Ciencia en la frontera: revista de ciencia y tecnología de la UACJ.
Volumen XIV, pp. 15-25, 2017 / Impresa en México

ISSN 2007-042X

Establishing a stochastic regression model to analyze diabetes incidence rates in the U.S.

Héctor A. Quevedo Urías,[1] Jorge A. Salas-Plata Mendoza,[1] Angelina Domínguez Chicas,[1] Michel Y. Montelongo Flores[2] y L. Susana Alonso López[1]*

Recibido: 26 de mayo de 2016; aceptado: 20 de enero de 2017.

RESUMEN

Este estudio aplica técnicas de regresión estadística para establecer un modelo de pronóstico estocástico en las tasas de incidencia de diabetes en Estados Unidos. La metodología utiliza una muestra aleatoria de treinta y cuatro años (1980-2014) de las tasas de incidencia de diabetes proporcionadas por los Centros para el Control y Prevención de Enfermedades (CDC, por sus siglas en inglés). El primer paso consistió en dibujar los datos de casos de diabetes (millones) contra el tiempo (años), para verificar el tipo de función de los registros. La consideración de un modelo de regresión lineal simple no fue aceptable, ya que su diagnóstico y medidas de precisión no fueron satisfactorios. La opción de un modelo de regresión polinomial fue más adecuado, pero no del todo. Sin embargo, el modelo de regresión logarítmico transformado fue más satisfactorio, porque sus diagnósticos objetivos y subjetivos fueron superiores. El modelo con transformaciones logarítmicas fue el mejor de todos los modelos analizados según los resultados obtenidos.

Palabras clave: modelo de regresión estadística, modelo de regresión polinomial, modelo de regresión logarítmico transformado, modelos de regresión para analizar tasas de incidencia de diabetes.

ABSTRACT

This study applied statistical regression techniques to establish a stochastic prognostic model for analyzing diabetes incidence rates in the U.S. The methodology used a 34-year random sample (1980-2014) of diabetes incidence rates apportioned by Centers for Disease Control and Prevention (CDC). The first step consisted in plotting the data of cases of diabetes (millions) *versus* time (years) to check on the type of function followed by the records. The fitting of a simple linear regression model was not acceptable because its diagnostics and measures of accuracy were not satisfactory. The fitting of a polynomial regression model was more satisfactory, but not quite right. However, the resulting logarithmic transformed regression model was even more satisfactory because its objective and subjective diagnostics were more acceptable. The model with logarithmic transformations was then the best candidate model according to the obtained results.

1 *Departamento de Ingeniería Civil y Ambiental, Universidad Autónoma de Ciudad Juárez. Ciudad Juárez, Chihuahua, México.*
2 Facultad de Ingeniería, Universidad Autónoma de Chihuahua.
* Autor para correspondencia: hquevedo@uacj.mx ; Instituto de Ingeniería y Tecnología, Av. del Charro 450 norte, Col. Partido Romero, CP 32310; Ciudad Juárez, Chihuahua, México; Tel. +52 (656) 688 48 46.

Keywords: regression statistical modeling, polynomial regression modeling, logarithmic transformed regression modeling, regression models to analyze diabetes incidence rates.

INTRODUCTION

This study applied statistical regression analyses aimed at establishing a stochastic model for analyzing diabetes incidence rates in the U.S. The procedure used a 34-year random sample (1980-2014) of diabetes incidence rates apportioned by Centers for Disease Control and Prevention (CDC). The first step consisted in plotting the data of diabetes cases (expressed in millions) *versus* time (years) to identify the type of function followed by these records. Next, the procedure consisted in fitting a simple linear regression model followed by its evaluation to assess its predictive quality. Within this approach the methodology also prepared time series graphical analyses. The third step consisted in fitting a quadratic polynomial regression model along with its objective and subjective evaluations to assess its fitting capability. This procedure included time series analyses with their evaluated measures of accuracy MAPE, MAD, and MSD. The fourth step consisted in fitting a time series logarithmic transformed model along with its complementary residual graphs. In addition, this approach included its objective and subjective diagnostics. All these steps were applied to control the experimental errors aimed to optimize the fitting quality of the selected model. The results showed that in all of the models tested, the time series logarithmic transformed model was the best candidate because its objective and subjective diagnostics were more acceptable.

