Covenant University Journal of Politics & International Affairs, Vol.9 No. 1, Special September, 2021
ISSN: p. 2354-3558 e. 2354-3493 DOI: xxxxxxxxxxxxxxx

Open Access Journal Available Online

Big Data Analysis using Multi Linear Regression and Simulation techniques in predicting the growth of a business

Gregory Emeka Anichebe
Department of Computer Science,
University of Nigeria, Nsukka,
Enugu State, Nigeria.
gregory.anichebe@unn.edu.ng

Nnaemeka Emeka Ogbene
Department of Computer Science,
University of Nigeria, Nsukka,
Enugu State, Nigeria.
nnaemeka.ogbene@unn.edu.ng

Emmanuel C. Ukekwe
Department of Computer Science,
University of Nigeria, Nsukka.
Enugu State, Nigeria.
emmanuel.ukekwe@unn.edu.ng

Received: 23.09.2020 Accepted: 25.07.21
Date of Publication: September, 2021

Abstract: This paper developed a framework for enabling both small and large business enterprises remain profitable in business by making use of multi linear regression technique in predicting her business growth from the Big Data collected from respondents about the various external factors affecting a business entity such as, (i) the purchasing power of the people, (ii) existence of similar competitive products/services, (iii) availability of retail outlets, (iv) means of advertisement, and (v) means of transportation. A simple linear regression model was formulated in respect of each of these external factors (which are the independent or predictor variables) and its causality effect to the growth of a business (which is the response variable). An algorithm written in Java was used to simulate 10,000 input values for both the predictor and response variables in order to determine the parameters for each of the regression models. The derived parameters from the respective models were used to formulate the multi linear regression model for predicting the overall growth of a business. Results showed that the model gave very good predictions of a business growth index whenever the values of the external factors were varied to represent a typical real-life scenario. Business managers for small, medium, and large enterprises will therefore find the developed framework highly invaluable for strategic planning in order to increase their profit margins.

Keywords: Small and large business enterprises, Big Data, simple and multiple linear regression analysis, simulation, prediction

Introduction

Anyone can venture into business, but not all can come out of it successfully. So many factors must therefore be put into consideration by a business proprietor in order to prepare adequately for such venture. This goes along with the English adage, “Look before you leap”. Kituu (2020) is strongly in support of this idea by saying “Every business owner wants to constantly develop and succeed; there’s no doubt about that; but to accomplish it, it is essential to comprehend the elements or factors influencing such growth”.

Internal and external factors can affect the growth of a business. The external factors (such as, the purchasing power of the people, existence of other competitive products/services, availability of retail outlets for marketing the product/service, availability of advertisement means for creating awareness of the product/service to the people, and availability of transport means for making the product/service accessible to the people) are customer-oriented factors which affect a business entity more adversely than internal factors that centre mostly on organizational structure and managerial capability. This is because, no matter how good the internal structures of a business may be, the business will hardly be successful without considering the external factors for its marketability. These external factors are therefore used for formulating the model developed in this work for predicting the growth of a business.

The use of statistical survey such as questionnaire, interview, observations, documentary sources, etc can be adopted by an entrepreneur in order to gather large amounts of data about these external factors. The more data that are gathered, the more information that can be retrieved from it. Such huge assemblage of data which can be of different varieties (such as text, sound, and video), and gathered from different sources (such as websites, emails, smart phones, sensors, camera, etc) and received at a very fast rate (or velocity) for immediate processing is what is referred to as Big Data, (Segal, 2019).

Big Data can be analyzed using machine learning techniques such as data mining, decision tree, fuzzy logic, regression analysis, etc for making very important predictions or forecasting. Many authors have written volumes of materials on how to deploy Big Data in many areas of human endeavour such as healthcare, science, engineering, business and finance, to mention a few. The business sector, just like any other sector, still poses a lot of challenges to business entrepreneurs on how to always remain profitable in business in the midst of so many challenging factors such as the external factors considered in this work. This work therefore focused on the use of multi regression technique in analyzing the Big Data generated through computer simulation technique with respect to the various external factors militating against the success of a business so that meaningful predictions can be made with the formulated regression model for decision making. “Regression models provide the scientist with a powerful tool, allowing predictions about past, present, or future events to be made with information obtained about past or present events” (Stockburger, 2015).

LITERATURE REVIEW

The term “Big Data” refers to data that is too voluminous [by virtue of its size which can run into terabytes, petabytes or even exabytes], too complex [by virtue of its structured or unstructured data type], of too many varieties [such as text, graphic, sound, and/or video], and too fast [by virtue of its arrival rate] to be processed with a standalone computer or server (“big data: what it is, and why it matters”, 2020).

