University of Economics, Prague

Faculty of Informatics and Statistics

REGRESSION ANALYSIS ON TESTS RESULTS FOR DIABETES DIAGNOSIS USING R

MASTER THESIS

Study programme: Quantitative Methods in Economics

Field of study: Quantitative Economic Analysis

Author: Bc. Gayrat Dadamirzaev

Supervisor: Ing. Karel Helman, Ph.D.

Prague, June 2020

Declaration

I hereby declare that I am the sole author of the thesis entitled “Regression analysis on tests results for diabetes diagnosis using R”. I duly marked out all quotations. The used literature and sources are stated in the attached list of references.

In Prague on …….

Signature

Bc. Gayrat Dadamirzaev

Acknowledgement

I hereby wish to express my appreciation and gratitude to the supervisor of my thesis, Ing. Karel Helman, Ph.D., for valuable advices and practical suggestions during the time of writing this thesis. I very much appreciate for profound belief in my research.

I would like to express my gratitude to all professors of QEA program for provided extensive knowledge in the field of quantitative methods of economic analysis and statistics.

My deepest gratitude to professors, whose courses inspired me to choose the topic of given thesis, to prof. Ing. Josef Arlt (Time series), Mgr. Milan Basta, Ph.D.(Regression), Ing. Zdenek Sulc,Ph.D.(Applied Multivariate Statistics), Ing. Tomas Formanek, Ph.D.(Advanced Econometrics) and other professors.

I would like to extend my sincere appreciation to coordinator of QEA program, Mgr. Veronika Brunerova, for unwavering support and professional guidance.

I'm extremely grateful to my family for everything.

Abstract

In multivariate analysis, such as multiple linear regression, unusual points in dataset may influence on fitting of regression model, i.e. may influence on overall estimation results of model and statistical significance of coefficients. Especially, when dataset is full of outliers, it is a question how to deal with such outlier points in order to avoid violations of regression assumptions and to keep model statistically significant. This master thesis is aimed to explore and answer to this question by using classical linear and predictive approach. Theoretical part of this thesis starts with introduction, where it is described the main idea and tasks of research as well as steps of implementation of assigned tasks. Theory of the thesis will introduce basic linear regression terms, such as determination coefficient, test of coefficients significance, regression assumptions and issues like multicollinearity, outliers, etc. Methods of outlier detection and predictive analysis are some of the most main topics in this thesis. Practical part is realization of thesis targets based on terms and methods, which are mentioned in theoretical part of the thesis. Using official R Studio software, four different scenarios or situations with data will be discovered individually and then will be compared between each other. As a conclusion, based on the outcomes of four identical regression models, the quality of these models will be determined.

Keywords

Classical linear regression, outliers, influential outliers, residual diagnostics, outlier detection, prediction

Content

Introduction 8
Theoretical part 11
1.1 Regression and correlation. 11
1.2 Correlation. 11
1.3 Regression. 13
1.4 Intercept and Slope. 15
1.5 Error term. 16
1.6 Regression matrix notation. 16
1.7 OLS(Ordinary Least Squares) 18
1.8 Regression assumptions. 19
1.9 Types of regression. 20
1.10 Predictive analysis. 21
1.11 Multicollinearity. 25
1.12 OUTLIERS 27
Practical part. 35
2.1 Introduction. 35
2.2 Initial data analysis. 37
2.3 CASE 1. NO OUTLIERS, LEVERAGES, INFLUENTIAL POINTS 40
2.4 CASE 2. NO EXTREME VALUES. 49
2.5 CASE 3. OUTLIERS INCLUDED. 52
2.6 CASE 4. NO INFLUENTIAL OUTLIERS. 56
Conclusion. 66
List of reference 69

List of Figures

Figure 1. Correlation graph. 12
Figure 2.Correlogram. 13
Figure 3. Simple regression fitted line. 14
Figure 4. Multiple regression fit (two independent variables, Income is dependent variable) 15
Figure 5.Boxplot and 3 sigma graph 28
Figure 6. Graph of leverage, dicrepancy and high influence 29
Figure 7. (Skewed normal distribution with normally distributed outliers) 29
Figure 8. Residual vs Fitted. 30
Figure 9. Residuals vs Leverage 30
Figure 10. Cook's distance influential points graph 31
Figure 11. View of diabets dataset. 37
Figure 12. Boxplots of all variables. 39
Figure 13. Distribution of "chol" 40
Figure 14. Qqplot of "chol". 40
Figure 15. Corrplot of variables 41
Figure 16. Residuals vs Fitted (Case 1) 43
Figure 17. Normal Q-Q plot (Case 1) 44
Figure 18. Scale Location graph (Case 1) 45
Figure 19. Residuals vs Leverage(Case 1) 45
Figure 20. Plots of residuals diagnostics(Case 2) 50
Figure 21. Residuals vs Fitted (Case 3) 52
Figure 22. Normal Q-Q plot of residuals (Case 3) 53
Figure 23. Scale-Location plot (Case 3) 53
Figure 24. Residuals vs Leverage(Cook's distance) (Case 3) 54
Figure 25. Residuals vs Leverage (Case 4) 56
Figure 26. Cook's distance, before removing influential points (Case 4) 57
Figure 27. Cook's distance, after removing influential points (Case 4) 58
Figure 28. Cook's distance, different repressentation(Case 4) 58
Figure 29. Residual diagnostic plots after removing three influential points (Case 4) 59
Figure 30. Bonferroni correction (Case 4) 61
Figure 31. Bonferroni correction and Studentized residuals (Case 4) 62
Figure 32. Mahalanobis distance scatter and qq plot(Case 4) 64

List of tables

Table 1. Introduction table 10
Table 2.ANOVA table 23
Table 3.Summary table 37
Table 4.More detailed summary table 38
Table 5. Different tests for multicollinearity 43
Table 6. T test of coefficients significance 46
Table 7. F test of significance 46
Table 8.These results are for first 10 patients. 47
Table 9.T test for Case 2 50
Table 10.F test for Case 2 51
Table 11. Predicted and actual values(Case 2) 51
Table 12. T test for Case 3 54
Table 13. F test for Case 3 55
Table 14. Actual and Fitted values for Case 3 55
Table 15. Patients with influential unusual values by Cook's distance 57
Table 16.Actual and Fitted by Cook's distance 60
Table 17.Bonferroni correction influential outliers 61
Table 18.Actual and Fitted values by Bonferroni correction 61
Table 19.13 outliers above upper cutoff: 62
Table 20.12 outliers below lower cutoff: 63
Table 21.Actual and Fitted values by Studentized residuals 63
Table 22.Outliers by Mahalanobis distance 64
Table 23.Actual and Fitted by Mahalanobis distance 65
Table 24. Methods comparison of Case 3 67
Table 25. Comparison of all cases 68

Introduction

The relevance of selecting given topic is explained by high popularity in statistics and it plays a tremendous role in quantitative analysis since 19th century.

The regression analysis is used to combine the practical methods of investigation of regression dependency between values(variables) on some data suitable for statistical analysis. Regression analysis is widely used in such statistical applications like Econometrics, Time series, Insurance and in every science or sector (e.g. business) where panel data, cross sectional data or time series data are available. Especially, in today's trends like Big Data and Machine learning, regression analysis is indispensable method that used for estimation of numeric value of parameters of regression function.

Building a regression function is inception of regression analysis. Therefore in theoretical part of my thesis, I will give an explanation what should be considered before building a model, as well as illustrate it on example.

This master thesis is the research based on data analysis by introducing to reader the main concept of regression and application of classical linear regression analysis on the real data. Following the data science principal there are six types of analysis: descriptive, exploratory, inferential, predictive, causal and mechanistic. This thesis includes descriptive, exploratory, inferential and predictive analysis.

The theoretical part or research outline will cover detail description of the process used in practical part, that consists of the term's explanation, steps to build the model, conducting the test hypothesis for statistical significance of coefficients and interpreting the final results.

The fundamental goal of the thesis and essential aim of practical part is to detect outliers and their influence on the model's fit and prediction accuracy. We will deal with several metrics of detection outliers, such as Cook's distance, Studentized residuals, Bonferroni correction and others.

In practical part we will investigate four cases or forms of dataset, where we either keep or delete outliers.

1st case: Remove outliers, leverages and influential points

We will employ predictive regression model on dataset, from which all possible unusual points were extracted.

2nd case: Exclude all extreme outliers.

We will remove outliers, which values are extreme, and leave the rest outliers with dataset.

3rd case: Conversely to 1st case.

None of outlier will be removed from dataset.

4th case: Remove only influential outliers

The regression model for each case will be consider the same.

We will use only one linear regression model for all cases, which will be considered in the beginning of analysis.

We will start with initial data analysis, i.e. general data overview, then we'll continue with regression data analysis (cross validation, checking weak set of assumptions) for each of cases. We'll predict values and calculate accuracy metrics like mean square error, mean absolute error, Akaike Information Criteria and others, for individual case. Finally, we'll make comparison of four cases outcomes between each other in order to make constructive conclusion.

The data will be randomly split into two parts: train and test. This separation is used to prevent the regression model from overfitting. The model will be first trained on train dataset, then the same trained model will be used to forecast dependent variables values for test dataset to determine how regression model works on unobserved data by comparing the percentage of results accuracy between train and test models.

An overall structure of the thesis is given by Introduction, brief theoretical background, case study and empirical research, interpretation of terms and results used in practical part and final conclusion.

Official R studio software is used for calculations the results of predictive regression analysis and for their graphical representations.

Table 1. Introduction table

Main parts	Key subjects	Summary
1. Theory	Statistical overview on terms	Covers brief explanation of basic statistical terms and terms used in regression analysis
2. Empirical Research	Plan and strategy Data preparation Initial data analysis	Theoretical discussion on how to begin with practical analysis. Data observation and start of empirical research
3. Solution of the task	Correlation analysis Regression analysis Fitting model Prediction Model quality and accuracy	Main practical part. Demonstration of calculation, graphs and values using R studio. Defining a fitted model, applying methods, training and testing data.
4. Results	Description of process Interpretation of results	Theoretical portrait of practical part, results overview and its meaning.
5. Conclusion	Key findings	Report of what was gained by thesis outcome and author's personal opinion.

Theoretical part

For this part it was considered to introduce the notion of basic terms and topics which are directly related to practical part. We will focus mainly on idea of linear predictive regression analysis, i.e. model quality, model accuracy, etc., and in addition, we will go through description of assumptions of classical linear regression, residual diagnostics, regression issues as well as methods to fix them and so forth.

1.1 Regression and correlation.

Aim of data analysis is to answer the research questions, for instance the research question for given thesis is: "What and how different will be regression results of different data, data containing outliers and data without outliers. Full-fledged data analysis consists of two parts: descriptive and inferential statistics. Descriptive statistics is discovering data, for e.g., checking data quality by determining the amount of observations, missing values, outliers, distributions plots, calculation of mean and variance, etc. It prepares data for further or main part of data analysis. Inferential statistics testing sample data and based on it makes conclusion about population, for e.g. hypothesis testing, regression analysis, prediction, etc.

There are three types of analysis: univariate, bivariate and multivariate analysis.

Univariate: analysis of data with one variable. (graphs, frequency table, etc.)

Bivariate: analysis of data with two variables. It includes contingency table analysis, mean comparison (ANOVA, t-test), correlation analysis and simple linear regression. (Z.Sulc, 2019)

Multivariate: analysis of data with more than two variables. It is similar to bivariate analysis, but instead of simple regression, multiple regression is employed.

Thus, this thesis research is made by multivariate data analysis.

1.2 Correlation.

When we talk about regression or correlation, we keep in mind some association or some relationship between two or more phenomenon. If there is a relationship, then there is a dependency.

Dependency, in terms of two variables, means how variation of one variable may depend on variation of another variable. Such dependency of pair variables is called correlation.

Correlation is a part of bivariate analysis that describes the extent of pair-wise dependency between variables. The measure of correlation coefficient is known as Pearson correlation rho:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma X \sigma Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)}{\sigma X \sigma Y} \; ; \; \rho_{X,Y} \in < -1; 1 > \qquad (1.1)$$

Cov(X,Y) - covarince between variables X and Y

σ is standard deviation

μ is mean

Positive and negative values of correlation coefficients depend on the type of relationship between two variables, i.e. linear negative or linear positive as on Figure 1 below.

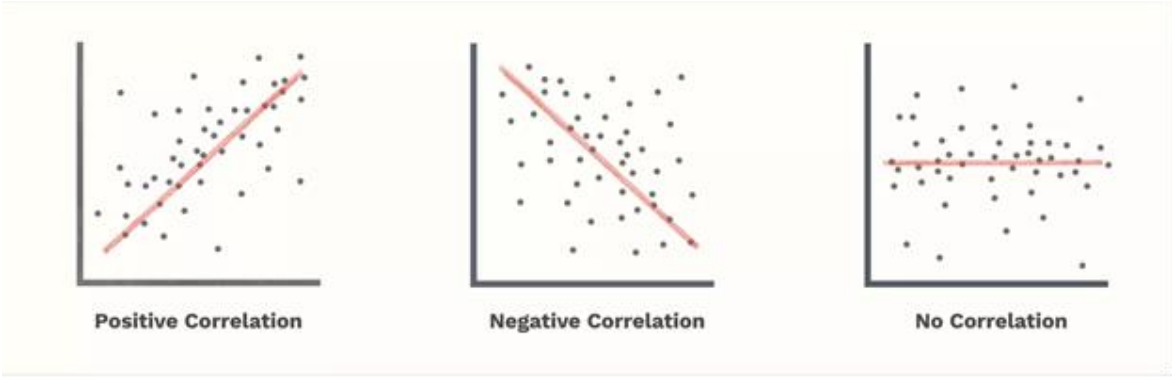


Figure 1. Correlation graph.

When *ρ=0,* or when there is no correlation between variables, then the line will be horizontal.

If *ρ=-1 or ρ=1* it indicates perfect correlation, consequently if *ρ* is closer to 0 indicates weak correlation and if *ρ* is closer to 1 - high correlation.

Alternative measure for non-parametric correlation (e.g. don't follow the assumption of normality or normal distribution) are Spearman's rho and Kendall's tau. The difference is that Pearson rho uses only continuous values of X and Y, while for non-parametric method ordinal and continuous values. Pearson rho is more efficient, because it has information about the mean and standard deviation, while other measures use only ranks and scores of pairs. This research is presumed to use only Pearson's correlation coefficient for main analysis.

In case if it is necessary to identify correlation coefficient for more than one pair, then it is useful to draw up a correlation table. This correlation table is also called as **correlation matrix**. Each variable is compared with all other variables to see the strength of correlation. In official R software, by using "corrplot" package it is possible to draw interactive tables, called correlogram. Figure 2.

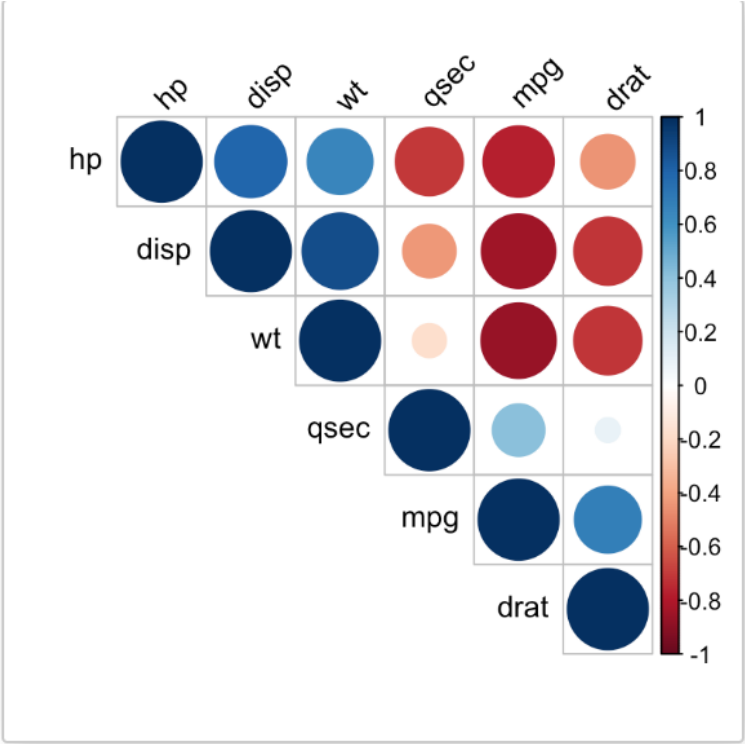


Figure 2.Correlogram.

By this correlation matrix one may conclude which variables are worth to use for further analysis (e.g regression), and which are rather to exclude, because they might cause a **multicollinearity.** This important term will be discussed further under multicollinearity topic.

Thus, correlation is useful and important part of preliminary analysis of regression.

1.3 Regression.

“Regression analysis is the hydrogen bomb of the statistics arsenal.” (Charles Wheelan, American professor)

Regression is one of the most essential tools of statistical analysis that evaluates the relationship between one and another or combination of variables, as well as allows to make prediction of future values of variable based on this relationship. It explains how one variable can be changed or explained by one or set of variables. The variables in regression equation are called independent and dependent variables. Dependent variable is the one which we would like to analyze (e.g. estimate, predict), therefore such variable in regression equation can be only one and located on the left size of equation. Independent variables can be one for simple regression and more for multiple regression. Such variables, consequently, located on the right size of equation.

$$y = \beta 0 + \beta 1 * x1 + \beta 2 * x2 + \beta 3 * x3 + \cdots + \beta i * xi + \varepsilon \quad (1.2)$$

y is dependent variable

x_i is i^{th} independent variable

β_i is regression coefficients

β_0 is intercept and ε is error term.

In this thesis we will use names “response”, “target” for dependent variable and denote it as Y and for independent – “regressor”,” predictor”, “parameter”, “explanatory variable” and sometimes just “variable”, we denote as X.

Task of regression analysis is to build such fitted model that could be used for relationship explanation or prediction. The fitted model called regression equation or regression function is being constructed in order to construct the fitted line of regression function. Fitted line is a line of best fit, being used for further estimation of parameters of regression equation. These parameters are the keys and are the objective that are necessary to reaching aim of regression analysis.

