Outbreak, Surveillance, Investigation & Response (OSIR) Journal

Field Epidemiology Training Program, Bureau of Epidemiology
Department of Disease Control, Ministry of Public Health, Thailand
Tel: +6625903894, Fax: +6625903845, Email: osireditor@osirjournal.net, http://www.osirjournal.net

Grammar of Science: Engines of Statistical Models

Jaranit Kaewkungwal*

Mahidol University, Thailand

* Corresponding author, email address: jaranit.kae@mahidol.ac.th

Introduction

An engine or motor is a machine designed to convert one form of energy into mechanical energy[1]. Heat engines burn a fuel; electric motors convert electrical energy; pneumatic motors compressed air; then the energy conversion from those engines is used to do the work. The same principle applies in biological systems; for example, molecular motors (e.g. myosins in muscles) use chemical energy to create forces and motion[1]. In statistics, I would say that there are two mathematical engines which are the driving forces underneath almost all statistical methods/models: “regression” and “correlation”.

There are varieties of regression and correlation. But we will focus on the classic engines of all, the “Linear regression” and the “Pearson’s correlation”. It is important to understand the mechanism of these engines because they are the foundations or driving forces of all other types of regression and correlation and used as basis for several other statistical models. We will take a close look at the development of linear regression model, the steps in derivation of Pearson’s correlation coefficient, and the mathematical linkage between the two statistical terms.

What are Correlation and Regression?

The two statistics are both similar and different. Regarding the meaning, correlation determines co-relationship or association of two variables while regression describes how an independent variable is numerically related to the dependent variable[2]. Correlation quantifies the degree to which two variables (say, x and y) are related. Regression identifies the “best” equation that predicts y from x. We can say that correlation does not distinguish the dependent variable (y) and Independent variable (x) but regression tends to do so[3-4].

Both statistics are based on linear relationship. Correlation assumes that the association is linear, that one variable increases or decreases a fixed amount for a unit increase or decrease in the other. Regression, on the other hand, involves estimating the best straight line to summarize the association between the variables. Therefore correlation coefficient infers the extent to which two variables are associated with each other while regression coefficient estimates the impact of a unit change in the variable (x) on the variable (y)[2,4].

A Brief History

The name “Pearson’s correlation” leads to believe that Karl Pearson (1857-1936) developed this statistical measure himself. Although he is the one who made correlation as currently known today, but history went back before his time.

Sir Francis Galton (1822-1911) is commonly regarded as the founder of the statistical techniques of correlation and linear regression[5-7]. Galton, a cousin of Charles Darwin (1809-1882) was a distinguish scientist in biology, psychology and applied statistics. His works on genetics and heredity provided the initial inspiration that led to regression and correlation[5,6]. As Galton's biographer, Pearson described interesting story of the discovery of the regression analysis. In 1875, Galton had distributed packets of sweet pea seeds to seven friends to harvest the seeds and return the next generations to him. Each friend received seeds of uniform weight but there was substantial variation across different packets. Galton then plotted the weights of the daughter seeds against the weights of the mother seeds, and he discovered a straight line relationship with positive slope of the two weights[5].

But Pearson also credited Auguste Bravais (1811-1863), a professor of astronomy and physics, as a founder of initial mathematical formulae for regression and correlation concepts. As noted by Pearson, Bravais wrote about "mathematical analysis on the probability of errors of a point" which is the fundamental theorems of the correlational calculus[7].

Some argued that regression and correlation went even further back to the legendary mathematician Carl Friedrich Gauss (1777-1855) and Adrien-Marie Legendre (1752-1833) who independently discovered the method of "least squares", the essential feature of linear regression[8].

