Ministry of Education and Science of the Republic of Kazakhstan
Suleyman Demirel University

Araily Khuandykh

The analysis of indexes quality of a healthcare in Kazakhstan

THESIS

Presented in Partial Fulfillment for the
Degree of Master of Science in Computing Systems and Software
(degree code: 6060100)
Department of Computer Sciences
Faculty of Engineering and Natural Sciences

Supervisor: **Aziza Aipenova**

Kaskelen, 2019

Abstract

In this study considered analysis for healthcare service quality. Study considered quantity of hospital facilities for every 10000 population. As well as considered the average number of medical staff per hospital.The results of study can be used in the shade of the number of medical institutions and the training of the necessary number of medical personnel in the healthcare system. For these used regression method and as software used SAS. Python and R programming languages. These methods study were used on the example of two cities of Kazakhstan Almaty and Nursultan.

Аңдатпа

Бұл зерттеуде медициналық қызметтердің сапасы талданады. Зерттеуде әрбір 10 000 адамға ауруханалар саны карастырылады. Сондай-ақ бір ауруханаға медициналық персоналдың орташа санын ескеріледі. Сондай-ақ халықтың жан басына шаққандағы медицина қызметкерлерінің саны зерттеледі. Зерттеу нәтижелері медициналық мекемелердің саны мен денсаулық сақтау жүйесіндегі медициналық қызметкерлердің қажетті санын дайындау істеріне пайдаланылуы мүмкін. Бұл мақсаттар үшін регрессия әдісі қолданылады, ал бағдарламалық қамтамасыз ету ретінде SAS, Python және R бағдарламалау тілдері қолданылады. Зерттеудің осы әдістері Қазақстанның екі қаласы-Алматы және Нұрсұлтан мысалында қолданылды.

Аннотация

В данном исследовании рассмотрен анализ качества медицинских услуг. В исследовании учитывалось количество больничных учреждений на каждые 10000 населения. А также учитывается средняя численность медицинского персонала на одну больницу.Результаты исследования могут быть использованы, чтобы заполнить количество медицинских учреждений и подготовки необходимого количества медицинского персонала в системе здравоохранения. Для этих целей используется регрессионный метод, а в качестве программного обеспечения используются языки программирования SAS, Python и R. Данные методы исследования были использованы на примере двух городов Казахстана Алматы и Нурсултан.

Acknowledgements

I would like yo express my sincere gratitude to my supervisor PhD Asst. Prof. Aziza Aipenova who patiently guided me during the evolution of the thesis. I am truly indebted and thankful to my family for understanding and motivations. I am also thankful to my colleagues and friends who helped and supported me.

To my family

Contents

1. Introduction

Healthiness is the greatest gift in human life. The human being is the basic unit of demography: according to this the healthcare organizations are the main investors in human life as demography developers. Demographic, social and health saving factors for the development of urban and rural areas, their life supporting systems depend on the organization of activities and optimal equipment of medical institutions. According to the data provided by worldwide healthcare organizations we will give following facts about life expectancy: men's index is equal to 80, 2 years and woman own 85 years lifetime. Also in Japan average lifespan about 90 years. But in our country this index is significantly lower which is only 65 years [1]. Nowadays measuring healthcare quality is the most important issue in medicine. Growing number of people, increasing the cost of medical service and growing request for healthcare system are motivate people to know the used medical services are qualitatively. Expanding the use of medical services will complicate the work of health organizations, increase their vulnerability and require quality control of medical services. In addition, modern society requires even more transparency, which requires the health sector -as well as other public sectors - to give insight into their activities. Different sector professionals can use information about the quality of care for different aims. Medical staff tries to improve the quality of their assessment. State agencies use information on the quality of medical care and supervision of medical care, and insurance companies use it to select hospitals and patients who want to conclude contracts for the treatment of hospitals for use. Since care quality information is used for different purposes, it is important that quality indicators are reliable and valid and that care quality is accurate. However, despite the rapidly evolving scientific literature on this subject, currently there is no consensus on how to measure the quality of care.

1.1 Motivation

One of the main directions of the country's national policy is the health of the nation, which has undergone multiple transformations and is planned to be further developed. Today, huge amounts of funding are being made to improve the healthcare sector, including the private sector, introducing new financing and management techniques, and a new HR policy is being developed to upgrade and attract doctors, all of which is developing harmoniously in the long run. If we look at history, together with gaining independence, Kazakhstan inherited a vast system of health care from the former Soviet Union. Many health clinics and hospitals have also material and human resources valuable, and have also used old-fashioned techniques. Despite this in the beginning of independence our country were very poor and meet a lot of financial problems which were affects on healthcare system, by cause of this situation during 1991 and 1996 years child and adult morality rate were increased. But the Ministry presented the first state program "People's Health" within the framework of the strategic development of the Republic of Kazakhstan "Kazakhstan-2030". The purpose of the program was to improve the health of the people of Kazakhstan. It was a 1998-2008 academic years. [2] It consisted of three basic stages and was based on the following principles:

1. maintaining the level of access to health care services for the population, enhancement of economic, legal and organizational activities and adaptation of the health care system to market conditions;

2. Creation of economic and legal prerequisites for the formation of the domestic market of medical services;

3. Ensuring the effectiveness of medical institutions.

During 26 years of independence our country achieves a lot of successes in healthcare sphere but there are many unresolved issues in the area of health. They:

1. lack of professionals in villages;

2. poor training of health professionals;

3. low salaries and low social security of health care workers;

4. technically poor supply of medical institutions in the regions;

5. corruption in public administration;

6. Insufficient responsibility of doctors for the medical errors they have made the difficulty of bringing them to justice and prosecution.[4]

In Kazakhstan, there are a lot of programs and laws are presented to improve healthcare service quality during independence years, but there are a lack of analysis that shown the results. These tasks are still reflecting today and require a comprehensive solution to the problem. Therefore our study converges to analyze lack of healthcare indicators. Also, in this study we will take two main mega cities (Almaty, Nursultan) of country to assess the quality indicators of healthcare. In addition, to study relationship between studied observations or regions we will use linear regression analysis, time series analysis and SAS programming language, correlation analysis which are very useful for predicting.

1.2 Aims and Objectives

In this study our aim is to briefly study the linear regression model, which can give us ability to make deep analyzing to medical care service system; use 3 tools to give the best regression model results and show one of them; to prepare data to work, and show the last linear regression model of quality of service and make analyses.

1.3 Thesis Outline

The first chapter is Introduction chapter. It is this one that you are currently reading. It gives insight into the work done.In the first chapter we will make literature review. In Chapter two we review regression model and formulate the problem to solve our analysis. Chapter three is describing the tools that we can use in our analysis to the problem.In chapter fore we will give results of examined two city. And in chapter five we conclude our conclusion.

2. Literature review

In this proposal we focus on the dimension of quality of hospital care for external aims. Specially, we will learn the connection among variables of 'quality indicators'. In this chapter several concepts connected to this subject will be introduced and as a result the exact study questions and content of the rest of the thesis will be obtained. A variety of definitions of quality of healthcare have been formulate over the years. The terminology 'quality of medical service' does not own exactly definition. The quality of medical care - a set of characteristics confirming the conformity of the indicated health care to the patient's needs expectations, the current level of medical science and technology (Lindenbraten A. L., Grigereva T. N., 2007). Also, one usually used definition is from the Organization for Economic Cooperation and Development (OECD): excellence of healthcare can be calculated on at least three key size: efficiency, individual centeredness and security.[6] Efficiency reflect the quantity to which procedure outcome error-freely in preferred results.[7] Individual centeredness or receptiveness means the level to which a system functions by placing the patient at the centre of its deliverance.[6,8] This component is often measured in terms of the care, communication and understanding experienced by the patient in the clinician-patient relationship.[9,10] The level to which health care processes avoid, and improve unfavorable outcomes, possibly resulting from the healthcare processes itself, is referred to as protection. Quality of care investigate has a long history. From our country Independence Day rules about healthcare and measuring the quality indicators changed many time. If we look to the Rules of organization and conduct of internal and external expertise of quality of medical services of Minister of Health and Social Development of the Republic of Kazakhstan dated March 27, 2015 № 173. In these Rules the following main concepts are used:

1. Quality indicators of medical services - the criteria, including the effective-

ness, completeness and quality of medical services, which are reflected in the healthcare standards;

2. Structural Indicators - criteria for providing with human, financial and technical resources;

3. Process indicators - criteria for evaluating the implementation of preventive, diagnostic, therapeutic and rehabilitation technologies;

Indicators of the results of medical services - criterion of assessment of health impact caused by healthcare organizations or physical persons rendering medical aid: In the mid-1800s Florence Nightingale consider quality of care by assess death and sickness rates in British armed forces hospitals during the Crimean War. In Austria, Ignaz Semmelweis calculate and compare death charge connected to puerperal fever with parenthood hospital in Vienna (1841-1846).[12] in the start of the 20th century, the American physician Ernest Codman introduce a union to imprison patient outcomes following surgical actions in US hospitals.[13]These quality assessments were all initiated by health care professionals. The interest in quality and protection of care revived after the journal of the Harvard Medical Practice study in 1991 and the later publication 'To Err Is Human: Building a Safer System' by the Institute of Medicine in America. Since then the significance given to patient safety and quality of care has increased. And specially the demand for quality assurance and transparency from external stakeholder, such as humanity and government, has become stronger.[1,14-16] Nowadays, information on quality of care is used for internal and external purposes. Internal aim include, for example, the initiation and assessment of quality development programs. External use of quality information involves the public evaluation of medical centers based on data concerning the hospitals' quality of care. Such external comparison may guide to quality development through assortment of the 'best' providers by patients or clients.[17,18] The obtainable scientists, who focused on the organization analysis among effectiveness and quality of healthcare service, mostly used linear regression analysis. Laine et al. [22] studied the organization among productive effectiveness and proven quality of institutional enduring care for the aged by applying stochastic creation limit approach. Their outcome revealed that there was no system connection among creative effectiveness and quality, but lower quality might have an force on the productivity efficiency from a long-term point

of view. Laine et al. [23] calculated the scientific effectiveness of institutional long-term care for the aged by DEA system. Meantime, they discussed the connection among quality and technological effectiveness by Mann-Whitney test and correlation coefficients investigation. This investigation found a important correlation between the scientific effectiveness and "unwanted dimensions of Quality". Gok and Sezen [24] estimated the effectiveness of 348 public hospitals in Turkey utilizing DEA method and analyzed the exchange among quality and effectiveness of healthcare service by multiple regression analysis. In the process of analysis, they took the efficiency values as explain variables and regarded healthcare service quality as descriptive variables. The outcome manifested that the hospital dimension could have an effect on the trade-off among quality and effectiveness. It can be seen that these researches did not integrate the quality and effectiveness to talk about their connection, and they took quality as a power factor of efficiency. In addition, the trade-off among efficiency and quality of healthcare has not got a qualitative or quantitative finish form the feature of combined effectiveness and quality. Healthcare service quality by mathematical studies establish in little quantities, between them Tao Du in his paper studies about quality of medical service. His study based on the analysis over, he calculate the family member quality index value Q of healthcare service of 31 provinces of mainland China by using method for order liking by match to a perfect result (TOPSIS) method, and then treat Q as the complete quality level during efficiency capacity by DEA model. TOPSIS method calculate DMUs' relation index values by building perfect and negative ideal solution, so its essential thought is alike to DEA method[]. In the healthcare service field, Nayar and Ozcan [25] first used quality-adjusted DEA model to calculate and measure up to hospitals' comprehensive performance in terms of technical effectiveness and quality. They completed the hospitals were with top effectiveness also with upper quality services, and there was no evidence to prove the survival of a trade-off among quality and effectiveness. This study directly puts 3 quality measurements as results to estimate effectiveness by DEA model. Quality-adjusted DEA model was developed by Sherman and Zhu [26], and it viewed quality as the production for effectiveness measurement. DEA measures the effectiveness through the ratio of inputs and outputs, so only taking excellence as the output(s) may not guide to effectiveness estimation completely.