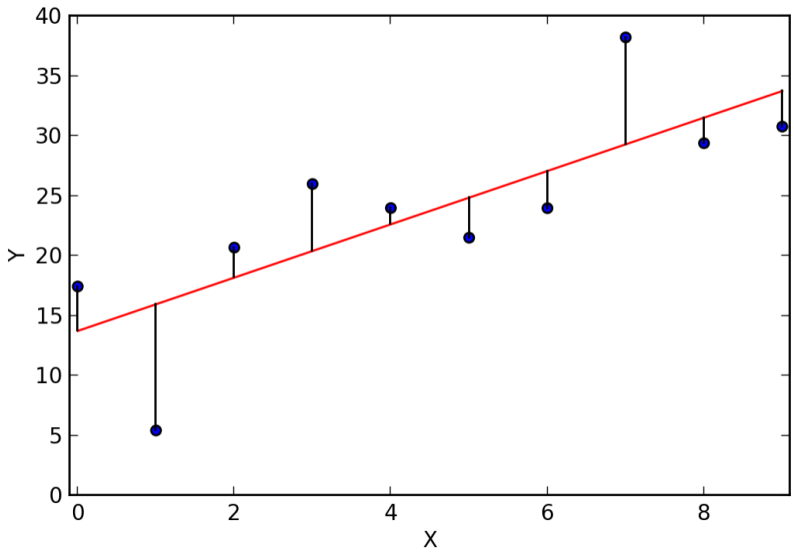


Figure 3. Simple regression fitted line.

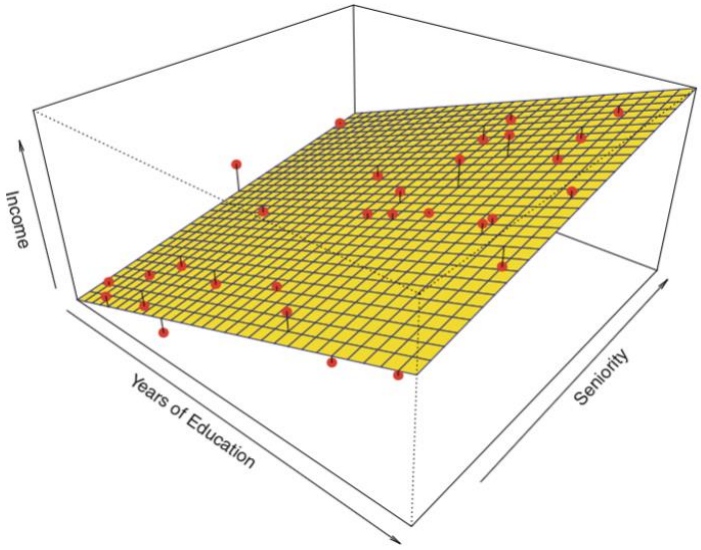


Figure 4. Multiple regression fit (two independent variables, Income is dependent variable)

1.4 Intercept and Slope.

Fitted line is illustration of relationship that best explain dependent variable. The distance between all points(dots) and line itself should be as much minimum as it possible. Points are cross observations between values of X and Y variables.

Intercept is any constant value of Y, it's is the predicted mean of Y when X equals zero. On the graph intercept equals to the distance on Y axis between zero and value of Y, where fitted line is crossing.

$$\beta = \bar{y} - \beta_1 \bar{x} \tag{1.3}$$

$\bar{y}$ – *mean value of dependent variable,*

$\bar{x}$ – *mean value of independent variable,*

β_1 – *slope*

Slope means on how many units Y will go up or down if each X increases by one. Slope can be positive or negative, zero slope means there is no relationship between Y and X.

$$\hat{\beta}_1 = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} \tag{1.4}$$

n-sample size

The values of slope and intercept are just estimation of true relationship.

1.5 Error term.

It is important to distinguish two types of errors, irreducible and residual error. Since the regression model deals with the sample, but not with whole population error terms have following differences:

Let's first denote ***e*** as residual error, and **ε** as irreducible error.

Residual or random error is a difference between real and predicted value of regression model of the sample. It is mainly needed for calculation of residual sum of squares, which used for drawing a fitted line.

$$e = y - \hat{y} \tag{1.5}$$

y actual value

$\hat{y}$ predicted value

Sample regression line or function usually is not the same as population regression function. The **irreducible or theoretic error** is a distance between observation and population regression line, therefore it cannot be calculated and consider to be a random value.

The property of theoretic errors is that the sum of errors is equal to zero or very close to zero.

$$y = \alpha + \beta x_i + \varepsilon_i \tag{1.6}$$

$$E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2 \tag{1.7}$$

α intercept

β slope

ε_i theoretical error term

σ standard deviation

1.6 Regression matrix notation.

Regression matrix is convenient way to calculate multiple linear regression, since the aim is estimation of coefficients, multiple regression has more coefficients than simple regression, because of number of predictors.

Similar to standard linear regression equation, matrix form model has the following:

$$Y_{n\times 1} = X_{n\times k}\beta_{k\times 1} + \varepsilon_{n\times 1} \tag{1.8}$$

$Y_{n\times 1}$ - *vector of* $n \times 1$ *dimension*

$X_{n\times k}$ - *vector of* $n \times (p+1)$ *dimension*

$\beta_{k\times 1}$ - *vector of* $(p+1) \times 1$ *dimension*

$\varepsilon_{n\times 1}$ - *vector of* $n \times 1$ *dimension*

$(p+1)$ - *because of intercept*

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, and\ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

n-number of observations of **k**th explanatory variable.

1 in matrix X denotes intercept.

To estimate of coefficients such as slope and intercepts it is needed to use transpose and inverse matrices of matrix X.

$$(X^T X)\hat{\beta} = X^T y \tag{1.9}$$

X^T -is transpose matrix.

Then,

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{1.10}$$

$X^T X$ is assumed to be invertible

Residual sum of squares is calculated as following:

$$RSS = e^T e$$

$$RSS = (y - \hat{y})^T (y - \hat{y})$$

$$RSS = (y - X\hat{\beta})^T (y - X\hat{\beta})$$

$$RSS = (y^T - X^T\hat{\beta}^T)(y - X\hat{\beta})$$

$$RSS = y^T y - y^T X\hat{\beta} - X^T\hat{\beta}^T y + X^T\hat{\beta}^T X\hat{\beta}) \tag{1.11}$$

$\hat{\beta}$ *is vector of coefficients that also includes intercept*

y^T *is transposed vector(matrix) of* y

1.7 OLS(Ordinary Least Squares).

The main method for linear regression to estimate the population parameters is Ordinary Least Squares method. Aim of OLS is to find such **a**(intercept) and $\mathbf{b_i}$(slope) that will minimize the sum of squared error $\mathbf{e_i}$:

$$\sum_{i=1}^{n} e^2 = \sum_{i=1}^{n}(Y_i - a - bX_i)^2 \tag{1.12}$$

$$b = \sum_{i=1}^{n} \frac{(X_i-\bar{X})(Y_i-\bar{Y})}{(X_i-\bar{X})^2} \qquad a = \sum_{i=1}^{n} \bar{Y} - b\bar{X}_i \tag{1.13}$$

a-intercept,

b-coefficient,

e-residual

$\bar{X}$*-average of X*

Since the residuals can get positive or negative values (because the fitted line is located between the points) and the sum of them will deliver to zero by canceling out each other, then in order to get some value, calculation of the residual sum of squares is needed.

Sum of squared errors (SSE) measures variability of dependent variable Y unexplained by the estimated regression function, it is unexplained deviation from point on fitted value to the real one. The smaller value of error, the closer model fits to data. That is why SSE tends to be minimized.

$$SSE = \sum_{i=1}^{m}(y - \hat{y}_i)^2 \tag{1.14}$$

The expected deviation from sample mean to the point of fitted line, called the **regression sum of squares (SSR)** and has the following formulae:

$$SSR = \sum_{i=1}^{m}(\hat{y} - \bar{y}_i)^2 \tag{1.15}$$

Sum of unexplained (SSE) and explained (SSR) deviation from the mean gives the total deviation or **total sum of square (SST).**

$$SST = SSE + SSR \tag{1.16}$$

SST is a measure of the variance in the target variable**.** (David Ziganto,2018)

These metrics, SSE, SSR and SST, are used to determine the R^2, which is the determination coefficient of regression function.

Coefficient of determination.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{1.17}$$

Coefficient of determination of regression, denoted as $\mathbf{R^2}$ is squared correlation between predicted and sample (real) data of response variables. Therefore, the outcome of $\mathbf{R^2}$ lies in range [0,1].

It measures percentage of variance(variation) of predicted outcome. It shows the quality of regression fit. If R^2 =1, it indicates best regression fit or 100 percent of variation model explains, and if 0, it means regression model absolutely doesn't fit the data.

In this thesis we will use R^2 for interpreting the regression model quality.

R^2 has a fix value(percentage) of certain regression model, however with extension model by adding extra variable, R^2 will always increase, even if added variable has no meaningful data or association to response variable. In order to recognize whether added variables into model has sense and are statistically significant, there exists the notion of **adjusted $\mathbf{R^2}$**.

In other words, R^2 is a adjusted by removing an impact of degrees of freedom. Adjusted R^2 gives a quantity that is more comparable than R^2

$$Adjusted\ R^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1} \tag{1.18}$$

n-number of observations or sample size, p-number of regressors.

(Rawlings,2006)

1.8 Regression assumptions.

There are five standard assumptions of OLS that regression model should meet in order to generate unbiased coefficient estimates that better perform true population values. We check main assumptions by residuals diagnostics.

1. Linearity. Linear in parameters and coefficients (i.e. linear relationship between predictors and response variable)

2. Normality. Normal distribution of residuals.

3. Homoscedasticity of residuals or constant variance

4. independent variables and uncorrelated residuals

5. No multicollinearity

6. No influential outliers

If there is some violation in regression assumptions such as normality or homoscedasticity, then have two options:

1. Find a robust(sandwich) estimator of the variances after performing OLS regression
2. Reweight data, to get efficient estimator

Hence, some of solutions to fix violation of assumptions are:
1. Transformations (log, Box-Cox)
2. Generalized least squares
3. Sandwich estimators.

1.9 Types of regression.

Many regression types have been invented and nevertheless, and time to time new types of regression are appearing. The main objective (i.e. prediction) remains the same for all of them, but they may be differ based on distribution of response variable and error term, by type of variable, either categorical or continues, etc. In other words, there are three distinctions that make each regression technique different in most cases: number of independent variables, shape of regression line, type of dependent variable. All regression technique(model) can be used by researcher to find such a model with the best fit, i.e. most expressive model with the highest prediction accuracy.

1. Linear regression.

2. Logistic regression.

   If the dependent variable is binary, then logistic regression best suits for modelling the relationship, while independent variables can be nominal, ordinal and interval variables. It is usually case for classification tasks. In classification task the probabilities are calculated for each observation to classify whether 1 or 0. The researcher will set up the threshold, for example if probability of X observation is lower than 40% than it is classified as 0 and if it is higher than 40% or let say equal 40% then it is considered as 1. It is mainly depending on specific of target variable as well as researcher's decision, however the standard probability proportion is 50 to 50 percent.

3. Polynomial regression.

   In most real cases the regression line may not be fitted well as a straight line, it rather be curve that would fits better. Such curve can be drawn up by polynomial regression equation. If the power of independent variable(s) is higher than 1, such equation called polynomial equation:

4. Ridge regression.

   The ridge regression reduces the variance, standard errors and solves the problem of multicollinearity by adding degree of bias to regression estimates. This made by adding to OLS so called a regularization term, as known as penalty term, where weights are involved.

5. Lasso regression.

   Lasso regression is very similar to ridge regression, except few differences. The biggest difference is that the regularization term of Lasso regression is the sum of absolute values of the model parameters, not the sum of squared values as for ridge regression.

6. Elastic Net Regression is hybrid of Lasso and Ridge regression.

1.10 Predictive analysis.

Basic statistical measures of linear regression model output.

The descriptive statistics that consist from mean, variance, median, mode, upper quartile, lower quartile and etc. are, by default, primarily considered as the most important statistical indicators.

Standard error is a measure of how different mean is varying from sample to sample. Standard deviation tells how data(sample) is distributed around the mean, while standard error tells how means of different samples are distributed around the mean of all these sample means. It is important indicator that shows how precise sample estimates the population parameter, as low sample standard error is as much it is precise to population mean.

$$se = \frac{\sigma}{\sqrt{n}} \tag{1.19}$$

Where **σ** – standard deviation of sample means and n is number of samples.

In regression standard error of mean response is:

$$s(\hat{\eta}_*) = \sqrt{\sigma^2(\hat{\eta}_*)} = s\sqrt{\left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)} \tag{1.20}$$

s -standard deviation
$\hat{\eta}_*$ is an unbiased estimator of η_*(estimated mean response)

The theoretical mean of Y at a given value of X is*:*

$$E(\mu(y)|x_*) = \eta_* = \beta_0 + \beta_1 x_* \tag{1.21}$$

The confidence interval of mean response:

$$E(\mu(y)|x_*) = \hat{\eta}_* \pm t_{\alpha/2} * SE(\hat{\eta}_*) \tag{1.22}$$

where d.f. is $n - 2$, $SE(\hat{\eta}_*)$*is standard error of the fit*
Estimator of Var($\hat{\eta}_*$):

$$Var(\hat{\eta}_*) = \sigma^2\left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right) \tag{1.23}$$

x* is the value of X we are choosing to estimate Y,xi every value in dataset.

Properties:

Unbiased. Let $\hat{\beta}$ be a point estimator for a parameter β, $\hat{\beta}$ is an unbiased estimator if E($\hat{\beta}$) = β. Otherwise if E($\hat{\beta}$) ≠ β, $\hat{\beta}$ is said to be biased. The bias of a point estimator $\hat{\beta}$ is given by:

$$B(\hat{\beta}) = E(\hat{\beta}) - \beta. \quad (1.24)$$

Consistent. As sample size n -> ∞, then. $\hat{\beta}$ ->β

Efficiency. The lower variance of the distribution of the estimator V($\hat{\beta}$) is better, indicating that in repeated sampling a majority of values of $\hat{\beta}$ will be "close" to β.

The mean square error MSE of a point estimator $\hat{\beta}$ is $MSE(\hat{\beta})$. Thus, MSE is a function of both its variance and its bias.

$$MSE(\hat{\beta}) = E\left[\left(\hat{\beta} - \beta\right)^2\right] = V(\hat{\beta}) + [B(\hat{\beta})]^2 \quad (1.25)$$

Prediction of new response.

The theoretical value of Y for a given value of X is:

$$\hat{y}_* = \hat{\eta}_* = \beta_0 + \beta_1 x_*$$

The prediction interval for a single value of Y at x* is:

$$\hat{y} \pm t_{\alpha/2} * SE\ of\ prediction(\hat{y})$$ where d.f. is n – 2

Estimated variance of predicted value:

$$Var(\hat{\eta}_*) = \sigma^2\left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right) + \sigma^2 \quad (1.26)$$

Estimated standard error

$$s(y_* - \hat{y}_*) = \sqrt{s^2(y_* - \hat{y}_*)} = s\sqrt{\left(1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)} \quad (1.27)$$

Properties of the error of prediction.

- Confidence interval of mean response Y is narrowest than confidence interval of prediction value.
- Confidence interval of mean response Y is the narrowest in the interval at the mean of X, and c.i. goes wider when X is far from mean.
- As sample size increase, confidence interval decreases.

ANOVA (Analysis of Variance). ANOVA table is the statistics of regression output that consists values of SSR (Sum of squares total), SSE (Sum of squares error) to form two mean squares for regression (treatment) model (MST) and for error term (MSE). ANOVA table represents statistics to test hypothesis about population mean.

ANOVA table.

Source of variability	Sum of squares	df	Mean squares (MS)	F statistic
Regression	SSR	k-1	MST=SSR/(k-1)	F=MST/MSE
Error	SSE	n-k	MSE=SSE/(n-k)	
Total	SST	n-1		

Table 2.ANOVA table

F statistic is used in **F test** to provide information whether any regression equation is statistically significant.

Hypothesis testing for F test:

1. State the null and alternative hypotheses:
   H0: $\beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$
   H1: $\beta_j \neq 0$
2. Compute F statistic:

$$F = \left(\frac{SSE(R) - SSE(F)}{df_R - df_F}\right) / \frac{SSE(F)}{df_F} \quad (1.28)$$

$$SSE(R) = \sum(y_i - \bar{y})^2$$

$$SSE(F) = \sum(y_i - \hat{y})^2$$

$$df_R = n - 1$$

$$df_F = n - 2$$

3. Reject H0 if $F > f_{\alpha,1,n-2}$, where $f_{\alpha,1,n-2}$ is percentile of F distribution (taken from F table)
   α is significant level, usually 0.05 is used, n-sample size.

Test to check hypothesis whether coefficients are statistically significant is t test.

1. State the null and alternative hypotheses:
   H_0: $\beta_j = 0$
   H_1: $\beta_j \neq 0$
2. Calculate T statistics:

$$T = \frac{\widehat{\beta_j}}{se(\widehat{\beta_j})} \quad (1.29)$$

$$se(\beta) = \sqrt{\frac{\frac{\sum_{i=1}^{n} {e_i}^2}{n-2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \quad (1.30)$$

3. If $|T| \leq t_{n-p-1,1-\alpha/2}$, then we not reject H_0

n-number of observations,

p-number of parameters,

α significance level,

t-critical value (percentile of t distribution)

Regression metrics.

Evaluation of regression performance in terms of prediction is driven by error **metrics**. Given that the task of OLS is minimizing error term(squared), correspondingly the best value for error metrics is lowest value. Thereby obviously, that metrics help with decision making of choosing single best model from many potentially appropriate models.

There are various metrics deal to evaluate the output of model prediction accuracy.

MAE is the Mean Absolute Error, defined as mean of absolute difference between actual target values and predicted. MAE is reliable to outliers, only if they don't require a special attention, because all individual differences are weighted prorate.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (1.31)$$

MAPE is Mean Absolute Percentage Error, same as MAE, but measured in percentage.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \quad (1.32)$$

MSE-is Mean Squared Error, defined as average of squared difference between actual target values and predicted. As error growth quadratically, it penalizes the outliers.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (1.33)$$

RMSE, Root Mean Squared Error, is the most widely used metric. It is the standard deviation of prediction errors.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{1.34}$$

AIC and BIC.

Akaike and Bayesian (Shwarz) Information criteria are based on information theory and Kullback–Leibler divergence, that intended to evaluate likelihood of model's forecasting by selecting optimal model from set of possible options. In other words, It used to find out which model will predict better with less or larger number of predictors (model complexity), or to determine which certain combination of predictors will have higher accuracy of matching fitted values with values of dependent variable.

Model may be considered not sufficiently good if it:

1. Doesn't predict well
2. Has large number of predictors, i.e. is complex

In linear regression, AIC and BIC penalize RSS (since if model not fitting well, then RSS is large) and the number of predictors *k,* by following formula:

$$AIC = n\ln\left(\frac{RSS}{n}\right) + 2k \tag{1.35}$$

$$BIC = n\ln\left(\frac{RSS}{n}\right) + \ln(n)k \tag{1.36}$$

k is the number of regression parameters including intercept.

n-number of observations.

RSS is residuals sum of square.

