# Teaching the Topic of Linear Regression

Koh Liang Cheung Daniel

**Abstract**

Correlation and Regression is a topic in the Statistics syllabus within Advanced "A" level H2 Mathematics. In this article, approaches to teaching various aspects of the topic of Linear Regression is introduced and discussed.

## 1 Introduction and Background

Correlation and Regression is a topic in the 2006 revised Advanced "A" Level H2 Mathematics curriculum. With the introduction of the Graphics Calculator (GC), calculation of the least squares estimates has become automated, de-emphasizing the formula describing these estimates. In addition, the emphasis of the syllabus is on having the student correctly use the graphics calculator to draw a scatter plot and obtain the line of best fit, in order to obtain a model to describe the relationship between the 2 variables of interest. Within the context of the question given students are also typically asked to make a prediction and to comment on the reliability of the value obtained. As a result, the teaching and learning of this topic is sometimes reduced to an exercise in GC manipulations.

Many good students in Mathematics comment that the computations performed by the graphics calculator appear rather mysterious. The formula for the least squares coefficients and the product moment correlation coefficients also seem complex and unintuitive. Even the terminology "regression" and "correlation" is unusual, as is the need to describe the least squares solutions as "estimates". In this article, some ideas are presented to address these concerns. These ideas were implemented with students in the 2009/2010 and 2011/2012 Raffles Academy Mathematics Program. We hope that it will be useful for Junior College (JC) students and teachers.

The article is organized as follows. In section 2, we motivate the discussion of the regression line as the least-square line which is obtained as the best approximate solution to the problem of solving a system of equations where the number of linear equations exceeds the number of unknowns. In section 3, using simple ideas about lines and planes taught to JC students we present the geometry of the least squares solutions. In section 4, we proceed more generally by appealing to students who have familiarity with Linear Algebra and the idea of orthogonal projections. In section 5, we finally take a more Statistical perspective, extending notions taught in Secondary School about Descriptive Statistics used to summarize data in one variable to the two variable case.

## 2 Method of Least Squares

**Systems of Equations (number of equations equal number of unknowns)**

In Secondary School, students are taught that given any pair of points $P(x_1, y_1)$ and $Q(x_2, y_2)$ with $x_1 \neq x_2$, there is a unique line passing through these points. The equation of the line can be easily obtained by solving the following equation for $y$ in terms of $x$,

$$\frac{y - y_1}{x - x_1} = \frac{y_2 - y_1}{x_2 - x_1}.$$

Equivalently, we can produce and solve a system of 2 linear equations in terms of $m$ and $b$.

$$y_1 = mx_1 + b$$
$$y_2 = mx_2 + b$$

The graphics calculator (simultaneous equation solver) can often help us in this process, especially when the values are not nice whole numbers but given in decimal expansion form.

In a similar way, if our task is to find a parabola of the form

$$y = a + bx + cx^2$$

that passes through 3 points, we can form a system of 3 linear equations and solve uniquely. Likewise, we can fit a polynomial of degree $n$ through $n$ given points.

We emphasize here that we can find a ***unique*** curve passing through the points when the number of data points matches the number of unknown parameters.

**Systems of Equations (number of equations *exceeds* number of unknowns)**

So the question we can ask now is, "What happens if we have more data points than we have unknown parameters?" Well, t*his is precisely the kind of situation the Method of Least Squares* can help to address*!!!* Consider the following question,

> Can we find a *line* passing through the points $(1,2),(2,6)$ and $(3,4)$ ?

In other words we want to solve the following system of linear equations

$$2 = a + b$$
$$6 = a + 2b$$
$$4 = a + 3b.$$

It should be clear that the above system of equations does not have a solution.

So the next best question we can ask is,

> Can we find a line *closest* to the points $(1,2),(2,6)$ and $(3,4)$ ?

Our goal now is to find solutions so that the LHS in the above system is close to its RHS.

That is, can we have

$$2 \approx a + b$$
$$6 \approx a + 2b$$
$$4 \approx a + 3b?$$

To answer this question requires us to establish a criterion for closeness (for individual points and the entire collection of points) and the natural choice, perhaps, is to take the *absolute difference* and minimize the *sum* of these deviations.
So our criteria becomes, find values of $a$ and $b$, such that the total error,

$$|2 - a - b| + |6 - a - 2b| + |4 - a - 3b| ,$$

is as small as possible.

Even in this simple scenario, there is no nice analytic approach to use, although it is a relatively simple matter for the computer. So for analytic tractability we consider minimizing the sum of squared errors or deviations,

$$(2 - a - b)^2 + (6 - a - 2b)^2 + (4 - a - 3b)^2 .$$

But we still have to resort to the multi-variable analogue of differentiating and setting to zero. However, this is an easy enough recipe to follow and may motivate interested students to read a bit more on multi-variable Calculus.

