Diabetes Ther (2019) 10:2079–2093
https://doi.org/10.1007/s13300-019-00684-1

ORIGINAL RESEARCH

Forecasting the Prevalence of Diabetes Mellitus Using Econometric Models

Assel Mukasheva · Nurbek Saparkhojayev · Zhanay Akanov · Amy Apon · Sanjay Kalra

Received: July 1, 2019 / Published online: September 13, 2019

ABSTRACT

Introduction: The prevalence of diabetes in Kazakhstan has reached epidemic proportions, and this disease is becoming a major financial burden. In this research, regression analysis methods were employed to build models for predicting the number of diabetic patients in Kazakhstan in 2019, as this should aid the costing and policy-making performed by medical institutions and governmental offices regarding diabetes prevention and treatment strategies.

Methods: A brief review of mathematical models that are potentially useful for the task of interest was performed, and the most suitable methods for building predictive models were selected. The chosen models were applied to explore the correlation between population growth and the number of patients with diabetes as well as the correlation between the increase in gross regional product and the growth in the number of patients with diabetes. Moreover, the relationship of population growth and gross domestic product with the growth in the number of patients with diabetes in Kazakhstan was determined. Our research made use of the scikit-learn library for the Python programming language and functions for regression analysis built into the Microsoft Excel software.

Results: The predictive models indicated that the prevalence of diabetes in Kazakhstan will increase in 2019.

Conclusion: Mathematical models were used to find patterns in a comprehensive statistical dataset on registered diabetes patients in Kazakhstan over the last 15 years, and these patterns were then used to build models that can accurately predict the prevalence of diabetes in Kazakhstan.

Enhanced Digital Features To view enhanced digital features for this article go to https://doi.org/10.6084/m9.figshare.9618371.

A. Mukasheva (✉)
Department of Cybersecurity, Data Processing and Storage, Satbayev University, Almaty, Kazakhstan
e-mail: mukasheva.a.82@gmail.com

N. Saparkhojayev
Dean of Engineering Faculty, Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkestan, Kazakhstan

Z. Akanov
President of Kazakh Society for Study of Diabetes, Member of AASD, Almaty, Kazakhstan

A. Apon
Professor, Chair of the Computer Science Division, Clemson University, Clemson, SC, USA

S. Kalra
Department of Diabetes and Endocrinology, Bharti Hospital, Karnal, India

Keywords: Data analysis; Diabetes mellitus; Forecasting; Python; Regression analysis; Scikit-learn; Statistics

INTRODUCTION

The increasing incidence of diabetes worldwide is of great concern and has attracted the attention of many researchers [1, 2]. All types of diabetes pose a high risk of premature death and are a serious problem. Official statistics published by the World Health Organization indicate that 422 million people [3] were suffering from this disease in 2014; this number is forecast to rise to more than 690 million people in 2045 [4]. In Kazakhstan alone, the number of diabetic patients is believed to exceed 300,000, and this figure only includes patients who were directly diagnosed by doctors [5]. There are serious problems in this country due to a lack of qualified specialists in diabetes, meaning that diabetes is often first treated during the advanced rather the early stages of the disease. These shortcomings have led to an increase in diabetic patients in Kazakhstan, and diabetes mellitus (DM) is currently the fourth most prevalent disease in the country [6]. DM is thus a growing public health problem that affects not only human health but the health care system overall and, indeed, the global economy [7].

The steady growth in the number of diabetes patients has prompted researchers around the world to explore methods permitting the prediction and early diagnosis of diabetes. For instance, the prevalence of chronic kidney disease in European patients with diabetes until 2025 was predicted in [8], and a demographic epidemiological model of Singapore was devised in [9] and then used to predict the overall prevalence of type 2 diabetes in Singapore until 2050. While measures to tackle the rising prevalence of diabetes can and are being taken at the state level, multisectoral efforts to treat this disease are needed to optimize socioeconomic productivity. The authors of [10] argue that the pandemic of diabetes cannot be solved without the participation of all of the stakeholders concerned, including diabetic patients and the community.

A model for predicting type 2 diabetes based on data-mining methods was proposed in [11]. This model consisted of an improved *k*-means algorithm and a logistic regression algorithm. Other researchers have developed a model for predicting the prevalence and incidence of obesity and diabetes as well as the direct costs of treating diabetes and its complications [12]. In another study, a population-based analysis of an elderly cohort was carried out to investigate whether oral antidiabetic agent use can decrease the risk of dementia in type 2 diabetes patients, and the correlation of the incidence of dementia to the duration of diabetes was explored [13]. In Taiwan, models for estimating the diabetes-associated risk of hospitalization and the risk of type 2 diabetes inpatient mortality were devised in order to facilitate the identification of at-risk patients [14]. Three forecasting machines were used in [15] to anticipate the glycemic impacts of various meals: a data assimilation machine, a model averaging the data assimilation results, and a machine utilizing dynamic Gaussian process model regression. The forecasted glycemic impacts were found to correlate well with glucose indicators, and the prediction accuracy of the technique was as good or better than expected. Computer simulations can also help researchers to understand chronic disease progression. In this context, the authors of [16] developed and used system dynamics to perform diabetes system modeling.