3. Method and methodology

3.1 The concept 'regression'

The concept of "regression" is associated with Francis Galton. In 1885 he published his scientific work "Regression in the direction of the overall average size of inheritance. of growth." Therein he came to the conclusion that the signs of parents are not fully inherited by children. and the more distant the ancestor. the to a lesser extent affect its properties on the descendant. Galton showed that children of very high or very low parents in average have less high or less low growth. In addition. the deviation in the growth of children is not as large as deviation of the growth of their parents from the average growth of the studied. This movement back in the direction to the mean Galton called regression (to regression - movement in the opposite direction). Galton wrote: "the Law of regression strongly testifies against full inheritance of any sign. From a large number children only a few will shy away from the average level of compared with the evasion of one of the parents. different their natural qualities. The brighter talent of one of parents. the less likely parents are to have the happiness of seeing that nature also generously endowed their sons. and even less often. to the endowment passed on to the next generation. The law is impartial and objective. It is evenly distributes the inheritance of good and bad traits. He destroys the excessive illusions of one gifted parent, cherishing the dream that his children will inherit all his abilities. The law also eliminates exaggerated concerns about the fact that children will pass all the weaknesses. shortcomings and diseases of parents Of course. these statements are contrary to the General theory that children talented parents are more likely to have any talents than children of parents with average abilities. Our reasoning only expresses that the fact that the most gifted of all the children of the few gifted parent pairs is not so talented as the most gifted of all the children

very many regression is due to Francis Galton. In 1885 he published his scientific work "Regression in the direction of the overall average size of inheritance. of growth." Therein he came to the conclusion that the signs of parents are not fully inherited by children. and the more distant the ancestor. the to a lesser extent affect its properties on the descendant. Galton showed that children of very high or very low parents in average have less high or less low growth. In addition. the deviation in the growth of children is not as large as deviation of the growth of their parents from the average growth of the studied. This movement back in the direction to the mean Galton called regression (to regression - movement in the opposite direction). Galton wrote: "the Law of regression strongly testifies against full inheritance of any sign. From a large number children only a few will shy away from the average level of compared with the evasion of one of the parents. different their natural qualities. The brighter talent of one of parents. the less likely parents are to have the happiness of seeing that nature also generously endowed their sons. and even less often. to the endowment passed on to the next generation. The law is impartial and objective. It is evenly distributes the inheritance of good and bad traits. He destroys the excessive illusions of one gifted parent. cherishing the dream that his children will inherit all his abilities. The law also eliminates exaggerated concerns about the fact that children will pass all the weaknesses. shortcomings and diseases of parents. Of course. these statements are not contrary to the General theory that children talented parents are more likely to have any talents than children of parents with average abilities. Our reasoning only expresses that the fact that the most gifted of all the children of the few gifted parent pairs is not so talented as the most gifted of all the children of so many couples with average abilities.» In the statistical interpretation of the regression is the change function depending on one or more changes arguments'. A function is defined as a variable that depends on another argument variable (independent variable.) Regression is a one-way statistical dependence process. With simple correlation study dependence between the variability of two variables X and Y. Using regression is an additional task: to establish how quantitatively, changing one variable while changing another (or others) per unit. If you investigate the dependence variable Y from X. then set the regression Y to X. If the same study the dependence of the variable X on Y. then determines the regression of X on Y. Purpose of the regression analysis - based on the values one variable selected

as an argument, predict the corresponding value of the other function). This is the first difference between the regression method and the correlation method. Second the difference is that the degree and nature of the regression can be set and with a small number of pairs of values of the dependent and independent variables.

3.2 Regression analysis tasks

In healthcare service analysis studies, regression analysis is used to solve the following tasks: 1. Establishing the form of dependence between variables (linear-nonlinear, negative-positive, etc.). 2. Definitions of regression function. It is important to find our what would be the effect on the dependent variable of the main factors if other factors did not change and if random elements were excluded. 3. Predictive evaluation of unknown values of the dependent variable. Using the regression function, you can reproduce the values of the dependent variable within the interval of the specified values of the independent variables (interpolation) or evaluate the process beyond the specified value interval (extrapolation).

3.3 Types of regression

With respect to the number of features taken into account, the regression can be: simple - between two variables, and multiple (or particular) - between the dependent variable Y and several independent (explanatory) variables: X1, X2, ...Xm; with respect to the form of the dependence - linear and nonlinear; with respect to the direction of the relationship - positive and negative. By the nature of the relationship between the dependent and independent variables, regression can be direct (the cause has a direct effect on the consequence), indirect (the independent variable acts through a third or a number of other causes on the dependent variable), and false (nonsense-regression - occurs in a formal approach without understanding of the reasons that cause this relationship).

3.4 Simple linear regression

Simple linear regression is understood as a one-way linear statistical dependence of a feature on only one independent variable. The analyzed characteristic is more often called dependent or resultant variable and denote by the symbol "y", and the factor-cause - independent or explanatory variable and denote by the symbol "x" (in the case of multiple regression - x_k , where k = 1.....m factors).Simple linear regression can be expressed:

- empirical regression line:
- regression equation and theoretical regression line:
- regression coefficient.

3.5 Empirical regression line

To build a regression line, you must have two data series. On the horizontal x-axis, the coordinate systems mark the values of the independent variable. On the vertical axis y - the values of the dependent variable corresponding to the values of the independent variable. Connecting all the dots line represents the regression line of Y on X. (see Figure 3.1 and)

3.6 The selection equation of the regression line

An empirical regression line is usually a more or less broken line. Despite the clear nature of the relationship between X and Y, it does not make it possible to accurately determine any value of Y for a given value of X. For this purpose, use the regression equation, which in general can be written as follows: $y_i - y = b(x_i - x) + c_i$ where y_i is the value of the i-th observation of the dependent variable (i = 1.....n); x_i - the value of the corresponding independent variable; x and y- n averages observations; b – coefficient of proportionality;c_i - error. The equation expresses certain dependence: after the deviation of $x_i$from the average for the variable X, the deviation of y_i from the average for the variable Y. The indicator b is the coefficient of proportionality, i.e. a measure that, on average, indicates a quantitative change in Y when X changes by a certain value. Moving y_i to the right side of equality, we get $y_i = y + b(x_i - x) + c_i$ If x is equal to zero,

Obs	year	population	med_person
1	[illegible]	[illegible]	[illegible]
2	[illegible]	[illegible]	[illegible]
3	[illegible]	[illegible]	[illegible]
4	[illegible]	[illegible]	[illegible]
5	[illegible]	1287246	[illegible]
6	[illegible]	1324736	12457
7	[illegible]	[illegible]	[illegible]
8	[illegible]	[illegible]	[illegible]
9	[illegible]	1414917	[illegible]
10	[illegible]	[illegible]	14152
11	[illegible]	1475579	[illegible]
12	2014	1507589	15945
13	2015	1552349	[illegible]
14	2016	[illegible]	[illegible]

Figure 3.1: Table of population and medical person in Almaty between 2003 and 2016

then x will be the initial value of Y. which should be started when constructing the regression line. when$x_i = 0$. Therefore. it is usually denoted by b_0 or a. Then the linear regression equation takes the form: $y_i = b_0 + b_1 x_i + \epsilon_i$ $y_i = a + b_1 x_i + \epsilon_i$ This equation is for simple linear regression. where f x_i - independent variable (cause-factor): a (or b_0) and b (or b_1) are the regression parameters to be evaluated. a (b_0) is the regression constant. It determines the point of intersection of the regression line with the y-axis. . a (b_0) is the mean of Y at $x_i = 0$. Therefore. biological interpretation of a (b_0)) is often difficult or even impossible. The constant performs in the regression equation alignment function. Thanks to it. the regression function is unbiased. b (b_1) is the coefficient of proportionality that characterizes the slope of the line to the abscissa axis: it is a measure of the influence of variable X on variable Y. or a measure of the dependence of variable Y on variable X. It indicates the average value of the change of variable Y when X changes by one unit. The sign at b (b_1) indicates the direction of this change. A positive value indicates the progressive nature of changes in the dependent variable with increasing values of the argument. At negative b (b_1) there is a negative regression - with increasing xi the values of the variable Y decrease. The regression parameters are not a dimensionless quantity. The regression equation

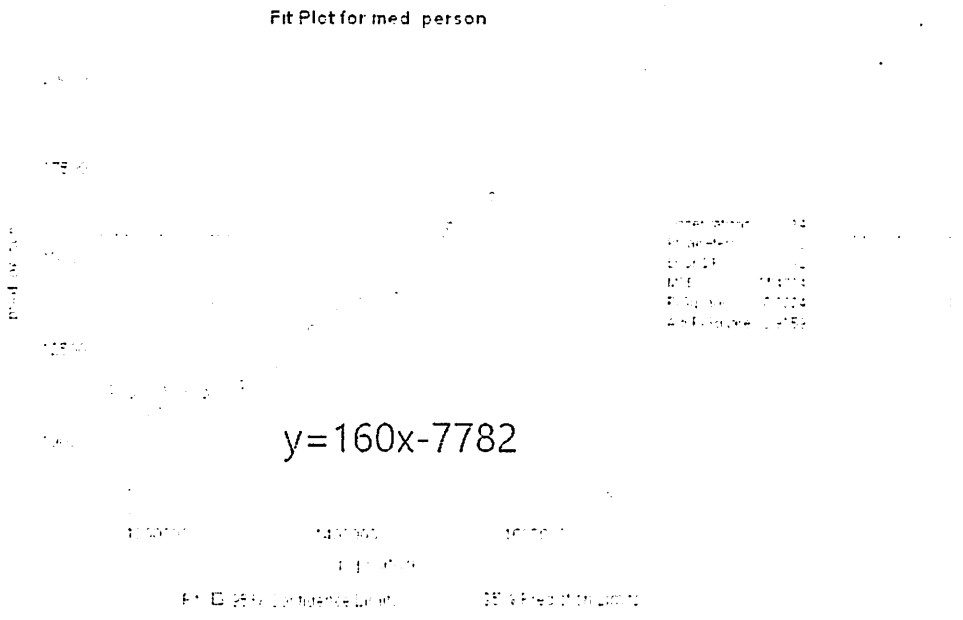


Figure 3.2: The estimation of regression analysis between medical person and population for 10000 people

constant. a (b_0) has the dimension of the dependent variable. Dimension b (b_1) is the ratio of the dimension of the dependent variable to the dimension of the independent variable. The parameters of the regression equation are unknown. Different values of a and b will correspond to different direct regressions. Therefore, the task of regression analysis is to find such estimates of these parameters (the selection of the line), which would be the most they were in good agreement with the actual data. To do this, use the method of least squares (LS). The system of normal equations of the LS-method for simple linear regression has the form:

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

The solution of the system gives unbiased estimates of b and a:

$$b = \frac{n\sum x_i y_i - \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$a = y - bx.$$

These estimates include in the equation for the selection of the line regression of Y on X:

$$y = a + bx.$$

For each value of the independent variable x_i (i =1.....n), the regression function – $(y_i)$is calculated using this equation. The values of the regression function (y_i) (i=1.....n) are called predicted or calculated values of the variable Y for fixed x_i. Then the observed value (y_i) can be represented as

$$y_i = (y_i) + e_i$$
$$= a + bx_i + e_i.$$

Where e_i is a perturbing variable or residue involving the influence of unaccounted factors (interpreted as an error). Predictive estimates $((y_i))$ are the best linear approximations (approximations) to the actual (empirical) values (y_i), because their standard error is minimized by the LS-method. The set of predicted values forms a theoretical regression line (see fig).

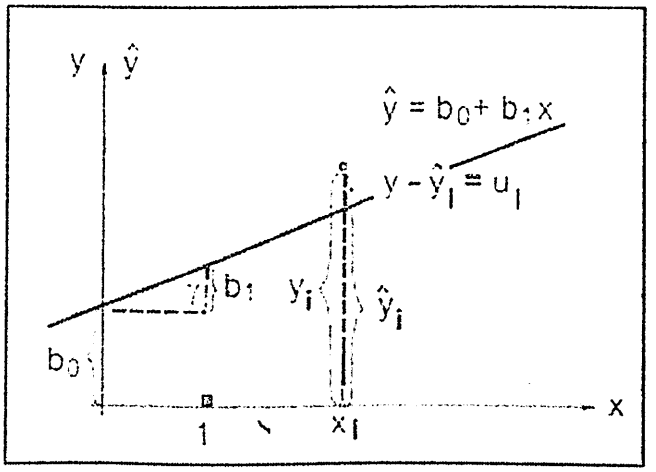


Figure 3.3:

Due to the modifying influence of unaccounted factors, several empirical values of y_i can be observed for each value of x_i. The values of the regression function $((y_i))$ are thus estimates of the average values of the variable Y for each fixed value of the variable X.

