Study of Regression Techniques with their Some Applications

Khine Wah Wah Latt[1], Mya Mya Thin[2], Hnin Moh Moh Tun[3]

Abstract

Regression techniques are most widely used in large variety of applied research. In this paper, the least square technique for linear regression is introduced firstly. Then some non-linear regression techniques are expressed with their applications. Finally, advantages and disadvantages of linear regression, exponential regression, power regression, polynomial regression and gompertz regression are presented.

Keywords: linear regression, exponential regression, power regression, polynomial regression, gompertz regression.

INTRODUCTION

The most common method of displaying a set of bivariate data is using a **scatter diagram**. The bivariate pairs as sets of (x, y) coordinates plot them on a graph to obtain a set of points. The observation pairs are said to be **positively(or directly)** and **negatively(or inversely) correlated**. Statistically, the correlation between two variables can be measured as a number between –1 and +1 inclusive, where:

$$\begin{cases} +1 \text{ signifies 'perfect' positive correlation} \\ 0 \text{ signifies no correlation} \\ -1 \text{ signifies 'perfect' negative correlation} \end{cases}$$

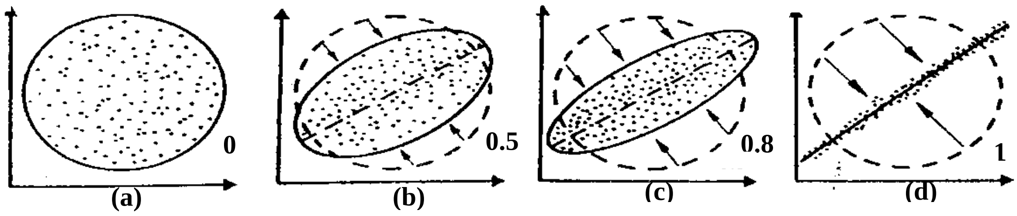

Figure 1

Where no correlation existed, the points on a scatter diagram could roughly be enclosed within a circle in Figure 1(a) and Figure 2(a). The situations for allocating numbers between 0 and 1 to varying degrees of positive correlation, and the perfect positive correlation would be obtained when the circle has been finally squashed into a straight line as in Figure 1(d). Next, depending on the degree of 'squash', an approximate number could be allocated between 0 and – 1. Figure 2 shows the perfect negative correlation with a coefficient of –1.

---

[1] Associate Professor, Dr, Department of Mathematics, Hinthada University
[2] Lecturer, Department of Mathematics, Hinthada University
[3] Tutor, Department of Mathematics, Hinthada University

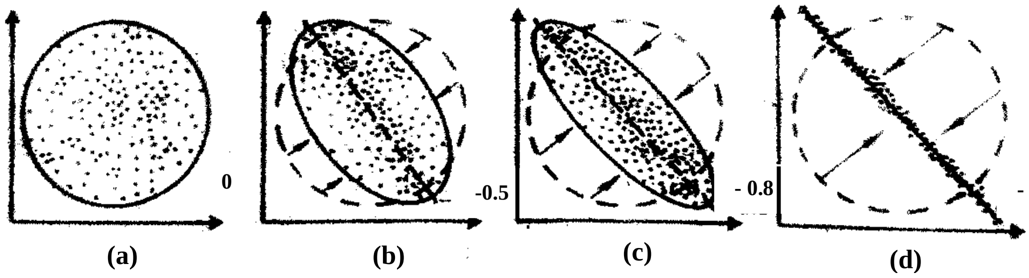

Figure 2

The Least Square Technique for Linear Regression

Definitions: Lines fitted to data, describing the mathematical relationship between variables are called **regression lines**. For any bivariate set of data, connecting variables x and y, whose points on a scatter diagram can be enclosed by an ellipse, there are always two uniquely defined regression lines: (i) take specific values of x, and draw the corresponding chords relating to relevant values of y will be vertical lines. Choose the chord mid-points as representative y values and finally join these up. As in Figure 3(a), the straight line obtained is the **regression line of y on x**: $y = a_1x + b_1$. **The least square regression line of y on x** is that line such that the quantity $\sum_{i=1}^{n} d_i^2$ is a minimum which can be seen in Figure 4(a).(ii)) take specific values of y, and draw the corresponding chords relating to relevant values of x will be horizontal lines. Choose the chord mid-points as representative x values and finally join these up. As in Figure 3(b), the straight line obtained is the **regression line of x on y**: $x = a_2y + b_2$.**The least square regression line of x on y** is that line such that the quantity $\sum_{i=1}^{n} e_i^2$ is a minimum which can be seen in Figure 5(a).