An important direction in the development of medical services for populations is the construction and implementation of various problem-oriented information systems that can utilize all of the heterogeneous information collected during the diagnosis and treatment of patients with diabetes and apply "big data" technology and cloud services as a toolkit. There is a high demand from modern medical institutions for such systems that use the latest information technologies to facilitate the diagnosis and treatment of diabetes mellitus [17–20].

The purpose of the study described below was to identify the most effective regression analysis method for predicting the growth in

the number of patients with diabetes in Kazakhstan using ins passive detection and real statistical data on such patients

METHODS

Statistical Analysis

Data on diabetic patients in Kazakhstan were provided by a public foundation, the Kazakh Society for the Study of Diabetes [5], which provided informed consent for the publication of all statistical data used in this study.

Data on gross domestic product (GDP) and the population of Kazakhstan were taken from the official website of the Statistical Agency of the Republic of Kazakhstan [21]. The aims of this study was to build a model that could predict the growth in the number of diabetes patients in Kazakhstan via passive detection and regression analysis methods, and to identify the most accurate experimental method for predicting diabetes. Data on patients with diabetes in Kazakhstan from 2004 to 2018 (see Table 1) were used.

Table 1 The total number of patients with diabetes mellitus (DM) in Kazakhstan during each of the last 15 years

Year	Number of patients with DM in Kazakhstan
2004	114,355
2005	117,563
2006	128,039
2007	147,717
2008	151,336
2009	162,012
2010	175,685
2011	190,682
2012	207,935
2013	226,202
2014	261,453
2015	272,629
2016	293,171
2017	310,114
2018	326,449

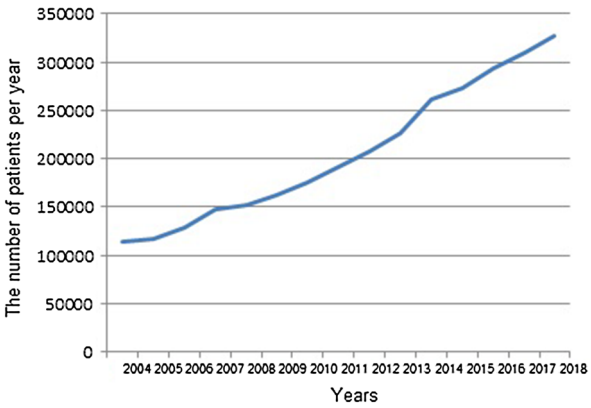

Fig. 1 Data from the register of patients with DM in Kazakhstan

From 2004 to 2018, the number of patients with diabetes increased from approximately 114,000 to approximately 326,000 (185.46%), as shown in Fig. 1.

Based on this graph, we can conclude that there is a positive growth trend in the number of diabetic patients in Kazakhstan. The largest jump in the number of diabetic patients was observed in 2014—an increase of 35,251 people (15.58%).

RESULTS

Correlating Population Growth with the Number of DM Patients and the Increase in the Gross Regional Product with the Growth in the Number of DM Patients in Kazakhstan

It is often necessary to explore the relationship between continuous variables. This can be probed using correlation analysis, as illustrated by Table 2, which presents the correlation between population growth and the number of DM patients in each region of Kazakhstan. The final result of correlation analysis is a correlation coefficient (r), the value of which can range from -1 to $+1$. A correlation coefficient of $+1$ indicates that there is a strong positive linear relationship between two variables, a

Table 2 The correlation coefficient for population growth versus number of DM patients in each region of Kazakhstan

Population growth in the region	Number of patients with DM in the region						
	Akmola	**Aktobe**	**Almaty**	**Atyrau**	**West Kazakhstan**	**Jambyl**	**Karaganda**
Akmola	− 0.2856						
Aktobe		0.96606					
Almaty			0.9023				
Atyrau				0.97796			
West Kazakhstan					0.9445		
Jambyl						0.95496	
Karaganda							0.98211
Kostanay							
Kyzylorda							
Mangistau							
South Kazakhstan							
Pavlodar							
North Kazakhstan							
East Kazakstan							
Astana city							
Almaty city							

Population growth in the region	Number of patients with DM in the region								
	Kostanay	**Kyzylorda**	**Mangistau**	**South Kazakhstan**	**Pavlodar**	**North Kazakhstan**	**West Kazakhstan**	**Astana city**	**Almaty city**
Akmola									
Aktobe									
Almaty									
Atyrau									

