King Fahd University of Petroleum & Minerals

DEPARTMENT OF MATHEMATICAL SCIENCES

Technical Report Series

TR 407

June 2009

A Gentle Approach to Linear Regression

Anwar H. Joarder and M. H. Omar

A Gentle Approach to Linear Regression

Anwar H. Joarder and M. H. Omar
Department of Mathematics and Statistics
King Fahd University of Petroleum & Minerals
Dhahran 31261, Saudi Arabia
Emails: anwar@kfupm.edu.sa, omarmh@kfupm.edu.sa

The line of best fit is motivated through a gentle introduction based on coordinate geometry and basic statistics. Least squares sense is explained by simple examples. Several mnemonically plausible points are proposed to help draw the line of best fit.

1. Introduction

In simple linear regression, the intercept and slope parameters are estimated by minimizing the sum of squared errors for every sample points. The concept is not easy to explain to students who are not that motivated in differential calculus. In this paper we try to motivate students by capitalizing on their background of elementary coordinate geometry and basic statistics. Several mnemonically plausible points are proposed to help draw the line of best fit.

2. The Use of All Points in Determining Slope and Intercept

From elementary coordinate geometry, a straight line $y = \alpha + \beta x$ is determined by finding β, the slope and then α, the y-intercept. The slope is determined by any two points on the line. If we have n pairs of points say (x_i, y_i), $i = 1, 2, \cdots, n$, all of which are on a straight line, then we need any two points to determine the slope of the line. If those two points are (x_i, y_i) and (x_k, y_k), then the slope of the line is given by

$$\beta = \frac{y_i - y_k}{x_i - x_k}, \ (i = 1, 2, \cdots, n; \ k = 1, 2, \cdots, n; \ i \neq k). \tag{2.1}$$

Then one can determine the y-intercept and obtain the line. Indeed if all the points fall on a line, one can check that the line also passes through $(\bar{x}, \bar{y})$. Hence, one can write the slope formula as

$$\beta = \frac{y_i - \bar{y}}{x_i - \bar{x}}, \quad (i = 1, 2, \cdots, n). \tag{2.2}$$

The numerator and denominator of the right hand side of (2.2) show the variation in y-values and x-values respectively around their respective means. It thus makes sense to use the ratio of standard deviations of y-values and x-values respectively to determine slope or gradient. In what follows we denote

$$s_{xx} = \sum (x - \bar{x})^2 = (n-1)\, s_x^2, \quad s_{yy} = \sum (y - \bar{y})^2 = (n-1)\, s_y^2$$
$$\text{and} \quad s_{xy} = \sum (x - \bar{x})(y - \bar{y}) \tag{2.3}$$

where s_x^2, s_y^2 and $s_{xy}/(n-1)$ are the sample variance of x, sample variance of y and sample covariance between x and y. If the points fall on a line with positive slope, the gradient of the line is given by s_y / s_x, but if the points fall on a line with negative slope, the gradient of the line is given by $-s_y / s_x$. Thus a general formula for the slope of the line would be $r\, s_y / s_x$ where the value of r is 1 or -1 depending on positive or negative slope of the line. This formula accommodates all sample points though algebraically it is equivalent to the usual formula in (2.1) or (2.2). Thus a slope formula of a line where all the points are on the line can be written as:

$$b = r\, \frac{s_y}{s_x}, \quad r = -1,\ 1. \tag{2.4}$$

3. Determining the Slope and Intercept of the Line of Best Fit

If the scatter diagram (scattergram) of the two variables do not form a line ($y = \alpha + \beta x$) but shows a linear trend with the model $y = \alpha + \beta x + \varepsilon$, where ε is error in y to settle to $\alpha + \beta x$ corresponding to x. How we can use all the points in the determination of the slope of the line of best fit? Observe that in (2.1) or (2.2), we are using the 'differences' as a criterion to determine the slope. Indeed the difference $y_i - \bar{y}$ or $x_i - \bar{x}$ $(i = 1, 2, \cdots, n)$ measures the variation in y-values or x-values around their means $\bar{y}$ and $\bar{x}$. In this case, we try to estimate a line of best fit through the concentration of points. In analogy with the case of perfect linearity, one can argue that the line of best fit passes through the average point $(\bar{x}, \bar{y})$.