Continuing with this introductory note, the American Diabetes Association (ADA) affirms that 1.4 million Americans are diagnosed with diabetes every year. This organization also affirms that symptomatic diabetes was the seventh leading cause of death in the United States in 2010 based on the 69,071 death certificates in which diabetes was listed as the underlying cause of death. In 2010, diabetes was mentioned as a cause of death in 234,051 certificates (ADA, 2015). Furthermore, the CDC gives some epidemiological estimations on racial and ethnic percentage of people aged 20 years or older with diagnosed diabetes by race/ethnicity in the United States period 2010-2012. For example, among non-Hispanic whites the incidence rate was 7.6%. Besides, among Asian-Americans the incidence rate was 9.0%. By the same token, among Hispanics the rate was 12.8%. Similarly, among non-Hispanic blacks the incidence rate was 13.2%. Finally, the CDC affirms that among American indians/Alaska natives the rate amounted to 15.9% (CDC, 2015). Likewise, a CDC national diabetes statistics (2014) affirms that 29.9 million people or 9.3% of the U.S. population have diabetes. Furthermore, in relation to ethnic differences among people aged 20 years or older, American indians/Alaska natives led the number of cases in 15.9%. Further reports of this source of information give some figures on the estimated diabetes costs in the United States, in 2012. For example, the total (direct and indirect) costs amount to US $245 billion. Direct medical costs amount to $176 billion after adjusting for population age and sex differences. Average medical expenditures among people with diagnosed diabetes were 2.3 times higher than people without diabetes. Also, indirect costs amounted to $69 billion due to disabilities, work losses, and premature deaths. Still further, the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) says that diabetes affects an estimated 29.1 million people in the United States and that it is the seventh leading cause of death. It says that diabetes can affect many parts of the body with its associated serious complications, such as heart diseases and strokes, blindness, kidney failure, and lower limb amputations. This NIDDK further affirms that type 1 diabetes affects approximately 5 percent of adults and the majority of children and youth with diagnosed diabetes. Moreover, this source says that type 2 diabetes is the most common form of the disease, accounting for about 90 to 95 percent of diagnosed diabetes cases in U.S. adults (NIDDK, 2015). Type 2 diabetes is also increasingly being diagnosed in children and adolescents, and disproportionately affects minority youth. Finally, source says that prediabetes affects an estimated 86 million adults in the United States. Those with prediabetes are at high risk of developing type 2 diabetes (CDC, 2014).

The National Institute of Diabetes of the United Kingdom (NIDUK) affirms that diabetes is the fastest growing health threat of our times and an urgent public health issue. It says that since 1996, the number of people living with diabetes has more than doubled. This source says that if nothing changes, it is estimated that over five million

people in the U.K. will have diabetes on the next years (NIDUK, 2014). Similarly, according to the International Diabetes Foundation (IDF), diabetes is a leading threat to global health and economic development. According to IDF, the disease now affects over 300 million people worldwide and will cost the global economy at least $376 billion in 2010, or 11.6% of the total world healthcare expenditure. A further 344 million people are at risk of developing type 2 diabetes, the most common form of the disease. If nothing is done to reverse this epidemic, IDF predicts that by 2030, 438 million people will live with diabetes at a cost projected to exceed $490 billion (IDF, 2010).

The International Diabetes Federation affirms that China now is the country with the largest number of people with diabetes. Previous estimates in the IDF's *Diabetes Atlas* Fourth Edition —published in October, 2009— put the number of people with diabetes in China at 43.2 million based on the best evidence available at the time (IDF, 2014). Now, it would appear that China has overtaken India becoming the global epicenter of the diabetes epidemic with 92.4 million adults with the disease (CDC, 2010).

According to CDC, the number of Americans with diabetes symptoms is projected to double or triple by the year of 2050. It is affirmed that as many as 1 in 3 U.S. adults could develop diabetes by 2050 if current trends continue, according to a new analysis made by the same Center. This office says that 1 in 10 U.S. adults has diabetes now. The prevalence is expected to rise sharply over the next 40 years due to an aging population more likely to develop type 2 diabetes, increases in minority groups that are at high risk for type 2 diabetes, and people with diabetes living longer, according to CDC projections published in the journal *Population Health Metrics*. Further on, the report predicts that the number of new diabetes cases each year will increase from 8 per 1,000 people in 2008 to 15 per 1,000 in 2050. Additionally, the report estimates that the number of Americans with diabetes will range from 1 in 3 to 1 in 5 by 2050. "These are alarming numbers that show how critical it is to change the course of type 2 diabetes," said Ann Albright, Ph.D., RD, director of CDC's Division of Diabetes Translation. "Successful programs to improve lifestyle choices on healthy eating and physical activity must be made more widely available, because the stakes are too high and the personal toll too devastating to fail" (Lebech-Cichosz, Johansen, & Hejlesen, 2015). Insofar, as predictive models to related to manage diabetes and its complications, in the *Journal of Diabetes Science and Technology* the investigators Lebech-Cichosz *et al.* (2015) of the Department of Health Science and Technology at Aalborg University, in Aalborg, Denmark, affirm that statistical models or complex pattern recognition models may be fused into predictive models that combine patient information and prognostic outcome results. They contend that such knowledge could be used in clinical decision support, disease surveillance, and public health management to improve patient care. These investigators further affirm that predictive models have been developed for management of diabetes and its complications, and the number of publications on such models has been growing over the past decade. They add that multiple logistic or linear regression models can be used for prediction model development, possibly owing to its transparent functionality (Lebech-Cichosz *et al.* (2015).