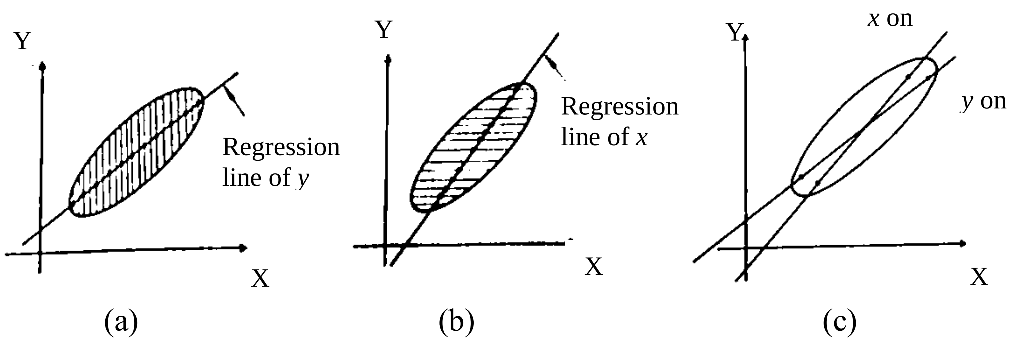

Figure 3

The above figure 3(c) shows these two regression lines superimposed on the same diagram, where it can be seen that the two lines are markedly different.

Theorem : Given the bivariate set $((x_1, y_1), (x_2, y_2),\dots , (x_n, y_n))$,

(a) the constants a_1 and b_1 in the least squares regression line of y on x : $y = a_1x + b_1$ can be determined by solving the simultaneous equations:

$$\left.\begin{aligned}\sum y &= a_1 \sum x + nb_1 \\ \sum xy &= a_1 \sum x^2 + b_1 \sum x\end{aligned}\right\} \text{ called the normal equations for } y \text{ on } x$$

(b) the constants a_2 and b_2 in the least square regression line of x on y : $x = a_2 y + b_2$ can be determined by solving the simultaneous equations:

$$\left.\begin{aligned}\sum x &= a_2 \sum y + nb_2 \\ \sum xy &= a_2 \sum y^2 + b_2 \sum y\end{aligned}\right\} \text{ called the normal equations for } x \text{ on } y$$

Proof: (a) The least square regression line of y on x is the line: $y = a_1 x + b_1$.

Let d_i be the minimum vertical distance between the point (x_i, y_i) and the line $y = a_1 x_i + b_1$.