Analogous to (2.4), a formula that takes care of all points to find the gradient (slope) of the line of 'best' fit is given by

$$b = r\, \frac{s_y}{s_x} \tag{3.1}$$

where r is given by $-1 \le r \le 1$. The intermediate values of r shows the strength of linearity of the points. Note that the slope formula in (3.1) simplifies to (2.1) in case all the points fall on a line. Since the line or the line of best fit passes through the mean point

(see 2.2), it would be reasonable to see how the values of x and y vary around their means. Thus the issue is how to incorporate $(x_i - \bar{x})$ and $(y_i - \bar{y})$ in determining the strength (r) of linearity of the points. Since the slopes depend on their signs, a formula for r may be

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{3.2}$$

which may be weighted by their standard deviations to squeeze it between -1 and 1. Another reason would be to give importance to the concentration of the data points. Thus a refined formula for r is given by

$$\sum_{i=1}^{n}\frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} = \frac{s_{xy}/(n-1)}{s_x s_y} \tag{3.3}$$

or

$$r = \frac{1}{n-1}\sum z_x z_y, \qquad -1 \le r \le 1 \tag{3.4}$$

where $z_x = (x - \bar{x})/s_x$ and $z_y = (y - \bar{y})/s_y$ are the z-scores of the variables x and y respectively. Note that by (2.3), the formula in (3.4) can also be written as in the following popular form

$$r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}}. \tag{3.5}$$

Since the line of best fit $y = a + bx$ passes through the center of gravity $(\bar{x}, \bar{y})$ with b as in (3.1), the intercept is $a = \bar{y} - b\bar{x}$. But some readers may prefer to join two points to draw the line of best fit. In Section 4, we will explore some mnemonically plausible points any two of which can be joined to form the line of best fit.

4. The Regression Line Simplified

The line of best fit is also popularly known as linear regression. It is a statistical method for determining the slope and intercept parameters for the equation of a line that "best fits" a set of data $y = \alpha + \beta x + \varepsilon$. The dependent variable y, better known as a response variable, is the one we want to predict in terms of the explanatory variable x. The parameters α and β respectively are also called regression coefficients, and ε is the error. The estimated line is

$$\hat{y} = a + bx \tag{4.1}$$

where a and b are given by (3.5) and (3.1) respectively. The word 'best' means β is estimated in such way that the sum of squared value of errors (or residuals), $y - \hat{y}$ would be minimized. Thus, this estimator is also called least squares estimator or simply LSE. Though the line of best fit passes through $(0, a)$ and $(-a/b,\ 0)$, the choice of zero or $-a/b$ for the value of x may not always be meaningful as these may be far away from the scatter diagram of the data. In the following proposition, we provide two points that stems out from the consideration of the concentration of the points. One advantage of knowing these two points is that instructors can easily use them to guide students in drawing the line of best fit. This should make the student feel more comfortable with the material as they can use their prior knowledge of high school analytic geometry.

Preposition: The estimated linear regression line passes through $(\bar{x}, \bar{y})$ and $(\bar{x} + cs_{xx},\ \bar{y} + cs_{xy})$ where c is so chosen such that $\bar{x} + cs_{xx}$ does not exceed the largest x in the sample.

If the line of best fit passes through $(\bar{x}, \bar{y})$, then for any x,

$$\hat{y} = a + bx = (\bar{y} - b\bar{x}) + bx = \bar{y} + b(x - \bar{x}).$$

Thus if $x = \bar{x} + cs_{xx}$, then $\hat{y} = \bar{y} + b(\bar{x} + cs_{xx} - \bar{x}) = cbs_{xx} = cs_{xy}$ meaning the line of best fit passes through $(\bar{x} + cs_{xx}, \bar{y} + cs_{xy})$. The line of best fit passes through the mean point and also a point which is cs_{xx} units away horizontally and cs_{xy} units away vertically from the mean point. It may be noted that the quantity c in the proposition can be chosen such that $\bar{x} + cs_{xx}$ is between the smallest $(x_{\min})$ and the largest $(x_{\min})$ values of x. That is $x_{\min} \le \bar{x} + cs_{xx} \le x_{\max}$ which means

$$\frac{x_{\min} - \bar{x}}{s_{xx}} \le c \le \frac{x_{\max} - \bar{x}}{s_{xx}}. \qquad (4.2)$$

There are also other interesting points on the regression line that are possibly in the data concentration. A table is prepared below:

x	$\hat{y}$	Comments
$\bar{x}$	$\bar{y}$	
$\bar{x} + cs_{xx}$	$\bar{y} + cs_{xy}$	$x_{\min} \le \bar{x} + cs_{xx} \le x_{\max}$
$\bar{x} + v$	$\bar{y} + bv$	$v = s_x / s_y$
$\bar{x} + r$	$\bar{y} + br$	r is any value from -1 to 1 and not necessarily the observed sample correlation coefficient.