A structured data is data that conforms to a predetermined format like in a relational database where data types, data items, data sizes, and data arrangements follow a given format. On the other hand, an unstructured data is data that appears in random format, and so does not fit into any predetermined format. Examples of such data are data downloaded from websites or received from social media. Such data consists of text, graphics, sound, and video in any order.

Big Data can be collected quickly from various sources such as sensors, camera, websites, emails, social media, online questionnaire, etc and then uploaded to a cloud-based database such as Hadoop, NoSQL, Amazon simple storage server (S3), etc for immediate storage and processing.

According to Margaret R. (2019),

> The computing power required to quickly process huge volumes and varieties of data can overwhelm a single server or server cluster. Organizations must apply adequate processing capacity to big data tasks in order to achieve the required velocity. This can potentially demand hundreds or thousands of servers that can distribute the processing work, and operate collaboratively in a clustered architecture [like cloud computing]

Any cloud computing company can readily provide Big Data services. A list of reputable cloud computing companies is provided by (“15 Top Cloud Computing service providers”, 2020).

According to Ben Kazora (2019), “In today’s data-driven environment, businesses utilize and make big profits from big data. Big data, in turn, empowers businesses to make decisions based on trends, facts, and statistical numbers”. Furthermore, Grow (2020) stated that, “Businesses can harness data to make decisions about: (i) finding new customers, (ii) increasing customer retention, (iii) improving customer service, (iv) better managing marketing efforts, (v) tracking social media interaction, and (vi) predicting sales trends”. In summary, the author stated that, “any business with a website, a social media, and accepts electronic payments of any kind, is actually collecting data about customers, user habits, web traffic, demographics, and much more”, (Grow, 2020). This shows that, “the best-run companies are data-driven, and this skill sets businesses apart from their competitors, (Tomasz T., 2016).

The major issue in Big Data is not actually about the huge volume of data it
contains, but on how to factor out key information from it for decision making. One way of achieving this, is through Regression technique, as demonstrated by (Sunghae et al, 2015). Regression analysis, as explained by (Samprit and Ali, 2012), is a statistical technique used for studying the relationships between variables so that a mathematical equation called a regression model can be developed and used for predicting the value of an unknown variable [called the dependent or

response variable] from one or more known variables [called the independent or predictor variables]. (Sunghae et al, 2015) showed how Big Data can be decomposed into various sub units to conform to the various variables of interest involved in a regression model.

A sample size that is appreciably large was then collected from each sub unit for easier data analysis using simple regression technique. The analysis derived from each of the sub units was finally used to formulate a multi regression equation. This approach is similar to the "stratification" technique
used in statistics whereby a population data is divided into various subgroups of heterogeneous data sets for statistical analysis. This ensures that different kinds of data in the population are well represented for proper statistical analysis. According to Moser and Kalton (1979), "stratification is usually required if certain nonhomogeneities are present in the population". He concluded by saying that stratification does not require the size of each stratum to be the same, rather the sample size for each stratum should be proportional to the population size of that stratum. This goes along with the "Divide and Conquer" technique used in computer programming whereby a major task (or Big Data in this case) is broken down into subtasks that are simple enough for easy programming or processing.

The computer simulation technique was used in this work to select the various samples required for the experiment. Simulation is experimentation with a model such that the behavior of the model imitates some salient aspects of the behavior of the system under study (White and Ingalis, 2009). Simulation is typically used in order to save time and/or cost of studying the real system inasmuch it closely mirrors the behavior of the real system (Odoh, 2011) and (Hossein, 2020).

METHODOLOGY

The computer Simulation technique was used to generate input data of size n = 10,000 for each of the following five external factors affecting the growth of a small or large business enterprise (as considered by the researcher):

i. purchasing power of the people
ii. existing number of competitive products/services
iii. existing number of retail outlets
iv. available means of advertisements
v. available means of transportation

Table 1 (which is provided in Appendix A) shows the format or structure for collecting data for each of the aforementioned factors.

The table shows the five major data entries and their structure for data collection. The 1st entry: "*Purchasing Power of the people*" is used to determine the number of people that belong to the various income categories so that their purchasing power can be classified according to the associated codes. The 2nd entry: "*Existing number of competitive products/services"* contains class intervals for determining the possible number of existing similar competitive products/services in the locality. The 3rd entry: "Existing n*umber of retail outlets*" contains class intervals for determining the possible number of retail outlets in the locality. The 4th entry: "*Available means of advertisements*" contains code numbers for classifying the possible advertisement means in the locality. Similarly, the 5th and last entry: "*Available means of transportation*" contains code numbers for classifying the transport means in the locality.

