Modern Applied Science; Vol. 14, No. 5; 2020
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

A Novel Method for Computationally Efficacious Linear and Polynomial Regression Analytics of Big Data in Medicine

Ahmed Al-Imam[1,2]

[1] Department of Anatomy and Cellular Biology, College of Medicine, University of Baghdad, Iraq

[2] CERVO Brain Research Centre, Faculty of Medicine, University of Laval, Canada

Correspondence: Ahmed Al-Imam, Department of Anatomy and Cellular Biology, College of Medicine, University of Baghdad, Iraq.

Received: March 19, 2020 Accepted: April 14, 2020 Online Published: April 16, 2020

doi:10.5539/mas.v14n5p1 URL: https://doi.org/10.5539/mas.v14n5p1

Abstract

Background:

Machine learning relies on a hybrid of analytics, including regression analyses. There have been no attempts to deploy a scale-down transformation of data to enhance linear regression models.

Objectives:

We aim to optimize linear regression models by implementing data transformation function to scale down all variables in an attempt to minimize the sum of squared error.

Materials and Methods:

We implemented non-Bayesian statistics using SPSS and MatLab. We used Excel to generate 40 trials of linear regression models, and each has 1,000 observations. We utilized SPSS to conduct regression analyses, Wilcoxon signed-rank test, and Cronbach's alpha statistics to evaluate the performance of the optimization model.

Results:

The scale-down transformation succeeded by significantly reducing the sum of squared errors [absolute Z-score=5.511, effect size=0.779, p-value<0.001, Wilcoxon signed-rank test]. Inter-item reliability testing confirmed the robust internal consistency of the model [Cronbach's alpha=0.993].

Conclusions:

The optimization model is valuable for high-impact research based on regression. It can reduce the computational processing demands for powerful real-time and predictive analytics of big data.

Keywords: artificial Intelligence, data Transformation, linear models, machine learning, polynomial models, predictive analytics, regression analysis

1. Introduction

Although data science and statistical modeling have been evolving for centuries, most analytics are not entirely accurate (Box, 1976). The British statistician, George EP Box, coined the epigram "*All models are wrong, but some are useful*" (Box, 1976; Field, 2015). The famous aphorism of statistics appeared in a paper published by George Box at the Journal of the American Statistical Association in 1976 (Berro, 2018; Box, 1976). Regression models describe a continuous response variable as a function of predictor variables that can predict the behavior of complex systems (Dawes, 1974; Searle, 1971). Regression analytics utilizes the least squares to model the causality relationships within data between explanatory and outcome variables (Grizzle, Starmer, & Koch, 1969). However, these methods are not sufficiently "bulletproof" in terms of statistical precision (Cohen, 1992; Hlaváč, Krajcarz, Hlaváčová, & Spadło, 2017). Sir Ronald Fisher, a British data scientist, introduced the modern regression model in 1922 (Edward, 2011; Efron, 1998; Halds, 1998). Ronald Fisher followed in the footsteps of contemporary statisticians, including Karl Pearson, a 19th-century English mathematician (Norton, 1978). Pearson innovated many statistical tests, including Pearson's correlation, which is connected to Fisher's regression models (Sedgwick, 2012). Simple regression examines the relationship between one predictor variable and one outcome variable in causality testing, while multiple regression, including linear and polynomial variants, examines a

multitude of explanatory variables for a higher predictive power (Al-Imam, 2020; Freedman, 1981; Godfrey, 1992; Zou, Tuncali, & Silverman, 2003). Machine learning, an application of artificial intelligence, relies on several methods, including regression models, non-biological neural networks, and classification trees (Al-Imam and Al-Lami, 2020; Al-Imam and Motyka, 2019; Al-Imam, Sahai, Al-Derzi, & Al-Shalchy, 2019; Jordan and Mitchell, 2015). The optimization of those analytics can positively impact machine learning technologies (Everitt, Goertzel, & Potapov, 2017; Peuvo, 2018; Schneider, 2016). As mentioned earlier, regression exploits the least-squares method to extrapolate a line that best fits the causality association (Chevreuil, Lebrun, Nouy, & Rai, 2015; Van Dao, Chaitusaney, & Nguyen, 2016). Perfecting the least squares is critical for a rigorous statistical inference and predictive modelling.