Table 2 continued

Population growth in the region	Number of patients with DM in the region								
	Kostanay	**Kyzylorda**	**Mangistau**	**South Kazakhstan**	**Pavlodar**	**North Kazakhstan**	**West Kazakhstan**	**Astana city**	**Almaty city**
West Kazakhstan									
Jambyl									
Karaganda									
Kostanay	− 0.7566								
Kyzylorda		0.94915							
Mangistau			0.89454						
South Kazakhstan				0.99043					
Pavlodar					0.84446				
North Kazakhstan						− 0.9056			
East Kazakstan							− 0.8861		
Astana city								0.99274	
Almaty city									0.992031

Table 3 The correlation coefficient for growth in GRP versus growth in the number of patients with DM in each region of Kazakhstan

Growth in GRP of region	Growth in the number of DM patients in the region							
	Akmola	Aktobe	Almaty	Atyrau	West Kazakhstan	Jambyl	Karaganda	Kostanay
Akmola	0.83894							
Aktobe		0.94669						
Almaty			0.99265					
Atyrau				0.94822				
West Kazakhstan					0.95174			
Jambyl						0.96482		
Karaganda							0.98211	
Kostanay								0.93
Kyzylorda								
Mangistau								
South Kazakhstan								
Pavlodar								
North Kazakhstan								
East Kazakstan								
Astana city								
Almaty city								

Growth in GRP of region	Growth in the number of DM patients in the region							
	Kyzylorda	Mangistau	South Kazakhstan	Pavlodar	North Kazakhstan	West Kazakhstan	Astana city	Almaty city
Akmola								
Aktobe								
Almaty								
Atyrau								
West Kazakhstan								

Table 3 continued

Growth in GRP of region	Growth in the number of DM patients in the region							
	Kyzylorda	**Mangistau**	**South Kazakhstan**	**Pavlodar**	**North Kazakhstan**	**West Kazakhstan**	**Astana city**	**Almaty city**
Jambyl								
Karaganda								
Kostanay								
Kyzylorda	0.86412							
Mangistau		0.92202						
South Kazakhstan			0.98541					
Pavlodar				0.98318				
North Kazakhstan					0.9804			
East Kazakstan						0.99302		
Astana city							0.99545	
Almaty city								0.994719

Table 4 The number of DM patients in each region of Kazakhstan in the year 2019, as predicted using three regression analysis methods

Numerical label for region	Region	Linear regression	Polynomial regression	Exponential regression
1	Akmola	16,110	16,187	17,957
2	Aktobe	14,107	15,189	16,150
3	Almaty	31,222	30,867	36,074
4	Atyrau	8882	9419	10,580
5	West Kazakhstan	9446	10,225	10,817
6	Jambyl	17,172	19,725	18,472
7	Karaganda	31,500	31,569	33,915
8	Kostanay	22,444	25,198	23,805
9	Kyzylorda	10,453	11,852	13,304
10	Mangistau	9399	10,609	11,624
11	South Kazakhstan	39,085	38,867	42,957
12	Pavlodar	18,612	20,445	20,082
13	North Kazakhstan	17,554	17,137	19,236
14	East Kazakhstan	34,932	35,993	37,296
15	Astana	15,254	16,240	17,976
16	Almaty	36,837	40,552	39,702
17	The Republic of Kazakhstan	333,010	350,074	369,945