Example 3.1 It is required to describe the linear dependence of the ratio medical person (x_i) and hospital (y_i):

$$\sum_{i=1}^{14} y_i = 191218$$
$$\sum_{i=1}^{14} x_i = 912$$
$$\sum_{i=1}^{14} (x_i)^2 = 61560$$
$$\sum_{i=1}^{14} (x_i y_i) = 12771057$$

Obs	year	med_person	Hospital
1	2003	11432	53
2	2004	11682	55
3	2005	11592	53
4	2006	11842	54
5	2007	11823	54
6	2008	12157	64
7	2009	13101	63
8	2010	13611	62
9	2011	13823	69
10	2012	14191	78
11	2013	14349	84
12	2014	15945	72
13	2015	16779	76
14	2016	20180	73

Figure 3.4: Table for population and hospital of Almaty between 2003 and 2016

Therefore,

$$b = \frac{14(12771057)-(191218)(912)}{14(61560)^{-}(912)^2} = \frac{4403982}{30096} = 146.3$$

And.

$$a = \frac{(191218-146.3(912))}{14} = 4128$$

making the least squares line

$$y = 4128 + 146.3x$$

These equations are used to calculate prognostic estimates of hospital (y_i) for medical person (x_i). For example. the number of hospitals in 2003 equal to 53. the prognostic assessment of an average medical person will make

$$y_1 = 4128 + 146.353 = 11882$$

or

$$y_i = 4128 + 146.379 = 15685.7$$

This predicted (theoretical) value is the best, in the sense of the LS-method, linear approximation (approximation) to the actual (empirical) value.$y_1 = 11432$, because the standard error of the forecast is minimized. The theoretical direct regression is based on prognostic values (Fig. 25). You can predict the average annual number of doctors for any hospital. Thus, with a 85 hospitals, the expected average med person for one hospital will be:

$$y_i = 4128 + (146.385) = 16564$$

3.7 Regression coefficient

In healthcare analysis studies, it is often not the direct regression itself that is of interest, but the effect that one variable has on the other. In such cases, the regression coefficient is calculated. The regression coefficient is the ratio of the covariance between the independent and dependent variables to the variant of the independent variable:

$$b_{xy} = \frac{\sigma_{xy}}{(\sigma_x)^2}$$

Previously, it was shown that

$$\sigma_{xy} = \frac{SP_{xy}}{n-1}$$

and

$$(\sigma_x)^2 = \frac{SS_x}{n-1}$$

where SP_{xy} is the sum of products of deviations from means:$\sum(x_i - x)(y_i - y)$: SS_x- sum of squares of deviations from the mean:$\sum(x_i - x)^2$.

Then the estimation of the regression coefficient on the sample can be calculate the following formulas:

$$b_{xy} = \frac{\frac{SP_{xy}}{n-1}}{\frac{SS_x}{n-1}} = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{\sum x_i y_i - n\overline{x}\cdot\overline{y}}{\sum x_i^2 - n\overline{x}^2}$$

Thus, the regression coefficient is nothing but the coefficient of proportionality (b) in regression equation:

$$y_i = a + bx_i.$$

Like the coefficient of proportionality, the regression coefficient is a measure of the dependence of the variable Y on the variable X. It shows the average value of the change in the variable Y when x changes by one unit. The sign in the regression coefficient determines the direction of this change. A positive value indicates the progressive nature of changes in the dependent variable with increasing values of the argument. When the regression coefficient is negative, there is a negative change, in which the values of the variable Y decrease with increasing X. Regression analysis provides more information than correlation analysis. It allows you to set the dependence of both the variable Y on the variable X, and versa - X on Y. Therefore, the regression coefficients can be two:

- changing the Y variable when changing the X variable

$$\hat{b_{xy}} = \frac{\sigma_{xy}}{(\sigma_y)^2};$$

- changing the X variable when changing the Y variable

$$\hat{b_{yx}} = \frac{\sigma_{xy}}{(\sigma_x)^2};$$

Exercise3.2 For the data of example 3.1 will receive

$$\sum_{i=1}^{11} y_i^2 = 2697731611;$$

$$\hat{b_{y,r}} = \frac{\sigma_{xy}}{(\sigma_y)^2} = \hat{b_{yx}} = \frac{\sigma_{xy}}{(\sigma_x)^2} = \frac{\sum x_i y_i - \overline{x}\overline{y}n}{\sum y - n(\overline{y})^2};$$

$$= \frac{12771057 - 14(13658)(65.14)}{191218 - 14(13658)^2} = \frac{315507.37}{-261138227.8} = -0.00012$$

The result shows that with each increase in the average hospital, the live medical stuffs number will decrease by 0.00012 people. From a scientific point of view, the dependence of number of hospitals on the medical stuff is meaningless. Here this dependence is given only to illustrate the calculations.

3.8 Null hypothesis and confidence interval

Deviation of the prognostic assessment $\hat{y}_i$ from the actual the values of the dependent variable (y_i) are called the remainder or forecast error:

$$y_i - y_i = e_i$$

Forecast errors have a distribution with mean equal to zero and variances $\frac{2}{e}$. This residual variance estimated according to the formula:

$$\hat{\sigma_i^2} = \frac{SS_e}{df} = \frac{\sum(y_i - \overline{y_i})}{n-(m+1)} = \frac{\sum(e_i^2)}{n-(m+1)}$$

where SS_i is the sum of squares of residues: df is the number of degrees of freedom: n is the sample size: m is the number of independent variables. The physical meaning of the residual variance is the portion of the total variance of the dependent variable Y ($\frac{2}{y}$), which can not be explained by the dependence of the variable Y on the variable X. The residual variant is used to calculate the variant due to regression or an independent variable ($\frac{2}{b}$):

$$\hat{\sigma}_b^2 = \frac{\hat{\sigma}_\varepsilon^2}{\sum(x_i-\overline{x_i})^2} = \frac{n\hat{\sigma}_\varepsilon^2}{\sum x_i^2 - n\overline{x}^2} = \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_x^2(n-1)}$$

and error estimates of the regression coefficient (m_b)

$$m_b = \sqrt{\hat{\sigma}_\varepsilon^2} = \frac{\hat{\sigma}_\varepsilon}{\hat{\sigma}_x\sqrt{n-1}}$$

where $_\varepsilon$ is the standard deviation of the residuals: SS_x and $_x$ - sum of squares and the standard deviation of the independent variable. respectively. Thus. the standard error of the sample regression coefficient estimation is directly proportional to the residual distance and inversely proportional to the scattering of the independent variable (argument) and the sample size.

To test the null hypothesis ($H_0 : b = 0$). t-statistics are calculated:

$$t_b = \frac{|\overline{b}-0|}{m_b}$$

If $t_b l_(:df)$ at the number of degrees of freedom $df = n - m - 1$. then the null hypothesis at the significance level is rejected: the regression coefficient is considered statistically significant.

Confidence interval for the true value of the regression coefficient:

$$(\hat{b} - t_{a;d}\, f m_b) < b < (\hat{b} + t_{a;d}\, f m_b)$$

It can be argued that the confidence probability $P = 1$ parameter in the population (the true value of b) will not go beyond these boundaries.

Exercise3.3. The test of the null hypothesis and the construction of the confidence interval are illustrated for the regression coefficient of the hospitals on medical person (example 3.1).

According to the regression function

$$y = 4128 + 146.3x$$

the calculated prognostic score yi (for example, for the 2003 year $y_1 = (4128 + 146.353 = 11882)$ and their errors ($e_1 = 11432 - 11882 = 450$):

According to prognostic estimates, a theoretical line is drawn regressions ( ).Variance and standard deviation of the residues:

$$\hat{\sigma}_\varepsilon^2 = \frac{SS_\varepsilon}{n-(m-1)} = \frac{3971396}{14-(1-1)} = 3309493;$$

$$\hat{\sigma}_\varepsilon = \sqrt{2229.95} = 1819;$$

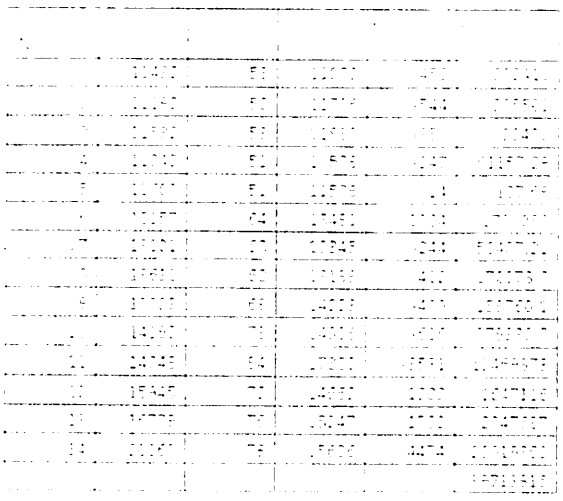


Figure 3.5: Table for population and hospital of Almaty between 2003 and 2016 and statistical analysis

Calculation SS_x.$_x^2$ and $_x$:

$$\sum_i^{11} x_i = 53 + 52 + 51 + ... + 76 + 79 = 912:$$

$$\overline{x} = \frac{912}{14} = 65.14;$$

$$\sum_1^{11} x_i^2 = 53^2 + 52^2 + 51^2 + ... + 76^2 + 79^2 = 61560:$$

$$SS_x = \sum (x_i - \overline{x_i})^2 = \sum x_i^2 - n\overline{x}^2 = 61560 - 14 * 61.14^2 = 9226.6$$

$$\hat{\sigma_x^2} = \frac{9226.6}{14-1} = 709.7:$$

$$\hat{\sigma_x} = \sqrt{709.7} = 26.64:$$

Variation and regression error:

$$\acute{\sigma_b^2} = \frac{\sigma_\varepsilon^2}{SS_x} = \frac{3309493}{9226.6} = 358.7:$$

$$m_b = \sqrt{\acute{\sigma_x^2}} = \sqrt{358.7} = 18.9:$$

$$= \frac{\sigma_\varepsilon}{\sigma_x\sqrt{n-1}} = \frac{1819}{26.64\sqrt{13}} = 18.9.$$

Hence,

$$b \pm m_b = 146.3 \pm 18.9.$$

Actual t-test :

$$t_b = \frac{b_{y.x}}{m_b} = \frac{+146.3}{18.9} = 7.74.$$

At $= 5\%$ and the number of degrees of freedom $df = 14 - (1 + 1) = 12$ critical value $t_(0.05; 12) = 2.179$(table. A. 8 of Annex A). Since, $t_b < t_(0.05; 12)$, the null hypothesis remains valid. The true value of the regression coefficient 95% confidence probability is in the range of estimates

$$(b - t_0.05;_1 2m_b) < b < (b + t_0.05;_1 2m + b). i.e.$$

From (146.8 − 2.17918.9) to (146.8 + 2.17918.9) or
From 105.6 to 187.98 and it can take a zero value.

Exercise3.4.

Regression analysis of population (X) and the number of medical stuff (Y) between 2003 and 2016 illustrates a different calculation technique. Here we will use per 10.000 people of the population.

Initial data and calculation of sums of squares and products:

1	115	1.14	0.22	-22.71	0.05	521.15	5.06
2	120	1.12	0.25	-20.75	0.06	430.56	4.65
3	121	1.15	-0.23	-19.72	0.05	276.67	2.54
4	125	1.13	-0.23	-12.78	0.05	163.41	2.96
5	134	1.16	-0.21	-8.75	0.04	76.59	1.82
6	135	1.20	-0.25	-5.76	0.06	33.18	0.77
7	136	1.31	-0.16	-1.75	0.03	3.06	0.36
8	139	1.35	0.09	1.51	0.00	2.28	-0.01
9	141	1.38	0.02	3.58	0.00	14.06	0.06
10	145	1.42	0.05	7.46	0.00	55.65	0.40
11	148	1.43	0.07	9.59	0.00	91.70	0.69
12	151	1.59	0.23	13.13	0.05	172.66	3.01
13	155	1.55	0.71	17.00	0.50	311.95	5.51
14	171	2.00	0.65	33.75	0.42	1139.21	21.94
	1926	19.13	0.00	0.00	SS_y =0.86	SS_x =3256.22	SP_{xy} =50.75

Figure 3.6: Initial data and calculation of sums of squares and products

$$\overline{x} = \frac{\sum x_i}{n} = \frac{1926}{14} = 137.8$$

$$\overline{y} = \frac{\sum y_i}{n} = \frac{19.13}{14} = 1.36$$

$$\hat{\sigma}_y^2 = \frac{SS_y}{n-1} = \frac{0.86}{14-1} = 0.066$$

$$\hat{\sigma}_x^2 = \frac{SS_x}{n-1} = \frac{3256.22}{14-1} = 250.48$$

$$\hat{\sigma}_{xy} = \frac{SP_{xy}}{n-1} = \frac{50.75}{13} = 3.90$$

Regression coefficient:

$$b_{y,x} = \frac{\sigma_{x,y}}{\sigma_x^2} = \frac{3.90}{250.18} = +0.016$$

$$b_{x,y} = \frac{\sigma_{x,y}}{\sigma_y^2} = \frac{3.90}{0.066} = +59.09$$

The regression coefficients show that with the increase of population increased number of medical stuff by an average of 59 people. On the other hand, the increase in population with 59 people increased number of medical person by an average of 0.016 people.

