Towards a relational understanding of the regression line

Terrence Mills
Deakin University, Vic.
tmmi@deakin.edu.au

Introduction

The sudden perception of a connection between ideas is exhilarating. In an instant, we see how things fit together; we see relationships more clearly; we have arrived at a 'relational understanding' which is a term that was first coined by Richard Skemp in 1976 (see Skemp, 1987, pp. 152–163). Barnes (2000) calls these moments 'magical'. We might call these moments 'Aha!' moments.

The purpose of this paper is to demonstrate see how several different ideas can come together in Year 12 mathematics.

The subject Further Mathematics in the Victorian Certificate of Education (VCE) is the Victorian adaptation of General Mathematics in the Australian Curriculum. I will explain how three disparate topics (parabolas, simultaneous equations, and linear regression) in VCE Further Mathematics fit together to lead students to arrive at a relational understanding of the regression line, or the line of best fit in the least squares sense. The ideas below may be pertinent to other subjects such as Mathematical Methods and some first year university subjects.

Before launching into the mathematical topic, let me offer a brief introduction to Richard Skemp and his work in mathematics education. Richard Skemp (1919–1995) was a mathematics teacher in Britain who was attracted to the study of psychology and returned to university to pursue this new interest. After completing a PhD in in psychology at the University of Manchester in 1959, Skemp pursued an academic career with a focus on teaching and learning mathematics in schools. The book *The psychology of learning mathematics* (Skemp, 1987) offers the reader a good coverage of Skemp's work and views in a single volume. Chapter 12 of this book is his most famous paper 'Relational understanding and instrumental understanding' which was first published in the British journal *Mathematics Teaching* in 1976 and then reprinted several times elsewhere (Skemp, 1976).

A recurring theme in Skemp's work is the concept of understanding. Skemp (1987, pp. 152–163) distinguishes between relational understanding and instrumental understanding. Relational understanding of a concept is a deep appreciation of the concept; instrumental understanding is shallow. Skemp illustrates the distinction by referring to division of fractions. Consider the problem of simplifying

$$\frac{2}{\left(\frac{1}{3}\right)}$$

Realising that 2 can be represented as

$$\frac{1}{3}+\frac{1}{3}+\frac{1}{3}+\frac{1}{3}+\frac{1}{3}+\frac{1}{3}$$

and hence

$$\frac{2}{\left(\frac{1}{3}\right)}=6$$

indicates a relational understanding of the process. Applying the rule 'invert and multiply' reflects an instrumental understanding of the process; it gets the job done. Skemp (1987, pp. 152–163) discusses which sort of understanding we should be striving for in the mathematics classroom.

A characteristic of Skemp's work is that he writes for classroom teachers. At one point, Skemp (1987, p. 85) describes presenting "meaningless rules" to students in a mathematics class as an insult to their intelligence: "Viewed in this light, one can begin to see why learners acquire not just a lack of enthusiasm for mathematics but a positive revulsion." One sympathises.

Now let me return to the mathematical part of this paper.

The context

The content of Further Mathematics in VCE is unrelentingly practical. There is no calculus, and students can use CAS calculators in all forms of assessment. Having said that, Further Mathematics has become the most popular Year 12 mathematics subject in the VCE. Hence this subject has an important place in the school curriculum in Victoria. *Further mathematics: Units 3 and 4* (Jones, Evans, Lipson & Staggard, 2016) is a popular text-book used in Further Mathematics; I will often refer to this book as 'the text'.

Simple linear regression is a topic in Further Mathematics. The mathematics underpinning the calculation of the equation of the straight line of best fit in the least squares sense (or regression line) is not explained in the text. Students are told that the regression line minimises the sums of squares of the residuals, and how to calculate it on a CAS calculator. When discussing the exact formula for the equation of the regression line in the text, the authors write only the following: "Fortunately the exact solution can be found using the techniques of calculus. Although the mathematics is beyond Further

Mathematics, we will use the results of this theory summarised below." (Jones et al., 2016, p. 314).

Using the terminology of Skemp (1987, pp. 152–163), this leads to an "instrumental understanding" of the line of best fit in the least squares sense rather than a "relational understanding".

The main point of this paper is to show that calculating the equation of the line of best fit in the least squares sense does not necessarily require calculus. The equation can be found using only the algebra of quadratic functions, and the method for solving two simultaneous linear equations. Here is an opportunity to bring together the concepts of parabolas, simultaneous equations, and regression, to reveal hidden connections across the curriculum (Goos et al., 2017, chapter 3; Sullivan, 2011, p. 26, Principle 2), with a view to developing a relational understanding of the regression line.

These three topics are discussed, separately, in the text; see Jones et al. (2016, chapters 22, 21 and 4 respectively). In the *Australian Curriculum: Mathematics*, they come under General Mathematics under the headings "simple non-linear algebraic expressions" (ACMGM010), "simultaneous linear equations and their applications" (ACMGM044), and "fitting a linear model to numerical data" (ACMGM057) respectively (Australian Curriculum, Assessment & Reporting Authority [ACARA], 2016). However, these topics are not linked with each other in the subject.

Below I will show how the equation of the regression line can be calculated using only the algebra of quadratic functions (or parabolas) and the method for solving two simultaneous linear equations in two unknowns.

The mathematics

This section is devoted to explaining the mathematics that underpins how the three concepts can be linked with a simple numerical example. The choice of example is deliberately simple with "low noise – clear embodiment of the concept, with little distracting detail" (Skemp 1987, p. 19).

Suppose that we have a set of bivariate data where x is the independent or explanatory variable, and y is dependent or response variable. Five data points $\{(x_i, y_i): i = 1, 2, 3, 4, 5\}$ are presented in Table 1.