Most importantly, these optimized models will not only be more accurate in terms of statistical inference but will also be economical in terms of the computational processing demands for the analyses of large-scale arrays of data (Schilling et al., 2005). Scientists use several data transformation techniques to boost a spectrum of statistical tests, including the Fourier transformation, the Log Base-10 (Log10) transformation, the natural logarithm (Ln) transformation, and inverse transformation, as well as the square root and cubic root transformers (O'Hara and Kotze, 2010; Takeda et al., 1982). The scale-down optimization of data can capitalize on powerful and economic computational processing for real-time analyses and predictive models (Schilling et al., 2005). In 1965, the British statistician, Austin Bradford Hill, proposed the nine-element criteria to provide evidence for causality between a presumed effect and an observed outcome (Phillips and Goodman, 2004). Hill proposed to analyze the effect size, the strength of association, the replicability of the results, the specificity of association, the temporality of causation, the biological gradient effect, as well as plausibility, coherence, experimentation, and analogy (Fedak et al., 2015; Phillips and Goodman, 2004). If researchers and data analysts integrate optimized linear or polynomial models, in combination with Hill's criteria, they can infer robust data that possess the least prediction error and the highest statistical power, while keeping the human resources and the requisite computational infrastructure to a minimum.

Our primary objective is to optimize linear and polynomial models, principally for analytics that are dependent on correlation and regression statistics, by implementing a scale-down transform function that significantly reduces the error of residuals by minimizing the sum of squared errors (SSE). Thereby achieving more powerful and externally valid models that apply to real-time analytics, as well as predictive models that are necessary for high-impact research, based on big data (Al-Imam, 2017; Al-Imam, 2019; Al-Imam et al., 2019).

2. Materials and Methods

2.1 Mathematical Simulations

We made multiple simulations based on a random number generator that follows a normal distribution [mean=0, standard deviation=1]. We created 40 trials (i.e., simulation models) for linear regression calculations [k=40], each test has a sample size of one thousand observations [n=1,000] for two variables as a predictor and an outcome (X and Y), thereby, summing to a grand sample size of 40,000 [n_{total}=40,000]. We transformed the two variables, by dividing, each observation to the maximum observation within the same variable, by using the "max" function in Excel 2016, thereby scaling them down. Within each linear model, we calculated correlation and regression statistics, including the sum of squares (SS), mean of squares (MS), F statistic [ANOVA], and *p-value* [regression]. We calculated the sum of squared errors (SSE) using the formula $SSE=\sum (y-\hat{y})^2$ to fulfill the regression equation $\hat{y}=b0+b1X$. Calculations were conducted twice, before [pre-optimization] and after deploying the scale-down transformation [post-optimization]. We statistically tested the performance of the scale-down optimization model using the Wilcoxon signed-rank test for non-parametric within-subjects statistical inference by comparing the pre-optimization versus post-optimization statistics. Ultimately, we further examined the optimization efficacy of our model by implementing Cronbach's alpha as a measure of the internal consistency of the summative optimized model.

2.2 Statistical Analysis, Ethics, and Level of Evidence

We implemented the Statistical Package for the Social Sciences [IBM-SPSS version 24] and Excel [Microsoft Office 2016] with integrated Data Analysis ToolPak. We made descriptive statistics using Excel and GNU-Octave version 5.1.0 [GNU's Not UNIX Project]. We implemented MatLab high-level programming language (HLL) version R2019a [MathWorks] for two-dimensional array transposition before exporting the data to SPSS for Cronbach's alpha calculations. We conducted an elaborate set of parametric and non-parametric models of non-Bayesian statistics, including linear and polynomial regression, Fisher's ANOVA, Wilcoxon signed-rank test for within-subjects study design, and Cronbach's alpha analytics for assessing the reliability and internal consistency of our proposed statistical model based on the scaling down of the data.

The authors conducted the work described in this article following the Code of Ethics of the World Medical Association (Declaration of Helsinki) on medical research involving human subjects, EU Directive (210/63/EU) on the protection of animals used for scientific purposes, uniform requirements for manuscripts submitted to biomedical journals, and the ethical principles defined in the Farmington Consensus of 1997. According to the Oxford Centre for Evidence-based Medicine (OCEBM), our research represents "Absolute Better-Value or Worse-Value Analyses" under the category "Economic and Decision Analyses" (Greenhalgh, Howick, & Maskrey, 2014; OCEBM Levels of Evidence, 2016). Accordingly, our study is of level-1c, which belongs to the top tier [level-1, Grade-A] of the categorization scheme rectified by the OCEBM (OCEBM Levels of Evidence, 2016).