Regression equation constants:

$$a = y - b_{y}x.x = 1.36 - 0.016 * 137.8 = -0.84$$

$$a = x - b_{x}yy = 137.8 - 59.09 * 1.36 = +57.44$$

The regression function of Y on X:

$$y_i = -0.81 + 0.016x_i.$$

and X by Y:

$$x_i = 57.44 + 59.09y_i.$$

The function of regression of population to number of medical stuffs:

$$y_i = -0.84 + 0.016x_i.$$

Calculation of prognostic estimates (y_i), residues (e_i), the sum of squared residuals (SS_e) and the sum of the squares of prognostic estimates (SS_y):

Actual (y_i) and the predictive estimate (y_i) held the actual and the theoretical regression line (see fig:picture2). Residual of variance:

$$\sigma_e^2 = \frac{SS_e}{n\ (m-1)} = \frac{0.068}{12} = 0.00567$$

Regression variance:

$$\sigma_b^2 = \frac{\sigma_e^2}{SS_x} = \frac{0.00567}{3256.22} = 0.0000002$$

The test of the null hypothesis for estimation of regression coefficient of number of medical stuffs (Y) for the population (X):

Error:

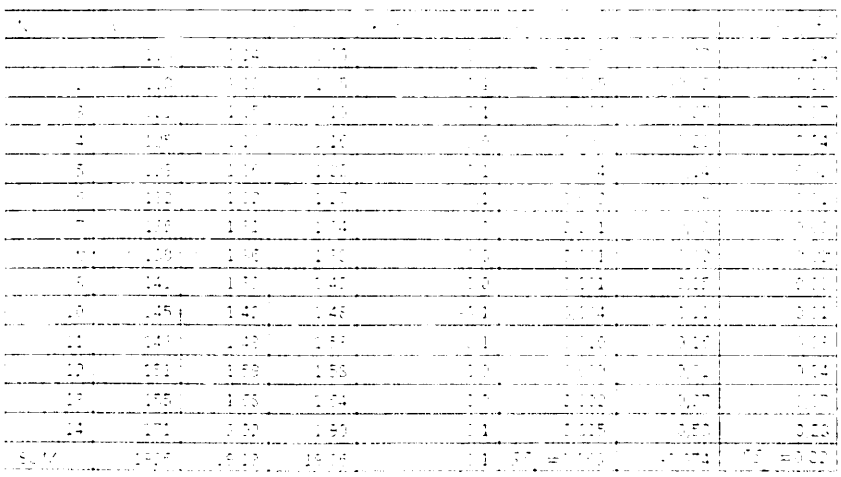


Figure 3.7: Initial data and calculation of sums of squares and products

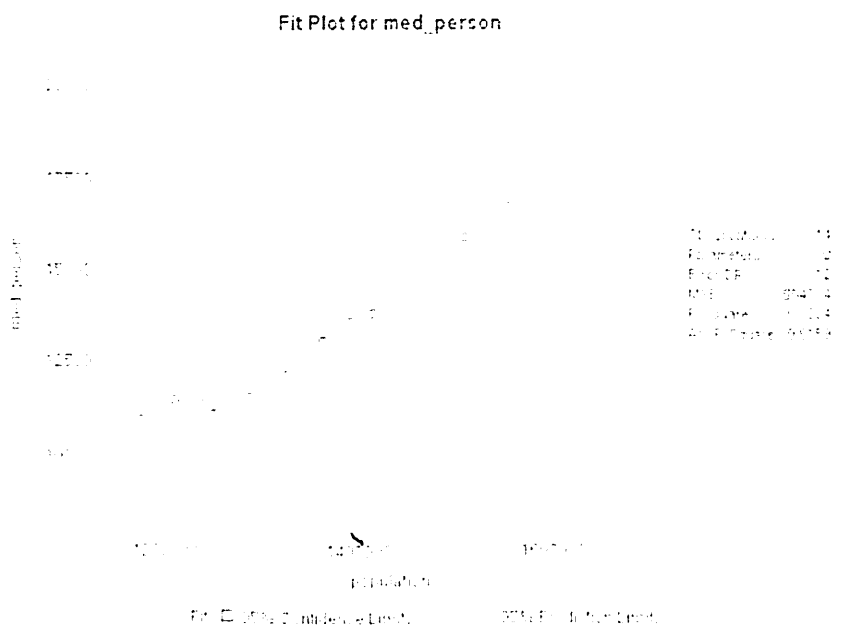


Figure 3.8: regression line between med person and population

$$m_b = \sqrt{\hat{\sigma}_b^2} = \pm\sqrt{0.000002} = 0.0045$$

Regression with error:

$$b \pm m_b = 0.015 \pm 0.0045.$$

Actual t-criterion:

$$t_b = \frac{b_{act}}{m_b} = \frac{0.016}{0.0045} = 3.56.$$

When = 10% and the number of degrees of freedom $df = 14 - 1 - 1 = 12$ table. A. 8 of Appendix A we find $t(0.10:12) = 1.78$. Since, $t_b < t(0.05:12)$ the null

hypothesis can be rejected. The true value of the regression coefficient with a confidence probability of 90% is in the range of estimates

$$(\hat{b} - t_0.10;_1\, 2m_b) < b < (\hat{+}t_0.10;_1\, 2m_b).\ i.e.\ ,$$

From $(0.016 - 1.780.0567)$ to $(0.016 + 1.780.0567)$ or
From -0.085 to $+0.117$ and it can take a zero value.
If we take $= 5\%$. then, $t_i 0.5; 12) = 1.18$ and the null hypothesis is not rejected. because $t_b < t_0.05$. It follows that the choice of significance level ( ) has a decisive influence on the acceptance or rejection of the null hypothesis. Therefore. to exclude subjectivity in the discussion of the results of the experiment. it is necessary to set the value in advance. before the results.

The analysis of the security of the population doctors per 10.000 populations

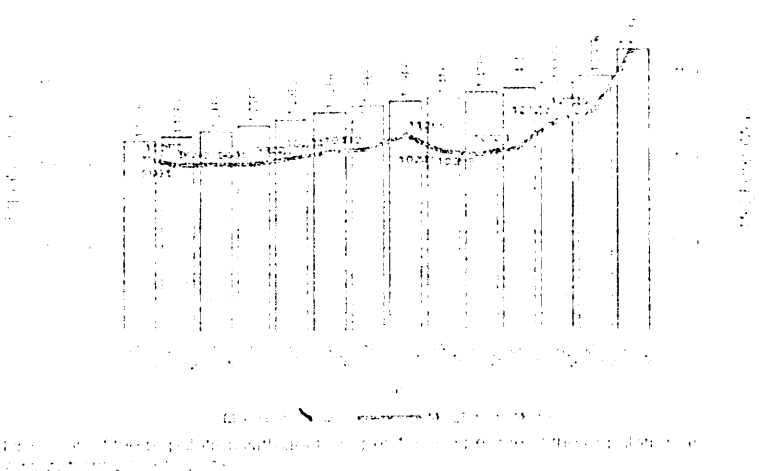


Figure 3.9:

in the region showed that from 2003–2016. there was a significant increase which is 62.4%. and it speaks about the quality of service in health care, because this reduces the workload of doctors. The growth rate in 2011 amounted to – 9.09% in relation to 2010. It should be borne in mind that the number doctors and high provision of 10 thousands of people are not a guarantee of quality medical assistance and good health outcomes. The decisive factors here are quality, training of specialists. effectiveness of financing, share of health care and labor costs the doctor.

3.9 Linear multiple regression

Any sign or phenomenon is determined (determined), as a rule, by a set of simultaneously and jointly acting causes. Therefore, one of the tasks of regression analysis is the study of dependence one variable Y from several explanatory or independent variables X1, X2, ... Xm in a specific place and time. This problem is solved by multiple (multifactor) regression analysis. In the presence of linear relations between variables, the General expression of the multiple regression equation is:

$$y = b_0 + b_1 x_1 + b_2 x_2 + + b_m x_m + \epsilon,$$

where $b_1, b_2, \ldots, b_m$ regression coefficients.
The linear multiple regression function is:

$$y = b_0 + b_1 x_1 + b_2 x_2 + + b_m x_m.$$

y_i (i = 1, ..., n) - the estimated values of the regression. They indicate the average values of the variable Y at the point i at fixed values $x_i k$(k = 0, ... m) - assuming that only these m variables are the cause of the variable Y.
Coefficients b_k (k = 0, ..., m) - regression parameters. The regression constant b_0 performs the alignment function in the regression equation. It defines the intersection point of the regression hypersurface with the ordinate axis.
Values $b_1, b_2, \ldots, b_m$ there are estimates of regression coefficients. The sub-index when the coefficient corresponds to the sub-index of the independent variable. Thus, b_1 indicates the average value of the change in Y when X_1 changes by one unit (provided that other variables remain unchanged); b_2 shows how many units on average will change Y if the variable X_1 changed by one (provided that the variables $X_k(k2)$ would remain unchanged), etc. While the regression function covers the cumulative simultaneous effect of the independent variables, the regression coefficient $b_k(k = 1, 2, \ldots, m)$ indicates the corresponding averaged partial effects of the variable X_k, assuming that the remaining independent variables remain constant.
Thus, from the point of view of statistical methodology, there is no difference between multiple and partial regression. Therefore, in the study of regression, there is no need to distinguish between partial and multiple regression. Therefore, in

the literature parameters $b_k (k = 1, 2, \ldots, m)$ are called as coefficients of multiple and partial regression. It should be noted that although multiple regression covers the simultaneous action of m independent variables, the regression coefficient b_k excludes the influence of other variables-factors (in simple linear regression, the influence of other unaccounted factors is partially reflected in the regression coefficient).

The task of multiple regression analysis is to evaluate the regression parameters based on the results of sample observations of the variables included in the analysis. If we introduce a dummy variable $x_{i0} = 1$ for the constant, b_k for all i 1.2. ...n, then the linear model of multiple regression can be represented as

$$y = b_0 + b_1 x_1 + b_2 x_2 + + b_m x_m + \epsilon$$

or in matrix form

$$y = Xb + \epsilon.$$

To estimate the unknown parameters of vector b, as in the case of simple linear regression, the least squares method is used. Normal equations that satisfy the requirement that the sum of the squared deviations of the empirical values from the calculated valuesregressions should be minimal, have the form

$$X^{'}Xb = X^{'}y.$$

If the $x'x$ matrix is invertible, it can be obtained as solution of the system of normal equations a column vector the unknown parameters of the regression:

$$b = (X^{'}X)^{(} - 1)X^{'}y.$$

The matrix $X^{'}X$ and vector $X^{'}Y$ have the following structure:

$$X'X = \begin{bmatrix} n & \sum x_{i1} \ldots & \sum x_{im} \\ \sum x_{i1} & \sum x_{i1}^2 \ldots & \sum x_{i1} x_{im} \\ \sum x_{im} & \sum x_{im} x_{i1} \ldots & \sum x_{im}^2 \end{bmatrix}$$

$$X'X = \begin{bmatrix} \sum y_i \\ \sum x_{i1} y_i \\ \cdot \\ \sum x_{im} y_i \end{bmatrix}.$$

3.10 The relationship of correlation coefficients and regression

There are two coefficients in regression analysis regressions. The correlation coefficient is a common measure the coupled variations of the two characteristics. It's more artificial than regression. In regression, one feature acts as an independent variable, the other as a dependent and conversely. These dependencies have a specific meaning.

Define the regression of Y on X as:

$$b_{y,x} = \frac{\sigma_{x,y}}{\sigma_x^2}.$$

We will multiply both part of equation with $\frac{\sigma_x}{\sigma_y}$:

$$\frac{\sigma_x}{\sigma_y} b_{x,y} = \frac{\sigma_{x,y}}{\sigma_x^2} \frac{\sigma_x}{\sigma_y}.$$

After simplifying the right side by $_x$ we obtain:

$$\frac{\sigma_x}{\sigma_y} b_{x,y} = \frac{\sigma_{x,y}}{\sigma_x} \frac{\sigma_x}{\sigma_y}$$

Or

$$\frac{\sigma_x}{\sigma_y} b_{y,x} = r.$$

Then:

$$b_{y,x} = r \frac{\sigma_y}{\sigma_x}$$

Respectively

$$b_{y,x} = r \frac{\sigma_x}{\sigma_y}.$$

Therefore, if the correlation coefficient is known, then standard deviations can be used to determine the required regression coefficient.

Then, if you multiply the two regression coefficients, get

$$b_{y,x} \times b_{x,y} = r^2.$$

From this relationship it follows that

$$r = \pm\sqrt{b_{y,x} \times b_{x,y}} = \pm\sqrt{0.016 \times 59.09} = \pm 0.95 =$$

$$= \sqrt{\frac{\sigma_{x,y}}{\sigma_x^2}\frac{\sigma_{x,y}}{\sigma_y^2}} =$$

$$= \pm\frac{\sigma_{x,y}}{\sigma_x\sigma_y} = \pm\frac{3.90}{0.26*15.8} = \pm 0.95.$$

that is, mathematically, the correlation coefficient is the geometric mean of the two regression coefficients.