Since AIC and BIC are relative measure, it is not necessarily to get absolute value. The lower AIC or BIC, the higher model quality or accuracy.

1.11 Multicollinearity.

Multicollinearity may occur in multiple regression. Once we get confidence with the selection of model explanatory variables, which were considered by their theoretical relevance along with empirical analysis (test of significance), it is necessary to check multicollinearity of those selected independent variables. Term "multicollinearity" refers only for independent variables of regression, meaning that there exists **high correlation** between two or more variables. If Pearson correlation coefficient is higher than 0.8, it indicates "severe" multicollinearity, while one indicates "perfect" multicollinearity. Both cases are issue, because it reduces the exactness of regression coefficient estimates, p values

of statistical significance are no longer reliable, which leads to incorrect interpretation about mean response change due to unit change of individual variable, while other variables are constant. It also leads that null hypothesis in hypothesis testing is fail to be rejected due to high variance of parameter estimators and eventually it will give large standard errors of linear model parameters.

Detection and dealing.

There is not a single way how to detect multicollinearity. The simple way of detecting is just removing any variable from the model and then check the standard error of remaining parameters. If the standard error was not significantly reduced, then we may assume that this variable does not have multicollinearity, otherwise it implies that this variable has multicollinearity with another variable.

Correlogram.

More widely used approach is creating a correlation matrix, also called correlogram. By formula below, we calculate correlation for each pair of variables and represent values in of *k x k* symmetric matrix form with diagonal values equal to 1. k-number of variables.

R=3x3 matrix, where $\rho_{xy} = \frac{cov(x,y)}{\sigma_x * \sigma_y}$, where cov(x,y)-covariance of x and y, σ_x and σ_y are standard deviation of x and y, where $\rho_{xy} \in$ <-1;1>

$$\mathbf{R}= \begin{matrix} 1 & \rho_{1,2} & \rho_{1,3} \\ \rho_{2,1} & 1 & \rho_{2,3} \\ \rho_{3,1} & \rho_{3,2} & 1 \end{matrix} \quad (1.37)$$

The cutoff point of correlation coefficients is 0.5, meaning that the value of coefficients can be at most 0.5, otherwise if it is higher and considered to be a multicollinearity. However, in most cases it is adopted to use cutoff of 0.8 for multicollinearity detection, meaning that if there are value or values of ρ_{xy} that equal to 0.8 or higher, it implies multicollinear relationship. In such case, from pair of multicollinear variables we decide to choose one that better suits (due to constructive reasons) to regression model, and we remove another one from the model. If we have more than two variables in the model that are highly correlated, we will continue to remove each from pair until we get a correlation matrix with "admissible" values of correlation coefficients.

Variance inflation factor

Correlation matrix is limited, because even if there is small correlation between pair of predictors, it is possible that one variable might have a high correlation within linear combination of two or more variables. Therefore, for the sake of clarity, we use variance inflation factor (VIF). VIF is a classical approach of multicollinearity detection. As we mentioned above, the variance of estimated coefficients is high in presence of multicollinearity, but in fact the variance is "inflated" by factor because of correlation among variables. Each independent variable in multiple regression may have inflated variance, consequently VIF.

If VIF is equal to 1, it means that variance of estimated coefficient is not inflated, therefore there is no correlation with combination of remaining regressors. Those independent variables with high VIFs than admissible threshold, are considered to have a high collinearity with remaining independent variables.

If VIF is higher than 4 and less than 10, further analysis is needed to determine how to deal with such variables. In case if VIF exceeds 10, then it indicates that there is high collinearity.

The formula of VIF calculation:

$$VIF_i = \frac{1}{1-R^2{}_i} \quad (1.38)$$

where $R^2{}_i$ *is determination coefficient of regression of* i^{th} *independent variable with combination of remaining independent variables. (e.g.* $X_2=X_1+X_3+X_4+e$*)*

Farrar-Glauber test.

Multicollinearity can be tested by Farrar-Glauber test. This consists of three tests, such as **chi-square test**, **F test and t test.**

1.Chi square test is used for detection the presence of multicollinearity.

2.If Chi square test detected multicollinearity, then F test is used to identify (locate) which

3. If F test is positive, then t test is used to test for the pattern of multicollinearity (variables that makes multicollinearity).

Dealing with multicollinearity.

The most popular technique to fix multicollinearity is by removing variable, that has high intercorrelation. This means for correlation matrix we remove one variable from pair of variables that have high correlation coefficient. For VIF, we remove one that has high VIF factor.

Other methods for dealing with this issue is applying so called Principal Component Analysis (PCA), Partial Least Squares (PLS) and Ridge regression.

1.12 OUTLIERS.

Discrepancy *or* ***outliers*** are extreme or unusual values of some vector or data. In terms of regression analysis, outliers refer to extremely high or extremely low point values of dependent variable. For independent variables, these extreme values called *leverage points(observations).*

By boxplot approach below, circle points are considered to be outliers, whereas point labeled as a star is extreme outlier.

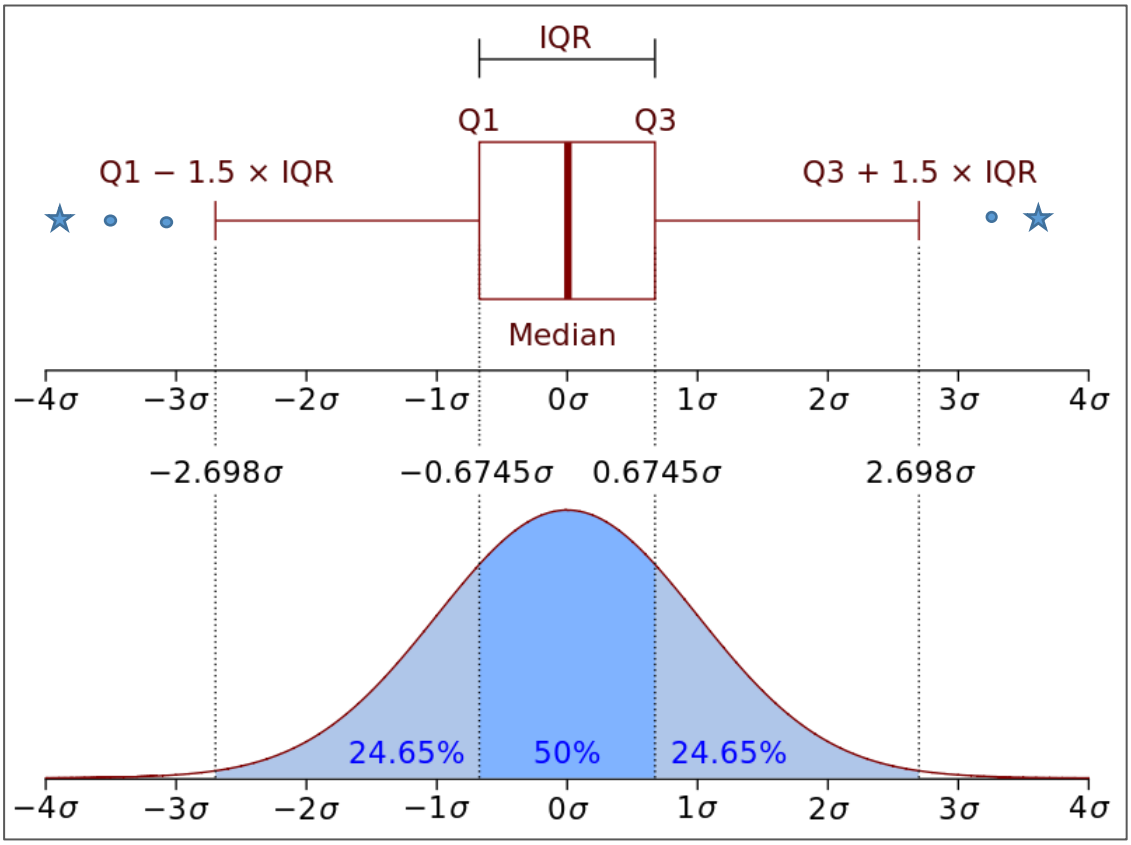


Figure 5.Boxplot and 3 sigma graph

The range where the points considered to be outliers is following:

$$< \tilde{x}_{0.25} - 1.5IQR, \tilde{x}_{0.25} - 3IQR > \cup < \tilde{x}_{0.75} + 1.5IQR, \tilde{x}_{0.75} + 3IQR > \quad (1.39)$$

IQR is interquartile range, $\tilde{x}_{0.25}$ lower quartile, $\tilde{x}_{0.75}$ upper quartile.

Similarly, the range where extreme points are considered to be extreme outliers, following:

$$(\tilde{x}_{0.25} - 3IQR, -\infty) \cup (\tilde{x}_{0.75} + 3IQR, +\infty) \quad (1.40)$$

Extreme outliers can be also determined by Three sigma rule, using mean and standard deviation of the data:

$$(x < \mu - 3\sigma;\ x > \mu + 3\sigma), \quad (1.41)$$

where x is extreme value. (Z.Sulc,2019)

Outliers and leverages may or may not affect to results of regression analysis. If after removing outlier or leverage, the regression fitted line changes and the residual standard errors are likely smaller than before, then these points are tending to be *influential points*.

Therefore, some of them may stay in data, but it is necessary to analyze them, e.g. graphically (scatter plots) or calculating measures such as studentized residuals, Cook's distance and etc.

Changes of regression fitted lines due to leverages, outliers(discrepancy) and influential points.

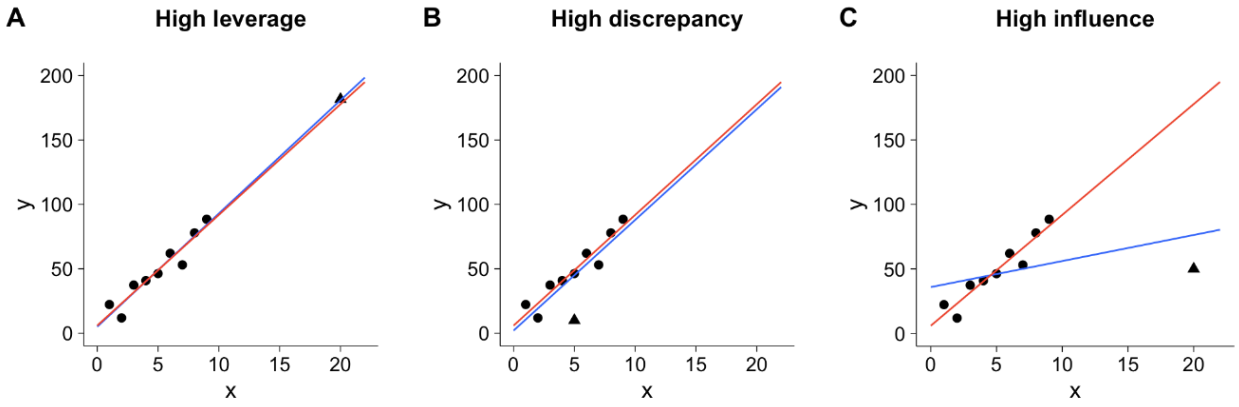


Figure 6. Graph of leverage, dicrepancy and high influence

Graphical approach.

Once we decide with regression model and we are confident that there is no high correlation between variables, we start with checking the regression assumptions of residuals, as well as outlier investigation. Distribution density or histograms of dependent or independent variable will show either there are outliers or not, like in picture#

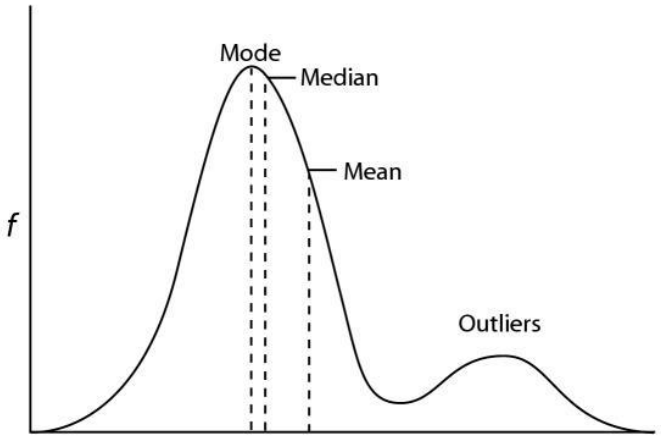


Figure 7. (Skewed normal distribution with normally distributed outliers)

In R, it is possible to quickly represent a scatter plot of residuals and standardized residuals, called "Residuals vs. Fitted" and "Residuals vs. Leverage".

"Residual vs. Fitted" is plot of residuals and regression fitted line, which gives information about linearity, homoscedasticity and **outliers**.

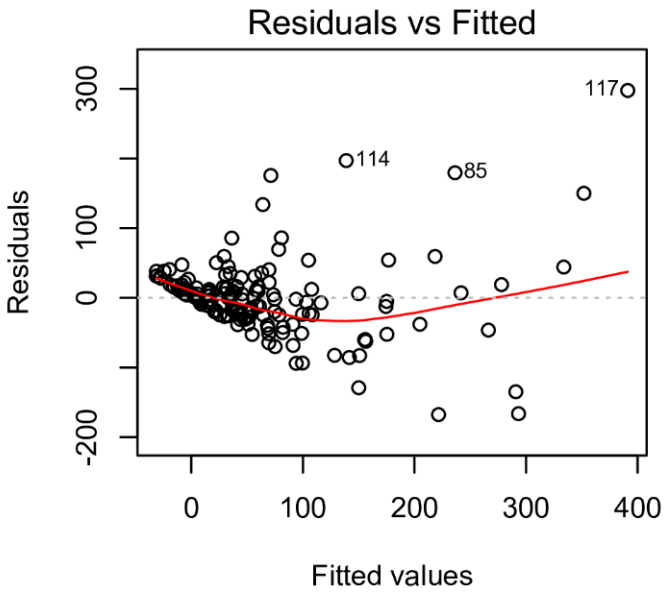


Figure 8. Residual vs Fitted.

Those residuals, that are located far from the fitted line or from majority of residuals, are supposed to be outliers. (e.g. point # 114,85,117).

"Residuals vs. Leverage"

This plot gives an information about influential leverage points that can affect to regression line. Such points are usually located inside of top right and bottom right corner, marked with red broken line, so-called *Cook's distance* line.

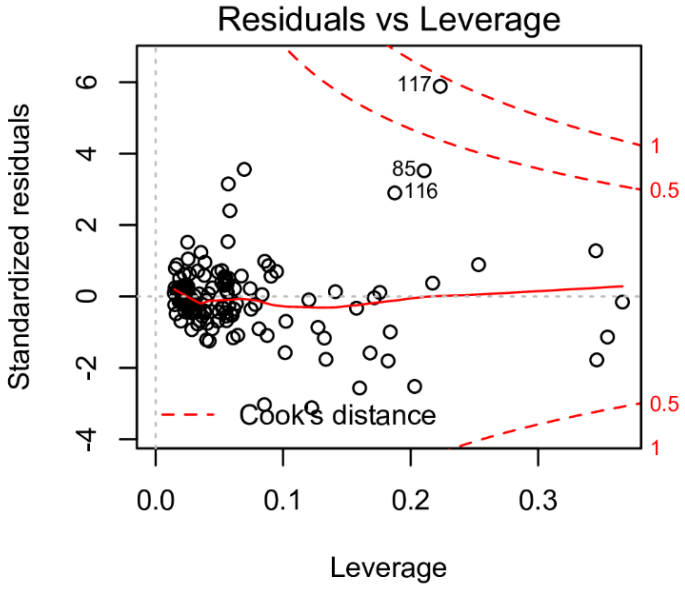


Figure 9. Residuals vs Leverage

In this picture we see one point inside of Cook's distance corner, that means that leverage point #117 is influential.

Detection of influential points based on calculation approach.

Cook's distance is a statistical measure that determines the effect (influence) of leverage, not all outlier or leverage can be influential as we mentioned before. If we check regression estimation coefficient with and without point, and if we get different coefficient values, then we expect large Cook's distance measure.

$$D_i = \frac{e_i^2}{p(\frac{1}{n-p})\sum_{j=1}^{n}\hat{\varepsilon}_j^2(1-h_{ii})}\left[\frac{h_{ii}}{(1-h_{ii})}\right] \tag{1.42}$$

D_i is a cook distance of i^{th} observation, p is the number of regression coefficients, n is sample size, h_{ii} is hat value of matrix H and $\hat{\varepsilon}_j^2$ is squared error term.

Other graphical representation of Cook's distance may look following:

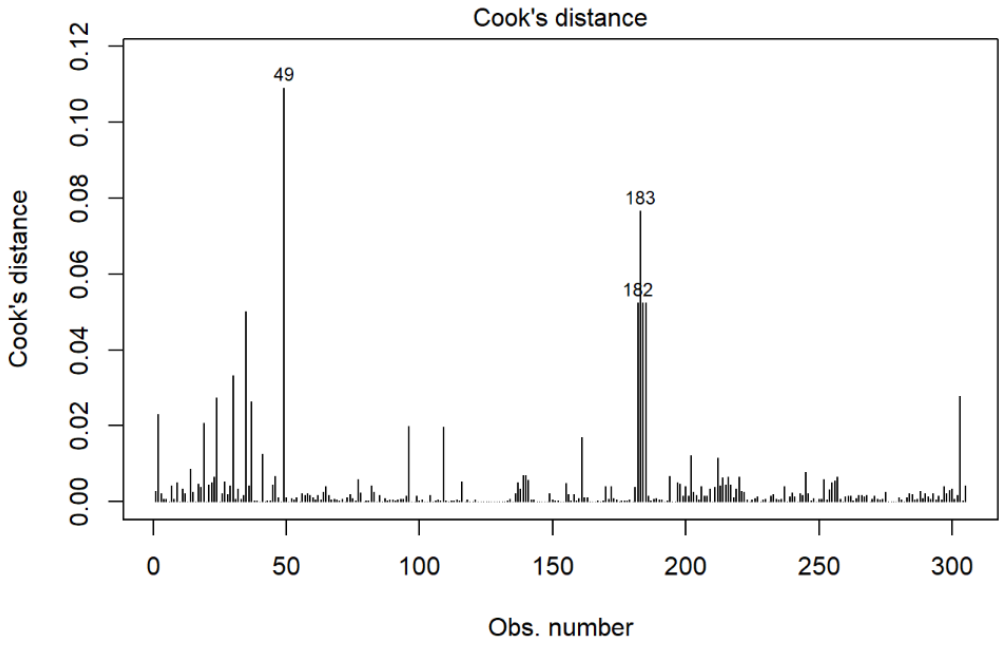


Figure 10. Cook's distance influential points graph

Influential observation or large Cook's distance is, if it satisfies following condition:

$$D_i > \frac{4}{(n-p-1)} \tag{1.43}$$

For the graph above we draw cutoff line and then we identify which observations has high cook's distance.