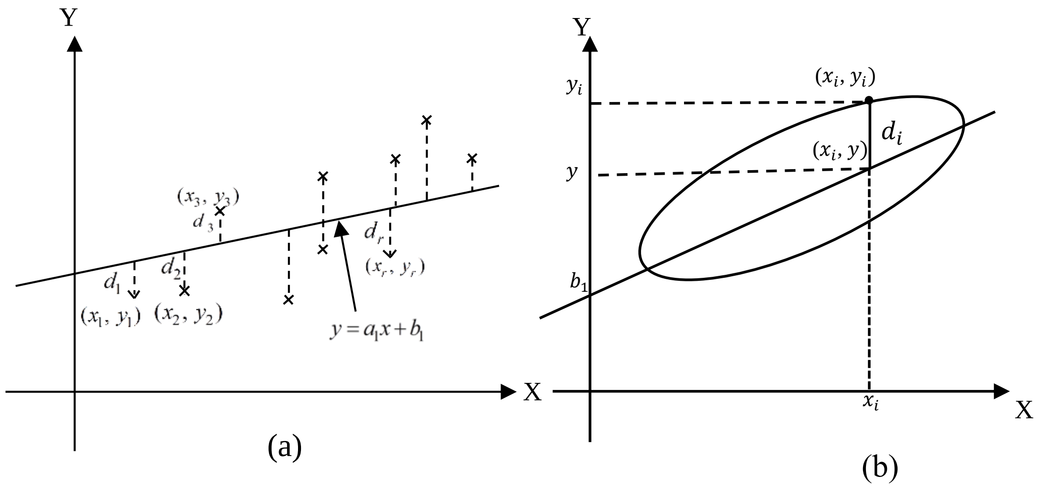

Figure 4

Therefore, $d_i = y_i - y = y_i - (a_1 x_i + b_1) = y_i - a_1 x_i - b_1$

$$\sum d_i^{\,2} = \sum_i (y_i - a_1 x_i - b_1)^2 = S \text{ (Say)}$$

To find the values of a_1 and b_1, that make S a minimum. We have to solve the two equations $\frac{\partial S}{\partial b_1} = 0$ and $\frac{\partial S}{\partial a_1} = 0$. Now, $\frac{\partial S}{\partial b_1} = \frac{\partial}{\partial b_1}\left\{\sum_i (y_i - a_1 x_i - b_1)^2\right\}$

$$= \sum_i \left\{\frac{\partial}{\partial b_1}(y_i - a_1 x_i - b_1)^2\right\}$$

$$= \sum_i \{2(y_i - a_1 x_i - b_1)(-1)\}$$

$$= \sum_i \{-2y_i + 2a_1 x_i + 2b_1)\}$$

$$= -2\sum y + 2a_1 \sum x + 2b_1 n .$$

Then, $\frac{\partial S}{\partial b_1} = 0$ gives, $-2\sum y + 2a_1 \sum x + 2b_1 n = 0$.

Therefore, $\sum y = a_1 \sum x + b_1 n$ is a required normal equation of *y* on *x*.

Next, $$\frac{\partial S}{\partial a_1} = \frac{\partial}{\partial a_1}\left\{\sum_i (y_i - a_1 x_i - b_1)^2\right\}$$
$$= \sum_i \left\{\frac{\partial}{\partial a_1}(y_i - a_1 x_i - b_1)^2\right\}$$
$$= \sum_i \{2(y_i - a_1 x_i - b_1)(-x_i)\}$$
$$= \sum_i \{-2x_i y_i + 2a_1 x_i^2 + 2b_1 x_i)\}$$
$$= -2\sum xy + 2a_1 \sum x^2 + 2b_1 \sum x .$$

Then, $\dfrac{\partial S}{\partial a_1} = 0$ gives, $-2\sum xy + 2a_1 \sum x^2 + 2b_1 \sum x = 0.$

Therefore, $\sum xy = a_1 \sum x^2 + b_1 \sum x$ is an another normal equation of *y* on *x*.

(a) The least square regression line of *x* on *y* is the line: $x = a_2 y + b_2$.

Let e_i be the minimum horizontal distance between the point (x_i, y_i) and the line $x = a_2 y + b_2$.