2.3 Systematic Review of the Literature

During September 2019, we conducted a pragmatic review of the databases of peer-reviewed literature, including the Cochrane Library [the Cochrane Database of Systematic Reviews | the Cochrane Collaboration], PubMed [the United States National Library of Medicine], and Embase [Elsevier]. We implemented an exhaustive set of keywords based on medical subject headings (MeSH), in addition to generic terms, while using Boolean expression operators and truncations. We deputized keywords of five main themes, including 1) machine learning and artificial intelligence, 2) real-time and predictive analytics, 3) real-time analytics and epidemiology, 4) data transform functions, and 5) an amalgamation of the previous four themes. The aim is to explore the existing literature for prior attempts of using scale-down data transformation for enhancing and optimizing linear models.

3. Results

For the optimization model, we applied the scale-down transform for 40 trials of linear regression analyses (Table 1).

Table 1. Optimization Model Analytics

Trial	Pre-Optimization				Post-Optimization				*P-value*
	R^2	SSE	F Score	*p-value*	R^2	SSE	F Score	*p-value*	[Pre-optimization vs. Post-optimization]
1	0.003	1.01E+09	2.960	0.086	0.003	81.656	2.960	0.086	
2	0.002	9.36E+08	1.543	0.214	0.002	86.240	1.543	0.214	
3	0.001	9.16E+08	0.930	0.335	0.002	81.523	0.930	0.335	
4	0.002	9.90E+08	2.219	0.137	0.002	115.759	2.219	0.137	
5	0.002	9.83E+08	1.898	0.169	0.002	86.326	1.898	0.169	
6	0.003	9.82E+08	2.752	0.097	0.003	111.207	2.752	0.097	
7	<0.001	1.03E+09	0.242	0.623	<0.001	112.320	0.242	0.623	
8	<0.001	1.01E+09	0.042	0.838	<0.001	122.491	0.042	0.838	
9	<0.001	9.75E+08	0.321	0.571	<0.001	72.631	0.321	0.571	
10	0.002	9.61E+08	1.968	0.161	0.002	91.278	1.968	0.161	Absolute Z-score=5.511 Effect size=0.779 *p-value*<0.001 †
11	0.001	9.71E+08	0.152	0.697	0.001	73.725	0.152	0.697	
12	0.002	9.51E+08	0.105	0.746	0.002	89.030	0.105	0.746	
13	0.002	9.60E+08	0.528	0.468	0.002	81.494	0.528	0.468	
14	0.001	9.76E+08	0.446	0.504	0.001	122.198	0.446	0.504	
15	0.004	1.03E+09	0.887	0.347	0.004	74.295	0.887	0.347	
16	0.001	9.46E+08	0.520	0.471	0.001	90.658	0.520	0.471	
17	0.003	1.02E+09	0.879	0.349	0.003	93.520	0.879	0.349	
18	0.003	9.60E+08	0.712	0.399	0.003	82.434	0.712	0.399	
19	0.001	9.86E+08	0.447	0.504	0.001	76.680	0.447	0.504	
20	0.003	1.02E+09	0.816	0.367	0.003	92.841	0.816	0.367	
21	0.002	1.03E+09	0.525	0.469	0.002	74.150	0.525	0.469	
22	0.001	9.38E+08	0.160	0.689	0.001	102.680	0.160	0.689	

23	0.001	9.79E+08	0.626	0.429	0.001	113.438	0.626	0.429
24	0.002	1.02E+09	0.425	0.515	0.002	78.490	0.425	0.515
25	0.002	9.84E+08	0.676	0.411	0.002	109.852	0.676	0.411
26	0.004	9.62E+08	0.459	0.498	0.004	80.366	0.459	0.498
27	0.002	9.98E+08	0.834	0.361	0.002	108.458	0.834	0.361
28	0.004	9.93E+08	0.673	0.412	0.004	78.787	0.673	0.412
29	0.003	9.73E+08	0.936	0.334	0.003	80.806	0.936	0.334
30	0.002	9.18E+08	0.267	0.605	0.002	92.948	0.267	0.605
31	0.003	1.01E+09	0.607	0.436	0.003	88.754	0.607	0.436
32	0.004	9.32E+08	0.016	0.899	0.004	95.714	0.016	0.899
33	0.001	1.02E+09	0.027	0.870	0.001	112.493	0.027	0.870
34	**0.002**	**9.96E+08**	**0.048**	**0.827**	**0.002**	**76.484**	**0.048**	**0.827**
35	0.002	1.01E+09	0.418	0.518	0.002	87.004	0.418	0.518
36	0.001	9.22E+08	0.444	0.505	0.001	81.209	0.444	0.505
37	0.004	9.81E+08	0.874	0.350	0.004	109.339	0.874	0.350
38	0.001	9.41E+08	0.926	0.336	0.001	113.928	0.926	0.336
39	0.003	1.02E+09	0.563	0.453	0.003	85.213	0.563	0.453
40	0.003	9.71E+08	0.152	0.697	0.003	92.430	0.152	0.697

† Wilcoxon signed-rank Statistics: Pre-optimization vs. Post-optimization [Sum of Squared Errors (SSE)].