Hat values.

The measure of leverage also known as *hat values*. It measures the distance between point value of independent variable and the mean of the values for all data points.

For simple regression it can be calculated as:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} ; \frac{1}{n} \leq h_{ii} \leq 1 \quad (1.44)$$

For multiple regression we calculate **H** matrix by following formula:

$$H = X(XTX) - 1XT \quad (1.45)$$

$$0 \leq h_{ii} \leq 1$$

Since, it is projection matrix, it puts some values(hat) on the observed response vector to get the predicted response vector:

$\hat{y} = Hy$ (1.46)

Potentially influential or extreme leverages, identified as 2 (sometimes 3) times higher than the mean of the diagonal elements of **H:**

$$h_{ii} > 2(\frac{p}{n}) \quad (1.47)$$

where *p* is number of coefficients, *n* number of observations.

Studentized residuals.

Influential points are attracting regression line to itself, therefore the observed value of dependent variable will tend to be closer to predicted value. After removing this influential data point, fitted line will get away from the observed response, that's why it will produce a *deleted* residual. Studentized residuals are used to detect outliers of dependent variable. It is a measure of *deleted* ith residual, based on estimated standard error. Those data points that represent high deleted residuals are regarded as influential.

$$t_i = \frac{e_i}{\sqrt{MSE_i(1-h_{ii})}} \quad (1.48)$$

Where, e_i deleted residual, MSE is ith mean squared error, h_{ii} is hat values or leverages.

If $|t_i| > 3$, then data point is outlier.

Bonferroni correction.

Since, for detection outliers, we would like to get highest absolute value of studentized residual of *n* test statistic, it is not proper to use t statistic *t with n-k-2 degrees of freedom* to determine p value for largest absolute studentized residual without refitting model each time for each observation. Bonferroni correction is adjustment of p value for the largest studentized residual, that perform simultaneous inference of taking the largest n test statistics as well as taking into consideration the dependency of residuals. Correction of p value can be achieved by multiplication of two tailed p value by n tests.

$$p=2p'n, \quad (1.49)$$

p is Bonferroni corrected p value for testing significance of maximal residual.

$$p' = Pr(t_{n-k-n} > |r_{max}|, \quad (1.50)$$

r_{max} *is maximal studentized residual*

Difference in Fits.

It is similar principle as measure of studentized residuals, where we delete observations once and then we refit regression with the remaining observations. Then we compare the results with n and n-1(deleted ith observation) observations to check what is extent of influence of ith observation.

$$DFFITS_i = \frac{(\hat{y}_i - \hat{y_{(i)}})}{\sqrt{MSE_{(i)} h_{ii}}} \quad (1.51)$$

$\hat{y}_i$ is predicted response value, where ith observation is included

$\hat{y}_{(i)}$ is predicted response value, where ith observation is excluded

Cutoff value for DFFITS is equal to $2\sqrt{\frac{k+2}{n-k-2}}$, *where k is the number of predictors, n-sample size*.

Consequently, absolute value of DFFITS that higher than cutoff value indicates as influential data point.

Mahalanobis distance and Bonferroni correction.

Mahalabonis(D2) distance assumes multivariate distributions (if *k>1)* with multivariate outliers. D2 is a measure of how far each data point is from the multivariate centroids(means) of all predictors. It relies on estimation of density or standard deviation of each distribution of variable.

$$\Delta_i^2 \equiv \tilde{x}_i^T S_{XX}^{-1} \tilde{x}_i \quad (1.52)$$

$\tilde{x}_i$ is sample mean of predictors

S_{XX} *is sample covariance matrix*

The region of the means of explanatory variables, which has constant Mahalanobis distance, have ellipsoid shape.

Points that cross the ellipse line or out from ellipse are potential multivariate outliers. As cutoff value we can take a quantile from chi-square distribution (e.g 99.9%). If MD is larger than the cutoff with *degrees of freedom = k* at a critical value .001, then exist a potential multivariate outlier(s).

Robust regression.

Robust regression can be used for data with outliers. It is different from standard OLS regression only in a way that it considers weights for each observation calculated by e.g., Huber or Bisquare function and etc. The principal property of using weights is that the absolute residuals goes down when the weights go up (max 1) and vice versa, when residuals goes up, weights go down (min 0). The weights of residuals for OLS regression are equal to one. In robust regression, if the majority weights will be close or equal to one, then we can assume that results (estimates) of robust regression will be closer to results of OLS regression.

Dealing with outliers.

Once we identified outliers, the question is how to deal with them.

1. Leave them in data, whether they are not influential.

2. Remove them, if they are not so important (both influential and non-influential)

3. Replace them with NA values, if NA exist in dataset.

4. Use log transformation of variables (dependent and independent)

5. Use outlier robust regression.

6. Enhance sample size

In the next section, i.e. in practical part we will use different techniques of dealing with unusual data points.

Practical part.

2.1 Introduction.

In practical part we will have classical linear regression analysis by practicing topics of theoretical part. Furthermore, we will investigate regression analysis taking into account outliers as well as ignoring them. Therefore, in order to conduct an efficient analysis and to achieve reliable results, it was necessarily to find such dataset that will be suitable for the investigation of research goal, namely, such dataset that will contain some outliers. That is why real dataset, named "*diabetes*", of Dr. John Schorling from Department of Medicine, University of Virginia School of Medicine, has been chosen as optimal dataset for given research, because almost each of indicator (variable), fortunately, has a lot of outliers.

"*Diabets"* dataset consists of 403 observed patients among different ages, gender and location (Buckingham, Louisa) with test results and physique (i.e. height, weight, etc.) for detection presence of diabetes illness. These following tests and physical indicators, we can call both of them as variables, are cholesterol, glycosylated hemoglobin, cholesterol/HDL ratio, waist, hip and many others. There are 19 variables in total. For our analysis we will use less of them, according to the reasons that will be mentioned further.

There are also amount of missing data (NA) almost for each variable and each patient. Such variables like, "bp.2s"(Second Systolic Blood Pressure)and "bp.2d" (Second Diastolic Blood Pressure) were initially decided to be removed from dataset due to huge volume of missing data. More than half of patients do not have values for both indicators. (65% , or for 262 patients data is missing) For the remaining dataset we will remove those patients that have at least one "NA" value in any variable that we will include in regression model.

According to the journal "Heathline", "diabetes and high cholesterol often occur together". Respectively, "cholesterol" will be selected as our variable of interest. However, our goal is not to determine either patient has diabetes or not, but to check how reliably our constructed regression model will perform results and how precisely it is capable to predict. So, the main goal of analysis is to make prediction of the cholesterol values by given other characteristics and calculation of prediction accuracy of regression for different cases.

Generally, there are four cases that we will focus on:

1st case: Remove outliers, leverages and influential points

We will employ predictive regression model on dataset, from which all possible unusual points were extracted.

2nd case: Exclude all extreme outliers.

We will remove outliers, which values are extreme, and leave the rest outliers with dataset.

3rd case: Conversely to 1st case.

None of outlier will be removed from dataset.

4th case: Remove only influential outliers

By calculation of Cook's distance, Mahalanobis distance, studentized and other residual measures, we will identify and delete influential points(unusual patient's results) from dataset, then employ regression model.

We will use only one linear regression model for all cases, which will be considered in the beginning of analysis.

We will start with initial data analysis, i.e. general data overview, then we'll continue with regression data analysis (cross validation, checking weak set of assumptions) for each of cases. We'll predict values and calculate accuracy metrics like mean square error, mean absolute error, Akaike Information Criteria and others, for individual case. Finally, we'll make comparison of four cases outcomes between each other in order to make constructive conclusion.

Official R studio software is used for calculations the results of predictive regression analysis and for their graphical representations.

2.2 Initial data analysis.

Let's look on dataset:

	chol	stab.glu	hdl	ratio	glyhb	location	age	gender	height	weight	bp.1s	bp.1d	waist	hip
1	203	82	56	3.6	4.31	Buckingham	46	female	62	121	118	59	29	38
2	165	97	24	6.9	4.44	Buckingham	29	female	64	218	112	68	46	48
3	228	92	37	6.2	4.64	Buckingham	58	female	61	256	190	92	49	57
4	78	93	12	6.5	4.63	Buckingham	67	male	67	119	110	50	33	38
5	249	90	28	8.9	7.72	Buckingham	64	male	68	183	138	80	44	41
6	248	94	69	3.6	4.81	Buckingham	34	male	71	190	132	86	36	42
7	195	92	41	4.8	4.84	Buckingham	30	male	69	191	161	112	46	49
8	227	75	44	5.2	3.94	Buckingham	37	male	59	170	*NA*	*NA*	34	39

Figure 11. View of diabets dataset.

Our dataset contains 14 columns(variables), we exclude other columns because they are considered as non-essential for further analysis (id, time, …) and 403 rows(lines) which are patients.

We have different data types in our data.

Discrete:
chol(cholesterol), **stab.glu**(Stabilized Glucose), **hdl**(High Density Lipoprotein),_**age**(years), **height**(inches), **weight**(pounds), **bp.1s**(First Systolic Blood Pressure), **bp.1d**(First Diastolic Blood Pressure), **waist**(inches), **hip**(inches)

Continuous: **ratio**(Cholesterol/HDL Ratio), **glyhb**(Glycosolated Hemoglobin)

Categorical: **location**(Buckingham, Louisa), **gender**(male,female)

Since we have categorical data with two factors, we will recode them as 1 and 2. For gender:1-male,2-female, for location: 1-Buckingham, 2-Louisa.

We can check summary statistics of individual variable by summary table that consists of moments and quantiles (i.e. mean, median, IQR. etc.)

```
      chol          stab.glu          hdl             ratio           glyhb       location      age       
 Min.   : 78.0   Min.   : 48.0   Min.   : 12.00   Min.   : 1.500   Min.   : 2.68   1:200    Min.   :19.00  
 1st Qu.:179.0   1st Qu.: 81.0   1st Qu.: 38.00   1st Qu.: 3.200   1st Qu.: 4.38   2:203    1st Qu.:34.00  
 Median :204.0   Median : 89.0   Median : 46.00   Median : 4.200   Median : 4.84             Median :45.00  
 Mean   :207.8   Mean   :106.7   Mean   : 50.45   Mean   : 4.522   Mean   : 5.59             Mean   :46.85  
 3rd Qu.:230.0   3rd Qu.:106.0   3rd Qu.: 59.00   3rd Qu.: 5.400   3rd Qu.: 5.60             3rd Qu.:60.00  
 Max.   :443.0   Max.   :385.0   Max.   :120.00   Max.   :19.300   Max.   :16.11             Max.   :92.00  
 NA's   :1                       NA's   :1        NA's   :1        NA's   :13                               
 gender      height          weight          bp.1s           bp.1d           waist            hip       
 1:200   Min.   :52.00   Min.   : 99.0   Min.   : 90.0   Min.   : 48.00   Min.   :26.0   Min.   :30.00  
 2:203   1st Qu.:63.00   1st Qu.:151.0   1st Qu.:121.2   1st Qu.: 75.00   1st Qu.:33.0   1st Qu.:39.00  
         Median :66.00   Median :172.5   Median :136.0   Median : 82.00   Median :37.0   Median :42.00  
         Mean   :66.02   Mean   :177.6   Mean   :136.9   Mean   : 83.32   Mean   :37.9   Mean   :43.04  
         3rd Qu.:69.00   3rd Qu.:200.0   3rd Qu.:146.8   3rd Qu.: 90.00   3rd Qu.:41.0   3rd Qu.:46.00  
         Max.   :76.00   Max.   :325.0   Max.   :250.0   Max.   :124.00   Max.   :56.0   Max.   :64.00  
         NA's   :5       NA's   :1       NA's   :5       NA's   :5        NA's   :2      NA's   :2      
```

Table 3.Summary table

The table above is important statistics, which have to be calculated in the first place, because without knowing locations and variability of vector(variable), we are not able to conduct any analysis.

NA=missing value

We see that most of variables have “NA” values and if we count them, we receive 36 “NA”s in total. Now, we inspect how many patients have at least one “NA” value for given columns by following command:

```
nrow(data[rownames(data[!complete.cases(data), ]),])
```

26 patients have at least one missing value for all set of variables.

There are several options how to treat with “NA”s, some of them are:

1. Replace by mean value or any random number that located in interquartile range
2. Replace by outliers
3. Regress them (i.e. predict)
4. Delete whole raw that contain it

If we have a small sample size, it is rather to consider a value to replace them. In case, if we have a large sample size and number of rows that contain missing value is not relatively large, then we can delete them. Our patients are 403 in general, patients with some “NA” in result are 26, so we may assume that after getting rid lines with “NA”s from dataset, will not significantly change our final results. Consequently, we delete them and get a new dataset with 377 patients, without any missing value.

Next table is description table, more advanced, which contain more details then previous one.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
chol	1	402	207.85	44.45	204.00	204.93	37.06	78.00	443.00	365.00	0.92	2.54	2.22
stab.glu	2	403	106.67	53.08	89.00	94.54	17.79	48.00	385.00	337.00	2.75	8.10	2.64
hdl	3	402	50.45	17.26	46.00	48.52	14.83	12.00	120.00	108.00	1.19	1.93	0.86
ratio	4	402	4.52	1.73	4.20	4.36	1.63	1.50	19.30	17.80	2.20	13.17	0.09
glyhb	5	390	5.59	2.24	4.84	5.11	0.83	2.68	16.11	13.43	2.23	4.98	0.11
gender*	6	403	1.50	0.50	2.00	1.50	0.00	1.00	2.00	1.00	-0.01	-2.00	0.02
height	7	398	66.02	3.92	66.00	65.98	4.45	52.00	76.00	24.00	0.03	-0.21	0.20
weight	8	402	177.59	40.34	172.50	174.81	37.81	99.00	325.00	226.00	0.72	0.67	2.01
bp.1s	9	398	136.90	22.74	136.00	134.87	20.76	90.00	250.00	160.00	1.10	2.38	1.14
bp.1d	10	398	83.32	13.59	82.00	82.97	11.86	48.00	124.00	76.00	0.27	0.04	0.68
waist	11	401	37.90	5.73	37.00	37.58	5.93	26.00	56.00	30.00	0.47	-0.17	0.29
hip	12	401	43.04	5.66	42.00	42.61	4.45	30.00	64.00	34.00	0.80	0.83	0.28

Table 4.More detailed summary table

Table 4 represents also skewness, kurtosis and standard errors, so we can imagine how distribution may look like. For example, “chol” seems to be Normally distributed, but skewed and platykurtic, because standard Normal distribution has 0 skewness(symmetry) and kurtosis equals 3.

After making quick overview of statistics and fixing issue with missing values, we may start to display boxplots to see outliers and then start with Case #1.

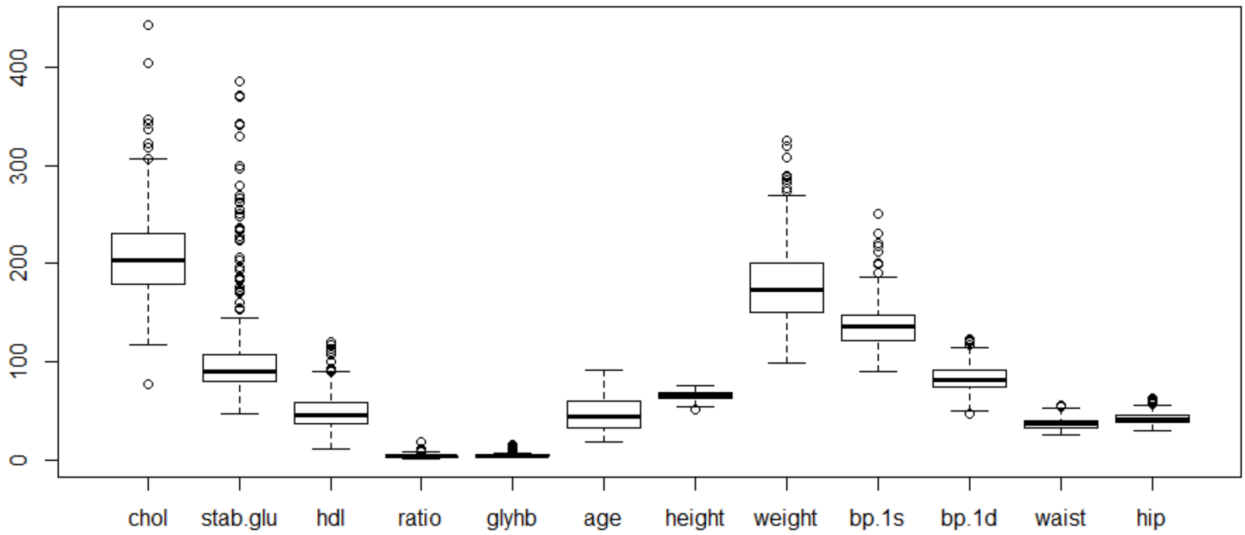


Figure 12. Boxplots of all variables.

Boxplots (Figure 12) shows that except "age", all variables contain number of outliers, some have fewer and some have larger. It is visible that "stab.glu" has extremely high and many unusual points, whereas "height" or "bp.1d" has very few and which are very close to the certain minimum or maximum point value. It may be interpreted in the way that these outliers are not much different than the other values of particular variable.

For instance, if we run R command:

```
data$chol[which(data2$chol %in% boxplot.stats(data2$chol)$out)]
```

we will get vector of numbers (78, 443, 318, 347, 342, 404, 307, 337, 322). This numbers are outlier values of "chol".

2.3 CASE 1. NO OUTLIERS, LEVERAGES, INFLUENTIAL POINTS

We will start with looking-over two assumptions of classical regression, normality of dependent variable and multicollinearity. Actually, the regression assumption is normality of regression residuals. However, since we don't have regression model yet, we may preliminary check our dependent variable to be Normally distributed, because if it is not Normally distributed, then it is hard to get normality of residuals.