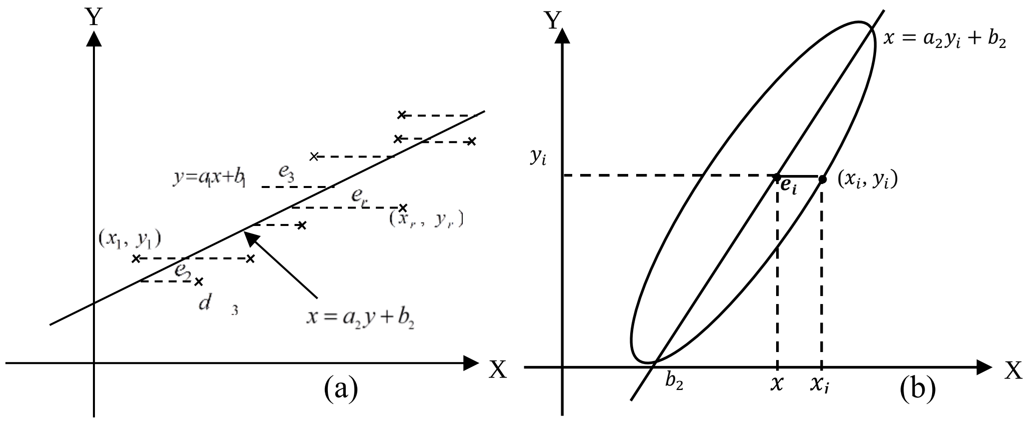

Figure 5

Therefore, $e_i = x_i - x \;= x_i - (a_2 y_i + b_2) \;= x_i - a_2 y_i - b_2$

$$\sum e_i^{\,2} = \sum_i (x_i - a_2 y_i - b_2)^2 = T \text{ (Say)}$$

To find the values of a_2 and b_2, that make *T* a minimum. We have to solve the two equations $\dfrac{\partial T}{\partial b_2} = 0$ and $\dfrac{\partial T}{\partial a_2} = 0$. Now, similar method is employed to obtain (b).

Definitions: Given the bivariate set$((x_1, y_1), (x_2, y_2),\ldots, (x_n, y_n))$and using the codings $X = \dfrac{x-a}{b}$; $Y = \dfrac{y-c}{d}$,can be calculated **the coded covariance** as $S_{XY} = \dfrac{1}{n}\sum XY - \overline{X}\,\overline{Y}$ and decoding $S_{xy} = bdS_{XY}$. The**regression coefficients** and **constants** for the least squares regression lines $Y = A_1 X + B_1$ (Y on X) and $X = A_2 Y + B_2$ (X on Y) can be calculate using

$A_1 = \dfrac{S_{XY}}{S_{X^2}}$, $B_1 = \overline{Y} - A_1\overline{X}$, $A_2 = \dfrac{S_{XY}}{S_{Y^2}}$ and $B_2 = \overline{X} - A_2\overline{Y}$. Moreover, the **product moment correlation coefficient** can be defined as $r_{xy} = r_{XY} = \dfrac{S_{XY}}{S_X S_Y}$. By the above definition,

written $r = \dfrac{\frac{1}{n}\sum XY - \overline{X}\,\overline{Y}}{\sqrt{\left(\frac{1}{n}\sum X^2 - \overline{X}^2\right)}\left(\sqrt{\frac{1}{n}\sum Y^2 - \overline{Y}^2}\right)}$.

Where $S_X = \sqrt{\frac{1}{n}\sum X^2 - \overline{X}^2}$ and $S_Y = \sqrt{\frac{1}{n}\sum Y^2 - \overline{Y}^2}$ are the standard deviations of X and Y values respectively.

Application 1: The following data gives the wheat yield (x) in millions of tons against the area planted (y) in millions of acres for Magway and Sagaing for the successive years 2010 and 2011:

x	3.7	4.1	3.4	3.8	3.4	3.3	4.2	4.7	4.7
y	2.2	2.5	2.2	2.3	2.4	2.1	2.5	2.7	2.8

The correlation coefficients can be calculated as:

x	y	$X = x - 3.4$	$Y = y - 2.5$	X^2	Y^2	XY
3.7	2.2	0.3	– 0.3	0.09	0.09	– 0.09
4.1	2.5	0.7	0	0.49	0	0
3.4	2.2	0	– 0.3	0	0.09	0
3.8	2.3	0.4	– 0.2	0.16	0.04	– 0.08
3.4	2.4	0	– 0.1	0	0.01	0
3.3	2.1	– 0.1	– 0.4	0.01	0.16	0.04
4.2	2.5	0.8	0	0.64	0	0
4.7	2.7	1.3	0.2	1.69	0.04	0.26
4.7	2.8	1.3	0.3	1.69	0.09	0.39
		$\sum X =$ 4.7	$\sum Y = -0.8$	$\sum X^2 = 4.77$	$\sum Y^2 = 0.52$	$\sum XY = 0.52$