Basics of Regression

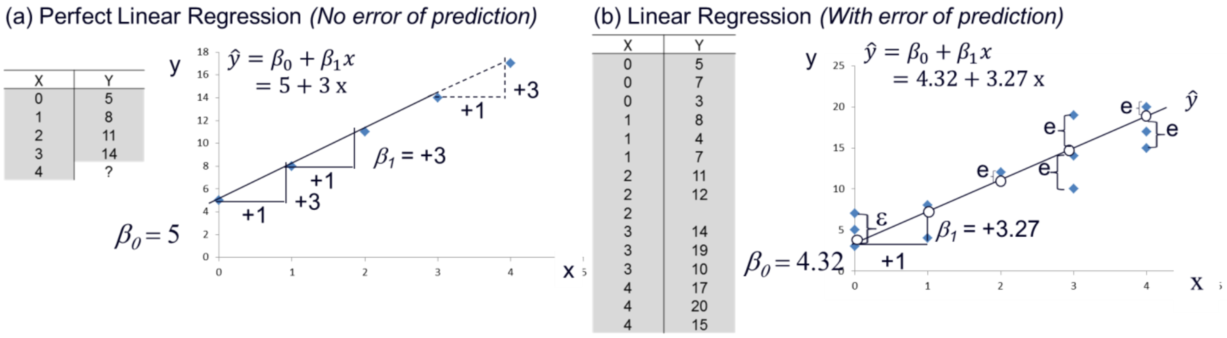
A simple way to explore relationships between the two variables is to construct a scatter diagram with one variable on the vertical scale and the other on the horizontal scale. In regression model the "dependent variable" is usually plotted on the vertical axis while the "independent variable" on the horizontal axis, or baseline[4]. The main purpose of regression analysis is to obtain an equation explaining the relationship between variables. Such equation is frequently used to predict the future (or unknown) value of the dependent variable, or to understand which factors (independent variables) cause or associate with an outcome (dependent variable)[8].

Back to the history of regression, in studying data on relative sizes of parents and their offspring in various species of plants and animals, Galton noted that a larger-than-average parent tends to produce a larger-than-average child, but the child is likely to be less large than the parent in terms of its relative position within its own generation[9]. Galton termed this phenomenon a "regression towards mediocrity", which in modern terms is a "regression to the mean". Regression to the mean can be expected in natural settings, for example, relative to others in the same class, your final exam score could be expected to be less good or bad than your midterm score[9]. However, the term "regression" later evolved and changed to the concept of slope determining the relationship between the independent variable(s) and dependent variable.

Regression is a statistical technique for estimating the change in the dependent variable (y) due to the change in one or more independent variables (x). The decision of which variable is dependent or independent variable must be pre-determined as the best-fit line will be different if you swap the two[3]. The simple regression line of y on x is expressed as: $\hat{y} = \beta_0 + \beta_1 x$ where, β_0 = constant (intercept), β_1 = regression coefficient (slope). The β_0 and β_1 are the two regression parameters in the equation. As shown in the hypothetical scenario (steps of a walking baby), in figure 1 (a); at Day 0 (baseline) a baby is able to walk 5 steps, and then 8, 11, 14 steps on Days 1, 2, and 3, respectively. This is one sample with a perfect linear relationship; the linear regression equation here is $\hat{y} = 5 + 3x$ where β_0 = 5 steps (intercept: when x=0) and β_1 = 3 (slope when x changes 1 unit-day, y changes 3 unit-steps). So, if this linear pattern holds, you can expect that the baby will walk 17 steps on Day 4.

But when the researcher collects data on walking steps from many babies with "not so perfect" linear relationship, the numbers will vary for each baby as shown in figure 1 (b), i.e. not all observed values fall on the straight line. "Error" as used in mathematical/statistical sense since 1726 is defined as "any deviation from accurate determination (or true value)" assuming that the accurate determination is obtainable[7]; see figure 1 (b), the distances from each of the observed values (y) collected from the study samples to its predicted value ($\hat{y}$) on the regression line. In order to find the best straight line that will represent the relationship between the two variables, the equation should be the one that gives the least "errors" of prediction.

There are many ways to minimize the error of your guess (prediction), but the "least squares" method optimizes by minimizing squared error. According to Pearson's approach, for linear regression if the slope is calculated from the least square method, then the observed x values predict the observed y values with the minimum possible sum of squared errors of prediction, $\sum(y - \hat{y})^2$[5]. The slope created from the



X	Y
0	5
1	8
2	11
3	14
4	?

X	Y
0	5
0	7
0	3
1	8
1	4
1	7
2	11
2	12
2	
3	14
3	19
3	10
4	17
4	20
4	15

Figure 1. Simple regression concept

least (minimum) errors is then considered as the best regression line with the best estimates of β_0 and β_1. This method is quite popular because it was comparatively easy to compute even manually to get the best guess for minimizing the squared error with the major assumption that the error is normally distributed[8].