For now we will just follow the recipe, and leave the rationale to the reader to find out herself. So differentiating with respect to $a$ and $b$, respectively,

$$2(2 - a - b)(-1) + 2(6 - a - 2b)(-1) + 2(4 - a - 3b)(-1) = 0$$
$$2(2 - a - b)(-1) + 2(6 - a - 2b)(-2) + 2(4 - a - 3b)(-3) = 0$$

And we can obtain

$$3a + 6b = 12$$
$$6a + 14b = 26$$

This gives the values $a = 2$ and $b = 1$, and we obtain our solutions

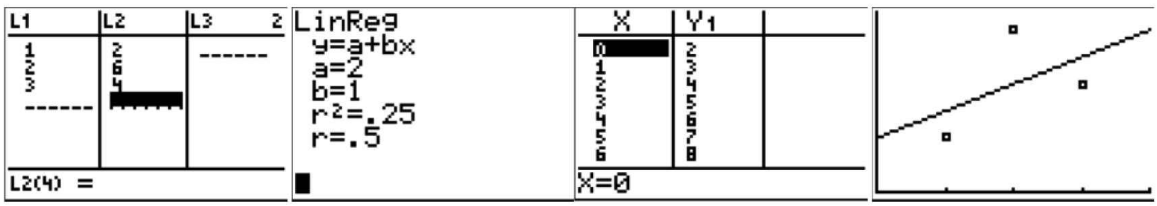
$$a + b = 2 + 1 = 3$$
$$a + 2b = 2 + 2 = 4$$
$$a + 3b = 2 + 3 = 5$$

So we have (1,3), (2,4) and (3,5) as the points closest in a difference of least squares sense to the corresponding points,

$$(1,2),(2,6) \text{ and } (3,4) .$$

By connecting the dots we interpolate and by extending the line beyond the data range we extrapolate and so we see that interpolation and extrapolation are actually ideas from a curve fitting (graphing) standpoint rather than viewed as Statistical methodology.

The graphics calculator is a useful tool to automate the entire process (Fig 1) similar to the way we use Polysimult2 to avoid solving a system of linear equations by hand.




Given points | Least squares solution | Fitted values | Graph Interpolation Extrapolation

Screen Shots from TI-84 Plus Family of Graphing Calculators

**Fig 1**

## 3 Geometry of the Least Squares Solution

The goal in this section will be to provide another perspective (a geometric one) to the ***reasonableness*** of using the method of least squares as a curve fitting procedure to obtain estimates rather than minimizing absolute deviations or other methods. In Section 2, we also did not justify that our algorithm producing the stationary point indeed yielded a unique solution that is a minimum.

We re-present our problem of finding the best approximate solution. In vector form, we ask whether we can find $a$ and $b$ such that,

$$\begin{pmatrix}2\\6\\4\end{pmatrix} \approx \begin{pmatrix}a+b\\a+2b\\a+3b\end{pmatrix} = a\begin{pmatrix}1\\1\\1\end{pmatrix} + b\begin{pmatrix}1\\2\\3\end{pmatrix}?$$

We will now utilize ideas about lines and planes within the H2 Mathematics Syllabus to tackle this problem.

It can easily be checked that $y = \begin{pmatrix}2\\6\\4\end{pmatrix}$ does not lie on the plane $\pi : \mathbf{r} = a\begin{pmatrix}1\\1\\1\end{pmatrix} + b\begin{pmatrix}1\\2\\3\end{pmatrix}, \quad a,b \in \mathbb{R}$ .

The key idea is the observation that the least squares solution, expressed as a vector $\hat{y} = \begin{pmatrix}3\\4\\5\end{pmatrix}$

is the ***projection*** of $y = \begin{pmatrix}2\\6\\4\end{pmatrix}$ onto the plane $\pi : \mathbf{r} = a\begin{pmatrix}1\\1\\1\end{pmatrix} + b\begin{pmatrix}1\\2\\3\end{pmatrix}, \quad a,b \in \mathbb{R}$ .

Phrased in a form familiar to H2 Mathematics students, we have the following example:

***Example*** The points $I$ and $X$ have position vectors $1=\begin{pmatrix}1\\1\\1\end{pmatrix}$ and $x=\begin{pmatrix}1\\2\\3\end{pmatrix}$ respectively, with respect to the origin $O$.

(i) Find the projection of $y=\begin{pmatrix}2\\6\\4\end{pmatrix}$ onto the plane $\pi:\mathbf{r}=a\begin{pmatrix}1\\1\\1\end{pmatrix}+b\begin{pmatrix}1\\2\\3\end{pmatrix}$, $a,b\in\mathbb{R}$;

(ii) Determine the shortest distance of the point $Y(2,6,4)$ to $\pi$.