METHODOLOGY

The methodology used a 34-year sample data of diabetes cases corresponding to the 1980-2014 period. Table 1 below shows the required information, where the term cases refer to the number of persons with diabetes symptoms expressed in millions (CDC, 2015 *bis*). The first step consisted in plotting the data of cases of diabetes (millions) *versus* time (years) to check on the type of function followed by the records. The second step consisted in fitting a simple linear regression model followed by its objective and subjective evaluations. This was followed by a time series graphical analysis. The third step consisted in fitting a quadratic polynomial regression model along with its objective and subjective evaluations. This procedure included time series analyses with their evaluated measures of accuracy MAPE, MAD, and MSD. The fourth step consisted in fitting a time series logarithmic transformed model along with its complementary residual graphs. Also, this approach included the objective and subjective diagnostics. The final step consisted in fitting a time series model with transformed values, excluding the outlier case of year 1996. All these steps were applied to control the experimental errors aimed in order to optimize the predicting quality of the selected model.

Table 1. *Table showing the time in years starting from 1980 to 2014 of symptomatic diabetes incidence rates expressed in millions of cases.*

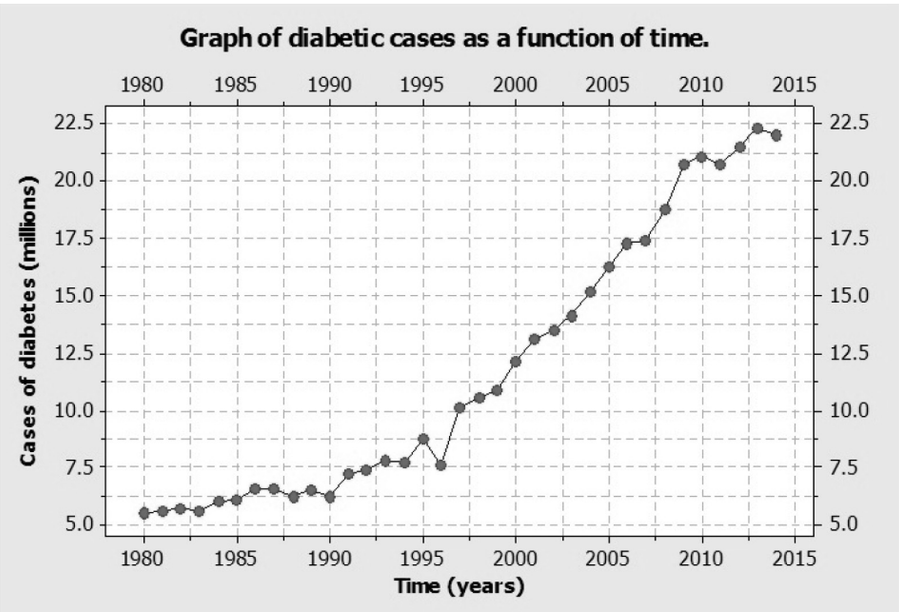
Years	Cases	Years	Cases	Years	Cases	Years	Cases	Years	Cases
1980	5.5	1987	6.6	1994	7.7	2001	13.1	2008	18.8
1981	5.6	1988	6.2	1995	8.7	2002	13.5	2009	20.7
1982	5.7	1989	6.5	1996	7.6	2003	14.1	2010	21.1
1983	5.6	1990	6.2	1997	10.1	2004	15.2	2011	20.7
1984	6	1991	7.2	1998	10.5	2005	16.3	2012	21.5
1985	6.1	1992	7.4	1999	10.9	2006	17.3	2013	22.3
1986	6.6	1993	7.8	2000	12.1	2007	17.4	2014	22.0

Data source: CDC, 2015 ***bis***. Annual Number (in Thousands) of New Cases of Diagnosed Diabetes Among Adults Aged 18-79 Years, United States, 1980-2014. National Center for Health Statistics, Division of Health Interview Statistics, data from the National Health Interview Survey. National Center for Chronic Disease Prevention and Health Promotion, data computed by personnel of the Division of Diabetes Translation.