†† Linear Model-of-Interest in Bold Font [Random Selection, Trial 34].

The model was triumphant in attaining a significant reduction of the sum of squared errors (SSE) for each trial following the application of the scale-down transform [absolute Z-score = 5.511, effect size = 0.779 (i.e., strong effect), *p-value* < 0.001 for the Wilcoxon signed-rank test] (Table 2). We utilized a non-parametric alternative of the dependent Student's t-test due to the violation of t-test assumptions, including the absence of statistical outliers, homoscedasticity, and the normality of distribution [Shapiro-Wilk test] (Table 2).

Table 2. Optimization Model Statistics: Normality testing and Wilcoxon signed-rank test

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Pre-optimization SSE	40	100.0%	0	0.0%	40	100.0%
Post-optimization SSE	40	100.0%	0	0.0%	40	100.0%

Descriptive Statistics

Pre-optimization SSE	Mean		980447924.50	5280931.108
	95% Confidence Interval for Mean	Lower Bound	969766233.10	
		Upper Bound	991129615.90	
	5% Trimmed Mean		980985772.20	
	Median		981555517.50	
	Std. Deviation		33399540.930	
	Range		118840731	
	Interquartile Range		50899902	
	Skewness		-.198	.374
	Kurtosis		-.818	.733
Post-optimization SSE	Mean		92.52123	2.359813
	95% Confidence Interval for Mean	Lower Bound	87.74805	
		Upper Bound	97.29440	
	5% Trimmed Mean		91.93900	
	Median		88.89200	
	Std. Deviation		14.924771	
	Range		49.860	
	Interquartile Range		28.212	
	Skewness		.574	.374
	Kurtosis		-.965	.733

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Pre-optimization SSE	.081	40	.200*	.968	40	.309
Post-optimization SSE	.148	40	.027	.907	40	.003

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Wilcoxon Signed Ranks Test

		N	Mean Rank	Sum of Ranks
Post-optimization SSE - Pre-optimization SSE	Negative Ranks	40[a]	20.50	820.00
	Positive Ranks	0[b]	.00	.00
	Ties	0[c]		
	Total	40		

a. Post-optimization SSE < Pre-optimization SSE

b. Post-optimization SSE > Pre-optimization SSE

c. Post-optimization SSE = Pre-optimization SSE

Test Statistics[a]

	Post-optimization SSE - Pre-optimization SSE
Z	-5.511[b]
Asymp. Sig. (2-tailed)	<0.001

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

On the other hand, there was no significant change in the coefficient of determination (R^2 score) and the F-score for the pre-optimized versus post-optimized trials, as we created each with a random number generator function using the Data Analysis ToolPak plugin in Excel. A randomly selected linear model, the 34th trial, manifested with a sum of squared errors of 9.96E+08 [pre-optimization] and 76.484 [post-optimization], confirming a significant SSE reduction and a better predictive model fitting. The scale-down transformation neither had a distortion nor an artefactual effect on the scattered correlates of the tested variables (Figure 1).