It should be noted that for the regression coefficient Y to X there are relations:

$$R^2 = b_{y,x}^2\frac{\sigma_x^2}{\sigma_y^2} =$$

$$= (r\frac{\sigma_y}{sigma_x})^2\frac{\sigma_x^2}{\sigma_y^2} =$$

$$= r^2.$$

Thus, the square of the correlation coefficient is equal to the coefficient of determination.

As the formula for coefficient of correlation and the formulas for the regression coefficients is Central covariance. It is, in fact, a true measure of the conjugate variation of the variables analyzed. Therefore, covariance is a link in correlation and regression analysis.

3.11 Regression analysis sequence

In General, the regression analysis procedure includes the following steps.

1. Formulation of the problem. Implies the specification.
   Bio technically phenomena and processes: the relationship between them is subject to evaluation.

2. Identification of variables. On the basis of professional theoretical considerations and biological sense determine a reasonable number of variables: classify them into dependent and independent.

3. Data collection. Based on the purpose and objectives, establish the principle of data selection and sample size. If the required amount of data cannot be provided for any phenomena, the first step should be revisited.

4. Specification of the regression equation (parameterization of the model).At this stage: - formulate hypotheses about the form of connection (linear or nonlinear, simple or multiple) and - checking the background.
   For the most part, the type of regression equation in the research process define step-by-step by eliminating variables that do not significant influence on the dependent variable, and the inclusion of analysis of new factors-causes with a check of their significance.

5. Estimation of regression function. Determine the numerical values of parameters of the regression equation.

6. Evaluation of the accuracy of the regression analysis. Calculate statistical indicators of accuracy regression analysis.

7. Bio technically interpretation of the results. Results compare with hypotheses, evaluate their likelihood with same points of view.

8. Prediction of unknown values of the dependent variable. The resulting regression function is used for predictive analysis. If the regression function is defined and it is biologically warranted, the predictive (theoretical) evaluation has sufficient reliability. In essence, they are average values that are more likely to be expected. The power of multifactorial biological phenomena and the multiplicity of expressions of individual actual (empirical) values are scattered around the means. Therefore, it is natural that the actual values of the dependent the variables will not coincide with the calculated ones, i.e. with the forecast. After this must be considered degree of scattering of observations around the theoretical regression line characterizes reliability obtained the regression equation of predictive assessments. The accuracy of the forecast is determined not only by the accuracy estimates of regression parameters, but also the extent to which future values of independent variables are reliably estimated at based on additional information. Source of such additional information may be professional theoretical considerations in accordance with zootechnical, tribal, economic and even social policy of the economy, region of the state. Therefore, the process of building statistical models should be accompanied by adjustment of parameter estimates regression and statistical characteristics according to expected changes in

the circumstances of their formation. The forecasting results on regression easier to meaningful interpretation than simple extrapolation trends, because you can better take into account the nature of the study phenomena. Because of this, the regression analysis is widely applied in solving problems of long-term planning.

Addition

Prerequisites of regression analysis

Basic assumptions of regression analysis:

1. . The perturbation (remainder) of e_i (or dependent variable y_i) is a random variable, and the explaining variable x_i is a non-random variable.

2. The expectation e_i is equal to zero:
   $E(e_i) = 0$ or $E(y_i) = b_0 + b_1 b_x$.

3. The perturbation variant of e_i or (y_i) is constant for any i (unaccounted factors have the same effect): $Var(e_i) = ^2_e$ or $Var(e_i) = ^2$.

4. Perturbations e_i and e_j (y_i and y_j) are not correlated: $E(e_i, e_j) = 0$ when, $i0$.

5. The perturbation e_i (or y_i) is normally distributed random variable.

6. The number of observations must exceed the number of parameters, otherwise impossible their assessment.

7. Explanatory variables X should not correlate with the disturbing variable e, i.e. $E(x_i k, e_i) = 0$ when $(k = 1 \ldots . m)$

8. Variables $x_k$explain variable y, but does not explain e it is assumed unilateral dependence of x_k and the lack of relationship.
   To obtain the regression equation, the first ones are sufficient four prerequisites. Requirement of the fifth prerequisites needed to estimate the accuracy of the equation regression and its parameters.
   In next paragraph we will show analysis for quality of healthcare service using linear and multiple linear regression. Also, will show two cities of Kazakhstan which are Almaty and Nur-Sultan, at the end we will compare service quality of two city.

4. Software and tools

4.1 Introduction

All professional want to use the top tools for their responsibilities. Veteran professionals include the knowledge from the experiences of their careers while inexperienced persons look for guidance. Many study offer preferences founded on fame, cost, accessibility, data handling, visual representations, advancements, and technical community support and career opportunities. The preferences are suitable; on the other hand, the articles often take in favoritism and qualifiers that are not measurable. In response, the research of this section will focus on scientific and qualifiable attributes to offer comparison of multiple tools with a focus on performance. The likely tools for a data scientist are various. The first choice was derived from community information as well as the tools in the Southern Methodist University (SMU) Master of Science in Data Science set of courses. The communal data included research from famous sites devoted to data science and data analysis, Burtch Works and KDNuggets. An article by KDNuggets incorporated Python, R, and SAS in the top 4 tools for analytics and data mining [1]. In 2017, Burtch Works conduct a flash study with above one-thousand data expert to get the preferences for Python, R, and SAS [2]. As a outcome of the survey, this section will focus on these tools. To compare these 3 methods we use our ready data sources, without data our tools are do not working, this could be easily summarized by a quote by Tim O'Reilly in his article "What is Web 2.0" [3]: "Without the data, the tools are useless; without the software, the data is unmanageable." The presentation comparisons will contain data wrangling, visualization, and linear regression tasks with capacity on code difficulty, computing time, and computing property.

4.2 Literature review

According to the 2017 Data Scientists story by Crowd Flower, above 50% of time is used up collect, group, cleaning and organize data (Fig. 4 shows full time allocation) [4]. With the tall proportion of occasion invested in the start of the procedure, the requirement to select the correct tool is dominant for the effectiveness of a data scientist. Over four decades ago, formulas were developed

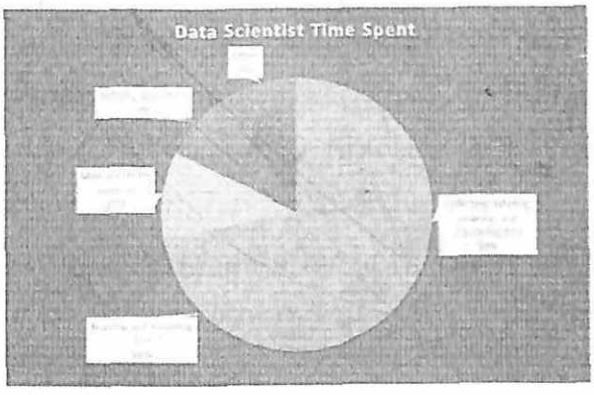


Figure 4.1: Time portion of data scientists derived from a survey conducted in February and March 2017 with 179 data scientists worldwide on behalf of varying companies.

to determine the difficulty of algorithms and languages. The found of this field is Maurice H. Halstead attributed with the metrics now known as the Halstead difficulty measures. Forceful investigation and testing were done on these capacity [5], [6], [7]. The number of operators (N1) and operands (N2) are recognized the length of with the unique operators (n1) and operands (n2). Calculations are then performed on these numbers to give the program terminology (n), program length (N), volume (V), difficulty (D), effort (E), and time (T).

$n = n_1 + n_2$ (1)

$N = N_1 + N_2$ (2)

$V = Nlog_2 n$ (3)

$D = \frac{n_1}{2} \times \frac{N_2}{n_2}$ (4)

$E = DV$ (5)

$T = \frac{E}{18}$ (6)

In this background, an operator has the skill to control and verify on the value of an operand; although an operand is either a numeric, text and/or Boolean values capable to be manipulate. There is not a strict gathering as to what defines an

operator and an operand consequently a single code script could return unlike counts of these attributes depending on the criterion used to select them [8]. The Cyclomatic difficulty model, also known as McCabe's complexity, was also developed in the 1970s [9]. The cyclomatic complexity focuses on the number of edges (e), vertices (nce), and connected components (p).

$$c(G) = e - n_c c + 2p \quad (7$$

Since the beginning of the models, criticism has been expressed on both models. One disquiet is that difficulty of code may stand for more than the complexity of the language. The complexity may stand for the difficulty of the tasks being perform by the code or less direct coding practice [10]. One more concern is the straight correlation between the difficulty and the lines of code [11]. In spite of the concerns, the complexity measurements by both Halstead and McCabe continue to be used.

4.3 Overview

4.3.1 Python

Python is an open resource general aim tool with applications for web, Internet, and software development; education and academia; numeric and scientific, to point out a few. Python—created by Guido Van Rossum as the successor of the ABC language and officially released in 1991—relies on the contribution of its wide community of users and developers self-identified as PUGs (Python User Groups)5 for its incessant development and increase. There is a scientific community of "well-established and growing group of scientists, engineers, and researchers using, extending, and promoting Python's use for scientific research" [12]. Python capabilities are extended through its healthy collection of packages. As of today, PyPI, also known as the "Cheese Shop."—the official package warehouse—has more than 100,000 junk mail stored 6 . Approximately explained, a package is a collection of modules that in turn contain definitions and statements to carry out functions or determine classes. In the field of data analysis some of the common packages [13] are: Pandas—ideal for data manipulation—; Statsmodels—for model and test—; scikit-learn—for categorization and machine learning tasks—; NumPy (Numerical Python)—for mathematical operations—and SciPy (Scientific

Python—for ordinary scientific tasks. A modern review [14] found that Python's NumPy, and SciPy packages were among the most favored ones for statistical analysis, while scikit-learn stood as a data mining favorite. Python also provide an extensive list of Integrated Development Environments (IDE). According to DataCamp [15] among of the top ones for data science: Spyder, a cross-platform IDE distributed through Anaconda (a "freemium" open source distribution for large-scale data); PyCharm integrates libraries such as NumPy and Matplotlib and provides support for JavaScript, KTML, CSS, Node.js, making it a good interface for web development; and Jupyter Notebook, before known as IPython, "offers an end-user environment for interactive work, a component to embed in other systems to provide an interactive control interface, and an abstraction of these ideas over the network for interactive distributed and parallel computing [16]."

4.3.2 R

The improvement of R was motivated by S with some programming influences from Scheme [17]. Two professors introduced the language to assist students with a more intuitive language, specifically lexical scoping which eliminates the necessity for global defining of variables [18]. Although the history R can find a foundation in FORTRAN, R is its own language. R is an interpreted language with code directly executed rather than compiled. Using a compiler, programmers can write interfaces for C, C++, and FORTRAN for efficiency. R is part of the GNU Project, which focused on free software allowing users the ability to run, redistribute, and improve the program [19]. Although initially criticized, R upgraded quickly with collaboration from around the globe. Since R is open source, the target audience is any user interested in statistical computing. R can be installed using Unix, Windows, or Mac. R is available for download via the Comprehensive R Archive Network (CRAN). The master site is in Austria; however, mirrored sites throughout the world distribute the load on the network. In addition to the software, the CRAN hosts provide supporting documentation and libraries with add-on packages. The open-source add-on packages, which are groups of functions developed by other users, are available on CRAN. As on January 27, 2017, the CRAN hosted more the 10,000 packages which does not include packages from other vendors [20]. Although no warranties are given by

R for any packages on CRAN, all the package contributions are reviewed by the CRAN team. Some packages in the libraries may restrict commercial use although the same packages may be openly available for education and research. RStudio is an IDE using packages (knitr and rmarkdown) to develop composed documents with the code and output from the R language. In addition, RStudio is an editor for LaTeX which is a markup language to produce high quality documents. A 2011 poll rated RStudio as the most used IDE with only the basic R console more frequently used [21].