The last row of the table includes points such as $(\bar{x} + 1, \bar{y} + b)$ or $(\bar{x} - 1, \bar{y} - b)$.

Example 4.1 Suppose that the amount of time (x) spent per week on a Statistics Course contributes to student grades (y). That is, the more the time spent, the higher is the grade. Suppose that (x, y) is a data point and we have 3 observations $A(4,80),\ B(5,83),\ C(6,88)$. It can be checked that

$$s_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 1263 - \frac{(15)(251)}{3} = 8,$$

$$s_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 77 - \frac{(15)^2}{3} = 2.$$

The sample means are $\bar{x} = 5$ and $\bar{y} = 83.67$ so that the regression line passes through $(\bar{x},\ \bar{y}) = (5,\ 83.67)$. In addition, it also passes through $(\bar{x} + cs_{xx},\ \bar{y} + cs_{xy})$, or, $(5 + 2c,\ 83.67 + 8c)$ where c can be chosen as 0.5 to make the time spent per week to be 6. The point is calculated to be $(5 + 2 \times 0.5,\ 83.67 + 8 \times 0.5)$ or $(6,\ 87.67)$. By plotting the above two points we get the regression line as shown in the Figure 1 below.. One can also check that the slope and the intercept are then given by $(87.67 - 83.67)/(6 - 5) = 4$ and 63.67 respectively, so that the line is given by

$$\hat{y} = 63.67 + 4x\ . \qquad (4.3)$$

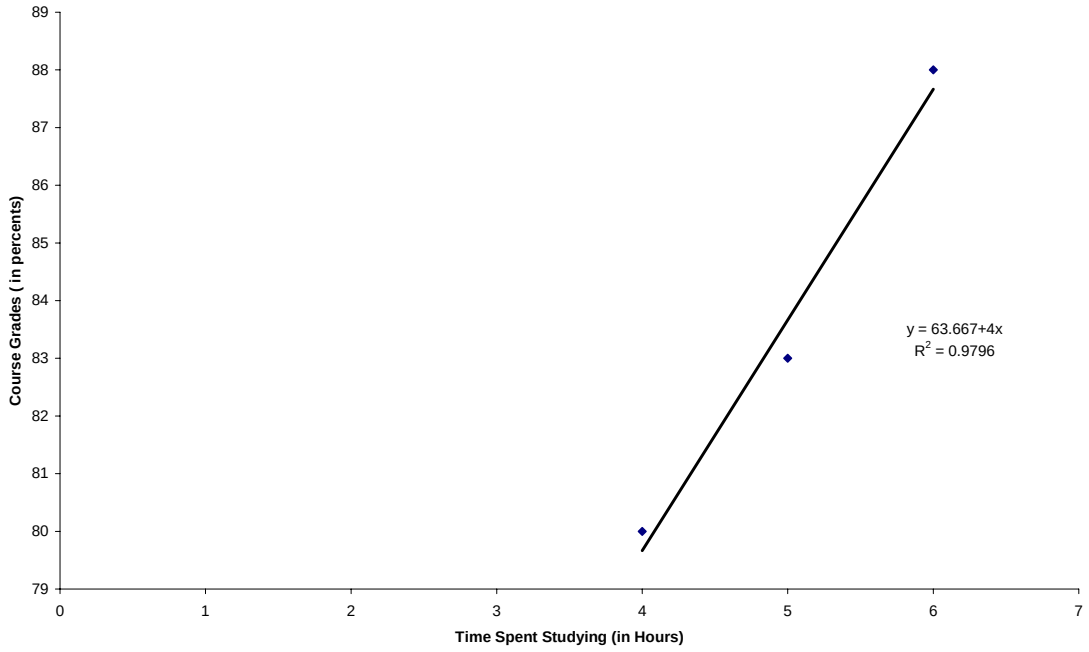

Figure 1. Time Spent on Studies Versus Course Grades

Hence we show that the line of best fit passes through the mean point and also a point which is $s_{xx} = 2$ units away horizontally and $s_{xy} = 8$ units away vertically from the mean point.