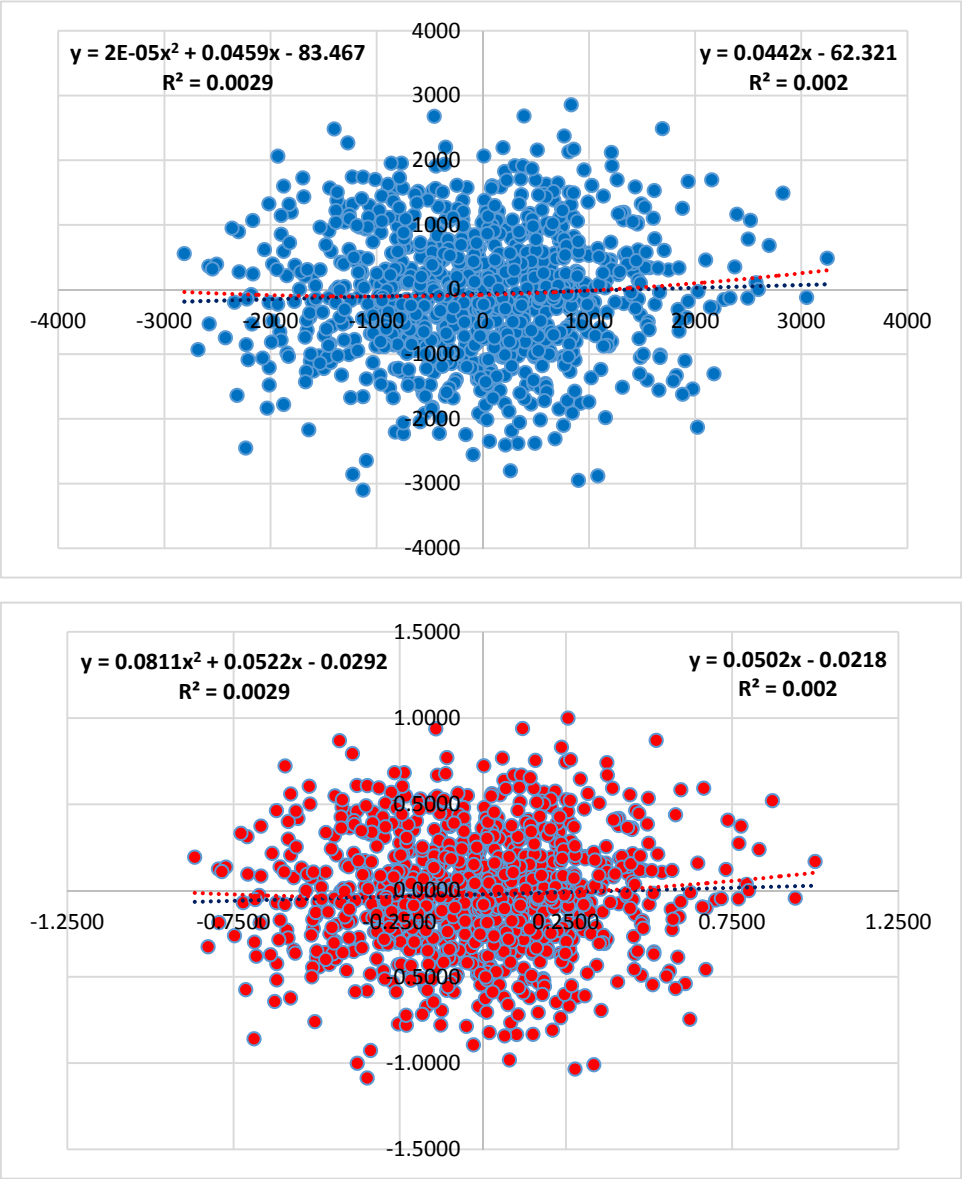


Figure 1. Randomly Selected Regression Model: Pre-optimization [upper, blue] and Post-optimization [lower, red]

† Linear Model-of-Interest: Random Selection [Trial 34, Table 1].

†† Coefficient of Determination & Regression Equation for Linear and Polynomial Models in Bold Fonts.

††† Identical, Superimposed Morphology, of the Scattered Distribution for Pre-optimization versus Post-optimization, i.e., No Deformation of Data.

Lastly, Cronbach's alpha analysis yielded collateral evidence and verified the internal consistency of the optimization model [Cronbach's alpha=0.993]. Deleting any trial from the optimization model had no effect on the inter-item reliability with an exception for five simulations [1st, 4th, 5th, 6th, and 10th], the deletion of which increases the internal consistency to 0.998, an almost perfect consistent model.

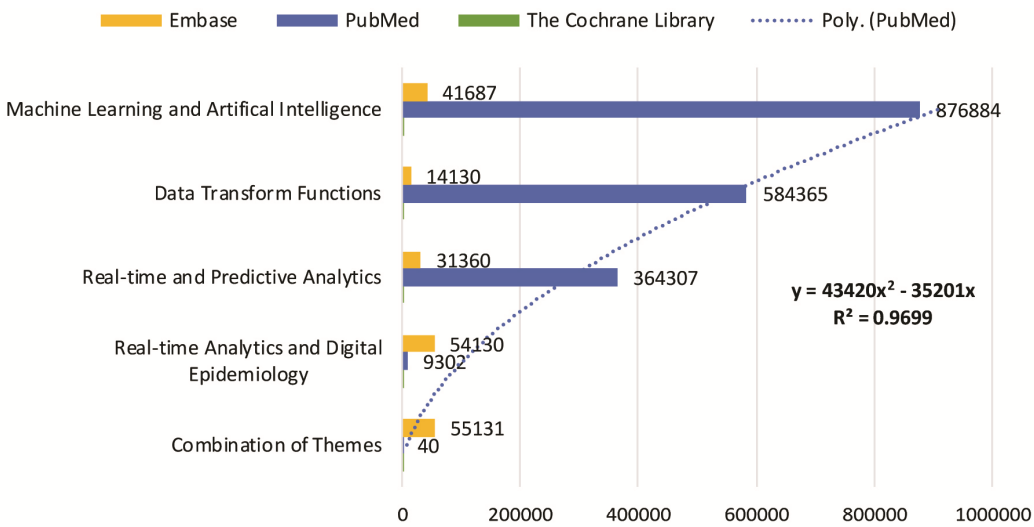
4. Discussion

Our optimization model applies to anticipated high-impact research that requires linear or polynomial model analyses (Figure 1), including anatomical sciences, dermatology, and medical research and practice.