4.3.3 SAS

SAS is a proprietary comprehensive statistical and data management tool developed by the SAS Institute: used internationally by government, private industry, and academia. "91 of the top 100 companies on the 2016 Fortune Global 500® are SAS customers."7 SAS is the largest privately-owned software company in the world. 8 Once an acronym for Statistical Analysis System: SAS has grown into much more than that and is no longer considered an acronym. It was originally created in 1966 for agricultural research work and later developed into a full-fledged system with the inception of SAS Institute. A study released in 2016 by MONEY and PayScale.com listed "Making Sense of Big Data" as the most valuable career skill now: with SAS as the top skill [22]. Current uses include business intelligence, and analysis of data in almost every business sector. SAS has various components and products that can be licensed along with Base SAS which is the core procedures and data management tool. SAS is a static typed language that uses the proprietary SAS dataset as the main table style data structure with only 2 data types numeric and character. SAS does not have the large user-written package library common to R and Python. There is however a huge user base that write and share code: it is just not included and centralized in the same way. SAS has its macro language which allows users to write code that can take various parameters and encapsulate code similar to functions in other languages. A user would then reference the code and call the macro. Unlike Python and R, this macro code is not compiled and does not get installed. It is SAS macro language and Base SAS code that a user can read and modify as needed. A good source for this type of code is GitHub. 9The SAS Global Forum, an annual conference for users by users, is a great source of SAS knowledge and code sharing. SAS

provides an extensive "Knowledge Base" on the support section of their website10 and has well-supported user support groups including the SAS-L list-serve hosted by the University of Georgia. 11 The main IDE for SAS is referred to as Foundation SAS or generally known as PC SAS. SAS Enterprise Guide was released several years ago: mainly known as a more point-and-click method of coding in SAS. More recently, SAS developed the SAS Studio which is a platform agnostic alternative that is based in Java and runs in a web browser from a licensed SAS install. SAS also recently introduced Jupyter integration where SAS can be run from Jupyter with a licensed install on the machine running Jupyter.

4.3.4 Ethics

The ethics adjoining this investigation are focused on the appearance of the facts and data of the tools without personal bias or influence. To eliminate potential bias in writing and results, all experiments and research was performed with a focus on the quantitative or qualitative data. An additional ethical perspective is added when reviewing the code of conduct from the ACM (Association for Computing Machinery) [23]. The code of conduct defined the contribution to society and human well-being as well as creating opportunities for members to learn the principles and limitations of computer systems. The research to assist less experienced individuals entering the field is an effort to assist with guidance by providing all of the resources applied on this paper as an educational tool. The source of any external code used in this paper was appropriately credited in the corresponding tool.

4.3.5 Quantifiable Experiments

The studied projects were identified to apply multiple traditional applications. The practices were performed on 1 machine. Each test run of the three (3) programming languages .

The projects we tried to test are represents a more usual data analysis task performing various analytical tasks. All data was obtained from the legal site of The Ministry of Health and Social Development of the Republic of Kazakhstan 13. A single file focusing on data about population, provision of medical stuff, provision of medical beds and hospitals, for the years between 2003 and 2016

Specifications	Machine
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
Ram	8 GB
OS	Windows 10 Pro

Figure 4.2: Machine specifications.

Tool	Version	installation
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
SAS	SAS Studio	Online

Figure 4.3: Software specifications for both test machines..

was used. The original file contained 224 observations and 10 variables for whole cities of Kazakhstan. In our study we used two mega cities which are Almaty and Nursultan. The data was preprocessed and cleaned before usage. Variables with less than half of the observations recorded were dropped, and unknown values were converted to actual missing values. The purpose of the task was the examining data analysis and multiple linear regression for healthcare service quality. At first, a model considering more than 5 regressors was defined to know the connection among variables and the most quantity of healthcare service quality involved in demography of population.

$$Y = _0 + _1x_1 + _2x_2 + + _kx_k + \quad (8)$$

In this section, the point of the model, though, was to evaluate its performance against other tools. The first code was wrote by Python and then replicated into R and SAS. The process started with the loading of the data. Firstly, data converted to csv file to upload. Table graphs was generated to imagine the information. A correlation table was developed, and a heatmap was constructed for imagination of the correlation analysis. To begin the linear regression, model variables were wrote for the factor variables. A linear regression was performed on 6 variables. The independent and dependent variables were chose based on ordinary least squares. For consistency every trial used the same six (6) variables selected from the SAS analysis. The final model with six (6) variables was trained using 80% of the data and tested with the remaining 20% of the data. The mean absolute error, mean squared error, root mean squared error, and variance were calculated to provide the analysis of the model.

4.3.6 Code Complexity

The rule defined by the Halstead Metrics was used to the programming for all three tools. The guiding principles were first defined for analysis C, so adaptations needed to be made. While programs exist for compute the metrics, the accessible applications do not used every time to all three programs. To make sure stability in the counting of operands and operators, the calculation was done by two or more persons with compromise on all tasks. The code was reviewed line by line with each appearance being identified alone. The total operators and operands were consolidated to provide a sum. The single operators and operands were then recognized. All calculations for other difficulty capacity were based on the operand and operator counts.

4.3.7 Qualifiable Research

There are definite considerations to be taken into explanation when choosing the correct application for executing data analysis tasks: A programming language for data analysis should be simply writeable and readable by people not by techniques only, able to handle a range of data types whether those are random in nature, have unusual options to run missing values correctly, and give at least basic mathematical and statistical functions such as the skill to produce arbitrary figures and probabilistic distributions, as well as high-level visualizations [24]. To assess the chosen tools, a matrix with these attributes was developed. All variables were incorporated except for the characteristic case. Simplicity was excluding for the reason that coding is a creative activity that may contrast in length and style. Also, previous programming experience would influence the awareness of simplicity.

4.3.8 Results

The experiments and investigate provided both experimental and qualifiable study as defined previously. Key aspects of the experiments and study were recognized to offer comparisons of the act. Ordinary dimensions in the experiments were number of lines in the code. The count eliminated all comments and white space and only alert on the needed code to perform the activities recognized. The

selection of lines and words was to demonstrate how the tools compare when looking at the quantity of code that is required to perform a task. Program running time measured at a variety of points and general was measured in "wall clock" time, the time it takes the process to finish on the given hardware. Machine glasses affect the performance in different ways relating on the tool. Using our machine allowed us to give a more balanced performance score of average machine. We measured running also on a Mac, however SAS is not accessible in local form or local install for the Mac. In general Mac users use Citrix or other similar virtual machine to run SAS; in which case the code is actually handing out on a remote engine and this would not be a high-quality comparison across tools. A lot of results were obtained from the experiments. The outline below highlights key outcome for the tools. The one test that we perform was to run the Healthcare system service data on the full Kazakhstan file with 2.5 million records. When we attempt this, SAS was the only one to complete the job. Python and R both threw errors once the data tired the RAM on the machine. This happened on other test machines also. This is by reason of the way data is stored while processing. Python and R both use RAM to store the data in a work space while processing. SAS, in comparison, uses the Hard Drive (HD) for work space data storage. This usually outcome in slower processing but the skill to handle more data on a usual machine.

4.3.9 Qualitative

The table below summarizes some of Huber's necessities of a statistical and data analysis tool. The first six share to universal aspects of the programming languages, while the last three match specifically to data analysis. According to Huber, a tool keen to data and statistical analysis should be extensible. R and Python take plus of the library of packages they sport to extend their capability outside their base language. Having users causative in the creation of packages is what keeps these tools in constant growth. On the other hand, caveats take in rapid criticism of packages, and lack of quality control in the development and use processes of the package creation. SAS has these extensions built-in in the base code avoiding the need to install and run an external application. Data handling capabilities in the case of this experiment are measured by the type of computational devices employed to perform this test, so in this background Python and

R's capabilities are restricted by the quantity of RAM. SAS is not forced by this feature as it runs directly on the hard drive. The three programs provide both on-line aid to the user at any time an error occurs, representing the sort of error and the correct location of the error in the code. Some packages in Python due to their design even provide alternatives to recover from an error. In this trial, during the data reading of the healthcare service files in csv in Python, an alert due to the overpassing of the default memory boundary was issued beside with a way to fix the error by changing the low_memory option from True to False. Python, R and SAS can handle quantities, characters, logical, compound, and random data types. All of them can be operated from either the command line or throughout one of their multiple IDEs. All three tools offer differ options to handle missing values. Python's package fancyimpute includes methods such as SimpleFill, MICE (Multivariate Imputation by Chained Equations) and SoftImpute; R offers the MICE, Amelia, Hmisc, mi libraries, while SAS deals with this through the procs MI and MIANALYZE. All three programs have methods that allow for the alteration or substitute of missing values by a experiment statistic or exact value as well as the reducing of rows or columns that enclose unidentified values in them. The three programming languages can calculate linear algebra functions, produce random information, do probabilistic distributions, and have high-level customizable image packages in the case of Python and R, while SAS has this feature integrated.

Attribute	Python	R	SAS
Packages Available	138,915	10,000	Integrated
Data Handling	RAM	RAM	Hard Drive
On-line Help	Yes	Yes	Yes
Error Log & Debug	Yes	Yes	Yes
Procedural & Object-oriented programming	CLI & IDE	CLI & IDE	CLI & IDE
Built-in Data Structures	Yes	Yes	Yes
Missing Values	Yes	Yes	Yes
Linear Algebra	Yes	Yes	Yes
Graphics	Yes	Yes	Yes

Figure 4.4: Qualitative attributes

5. Preparation of data

5.1 Analysis for Almaty

Indicators of quality management in health care have a direct relationship with the completeness and complexity of the system. In this regard, this thesis analyzes the main indicators that affect the quality of services, which are the figures of population, medical person, medical beds and, hospitals.

Regional Health:

1. Provision of the population with hospital beds (per 10,000 people):
2. Provision of the population with doctors (per 10,000 people of the population):
3. Provision of population with hospitals (per 10,000 people of the population):
4. Number of nursing staff per 1000 people.

General information about dataset: overall two regions own 14 observations and 5 variables. Observations are the years between 2003 and 2016 for each city. Variables are the indicators of quality services, namely: years, population, med_person, medical_bed and hospitals. (see)

5.1.1 Analysis of Correlation for indictors of quality services in Almaty

The first two tables tell us what variables were analyzed and their descriptive statistics for Almaty.

The third table contains the Pearson correlation coefficients and test results.

Obs	year	population	med_person	medical_bed	hospital
1	2003	1149641	11492	10947	53
2	2004	1175236	11162	10959	52
3	2005	1203485	11532	11029	53
4	2006	1247696	11342	11040	51
5	2007	1287246	11603	11337	51
6	2008	1324739	12157	10756	64
7	2009	1361677	13101	12201	63
8	2010	1391095	13611	10159	62
9	2011	1414017	13803	10471	69
10	2012	1450327	14192	10652	73
11	2013	1475579	14349	10080	94
12	2014	1507509	15945	10989	72
13	2015	1552349	16779	11011	76
14	2016	1713220	20160	11480	79

Figure 5.1: Table10.The indicators of service quality in Almaty

The CORR Procedure

5 Variables: year population med_person medical_bed hospital

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
year	14	2010	4.18330	28133	2003	2016
population	14	1375729	158285	19260158	1149641	1713220
med_person	14	13858	2504	194018	11192	20160
medical_bed	14	10936	537.27412	153109	10080	12201
hospital	14	68.14286	12.85934	954.00000	51.00000	94.00000

Pearson Correlation Coefficients – These numbers measure the strength and direction of the linear relationship between the two variables. The correlation coefficient can range from -1 to + 1. with -1 indicating a perfect negative correlation. + 1 indicating a perfect positive correlation. and 0 indicating no correlation at all. (A variable correlated with itself will always have a correlation coefficient of 1.) You can think of the correlation coefficient as telling you the extent to which you can guess the value of one variable given a value of the other variable. From the scatter plot of the variables read and write below, we can see that the points tend along a line going from the bottom left to the upper right, which is the same as saying that the correlation is positive. The 0.97965 is the numerical description of how tightly around the imaginary line the points lay. If the correlation was

Pearson Correlation Coefficients, N = [illegible]
Prob > |r| under H0: Rho=0

	year	population	med_person	medical_bed	hospital
year	1.00000	0.97865 <.0001	0.90980 <.0001	-0.10184 0.7316	0.85854 <.0001
population	0.97865 <.0001	1.00000	0.90040 <.0001	[illegible]	0.91884 0.0004
med_person	0.90980 <.0001	0.90040 <.0001	1.00000	0.07431 [illegible]	0.73269 0.0029
medical_bed	-0.10184 0.7016	-0.01095 0.9704	0.07431 0.8007	1.00000	-0.32027 0.2643
hospital	0.85854 <.0001	0.91884 0.0004	0.73269 0.0029	-0.32027 0.2643	1.00000

higher. the points would tend to be closer to the line: if it was smaller. they would tend to be further away from the line.

Also we the strength can be assessed by these general guidelines:

- .1 < $|r|$ < .3 small weak correlation
- .3 < $|r|$ < .5 medium / moderate correlation
- .5 < $|r|$ < large strong correlation

For example the variables med_person and population has a strong positive correlation with year variable. Also. medical_beds with med_person gave the large correlation. also between coefficients of medical_beds and hospital negative correlation.

5.1.2 Regression analysis for Almaty

This paragraph made analysis of medical stuff per people. According to the obtained estimations at the Table2 the linear equation $y = 160x - 7782$ show the average increase is (-0.016) doctors. which indicates a trend reduction in medical stuff capacity. The coefficient of determination is equal to 0.92 which is closer to 1: it means the model can work correctly. For example. if we predict that next year the demography of city will increase by 1900 million people then the number of needed medical person equal to 21820. The coefficient of variation is 5.45%. (< 33%) confirming the homogeneity of the model. It is observed that the lowest staff ensured years were between 2003 and 2005. Analysis of changes in the indicator allows highlight the peak of significant growth in medical stuff in 2016 which is bigger than for 17% comparing by 2016 year.