The quantity r is, indeed, defined as the linear correlation coefficient between x and y. If the random variables are perfectly correlated, the gradient of the line of best fit is given by s_y / s_x for perfect positive correlation, and $-s_y / s_x$ for perfect negative correlation (Mayer and Sykes, 1996, p 35). There is a tendency among some applied statisticians to use the above gradient for the line of the best fit when the variables are very strongly (but not perfectly) correlated. This results in a line called the SD (Standard Deviation) line. In other words, it is the line around which the points are clustered when they are highly correlated.

Suppose that $b = s_y / s_x$ (i.e., $r = 1$), and $x = \bar{x} + cs_x$, then $\hat{y} = a + bx = a + b(\bar{x} + cs_x) = \bar{y} + cs_y$. This means if the line of best fit has a positive slope $b = s_y / s_x$, then 'SD' (Standard Deviation) line passes through $(\bar{x} + cs_x, \bar{y} + cs_y)$. That is SD line passes through the average point and a point which is cs_x units away horizontally and cs_y units away vertically. Alternatively, suppose that $b = -s_y / s_x$ (i.e., $r = -1$), and $x = \bar{x} + cs_x$, then $\hat{y} = a + bx = a + b(\bar{x} + cs_x) = \bar{y} - cs_y$. This means if the line of best fit has a negative slope $b = -s_y / s_x$, then 'SD line' passes through $(\bar{x} + cs_x, \bar{y} - cs_y)$. That is 'SD line' passes through the average point and a point which is cs_x units away horizontally and $-cs_y$ units away vertically.

5. The Average Sense of Slope

In this section we provide an example for the average sense of the slope of the line of best fit which is due to Lipovetsky and Conklin (2001).

Example 5.1 Consider estimating a regression line through following data set $A(4,80),\ B(5,83),\ C(6,88)$. We will use the notation $b(LMN)$ to mean the slope of the line that best fits the set of data points L, M and N. It is checked in Section 3 that

$$b(ABC) = \frac{s_{xy}}{s_{xx}} = \frac{8}{2} = 4$$

which is the slope of an estimated regression line through the data points of A, B and C. Now we demonstrate that the average of the slopes of lines based on pairs of points *(A & B), (A & C) & (B & C)* is equal to the slope parameter calculated above. The slopes

of the lines passing through $A(4,80)$ and $B(5,\ 83)$, $A(4,\ 80)$ and $C(6,\ 88)$, and, $B(5,\ 83)$ and $C(6,\ 88)$ are given by

$$b(AB)=\frac{83-80}{5-4}=3,\ b(AC)=\frac{88-80}{6-4}=4 \text{ and } b(BC)=\frac{88-83}{6-5}=5$$

respectively. The average of these 3 slopes gives the least squares estimate of the slope:

$$\frac{b(AB)+b(AC)+b(BC)}{3}=\frac{3+4+5}{3}=4=b.$$

6. The Least Squares Sense

In this section we provide a sense of least squares by an example.

Example 6.1 If the intercept and the slope parameter is estimated by only two points $A(4,\ 80)$ and $B(5,\ 83)$, the slope is given by $b(AB)=\frac{83-80}{5-4}=3$, and hence the intercept is 68. The sum of squared errors is calculated below.

Sample	$\hat{y}(AB)=68+3x$	e	e^2
$A(4,\ 80)$	80	0	0
$B(5,\ 83)$	83	0	0
$C(6,\ 88)$	86	2	4
SSE			4

The sum of squared errors when the intercept and the slope parameter are estimated by only two points $A(4,\ 80)$ and $C(6,\ 88)$ is calculated below. The slope is given by $b(AC)=\frac{88-80}{6-4}=4$, and hence the intercept is 64.

Sample	$\hat{y}(AC)=64+4x$	e	e^2
$A(4,\ 80)$	80	0	0
$B(5,\ 83)$	84	-1	1
$C(6,\ 88)$	88	0	1
SSE			1

If the intercept and the slope parameter are estimated by only two points $B(5,\ 83)$ and $C(6,\ 88)$, the slope is given by $b(BC)=\frac{88-83}{6-5}=5$, and hence the intercept is 64. The sum of squared errors is calculated below.

Sample	$\hat{y}(BC) = 58 + 5x$	e	e^2
$A(4,\ 80)$	78	2	4
$B(5,\ 83)$	83	0	0
$C(6,\ 88)$	88	0	0
SSE			4

The sum of squared errors when the intercept and the slope parameter is estimated by all three points A and B is calculated below.