**Solution**

An approach that a student taking H2 Mathematics would take is to find the position vector of the foot of the Perpendicular from $Y(2,6,4)$ to $\pi$ by intersecting the line passing through $Y$ in the direction of the normal of the plane $\pi$, with the plane $\pi$ itself. In this manner, we obtain the projection vector, $\hat{y}=\begin{pmatrix}3\\4\\5\end{pmatrix}$ and we can easily deduce the value of $a$ and $b$ from

$$\begin{pmatrix}3\\4\\5\end{pmatrix}=a\begin{pmatrix}1\\1\\1\end{pmatrix}+b\begin{pmatrix}1\\2\\3\end{pmatrix},\quad a,b\in\mathbb{R}.$$

We have

$$a=2,\ b=1.$$

The error or residual is now measured as the shortest distance from $Y$ to $\pi$ given by

$$\left|y-\hat{y}\right|=\sqrt{(y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2+(y_3-\hat{y}_3)^2}=\sqrt{6}\text{ units}.$$

To conclude, we see that $a=2,\ b=1$ *minimizes* $\left|Y-a1+bX\right|$ or equivalently minimizes $\left|Y-a1+bX\right|^2=\sum_i(y_i-a+bx_i)^2$ over $a,b\in\mathbb{R}$ which justifies the solution obtained in Section 2 is indeed a minimum.

## 4 Least Squares Estimates and Correlation Coefficient

In this section, we proceed in a way that generalizes easily to the situation when we have $n$ equations and $k$ unknowns ($n > k$). We also gain further insight into the formulae for the least squares estimates and the product moment correlation coefficient. The correlation coefficient provides a quantitative measure of the strength of linear relationship. This section was discussed with students who had taken Linear Algebra as a H3 subject but should hopefully still be accessible for good H2 Mathematics students.

In H2 Mathematics students learn about the projection of $x$ onto $a$. Questions focus on finding the length of projection, $\left|x\cdot\hat{a}\right|$. Here, we emphasize the *projection vector*.

If $a$ and $x$ are two non-zero non-parallel vectors, and the vector $b$ is given by $b = x - (x \cdot \hat{a})\hat{a}$ where $\hat{a}$ is the unit vector along $a$, then $a$ and $b$ are perpendicular.

**Solution**

It suffices to show that $a \cdot b = 0$.

$$\begin{aligned} a \cdot b &= a \cdot \left(x - (x \cdot \hat{a})\hat{a}\right) \\ &= a \cdot x - (x \cdot \hat{a})|a| \qquad \hat{a} = \frac{a}{|a|} \\ &= 0 \end{aligned}$$

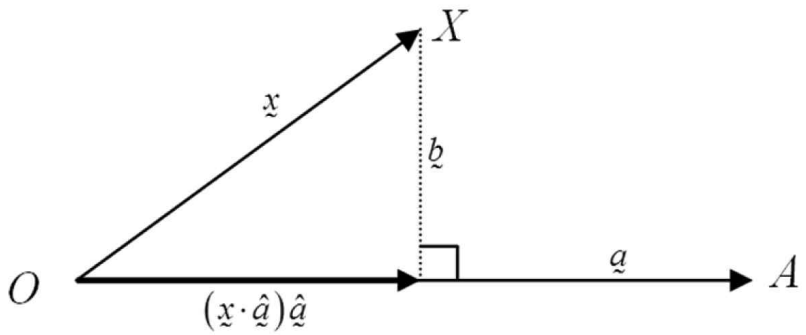



**Fig 2**

The vector $(x \cdot \hat{a})\hat{a}$ is the projection of $x$ onto $a$ (Fig 2). The vector $b$ is perpendicular to $a$ and its length is the minimum (perpendicular) distance of the point $X$ from the line through the origin in the direction of $a$.

We now generalize the previous discussion. First, we need a definition to generalize the idea of non-parallel vectors for an arbitrary collection of vectors.

Vectors $v_1, v_2, ..., v_n$ are said to be *linearly independent* if, for $a_i \in \mathbb{R}, i = 1,2,...,n$,

$$a_1 v_1 + a_2 v_2 + ... + a_n v_n = 0 \Rightarrow a_i = 0 \text{ for } i = 1,2,...,n.$$

**Lemma**

If $a$, $x$ and $y$ are three linearly independent vectors, and the vectors $b$ and $c$ are given by $b = x - (x \cdot \hat{a})\hat{a}$ and $c = y - (y \cdot \hat{a})\hat{a} - (y \cdot \hat{b})\hat{b}$, where $\hat{a}$ and $\hat{b}$ are unit vectors along $a$ and $b$ respectively, then $a$, $b$ and $c$ are mutually perpendicular.

**Solution**

$a$ and $b$ are mutually perpendicular as before.