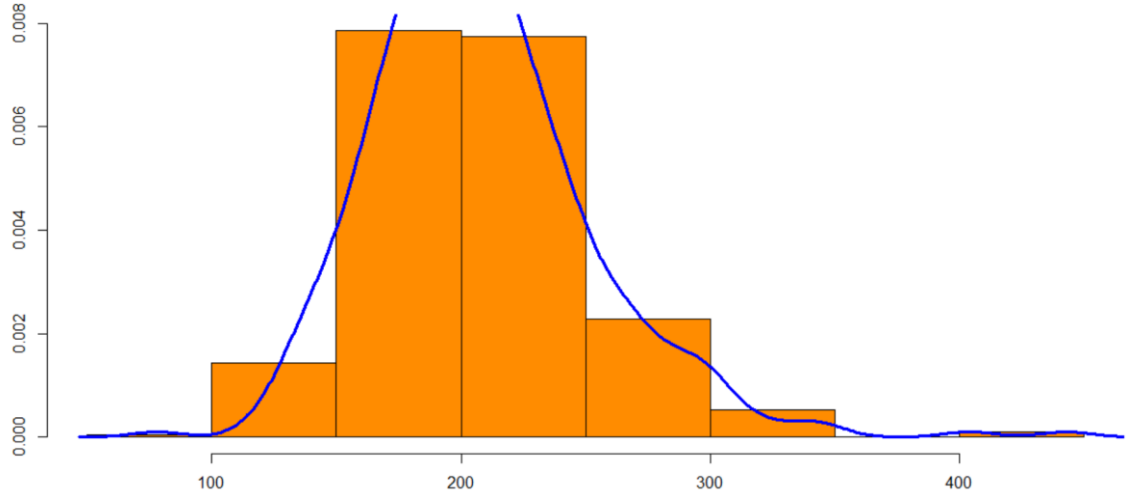


Figure 13. Distribution of "chol"

The distribution of "chol" is approximately Normal and it is noticeable that right tail is longer than in standard Normal distribution and a little skewed, this is due to values that are different from majority of values in dataset (i.e. outliers). We may suppose that without outliers, distribution may look more symmetrical with normal tails.

Similarly, it is possible to check normality by qqplot, below.

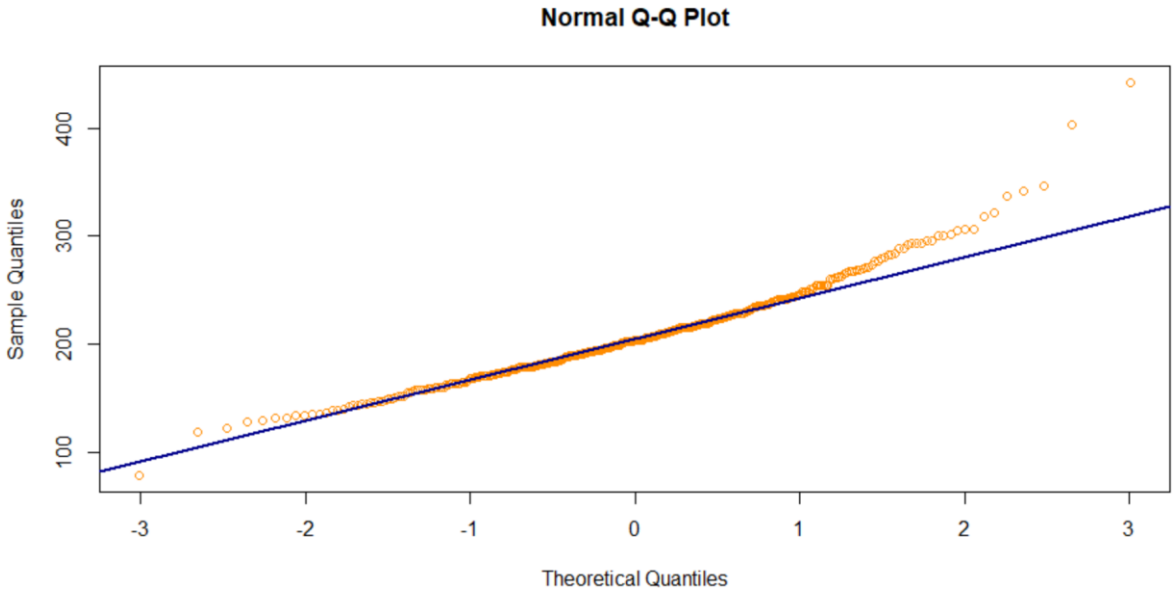


Figure 14. Qqplot of "chol".

Correlogram. By plotting correlogram we will discover variables that are highly correlated with each other.

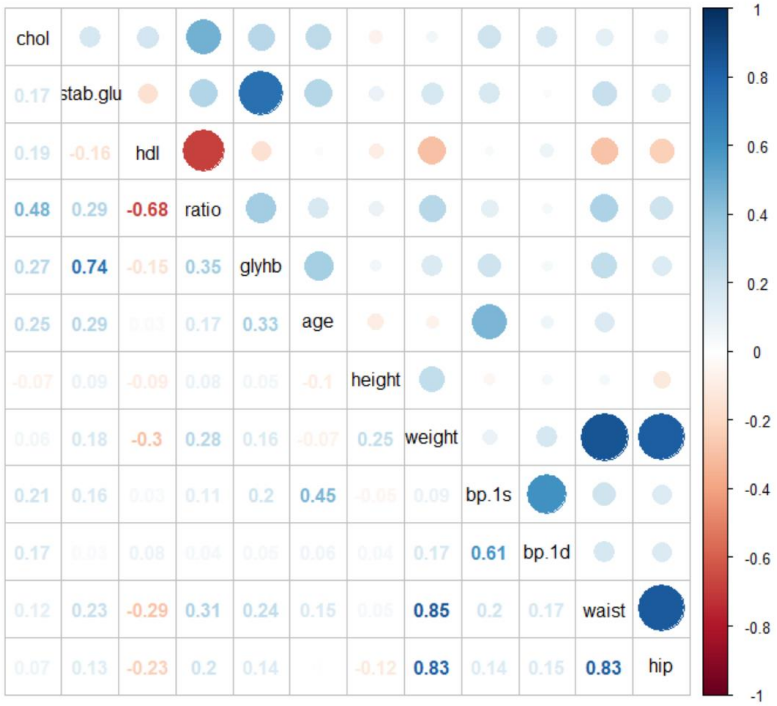


Figure 15. Corrplot of variables

On Figure 15, we have four pairs of highly correlated (>0.7) variables: (stab.glu,glyhb),(weight,waist),(weight,hip),(waist,hip).

This means that for regression analysis we can include only one variable from each pair to the model (if it is necessary). Final decision, regrading multicollinearity, we will make after VIF calculation in regression model.

Construction regression model.

There are three basic methods of selection variables: Forward selection, Backward elimination and Stepwise selection.

Forward selection: starts without any independent variable and add one by one contributive variable to the model until the improvement is no longer statistically significant.

Backward elimination: includes all independent variables in the model and then starts to eliminate one by one less contributive until the model is statistically significant. (Zdenek Sulc,2019)

Stepwise selection: starts with forward selection and then continues with backward elimination.

The aim of building good fit is to get statistically significant predictors which will minimize Akaike Information criteria or Shwarz Information criteria.

In our example we are going to use stepwise selection of predictors. In R, the function is

```
step<-stepAIC(model, direction = "both")
```
Stepwise elimination suggests the following model:

```
data$chol ~ hdl + ratio + location + height + bp.1d + hip #model1
```
Model above has minimum AIC (2379.158) from other AIC of possible variants of model and each independent variable supposed to be statistically significant.

There are no pairs of variables in the model that we considered as highly correlated based on correlation matrix. But let's now look what information Variance Inflation Factor will give us.

```
vif(model1) #location excluded
```

```
hdl    ratio   height    bp.1d      hip
1.978449 1.913873 1.036196 1.058200 1.120362
```
VIF shows that model does not suffer from multicollinearity issue, because VIF value of each variable is less than 5 and even less than 2.

The model above took into account the presence of outliers. But what changes may occur if we remove all outliers from the variables that are involved in the model? Will model be still the same and significant?

Let's write a function in R that will delete all outliers:

```
remove_outliers <- function(x) {
    R <- 1.5 * IQR(x)
    y <- x
    y[x < (quantile(x, probs = 0.25) - R)] <- NA
    y[x > (quantile(x, probs = 0.75) + R)] <- NA
    y
  }
```

Detected outliers' values of each involved variable of suggested above model will be replaced by NA values. And same like we were dealing with NAs before, we remove from dataset those patients, that will have at least one NA (in this case it is outlier) for given variables of model.

Before deletion we had 377 patients, after deletion the number of patients was decreased to 333.

We again run stepwise selection algorithm for sample size of 333.

After running corresponding command, the suggested model looks a bit different:

```
data$chol ~ hdl + ratio + height + hip #model2
```
The first important difference is lower AIC for model without outliers. It shows 1868.5, while in "model1" 2379.1. R^2 has different coefficient for both models, for first it is 0.73, for the second 0.81. According to this results, we can conclude that removed outliers contained influential data points.

VIF didn't detect any high collinearity and if we run different multicollinearity diagnostic measures, including Farrar Glauber test by

`omcdiag(model2)`
then the outcome is following:

```
Overall Multicollinearity Diagnostics

                       MC Results detection
Determinant |X'X|:         0.3787         0
Farrar Chi-Square:       320.3097         1
Red Indicator:             0.3465         0
Sum of Lambda Inverse:     6.9880         0
Theil's Method:           -1.1564         0
Condition Number:         44.8812         1

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test
```

Table 5. Different tests for multicollinearity

Only two measures from overall six, Farrar Chi-square and Condition Number, detected multicollinearity, whereas the rest four tests did not detect any collinearity. That is why we may reckon the lack of multicollinearity of "model2".

We will choose model: `data$chol ~ hdl + ratio + height + hip,` as the main model for analysis of all cases. It doesn't mean that this model will be optimal for all cases, but we will investigate the difference of regression results with regard to each case. We rename "model2" to "mymodel" just to avoid confuses.

Residual diagnostics.

By command `plot(mymodel)` we test weak set of regression assumption. Residual diagnostics are important in regression analysis, they are assumed to be normally distributed and to have a constant variance.

Linearity.

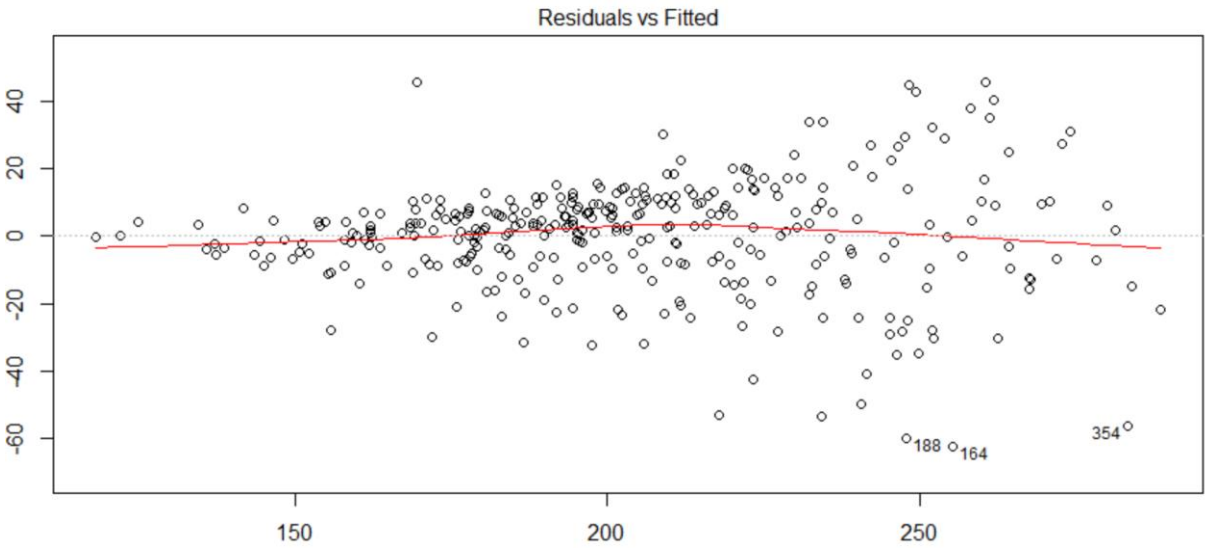


Figure 16. Residuals vs Fitted (Case 1)

Here we may conclude linear relationship between our predictors and response variable, because the line is horizontal and there is no pattern in residual plot.

Normality.

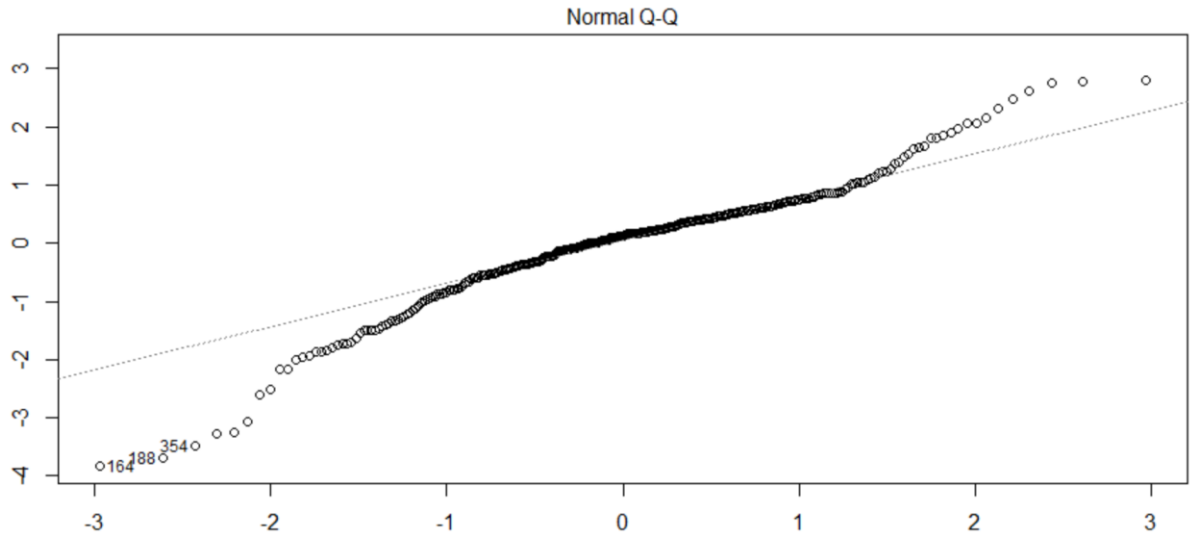


Figure 17. Normal Q-Q plot (Case 1)

If residuals follow straight line, it indicates that residuals are normally distributed. In our case, not all residuals are laying on the line, therefore they are likely having not Normal distribution.

For the sake of clarity, we will use Shapiro Wilk test for normality:

```
shapiro.test(studres(mymodel))
```

p value is less than significance level 0.05, therefore we reject null hypothesis that residuals are normally distributed.

Homoscedasticity.

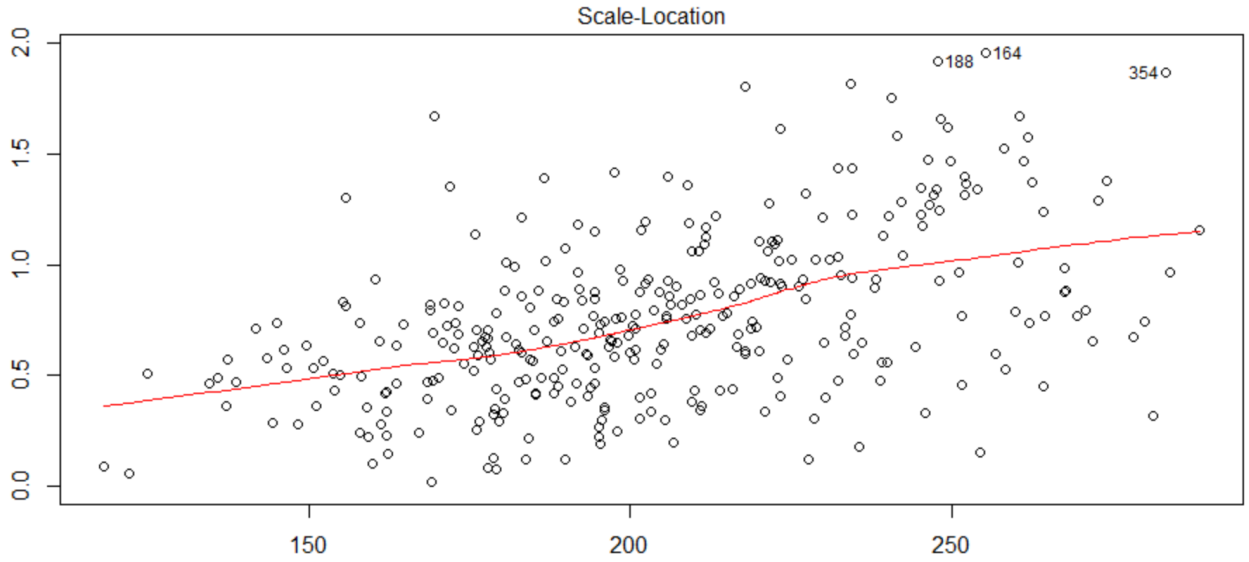


Figure 18. Scale Location graph (Case 1)

For homoscedasticity we would like to achieve fitted line in Figure 18 to be horizontal, however this is not in our case, therefore there is probably heteroscedasticity.

To make sure, whether our assumption is true, we will use non-constant error variance test and studentized Breusch-Pagan test, `ncvTest(mymodel)` and `bptest(mymodel)`, respectively.

p value of both tests is less than 0.05, we reject the null hypothesis. Our assumption on heteroscedasticity based on graph was true.

Outliers.

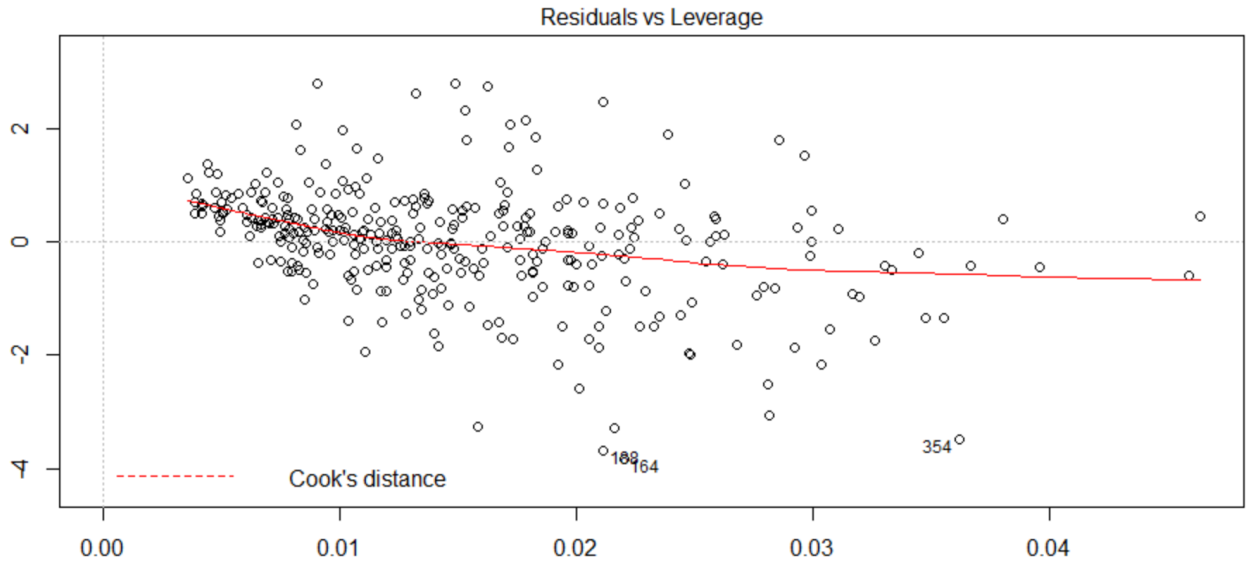


Figure 19. Residuals vs Leverage(Case 1)

Since we deleted all outliers, the graph above doesn't show any influential data points, therefore Cook's distance line and corners are missing.