RESULTS AND DISCUSSION

The first step used in the methodology consisted in graphing the original diabetic incidence rates to visualize the type of function followed by the data. Figure 1 below shows this situation.

Figure 1. *Graph showing the diabetic incidence rates among adults aged 18-79 years, United States, 1980-2014.*



Data source: own elaboration

As seen in figure 1 above there is an outlying case that occurred in the year of 1996. Similar situations occurred in the years of 2011 and 2014. These events probably occurred because the participating subjects were not blocked by similar characteristics.

The second step consisted in fitting a simple linear regression model by evaluating its utility through objective and subjective diagnostics. Also, our methodology prepared time series graphical analysis. Table 2 below shows the printed results. Likewise, figure 2 below depicts the time series graphical analysis.

Table 2. *Diabetic incidence rates among adults aged 18-79 years, United States, 1980-2014, after fitting a simple linear regression model.*

Predictor	Coef	SE	T	P	VIF
Constant	-1087.29	56.72	-19.17	0.000	
Time (years)	0.55034	0.02840	19.38	0.000	1.000
The regression equation is: Cases of diabetes (millions) = -1087 + 0.550 time (years) S = 1.69697 R-sq = 91.9% R-sq(adj) = 91.7% PRESS = 108.181 R-sq(pred) = 90.80%					
Analysis of variance					
Source	DF	SS	MS	F	P
Regression	1	1081.2	1081.2	375.47	0.000
Residual error	33	95.0	2.9		
Total	34	1176.3			

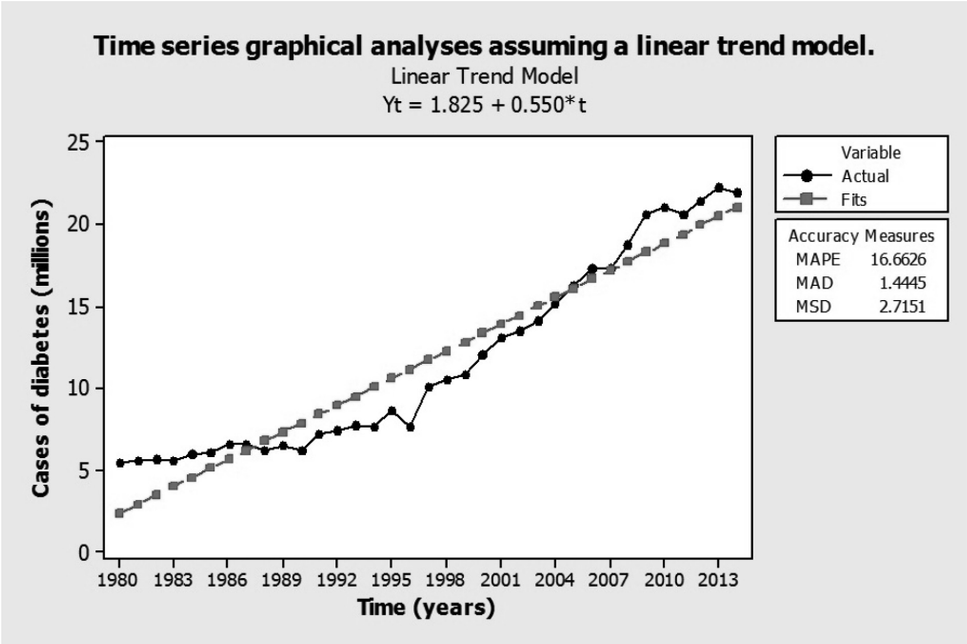
Data source: own elaboration

Note: The S stands for standard error of estimate and it represents the average distance that the observed values fall away from the regression line. Conveniently, it tells you how biased the regression model is on average using the units of the response variable. Smaller values are better because it indicates that the observations are closer to the fitted line.

R-sq stands for the coefficient of determination (R^2). It is a value between 0 (0 percent) and 1 (100 percent). The higher the value, the better the degree in which the X variable explains the variance of the Y variable.