We now show that $a \cdot c = 0$. ($b \cdot c = 0$ is shown similarly.)

$$\begin{aligned} a \cdot c &= a \cdot \left(y - (y \cdot \hat{a})\hat{a} - (y \cdot \hat{b})\hat{b}\right) \\ &= a \cdot y - y \cdot a - 0 \qquad a = |a|\hat{a} \text{ and } a \cdot b = 0 \\ &= 0 \end{aligned}$$

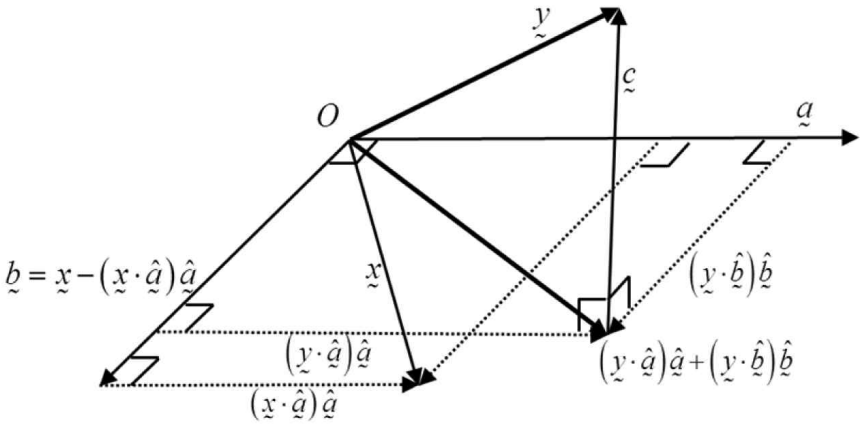



**Fig 3**

Analogously, we see that the vector $(y\cdot\hat{a})\hat{a}+(y\cdot\hat{b})\hat{b}$ is obtained by the projection of $y$ onto the space spanned by $a$ and $b$. The vector $c$ is perpendicular to $a$ and $b$ and its length is the minimum (perpendicular) distance of the point $Y$ to the plane through the origin spanned by the vectors $a$ and $b$.

**Application to Least Squares Line**

First some notation:

$$a = 1 = \begin{pmatrix}1\\1\\\\1\end{pmatrix}, x = \begin{pmatrix}x_1\\x_2\\\\x_n\end{pmatrix}, y = \begin{pmatrix}y_1\\y_2\\\\y_n\end{pmatrix}, \hat{y} = \begin{pmatrix}\hat{y}_1\\\hat{y}_2\\\\\hat{y}_n\end{pmatrix}, \overline{x} = \frac{x_1 + x_2 + \quad + x_n}{n} \text{ and } \overline{y} = \frac{y_1 + y_2 + \quad + y_n}{n}.$$

We want to form the linear space "spanned" by the column vectors $1$ and $x$, that is, all linear combinations of the form $a1 + bx$ where $a,b \in \mathbb{R}$.

The column space of the matrix $\begin{bmatrix}1 & x\end{bmatrix}$, is a plane $\pi : \mathbf{r} = a1 + bx, \quad a,b \in \mathbb{R}$.

We are given that $y$ does not lie on the column space so we want to find the orthogonal projection instead.

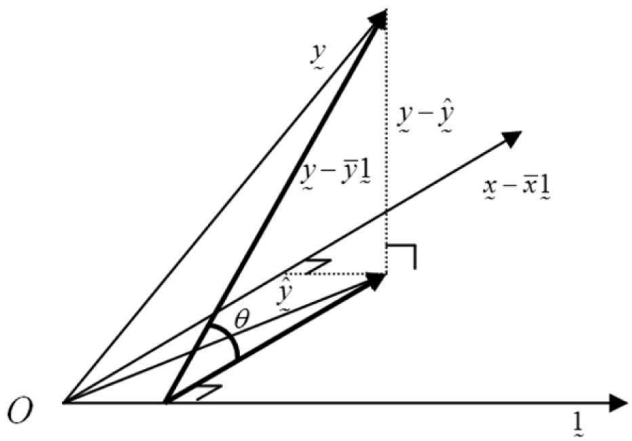



**Fig 4**

We can now compute.

$$\hat{a} = \frac{1}{\sqrt{n}}1 \; ; \; \left(x \cdot \hat{a}\right)\hat{a} = \bar{x}1 ; \; \left(y \cdot \hat{a}\right)\hat{a} = \bar{y}1$$

$$b = x - \left(x \cdot \hat{a}\right)\hat{a} = x - \bar{x}1 ; \; \hat{b} = \frac{1}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}\left(x - \bar{x}1\right)$$

The projection vector $\left(y \cdot \hat{a}\right)\hat{a} + \left(y \cdot \hat{b}\right)\hat{b}$ is denoted by $\hat{y}$ and given as follows

$$\hat{y} = \bar{y}1 + \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\left(x - \bar{x}1\right) . \tag{4.1}$$