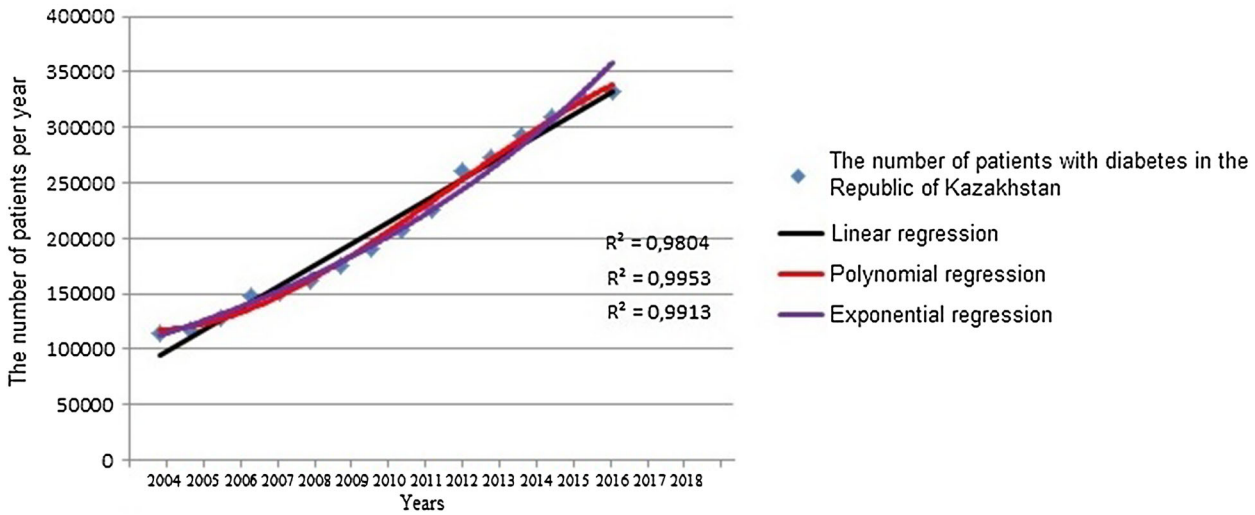

Fig. 2 Plot showing the number of DM patients in Kazakhstan each year from 2004 to 2018, as well as three different regression lines fitted to the data. The regression lines were used to predict the number of DM patients in Kazakhstan in 2019

correlation coefficient of − 1 indicates that the variables have a strong negative linear relationship, and a correlation coefficient of 0 means that there is no linear relationship between the variables [21–24].

The regions of Kostanay, North Kazakhstan, and East Kazakhstan were all found to show relatively strong negative correlations between population growth and the number of DM patients, while the Akmola region showed a weak negative correlation between those parameters. This shows that there are strong negative linear relationships between population growth and the number of DM patients in Kostanay, North Kazakhstan, and East Kazakhstan, and that there is a weak negative linear relationship between these parameters for the Akmola region. Positive correlations between the two variables are seen for the other regions of Kazakhstan.

The correlation between the growth in the gross regional product (GRP) and the growth in the number of diabetic patients in each region of the country was also analyzed; the corresponding correlation coefficients are shown in Table 3. GRP is a general indicator of the economic activity of the region, i.e., the amount of goods and services produced in that region [25].

According to Table 3, there is a strong positive linear relationship between the growth in GRP and the growth in the number of patients with DM in each region of Kazakhstan.

Literature Review of Methods Used to Construct Predictive Models

The standard tool used in medical research (indeed, in all areas of research) to explore correlations between variables is regression analysis [26, 27]. For instance, the authors of [28] performed studies to detect anomalies in surveillance data and concluded that while the number of studies that use more sophisticated methods such as machine learning methods and hidden Markov models is increasing, studies that use traditional methods such as control charts and linear regression remain more popular.

In [29], predictive methods based on machine learning were compared with those based on traditional statistical methods. The empirical results of this comparison highlighted the need for objective and unbiased approaches to testing the performance of forecasting methods. This can be achieved by comparing the predictions afforded by the various forecasting methods when they are all applied to the same task, and by analyzing a large dataset (e.g., a large number of time series in the present work), as this should lead to fair and meaningful comparisons and definite conclusions.

The application of methods based on regression analysis to build predictive models will be successful if there is a known correlation between two variables of interest. Our correlation analysis revealed that there was a strong positive linear relationship between growth in GRP and growth in the number of patients with DM in Kazakhstan. The next step was to apply three types of regression analysis to predict the growth in the number of patients with diabetes mellitus based on passive detection: linear regression, polynomial regression, and exponential regression. If the value of one of the parameters considered is known to a high level of accuracy, we can use these three regression equations to determine the value of another parameter that is related to the first parameter [30].

Forecasting the Growth in the Number of DM Patients in Kazakhstan in 2019 Using Three Types of Regression Analysis

Regression methods are statistical methods for studying the distribution of a dependent variable in relation to one or more independent variables [31]. The aim of regression analysis is to build a mathematical model that allows the value of a dependent variable to be estimated from the values of independent variables [32]. Such a model incorporates regression coefficients that are identified by constructing a regression line—a line of best fit to the distribution of the dependent variable in relation to the independent variable(s). In the present work, we used various types of regression lines—linear, third-degree polynomial, and exponential—to achieve the best fit to the distribution. In each case, the best variant of the regression equation was chosen by identifying the variant with highest coefficient of determination

R^2 [30]. Many methods of determining the parametric relationship between a dependent variable and independent variables have been developed. These methods usually differ in the shape of the function used in parametric regression and the distribution of the error term in the regression model. Examples include linear regression, logistic regression, and Poisson regression [33].

In the present work, we applied the three regression methods to a situation with one dependent and one independent variable using the machine-learning library scikit-learn of the programming language Python. Particular attention was paid to verifying that the conditions required for the appropriate application of the methods were present.

1. In linear regression analysis, the parameters of a straight line that can be used to accurately predict the value of one variable based on the value of the other variable are predicted.