Table 1. Bivariate data.

i	1	2	3	4	5
x_i	1	2	3	4	5
y_i	4	6	11	12	16

To find the equation of the line of best fit in the least squares sense to the data, $y = a + bx$, we must find values of a, b that minimise the sum of the squares of the residuals which we can express as:

$$E = \sum_{i=1}^{5}\left(y_i - \left(a + bx_i\right)\right)^2$$

The most common approach is to use calculus to minimise E by applying partial differentiation as suggested by Jones et al. (2016, p. 314). However, on closer inspection, we see that E is a quadratic function of a, and a quadratic function of b. Now we can minimise a quadratic function by finding the vertex of the corresponding parabola. So, we ought to be able to minimise E without using calculus. Indeed, this is done in Moroney (1956, pp. 277–283).

Table 2. Calculation of the sum of squares of the residuals.

	observed	predicted	residual	(residual)^2
x_i	y_i	$a + bx_i$	$y_i - (a + bx_i)$	$(y_i - (a + bx_i))^2$
1	4	$a + b$	$4 - (a + b)$	$16 + a^2 + b^2 - 8a - 8b + 2ab$
2	6	$a + 2b$	$6 - (a + 2b)$	$36 + a^2 + 4b^2 - 12a - 24b + 4ab$
3	11	$a + 3b$	$11 - (a + 3b)$	$121 + a^2 + 9b^2 - 22a - 66b + 6ab$
4	12	$a + 4b$	$12 - (a + 4b)$	$144 + a^2 + 16b2 - 24a - 96b + 8ab$
5	16	$a + 5b$	$16 - (a + 5b)$	$256 + a^2 + 25b2 - 32a - 160b + 10ab$
			Total (E) =	$573 + 5a^2 + 55b^2 - 98a - 354b + 30ab$

Starting with the data in Table 1, we see from the calculations in Table 2 that

$$\mathrm{E} = 573 + 5a^2 + 55b^2 - 98a - 354b + 30ab \qquad (1)$$

In other words, E is a quadratic function in each of the two variables a and b. Our mission is to minimise E. By re-writing E as a quadratic function of a, and as a quadratic function of b, we obtain

$$E = E(a) = 5a^2 + (-98 + 30b)\,a + (55b^2 - 354b + 573) \qquad (2)$$

and

$$E = E(b) = 55b^2 + (-354 + 30a)\,b + (5a^2 - 98b + 573) \qquad (3)$$

Let us recall a basic fact about quadratic functions and parabolas. If $E = \alpha z^2 + \beta z + \gamma$, $(\alpha > 0)$, then the minimum value of E occurs at the vertex of the parabola where

$$z = \frac{-\beta}{2\alpha} \qquad (4)$$

This can be proved by completing the square; calculus is not required.

To minimise E, we apply equation (4) to equations (2) and (3) separately. When we apply equation (4) to equations (2) and (3), we obtain two simultaneous linear equations in a and b, which, after simplification, can be written as $5a + 15b = 49$ and $15a + 55b = 177$, which, in turn, lead to $(a, b) = (0.8, 3)$.

These are the required least squares estimates of a and b that minimise E in equation (1). Thus, the equation of the regression line is $y = 0.8 + 3x$.

Conclusion

In this paper we have seen how disparate ideas can come together in a topic in Year 12 mathematics. One can calculate the equation of the regression line by using quadratic functions and parabolas, and simultaneous linear equations. All these topics are in VCE Further Mathematics. Only algebra and arithmetic are required for the calculation of the regression equation. Calculus, which is not a topic in Further Mathematics, is not necessary.

Although the calculation was illustrated with a simple numerical example, one could write it out in full generality. However, the accompanying algebraic notation would obscure the main point of this article.

When the links between these concepts are revealed, one arrives at a relational understanding of the line of best fit in the least squares sense. The subject, Further Mathematics, becomes more cohesive.

Most importantly, in future applications, one can proceed to use CAS calculators or computers in practical regression problems, confident that one has developed a relational understanding of the underlying process.

Acknowledgements

This paper developed from an assignment in ESM725 in the Master of Teaching at Deakin University in 2017. I thank Gaye Williams and Sandra Herbert who were the lecturers in the subject. I also thank those who reviewed the paper; their suggestions and comments assisted to me make several improvements.

References

Australian Curriculum, Assessment and Reporting Authority [ACARA]. (2016). *Australian Curriculum*, v8.3. Retrieved 27 April 2018 from http://www. australiancurriculum.edu.au

Barnes, M. (2000). 'Magical' moments in mathematics: Insights into the process of coming to know. *For the Learning of Mathematics, 20*(1), 33–43.

Goos, M., Vale, C. & Stillman, G., with Makar, K., Herbert, S. & Geiger, V. (2017). *Teaching secondary school mathematics: Research and practice for the 21st century* (2nd ed.). Crows Nest: Allen and Unwin.

Jones, P., Evans, M., Lipson, K. & Staggard, K. (2016). *Further mathematics: Units 3 & 4* (Rev.). Port Melbourne: Cambridge University Press.

Moroney, M. (1956). *Facts from figures* (3rd ed.). Harmondsworth: Penguin Books.

Skemp, R. (1976). Relational understanding and instrumental understanding. *Mathematics Teaching, 77*, 20–26.

Skemp, R. (1987). *The psychology of learning mathematics* (Expanded American ed.). Hillsdale NJ: Lawrence Erlbaum Associates.

Sullivan, P. (2011). Teaching mathematics: Using research-informed strategies. *Australian Education Review, 59*. Camberwell: ACER Publications.