The components correspond to the fitted values,

$$\hat{y}_i = \bar{y} + \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\left(x_i - \bar{x}\right) . \tag{4.2}$$

Finally, we have the formula for the least squares solutions

$$y = a + bx \;\; \text{where } b = \frac{\sum_{i}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right)}{\sum_{i}\left(x_i - \bar{x}\right)^2}, a = \bar{y} - b\bar{x} . \tag{4.3}$$

We can define the angle $\theta$ between $y - \bar{y}1$ and $x - \bar{x}1$ (Fig 4) or equivalently the ***product moment correlation coefficient*** which we denote by $r$ as follows

$$\begin{aligned} r &= \cos\theta \\ &= \frac{\left(y - \bar{y}1\right)\cdot\left(x - \bar{x}1\right)}{\left|y - \bar{y}1\right|\left|x - \bar{x}1\right|} \\ &= \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}} . \end{aligned} \tag{4.4}$$

We can measure the strength of a linear relationship by observing this value of $|r|$.

The following properties of $r$ follow immediately.

1. $|r| \le 1$ with equality when $y - \bar{y}1 = k\left(x - \bar{x}1\right)$ for some non-zero $k \in \mathbb{R}$ (4.5a)
2. $\left|y - \hat{y}\right| = \left|y - \bar{y}1\right|\sin\theta = \left|y - \bar{y}1\right|\sqrt{1 - r^2}$ (4.5b)
3. $\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \left(1 - r^2\right)\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$ (4.5c)

## 5 A Statistical Perspective

The previous sections may appeal to students with a stronger mathematical background. In this section we take a statistical approach illustrating Correlation and Regression as an exploratory data analysis tool useful for exploratory analysis of bivariate data, similar in spirit to the way mean and standard deviation is introduced in Secondary School for univariate data.

We recall Descriptive Statistics from Secondary School involving data in one variable. Given a set of data, we would be interested in *shape, centre, spread*. Students would be taught to draw a histogram and find summary Statistics such as the mean and variance.

### A model for errors

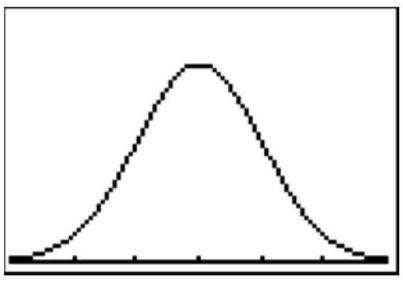
In H2 Mathematics, students learn about the normal distribution which is often used as a model for the error distribution. We say that $X$ is a ***normal random variable***, or $X$ is normally distributed with mean $\mu$ and variance $\sigma^2$, and denoted $X \sim \mathrm{N}(\mu,\sigma^2)$ if

$$P(a \le X \le b) = \int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx$$

The probability density function is given by the integrand,

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}.$$

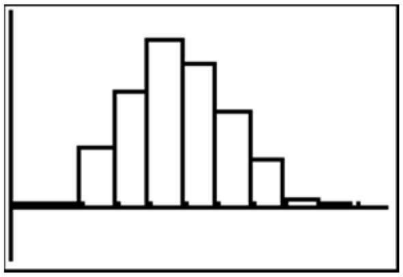
It has the familiar symmetric bell-shape (Fig 5).



Screen Shot from TI-84 Plus Family of Graphing Calculators
**Fig 5**

We can compute for the Normal $P(-2\sigma < X - \mu < 2\sigma) = 0.954$.

So if data comes from a normal distribution the shape of the histogram should not deviate too much from this mound shape and approximately 95% of observations will lie within 2 standard deviations of the mean (Fig 6).



Screen Shot from TI-84 Plus Family of Graphing Calculators
**Fig 6**

Now we will discuss data involving 2 variables and a natural consideration would be in the *relationship, direction* and *strength* between the variables of interest. As in the case with one variable data, we begin with a plot of the data, which is referred to as a Scatter Diagram or Scatter Plot.

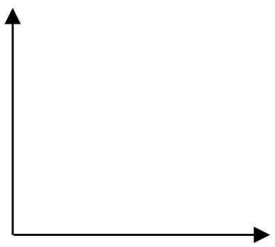
**Scatter Diagrams**

Suppose $X$ measures the weight of a sample of water and $Y$ measures the volume of the same sample of water. Clearly there is a strong relationship between $X$ and $Y$. If $(X,Y)$ pairs are measured on several different samples and the observed data pairs are plotted, the data points should fall on a straight line because of the physical relationship between $X$ and $Y$. They will not fit exactly for all pairs because of measurement errors, impurities in the water etc, but with careful laboratory technique the data points will fall very nearly on a straight line.