DATA ANALYSIS

Multi Linear Regression technique was used to analyze the Big Data collected from the five input variables through simulation technique. Regression Analysis requires us to have a pre-knowledge of the dataset (Y,X) where X is the set of independent variables called the *predictor variables*, and Y is the dependent variable called the *response variable*. When the pre-knowledge of the dataset (Y,X) is known, the future value of Y can be predicted whenever there are some variations in the predictor variables. The steps involved in using multi linear Regression analysis, according to (Samprit and Ali, 2012), are as follows:

1) statement of problem (that is, formulate an appropriate question relating the dependent and independent variables)
2) identify the set of independent variables that will be used for predicting the dependent variable
3) make appropriate assumptions about the type of relationships (direct or inverse) between the dependent and independent variables
4) collect data about the dependent and independent variables
5) use 'least squares method' to determine the parameters in the regression model based on the collected data
6) substitute the calculated parameters into the regression model, and then use the model to make predictions about the dependent variable for any change in the independent variables.
7) compare the predicted value with actual results

The equation of a Multi Linear Regression equation, according to (Samprit and Ali, p.57) is given as,

$$Y^{\prime} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p + \epsilon \qquad \textbf{(3.1)}$$

Where,

$Y^{\prime}$ is the predicted value of the response variable

X_i (i = 1 to p) are the independent or predictor variables used for predicting $Y^{\prime}$

β_0 is a constant coefficient. It is the estimated value of $Y^{\prime}$ when all the
predictor variables are equal to zero (that is, $X_1 = X_2 = X_3 = \ldots = X_p = 0$). In other words, β_0 is the estimated value of $Y^{\prime}$ when all the predictor variables are
ignored (or not taken into consideration)

β_i, (i = 1 to p) are the regression coefficients used in determining the change in $Y^{\prime}$ that corresponds to one unit change in X_i when all the other predictor variables are held constant. For instance, **β_1** is the change in $Y^{\prime}$ to one unit change in X_1 when X_i (i = 2 to p) are held constant.

ϵ is the error term incurred in the predicted value of $Y^{\prime}$. However, since we are dealing with big data, the error term tends to zero if the sample size is very large.

Note1: the interpretation of β_i is ideal when all the predictor variables are not correlated (that is, there are independent of one another), otherwise (that is, if there are correlated or dependent of each other) β_i becomes a partial regression coefficient since its value is jointly contributed by all the X_i. In this work, the predictor variables, X_i are not correlated, as shown in table 2 (of Appendix A). This enables us to determine the independent effect on Y by each of the predictor variables, X_i. (See Table 2)

Note2: when the number of predictor variables is equal to 1, the regression equation of (eqn. 3.1) becomes a simple linear regression equation which can be written as,

$$Y' = \beta_0 + \beta_1 X_1 + \epsilon$$

Thus, a simple linear regression equation is a special case of a multi linear regression equation. In other words, a multi linear regression equation is an extension of a simple linear regression equation.

Equation (3.1) will be applied in this work for predicting the growth index of a business entity.

In this case, the equation can be rewritten as,

$$Y' = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon \qquad \textbf{(3.2)}$$

Where:

Y' = the predicted growth index of a business entity whose value lies between 1 and 10 (where, 1 is the smallest growth index, and 10 is the highest)

X_1 = purchasing power of the people

X_2 = existing number of competitive products/services

X_3 = existing number of retail outlets

X_4 = available means of advertisement

X_5 = available means of transportation

β_0 = the minimum growth index of a business when X_i (i = 1 to 5) is equal to zero

β_1 = the change in Y' to 1 unit change in X_1 when X_i (i = 2 to 5) are held constant

β_2 = the change in Y' to 1 unit change in X_2 when $X_{i,\ i\neq 2}$ (i = 1 to 5) are held constant

β_3 = the change in Y' to 1 unit change in X_3 when $X_{i,i\neq 3}$ (i = 1 to 5)
are held constant

β_4 = the change in Y' to 1 unit change in X_1 when $X_{i,i\neq 4}$ (i = 1 to 5) are held constant

β_5 = the change in Y' to 1 unit change in X_1 when X_i (i = 1 to 4) are held constant

ϵ = the error incurred in the predicted value of Y'

Now, since the predictor variables, X_i (i = 1 to 5) are independent of each other, equation (3.2) can be transformed into a series of simple linear regression equations as follows:

Given, $Y' = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$ **(3.2)**

$$\text{Then,}\quad \left.\begin{array}{l} Y_1' = \beta_{01} + \beta_1 X_1 + \epsilon \\ Y_2' = \beta_{02} + \beta_2 X_2 + \epsilon \\ Y_3' = \beta_{03} + \beta_3 X_3 + \epsilon \\ Y_4' = \beta_{04} + \beta_4 X_4 + \epsilon \\ Y_5' = \beta_{05} + \beta_5 X_5 + \epsilon \end{array}\right\} \qquad \textbf{(3.3)}$$

Thus, equation (3.2) can be approximated by equation (3.3) where,

Y_1', Y_2', Y_3', Y_4', and Y_5' are the estimators of Y'

And $\beta_0 \approx (\beta_{01} + \beta_{02} + \beta_{03} + \beta_{04} + \beta_{05}) / 5$

We shall therefore conduct five different sets of regression analysis involving each of the simple linear regression equations in equation (3.3) in order to determine the coefficients: β_0, β_1, β_2, β_3, β_4, and β_5, and then combine them into one multi linear regression equation to give us back equation (3.2) that will be used to predict the cumulative growth index of a business

- The first regression equation: $Y_1' = \beta_{01} + \beta_1 X_1 + \epsilon$ requires us to estimate the growth index, Y_1' of a business entity with respect to X_1 = the purchasing power of the people.
- The second regression equation, $Y_2' = \beta_{02} + \beta_2 X_2 + \epsilon$ requires us to estimate the growth index, Y_2' of a business entity with respect to X_2 = existing number of competitive products/services in the locality.
- The third regression equation, $Y_3' = \beta_{03} + \beta_2 X_3 + \epsilon$ requires us to estimate the growth index, Y_3' of a business entity with respect to X_3 = existing number of retail outlets.
- The fourth regression equation, $Y_4' = \beta_{04} + \beta_4 X_4 + \epsilon$ requires us to estimate the growth index, Y_4' of a business entity with respect to X_4 = available means of advertisement in the locality.
- The fifth regression equation, $Y_5' = \beta_{05} + \beta_5 X_5 + \epsilon$ requires us to estimate the growth index, Y_5' of a business entity with respect to X_5 = available means of transportation in the locality.

Table3 (of Appendix A) shows the type of relationship between the predictor variables and response variables, as well as the *assumed growth index values* of the response variables for each of the predictor variables.

From the first simple linear regression equation in (3.3), $Y_1' = \beta_{01} + \beta_1 X_1 + \epsilon$, the coefficient, β_1 can be determined from the method of least squares,

according to (Lawrence, 1980) by the formula,

$\beta_1 = [\sum XY - nX'Y'] / [\sum X^2 - nX'^2]$ **(3.4)**

where,

$X' = (\sum X)/n$ and $Y' = (\sum Y)/n$

Similarly, the constant, β_{01} is given by the formula, $\beta_{01} = Y' - \beta_1 X'$ **(3.5)**

Thus, if we collect a sample of say n = 10,000 about the purchasing power of people in a business locality, we use the sample to determine the values of the constants, β_1 and β_{01} from equations (3.4)

and (3.5). The same technique is applied in determining the values of the other constants: β_2 and β_{02} for the second regression equation; β_3 and β_{03} for the third regression equation; β_4 and β_{04} for the fourth regression equation; β_5 and β_{05} for the fifth regression equation whose sample size for each of them is n =10,000.
The Java code in Appendix B shows how the parameters (**β_{01} ,β_1, β_{02} , β_2 , β_{03} , β_3 , β_{04} , β_4 , β_{05} , β_5)** for each of the respective simple linear regression equations in (3.3) were derived from the simulated data of 10,000 input values for X and Y. Table 4 (of Appendix A) shows the summary of the values obtained from Appendix B when executed with a computer.

We now determine β_0 as (β_{01} + β_{02} + β_{03} + β_{04} + β_{05}) / 5. That is, β_0= (1.169 – 1.197 + 0.612 + 1.254 + 1.249) / 5 = (3.087) / 5 = 0.6174.

Now, substituting the values of β_1, β_2, β_3, β_4 and β_5 from table 4 into the multi regression equation of (3.2) we obtain the following regression model:-

$$Y' = 0.6174 + 1.142X_1 - 8.482X_2 + 3.812X_3 + 3.481X_4 + 1.959X_5 + \epsilon \quad \textbf{(3.6)}$$

Equation (3.6) can now be used to predict the growth index (Y') of a business entity for various values of X_1, X_2, X_3, X_4, and X_5 (being the external factors). A random number function such as, *randomNumbers.nextInt()* used by JAVA programming language can be used for varying the values of X_i(i = 1 to 5) for predicting Y' with respect to the change in X_i. Table 5 shows the randomNumbers.nextInt() function for varying X_i according to its data range.