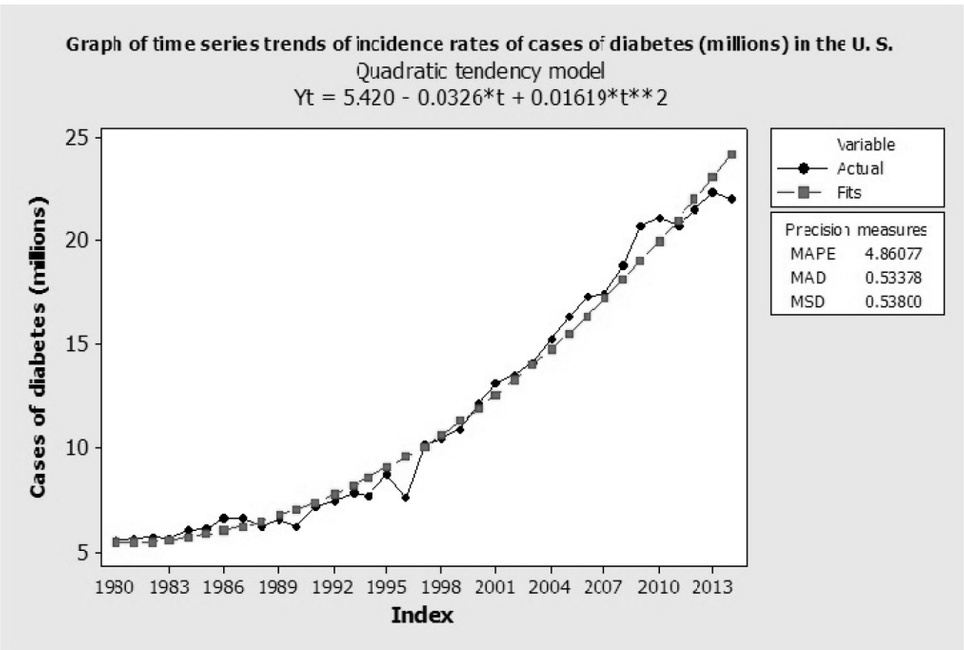
The PRESS value stands for predicted sum of squares. This statistic is used to assess the model's predictive capability. Usually, the smaller the PRESS value, the better the model's predictive ability.

Figure 2. *Graph showing the time series trend analysis assuming a linear trend model with its corresponding accuracy measures of MAPE, MAD, and MSD.*



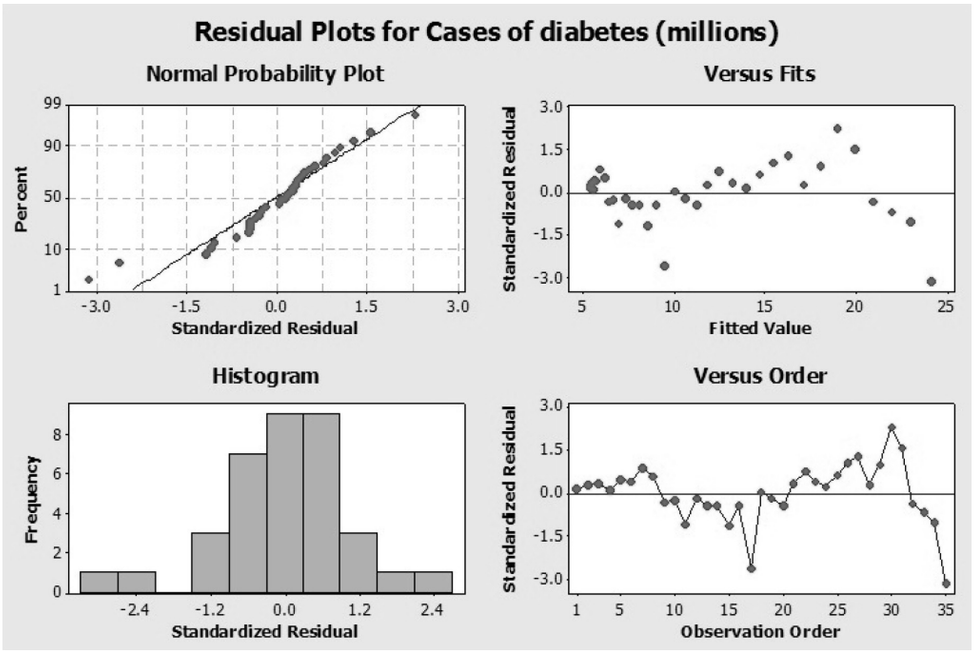
Data source: own elaboration

Figure 3. *Graph showing the time series graphical analysis for the adjusted quadratic polynomial regression model along with its measures of precision.*



Data source: own elaboration

Figure 3a. *Graphs showing the subjective residual plots for the cases of diabetes assuming a quadratic polynomial regression model.*



Data source: own elaboration

The time series graphical analysis is shown in figure 3 below. In figure 3, the terms MAPE, MAD, and MSD are measures of fitting accuracy and are used to evaluate the fitted accuracy of the model. For example, the acronym MAPE (Mean absolute percentage error) expresses the fitted accuracy as a percentage. Here, the lower its value, the more accurate the model will be. Likewise, MAD (Mean absolute deviation) helps to conceptualize the average amount of error in absolute value. Too, MSD (Mean squared deviation) is a measure of the precision of the adjusted values. In general, as the values of these statistics decrease, the more precise the model will be.