$$\overline{X} = \frac{\sum X}{n} = \frac{4.7}{9} = 0.52 \;,\; \overline{Y} = \frac{\sum Y}{n} = \frac{-0.8}{9} = -0.09$$

$$S_{XY} = \frac{1}{n}\sum XY - \overline{X}\,\overline{Y} = \frac{1}{9}(0.52) - (0.52)(-0.09) = 0.10$$

$$S_X = \sqrt{\frac{1}{n}\sum X^2 - \overline{X}^2} = \sqrt{\frac{1}{9}(4.77) - (0.52)^2} = \sqrt{0.53 - 0.27} = 0.51$$

$$S_Y = \sqrt{\frac{1}{n}\sum Y^2 - \overline{Y}^2} = \sqrt{\frac{1}{9}(0.52) - (-0.09)^2} = \sqrt{0.06 - 0.01} = 0.22$$

Therefore the correlation coefficient is

$$r_{xy} = r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{0.10}{(0.51)(0.22)} = 0.89$$

Definitions: The given data is in a more complex form and its volume will be greater, the bivariate frequency distribution given as:

	x_1	x_2	$\cdots$	x_n	
y_1	f_{11}	f_{12}	$\cdots$	f_{1n}	f_{y_1}
y_2	f_{21}	f_{22}	$\cdots$	f_{2n}	f_{y_2}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
y_m	f_{m1}	f_{m2}	$\cdots$	f_{mn}	f_{y_m}
	f_{x_1}	f_{x_2}	$\cdots$	f_{x_n}	

where f_{y_1} , $f_{y_2}, \ldots, f_{y_m}$ are **row totals** and f_{x_1}, $f_{x_2}, \ldots, f_{x_n}$ are **column totals**. Referring to the given model, $\sum fy$ and $\sum fx$ would be calculated as: $f_{y_1}y_1 + f_{y_2}y_2 + \cdots + f_{y_m}y_m = \sum_{i=1}^{m} f_{y_i}y_i$ and $f_{x_1}x_1 + f_{x_2}x_2 + \cdots + f_{x_n}x_n = \sum_{i=1}^{n} f_{x_i}x_i$. The grand total of the table frequencies $\sum f$ is calculated as $\sum f_{y_i} = \sum f_{x_i}$. Also, $\sum fX^2 = \sum_{i=1}^{n} f_{x_i}x_i^2$ and $\sum fY^2 = \sum_{i=1}^{m} f_{y_i}y_i^2$. Furthermore, the most complicated quantity $\sum fXY$ can be fined by:

$$\sum fXY = f_{11}x_1y_1 + f_{12}x_2y_1 + f_{13}x_3y_1 + \cdots + f_{mn}x_ny_m .$$

Also the **covariance** as: $S_{XY} = \frac{\sum fXY}{\sum f} - \left(\frac{\sum fX}{\sum f}\right)\left(\frac{\sum fY}{\sum f}\right)$ and the **standard deviations of *X* and *Y* values** are $S_X = \sqrt{\frac{\sum fX^2}{\sum f} - \left(\frac{\sum fX}{\sum f}\right)^2}$ and $S_Y = \sqrt{\frac{\sum fY^2}{\sum f} - \left(\frac{\sum fY}{\sum f}\right)^2}$.

Application 2: The results of both an English and Mathematics examination for 73 students are tabulated below:

		Maths marks (x)					
		0 - 19	20-39	40-59	60-79	80-99	Total
	0 - 19	0	0	1	2	0	3
English	20 - 39	0	3	5	2	1	11
Marks	40 - 59	1	1	4	6	4	16
(y)	60 - 79	0	2	8	14	6	30
	80 - 99	0	0	4	7	2	13
	Total	1	6	22	31	13	73

Coding: $X = \dfrac{x-49.5}{20}$, $Y = \dfrac{y-49.5}{20}$

The boxes in the main body of the table into 3 sections. The middle figure is the frequency shown in the original table, the figure at top left is the product of the respective x and y values and the figure at bottom right is the product of the xy value and the frequency. That is in the form:

xy
f
fxy

To find the regression line of *y* on *x*, *x* on *y* and product moment coefficient of correlation, the frequency table can be constructed as follows:

