

Introducing the Least Squares Regression Principle with Computer Technologies

Patricia A Forster, Edith Cowan University
p.forster@ecu.edu.au

This paper is about the use of computer technologies for teaching and learning the least squares regression principle. I report the introduction of the principle with two Java applets in a Year 12 (upper-secondary) class, and the use of scatter plots on an Excel spreadsheet for the introduction in a second Year 12 class. The approaches offered different advantages for learning. The paper includes a review of the research literature on students' identification and application of trend and regression relationships.

Introduction

The syllabus for Year 12 Applicable Mathematics in Western Australia (Curriculum Council, 2005) specifies that students should “find the least squares regression line, examine it's properties and draw its graph” and “students should also learn that the regression line minimises the sum of the squares of the residuals, but the derivation of this result is beyond the scope of this subject” (pp. 48-49). This paper is about teaching that led to identification of the sum of the squares relationship in two Applicable Mathematics classes in different schools in Perth, Western Australia. Applicable Mathematics is a tertiary entrance examination subject for which graphics calculators are mandated.

Research findings from several studies provide background for the paper. They address students' intuitive analysis of scatter plots, intuitive determination of trend models, and use of curve fitting technologies. The findings in the two Year 12 classes complement existing research in that technology-based approaches used for introducing the least squares relationship do not seem to have been previously reported.

Review of the literature

Intuitive analysis of scatter plots

When analysing bivariate data, students may/may not perceive a trend that is present, and the properties of data can explain their responses. For example, Ainley (2000) reports that elementary-school students recognised a positive trend in (age, height) data on a scatter plot and inferred results from the graph using interpolation. The data facilitated the recognition of a trend relationship because: the covariance between adjacent points was always positive (see Figure 1a); students were familiar with height increasing with age in real life; and there was a physical resemblance between the height being represented and vertical height of points above the axis.

On the other hand, Ben-Zvi and Arcavi (2001) report that junior high-school students were slow in perceiving a negative trend in Olympic record times for the men's one hundred meter sprint and were slow in recognising that the trend was relevant for prediction. The Olympic-records data showed local irregularity (covariance between adjacent points was sometimes positive and sometimes negative, see Figure 1b), which mediated against students' discernment of trend. The data were discrete in that Olympics occur at four yearly intervals, and this property and the irregularity were the main reasons for students' reluctance to predict from the graph.



Figure 1. Data showing a (a) regular increase and (b) local irregularity. Data in (c) an ellipse shape and (d) stacks.

Intuitive determination of trend models

When students recognise trend, they exhibit a variety of approaches when producing a trend model. For example, Buffler, Allie, Lubben, and Campbell (2001) report that first-year undergraduate physics students, in drawing a trend line: connected the points to the extreme left and right of the scatter plot; drew a line through as many points as they could; drew a line to go through the origin and then through the middle of the points; or drew an undulating curve through all the points. Others captured the idea of overall trend by drawing a line so that: the number of points above it was approximately equal to the number of points below it; or the distances between the data points and the line were more-or-less minimised.

Cobb, McClain, and Gravemeijer (2003) citing Konold report that if students fit a line by eye to minimise the distances, they frequently focus on the diagonal distances of points from the line, rather than on deviations in the vertical direction. In addition, Cobb et al. found that Year 8 students recognised a positive overall trend in data in an ellipse shape (see Figure 1c) and they visualised a trend line by ‘eyeballing’ the data. However, they did not assign any particular mathematical meaning to the location of points above and below the line, and did not relate variation (or deviation) from the trend to the situation from which data were generated. Cobb et al. concluded that the students were not doing statistics because they did not manage uncertainty in data.

When analysing stacked data (see Figure 1d) the same students inferred a positive trend but did not seem to assign any meaning to the location of data in the stacks (Cobb et al., 2003). However, they identified distribution properties of the data (e.g., that the majority of data were in the centre of the stacks) when asked to predict data if collected again. Cobb et al. concluded that stacked data could help students realise that a trend line indicates a covariation relationship about which data are distributed.

Use of curve fitting technologies

Several studies have investigated students’ use of curve fitting technologies. For example, Zbiek (1998) reports that pre-service teachers adopted one of four approaches in deciding models to fit data. They: fitted multiple models and relied on the goodness of fit statistics to choose the best model, without inspecting the plot; or they generated a linear model using particular points and did not use the curve fitter that was available. More appropriately, they: used the curve fitter and based their choice of model on fit statistics, inspection of the graph, and on how well they perceived a model reflected the real world situation (e.g., whether or not the y intercept was appropriate); or they generated function equations themselves, overlaid them on the scatterplot, and relied on qualitative judgements to choose the best model.