The third step consisted in fitting a quadratic polynomial regression model along with its objective and subjective evaluations. Figures 3 and 3a, and table 3 below show this tactic.

Figure 4a above shows the normal probability graph with most of the values very close to the least square line, except two values which corresponded to the years of 1996 and 2014. Moreover, the graph of the fits shows a bit of temporal autocorrelation. Also, analyzing the histogram looks reasonably symmetric.

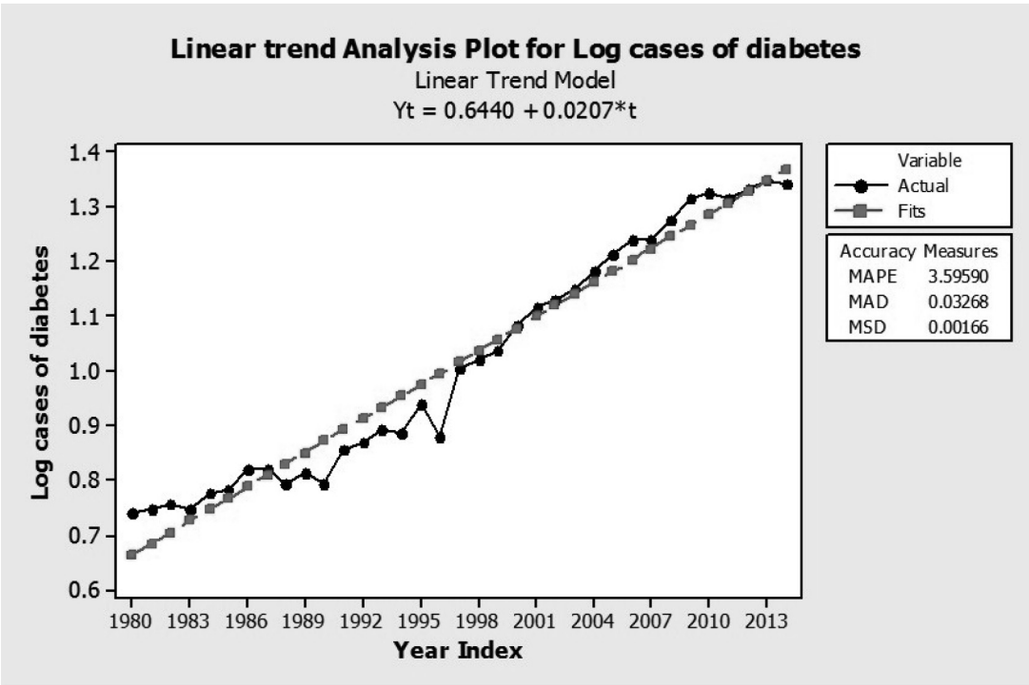
Table 3. *Table showing the diabetic incidence rates among adults aged 18-79 years, United States, 1980-2014, after fitting the adjusted quadratic polynomial regression model.*

Predictor	Coef	Of EE	T	P	VIF
Constant	5.4203	0.4123	13.15	0.000	
Time (index)	-0.03262	0.05281	-0.62	0.541	16.921
xsqr time index	0.016193	0.001423	11.38	0.000	16.921
Regression equation: Cases of diabetes (millions) = 5.42 - 0.0326 (time index) + 0.0162 (time index)^2					
S = 0.767094 R^2 = 98.4% R^2(adjusted) = 98.3% PRESS = 24.2893 R^2(pred) = 97.94%					
Analysis of variance					
Source	GL	SC	MC	F	P
Regression	2	1157.45	578.72	983.50	0.000
Error	32	18.83	0.59		
Total	34	1176.28			
Durbin-Watson statistics = 0.819057					