Now consider another experiment in which $X$ and $Y$ are the body weight and height of the same person. Clearly there is a relationship but it is not nearly as strong. The $(X,Y)$ points will cluster less tightly around a line. If most or all the points in a scatter diagram seem to lie near a straight line, we say there is a **linear relationship** between the variables.

Some examples are as follows and a qualitative description accompanies the plot of the data.

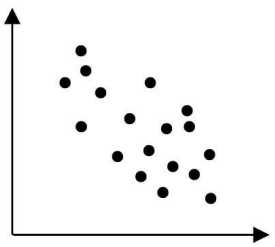
**Example 1**



Both the variables increase together. The points lie close to a straight line.

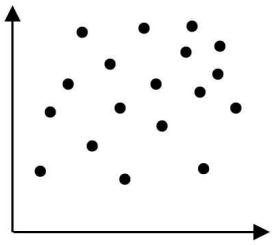
We say that the variables have a strong positive correlation.

**Example 2**



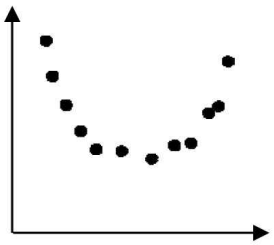
As one variable increase, the other decreases. There is a negative correlation between the variables.
The points are more disperse from a straight line as compared to the first case. So the correlation here is not as strong as the first case.

**Example 3**



There appears to be no relation between the variables.

**Example 4**



There is a non-linear relation between the variables.

**Summary Statistics**

A scatter diagram shows the relation between two variables in a diagram, which we can describe qualitatively. The graph appears as a cloud of points, which we now want to summarize numerically. We also want to quantify in some sense what we mean by "close to a line". The first step in our data analysis of the data of 2 variables, would be to identify and indicate the centre of the data set, $(\overline{x},\overline{y})$.

Next we can measure and indicate the variability in the horizontal and vertical directions about $\overline{x}$ and $\overline{y}$, using the standard deviations $\sigma_X = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n}}$ and $\sigma_Y = \sqrt{\dfrac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{n}}$.

We know that if the data is approximately "normal", then approximately 95% of $x_i$ values would lie within $\overline{x}\pm 2\sigma_X$ and 95% of $y_i$ values within $\overline{y}\pm 2\sigma_Y$. Now we need a summary statistics to relate the 2 variables of interest.

**Product Moment Correlation Coefficient**

We introduce a measure of the strength of the linear relation between the variables. First, convert each variable to standardized units $\left(\dfrac{x_i-\overline{x}}{\sigma_X},\dfrac{y_i-\overline{y}}{\sigma_Y}\right)$ then average the products to obtain,

$$\frac{\sum_{i=1}^{n}\left(\dfrac{x_i-\overline{x}}{\sigma_X}\right)\left(\dfrac{y_i-\overline{y}}{\sigma_Y}\right)}{n} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2\sum_{i=1}^{n}(x_i-\overline{x})^2}} \qquad (5.1)$$

which we recognize from (4.4) as the product moment correlation coefficient, *r*. We first examine the sign of this quantity. Later, we will discover in what sense *r* **quantifies the linear relation** or the amount of clustering (equivalently dispersion) around a line.

**Sign of *r* as a measure of association**

The product moment correlation coefficient is a measure of association in the following manner.
Begin by marking out the centre of the data set and form 4 quadrants relative to this point.
Consider where the data values are positioned on the scatter diagram relative to the quadrants.

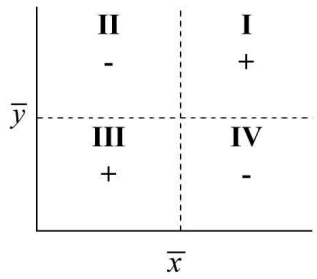
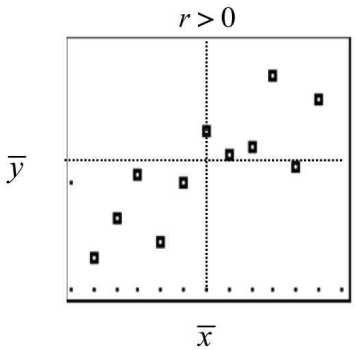



**Fig 7**

For each data pair $(x,y)$, if both $x$ and $y$ vary on the same side of their respective means (Quadrant I and III) then the product of their deviation is positive.