Y \ X	–2	–1	0	1	2	f	fY	fY^2
–2	4 0 0	2 0 0	0 1 0	–2 2 –4	–4 0 0	3	– 6	12
– 1	2 0 0	1 3 3	0 5 0	–1 2 –2	–2 1 –2	11	– 11	11
0	0 1 0	0 1 0	0 4 0	0 6 0	0 4 0	16	0	0
1	–2 0 0	–1 2 –2	0 8 0	1 14 14	2 6 12	30	30	30
2	–4 0 0	–2 0 0	0 4 0	2 7 14	4 2 8	13	26	52
f	1	6	22	31	13	73	39	105
fX	– 2	– 6	0	31	26	49	$\sum fXY = 41$	
fX^2	4	6	0	31	52	93		

From the table, $\sum f = 73$, $\sum fX = 49$, $\sum fY = 39$, $\sum fX^2 = 93$, $\sum fY^2 = 105$ and $\sum fXY = 41$.

$$S_{XY} = \frac{\sum fXY}{\sum f} - \left(\frac{\sum fX}{\sum f}\right)\left(\frac{\sum fY}{\sum f}\right) = \frac{41}{73} - \left(\frac{49}{73}\right)\left(\frac{39}{73}\right) = 0.2030$$

$$S_{X^2} = \frac{\sum fX^2}{\sum f}\left(\frac{\sum fX}{\sum f}\right)^2 = \frac{93}{73} - \left(\frac{49}{73}\right)^2 = 0.8234$$

$$S_{Y^2} = \frac{\sum fY^2}{\sum f}\left(\frac{\sum fY}{\sum f}\right)^2 = \frac{105}{73} - \left(\frac{39}{73}\right)^2 = 1.1529$$

Also, $S_X = 0.9072$ and $S_Y = 1.0738$

$$A_1 = \frac{S_{XY}}{S_{X^2}} = \frac{0.2030}{0.8234} = 0.2465\,,\ A_2 = \frac{S_{XY}}{S_{Y^2}} = \frac{0.2030}{1.1529} = 0.1761\,,$$

$$B_1 = \overline{Y} - A_1\overline{X} = \frac{\sum fY}{\sum f} - 0.2465\left(\frac{\sum fX}{\sum f}\right) = 0.5342 - (0.2465)(0.6712) = 0.3687$$

$$B_2 = \overline{X} - A_2 Y = \frac{\sum fX}{\sum f} - (0.1761)\left(\frac{\sum fY}{\sum f}\right) = 0.6712(-0.1761)(0.5342) = 0.5771$$ The correlation coefficient for Mathematics (x) and English (y) marks is

$$r_{xy} = r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{0.2030}{(0.9072)(1.0738)} = 0.2084$$

The regression line of y on x is $Y = A_1 X + B_1$

$$\frac{y-49.5}{20} = 0.2465\left(\frac{x-49.5}{20}\right) + 0.3687$$

$$y - 49.5 = 20\left(0.0123x - 0.6101 + 0.3687\right)$$

$$y = 0.2465x - 4.8280 + 49.5\text{ f}$$

Therefore $y = 0.2465x + 44.5720\,.$

If a student was absent from the English exam, but obtained 80 marks in Mathematics. So, for Mathematics (x) marks of 80, English marks (y) can be estimated using the above data by:

$$y = (0.2465)\ (80) + 44.6720 = 64.3920$$

Next, the regression line of x on y is $X = A_2 Y + B_2$

$$\frac{x-49.5}{20} = 0.1761\left(\frac{y-49.5}{20}\right) + 0.5771$$

$$x - 49.5 = 20\left(0.0088y - 0.4358 + 0.5771\right)$$

$$x = 0.1761y + 2.8260 + 49.5$$

Hence $x = 0.1761y + 52.3260\,.$

Another applications of linear regression: (1) **Epidemiology** - The branch of medicine which deals with the incidence, distribution and possible control of diseases and other factors relating to health. For example, the linear regression can be constructed in which cigarette smoking (the explanatory variable) and the lifespan of an individual (the dependent variable). (2) **Finance** - For evaluation of the systematic risk of funding, the linear regression model can be used that relates the return on the investment to the return on all risky assets. (3) **Econometric** - The independent and dependent variables can be fitted the line of linear regression estimating model which is used in wide variety of economic relationship. (4) **Environmental Science** - the most widely use in the life and earth science. For example, linear regression can be used to test whether temperature (the explanatory variable) is a good predictor of plant height (the response variable).