Boosted regression models are of utmost importance in the exponentially growing field of machine learning and artificial intelligence. The applications are not limited to psychoactive and novel psychoactive substances research, an emerging subdiscipline of addiction neuroscience and behavioral psychiatry. Optimized regression analytics are priceless when it comes to applications with extensive data analytics and bioinformatics, comprehensive genomic analyses, and analytics based on extracting information from open-source deposits of big data, for instance, Google Trends and Google Analytics databases. Optimum linear and polynomial models not only will reinforce the hypothesis testing for more powerful inferences but also will lessen the computational processing power and the human resources allocated for demanding real-time and predictive analyses. If our optimization model integrates with the anticipated quantum computing, the benefits will be monumental concerning the precision of analytics and the efficacy of the computational processing.

Machine learning relies upon the analyses of big data using a plethora of well-established techniques of mathematical and data science models, including artificial neural networks, regression analysis, and decision trees (Jordan and Mitchell, 2015). Artificial intelligence techniques attempt to reach the lowest achievable error rates of mathematically interpreted predictions for causality associations (Everitt, Goertzel, & Potapov, 2017). Machine learning is mandatory for unwitnessed benefits when it comes to applications related to spatio-temporal description and prediction of phenomena of interest, including epidemiological and digital epidemiological investigations (Everitt, Goertzel, & Potapov, 2017; Jordan and Mitchell, 2015). The infrastructure of big data upon which machine learning algorithms operate is the same as those designated for classical epidemiology and digital epidemiological research (Rothman, Greenland, & Lash, 2008). Researchers can retrieve data from the databases using survey tools, internet snapshots, longitudinal studies, cross-sectional studies, analyses of web-based social networks, and electronic commerce website analytics of the surface web as well as the deep web, including the infamous Darknet hypermarket (Al-Imam A and Al-Shalchi, 2019; Al-Imam, 2017; Motyka and Al-Imam, 2019; Rothman, Greenland, & Lash, 2008).

We reviewed the literature using a combination of thematic keywords search. There were 55,288 publications indexed in the Cochrane Library (117, 0.21%), PubMed (40, 0.07%), and Embase (55,131, 99.71%) (Figure 2).



	Combination of Themes	Real-time Analytics and Digital Epidemiology	Real-time and Predictive Analytics	Data Transform Functions	Machine Learning and Artifical Intelligence
Embase	55131	54130	31360	14130	41687
PubMed	40	9302	364307	584365	876884
The Cochrane Library	117	979	2599	1003	3254

Figure 2. Keywords-Based Systematic Review of the Databases of Literature

† Date of Review of Literature: 20th of September 2019.

Following a full-text retrieval of papers of interest, only fifteen publications (0.03%) indexed in the national library of medicine were found relevant to the primary objective. However, none of these studies implemented our data transform method to boost linear or polynomial regression models. Since the last decade, there have been several attempts in the existing peer-reviewed literature to implement linear models as well as other machine-learning methods in combination with the data transform function, including logistic regression, regression trees and Fourier transform, logistic regression with Log10 transformation, logistic regression with Ln transformation, multiple linear regression with log10 transformation, cycling regression model with Fourier transform, proportional hazards Cox regression model, time-series analytics regression with Fourier transform, logistic regression with square root and log10 transformation, and proportional hazards model in combination with logistic regression (Lorenz et al., 2017; Menotti, Puddu, & Lanti, 2002; Shaban-Nejad, Michalowski, & Buckeridge, 2018).

5. Conclusion

Our novel transform and optimization method serves three primary purposes: 1) Reducing the sum of squared errors (SSE), which will provide a better line of best fit. 2) The scale-down transformation will significantly reduce the computational processing demands for mathematical calculations for big data with an extensive list of variables, as well as an extended number of observations for each variable that is tangible in multiple polynomial regression analyses. 3) Real-time processing of correlations and regression among exhaustive multidimensional arrays of data will even be more consuming in terms of the requirement of computational processing power that can burden supercomputers existing today and the near future. The optimization will transform all variables into a narrower range with limited decimal places and without deforming the original correlation of variables, which can be economical for subsequent mathematical and computational processing.