Basics of Correlation

The term correlation composes of 'Co' (together) and relation (connection) between two quantities[2]. "Correlation coefficient", denoted by r. is measured on a scale that varies between +1 and -1. Complete correlation between two variables is expressed by either positive direction (+1) or negative direction (-1). Positive relationship occurs when one variable increases as the other increases; while negative relationship occurs when one decreases as the other increases. When there is no connection between the two variables, the correlation is 0[2,4].

The Pearson's correlation is calculated using statistics variance and covariance. Variance refers to the spread of data points around its mean, while a covariance refers to the measure of the directional relationship between two random variables[11,12]. Variance is the average of the squared deviations from the expected value (mean) for a single variable (x): $\sigma_x^2 = \sum(x-\bar{x})^2/n = [\sum(x-\bar{x})(x-\bar{x})]/n$. The larger the variance means the data scatter widely and at large distance from the mean[11]. A covariance refers to the measure of how two random variables (x and y) will change when they are compared to each other. In other words, covariance is an average measure of the deviations from both means $\sigma_{xy} = \sum(x-\bar{x})(y-\bar{y})/n$.

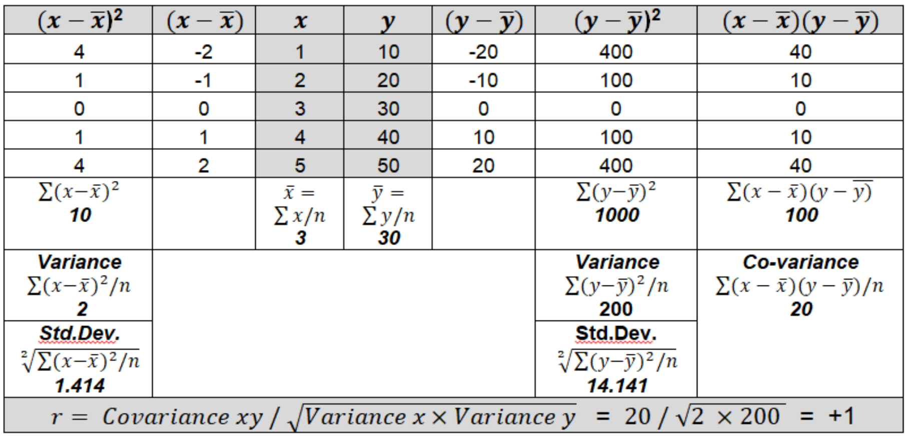
A positive covariance means the two variables move upward or downward in the same direction at the same time, while a negative covariance means the values of the two variables move in opposite direction from each other. Note that covariance is the measure that indicates the direction, but not the degree of the movements of two variables[11].

Correlation coefficient is the comparison of covariance with the variances of the two variables. That is, $r = Covariance\ xy\ /\ \sqrt{Variance\ x \times Variance\ y}$.

Figure 2 illustrates how correlations are calculated using this formula.

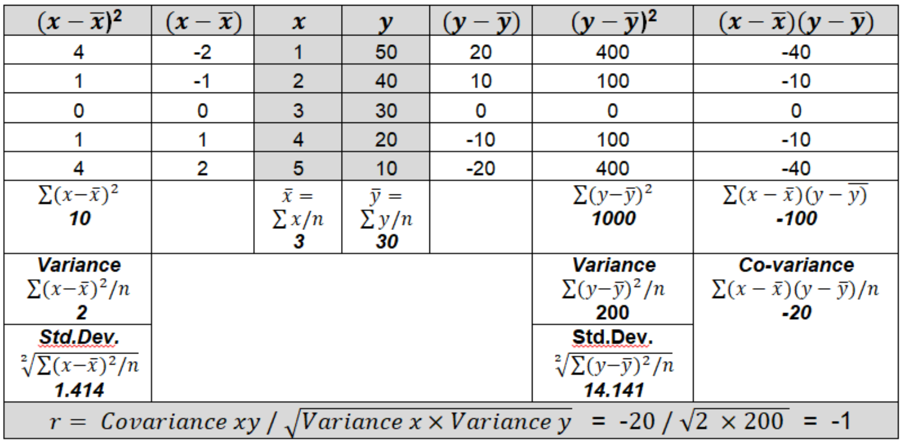
Major assumptions of the Pearson correlation coefficient are: (1) both variables are normally distributed; (2) the sample is randomly selected; (3)