Some Non-Linear Regression Techniques with their Applications

Exponential Regression: The process of exponential regression involves finding the best fits a set of data in the canonical exponential regression model $y = ab^x$, where $a \neq 0$. Figure 6 shows the exponential regression.

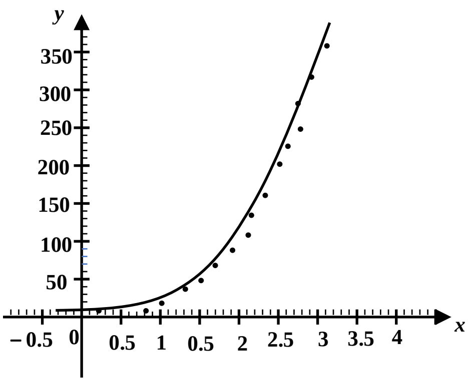

Figure 6

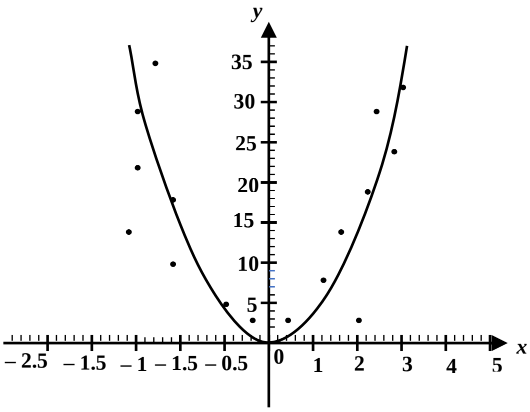

Figure 7

For example, $3e^2$ denotes the relative predictive power of the exponential model. Its value is between 0 and 1. If the value is close to 1, the model is more accurate.

Applications of exponential regression: (1) **Population Growth** - If the population grows twice every 3 days for a certain number of micro-organisms or animals, this can be represented as an exponential function. (2) **Exponential Decay** - For the population size may decrease instead of increasing, the exponential regression can be used with negative exponential power.

Power Regression: If the output variable is proportional to the input variable raised to a power, the power regression can be used. Power regression formula is of the form $y = x^b$, where $x \neq 0$ is the explanatory variables and b is constant. The power regression equation will be used to predict output values that lies within (interpolate) or outside (extrapolate) the plotted values. Figure 7 shows the power regression curve of $y = x^2$.

Applications of power regression:(1) **Weather Forecasting -**Highly used in weather forecasting to predict the rain as it provides a model which is very closed to the actual model. (2) **Physiotherapy** - To find the oxygen uptake by a person during cycling, running and walking in real-time.

Polynomial Regression: A special type of multiple linear regression is the polynomial regression. The output variable y is modelled as an n^{th} degree polynomial in terms of x as in equation $y = a_0 + a_1x^1 + a_2x^2 + \cdots + a_nx^n$, where $a_0, a_1, a_2, \ . \ . \ . \ . \ , a_n$ are constant coefficients. Figure 8 shows the polynomial regression curve.