5.1 Availability of Data

Our data are available upon request from the corresponding author.

5.2 Conflict of Interest

The authors declare that they have no conflict of interest.

5.3 Source of Funding

The authors self-funded this study.

References

Al-Imam, A. (2020). Novel Psychoactive Substances Research: On the Necessity of Real-time Analytics and Predictive Modelling. *Research and Advances in Psychiatry, 7*(1), [in press].

Al-Imam, A. (2017). *Monitoring and Analysis of Novel Psychoactive Substances in Trends Databases, Surface Web and the Deep Web, with Special Interest and Geo-Mapping of the Middle East*. (Master s thesis, University of Hertfordshire, Hertfordshire, United Kingdom). Retrieved from https://uhra.herts.ac.uk/handle/2299/19462

Al-Imam, A. (2019). Inferential Analysis of Big Data in Real-Time: One Giant Leap for Spatiotemporal Digital Epidemiology in Dentistry. *Odontostomatology Research Anatomy Learning and Implantology, 12*(1), 1-14.

Al-Imam, A., & Al-Lami, F. (2020). Machine learning for potent dermatology research and practice. *Journal of Dermatology and Dermatologic Surgery, 24*(1), 1-4. https://doi.org/10.4103/jdds.jdds_54_19

Al-Imam, A., & Al-Shalchi, A. (2019). Ekbom's Delusional Parasitosis: A Systematic Review. *Egyptian Journal of Dermatology and Venerology, 39*(1), 5-13. https://doi.org/10.4103/ejdv.ejdv_53_15

Al-Imam, A., & Motyka, M. (2019). On the necessity for paradigm shift in psychoactive substances research: the implementation of machine learning and artificial intelligence. *Alcoholism and Drug Addiction/Alkoholizm i Narkomania, 32*(3), 237-242. https://doi.org/10.5114/ain.2019.91004

Al-Imam, A., Khalid, U., Al-Hadithi, N., & Kaouche, D. (2018). Real-Time Inferential Analytics Based on Online Databases of Trends: A Breakthrough within the Discipline of Digital Epidemiology in Dentistry and Dental Anatomy. *Modern Applied Science, 13*(2), 81-94. https://doi.org/10.5539/mas.v13n2p81

Al-Imam, A., Sahai, A., Al-Derzi, A. R., Al-Shalchy, A., & Abdullah, F. (2020). All models are wrong, but some are useful: On the non-bayesian statistical robustness of Hilton's law. *European Journal of Anatomy, 24*(1), 75-78.

Berro, J. (2018). Essentially, all models are wrong, but some are useful—a cross-disciplinary agenda for building useful models in cell biology and biophysics. *Biophysical Reviews, 10*(6), 1637-1647. https://doi.org/10.1007/s12551-018-0478-4

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association, 71*(356), 791-799. https://doi.org/10.1080/01621459.1976.10480949

Chevreuil, M., Lebrun, R., Nouy, A., & Rai, P. (2015). A least-squares method for sparse low rank approximation of multivariate functions. *SIAM/ASA Journal on Uncertainty Quantification, 3*(1), 897-921. https://doi.org/10.1137/13091899X

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1*(3), 98-101. https://doi.org/10.1111/1467-8721.ep10768783

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*(2), 95. https://doi.org/10.1037/h0037613

Edwards, A. W. F. (2011). Mathematizing darwin. *Behavioral Ecology and Sociobiology, 65*(3), 421-430. https://doi.org/10.1007/s00265-010-1122-x

Efron, B. (1998). RA Fisher in the 21st century. *Statistical Science, 13*(2), 95-114. https://doi.org/10.1214/ss/1028905930

Everitt, T., Goertzel, B., & Potapov, A. (2017). Artificial general intelligence. *Lecture Notes in Artificial Intelligence*. Heidelberg: Springer. https://doi.org/10.1007/978-3-319-63703-7

Fedak, K. M., Bernal, A., Capshaw, Z. A., & Gross, S. (2015). Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology. *Emerging Themes in Epidemiology, 12*(1), 14. https://doi.org/10.1186/s12982-015-0037-4

Field, E. H. (2015). All models are wrong, but some are useful. *Seismological Research Letters, 76*(2A), 291-293. https://doi.org/10.1785/02201401213

Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics, 9*(6), 1218-1228. https://doi.org/10.1214/aos/1176345638

Godfrey, K. (1992). Simple linear regression in medical research. *Medical Uses of Statistics*. NEJM Books, Boston.

