Data Analysis in Precalculus and Calculus

Daniel J Teague - Durham, North Carolina, USA

1. Motivation

Data analysis can play an important role in bridging the gap between the world of mathematics and the student's world of experience. Students study functions in class, but seldom have the opportunity to see these functions and their interactions exhibited in the world around them. As the students study the behaviour of functions in calculus and precalculus courses, they learn how things should happen in theory. Through data analysis, the theory can be motivated and realised in the actual.

The principles of curve fitting, re-expression, and residual analysis, offer a very exciting and enlightening basis for the motivation and derivation of many of the functions and functional concepts taught in high school algebra and in calculus. The Mathematics Department at the North Carolina School of Science and Mathematics has created, tested, and published an innovative data-driven precalculus text and is presently writing a calculus course involving many laboratory experiences from which the examples in this article are taken.

2. Data from functions

When beginning the study of data analysis at the precalculus level, clean "well-mannered" data sets are helpful. An exceptionally good source of such data comes from the approximate solutions to the traditional max-min problems generated with computer and calculator tools (NCSSM, 1988). Consider the standard problem of finding the minimum surface area of a right circular cylinder of fixed volume, say 314cc. The equation, $A = 2\pi r^2 + 2(314)/r$, can be derived by secondary school algebra students. The students can then approximate the minimum value by using a graphing calculator. An approximate solution of $r \approx 3.68$ is quickly found. The right circular cylinder with a volume of 314cc which has the minimum surface area has a radius of approximately 3.68cm. For a given volume, a particular radius will minimise the surface area. Clearly

then, the radius which minimises the surface area is a function of the volume. But what function? What is the relationship between volume and radius which minimises the surface area? To find out, we can assign each student one or two specific volumes for which the radius which minimises the surface area is approximated. The assigned volumes and the resulting solutions for the radius become the data set (V,r). What does this data set look like? (See Figure 1.)

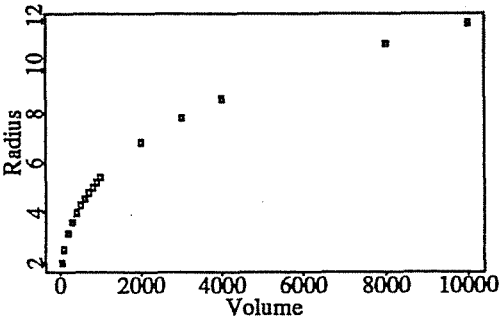

FIGURE 1

After viewing the data set, students often guess that a square root function, $r = a\sqrt{V}$, expresses the relationship between V and r. If the function is, in fact, a member of the square root family, then graphing either the ordered pairs (V,r^2) or $(\sqrt{V},r)$ should linearise the data. Below is the re-expression $(\sqrt{V},r)$ chosen by my class (Figure 2). The transformed data clearly has less curvature and the correlation coefficient is a healthy .933, but can the re-expressed data be called linear?

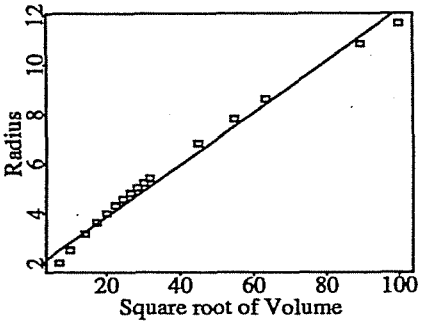

FIGURE 2

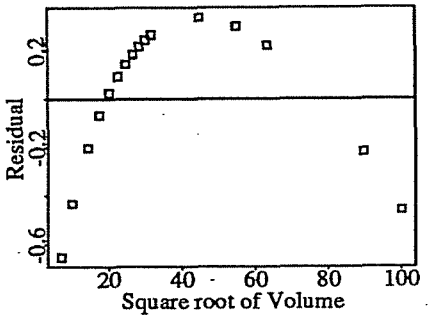

FIGURE 3

An analysis of the residuals yields a great deal of information about the fit. The residual plot (Figure 3) indicates that the linear fit partitioned the data into a group of data points below the line $x = 0$, another group above, and a third group again below. The pattern created by this partition gives evidence to the nature and direction of the curvature. Detecting such patterns is the key to linearising data. As both the original data set and the re-expressed data set are concave down, the re-expression as a square root is not sufficiently strong to linearise the data. It is necessary to try a re-expression

which is stronger than the square root. (If the concavity of the re-expressed data is opposite that of the original data set, then the chosen re-expression was too strong, and a weaker function should be investigated.) The cube root might be the next choice.