(a) Perfect Positive Correlation



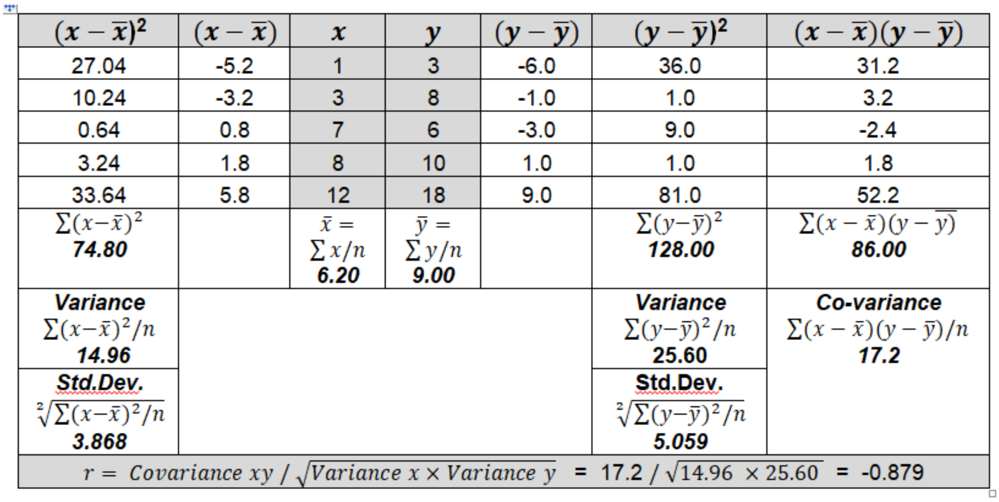
$(x-\bar{x})^2$	$(x-\bar{x})$	x	y	$(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
4	-2	1	10	-20	400	40
1	-1	2	20	-10	100	10
0	0	3	30	0	0	0
1	1	4	40	10	100	10
4	2	5	50	20	400	40
$\sum(x-\bar{x})^2$ *10*		$\bar{x} =$ $\sum x/n$ *3*	$\bar{y} =$ $\sum y/n$ *30*		$\sum(y-\bar{y})^2$ *1000*	$\sum(x-\bar{x})(y-\bar{y})$ *100*
Variance $\sum(x-\bar{x})^2/n$ *2*					**Variance** $\sum(y-\bar{y})^2/n$ *200*	**Co-variance** $\sum(x-\bar{x})(y-\bar{y})/n$ *20*
Std.Dev. $\sqrt[2]{\sum(x-\bar{x})^2/n}$ *1.414*					**Std.Dev.** $\sqrt[2]{\sum(y-\bar{y})^2/n}$ *14.141*	
$r = Covariance\ xy\ /\ \sqrt{Variance\ x \times Variance\ y}$ = 20 / $\sqrt{2 \times 200}$ = +1						

(b) Perfect Negative Correlation



$(x-\bar{x})^2$	$(x-\bar{x})$	x	y	$(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
4	-2	1	50	20	400	-40
1	-1	2	40	10	100	-10
0	0	3	30	0	0	0
1	1	4	20	-10	100	-10
4	2	5	10	-20	400	-40
$\sum(x-\bar{x})^2$ *10*		$\bar{x} =$ $\sum x/n$ *3*	$\bar{y} =$ $\sum y/n$ *30*		$\sum(y-\bar{y})^2$ *1000*	$\sum(x-\bar{x})(y-\bar{y})$ *-100*
Variance $\sum(x-\bar{x})^2/n$ *2*					**Variance** $\sum(y-\bar{y})^2/n$ *200*	**Co-variance** $\sum(x-\bar{x})(y-\bar{y})/n$ *-20*
Std.Dev. $\sqrt[2]{\sum(x-\bar{x})^2/n}$ *1.414*					**Std.Dev.** $\sqrt[2]{\sum(y-\bar{y})^2/n}$ *14.141*	
$r = Covariance\ xy\ /\ \sqrt{Variance\ x \times Variance\ y}$ = -20 / $\sqrt{2 \times 200}$ = -1						