RESULTS AND DISCUSSION

Refer to the multi linear regression equation of (3.6).
Suppose X_1, X_2, X_3, X_4, and X_5 are given as follows (according to the data in table 2):-

X_1 = 4 (small scale enterprise worker)
X_2 = 3 (existing competitive products that lie in the range 11-20)
X_3 = 4 (available retail outlets that lie in the range 51-100)
X_4 = 3 (means of advertisements being Radio and TV)
X_5 = 1 (poor road transportation)

Y' (the growth index of a business) can be predicted as follows:

$$Y' = 0.6174 + 1.142 * 4 - 8.482 * 3 + 3.812 * 4 + 3.481 * 3 + 1.959 * 1$$
$$= 7.3894$$
$$= 74\% \text{ (approx.)}$$

This figure is *appreciably Good* for the business entity.
Now, suppose X_1 remains at 4, but X_2 increases to 5 (which is the range 31-40 due to increased number of competitive products/services), and X_3 increases to 7 (which is the range 301-400 for increased number of retailers), and X_4 increased to 4 (so as to create room for more advertisements), while X_5 remains at 1,

(that is, poor access by road), then the new growth index of the business becomes,

$Y' = 0.6174 + 1.142 * 4 – 8.482 * 5 + 3.812 * 7 + 3.481 * 4 + 1.959 * 1$
$= 5.3424$
$= 53\%$ (approx.)

This figure is not very good for the business entity. But if the business entity increases X_3 (the number of retailers) to 8 (so as to be in the range 401-500), then the new growth index of the business now becomes,

$Y' = 0.6174 + 1.142 * 4 – 8.482 * 5 + 3.812 * 8 + 3.481 * 4 + 1.959 * 1$
$= 9.1544$
$= 91\%$ (approx.)

This figure is Excellent for the business entity. The company should keep working assiduously to maintain such growth index or even higher.

CONCLUSION / RECOMMENDATION

Multi linear regression analysis is one of the machine learning techniques used in the analysis of Big Data for making very meaningful predictions bothering on various issues on any sector of the economy such as business, healthcare, education, government, and a host of others. The technique, however, deals with structured quantitative data; therefore, any big data that is intended to be analyzed with regression analysis should be collected from the respondents in a structured manner for it to be processed efficiently for proper decision making.

Furthermore, the simulation technique employed in this work together with the Regression model
developed helps a business manager (whether small or large scale enterprise) make realistic predictions about her business growth in order to take proactive measures in ensuring that her business entity remains solvent at all times.

Lastly, the external factors affecting the growth of a business entity (both small and large enterprise), as mentioned in this work, are not exhaustive. Such factors are therefore subject to further research in order to enable a business entity grow from strength to strength for the betterment of the society.

APPENDIX A

Input factors/variables	Sample size (n)	Code number/classes used	
Purchasing power of the people	10,000	Jobless	1
		Student	2
		Junior civil servant	3
		Small scale enterprise worker	4
		Senior civil servant	5
		Big enterprise worker	6
Existing number of competitive products/services	10,000	0 – 5	1
		6 – 10	2
		11 – 20	3
		21 – 30	4
		31 – 40	5
Existing number of retail outlets	10,000	0 – 10	1
		11 – 30	2
		31 – 50	3
		51 – 100	4
		101 – 200	5
		201 – 300	6
		301 – 400	7
		401 – 500	8
Available means of advertisement	10,000	Radio	1
		TV/cable TV	2
		Mobile phone	3
		Bill board	4
Available means of transportation	10,000	Poorly accessible by road	1
		Accessible by water only	2
		Very accessible by road	3
		Accessible by both road and rail	4
		Accessible by both road and water	5
		Accessible by water, road, and rail	6

Table 2. Test for correlation between the predictor variables

predictor variables	**relationship**	**predictor variables**	**relationship**	**predictor variables**	**relationship**
X_1 vs X_2	insignificant relationship	X_2 vs X_3	insignificant relationship	X_3 vs X_4	insignificant relationship
X_1 vs X_3	insignificant relationship	X_2 vs X_4	insignificant relationship	X_3 vs X_5	insignificant relationship
X_1 vs X_4	insignificant relationship	X_2 vs X_5	insignificant relationship		
X_1 vs X_5	insignificant relationship				

Key: X_1 = purchasing power of the people, X_2 = existing number of competitive products
X_3 = existing number of retail outlets, X_4 = available means of advertisement
X_5 = available means of transportation

Table 3. Type of relation between the response and predictor variables, and the assumed values assigned to the response variables