Screen Shot from TI-84 Plus Family of Graphing Calculators

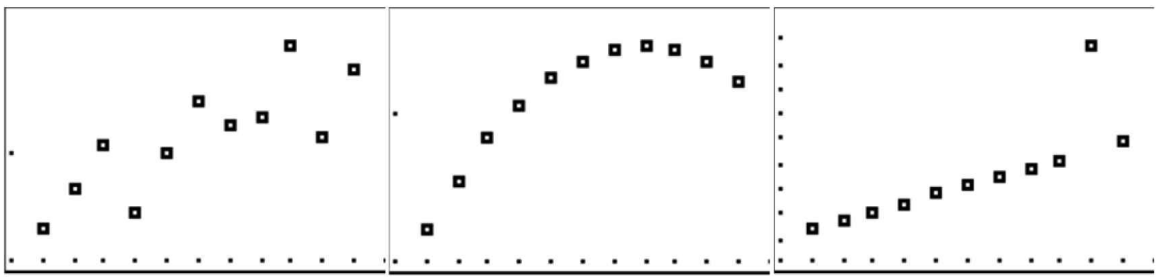
**Fig 8**

If most of the points lies in Quadrant I and III, then the average product of these deviations is also positive (Example 1) and $r > 0$. Conversely, if they tend to vary on opposite sides of their respective means at the same time (Quadrant II and IV), then the average product of their deviations is negative (Example 2) and $r < 0$. If the points are "evenly spread out" over the 4 quadrants then we expect that $r = 0$ (Example 3)

**Non-linear Relationships and Outliers**

We have to be careful when we try to interpret raw numerical summary statistics from a data set. The correlation coefficient is sensitive to extreme value and is also unable to detect non-linear relationships.
The scatter plots of the 3 data sets we see below can have the same basic 5 summary statistics. When we think of the product moment correlation as measure of strength of linearity we may have a mental picture of the 1st diagram. But without actually looking at the scatter plot we could be mistaken, and the true situation could turn out very different. So data analysis should

incorporate both the qualitative (graphical) and quantitative (numerical) summary statistics aspects.



Positive linear Relationship | Strong non-linear Relationship | Strong Positive linear Relationship with Influential point

Screen Shots from TI-84 Plus Family of Graphing Calculators

**Fig 9**

**Translation and Scale Invariance**

A careful examination of the formula for $r$ reveals that it is not affected by changes in the measurement units. Suppose there is a change in measurement units in both $x$ and $y$ variables,

$$u_i = ax_i + b, \quad a > 0 \; ; \qquad v_i = cy_i + d, \quad c > 0$$

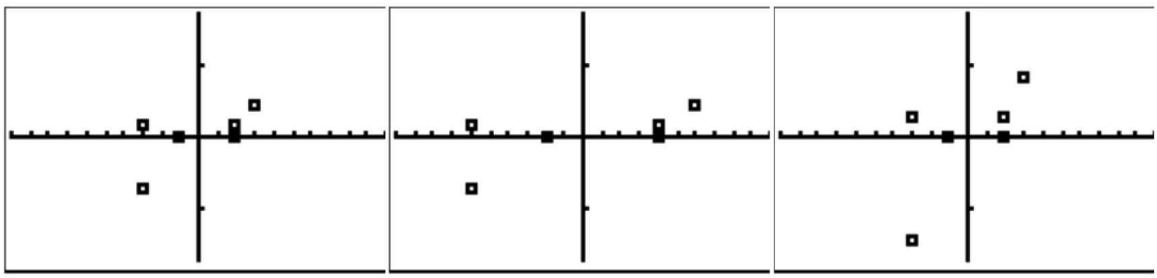
Observe that,

$$\frac{u_i - \overline{u}}{\sigma_U} = \frac{x_i - \overline{x}}{\sigma_X} \; ; \frac{v_i - \overline{v}}{\sigma_V} = \frac{y_i - \overline{y}}{\sigma_Y} \; .$$

Hence, the correlation coefficient between $u$ and $v$ is the same as that between $x$ and $y$.

This cautions us about the difficulty of ***visually*** assessing the quality of fit due to changes in scale (and adjusting window settings). The following scatter diagrams from 3 data sets (modified by scaling in *x and y* directions respectively) have the ***same*** correlation coefficient value.

In our classroom discussion, we asked students to make a "guess" as to which of the data sets they perceived had the strongest linear relationship and it was interesting to hear their reasons.



Original data | Scaled in horizontal direction | Scaled in vertical direction

Screen Shots from TI-84 Plus Family of Graphing Calculators

**Fig 10**

**Least Squares Line and Correlation Coefficient**

In the previous section, we obtained a derivation of the least-squared regression line to a set of data. We re-write (4.2) in a statistical manner incorporating the correlation coefficient $r$ to obtain,

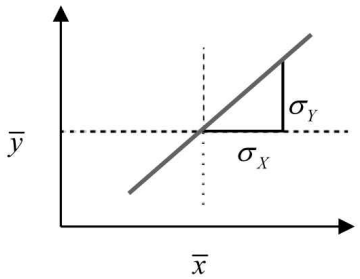
$$\hat{y} = \overline{y} + r\frac{\sigma_Y}{\sigma_X}(x - \overline{x}) \,. \tag{5.2}$$

Rearranging and "standardizing the variables", observe that,

$$\frac{\hat{y} - \overline{y}}{\sigma_Y} = r\frac{x - \overline{x}}{\sigma_X} \Leftrightarrow r = \left(\frac{\hat{y} - \overline{y}}{\sigma_Y}\right) \Big/ \left(\frac{x - \overline{x}}{\sigma_X}\right).$$

From this we see that the standardized variables would have least squares line passing through the origin with slope $r$.