Greenhalgh, T., Howick, J., & Maskrey, N. (2014). Evidence based medicine: a movement in crisis? *BMJ, 348*, g3725. https://doi.org/10.1136/bmj.g3725

Grizzle, J. E., Starmer, C. F., & Koch, G. G. (1969). Analysis of categorical data by linear models. *Biometrics,* 489-504. https://doi.org/10.2307/2528901

Hald, A. (1998). *A History of Mathematical Statistics from 1750 to 1930*. New York: Wiley.

Hlaváč, L. M., Krajcarz, D., Hlaváčová, I. M., & Spadło, S. (2017). Precision comparison of analytical and statistical-regression models for AWJ cutting. *Precision Engineering, 50*, 148-159. https://doi.org/10.1016/j.precisioneng.2017.05.002

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255-260. https://doi.org/10.1126/science.aaa8415

Lorenz, M. W., Abdi, N. A., Scheckenbach, F., Pflug, A., Bülbül, A., Catapano, A. L., & Orth, A. (2017). Automatic identification of variables in epidemiological datasets using logic regression. *BMC Medical Informatics and Decision Making, 17*(1), 1-11. https://doi.org/10.1186/s12911-017-0429-1

Menotti, A., Puddu, P. E., & Lanti, M. (2002). The estimate of cardiovascular risk. Theory, tools and problems. *Annali Italiani Di Medicina Interna: Organo Ufficiale Della Societã Italiana Di Medicina Interna, 17*(2), 81-94.

Motyka, M. A., & Al-Imam, A. (2019). Musical preference and drug use among youth: an empirical study. *Research and Advances in Psychiatry, 6*(2), 50-57.

Norton, B. J. (1978). Karl Pearson and statistics: The social origins of scientific innovation. *Social Studies of Science, 8*(1), 3-34. https://doi.org/10.1177/030631277800800101

O'hara, R. B., & Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution, 1*(2), 118-122. https://doi.org/10.1111/j.2041-210X.2010.00021.x

OCEBM Levels of Evidence. Retrieved from https://www.cebm.net/2016/05/ocebm-levels-of-evidence/ (accessed 2020).

Phillips, C. V., & Goodman, K. J. (2004). The missed lessons of sir Austin Bradford Hill. *Epidemiologic Perspectives & Innovations, 1*(1), 1-5. https://doi.org/10.1186/1742-5573-1-1. https://doi.org/10.1186/1742-5573-1-3

Pueyo, S. (2018). Growth, degrowth, and the challenge of artificial superintelligence. *Journal of Cleaner Production, 197*, 1731-1736. https://doi.org/10.1016/j.jclepro.2016.12.138

Rothman, K. J., Greenland, S., & Lash, T. L. (Eds.). (2008). *Modern Epidemiology*. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins.

Schilling, M., Maiwald, T., Bohl, S., Kollmann, M., Kreutz, C., Timmer, J., & Klingmüller, U. (2005). Computational processing and error reduction strategies for standardized quantitative data in biological networks. *The FEBS Journal, 272*(24), 6400-6411. https://doi.org/10.1111/j.1742-4658.2005.05037.x

Schneider, S. (Ed.). (2016). *Science Fiction and Philosophy: From Time Travel to Superintelligence*. John Wiley & Sons. https://doi.org/10.1002/9781118922590

Searle, S.R., & Gruber, M.H. (1971). *Linear Models*. New York: Wiley.

Sedgwick, P. (2012). Pearson's correlation coefficient. *BMJ, 345*, e4483. https://doi.org/10.1136/bmj.e4483

Shaban-Nejad, A., Michalowski, M., & Buckeridge, D. L. (2018). Health intelligence: how artificial intelligence transforms population and personalized health. *npj Digital Medicine, 2018*(1), 53. https://doi.org/10.1038/s41746-018-0058-9

Takeda, M., Ina, H., & Kobayashi, S. (1982). Fourier-transform method of fringe-pattern analysis for computer-based topography and interferometry. *Journal of the Optical Society of America, 72*(1), 156-160. https://doi.org/10.1364/JOSA.72.000156

Van Dao, T., Chaitusaney, S., & Nguyen, H. T. N. (2016). Linear least-squares method for conservation voltage reduction in distribution systems with photovoltaic inverters. *IEEE Transactions on Smart Grid, 8*(3), 1252-1263. https://doi.org/10.1109/TSG.2016.2536782

Zou, K. H., Tuncali, K., & Silverman, S. G. (2003). Correlation and simple linear regression. *Radiology, 227*(3), 617-628. https://doi.org/10.1148/radiol.2273011499