Chu (1996) and Lingefjärd (2002) also report on computer-based curve fitting. Chu identifies several cycles in the analysis by university students of data on the price of diamond rings with different carat weights of diamond. Limitations in students’ initial approaches were they invariably accepted a linear regression model because of a high correlation coefficient, without testing other models, and most were non-critical of a negative intercept, which indicated that a ring with no diamond had a negative price. Lingefjärd reports similar results in a preservice teacher setting: in a complex modelling task many students accepted the model that they tried first, without trying others, and did not check their results against real world knowledge. Lingefjärd recommended that asking students “to explain and argue for their models” (p. 141) forces them to be critical and leads to the disclosure of inaccuracies and misunderstandings that may otherwise remain hidden.

The study in the Applicable Mathematics classes

The two Year 12 classes ($n=22$, $n=23$) were in different all-girls’ schools. The teachers had extensive experience in teaching with technologies and each had a computer, data projector, graphics-calculator, and calculator view-screen available for classroom use. All students owned Hewlett Packard HP39 or 39+ graphics calculators. Students in both classes had drawn scatterplots and fitted lines by eye in lower-secondary school but hadn’t used a technology for curve fitting (except possibly on their own initiative). I attended the two classes for 17 and 16 consecutive lessons respectively, during the teaching of descriptive statistics. I observed whole-class discussion and discussed students’ work with

them during small-group and individual work. Multiple types of data were generated including video-recordings and copies of students' written work.

Introduction of the least squares principle in the first class

The class was discussing data on the (height, thickness) of scallop shells ($n=7$) when the teacher led students in producing a scatter plot on their graphics calculators and asked them how to predict 'thickness' if 'height' was known. A student said to draw a line. Without any further discussion, the teacher connected his laptop to the Internet and projected onto the whiteboard the regression applet from (http://matti.usu.edu/nlvm/nav/category_g_4_t_5.html) -the National Library for Virtual Manipulatives. The display showed a scatter plot, with the regression line for the data. The mean point (\bar{x} , \bar{y}) was marked clearly on the line. Summary statistics (e.g., n , r , \bar{x} , \bar{y} , and the equation for the line) were shown below the graph. The teacher added points to the graph--on, near, and distant from the line. The line and statistics were updated automatically by the technology, and the teacher asked "How do you calculate the line?". Various students responded: to draw the line through the mean point, calculate its gradient, and calculate it so the distances were minimised. When asked, they did not know which points to use to calculate the gradient and could not explain which distances were involved.

The teacher drew a scatter plot on the whiteboard, with a line through the data, and he drew the residuals from the points to the line. Referring to the residuals, he said vertical distances were used for calculating the line and he asked again how to do the calculation. A student suggested minimising the absolute distances to the line. The teacher said conventionally a different method was used and to think of standard deviation. Students suggested to square the distances, and sum them.

Next, the teacher accessed the 'Least squares' applet on the National Council of Teachers website (<http://standards.nctm.org/document/examples/chap7/7.4/index.htm>). The display showed a scatter plot, line through the data, residuals to the line, squares drawn on the residuals, and the numerical expression for the sum of the squares was given under the graph. The teacher dragged points on the graph and dragged the line to change its gradient and y -intercept. The sum of the squares expression was updated automatically. The teacher asked how to obtain the best line. Students said to 'lessen' and 'minimise' the sum. When the teacher asked how to do this they gave suggestions (increase the gradient of the line etc.).

Discussion

The approach used by the teacher attracted widespread interest and gave students the opportunity to partially define the least squares method themselves. The demonstration with the first applet started students thinking about how to calculate the line, and they were confronted with not being able to come up with a method. The teacher defined the convention of using vertical distances using his hand-drawn scatter graph, and still offered students the opportunity to decide how to calculate the line. Finally, the teacher used the NCTM applet to illustrate the least squares principle, and continued to give students the opportunity to decide the entailments of the line. Using the NCTM applet only would have been quicker, but would not have allowed students as many opportunities to think through the regression calculation.