The straight line has the formula

$$y = \beta_0 + \beta_1 x,$$

where y is the value of one of the variables, β_0 is the point at which the straight line crosses the y-axis, β_1 is the slope of the line, and x is the value of the other variable. Linear regression analysis is performed if correlation analysis reveals a relationship between the variables [24, 34]. The linear regression equation that was used as a model for predicting the number of DM patients took the following form:

$$y = 15915x + 78368, \text{with } R^2 = 0.9804.$$

2. Polynomials are widely used in situations where a curvilinear response is observed. Even the most complex nonlinear relationships can be adequately modeled by polynomials across a fairly narrow range of x values.

A regression equation based on a third-degree polynomial takes the following form:

$$y = ax^3 + bx^2 + cx + d,$$

where the number of extrema (maxima, minima, and inflection points) presented by the curve is determined by the degree of the polynomial [34, 35]. The polynomial regression equation that was used as a model for predicting the number of DM patients took the following form:

$$y = -38.378x_3 + 1487.6x_2 + 780.81x + 113349, \text{ with } R^2 = 0.9964.$$

3. Exponential regression involves regression functions of the following form:

$$y = a^* m^x = a^* (e^{\ln(m)})^x = a^* e^{x*\ln(m)} = a^* e^{bx}, \text{ where } b = \ln(m).$$

The exponential regression equation used as a model for predicting the number of DM patients took the following form [36]:

$$y = 102666e^{0.0796x}, \text{ where } R^2 = 0.995.$$

After calculating the regression equations, they were used to predict the number of DM patients in each region of Kazakhstan in the year 2019; these data are presented in Table 4.

According to the predicted data for 2019 obtained using linear regression, there will be 333,010 DM patients in Kazakhstan. According to polynomial regression, there will be 350,074 DM patients in Kazakhstan in 2019, but, according to exponential regression, there will be 369,945 DM patients. After obtaining these data, the regression model plot shown in Fig. 2 was generated.

As shown in Fig. 2, all three types of regression had high coefficients of determination, i.e., R^2 was always above 0.9, although polynomial regression yielded the highest R^2 value. From this, it follows that the polynomial model is best suited for use as a model for predicting the number of DM patients.

Relationship of Population Growth and GDP to the Growth in the Number of Patients with DM

A regression analysis was performed to determine the relationship of population growth and GDP to the growth in the number of diabetic patients in Kazakhstan. The model used took the following form:

$$y = a + b_1x_1 + b_2x_2 + \varepsilon.$$

More precisely, it was found to be

$$y = -128438 + 14.47913x_1 + 0.003268x_2 + \varepsilon,$$

where y is the number of patients with DM, and $R^2 = 0.98$ is the coefficient of determination. This is the proportion of variance of the dependent variable, explained by the model of dependence under consideration, i.e., the explanatory variables. The determination coefficient (R^2; $0 \le R^2 \le 1$) is a measure of the quality of the regression model; i.e., how well it describes the relationship between the dependent and independent variables of the model. The closer the value of the coefficient of determination is to 1, the better the model. If $R^2 = 1$, then the empirical points (x_i; y_i) lie exactly on the regression line and there is a linear functional relationship between variables Y and X. If $R^2 = 0$, then all of the variation of the dependent variable is due to factors not taken into account in the model.

In the present research, the model shows the relationship between growth in GDP and the DM population. x_1 is the population. We used the F test to determine the statistical significance of all the coefficients. F was calculated to be 496.4881, meaning that all of the coefficients were statistically significant. x_2 is the GDP, and the constant $a = -128{,}438$. There can be functions where one variable depends on the values of two or more other variables, where x_1 and x_2 together determine the value of y. The value of a shows that if x_1 and x_2 are equal to zero then y will equal to zero too. In the present research, a has a negative value, which means that if there is no population growth and the economy expands, the number of patients with DM will decrease. $b_1 = 14.47913$, which shows that iif the economy does not expand, the number of DM patients will increase by approximately 14,000 people as compared with the number in the previous year. $b_2 = 0.003268$, which shows that if the population continues to grow at the same rate the following year and the rate of GDP increases by 1 million tenge, the number of DM patients will be increasing by this ratio.

Using the scikit-learn Library of Python to Model Regression Methods

One of the most popular programming languages, Python, can be used to implement machine learning algorithms [37]. Python has a well-documented library named scikit-learn [38] that can be employed for machine learning. The scikit-learn library can be applied to tasks such as clustering, cross-validation, correlation, dimension reduction, algorithmic compositions, feature extraction, feature selection, optimization of algorithm parameters, and multiple learning.