Predictor variables X	Code number/classes		Response variable Y (and its assumed growth index with respect to X)	Type of Relationship Between X & Y
			Y_1'	
Purchasing power of the people (X_1)	Jobless	1	2	
	Student	2	4	
	Junior civil servant	3	4.5	Direct
	Small scale enterprise worker	4	6	
	Senior civil servant	5	7	
	Big enterprise worker	6	8	
			Y_2'	
Existing number of competitive products (X_2)	0 – 5	1	7.5	Inverse
	6 – 10	2	6	
	11 – 20	3	4.5	
	21 – 30	4	4	
	31 – 40	5	2.5	
			Y_3'	
Number of retail outlets(X_3)	0 – 10	1	4	
	11 – 30	2	5	
	31 – 50	3	6	
	51 – 100	4	6.5	Direct
	101 – 200	5	7	
	201 – 300	6	7.5	
	301 – 400	7	8	
	401 – 500	8	8.5	
			Y_4'	
Means of advertisements (X_4)	Radio only	1	4.5	
	TV only	2	6.5	Direct
	Radio and TV	3	7	
	Radio, TV, and Bill board	4	8.5	
			Y_5'	
Means of transport (X_5)	Poorly accessible by road	1	3	
	Accessible by water only	2	3.5	
	Very accessible by road	3	7	Direct
	Accessible by both road and rail	4	8	
	Accessible by both road and water	5	8.5	
	Accessible by water, road, and rail	6	9	

Table 4. Summary of values obtained from calculations made on the simulated data in Appendix B

Sample Size, n	**X**	**Data code for X**	**Y'**	**constants**	
				β_0	**β**
10,000	X_1	1	2	**$\beta_{01} = 1.169$**	**$\beta_1 = 1.142$**
		2	4		
		3	4.5		
		4	6		
		5	7		
		6	8		
10,000	X_2	1	7.5	**$\beta_{02} = -1.197$**	**$\beta_2 = 8.482$**
		2	6		
		3	4.5		
		4	4		
		5	2.5		
10,000	X_3	1	4	**$\beta_{03} = 0.612$**	**$\beta_3 = 3.812$**
		2	5		
		3	6		
		4	6.5		
		5	7		
		6	7.5		
		7	8		
		8	8.5		
10,000	X_4	1	4.5	**$\beta_{04} = 1.254$**	**$\beta_4 = 3.481$**
		2	6.5		
		3	7		
		4	8.5		
10,000	X_5	1	3	**$\beta_{05} = 1.249$**	**$\beta_5 = 1.959$**

Table 5. Random function for varying X_i

X	Data range for X	Randomfunction
X_1	1 to 6	1 + randomNumber.nextInt(6)
X_2	1 to 5	1 + randomNumber.nextInt(5)
X_3	1 to 8	1 + randomNumber.nextInt(8)
X_4	1 to 4	1 + randomNumber.nextInt(4)
X_5	1 to 6	1 + randomNumber.nextInt(6)

Note: *randomNumbers.nextInt(k) returns integer numbers randomly between 0 and k – 1. For instance, randomNumbers.nextInt(31) returns integer random numbers in the range 0 to 30, inclusive*