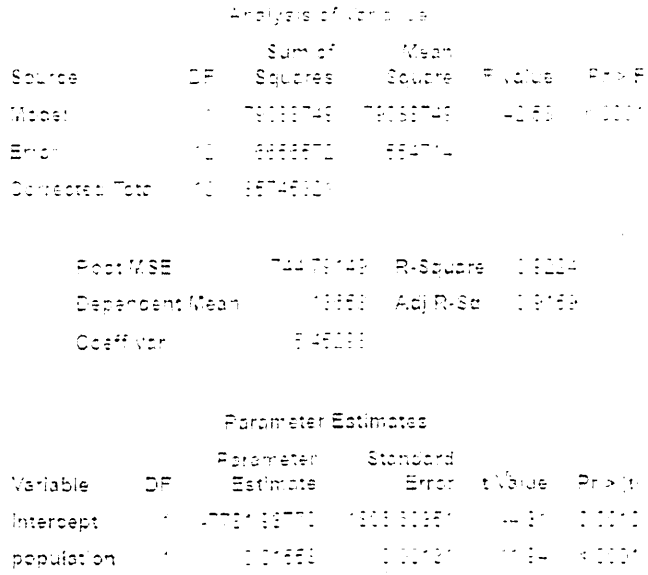


Figure 5.2: The estimation of regression analysis between medical person and population for 10000 people

In the analysis of the comparison. between 2003 and 2016 in healthcare area filled with 23 hospitals during. it means in every year given 2 hospitals to use. The peak number was in 2013 which is 94 hospitals. also in 2003 the figure of hospital where only 53. which is the lowest among studied years. Such a decrease and increase among number of healthcare houses related with politics and. can characterize high quality prophylactic activities of medical institutions. As part of a statistical analysis of the dynamics of indicators and building a linear model revealed the following trends. As shown at the table3 a linear model of the type y−146x− 1126 approximates that to each opened medical houses needed 146 medical people in average. and if there are no any opened hospital. then there are free 4126 medical person who might be work in different private hospitals and diagnostic centers. If next 2017 year there is given to use 2 hospitals. we need approximately 15950 medical person. In the next paragraph considered analyze between hospital and population.

The studied years the number of population only increased. From 2015 to 2016 there was big incensement in demography, the number of population filled directly with 160 thousand people. Within statistical analysis of the dynamics of indicators and build the linear model revealed the following trends. A linear view model $y = 0.000067x - 26.2$ approximates if the number of people in Almaty region increase by 1.8 million people then. there required 94 hospitals. Also by

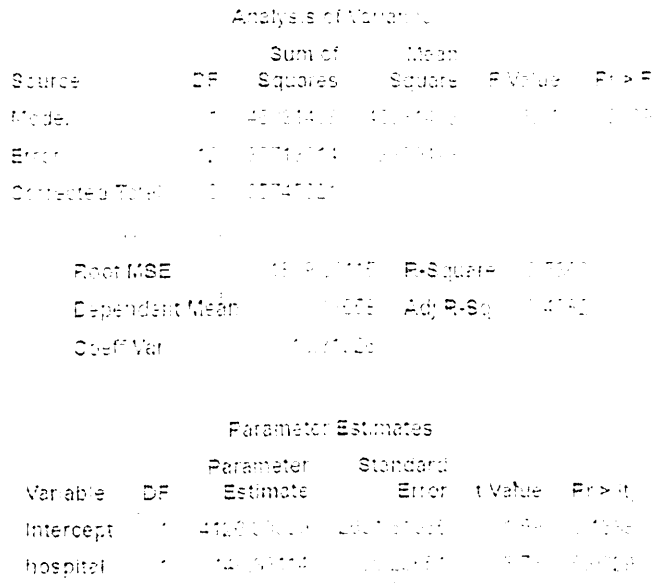


Figure 5.3: The estimation of regression analysis between doctors and hospital

regression model if we estimate $0.000067x = 1$ then to every 14925 people required 1 hospital facilities. In the next paragraph we will consider the multiple regression between number of hospital, number of medical bed and population.

If we look the regression model $y = -21.5x_1 + 0.25x_(2) + 0.017x_3 - 11039$ (x_1 is hospital and x_2 is medical bed and x_3 is population) predicted that the region has a lack of a hospitals, however medical bed is sufficient for 24%. If next year these three variables will increase for 10%, it requires medical doctors should also increase by 10%. Let's calculate: $y = -21.5 * 87 + 0.25 * 12628 + 0.017 * 1884542 - 11039 = 22285$. It is satisfied with increased medical stuff numbers. So there is no any luck with healthcare services. Coefficient variations equal to 5.5%, since this indicator is not more than 33%, it means that the aggregate is quantitatively homogeneous.

5.2 Correlation analysis for indicators of quality services of Nur-Sultan

The Table 6 described the general used datasets for Astana city and it has 14 observation time series from 2003 to 2016, also has the indicator variables as year, population, medical person, medical bed and hospital.

The first two tables tell us what variables were analyzed and their descriptive statistics for Nur-Sultan.

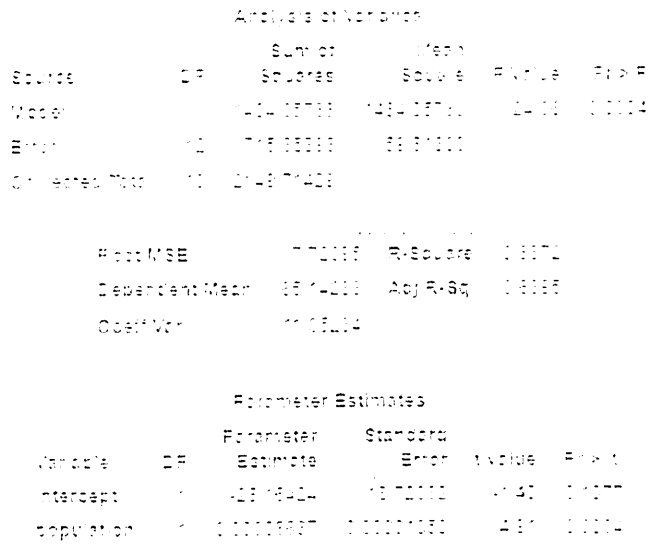


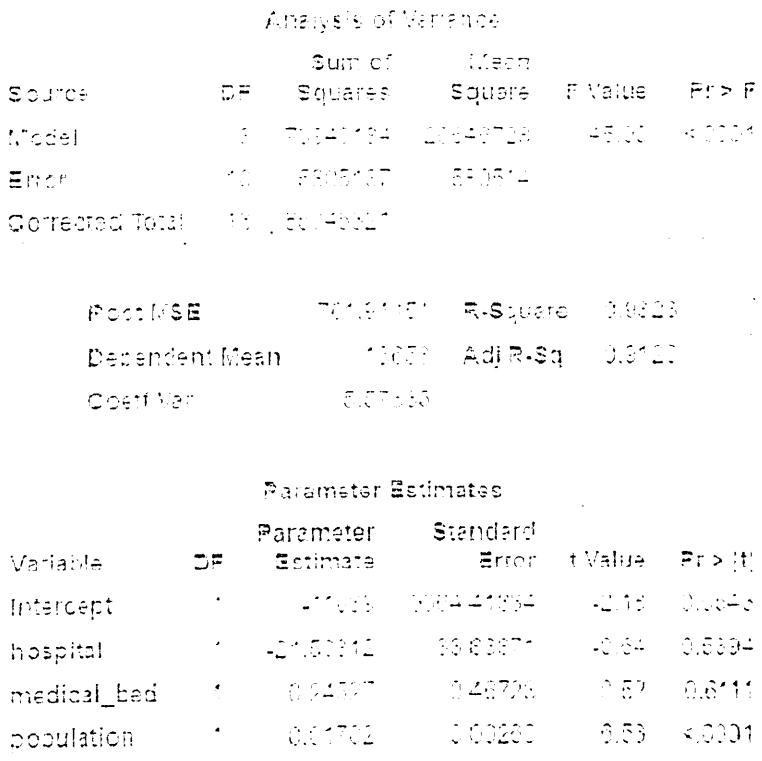
Figure 5.1: The estimation of regression analysis between medical stuff and hospital

The third table contains the Pearson correlation coefficients and test results. This table presents us the relationship between studied observations. We must check the correlation for coefficient of population to coefficients of med_person, medical_bed, and hospital. Third table present us that:

- Population and med_person (0.92) has strong positive relationship
- Population for medical_bed share also strong positive ration which is equal to 0.89 and 0.
- Population with hospital has a moderate relation (0.58).

5.2.1 Regression Analysis for Nur-Sultan

This paragraph made analysis of medical stuff per population. According to the obtained estimations at the Table7 the linear equation $y = 0.0089x - 413$ show the average increase of doctors for 0.89%. The coefficient of determination is equal to 0.85 which is closer to 1: it means the model can work correctly. For example, if we predict that next year the demography of city will increase by 900.000 people then the number of required medical person is approximately 7597. The coefficient of variation is 12.9%. ($< 33\%$) confirming the homogeneity of the model. It is observed that the lowest staff ensured years were 2003 and the highest is 2016.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	[illegible]	[illegible]	[illegible]	[illegible]
Error	[illegible]	[illegible]	[illegible]		
Corrected Total	[illegible]	[illegible]			

Root MSE	[illegible]	R-Square	[illegible]
Dependent Mean	[illegible]	Adj R-Sq	[illegible]
Coeff Var	[illegible]		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	[illegible]	[illegible]	[illegible]	[illegible]
hospital	1	[illegible]	[illegible]	[illegible]	[illegible]
medical_bed	1	[illegible]	[illegible]	[illegible]	[illegible]
population	1	[illegible]	[illegible]	[illegible]	[illegible]

Figure 5.5: The estimation of multiple regression analysis for independent variable medical person.

Figure 5.6: The indicators of service quality in Nur-Sultan region

Analysis of changes in the indicator allows highlight the peak of significant growth in medical stuff in 2016 which is bigger than for 64% comparing by 2003 year. (see)

In the analysis of the comparison. between 2003 and 2016 in healthcare area filled with 12 hospitals. The peak number was in 2016 which is 31 hospitals. also in 2003 the figure of hospital where only 19. which is the lowest among studied years. Such a decrease and increase among number of healthcare houses related with politics and. can characterize high quality prophylactic activities of medical institutions. As part of a statistical analysis of the dynamics of indicators and building a linear model revealed the following trends. As shown at the table8 a linear model of the type $y = 314x - 2675$ approximates that to each opened medical houses needed 314 medical people in average, and if there are no any opened hospital. then there are luck of medical person. If next year ministry gives hospitals to use. we need approximately 7687 medical person. In the next paragraph considered analyze between hospital and population. The time series

The CORR Procedure

5 Variables: year population med_person medical_bed hospital

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
year	14	2010	4.18330	28133	2003	2016
population	14	641576	175115	8982206	201888	872655
med_person	14	5255	1651	73618	2784	7780
medical_bed	14	5242	1214	73389	3329	6777
hospital	14	24.57143	4.23746	344.00000	19.00000	31.00000

Pearson Correlation Coefficients, N = 14
Prob > |r| under H0: Rho=0

	year	population	med_person	medical_bed	hospital
year	1.00000	0.98530 <.0001	0.99304 <.0001	0.95948 <.0001	0.60752 0.0212
population	0.98530 <.0001	1.00000	0.98080 <.0001	0.89156 <.0001	0.58548 0.0278
med_person	0.99304 <.0001	0.98080 <.0001	1.00000	0.93556 <.0001	0.59109 0.0260
medical_bed	0.95948 <.0001	0.89156 <.0001	0.93556 <.0001	1.00000	0.68503 0.0069
hospital	0.60752 0.0212	0.58548 0.0278	0.59109 0.0260	0.68503 0.0069	1.00000

analysis outlined that the staffs' rate only increased and no fluctuation among studied years. In 2007 and 2008 the numbers of medics hold constant 4675. In 2016 the figure of medics reached the peak and showed 7780 doctors, comparing by 2003 it shared 179% of changing.

