the lowest variance. When the shape of heteroskedasticity is known, the weighted least squares (WLS) estimation can be applied as an alternative to the OLS. More commonly, the function of the heteroskedasticity is unknown and must be estimated before using the WLS. The resulting feasible WLS estimator is no longer unbiased, but it converges in probability to the true population parameter as the sample size increases. When the error term is normally distributed, the test statistics from the FWLS estimation procedure are valid, assuming the heteroskedasticity has been appropriately modeled. Since the low-income areas are the main cause of high residual variance, their impact on the parameter estimates was reduced by assigning lower weight to these countries in the FWLS approach.

Apart from the quantitative data, the use of the qualitative information was shown in the estimation of the mean life expectancy at birth by incorporating the categorical variable *income level*. All the estimates on this dummy variable were interpreted relative to the benchmark group of high-income countries, for which no dummy variable was included in the model.

Furthermore, we have examined the topic of statistical inference, which allows inferring statements about the population from a random sample. We tested hypotheses about the statistical significance of the model and single parameters using the F-test and two-tailed t-tests, respectively. In traditional hypothesis testing, one first chooses a significance level, which, along with the alternative hypothesis and degrees of freedom, defines the critical value against which the associated test statistics are compared. We have discussed various ways for construction of the confidence intervals, both parametric (univariate and simultaneous CI, joint confidence regions) and non-parametric (using the bootstrapped standard errors).

Additionally, we have distinguished several types of outlying observations which can substantially affect the OLS estimates, primarily in small samples. It is essential to identify outliers based on some measures and generally accepted rules along with the knowledge about the process and then to re-estimate models without the suspected outliers. It was shown that Monaco and Switzerland have the highest scores for almost all measures of outlyingness, which differentiates them from the rest of the observed countries and confirms the fact that their presence (or absence) in the regression modeling procedure significantly affects the parameter estimates and fitted values.

Finally, we have concisely discussed several variable selection procedures which may serve as general guidelines for identification of the active and inactive (unimportant) variables in the regression model. The cross-validation resampling techniques were then conducted in order to asses the ability of the initial model and models proposed by the automated selection procedures to predict new data that did not participate in the estimation stage.

References

Anderson, T.W., 1962. *On the Distribution of the Two-Sample Cramér-von Mises Criterion.* The Annals of Mathematical Statistics 33, 1148–1159.

Anderson, T.W., Darling, D.A., 1954. *A Test of Goodness of Fit.* Journal of the American Statistical Association 49, 765–769. https://doi.org/10.2307/2281537

Bašta, M., 2017. *Course 4ST616 Regression.* Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics in Prague.

Bašta, M., 2018. *Properties of Backward Elimination and Forward Selection in Linear Regression.* Presented at the The 12th International Days of Statistics and Economics, Prague, , Czech Republic.

Belsley, D.A., Kuh, E., Welsch, R.E., 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* John Wiley & Sons. http://doi.org/10.1002/0471725153

Blatná, D., 2006. *Outliers in Regression*, in: Applications of Mathematics and Statistics in Economy. Presented at the 9th International Scientific Conference, Trutnov, Czech Republic.

Breusch, T.S., Pagan, A.R., 1980. *The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics.* The Review of Economic Studies 47, 239–253. https://doi.org/10.2307/2297111

Cook, R.D., 2000. *Detection of Influential Observation in Linear Regression.* Technometrics 42, 65–68. https://doi.org/10.2307/1271434

Davidson, R., MacKinnon, J.G., 2003. *Econometric Theory and Methods.* Oxford University Press Inc, New York, United States. https://doi.org/10.1017/S0266466605000356

Environmental Protection Agency, 2013. Federal Register, The Daily Journal of the United States Government. *National Ambient Air Quality Standards for Particulate Matter*; Final Rule 78.

Fox, J., Weisberg, S., 2011. *An R Companion to Applied Regression*, 2nd ed. Publications, Inc, Thousand Oaks, California, United States.

Franzese, R.J., Kam, C., 2009. *Modeling and Interpreting Interactive Hypotheses in Regression Analysis.* University of Michigan Press. https://doi.org/10.3998/mpub.206871

Greene, W.H., 2003. *Econometric analysis*, 5th ed. ed. Prentice Hall, Upper Saddle River, New Jersey, United States.

Gujarati, D.N., 2018. *Linear Regression: A Mathematical Introduction*. SAGE Publications.

Harrell, F.E., 2001. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.* Springer Science & Business Media. http://doi.org/10.1007/978-1-4757-3462-1

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. *An Introduction to Statistical Learning,* Springer Texts in Statistics. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4614-7138-7

Jarque, C.M., Bera, A.K., 1987. *A Test for Normality of Observations and Regression Residuals.* International Statistical Review / Revue Internationale de Statistique 55, 163–172. https://doi.org/10.2307/1403192

Lilliefors, H.W., 1967. *On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown.* Journal of the American Statistical Association 62, 399–402. https://doi.org/10.2307/2283970

McDonald, B., 2002. *A Teaching Note on Cook's Distance - A Guideline.* Research Letters in the Information and Mathematical Sciences 3, 127–128.

Montgomery, D.C., Peck, E.A., Vining, G.G., 2012. *Introduction to Linear Regression Analysis.* John Wiley & Sons, Incorporated, Hoboken, United States.

OECD (Ed.), 2017. *Health at a glance 2017: OECD indicators.* OECD, Paris. https://doi.org/10.1787/19991312

R Core Team (2018). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Rawlings, J.O., Pantula, S.G., Dickey, D.A., 1998. *Applied regression analysis: a research tool,* 2nd ed. ed, Springer texts in statistics. Springer, New York. http://doi.org/10.1007/b98890

Shapiro, S.S., Wilk, M.B., 1965. *An Analysis of Variance Test for Normality (Complete Samples).* Biometrika 52, 591–611. https://doi.org/10.2307/2333709

Stekhoven, D.J., Bühlmann, P., 2012. *MissForest - nonparametric missing value imputation for mixed-type data.* Bioinformatics 28, 112–118. https://doi.org/10.1093/bioinformatics/btr597

Stephens, M.A., 1974. *EDF Statistics for Goodness of Fit and Some Comparisons.* Journal of the American Statistical Association 69, 730–737. https://doi.org/10.2307/2286009

White, H., 1980. *A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.* Econometrica 48, 817–838. https://doi.org/10.2307/1912934

WHO (2019). Body mass index - BMI. [online] WHO/Europe. Available at: http://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi

WHO (n.d.). *Global Health Observatory data repository.* [online] World Health Organisation. Available at: http://apps.who.int/gho/data/node.home

Wooldridge, J.M., 2015. *Introductory Econometrics: A Modern Approach*, 6th ed. Cengage Learning.

World Bank (2019). *Country Classification: How are the income group thresholds determined?*. [online] Available at: https://datahelpdesk.worldbank.org/knowledgebase/articles/378833-how-are-the-income-group-thresholds-determined

World Bank (n.d.). *DataBank | World Development Indicators.* [online] Available at: https://databank.worldbank.org/source/world-development-indicators