Significance of regression coefficient (t test and F test)

By using t test and F test, we check whether predictors in our model are significant by 95% of confidence.

Command for t test in R: `summary(mymodel)`
Command for F test in R: `anova(mymodel)`

t test

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -81.80222   20.75692  -3.941 9.92e-05 ***
hdl1          3.11807    0.09879  31.562  < 2e-16 ***
ratio1       36.32164    0.98632  36.825  < 2e-16 ***
height1      -0.70788    0.23429  -3.021  0.00271 **
hip1          0.40768    0.19021   2.143  0.03282 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.41 on 328 degrees of freedom
Multiple R-squared:  0.814,     Adjusted R-squared:  0.8117
F-statistic: 358.9 on 4 and 328 DF,  p-value: < 2.2e-16
```

Table 6. T test of coefficients significance

F test

```
Analysis of Variance Table

Response: no1$chol1
           Df Sum Sq Mean Sq   F value    Pr(>F)
hdl1        1   7429    7429   27.5783 2.719e-07 ***
ratio1      1 375088  375088 1392.3626 < 2.2e-16 ***
height1     1   2978    2978   11.0564 0.0009839 ***
hip1        1   1237    1237    4.5937 0.0328249 *
Residuals 328  88360     269
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 7. F test of significance

In spite of regression assumption violations, both t test and F test show significance of regression coefficients ($p<0.05$) as well as significance of relationship of regression model, because p value of F statistic is less than 0.05.

In this research we will not try to adjust our models in order to meet regression assumptions, since our goal is to test and predict by models with and without affect of outliers.

Prediction analysis.

Cross validation is a technique of testing effectiveness of model prediction and model evaluation by resampling procedure with replacement. We will split our data sample into two samples, train and test samples, with proportion of 0.8 and 0.2. The data for train and test will be chosen from main dataset randomly. The proportion of splits will be the same for all cases datasets.

Cross validation in R looks following:

```
set.seed(777)
part<-sample(2,nrow(data),replace = T, prob = c(0.8,0.2))
 train<-data[part==1,]
 test<-data[part==2,]
```

Train dataset contain 273 patients and test dataset 60 patients, respectively. The idea of such split is to predict "chol" values of test dataset by learning train dataset. Then we compare how actual values of test dataset are differ from predicted values.

Prediction in R can be made by: `predict(train.ols,test, interval = "predict")`

Where "train.ols" is regression on train data.

By prediction function we will get the actual and fitted values of response variable and regression confidence intervals.

	actual	fit	lwr	upr
7	195	191	158	223
12	238	245	212	278
18	196	194	161	227
21	203	195	162	228
34	194	200	167	233
36	182	175	142	207
38	218	206	173	238
54	237	222	190	255
55	296	257	224	290
56	178	178	145	210

Table 8.These results are for first 10 patients.

Finally, we calculate accuracy metrics, such as R^2, RMSE, MSE, MAE, MAPE, AIC, BIC, that tells us the quality of the model and accuracy of model prediction.

R2	RMSE	MSE	MAE	MAPE	AIC	BIC
0.8399785	16.22736	263.3273	11.50504	0.05399865	2311.409	2333.066

This metrics, except R^2, should be as much minimum as it possible. We will use them for comparing to another metrics from other cases.

2.4 CASE 2. NO EXTREME VALUES.

Model: `data$chol ~ hdl + ratio + height + hip`

Case #2 ,#3 and #4 will be investigated reciprocally as case#1 with the initially considered model , the only difference is adjusted dataset with regard to case task.

In this case, the dataset will contain outliers, but will not contain extreme outliers.

First step is to detect which outliers are extreme (i.e. outside from $< \mu - 3\sigma, \mu + 3\sigma >$) in our dataset, for given variables by 3 sigma rule.

For detection such extreme values, we will run function detect:

```
detect <- function(x) {
   S <- 3 * sd(x)
   M <- mean(x)
   lower<-which(x < (M - S))
   cat(c("lower:",lower,"\n"))
   upper<-which(x > (M + S))
   cat(c("upper:",upper))
 }
```

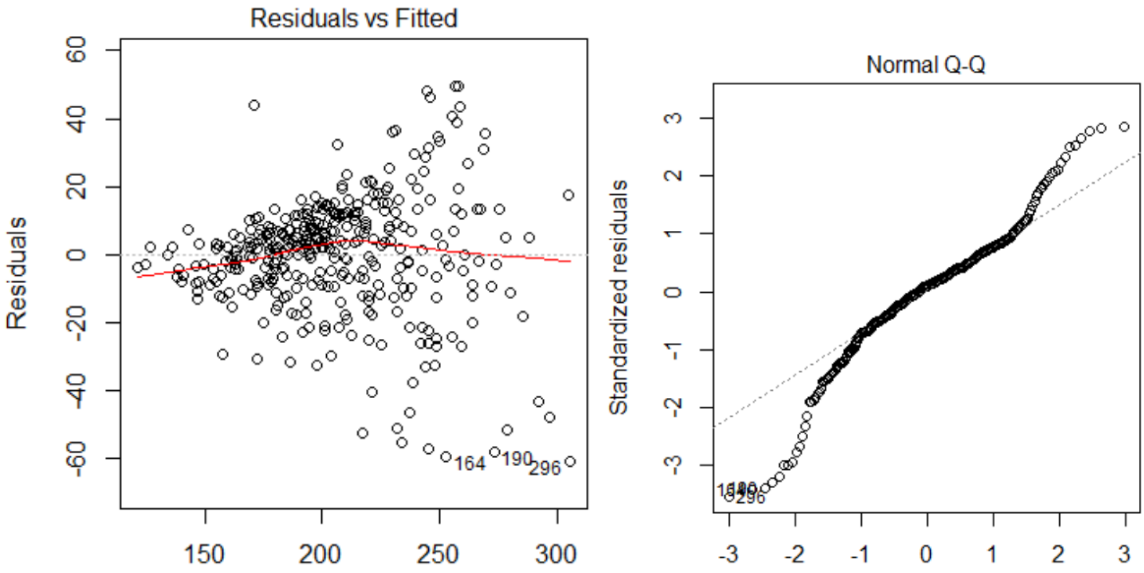
Same as for case 1, we assign to these values “NA” and by function `“complete.cases”` we remove all patient lines that has at least one “NA” (in this case extreme outlier).

We’ve got sample size 358 from 377, after removing extreme values.

Vector inflation factors are not indicating highly correlated variables:

```
  hdl           ratio         height           hip
2.348295      2.301208      1.027219      1.095676
```

Residual diagnostics.



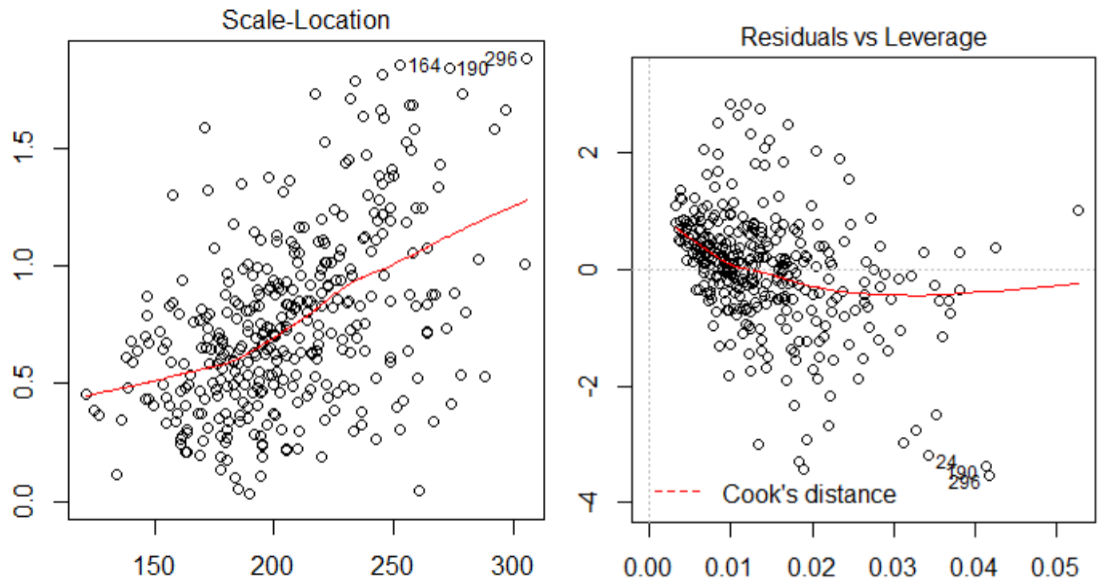


Figure 20. Plots of residuals diagnostics(Case 2)

Looking on four graphs from Figure 20, we may notice that they look pretty much the same as in case 1. Residuals supposed to be not normally distributed and it is likely heteroscedasticity.

Shapiro-Wilk test confirms that residuals have not Normal distribution.

Breusch-Pagan and Non-Constant Error Variance test also confirm non-homoscedasticity.

T test and F test, both display significance of coefficient likewise linear relationship(mode).

Test of significance.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -60.9074    21.0779  -2.890 0.004095 **
hdl           2.9637     0.0932  31.801  < 2e-16 ***
ratio        34.6269     0.9527  36.347  < 2e-16 ***
height       -0.8152     0.2429  -3.356 0.000876 ***
hip           0.4227     0.1813   2.332 0.020263 *
```

Table 9.T test for Case 2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
hdl	1	14948	14948	48.5942	1.555e-11	***
ratio	1	413252	413252	1343.4238	< 2.2e-16	***
height	1	4268	4268	13.8760	0.0002273	***
hip	1	1673	1673	5.4382	0.0202629	*
Residuals	353	108587	308			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

Table 10.F test for Case 2

R^2 has coefficient 0.779, which is lower than in first case.

Prediction analysis.

Again, we split our dataset without extreme values on train and test dataset, then predict values of “chol” from test dataset. We will get, following results for first 10 patients:

	actual	fit	lwr	upr
7	195	191	156	225
20	234	238	203	272
29	182	171	137	205
35	173	170	135	204
37	136	147	113	181
47	237	228	194	263
64	164	163	128	197
71	217	205	171	240
73	217	202	168	237
75	218	206	172	241

Table 11. Predicted and actual values(Case 2)

And final accuracy metrics:

R2	RMSE	MSE	MAE	MAPE	AIC	BIC
0.7653393	18.40617	338.787	12.73265	0.06363725	2441.613	2463.527

2.5 CASE 3. OUTLIERS INCLUDED.

Model: `data$chol ~ hdl + ratio + height + hip`

Main target of case 3 is to use dataset for regression without removing any outlier, influential points or extreme value.

In this case we have 395 patients in dataset.

According to VIF, multicollinearity is not existing:

```
  hdl        ratio     height        hip
1.926186 1.902294   1.029476    1.074994
```

Residuals diagnostics.

Linearity.

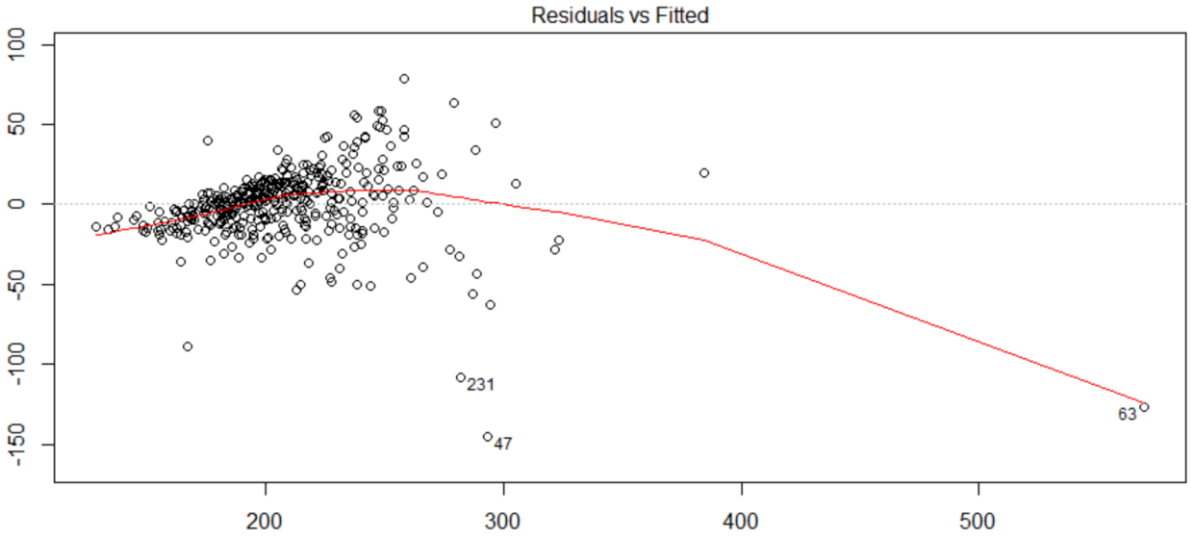


Figure 21. Residuals vs Fitted (Case 3)

The graph of Figure 21, may be accepted as no distinctive fitted pattern, i.e. relationship is still linear, however point 63 is very influential, because it pulls the line itself.

Normality.

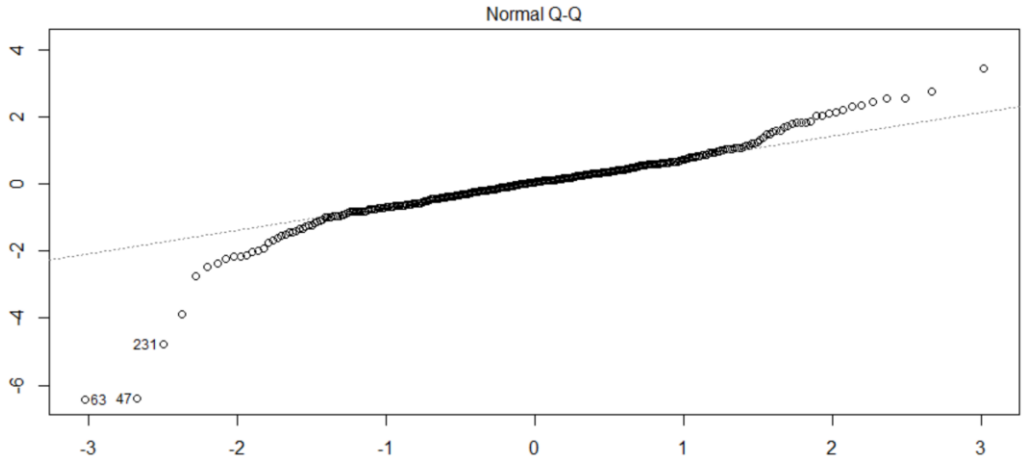


Figure 22. Normal Q-Q plot of residuals (Case 3)

The majority data points are laying on line in qqplot, it's disputable to say either Normal distribution or not. Therefore, as we did before, we check Shapiro-Wilk test:

p-value < 2.2e-16, means that residuals, do not have Normal distribution.

Heteroscedasticity.

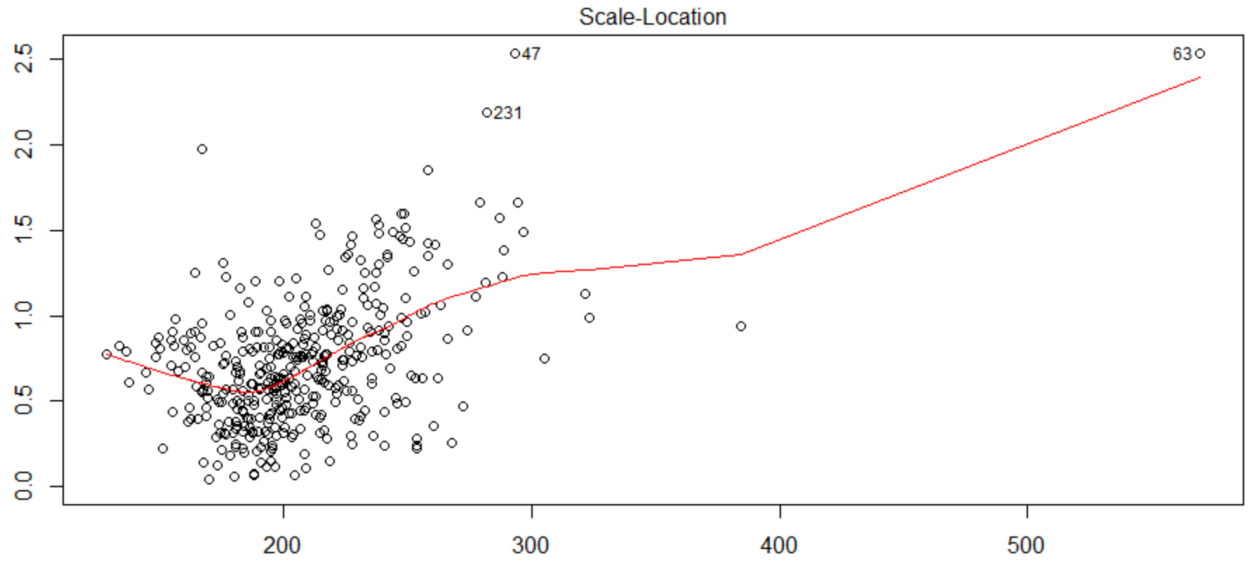


Figure 23. Scale-Location plot (Case 3)

It is obvious that residuals variance is not constant and tests like, Breusch-Pagan and non-constant error variance tests confirm it.

Outliers.

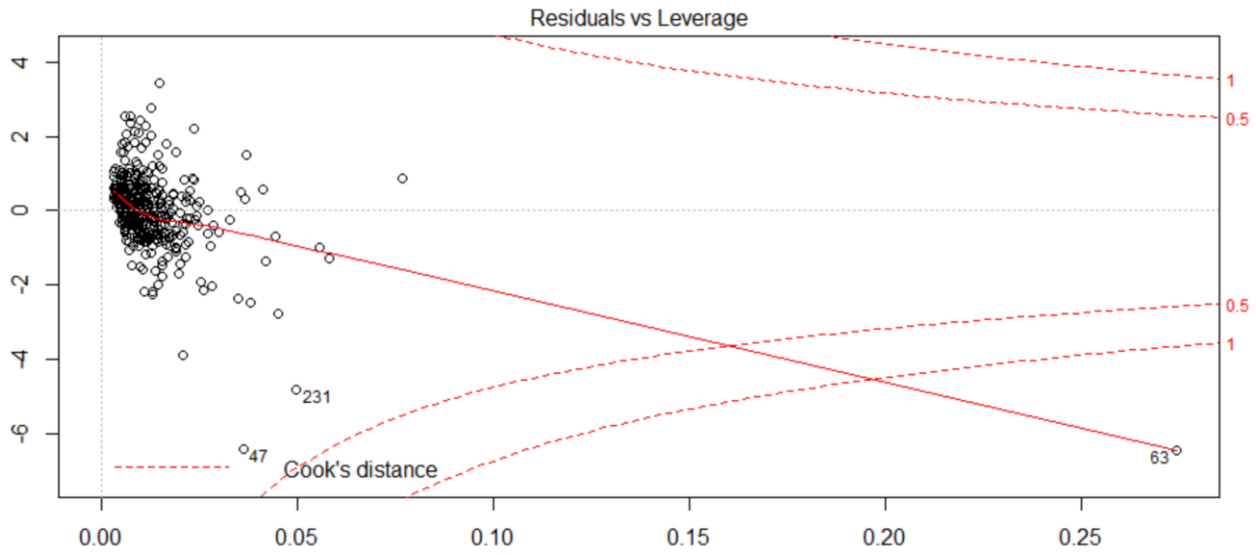


Figure 24. Residuals vs Leverage(Cook's distance) (Case 3)

The "Residuals vs Leverage" plot(Figure 24) is greatly differing from same previous graphs. As we noticed before, point 63 is problematic and influential, because it is located in Cook's distance corner. Of course, it is recommended to remove it, but not in this case, since our goal is to analyze data with presence of outliers and influential points.

Test of significance.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -18.87818   24.61714  -0.767   0.4436
hdl           2.49329    0.09325  26.737   <2e-16 ***
ratio        29.17207    0.92406  31.569   <2e-16 ***
height       -0.72818    0.30047  -2.423   0.0158 *
hip           0.39751    0.21411   1.857   0.0641 .
```

Table 12. T test for Case 3

All regression parameters and model are remaining significant, but not intercept. This might be issue, because intercept is important for estimation and prediction of response variable. Therefore, we will expect low prediction accuracy, i.e. high metrics coefficients or criteria.