To demonstrate the utilization of this library, let us consider an example in which it is used to determine the number of patients with diabetes based on data obtained from the statistical data register of the Republic of Kazakhstan. The first step is to download the required data. The scikit-learn library is used to model data, not to download the data. However, the Pandas library [39] can be used to download data as it has convenient functions for I/O and the processing of tabular data, and this library can also perform primary data analysis.

Data loading code:

```
data = pd.read_csv('registr_bolnyh', header=None, na_values='?')
X = data.drop([0], axis=1)
Y = data[0]
```

The next step is to work with arrays, where X is an array of signs Y is an array of classes. The subsequent step is to normalize features, since most machine-learning algorithms are based on gradient methods. After downloading the necessary data, researchers can use the capabilities of machine-learning algorithms. The scikit-learn

library implements a variety of algorithms, such as logistic regression, linear regression, naive Bayes, *k*-nearest neighbors, decision trees, and the support vector method. The scikit-learn library can easily be integrated into applications that perform traditional statistical data analysis, as well as other types of applications, as it relies on the Python scientific software ecosystem. Algorithms implemented in a high-level language can be used as building blocks in a range of applications, such as in medical imaging [40].

The first experiment, which used a linear regression algorithm, employed the following code:

```
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
x = x[:, np.newaxis]
y = y[:, np.newaxis]
model = LinearRegression()
model.fit(x, y)
y_pred = model.predict(x)
plt.scatter(x, y, s=10)
plt.plot(x, y_pred, color='r')
plt.show()
```

After training the model, it is easy to predict the number of patients according to the input attribute using the *predict* method. In the second experiment, the following polynomial regression algorithm was used:

```
import operator
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.preprocessing import PolynomialFeatures
x = x[:, np.newaxis]
y = y[:, np.newaxis]
polynomial_features= PolynomialFeatures(degree=3)
x_poly = polynomial_features.fit_transform(x)
model = LinearRegression()
model.fit(x_poly, y)
y_poly_pred = model.predict(x_poly)
rmse = np.sqrt(mean_squared_error(y,y_poly_pred))
r2 = r2_score(y,y_poly_pred)
print(rmse)
print(r2)
plt.scatter(x, y, s=10)
# sort the values of x before line plot
sort_axis = operator.itemgetter(0)
sorted_zip = sorted(zip(x,y_poly_pred), key=sort_axis)
x, y_poly_pred = zip(*sorted_zip)
plt.plot(x, y_poly_pred, color='m')
plt.show()
```

Finally, a third experiment that applied the following exponential regression algorithm was carried out:

```
import numpy as np
>>> x = np.array([1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15])
>>> y = np.array([114355, 117563, 128039, 147717, 151336, 162012, 175685, 190682,
207935, 226202, 261453, 272629, 293171, 310114, 326449])
>>> p1 = np.polyfit(x, np.log(y), 1)
#    y ≈ exp(-0.401) * exp(0.105 * x) = 0.670 * exp(0.105 * x)
# (^ biased towards small values)
print (p1)
#    y ≈ exp(1.42) * exp(0.0601 * x) = 4.12 * exp(0.0601 * x)
# (^ not so biased)
```

DISCUSSION

The authors examined the main features of the scikit-learn library that were used to solve machine-learning problems. Results obtained from Python were then compared with those obtained using Excel. In comparison, it was found that all three regression methods yielded similar predicted values regardless of whether Python or Excel was used, although there was a difference of 16 patients between the results obtained with exponential regression using Python (369,961 patients) and Excel (369,945 patients). Since the values predicted using the different regression analysis methods and Excel or Python are rather similar to each other and are quite close to the actual DM populations reported for Kazakhstan in recent years, and given the approximate nature of these forecasting techniques, it appears that it is feasible to use regression analysis methods to accurately predict the DM population in Kazakhstan.

The regression model would be more reliable if the statistical data for DM patients in Kazakhstan were obtained on a monthly basis rather than an annual basis. This is one limitation of this study.

CONCLUSION

In this work, we reviewed many studies that used regression analysis for forecasting purposes. This review led us to conclude that regression analysis methods are effective techniques for solving various problems, including many in the field of medicine. We therefore tested three different regression models as possible tools for predicting the number of patients with diabetes in Kazakhstan in 2019. All of the models indicated that the number of DM patients will increase, which is concerning. Strong correlations of population growth and GDP with the growth in the number of patients with diabetes in Kazakhstan were observed, and the relationship between population growth and the number of diabetics was determined. A correlation between the growth in GRP and the growth in the number of patients with diabetes was also discerned. The main features of the scikit-learn library that can be applied to machine learning problems were considered in the context of predicting the number of patients with diabetes in Kazakhstan using regression analysis methods. This research is part of a larger research project that focuses not only on the prediction of the number of diabetic patients in 2019 but also on the diagnosis and study of diabetes using big data technologies. Although some results of this research have been studied and published, investigations into the use of big data technology in the health sector are ongoing.