APPENDIX B

The JAVA code for deriving the parameters: **(β_{01} , β_1 , β_{02} , β_2 , β_{03} , β_3 , β_{04} , β_4 , β_{05} , β_5)** for each of the respective Simple Linear Regression equations of (3.3) from a simulated data of 10,000 input values for the independent and dependent variables, (X,Y), respectively.

```
import java.util.Scanner;
import java.util.Random;

public class CUCEN2020 {

   //function for processing simple linear regression model
   private static void determineRegressionCoeff(double arrayX[], double arrayY[], int N)
   {
      //let Xmean = (SumX)/N
      //let Ymean = (SumY)/N
      //let XYsum = Sum(X*Y)
      //let XYmeansum = Xmean * Ymean
      //let Xsqsum = Sum(X*X)

      //initialize variables
      double Xsum = 0;
      double Ysum = 0;
      double XYsum = 0;
      double Xsqsum = 0;

      //perform calculations
      for(int i=0; i<N; ++i)
     {
           Xsum = Xsum + arrayX[i];
           Ysum = Ysum + arrayY[i];
```

```
            XYsum = XYsum + arrayX[i] * arrayY[i];
Xsqsum = Xsqsum + arrayX[i] * arrayX[i];
 }
      double Xmean = Xsum/N;
      double Ymean = Ysum/N;
      double b0 = (XYsum - N * Xmean * Ymean)/(Xsqsum - N * Xmean * Xmean);
      double b = Ymean - b0 * Xmean;

      //show results
      System.out.println("\nRegression Analysis Results");
      System.out.println("b0 = " + b0);
      System.out.println("b = " + b);
      System.out.println("\nThe Simple Linear Regression model is shown below.");
      System.out.println("Y = " + b0 + " + " + b + "X");
   }

// The MAIN PROGRAM starts here
   public static void main(String[] args) {

      Random randomNo = new Random();

      //processing the 1st linear regression model for X1 and Y1
      Scanner inputn1 = new Scanner(System.in);
      System.out.print("\nEnter the sample size for X1 values: ");
      int n1 = inputn1.nextInt();

      //declare arrays for X1 and Y1
     double[] arrayX1 = new double[n1];
     double[] arrayY1 = new double[n1];

      //simulate values for X1
      int samplesize = n1;
      int count = 0;
      while(count <samplesize)
      {
         int Xvalue = 1 + randomNo.nextInt(6);   //X1 ranges from 1 to 6
         arrayX1[count] = Xvalue;
         if(Xvalue == 1)
         {
            arrayY1[count] = 2;
         }
         else if(Xvalue == 2)
         {
            arrayY1[count] = 4;
         }
         else if(Xvalue == 3)
         {
            arrayY1[count] = 4.5;
         }
         else if(Xvalue == 4)
         {
arrayY1[count] = 6;
         }
         else if(Xvalue == 5)
         {
            arrayY1[count] = 7;
         }
         else if(Xvalue == 6)
         {
            arrayY1[count] = 8;
         }
         ++count;
      }
      determineRegressionCoeff(arrayX1, arrayY1, n1);

      //processing the 2nd linear regression model for X2 and Y2
      Scanner inputn2 = new Scanner(System.in);
      System.out.print("\nEnter the sample size for X2 values: ");
      int n2 = inputn2.nextInt();
      //declare arrays for X2 and Y2
     double[] arrayX2 = new double[n2];
     double[] arrayY2 = new double[n2];

      //simulate values for X2
      samplesize = n2;
      count = 0;
      while(count <samplesize)
      {
         int Xvalue = 1 + randomNo.nextInt(5);   //X2 ranges from 1 to 5
         arrayX2[count] = Xvalue;
         if(Xvalue == 1)
         {
            arrayY2[count] = 7.5;
         }
         else if(Xvalue == 2)
         {
            arrayY2[count] = 6;
         }
         else if(Xvalue == 3)
         {
            arrayY2[count] = 4.5;
         }
         else if(Xvalue == 4)
         {
            arrayY2[count] = 4;
         }
```

```
          else if(Xvalue == 5)
          {
             arrayY2[count] = 2.5;
          }
++count;

       }
       determineRegressionCoeff(arrayX2, arrayY2, n2);

       //processing the 3rd linear regression model for X3 and Y3
       Scanner inputn3 = new Scanner(System.in);
       System.out.print("\nEnter the sample size for X3 values: ");
       int n3 = inputn3.nextInt();

        //declare arrays for X3 and Y3
       double[] arrayX3 = new double[n3];
       double[] arrayY3 = new double[n3];

       //simulate values for X3
       samplesize = n3;
       count = 0;
       while(count <samplesize)
       {
          int Xvalue = 1 + randomNo.nextInt(8);   //X1 ranges from 1 to 8
          arrayX3[count] = Xvalue;
          if(Xvalue == 1)
          {
             arrayY3[count] = 4;
          }
          else if(Xvalue == 2)
          {
             arrayY3[count] = 5;
          }
          else if(Xvalue == 3)
          {
             arrayY3[count] = 6;
          }
          else if(Xvalue == 4)
          {
             arrayY3[count] = 6.5;
          }
          else if(Xvalue == 5)
          {
             arrayY3[count] = 7;
          }
          else if(Xvalue == 6)
          {
             arrayY3[count] = 7.5;
```

```
    }
    else if(Xvalue == 7)
    {
        arrayY3[count] = 8;
    }
    else if(Xvalue == 8)
    {
        arrayY3[count] = 8.5;
    }
    ++count;
}
determineRegressionCoeff(arrayX3, arrayY3, n3);

 //processing the 4th linear regression model for X4 and Y4
Scanner inputn4 = new Scanner(System.in);
System.out.print("\nEnter the sample size for X4 values: ");
int n4 = inputn4.nextInt();
//declare arrays for X4 and Y4
double[] arrayX4 = new double[n4];
double[] arrayY4 = new double[n4];
//simulate values for X4
samplesize = n4;
count = 0;
while(count <samplesize)
{
    int Xvalue = 1 + randomNo.nextInt(4);   //X4 ranges from 1 to 4
    arrayX4[count] = Xvalue;
    if(Xvalue == 1)
    {
        arrayY4[count] = 4.5;
    }
    else if(Xvalue == 2)
    {
        arrayY4[count] = 6.5;
    }
    else if(Xvalue == 3)
    {
        arrayY4[count] = 7;
    }
    else if(Xvalue == 4)
    {
        arrayY4[count] = 8.5;
    }
    ++count;
}
 determineRegressionCoeff(arrayX4, arrayY4, n4);

 //processing the 5th linear regression model for X5 and Y5
```

```
        Scanner inputn5 = new Scanner(System.in);
        System.out.print("\nEnter the sample size for X5 values: ");
        int n5 = inputn5.nextInt();
        //declare arrays for X5 and Y5
double[] arrayX5 = new double[n5];
        double[] arrayY5 = new double[n5];

        //simulate values for X5
        samplesize = n5;
        count = 0;
        while(count <samplesize)
        {
            int Xvalue = 1 + randomNo.nextInt(6);  //X5 ranges from 1 to 6
            arrayX5[count] = Xvalue;
            if(Xvalue == 1)
           {
              arrayY5[count] = 3;
           }
           else if(Xvalue == 2)
           {
              arrayY5[count] = 3.5;
           }
           else if(Xvalue == 3)
           {
              arrayY5[count] = 7;
           }
           else if(Xvalue == 4)
           {
              arrayY5[count] = 8;
           }
           else if(Xvalue == 5)
           {
              arrayY5[count] = 7.5;
           }
           else if(Xvalue == 6)
           {
              arrayY5[count] = 9;
           }
           ++count;
        }
        determineRegressionCoeff(arrayX5, arrayY5, n5);
    }

}
```