The calculating power behind the applets made the visual approach possible. Aspects that supported visual learning were: the applet graphs could be manipulated; changes in summary statistics could be seen on the first applet; the residuals on the hand-drawn graph and squares on the NCTM applet pointed to the regression principle; and the numerical summing of the squares could be seen. Limitations of the approach were that data were treated merely as points in space and were decontextualised, although the activities were motivated originally by the prediction of scallop-shell data. (A more detailed analysis of the activity is provided in Forster, in press).

The approach in second class

The teacher in the second class set a problem involving the diamond ring data provided by Chu (1996). The dataset comprises 48 points each indicating (weight of a diamond in carats, price of a ring

containing the diamond). Prices are in Singapore dollars and are from 1992. There are several different prices for some carat weights, so the data form stacks when plotted (see Figure 1d).

The teacher provided the diamond ring data on a worksheet together with the information that price was determined by caratage, cut, colour and clarity of the diamonds. She posed the situation that a couple were planning to purchase a 0.25 carat diamond ring, and asked how much they could expect to pay. The task immediately captured students' attention, and discussion to establish what was asked revealed that some students knew a lot about diamond rings. They set about solving the problem in discussion with each other.

Subsequent whole-class discussion revealed that students: calculated the mean price for 0.25 carat diamond rings in the list of data (prices \$642, \$750, \$655, \$678, \$675); calculated the mean price for the 0.25 carat diamond rings without the outlier (\$750); or they graphed the data on the graphics calculator to determine if any prices were outliers, then calculated the mean price for the 0.25 carat diamond rings without the outlier. The class debated excluding the outlier, with some students saying it was "too different from the rest" and it was "outrageous", but the prevailing view was that a range of prices is normal because cut, clarity and colour of diamonds differ as does the workmanship in the producing a ring.

Next, the teacher asked: "What say on the way they changed their minds about the weight of the diamond? Think up a method that is more adaptable". Students suggested to: calculate the mean cost for rings with diamonds whose average weight is the required weight, and find the average gap between the mean prices (for rings with each weight of diamond) and use it to calculate the expected price. The end of the lesson was near and the teacher said to try other approaches for homework and that "a visual display can convey a lot of information".

Class discussion next day revealed that students had: produced a scatter plot for all data on their calculators, with the possibility of eyeballing the plot for the purpose of prediction. Otherwise they drew on graph paper: a column graph showing the mean price for each carat weight; or a scatter plot for all data, with a line drawn by eye (method unspecified); or a scatter plot of (carat weight, mean price) points, with a line passing through the point at the left of the graph and with gradient equal to the average of the gradients between adjacent points. Or they drew a scatter plot for all data, with trend lines for the bottom and top half of data. When asked how they had decided where to put the lines, a student said she had "looked at the average distance between the points and the lines".

The teacher projected onto the whiteboard a scatterplot that she had created in Excel for all the (carat weight, price) data and referred to it when discussing the students' methods. When they mentioned mean prices, the teacher pasted the (carat weight, mean price) points onto the graph and said of the data in stacks: "We could work out the standard deviations and look at the dispersion about the mean of the values we have got. We could look at these as one-variable data sets and deal with the individual calculations". As well, she drew a trend line, and two trend lines, on the projected display when students mentioned them.

When 'average distance between points and the line' was mentioned, the teacher asked the class: "What would we want the average distance to be?". Students suggested as small as possible and the teacher said that one method was squaring the distances from the line and choosing the line so the average squared-distance was minimised. She asked the class how she could improve the line that she had drawn through the data and they suggested to "pivot it" and "reduce the gradient". As well, the teacher asked how to minimise the sum using a calculation and a student suggested differentiating.

The teacher activated the regression line on the Excel graph and asked: "Can anyone think of one ordered pair that always lies on the line of best fit?". A student suggested (0, 0), which was disputed by others for it clearly was not true. Discussion on the pitfalls of extrapolating beyond the data followed, in particular in relation to the x and y intercepts that indicated zero and negative prices, respectively, for low carat diamonds. The teacher pointed out that the x intercept of the line implied a zero price and this indicated "they are going to give them [the diamond rings] to us", and the negative prices implied by the line indicated "they're going to pay us to take them away".

The teacher asked again 'what ordered pair always lies on the line' but no one offered a suggestion. She pasted the (overall mean carat weight, overall mean price) point onto the graph to illustrate that it lay on the line.