(c) Positive Correlation



$(x-\bar{x})^2$	$(x-\bar{x})$	x	y	$(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
27.04	-5.2	1	3	-6.0	36.0	31.2
10.24	-3.2	3	8	-1.0	1.0	3.2
0.64	0.8	7	6	-3.0	9.0	-2.4
3.24	1.8	8	10	1.0	1.0	1.8
33.64	5.8	12	18	9.0	81.0	52.2
$\sum(x-\bar{x})^2$ ***74.80***		$\bar{x} =$ $\sum x/n$ ***6.20***	$\bar{y} =$ $\sum y/n$ ***9.00***		$\sum(y-\bar{y})^2$ ***128.00***	$\sum(x-\bar{x})(y-\overline{y})$ ***86.00***
Variance $\sum(x-\bar{x})^2/n$ ***14.96***					***Variance*** $\sum(y-\bar{y})^2/n$ **25.60**	**Co-variance** $\sum(x-\bar{x})(y-\bar{y})/n$ ***17.2***
Std.Dev. $\sqrt[2]{\sum(x-\bar{x})^2/n}$ ***3.868***					**Std.Dev.** $\sqrt[2]{\sum(y-\bar{y})^2/n}$ ***5.059***	
$r = Covariance\ xy\ /\ \sqrt{Variance\ x \times Variance\ y}$ = 17.2 / $\sqrt{14.96 \times 25.60}$ = -0.879						

Figure 2. Calculation of Pearson Product Moment Correlations

each pair of the observations are independent of one another; (4) two variables is linearly related[4,12]. Note that other types of correlation (e.g., Spearman correlation, Biserial correlation) slightly relax some of these assumptions. Moreover, you should avoid common misconception stating that correlation implies causation. Actually correlation does not imply causation but it could be a pre-condition, but not necessary, for measuring causation[12].

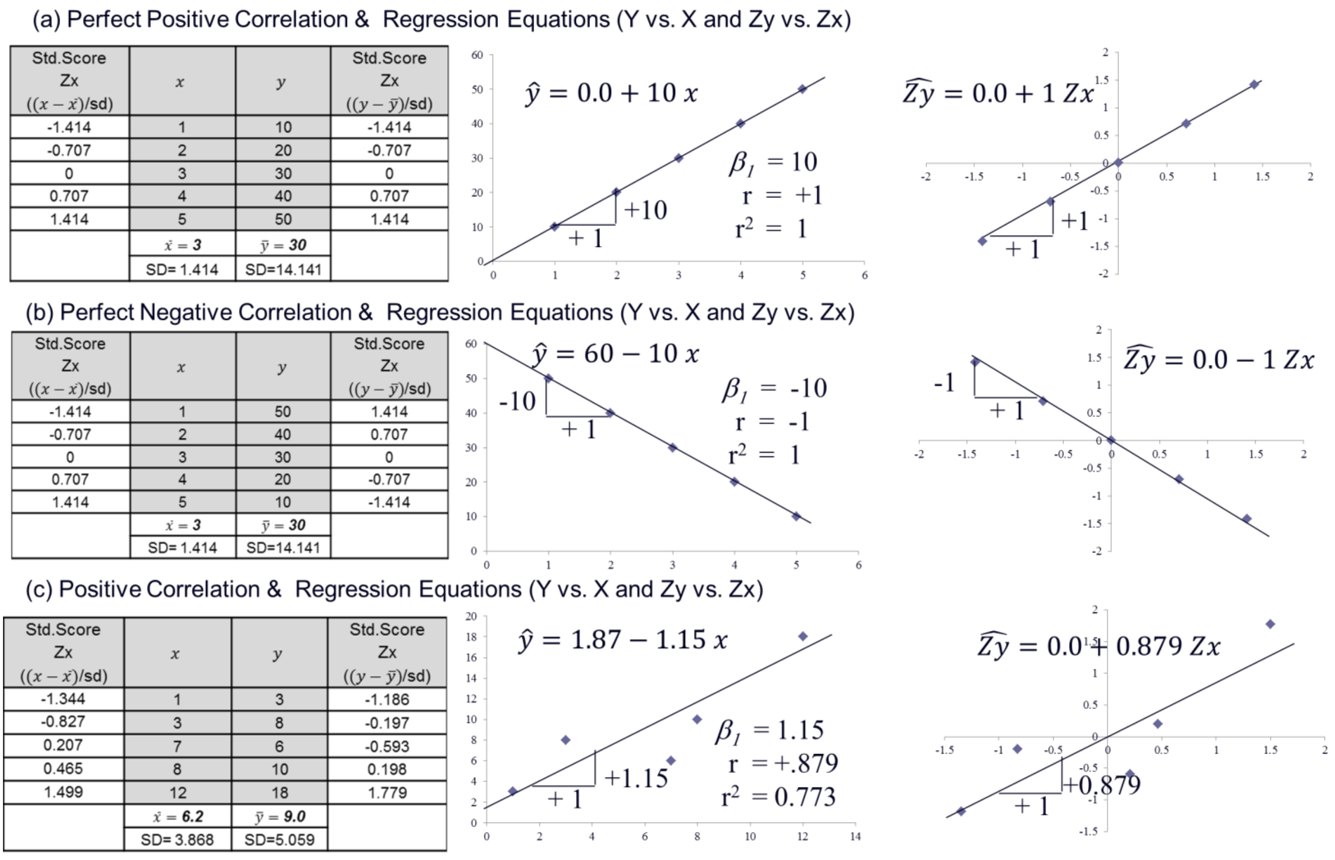
Mathematical Link between Regression and Correlation