This re-expression (Figure 4) appears quite linear and the residuals are nicely scattered, so the equation $r = .542\ V^{(1/3)} - .002$ can be used as the initial model. It should be noted that the correlation coefficient for this model is .941, only marginally better than the previous value of .938. However, comparing residual plots (Figure 5) gives dramatic evidence in support of the cube root model.

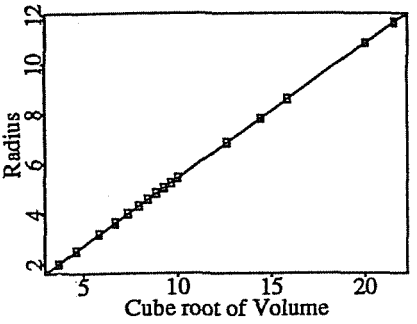

FIGURE 4

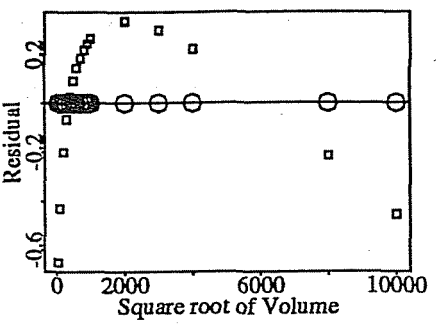

FIGURE 5

We must always decide what to do with the y-intercept generated by the fit, is it real or a residue of the fitting process? Arguing that zero volume should result in a zero radius, we conclude that the radius-intercept of -.00195 $\rightarrow$ 0 and our model for the relationship between V and r is given by $r = .542\ V^{1/3}$ (Figure 6). Given a volume, we can now approximate the radius which would minimise the surface area. (From calculus, $dA/dr = 0$ implies that $r = (V/2\pi)^{1/3}$. The cube root of $1/(2\pi) \approx .5419$.)

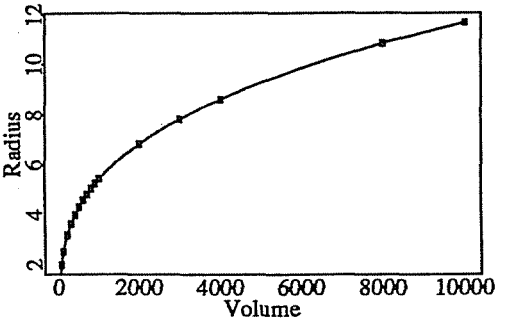

FIGURE 6

Another perhaps even more powerful example comes from a problem in blood testing (Meyer, 1970). Suppose that N samples of blood need to be tested for a certain characteristic, each test result being either positive or negative. If the blood samples could be pooled by grouping a portion of k samples together and then testing the pooled sample, the number of tests could be reduced. If the pooled sample is negative, then all

the individuals in the pool are negative, and we have checked k people with one test. If, however, the pooled sample is positive, we know only that at least one of the individuals in the sample will test positive. Each member of the sample must then be retested individually and a total of k+1 tests will be necessary to do the job. The relationship between the number of tests (T), the size of the population (N), the probability of testing positive for each individual (p), and the group size (k) is given by:

$$T = N/k + (1 - (1-p)^k)\,(N/k)\,k = N(1/k + [1 - (1-p)^k]).$$

What is the relationship between p and k which minimises T? The variable of interest, k, is both algebraic and transcendental in this equation, and calculus is of no help. However, by using a graphing calculator to approximate the value of k which minimises T for various values of p, a data set can be created:

p	.3	.25	.2	.15	.10	.05	.04	.03	.02	.01	.005	.001	.0005
k	2.7	2.8	2.9	3.2	3.8	5.0	5.6	6.3	7.6	10.5	14.9	32	45

Look at the graph of this data set. As $p \rightarrow 0$, the size of the group increases rapidly. Students can argue for a vertical asymptote at $p = 0$. Could the relationship between p and k which minimises T be a simple reciprocal function?