The teacher led students through fitting a line to data on their graphics calculators, then asked the class “the straight line took us down to where they were going to pay us to take the diamonds away, it just kept going [this had been visible on the Excel graph but was not evident on the calculator graphs], so what other graph could we perhaps try?”. Students suggested a “parabola” and an “exponential graph”. These were plotted and both were judged appropriate for describing the data in that they did not cut the x axis and had positive y intercepts but the parabola (quadratic) model fitted the data more closely at the top end. The teacher showed the class how to predict using their calculators.

Next lesson the teacher provided students with a copy of the Excel graphs that had been discussed and provided them with Anscombe’s four datasets (cited in Sowe, 2001). The data in each set have the same mean, standard deviation, and regression equation, but plot to show very different trends. The teacher used these to emphasise the necessity of checking a scatterplot before reaching any conclusions about best fit models. Causes for variation in diamond prices were revisited (clarity etc.), and the teacher pointed out the regression line gave an approximate price for a given carat weight. She proceeded to address reliability of the approximation as indicated by residual plots and the correlation coefficient.

In summary, the teacher engaged students’ interest and built on students’ suggestions as part of specifying the least squares method. The scatterplot on the Excel spreadsheet was valuable in that it supported discussion. Points and the trend line could be added to it, and numbered axes assisted interpretation. The value of the graphics calculator was it could be used to fit and display different regression models for the data.

Furthermore, the conversation in the class confirmed findings in the literature (e.g., Ben-Zvi & Arcavi, 2001; Buffler et al., 2001; Cobb et al., 2003). In particular: students initially considered some and not all data when estimating a diamond ring price; the choice of data influenced students’ subsequent intuitive analysis of trend and how to determine or calculate the trend line; and the diamond ring context which yielded data in stacks supported identification of variation in data (as distinct from covariation). I propose also that the discussion on different models prepared the class for using regression models appropriately, so they might avoid pitfalls described early in this paper (see Chu, 1996, Lingefjård, 2002; Zbiek, 1998).

In conclusion, I suggest that the approaches in both classes addressed well the Applicable Mathematics syllabus objective that students should learn that the regression line minimises the sum of the squares of the residuals. I also suggest that combining the two approaches could benefit learning: students could articulate their thoughts on how to draw trend lines for real data, the teacher could use Excel or another technology to display their suggestions, and the least squares principal could be established and illustrated with the Java applets.

Acknowledgement

This paper is part of an Australian Research Council funded project (DP0345843). The co-operation of the Year 12 teachers, Craig Davis of St Hilda’s Anglican School for Girls and Romaine Saunders, Presbyterian Ladies’ College, Perth is acknowledged.

References

- Ainley, J. (2000). Transparency in graphs and graphing tasks: An iterative design process. *Journal of Mathematical Behavior*, 19, 365-384.
- Ben-Zvi, D., & Arcavi, A. (2001). Junior high school students' construction of global views of data and data representations. *Educational Studies in Mathematics*, 45(1), 35-65.
- Buffler, A., Allie, S., Lubben, F., & Campbell, B. (2001). The development of first year physics students’ ideas about measurement in terms of point and set paradigms. *International Journal of Science Education*, 23(11), 1137-1156.
- Chu, S. (1996). Diamond ring pricing using linear regression. *Journal of Statistics Education*, 4(3) [on-line].
- Cobb, P., McClain, K., & Gravemeijer, K. (2003). Learning about statistical variation. *Cognition and Instruction*, 21(1), 1-78.

The Mathematics Education into the 21st Century Project
Universiti Teknologi Malaysia
Reform, Revolution and Paradigm Shifts in Mathematics Education
Johor Bahru, Malaysia, Nov 25th – Dec 1st 2005

- Curriculum Council (2005). Syllabus manual for year 11 and year 12 accredited courses. Osborne Park, Western Australia: Curriculum Council.
- Forster, P. A. (in press). Assessing technology-based approaches for teaching and learning mathematics. *International Journal of Mathematical Education in Science and Technology*.
- Lingefjård, T. (2002). Mathematical modelling for preservice teachers: A problem from anesthesiology, *International Journal of Computers for Mathematical learning*, 7(2), 117-143.
- Sowey, E. R. (2001). Striking demonstrations in teaching statistics. *Journal of Statistics Education*, 9(1) [on-line].
- Stanton, J. M. (2001). Galton, Pearson, and the peas: a brief history of linear regression for statistics instructors. *Journal of Statistics Education*, 9(3) [online].
- Zbiek, R. M. (1998). Prospective teachers' use of computing tools to develop and validate functions as mathematical models. *Journal for Research in Mathematics Education*, 29(2), 184-201.