Surprisingly, R^2 is 0.7321, it is almost the same as R^2 in case 2, but less low.

```
Response: data$chol
            Df Sum Sq Mean Sq  F value    Pr(>F)
hdl          1  26412   26412   49.579 8.652e-12 ***
ratio        1 535710  535710 1005.595 < 2.2e-16 ***
height       1   3875    3875    7.274   0.00730 **
hip          1   1836    1836    3.447   0.06412 .
Residuals  390 207764     533
```

Table 13. F test for Case 3

F test confirms significance of independent variables.

Prediction analysis.

Table 14 contain actual, predicted(fit), lower limit, upper limit values:

	actual	fit	lwr	upr
4	78	171	129	214
9	177	174	132	217
15	191	210	168	253
18	230	212	170	254
30	215	200	158	242
32	182	177	135	219
34	182	184	142	227
47	148	297	254	340
48	128	167	124	209
49	169	175	133	217
50	157	167	124	209

Table 14. Actual and Fitted values for Case 3

Model quality and accuracy metrics:

R2	RMSE	MSE	MAE	MAPE	AIC	BIC
0.5378645	29.87419	892.4673	17.83299	0.09996615	2930.122	2952.844

As we may expected, accuracy metrics for this case are much higher, then previous cases, and R^2 is respectively lower, due to low strength of linear relationship.

2.6 CASE 4. NO INFLUENTIAL OUTLIERS.

Model: `data$chol ~ hdl + ratio + height + hip`

Sample size: 395

In this case we will select and remove only those outliers or leverages that are or potentially influenced. The rest outliers and leverages will remain in dataset. For detection such data points, we will use Cook's distance, studentized residuals, Mahalanobis distance and Bonferroni correction. Therefore, first we will start with outlier analysis and after we will do residual revision.

Afterwards, we will select best outlier detector for this case #4, that improve our model quality and accuracy.

Cook' s distance.

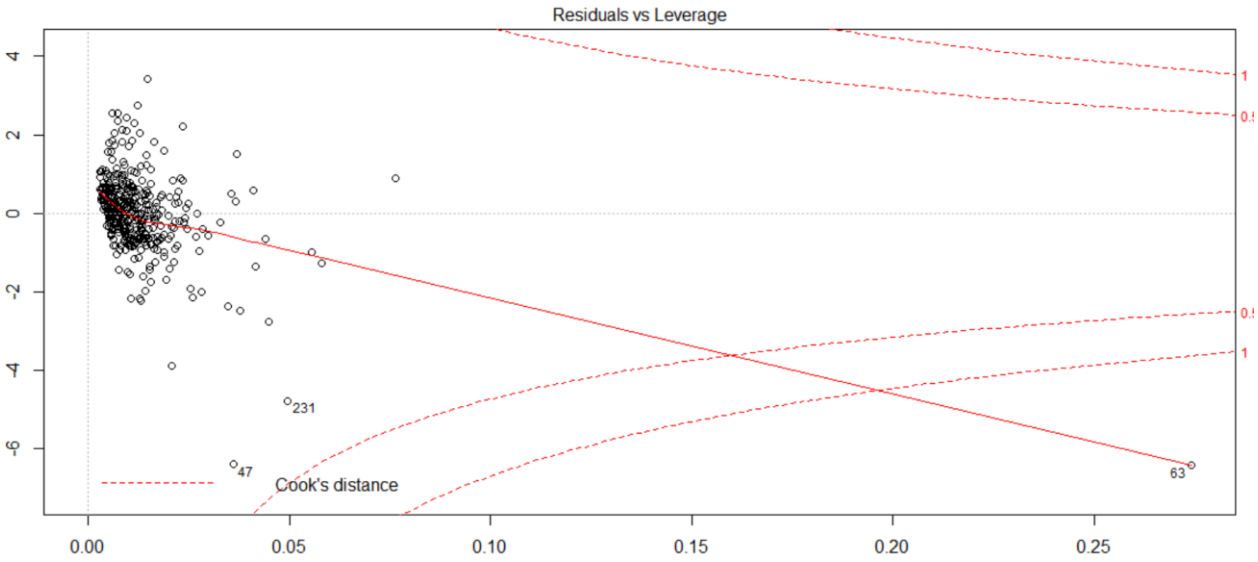


Figure 25. Residuals vs Leverage (Case 4)

By this graph we can see only one influential point (63), let take a look on different graph.

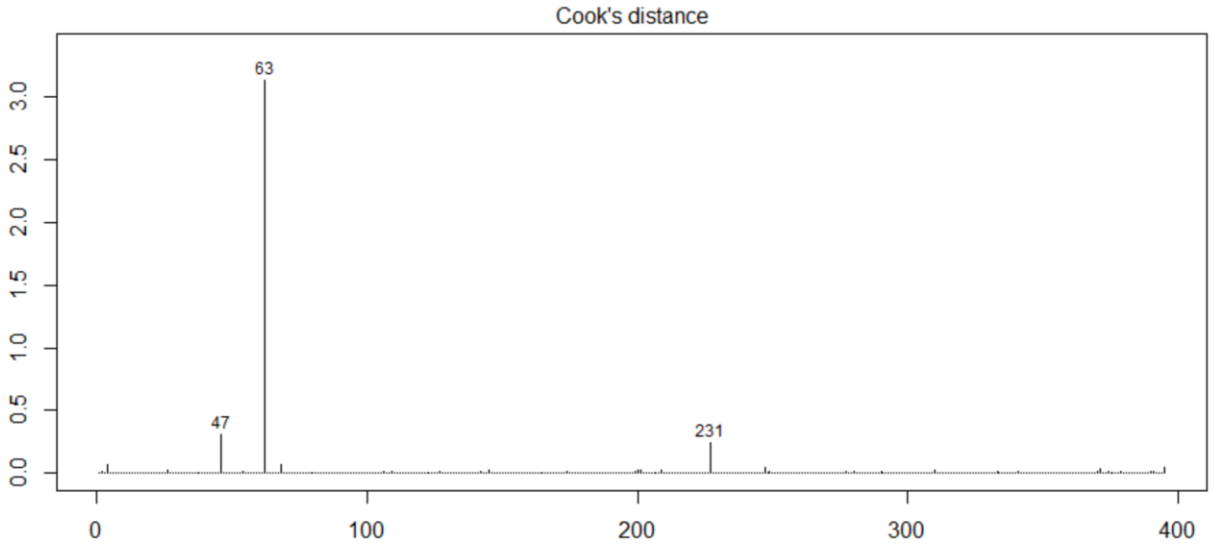


Figure 26. Cook's distance, before removing influential points (Case 4)

This graph shows three very extreme points 63,47 and 231, that even the rest of points are not well visible on graph, because these three points are very differ from all the rest:

ID	chol	hdl	ratio	height	hip
63	443	23	19.3	70	48
47	148	14	10.06	67	42
321	174	117	1.5	70	41

Table 15. Patients with influential unusual values by Cook's distance

Patient #63 has extremely high value result of "chol" and high "ratio".

Patient #47 has high "ratio" and patient #231 has extremely high "hdl" with low "ratio".

Hence, we decide to delete them and check again.

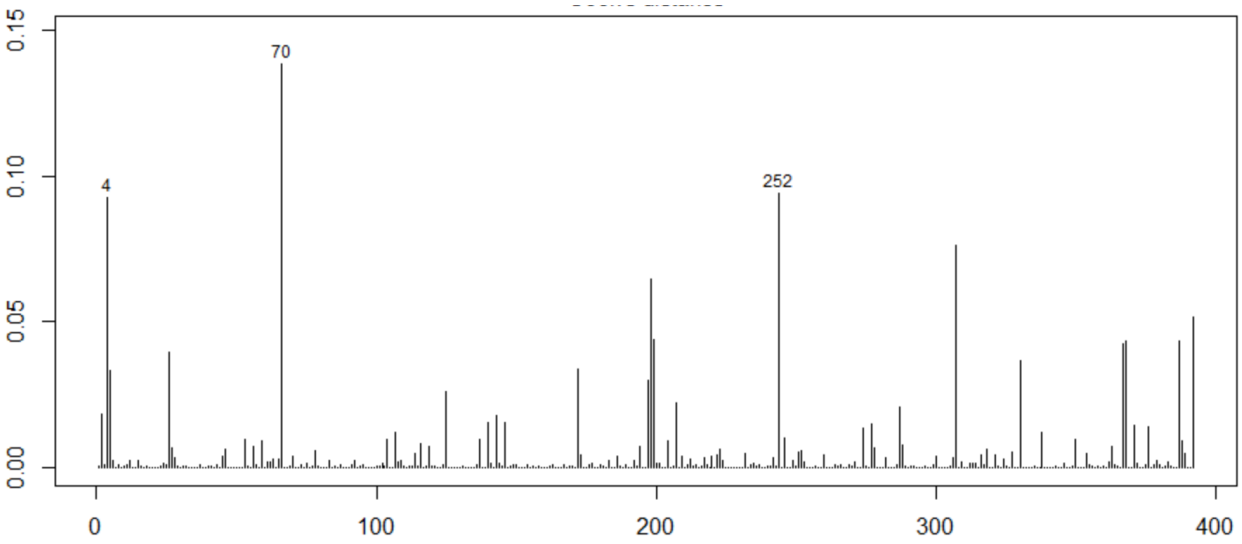


Figure 27. Cook's distance, after removing influential points (Case 4)

Spikes on graph are better visible then on previous, we can see almost all spikes of residuals. For better view, we can use graph below:

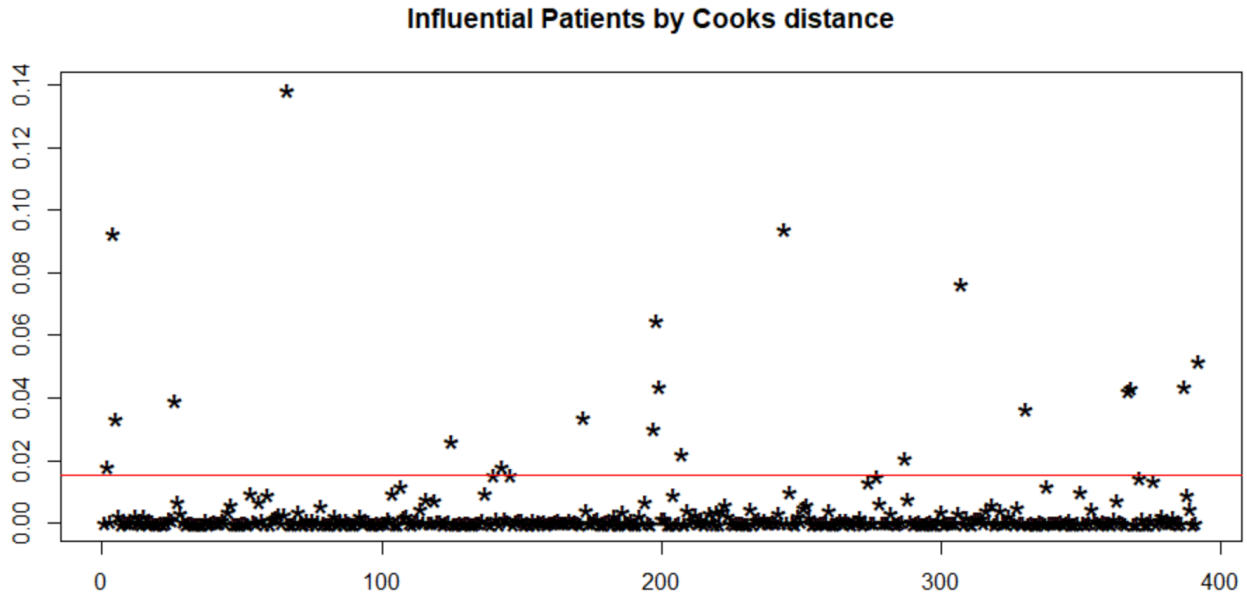


Figure 28. Cook's distance, different repressentation(Case 4)

On this graph we draw cutoff line. Stars which are higher than cutoff line are considered influential by Cook's distance. We can represent them, by command:

```
influential <- as.numeric(names(cooksd)[(cooksd > 4*mean(cooksd, na.rm=T
))])
View(influential)
ID's of influential patients:
2, 4, 5, 26, 70, 130, 145, 148, 177, 203, 204, 205, 213, 252, 295, 315 34
0, 377, 378, 398, 403
Total: 21
```

We will carry out regression analysis on reduced from influential outliers' dataset with 371 patients.

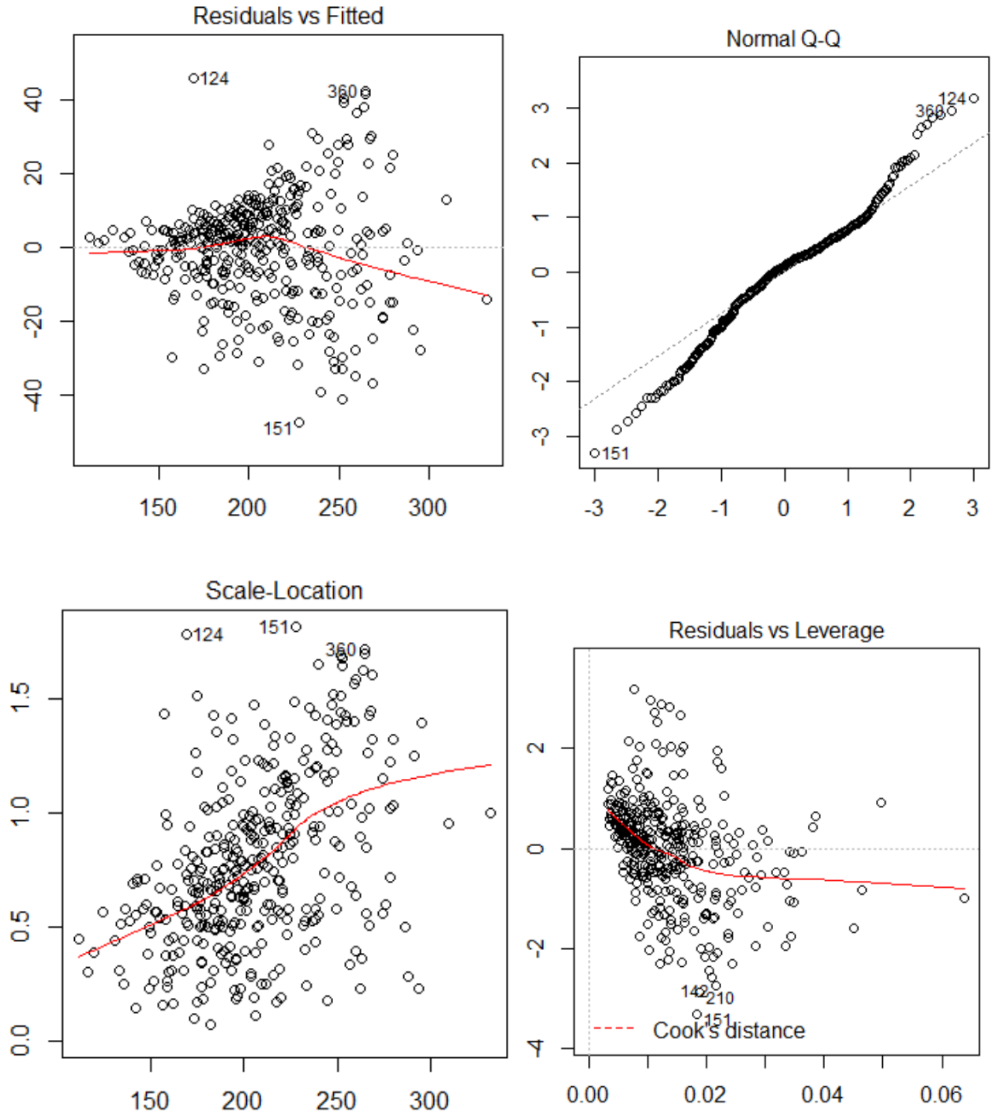


Figure 29. Residual diagnostic plots after removing three influential points (Case 4)

Again, assumption of normality and constant variance are violated, but t test and F test result a significance of coefficients, including intercept, and model significance.