ACKNOWLEDGEMENTS

Funding. No funding or sponsorship was received for this study or publication of this article.

Authorship. All named authors meet the International Committee of Medical Journal Editors (ICMJE) criteria for authorship for this article, take responsibility for the integrity of the work as a whole, and have given their approval for this version to be published.

Disclosures. Sanjay Kalra is a member of the journal's Editorial Board. Assel Mukasheva, Nurbek Saparkhojayev, Zhanay Akanov, and Amy Apon have nothing to disclose.

Compliance with Ethics Guidelines. Informed consent was obtained from the Kazakhstan Society for the Study of Diabetes (a public fund) for the publication of all statistical data used in this study.

Data Availability. The analyzed datasets are available from the corresponding author on reasonable request.

REFERENCES

1. Gao Y, Wang Y, Zhai X, He Y, Chen R, Zhou J, et al. Publication trends of research on diabetes mellitus and T cells (1997–2016): a 20-year bibliometric study. PLoS ONE. 2017;12(9):e0184869.

2. Abutaleb MH. Diabetes mellitus: an overview. Pharm Pharmacol Int J. 2016;4(5):406–11. https://doi.org/10.15406/ppij.2016.04.00087.

3. World Health Organization. Health topics: diabetes. https://www.who.int/diabetes/en/.

4. International Diabetes Federation. IDF diabetes atlas. 8th ed. Brussels: International Diabetes Federation; 2017.

5. Kazakhstan Society for the Study of Diabetes. Official website. https://www.kssd.site/.

6. World Health Organization. Diabetes profiles in countries, 2016. Weblink: https://www.who.int/diabetes/country-profiles/kaz_ru.pdf?ua=1.

7. Deepthi B, Sowjanya K, Lidiya B, et al. A modern review of diabetes mellitus: an annihilatory metabolic disorder. J In Silico In Vitro Pharmacol. 2017;3:1.

8. Kainz A, Hronsky M, Stel VS, Jager KJ, Geroldinger A, Dunkler D, Heinze G, Tripepi G, Oberbauer R. Prediction of prevalence of chronic kidney disease in diabetic patients in countries of the European Union up to 2025. Nephrol Dial Transpl. 2015;30(4):113–8. https://doi.org/10.1093/ndt/gfv073.

9. Phan TP, Alkema L, Tai ES, et al. Forecasting the burden of type 2 diabetes in Singapore using a demographic epidemiological model of Singapore. BMJ Open Diabetes Res Care. 2014;2:e000012. https://doi.org/10.1136/bmjdrc-2013-000012.

10. Kalra S, Akanov Z, Pleshkova A. Thoughts, words, action: the alma-ata declaration to diabetes care transformation. Diabetes Ther. 2018;9(3):873–6. https://doi.org/10.1007/s13300-018-0440-2.

11. Wu H, Yang S, Huang Z, He J, Wang X. Type 2 diabetes mellitus prediction model based on data mining. Informatics in medicine unlocked, vol. 10. Amsterdam: Elsevier; 2018. https://doi.org/10.1016/j.imu.2017.12.006.

12. Huang ES, Basu A, O'Grady M, Capretta JC. Projecting the future diabetes population size and related costs for the US. Diabetes Care. 2009;32(12):2225–9. https://doi.org/10.2337/dc09-0459.

13. Kim JY, Ku YS, Kim HJ, Trinh NT, Kim W, Jeong B, Lee EK. Oral diabetes medication and risk of dementia in elderly patients with type 2 diabetes. Diabetes Res Clin Pract. 2019;154:116–23. https://doi.org/10.1016/j.diabres.2019.07.004.

14. Li TC, Li CI, Liu CS, Lin WY, Lin CH, Yang SY, Chiang JH, Lin CC. Development and validation of prediction models for the risks of diabetes-related hospitalization and in hospital mortality in patients with type 2 diabetes. Metabolism. 2018;85:38–47. https://doi.org/10.1016/j.metabol.2018.02.003.

15. Albers DJ, Levine M, Gluckman B, Ginsberg H, Hripcsak G, Mamykina L. Personalized glucose forecasting for type 2 diabetes using data assimilation. PLoS Comput Biol. 2017;13(4):e1005232. https://doi.org/10.1371/journal.pcbi.1005232.

16. Mishra V, Samuel C, Sharma SK. System modeling for forecasting of diabetes prevalence. Indian J Public Health Res Dev. 2018;9(7). https://doi.org/10.5958/0976-5506.2018.00628.9.

17. Saparkhojayev N, Mukasheva A (2018) The development of information system of formation and use of information resources for evaluation of parameters and evaluation of recommendations based on big data technology tools: work with Mongo DB. In: International Conference on Cyber Security and Computer Science (ICONCS'18); 2018 Oct 18–20; Safranbolu, Turkey.