References

Ben Kazora, (2019), "Role of Big Data in Today's Business Environment". Available at: https://benkazora.medium.com/role-of-big-data-in-todays-business-environment-4c647285e434 . Accessed on March 25, 2020

"Big Data: What it is, and why it matters", (2020). Retrieved from https://www.sas.com/en_us/insights/big-data/what-is-big-data.html . Accessed Feb.18, 2020

Grow, (2020), "Why is Data important for Business?" Available at: https://www.grow.com/blog/data-important-business . Accessed on March 22, 2020

Hossein, A., (2020), modeling & Simulation in "Systems Simulation: The Shortest Route to Applications", available at: http://home.ubalt.edu/ntsbarsh/simulation/sim.htm#rwis (accessed March 19, 2020)

Kituu, S. (2020) "Factors influencing small business growth". Retrieved from https://pocreative.com/factors-influencing-small-business-growth/Accessed Feb. 19, 2020

Lawrence, L.(1980). Statistics – Meaning and Methods, (2nd ed.), Harcourt Brace Jovanovich inc., p.292

Margaret, R. (2020). "What is big data and why it is important". Retrieved from https://searchdatamanagement.techtarget.com/definition/big-data under "How big data is stored and processed". Accessed Jan. 29, 2020

Moser, C.A. &Kalton, G. (1979). Survey Methods in Social Investigation, (2nd ed.), Heinemann Educational Books, ltd., p.85

Odoh, L.C. (2011). Quantitative Techniques in management accounting for Business Decisions, De-Verge agencies Ltd., p.212

Samprit, C. & Ali, S.H. (2012). Regression Analysis by Example,(5th ed.), John Wiley& sons Inc., p.13

Stockburger, D.W. (2015). "Regression Models: Introductory Statistics: Concepts, Models, and Applications". Retrieved from http://www.psychstat.missouristate.edu/introbook/sbk16.htm Accessed Feb.11, 2020

Segal, T. "Big Data definition" (2020). Retrieved from https://www.investopedia.com/terms/b/big-data.asp Accessed Feb. 17, 2020

Sunghae, J., Seung-Joo, L., &Jea-Bok, R. (2015). "A divided Regression Analysis for Big data", International Journal of Software Engineering and Its Applications, vol.9, No.5 (2015), pp. 21-52, http://dx.doi.org/10.14257/ijseia.2015.9.5.03

Tomasz Tunguz, (2016), "Winning with Data". Available at, https://tomtunguz.com/winning-with-data/ . Accessed on March 11, 2020

White, P.K., and Ingalis, R.G. (2009), "Introduction to Simulation", conference: Proceedings of the 2009 Winter Simulation Conference, Dec., 2009, DOI: 10.1109/WSC.2009.5429315

"15 Top Cloud Computing Service provider companies". Retrieved from https://www.softwaretestinghelp.com/cloud-computing-service-provider/ Accessed on April 16, 2020