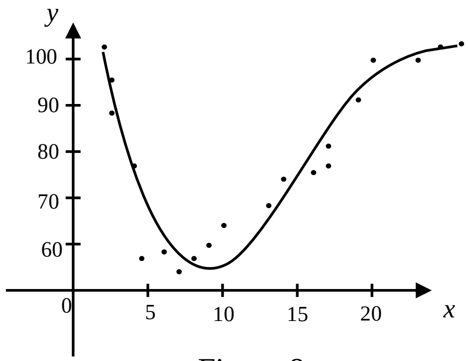

Figure 8

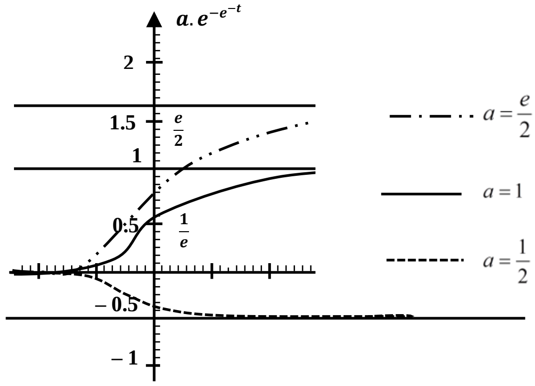

Figure 9

Applications of polynomial regression: (1) **Medicine** - To study the isotopes of the sediments, the rise of different diseases within any population and the generation of any synthesis. (2) **Environmental Study** - To predict the occurrence of events such as tsunami, thunderstorms, and sandstorms in advance to timely avoid their effects. (3) **Archaeology** - To predict the age of artefacts and how many years ago ancient civilizations were present.

Gompertz Regression: A mathematical model used for a time series, in which the sigmoid function shows growth as being slowest at the beginning and the end of a given period but very fast in the middle. Non-linear function, $y = ae^{-be^{-cx}}$ is the form of gompertz regression. In gompertz regression function, it has three parameters. The parameter a (asymptote) is obtained by applying limit of the function as x approaches infinity and the parameter b is the displacement along the x-axis, and finally, the growth rate is described by the parameter c which can be seen in Figure 9.

Application of gompertz regression: (1) **Biology** - It is useful for many phenomena such as the increase in the number of cancerous tumours limited to an organ without, the increase in the number of individuals in a population. (2) **Geology** - It is used to forecast the total natural gas consumption and to compare our results with those obtained.

Advantages and Disadvantages of Regression Techniques

Advantages: (i) **Linear regression:** The mathematical equation of the linear regression technique is very easy to understand. It can be used to model linearly separable data sets and to find the nature of the relationship among variables. (ii) **Exponential regression:** Only three data pieces are required for exponential regression, so it is easy to understand and apply to get accurate forecasts. (iii) **Power regression:** Power regression model is much close to real problems having the minimum error. It can be used to obtain better results. (iv) **Polynomial regression:** Non-linear problems are solved with good accuracy. (v) **Gompertz regression:** Because of its highly exponential behavior, it gives a more accurate solution for rapid phenomena of real-word problems such as radioactive decay, bacterial growth, and gas expansion.

Disadvantages: (i) **Linear regression:** It fails to fit complex data sets property because most of the naturally occurring phenomena are non-linear. (ii) **Exponential regression:** It is best utilized for short-term forecasting. (iii) **Power regression:** It is very hard to implement and complex to understand as its parametric form. (iv) **Polynomial regression:** If one or two outliers are occurred in the data set, its technique is sensitive to determine the accurate value. (v) **Gompertz regression:** Because of its high exponential nature, it does not give exact value for some conditions.

CONCLUSION

The regression technique are types of predictive modelling techniques that are useful statistical methods to estimate upon the independent variables and dependent variables. Many regression techniques have been developed and many more are in process for real-word situations. Depending upon the type of problems, choose the technique to get a robust solution.

Acknowledgements

I would like to *express* our gratitude to Dr Theingi Shwe, Rector, Hninthada University, Dr Yee Yee Than Pro-Rector, Hninthada University, and Dr Cho Kyi Than Pro-Rector, Hninthada University for their permission to summit this research paper. Special thanks are also due to Dr Myo Nilar, Professor and Head, Department of Mathematics, Hinthada University, for her valuable suggestions and guidance.

References

Francis, A., *Advance Level Statistics*, Stanley Thornes Ltd., London, 1979.

Muhammad, A. I., Application of Regression Techniques with their Advantages and Disadvantages, Elektron Magazine, 2020, 11-16.

Stefan, W., & Steven, R.Contenoble, *Finite Mathematics + Applied Calculus*, Cengage Learning products, Boston, USA, 2011.