Data source: own elaboration

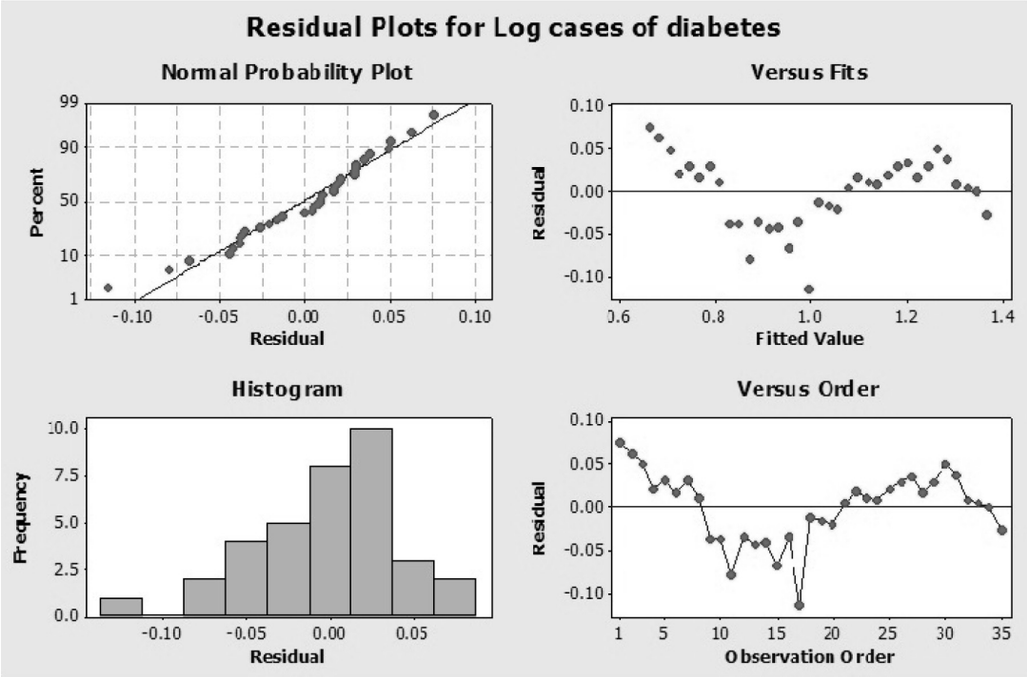
The fourth step consisted in fitting a time series logarithmic transformed model along with its complementary residual graphs. Also, this approach included the objective and subjective diagnostics. This methodology is depicted in figures 4 and 4a, and table 4 below.

Figure 4. *Time series trend analysis using log transformed values.*



Data source: own elaboration

Figure 4a. *Complementary residual plots using logarithmic transformations of cases of diabetes.*



Data source: own elaboration

Table 4. *Table showing the diabetic incidence rates among adults aged 18-79 years, United States, 1980-2014, after fitting a logarithmic regression model.*

Predictor	Coef	SE	T	P	VIF
Constant	-40.336	1.402	-28.78	0.000	
Time (years)	0.0207077	0.0007019	29.50	0.000	1.000
The regression equation is: Log (cases of diabetes) = -40.3 + 0.0207 (time, years) S = 0.0419398 R-sq = 96.3% R-sq(adj) = 96.2% PRESS = 0.0652005 R-sq(pred) = 95.90%					
Analysis of variance					
Source	DF	SS	MS	F	P
Regression	1	1.5308	1.5308	870.32	0.000
Residual error	33	0.0580	0.0018		
Total	34	1.5889			
Durbin-Watson statistic = 0.525540					

Data source: own elaboration

The graphing of the original diabetic incidence rates for checking the type of function tracked by the data shows that the values of the diabetic incidence rates resulted in a steadily increasing trend, as shown in figure 1. However, by closely examining this figure there seems to be an outlying value of 6.2, which corresponded to the year of 1996. Also, to a lesser extent, there is an additional outlier of 7.6, which corresponded to the year of 1990. Likewise, there is another one corresponding to the year of 2010. Again, this outlying value was probably due to experimental errors because the participating subjects were not grouped by similar characteristics as age, sex, weight, clinical back-

grounds, etc., from the experimental design point of view. These experimental errors were probably owed to the fact that there was no blocking, that is, there was no grouping of the subjects by similar characteristics in the sampling procedure done by the authors of the Disease Control and Prevention who prepared this sampling scheme.