Sketching the graph of the re-expressed data set (1/p, k) reveals a concavity which is different from the original data set. This suggests a weaker re-expression. The data set $(1/\sqrt{p}, k)$ is strikingly linear with a slope of .997 and a y-intercept of .699. An approximating model of $k \approx 1/\sqrt{p} + .7$ takes us directly from the probability p to the group size k which minimises T. Sketch the model against the data to emphasise the quality of the fit.

3. The use of calculus

Once the student begins the study of calculus, data analysis can take on an even greater role. The students can investigate all of the data sets generated in precalculus from a new perspective. In precalculus, students learned that a function can be defined by its data set. In calculus, they learn that a function can also be defined by how the data set changes (and later, by how the data set accumulates). To study the changes in the data, a new data set is derived from the original. To derive the new set, replace each y-value with the local average rate of change of the function. That is, replace y_i with $(y_{i+1} - y_{i-1})/(x_{i+1} - x_{i-1})$. This value is called the symmetric difference. The original n data points generate a derived set of n-2 data points. When the derived data sets are compared to the original data sets, something exceptional happens. Every quadratic data set collected in precalculus generates a linear derived data set. Every exponential data set generates another exponential data set. The derived set is characteristic of the original data set. It doesn't take students long to realise that if they know how the derived set behaves, through the usual re-expression techniques of data analysis, then they also know, up to a constant, the behaviour of the original data set.

Take, for example, the data set which came from minimising the surface area of the can. Our precalculus analysis determined the relationship to be $r = .542\ (V)^{1/3}$. If we generate the derived set of symmetric differences (sd), we have the data set given below.

V	10	50	100	200	400	600	800	1000	2000	3000
r	1.17	2	2.52	3.16	3.99	4.57	5.03	5.42	6.83	7.82
sd	**	.015	.00773	.0049	.00353	.00315	.00213	.00150	.00120	**

By re-expressing the data, it can be seen that the derived data set (V, sd) is a negative two-thirds power function. The average change in a cube root function seems to be a negative two-thirds power function!

Similar results can be generated from all of the data sets in the precalculus course. After generating these approximate results and making conjectures about the relationship between the original data set and the derived set, the conjectures can be verified with the development of the traditional differential calculus.

In another example, consider the logistic curve $P = M/(1 + B\ e^{Mkt})$, where $M = 1000$, $B = 39$, and $k = .001$. The early portion of the curve is given below with the goal of the data analysis to estimate M.

t	0	.29	.57	.86	1.14	1.43	1.71	2.00	2.29	2.57	2.86	3.14	3.43
P	20	26.4	34.9	45.9	60.1	78.5	101.7	131.0	167.1	210.8	262.2	321.0	386.2
sd	**	26.0	34.0	44.2	57.0	72.9	92.0	114.4	139.5	166.3	193.0	217.1	**

The logistic curve is defined by $dP/dt = kP[M-P]$. This means that the graph of the symmetric differences against P (P, sd) should be quadratic. Moreover, the ratio of the symmetric difference to P graphed against P (P, sd/P) should be linear, since $(dP/dt)/P = kM - kP$. If we fit a line to this re-expression, we should be able to find k, from the slope, and also M, from the y-intercept.

By fitting a least squares line to the re-expressed data, we find that the slope is -0.00103 and the y-intercept is 1.00939. The slope of -0.00103 implies that $k \approx .001$ and the intercept of 1.00939 implies that $M \approx 1000$, as expected.

The data analysis materials described in this article were designed to illustrate the interplay between secondary school mathematics and data. They were not designed solely for the students at NCSSM, a statewide, tuition free, public, residential high school for students with a demonstrated talent and interest in science and mathematics. The materials are being taught at secondary schools throughout the country. To use data analysis as a principal component of the precalculus and calculus curriculum requires a great investment of time and thoughtful effort from the students, a commitment to the use of emerging technology from the school, and, most importantly, a willingness from the teacher to let the students investigate, discuss (even argue), speculate, and justify, rather than memorise and remember.

References

Meyer, Paul L (1970) *Introductory Probability and Statistical Applications.* Addison-Wesley Publishing Company, Reading, Massachusetts.

The North Carolina School of Science and Mathematics (1991) *Contemporary Precalculus Through Applications.* Janson Publishing Company, Providence, Rhode Island.

The North Carolina School of Science and Mathematics (1988) *New Topics for Secondary School Mathematics : Data Analysis.* The National Council of Teachers of Mathematics, Reston, Virginia.