If the intercept and the slope parameter are estimated by all three points $A(4,80)$, $B(5,83)$, $C(6,88)$, the sum of squared errors is calculated below. From (4.3), we have $\hat{y} = 63.67 + 4x$.

Sample	$\hat{y}(BC) = 63.67 + 4x$	e	e^2
$A(4,\ 80)$	79.67	0.33	0.11
$B(5,\ 83)$	83.67	-0.67	0.45
$C(6,\ 88)$	87.67	0.33	0.11
SSE			0.67

The average of the SSE of the 3 lines are $(1+4+4)/3 = 3$ while the SSE of the line of best fit is 0.67 approximately.

Appendix: Algebraic Approach to Derive Slope and Intercept

Though the derivation based on differential calculus is very popular, it may not be easily understandable at the introductory level. However, for some avid readers, we present an algebraic method that does not depend on differential calculus in this section to estimate parameters of the line of best fit.

If the regression line $y = \alpha + \beta x + \varepsilon$ is estimated by the line of best fit $\hat{y}_i = a + bx$, then SSE is given by

$$L = \sum[y_i - (a + bx_i)]^2$$
$$= \sum[(y_i - \bar{y}) + (\bar{y} - b\bar{x} - a) - b(x_i - \bar{x})]^2.$$

The above can be expanded as

$$L = \sum[(y_i - \bar{y})^2 + (\bar{y} - b\bar{x} - a)^2 + b^2(x_i - \bar{x})^2 + 2(y_i - \bar{y})(\bar{y} - b\bar{x} - a)$$
$$- 2b(y_i - \bar{y})(x_i - \bar{x}) - 2b(\bar{y} - b\bar{x} - a)(x_i - \bar{x})]$$
$$= s_{yy} + n(\bar{y} - b\bar{x} - a)^2 + b^2 s_{xx} - 2bs_{xy}.$$

This can be manipulated as

$$L = s_{yy} - \frac{s^2_{xy}}{s_{xx}} + s_{xx}\left(b - \frac{s_{xy}}{s_{xx}}\right)^2 + n(\bar{y} - b\bar{x} - a)^2$$
$$= s_{yy}(1 - r^2) + s_{xx}\left(b - \frac{s_{xy}}{s_{xx}}\right)^2 + n[a - (\bar{y} - b\bar{x})]^2.$$

Since $s_{yy} \geq 0$, $1 - r^2 \geq 0$, the quantity L attains the minimum value when $b - \frac{s_{xy}}{s_{xx}} = 0$ and $a - (\bar{y} - b\bar{x}) = 0$ simultaneously. Hence the least squares estimates of the slope and intercept are $b = \frac{s_{xy}}{s_{xx}}$ and $a = \bar{y} - b\bar{x}$ respectively.

Suppose that the regression line passes through the origin so that $y = \beta x + \varepsilon$. Let the estimated line be $\hat{y} = bx$, where b is the slope of the estimated line and the intercept is zero. Then SSE is given by $L = \sum(y_i - bx_i)^2$ which can be written as $L = \sum y_i^2 - 2b(\sum x_i y_i) + b^2 \sum x_i^2$.

Denoting $s_{yy} = \sum y_i^2$, $s_{xx} = \sum x_i^2$, $s_{xy} = \sum x_i y_i$, the quantity L can be written as

$$L = s_{yy} - 2bs_{xy} + b^2 s_{xx}$$
$$= s_{yy} - \frac{s^2_{xy}}{s_{xx}} + s_{xx}\left(b - \frac{s_{xy}}{s_{xx}}\right)^2$$
$$= s_{yy}(1 - r^2) + s_{xx}\left(b - \frac{s_{xy}}{s_{xx}}\right)^2,$$

where $r = \frac{s_{xy}}{\sqrt{s_{xx} s_{yy}}}$. The quantity L is minimized if $b - \frac{s_{xy}}{s_{xx}} = 0$, i.e., $b = \frac{s_{xy}}{s_{xx}}$.

Acknowledgements

The authors acknowledge the excellent research facilities available at King Fahd University of Petroleum and Minerals.

References

Lipovetsky, Stan and Conklin, W. Michael (2001). Regression as weighted mean of partial lines: interpretation, properties and extensions. *International Journal of Mathematical Education in Science and Technology*, 32 (5), 697-706.

Mayer, A.D. and Sykes, A.M. (1996). *Statistics.* London, Arnold.