About the fitting of a simple linear regression model (figure 3) this procedure was not satisfactory because the resulting diagnostics and measure of accuracy of MAPE, MAD, and MSD were not acceptable. For example, the MAPE expressing the percentage of error was equal to 16.6626. This means that the prediction capability of a simple linear regression model could be in error by 16.66%. Similarly, the MAD and MSD values were a bit too high. Besides as judged by figure 2, the standard error of estimate, S and the PRESS value were too high. In addition, the Durbin-Watson statistics equal to 0.1743 indicates the data is too skewed. All these observations flagged experimental errors that degraded the capability of this type of fit. In a similar fashion, the fitting of a polynomial regression model was not very satisfactory because the MAPE, MAD, and MSD accuracy measures were not altogether satisfactory. The value of MAPE equal to 4.86 means the fitting capability of the model could be off by 4.86%. Besides the VIF (Variance of inflection factors) equal to 16 are waning collinearity problems that could give to experimental errors that degrades the fitting model capability. Finally, the resulting logarithmic transformed model was even more satisfactory because its objective and subjective diagnostics were more acceptable (figures 4 and 4a). This assertion is sustained by the resulting accuracy measures of MAPE, MAD, and MSD, which had the lowest values of all the models tested. For example, in figure 4, the value of the MAPE equal to 3.595 means the fitting percentage error is about 3.59%. Also, the values of MAD equal to 0.03 and of MSD equal to 0.0016 were the lowest one recorded of all the previous models tested. Though the R 2 = 98.4% of the adjusted quadratic model versus R 2 = 96.3% of the logarithmic transformed value was lower. Though the R^2 = 98.4% of the adjusted quadratic model *versus* R^2 = 96.3% of the logarithmic transformed value was lower. This disadvantage was offset by the much lower values of PRESS of the logarithmic transformed model. The regression equation of this model was: Log (cases of diabetes) = -40.3 + 0.0207 (time).

CONCLUSIONS

It is concluded that of the three models tested, the logarithmic transformed model, whose equation is Log (cases diabetes) = -40.3 + 0.0207, is the best candidate model. Though the errors variance were not constant in some instances, this could be since the participating subjects were not grouped by similar characteristics as age, sex, weight, clinical backgrounds, etc. Even though with these pitfalls, we can reasonably conclude that the results are correct with only a 3.59% error, as suggested by the MAPE value of 3.59. This is also supported by a small value of the standard error of estimate of S = 0.0419398, a determination coefficient of R^2 = 96.3%, a PRESS value of 0.0652, and very significant values of P.

ACKNOWLEDGEMENTS

The authors of this work are grateful beyond measure to the scientists of CDC.

BIBLIOGRAPHICAL REFERENCES

American Diabetes Association (ADA) (2015). Statistics about Diabetes. Retrieved April 1, 2016, from http://www.niddk.nih.gov/about-niddk/research-areas/diabetes/Pages/diabetes.aspx

Centers for Disease Control and Prevention (CDC) (2010). Number of Americans with Diabetes Projected to Double or Triple by 2050. Retrieved April 1, 2016, from www.cdc.gov/media/pressrel/2010/r101022.html

—— (2014). Report Estimates of Diabetes and its Burden in the United States. National Diabetes Statistics, 2014. Retrieved July 1, 2016, from http://www.cdc.gov/diabetes/pdfs/data/2014-report-estimates-of-diabetes-and-its-burden-in-the-united-states.pdf

—— (2015). National Diabetes Statistics, 2014. National Center for Chronic Disease Prevention and Health Promotion, Division of Diabetes Translation. Retrieved April 1, 2016, from http://www.cdc.gov/diabetes/pubs/statsreport14/national-diabetes-report-web.pdf

—— (2015 *bis*). Annual Number (in Thousands) of New Cases of Diagnosed Diabetes among Adults Aged 18-79 Years, United States, 1980-2014. National Center for Health Statistics, Division of Health Interview Statistics. Re-

trieved August 1, 2016, from http://www.cdc.gov/diabetes/statistics/incidence/fig1.htm

International Diabetes Federation (IDF) (2010). Millions Unite for Diabetes Awareness on World Diabetes Day 2010. Retrieved April 1, 2016, from www.idf.org/millions-unite-diabetes-awareness-world-diabetes-day-2010

—— (2014). New Diabetes Figures in China: IDF Press Statement. Retrieved April 1, 2016, from http://www.idf.org/press-releases/idf-press-statement-china-study

Lebech-Cichosz, S., Johansen, M. D., & Hejlesen, O. (2015). Toward Big Data Analytics. Review of Predictive Models in Management of Diabetes and its Complications. *Journal of Diabetes Science and Technology*. October 14, 2015. Retrieved April 1, 2016 from http://dst.sagepub.com/content/early/2015/10/14/1932296815611680.abstract

National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) (2015). Diabetes. Retrieved April 6, 2016, from http://www.niddk.nih.gov/about-niddk/research-areas/diabetes/Pages/diabetes.aspx

National Institute of Diabetes of the United Kingdom (NIDUK) (2014). Facts & Figures. Retrieved April 1, 2016, from https://www.diabetes.org.uk/Professionals/Position-statements-reports/Statistics/