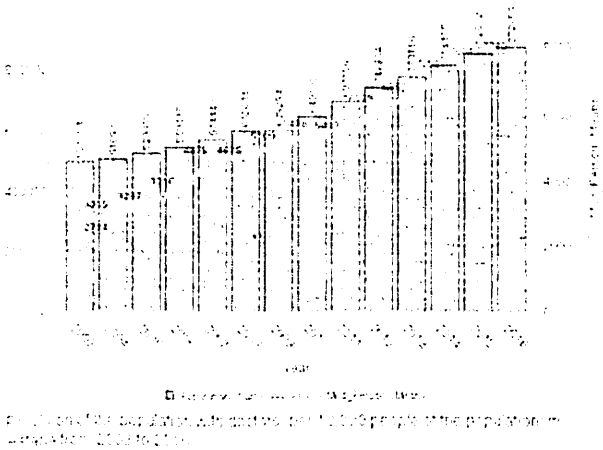
The studied years the number of population only increased among 2003 and 2016 there was 77% change. From 2003 to 2004 there was big incensement in demography. the number of population filled directly with 300 thousand people. Within statistical analysis of the dynamics of indicators and build the linear model revealed the following trends. A linear view model $y = 0.000019x + 12.5$ approximates if the number of people in Nursultan region increase by 900 thousand people next year then, there required 30 hospitals. It means there is no luck of hospital. Also by regression model if we estimate $0.000019x = 1$ then to every 52632 people required 1 hospital facilities. In the next paragraph we will consider the multiple regression between number of hospital. number of medical bed and population.

If we look the regression model $y = -5.7x_1 + 0.76x_(2) + 0.004x_3 - 1450$ (x_1 is hospital and x_2 is medical bed and x_3 is population) predicted that the region

Pearson Correlation Coefficients, N = 14
Prob > |r| under H0: Rho=0

	year	population	med_person	medical_bed	hospital
year	1.00000	0.98530	0.99304	0.95948	0.60752
		<.0001	<.0001	<.0001	0.0212
population	0.98530	1.00000	0.92080	0.89158	0.58548
	<.0001		<.0001	<.0001	0.0278
med_person	0.99304	0.92080	1.00000	0.93556	0.59109
	<.0001	<.0001		<.0001	0.0260
medical_bed	0.95948	0.89158	0.93556	1.00000	0.68503
	<.0001	<.0001	<.0001		0.0069
hospital	0.60752	0.58548	0.59109	0.68503	1.00000
	0.0212	0.0278	0.0260	0.0069	

Figure 5.7: The estimation of regression analysis between medical person and population for Nur-Sultan



has a lack of a hospitals, however medical bed is sufficient for 75%. If next year these three variables will increase for 10%, it requires medical doctors should also increase by 10%. Let's calculate: $y = -5.7 * 34 + 0.76 * 7455 + 0.004 * 959921 - 1450 = 6192$. It is not satisfied with increased medical stuff numbers must be 8558. So there is luck of medical stuffs in Nur-Sultan. Coefficient variations equal to 10.86%, since this indicator is not more than 33%, it means that the aggregate is quantitatively homogeneous.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	25085926	25085926	25.82	0.0003
Error	12	11657879	971490		
Corrected Total	13	36743805			

Root MSE	985.64169	R-Square	0.6827
Dependent Mean	5258.42857	Adj R-Sq	0.6563
Coeff Var	18.74404		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-2374.65247	1383.22346	-1.69	0.1169
hospital	1	313.73767	61.74059	5.06	0.0003

Figure 5.8: The estimation of regression analysis between doctors and hospital

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	158.88813	158.88813	19.87	0.0008
Error	12	95.99901	7.99992		
Corrected Total	13	254.88714			

Root MSE	2.82797	R-Square	0.6234
Dependent Mean	25.28571	Adj R-Sq	0.5921
Coeff Var	11.18405		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	12.47713	2.97186	4.20	0.0012
population	1	0.00001996	0.00000446	4.46	0.0008

Figure 5.9: The estimation of regression analysis between medical stuff and hospital

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	33509925	11169642	34.53	< .0001
Error	10	3234880	323488		
Corrected Total	13	36743805			

Root MSE	568.78005	R-Square	0.9120
Dependent Mean	5268.42857	Adj R-Sq	0.8855
Coeff Var	10.79616		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1450.51535	944.57442	-1.54	0.1556
hospital	1	5.63877	73.73656	0.08	0.9400
medical_bed	1	0.75659	0.36447	2.08	0.0646
population	1	0.00406	0.00199	2.03	0.0692

Figure 5.10: The estimation of multiple regression analysis for independent variable medical person.

6. References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] Strategy "Kazakhstan-2050" New political course of the established state. https: 'primeminister.kz 'cnpage/article-101

[3] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[4] S. Y. Wong ; K. L. Tsui ; K. S. Chin ; M. Xu. A simulation study to achieve healthcare service quality improvement in accident amp; emergency department (AED). 2011 IEEE International Conference on Quality and Reliability. Year: 2011. Pages: 259 - 263

[5] Tan-Hsu Tan ; Munkhjargal Gochoo ; Sukhbaatar Bilgee ; Ching-Su Chang ; Jin-Jia Hu ; Yung-Fu Chen ; JohnY. Chiang ; Yung-Fa Huang ; Ming-Hui Lee ; Yung-Nian Hsu ; Jin-Chyr Hsu. Development of an emergency medical service system based on wireless networks and real-time traffic information. 2012 International Conference on Computerized Healthcare (ICCH). Year: 2012. Pages: 35 - 42.

[6] S. A. M. A. Junid, S. Aljunid, S. Sulong, S. M. Hazmah, A. M. Nur and H. R. Mustafa. "Cloud computing: Feasibility of developing web-based UNU Casemix Grouper," 2012 IEEE Business, Engineering Industrial Applications Colloquium (BEIAC), Kuala Lumpur, 2012, pp. 399-403.

[7] V. Kannan, J. C. Fish and D. L. Willett, "Agile model driven development of electronic health record-based specialty population registries." 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Las Vegas, NV, 2016, pp. 465-468.

[8] V. Georgiana, D. Kartawiguna and Indrajani, "Evaluation of radiology data

warehouse implementation on education, research, and quality assurance." 2016 International Conference on Information Management and Technology (ICIMTech). Bandung, 2016, pp. 277-280.

[9] A. A. Samah et al., "Decision Support System Using System Dynamics Simulation Modelling for Projection of Dentist Supply." 2014 International Conference on Computer Assisted System in Health, Kuala Lumpur, 2014, pp. 22-25.

[10] W. Liu, T. Mundie, U. Krieger, E. K. Park and S. S. Zhu, "Rapid delivery e-Health service (RDeHS) platform," 2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom), Munich, 2016, pp. 1-6.

[11] S. Wang, "Medical Service Quality Evaluation Based on Lingual Description," 2010 International Conference on Management and Service Science, Wuhan, 2010, pp. 1-3.

[12] Sen Wang, Ying Chen, Miao Song and Lu-lin Zhou, "Research on the influence elements of medical service quality based on DEMATEL," 2011 International Conference on Computer Science and Service System (CSSS), Nanjing, 2011, pp. 1646-1648.

[13] Chaang-Yung Kung, Tzung-Ming Yan and Chih-Cheng Huang, "Applying GRM on the decision process of service quality in Medical Center," 2007 International Conference on Intelligent and Advanced Systems, Kuala Lumpur, 2007, pp. 286-289.

[14] L. Min and G. Yubing, "To study medical service quality with structural-equation model," 2011 International Conference on E-Business and E-Government (ICEE), Shanghai, China, 2011, pp. 1-4.

[15] P. Lin, C. Huang, C. Chan and Y. Yang, "Location-awareness Communications System for Improving Medical Service Quality," 2008 10th International Conference on Advanced Communication Technology, Gangwon-Do, 2008, pp. 682-686.

[16] Z. Lu-lin, C. Ying and H. Xi-jian, "Research on the construction of medical service quality evaluation index system in the perspective of patients," 2010 International Conference on Management Science Engineering 17th Annual Conference Proceedings, Melbourne, VIC, 2010, pp. 305-310.

[17] J. P. Bigus et al., "Information technology for healthcare transformation," in IBM Journal of Research and Development, vol. 55, no. 5, pp. 6:1-6:14, Sept.-

Oct. 2011.

[18] G. Corotinschi and V. G. Gaitan. "The use of IoT technologies for providing high-quality medical services." 2017 21st International Conference on System Theory, Control and Computing (ICSTCC). Sinaia, 2017, pp. 285-290.

[19] N. Maglaveras et al.. "The citizen health system (CHS): a Modular medical contact center providing quality telemedicine services." in IEEE Transactions on Information Technology in Biomedicine, vol. 9, no. 3, pp. 353-362, Sept. 2005.

[20] C. Wu, I. Khoury and H. Shah. "Optimizing Medical Data Quality Based on Multiagent Web Service Framework," in IEEE Transactions on Information Technology in Biomedicine, vol. 16, no. 4, pp. 745-757, July 2012.

[21] Ying Su and L. Al-Hakim. "Assuring information quality in medical platform for U-Healthcare service." 2010 IEEE International Conference on Wireless Communications, Networking and Information Security, Beijing, 2010, pp. 664-668.

[22] N. Philip and R. S. H. Istepanian. "Medical Quality of Service for Wireless Ultrasound Streaming in Robotic Tele-Ultrasonography System," 2007 IEEE International Conference on Networking, Sensing and Control, London, 2007, pp. 245-250.

[23] R. Rahman, M. R. Rifat, S. Moutushy and H. Shahid Ferdous. "Toward a customizable effective patient management system for ensuring quality medical service," 2011 IEEE Student Conference on Research and Development, Cyberjaya, 2011, pp. 415-419.

[24] K. Ganapathy and V. Vaidehi, "Medical intelligence for quality improvement in Service Oriented Architecture," 2011 International Conference on Recent Trends in Information Technology (ICRTIT), Chennai, Tamil Nadu, 2011, pp. 161-166.

[25] D. Guimin, W. Yu, Z. Li and H. Ma, "Improving medical service quality based on the Critical Incident Technique." ICSSSM12, Shanghai, 2012, pp. 31-36.

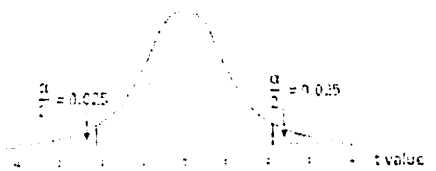
7. Conclusion

In this study considered analysis for healthcare service quality. Study considered quantity of hospital facilities for every 10000 population. As well as considered the average number of medical staff per hospital. Also studied medical staff per population. The results of study can be used in the shade of the number of medical institutions and the training of the necessary number of medical personnel in the healthcare system. For these used regression method and as software used SAS. Python and R programming languages. These methods study were used on the example of two cities of Kazakhstan Almaty and Nursultan.

A. Appendix

Student's t Distribution Table

For example the t value for 18 degrees of freedom is 2.101 for 95% confidence interval (2-Tail $\alpha = 0.05$).



	90%	95%	97.5%	99%	99.5%	99.95%	1-Tail Confidence Level
	80%	90%	95%	98%	99%	99.9%	2-Tail Confidence Level
	0.100	0.050	0.025	0.010	0.005	0.0005	1-Tail Alpha
df	0.20	0.10	0.05	0.02	0.01	0.001	2-Tail Alpha
1	3.0777	6.3138	12.7062	31.8205	63.6567	636.6192	
2	1.8856	2.9200	4.3027	6.9646	9.9248	31.5991	
3	1.6377	2.3534	3.1824	4.5407	5.8409	12.9240	
4	1.5332	2.1318	2.7764	3.7469	4.6041	8.6103	
5	1.4759	2.0150	2.5706	3.3649	4.0321	6.8688	
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.9588	
7	1.4149	1.8946	2.3646	2.9980	3.4995	5.4079	
8	1.3968	1.8595	2.3060	2.8965	3.3554	5.0413	
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.7809	
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.5869	
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.4370	
12	1.3562	1.7823	2.1788	2.6810	3.0545	4.3178	
13	1.3502	1.7709	2.1604	2.6503	3.0123	4.2208	
14	1.3450	1.7613	2.1448	2.6245	2.9768	4.1405	
15	1.3406	1.7531	2.1314	2.6025	2.9467	4.0728	
16	1.3368	1.7459	2.1199	2.5835	2.9208	4.0150	
17	1.3334	1.7396	2.1098	2.5669	2.8982	3.9651	
18	1.3304	1.7341	2.1009	2.5524	2.8784	3.9216	
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.8834	
20	1.3253	1.7247	2.0860	2.5280	2.8453	3.8495	
21	1.3232	1.7207	2.0796	2.5176	2.8314	3.8193	
22	1.3212	1.7171	2.0739	2.5083	2.8188	3.7921	
23	1.3195	1.7139	2.0687	2.4999	2.8073	3.7676	
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.7454	
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.7251	
26	1.3150	1.7056	2.0555	2.4786	2.7787	3.7066	
27	1.3137	1.7033	2.0518	2.4727	2.7707	3.6896	
28	1.3125	1.7011	2.0484	2.4671	2.7633	3.6739	
29	1.3114	1.6991	2.0452	2.4620	2.7564	3.6594	
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.6460	