Prediction analysis.

	actual	fit	lwr	upr
7	195	193	164	221
12	215	235	206	263
18	230	218	190	247
21	186	178	149	206
34	182	182	154	211
36	183	188	159	216
38	190	205	176	234
52	237	233	205	262
53	212	205	177	234
54	233	234	206	263
55	289	266	238	295

Table 16.Actual and Fitted by Cook's distance

Model quality and prediction accuracy.

R2	RMSE	MSE	MAE	MAPE	AIC	BIC
0.8444211	14.89495	221.8594	11.57017	0.0535711	2508.044	2530.386

This model is better than previous, R^2 is higher and accuracy metrics are lower.

Bonferroni correction and Studentized residuals.

Bonferroni correction.

To see visually residuals and cutoff lines we run following command in R:

```
n=nrow(data)
p=length(mymodel$coefficients)
plot(student<-rstudent(mymodel))
abline(h=-qt(1 - 0.05/(2*n), n - p - 1),col="red")
abline(h=qt(1 - 0.05/(2*n), n - p - 1),col="red")
```

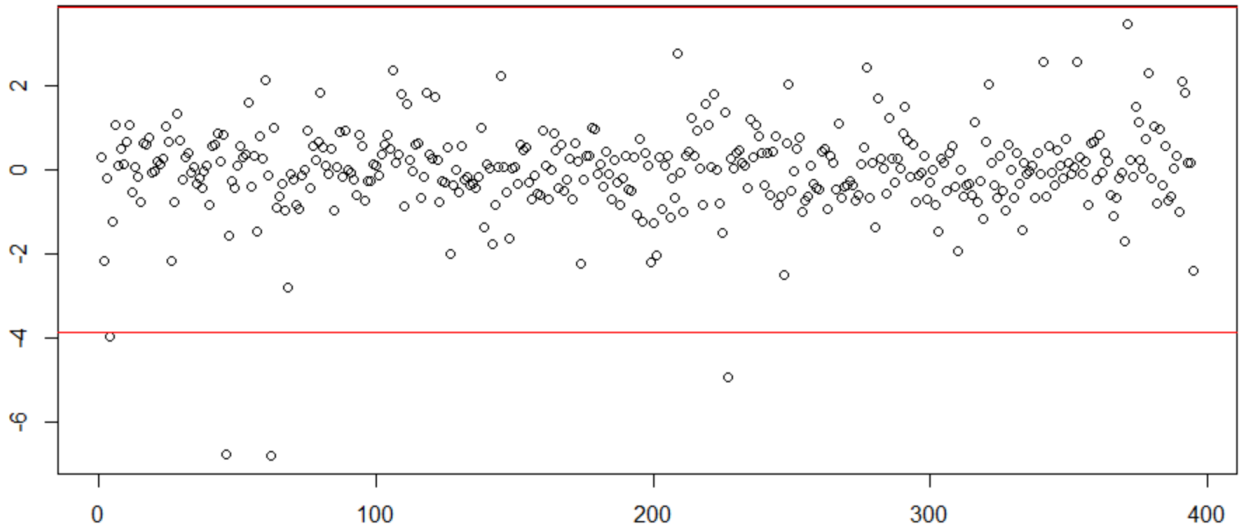


Figure 30. Bonferroni correction (Case 4)

By plot of Bonferroni correction there are only four influential outliers below lower cutoff.

We remove following outliers.

ID	chol	hdl	ratio	height	hip
4	78	12	6.5	67	38
47	148	14	10.06	67	42
63	443	23	19.3	70	48
231	174	117	1.5	70	41

Table 17.Bonferroni correction influential outliers

Significance of coefficients and violation of assumptions are similar as before.

Prediction analysis.

	actual	fit	lwr	upr
5	249	298	261	336
10	263	257	220	295
16	213	199	162	236
19	194	198	161	235
31	177	181	144	218
33	265	272	234	309
35	199	196	159	233
49	169	169	131	206
50	157	159	121	196
51	196	191	154	228

Table 18.Actual and Fitted values by Bonferroni correction

Model quality and accuracy.

```
     R2      RMSE       MSE       MAE        MAPE     AIC      BIC
0.8368254 19.88586 395.4476 13.79275 0.05956146 2826.83 2849.515
```

Studentized residuals.

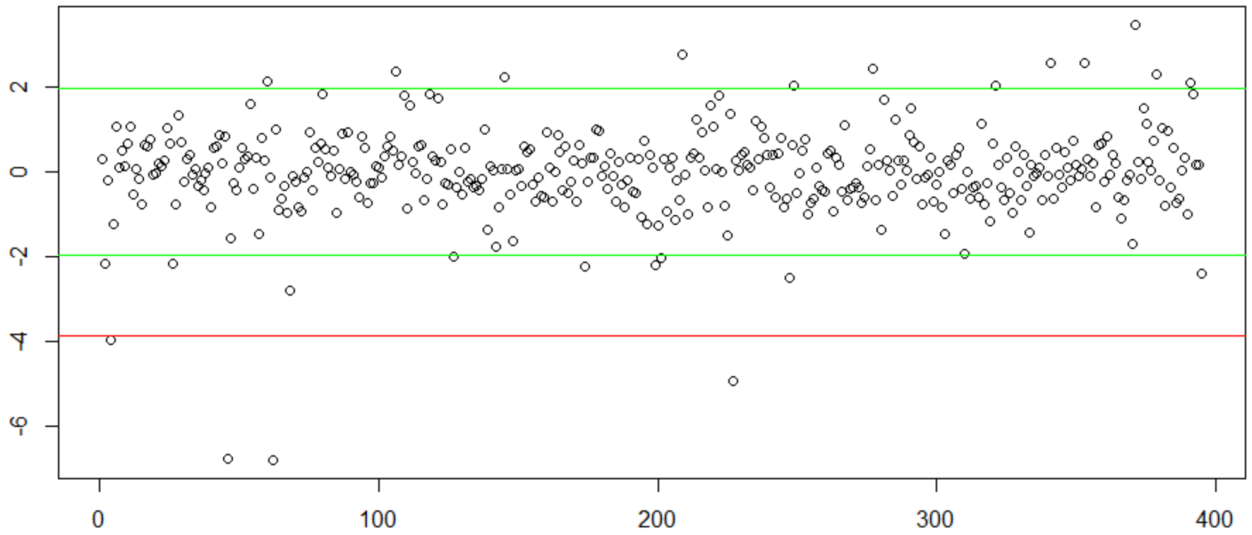


Figure 31. Bonferroni correction and Studentized residuals (Case 4)

Here, on graph, we can see how cutoffs are different for two outlier detection tests. The green lines are cutoffs of studentized residuals and red line is Bonferroni correction cutoff for outliers.

	chol	hdl	ratio	height	hip
61	296	60	4.9	63	48
109	292	55	5.3	70	41
148	347	42	8.3	70	49
213	342	48	7.1	65	46
254	305	44	6.9	71	45
282	293	54	5.4	71	39
327	298	50	6.0	66	46
348	306	56	5.5	69	41
360	307	58	5.3	67	42
378	337	62	5.4	72	44
386	302	57	5.3	67	51
399	296	46	6.4	69	39

Table 19.13 outliers above upper cutoff:

	chol	hdl	ratio	height	hip
2	165	24	6.9	64	48
4	78	12	6.5	67	38
26	179	92	1.9	72	36
47	148	14	10.6	67	42
63	443	23	19.3	70	48
70	232	114	2.0	61	38
130	181	24	7.5	71	47
177	193	24	8.0	66	45
203	188	24	7.8	68	48
205	215	100	2.2	65	34
231	174	117	1.5	70	41
252	231	110	2.1	63	41
403	159	79	2.0	64	58

Table 20.12 outliers below lower cutoff:

We reduce all mentioned above outliers.

Prediction analysis:

	actual	fit	lwr	upr
6	248	225	196	255
11	242	220	190	249
17	255	248	218	278
20	196	195	165	225
33	265	265	235	295
35	199	197	167	227
37	194	199	169	228
51	196	193	163	223
52	237	228	198	258
53	212	205	175	235
54	233	228	199	258
66	146	163	133	193

Table 21.Actual and Fitted values by Studentized residuals

Model quality and accuracy.

R2	RMSE	MSE	MAE	MAPE	AIC	BIC
0.8614337	14.46572	209.257	10.68823	0.05016209	2527.641	2549.963

Mahalanobis distance.

Detection of influential outliers by Mahalanobis distance in R looks following:

```
mahal<- Moutlier(data, quantile = 0.99, plot = FALSE)
```
To view cutoff value, we use

```
mahal$cutoff
```
The cutoff outcome is 3.88, hence data points which are higher than 3.88 are considered as influential.

We can plot scatter plot and qq plot of residuals to see outliers:

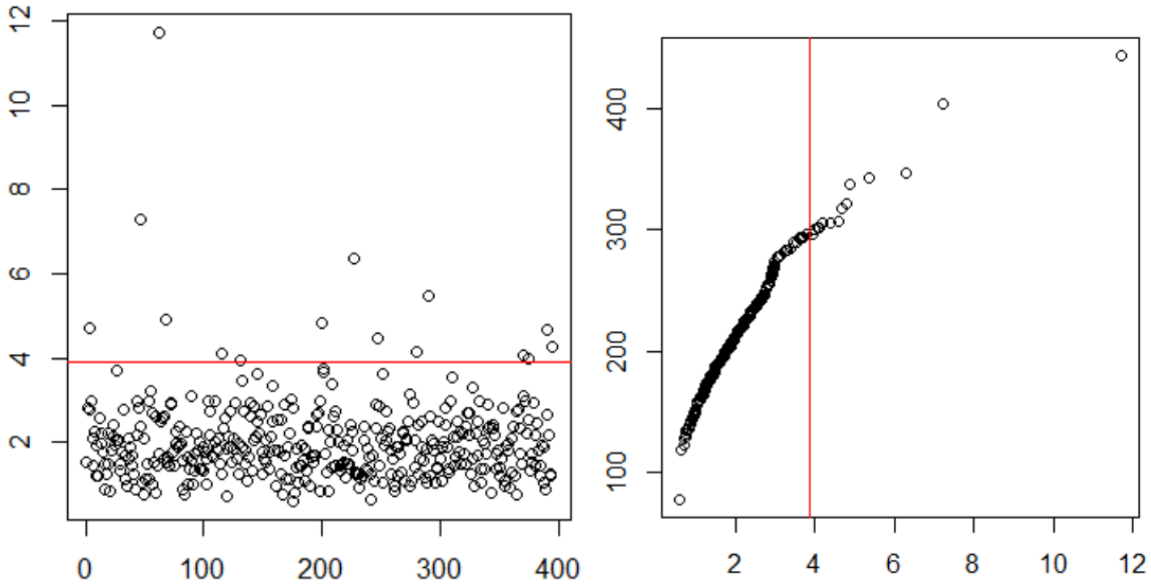


Figure 32. Mahalanobis distance scatter and qq plot(Case 4)

Let's list outlier patients and then delete them from dataset.

	chol	hdl	ratio	height	hip
4	78	12	6.5	67	38
47	148	14	10.6	67	42
63	443	23	19.3	70	48
70	232	114	2.0	61	38
119	174	34	5.1	70	64
134	318	108	2.9	65	44
204	293	120	2.4	64	45
231	174	117	1.5	70	41
252	231	110	2.1	63	41
285	152	32	4.8	52	49
295	404	33	12.2	69	39
378	337	62	5.4	72	44
381	322	92	3.5	56	41
398	301	118	2.6	61	41
403	159	79	2.0	64	58

Table 22.Outliers by Mahalanobis distance

Prediction.

	actual	fit	lwr	upr
5	249	297	261	333
10	263	258	222	293
16	213	199	164	234
19	194	197	161	232
31	177	180	145	216
33	265	271	235	307
35	199	197	162	233
49	169	169	134	205
50	157	160	124	195
51	196	193	157	228

Table 23.Actual and Fitted by Mahalanobis distance

Model quality and accuracy.

	R2	RMSE	MSE	MAE	MAPE	AIC	BIC
1	0.8553957	16.01986	256.6358	11.27954	0.05107732	2709.47	2731.967

Conclusion.

Regression analysis of given task gave expectable and non-expectable results.

We used same regression model for four different cases and we get four different results.

It is important to note, that the quality of analysis depends on quality of data characteristics as well as properly chosen techniques to investigate this data. Such characteristics or features can be data without missing values, data with large sample size, minimum or lack of unusual data values in data, a lot of variables and many others. But not always data features are as described. Occasionally, we have to face with the situation when data has poor quality or even is "problematic", similarly to our "*diabetes"* dataset. One possible approach is transformation of data in a way to get opportunity for exploration. According to my research, I've concluded, that it is important to inspect such data samples deeply, as we did in this thesis by analyzing different situations and cases, because each case is specific and requires individual approach as we made sure due to receiving distinctive results.

It is also important to note, that the research techniques are not limited by only techniques we used. Our analysis can be extended (continued) by applying other techniques of multivariate analysis, such as principal component analysis, factor analysis and etc. Since research topic deals with classical regression approach for prediction, we used only essential and basic classical methods for estimation regression model and for prediction values. We were testing a weak set of regression assumption by graphical review and also, we used different tests of hypothesis to make sure that our reasonings regarding graphs are matching with the test's hypothesis. We test statistical significance of model and it's coefficients by t test and F test. Additionally, we have been employed methods for detection influence of different types of outliers, such as extreme outliers, leverages and influential outliers. Finally, we were predicting values of "test" dataset by making regression analysis on "training" datasets. We've evaluated model quality and model prediction accuracy.

The four discovered cases represented us informative picture of how model efficiency may vary based on different metrics of model evaluation and prediction accuracy metrics. Moreover, we have witnessed of several regularities in our research:

1. Regression residuals remain same properties even after reducing some data from dataset. In all cases, we were either deleting or keeping outliers in dataset, but residuals distribution and variance were remaining similar for all cases.
2. For adjusted dataset (reduced or transformed) using same model may or may not be a proper decision. Some coefficients for "constant" model may be no longer statistically significant as it was for non-adjusted dataset (i.e. p values are changing with each case). Good example from thesis is case#3 when we include all outliers to the dataset.
3. Correlation of model variables varies. In our cases, we didn't have multicollinearity issue, however for each case correlation coefficients or variance inflation factors between variables were not the same.

4. Higher variation of model relationship(R^2), lower error measures (e.g. RMSE, MAE, etc.). From case #3, we concluded that if intercept is not statistically significant, then R^2 coefficient decreases, consequently error or accuracy metrics increase.
5. Violation of assumptions does not mean the model is wrong. Assumptions of classical linear regression for all cases were violated, however coefficients and model were statistically significant.

Considering these regularities, we received corresponding results of analysis. Main target of the thesis was to compare different scenarios with regard to outliers.

Before comparison, I would like to make summary of all gained results.

CASE #4.

	R2	RMSE	MSE	MAE	MAPE	AIC	BIC
Cook's distance	0.84	14.89	221.85	11.57	0.05	2508.04	2530.38
Bonferroni correction	0.83	19.88	395.44	13.79	0.05	2826.83	2849.51
Studentized residuals	0.86	14.46	209.25	10.68	0.05	2527.64	2549.96
Mahalanobis distance	0.85	16.01	256.63	11.27	0.05	2709.47	2731.96

Table 24. Methods comparison of Case 3

Culmination of Case #4 is that results of methods (Cook's distance, Mahalanobis, etc) for influential outlier detection were not identical. Some of the methods suggested more, some of them less quantity of outliers. But relying to model results, I've concluded that Cook's distance and Studentized residuals methods performed better than Bonferroni correction and Mahalanobis distance, for particular this research. We used the same dataset for all methods, and it means that Cook's distance and Studentized residuals are better suited for given dataset. Model strength of Cook's distance is lower and consequently error measures RMSE, MSE, MAE are larger than same measures of Studentized residuals. But interestingly, AIC and BIC are lower in Cooks distance, in spite of it has lower R^2. Hence, I can conclude here, that prediction accuracy metrics are not necessarily depend on R^2. Since, thesis is dealing with predictive analysis, it is worth to make a choice between Cook's distance and Studentized residuals, in favor to one that has minimum AIC and BIC, therefore we choose Cook's distance method, for Case #4.

Case comparison.

	R2	RMSE	MSE	MAE	MAPE	AIC	BIC
CASE #1	0.84	16.22	263.32	11.50	0.05	2311.40	2333.06
CASE #2	0.76	18.40	338.78	12.73	0.06	2441.61	2463.52
CASE #3	0.53	29.87	892.46	17.83	0.09	2930.12	2952.84
CASE #4	0.84	14.89	221.85	11.57	0.05	2508.04	2530.38

Table 25. Comparison of all cases

Case #1 is a leader, as we might expect, regression on data without outliers predicts better and has high determination coefficient. But some error metrics are relatively higher than in Case#4.

Dataset for Case#1 was reduced from all outliers, while in Case #2 only for influential points. Hence, I can conclude here, that outlier in the dataset, if its not influential may be helpful improve model accuracy, as we may see AIC and BIC from Case #2 are not relatively high. Another reason, that sample size of Case #1 is lower than in Case #2, therefore it produces higher error metrics.

The worst case is Case#3 by all metrics, where we kept all outliers in dataset.

Based on these results, my final conclusion is that it is rather to make analysis on removed from outlier's dataset, no matter if they are influential or not. But when some outlier carries out an important information, i.e. if it is not recommended to remove it, then it worth to use 3 sigma rule or method of Cook's distance, first one may have higher prediction accuracy (according on the practice we've observed) .

Even, in spite of violations of some regression assumptions, our classical regression approach showed good results of prediction values and their confidence intervals, some values of "test" dataset were precisely predicted.

List of reference

Applied Regression Analysis: A Research Tool, Second Edition John O. Rawlings Sastry G. Pantula David A. Dickey ISBN 0-387-98454-2

Cohen, J., Cohen P., West, S.G., & Aiken, L.S. (2002). Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.). Psychology Press. ISBN 0-8058-2223-2.

Douglas C. Montgomery ,2013. Introduction to Linear Regression Analysis, Fifth Edition Set 5th Edition, ISBN-10: 9781118780572

John Fox, 2015. Applied Regression Analysis and Generalized Linear Models: Edition 3 ISBN 9781483321318

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

Faraway, J. (2002). Practical Regression and Anova Using R. http://www.stat.lsa.umich.edu/ faraway/book/.

Hocking RR. A biometrics invited paper. The analysis and selection of variables in linear regression. Biometrics 1976;32:1-49. 10.2307/2529336

Kurt Varmuza, Peter Filzmoser, 2016. Introduction to Multivariate Statistical Analysis in Chemometrics. ISBN 9781420059

John Fox, 2002, An R and S-Plus Companion to Applied Regression Paperback, ISBN978-0761922803