Using the same data to calculate correlation and linear regression, you will get different statistics. Based on regression equation, you will get β_1 which is the slope indicating association between x and y (i.e., when x changes one unit of x, y will change β_1 unit of y). But correlation will give you the r which is the degree of linear association; r is simply a coefficient without unit attached. However, if you convert x and y to standard scores (Z-scores) and regress Zy from Zx, then β_1 calculated from the least square method will be the same value as correlation r while β_0 becomes 0. As shown in figure 3, β_1 the slope of Zx is equal to the correlation coefficient (r).

As for a note about Z-score, the Z-score is the number of standard deviations from the mean where a data point is located [i.e., Z-score = $(x-\bar{x})/\ sd$]. It is a measure of how many standard deviations below or above the population mean; it ranges from +3 standard deviations on the normal distribution curve[13]. Z-score is a useful way to compare observed data collected from a "normal" population. With raw score data, the values and units sometimes may not be informative. For example, your exam score of 80 from a total mark of 100 might sound good but where your position is when comparing your score to the average of the class is unknown; calculating your Z-score [(your score of 80 – average score of the class) / SD of the class] can then tell you where you are, compared to the rest of the class. That is, the Z score tells you how many standard deviations from the mean to where your score is[13]. Z-score has no unit attached, so does correlation r.

When you perform linear regression analysis, the model will quantify its goodness of fit with the coefficient of determination (r^2) which will be the same number as the square of correlation coefficient (r). In fact, the concept behind r^2 in linear regression is not quite the same as r from correlation analysis but interpretation of r^2 is useful to consider, in both regression and correlation context. The r^2 is a proportion (unlike r) as it is in effect measuring the proportion of explained/predicted variation compared to the total variation[3]. When all the observed variation is accounted for by the predicted portion (the line of best fit which is equivalent to perfect correlation), the r^2 is 1[12].

Many statistical models are based on correlation coefficient (r) among variables collected in the study. The interpretation of the r in those statistical models may not always be quantified as r^2, the coefficient of determination for the goodness of fit of the model. However, it is possible to provide a readily understandable interpretation by using the square of the correlation as the determinant of best fit of regression model. A correlation of 0.5 could mean that only 25% of the variability is accounted for by the correlation model[12].



(a) Perfect Positive Correlation & Regression Equations (Y vs. X and Zy vs. Zx)

Std.Score Zx $((x-\bar{x})/sd)$	x	y	Std.Score Zx $((y-\bar{y})/sd)$
-1.414	1	10	-1.414
-0.707	2	20	-0.707
0	3	30	0
0.707	4	40	0.707
1.414	5	50	1.414
	$\bar{x} = 3$	$\bar{y} = 30$	
	SD= 1.414	SD=14.141	

(b) Perfect Negative Correlation & Regression Equations (Y vs. X and Zy vs. Zx)