Returning back to our regression line, if $r = 1$ we observe $\hat{y} = \overline{y} + \frac{\sigma_Y}{\sigma_X}(x - \overline{x})$ which is a line passing through $(\overline{x}, \overline{y})$ with slope $\frac{\sigma_Y}{\sigma_X}$.




On this line, if $x$ lies $k$ standard deviations from $\overline{x}$, then $y$ lies $k$ standard deviations from $\overline{y}$. Hence, if all the points are tightly clustered along a line, the equation of this line is,

$\hat{y} = \overline{y} + \frac{\sigma_Y}{\sigma_X}(x - \overline{x})$ if there is a positive relationship between the 2 variables, and

$\hat{y} = \overline{y} - \frac{\sigma_Y}{\sigma_X}(x - \overline{x})$ if there is a negative relationship between the 2 variables.

**Regression to the Mean**

In general, our data set will not line up so perfectly, so our best fit predicted line is

$$\hat{y} = \overline{y} + r\frac{\sigma_Y}{\sigma_X}(x - \overline{x}) \text{ where } |r| < 1 \,.$$

Interpreting this equation, if a data point $x$ lies $k$ standard deviations from $\overline{x}$, then our predicted value for $y$ is less than $k$ standard deviations from $\overline{y}$. Historically, Galton noted the phenomenon which is sometimes referred to as the "regression effect" or "regression to the

mean" when he observed that very tall parents had above average height but shorter offspring, and very short parents had below average height children that were generally taller than them.

A further note, we are able to use the value of $|r|$ to measure the prediction error. We now interpret the formula (4.5c). By dividing by $n$ we obtain

$$\frac{\sum_i (y_i - \hat{y}_i)^2}{n} = (1 - r^2)\sigma_Y^2 . \qquad (5.3)$$

If the data is obtained from a normal distribution, then approximately 95% of the data is scattered within $\pm 2\sqrt{(1-r^2)}\sigma_Y$ of the regression line.

## 6 Further Discussion

### Linear Model with Normal Errors

At this introductory juncture it may not be suitable for JC students to proceed to a formal statistical discussion of a linear model with normal errors and/or discuss important statistical notions like maximum likelihood. This approach requires a discussion of joint random variables and their distributions which takes us too far away from the main syllabus. Such an approach is necessary if we are to also define the correlation coefficient between 2 random variables, and to justify that the least squares solutions are *bona fide* statistical estimates. With a linear (statistical) model and a specification of the error form and distribution, the point estimators will have a distribution and hence issues of statistical inference (confidence intervals and hypothesis testing) can truly proceed.

### Role of Examples

The nature of Statistics is such that when data is produced they are not just simply numerical quantities but numbers which are collected from some context and for some purpose. In teaching statistical methodology, mathematics can be utilized for explanatory purposes. The author has benefitted greatly from books by Freedman (1991) and Moore (2006) which approach Statistics from a less technical perspective. The former has an extensive discussion about Correlation and Regression which we have only touched on briefly here. The latter approaches Statistics as a liberal arts discipline. Both are useful as they contain many interesting examples and discussion about the practice of Statistics as a methodological discipline which may be appealing to a broad range of students.

## 7 Conclusion

The topic of Correlation and Regression can be made more attractive and appealing to JC Mathematics students if the treatment of the topic takes greater advantage of students' quantitative ability, and by drawing links and connections with other topics within the H2 Mathematics syllabus.

## 8 Acknowledgement

The author would like to express his thanks to his colleagues Lim Hway Kiong and Mok Hoi Nam from the 2009/2010 and 2011/2012 Raffles Academy Mathematics programme, and the anonymous referee for their invaluable suggestions and comments.

**References**

[1] D. Freedman, R. Pisani, R. Purves, and A. Adhikari, *Statistics* (2nd ed.), New York: W. W. Norton and Company, 1991.

[2] R. V. Hogg, J. W. McKean, and A. T. Craig, *Introduction to Mathematical Statistics* (6th ed.), New Jersey: Pearson Prentice Hall, 2005.

[3] D. S. Moore, W. I. Notz, *Statistics: Concepts and Controversies* (6th ed.), New York: W. H. Freedman and Company, 2006.

The author is a teacher and Assistant Department Head, Mathematics at Raffles Institution (JC).