18. Saparkhojayev N, Mukasheva A, Saparkhojayev P (2017) The concept of monetization of IoT-based project: case of medical system in Kazakhstan. In: 15th International Scientific Conference on Information Technologies and Management; 2017 Apr 27–28; ISMA University, Riga, Latvia.

19. Saparkhojayev N, Mukasheva A (2018) Introduction to BigData technology for diagnosis of diabetes. In: 16th International Scientific Conference on Information Technologies and Management; 2018 Apr 26–27; ISMA University, Riga, Latvia.

20. Saparkhojayev N, Mukasheva A, Tussupova B, Zimin I. Development of the information system based on BigData technology to support endocrinologist-doctors for diagnosis and treatment of diabetes in Kazakhstan. In: 6th International Smartcity Symposium; 2018 Oct 15–17; Palm Garden Hotel, Putrajaya, Malaysia.

21. Committee on Statistics, Ministry of National Economy of the Republic of Kazakhstan. GDP data on official website. http://stat.gov.kz/faces/homePage?_adf.ctrl-state=2fn371p1u_4&lang=ru&_afrLoop=8562077751869222.

22. Bingham NH, Fry JM. Regression. Linear models in statistics. London: Springer; 2010. https://doi.org/10.1007/978-1-84882-969-5 **(ISBN 978-1-84882-968-8)**.

23. Gogtay NJ, Thatte UM. Principles of correlation analysis. J Assoc Physicians India. 2017;65:78–81.

24. Rumyantsev PO, Saenko VA, Rumyantseva UV. Statisticheskie metody analiza v klinicheskoy praktike. Chast' I. Odnomernyy statisticheskiy analiz [Statistical methods of the analysis in clinical practice. Part I. One-dimensional statistical analysis]. Problemy endokrinologii – Endocrinology Problems. 2009;55(5):48–55.

25. Committee on Statistics, Ministry of National Economy of the Republic of Kazakhstan. GRP data on official website: http://stat.gov.kz/faces/wcnav_externalId/homeNationalAccountIntegrated?lang=ru&_afrLoop=9050851751248091#%40%3F_afrLoop%3D9050851751248091%26lang%3Dru%26_adf.ctrl-state%3Dzudoooxo_4.

26. Vach W. Regression models as a tool in medical research. Abingdon: Taylor & Francis; 2013 **(ISBN-13: 978-1-4665-1749-3)**.

27. Goldberg MA, Cho HA. Introduction to regression analysis. Southampton: WIT Press; 2004 **(ISBN-13: 978-1853126246)**.

28. Yuan M, Boston-Fisher N, Luo Y, Verma A, Buckeridge DL. A systematic review of aberration detection algorithms used in public health surveillance. J Biomed Inform. 2019;94:103181. https://doi.org/10.1016/j.jbi.2019.103181.

29. Makridakis S, Spiliotis E, Assimakopoulos V. Statistical and machine learning forecasting methods: concerns and ways forward. PLoS ONE. 2018;13(3):e0194889. https://doi.org/10.1371/journal.pone.0194889.

30. Sindeeva LV, Medvedeva NN, Nikolaev VG, Strelkovich NN, Orlova II (2013) Application of regression analysis methods in the biomedical researches. Bull New Med Technol. 2013;20(2):S216–S219.

31. Andersen PK, Skovgaard LT. Regression with linear predictors. New York: Springer; 2010. https://doi.org/10.1007/978-1-4419-7170-8 **(ISBN 978-1-4419-7169-2)**.

32. Cherkashina YA (2015) Application of regression analysis for solving diagnosis problem of children's health. Mod Prob Sci Educ. 2015;1(1).

33. Yan X, Su X. Linear regression analysis: Theory and computing. Singapore: World Scientific; 2009. ISBN-13: 978-981-283-410-2. https://doi.org/10.1142/6986

34. Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis, vol. 5. Hoboken: Wiley; 2012. p. 672 **(ISBN: 978-0-470-54281-1)**.

35. 4analytics. 3 ways to calculate a polynomial in Excel. https://4analytics.ru/trendi/3-sposoba-rascheta-polinoma-v-excel.html.

36. Excel2. Least square method: exponential dependence in MS Excel. https://excel2.ru/articles/mnk-eksponencialnaya-zavisimost-v-ms-excel.

37. Python Software Foundation. The official home of the Python programming language. https://www.python.org/.

38. Scikit-learn Authors. Scikit-learn: Machine learning in Python. https://scikit-learn.org/stable/.

39. PANDAS Project Core Team. Python Data Analysis Library. https://pandas.pydata.org/.

40. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.