Std.Score Zx $((x-\bar{x})/sd)$	x	y	Std.Score Zx $((y-\bar{y})/sd)$
-1.414	1	50	1.414
-0.707	2	40	0.707
0	3	30	0
0.707	4	20	-0.707
1.414	5	10	-1.414
	$\bar{x} = 3$	$\bar{y} = 30$	
	SD= 1.414	SD=14.141	

(c) Positive Correlation & Regression Equations (Y vs. X and Zy vs. Zx)

Std.Score Zx $((x-\bar{x})/sd)$	x	y	Std.Score Zx $((y-\bar{y})/sd)$
-1.344	1	3	-1.186
-0.827	3	8	-0.197
0.207	7	6	-0.593
0.465	8	10	0.198
1.499	12	18	1.779
	$\bar{x} = 6.2$	$\bar{y} = 9.0$	
	SD= 3.868	SD=5.059	

Figure 3. Regression and correlation models

Final Thought

It is not overstated to say that both correlation and regression are the engines of statistical models. Several models (e.g., factor analysis, structural equation model, generalized linear models, etc.) are based on these two engines[14-15]. A historian of statistics, Stephen M. Stigler (1941-) calls them the "automobile" of statistical analysis, though he also stated that "... despite its limitations, occasional accidents, and incidental pollution, it and its numerous variations, extensions, and related conveyances carry the bulk of statistical analyses, and are known and valued by nearly all"[8].

Knowing your engines, now you are up to speed on your journey wisely.

Suggested Citation

Kaewkungwal J. Grammar of science: engines of statistical models. OSIR. 2019 Mar;12(1):32-7.

References

1. Wikipedia. Engine. 2019 Feb 22 [cited 2019 Mar 2]. <https://en.wikipedia.org/wiki/Engine>.

2. Surbhi S. Difference between correlation and regression. 2016 May 3 [cited 2019 Mar 2]. <https://keydifferences.com/difference-between-correlation-and-regression.html>.

3. Graphpad. What is the difference between correlation and linear regression? 2009 Jan 1 [cited 2019 Mar 2]. <https://www.graphpad.com/support/faq/what-is-the-difference-between-correlation-and-linear-regression/>.

4. The BMJ. 11. Correlation and regression [cited 2019 Mar 2]. <https:/www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression>.

1. Stanton JM. Galton, Pearson, and the Peas: a brief history of linear regression for statistics instructors. Journal of Statistics Education. 2001;9:3.

2. Gillard JW. An historical overview of linear regression with errors in both variables. 2006 October [cited 2019 Mar 2]. <http://mathsdemo.cf.ac.uk/maths/resources/Gillard_Tech_Report.pdf>.

5. Denis DJ. The origins of correlation and regression: Francis Galton or Auguste Bravais and the error theorists? Proceedings of the 61st Annual Convention of the

Canadian Psychological Association; 2000 Jun 29; Ottawa, Canada [cited 2019 Mar 2]. <https://www.york.ac.uk/depts/maths/histstat/bravais.htm>.

6. Kopf D. The discovery of statistical regression. 2015 Nov 6 [cited 2019 Mar 2]. <https://priceonomics.com/the-discovery-of-statistical-regression/>.

7. Nau R. Introduction to linear regression analysis [cited 2019 Mar 2]. <http://people.duke.edu/~rnau/regintro.htm>.

8. Wikipedia. Moment (mathematics). 2019 Feb 7 [cited 2019 Mar 2]. <https://en.wikipedia.org/wiki/Moment_(mathematics)>.

9. Hall M. Variance vs. Covariance: What's the Difference? 2019 Feb 20 [cited 2019 Mar 2]. <https://www.investopedia.com/ask/answers/041515/what-difference-between-variance-and-covariance.asp>.

10. Beaumont R. An introduction to statistics correlation. 2012 Sep 19 [cited 2019 Mar 2]. <http://www.robin-beaumont.co.uk/virtualclassroom/contents.html>.

11. Theme Horse. Z-Score: definition, formula and calculation - statistics how to [cited 2019 Mar 2]. <https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/>.

12. Schumacker RE, Lomax RG. A beginner’s guide to structural equation modeling. 3rd ed. New York: Routledge Taylor & Francis Group; 2010.

13. McCullagh P, Nelder JA. Generalized linear models. 2nd ed. London: Taylor & Francis Group; 1989.