MODELING WITH MATRICES: A DATA-DRIVEN APPROACH

Jerry L. Moreno
John Carroll University
USA

As statistics continues to increase its presence in the school curriculum, particularly the mathematics one, it becomes increasingly more difficult for teachers to be able to fit everything in. They complain that if statistics must be included, then something must go. One suggestion to solve their problem is to combine the topics of statistics and mathematics so that both are presented together. The NSF-funded project Data-Driven Mathematics has done precisely that. The series of eleven modules motivates mathematics topics found in pre-algebra, algebra, geometry, advanced algebra, and advanced mathematics from a data point of view that involves students. This paper presents some insight as to how this may be done with the advanced mathematics topic of matrices (See Burrill, Burrill, Landwehr, & Witmer, 1998).

INTRODUCTION

The National Council of Teachers of Mathematics (NCTM), the professional organization for teachers in the United States, published a set of curriculum standards in 1989. It was called *Curriculum and Evaluation Standards for School Mathematics* and was revised in 2000 under the name *Principles and Standards for School Mathematics.* Although school districts do not have to follow the recommendations, most states have adopted them as guidelines for various forms of state-determined assessment of local districts' mathematics curricula. So, in essence PSSM is being followed throughout the nation. The most exciting part of these documents has been the integration of statistics and probability into the mathematics curriculum at every grade level. Over the past decade, textbooks for the elementary grades have increased the inclusion of statistical graphs and basic probability ideas. Middle school and high school texts are following suit with a few series having been right on target shortly after the NCTM *Standards* document was first published. Serious attention however will not be paid to statistics and probability in the curriculum until the various assessment tools at state and national levels begin to include meaningful statistics and probability questions. That has not yet happened generally speaking.

Much more could be said about to what extent and how statistics and probability will play into the school curriculum, but that is a topic for another time. It should be noted that there is thought being given to how mathematics instruction can be made more interesting so as to increase students' lifelong retention of useful mathematical ideas including statistics and probability. This movement has its basis in *Mathematics and Democracy: A Case for Quantitative Literacy* (Steen, 2001). There is also a very interesting project that is looking at revising the traditional algebra-geometry-advanced algebra-precalculus sequence of the order in which mathematics has been taught. The project focuses on creating four strands 1) algebra and trigonometry, 2) statistics and probability, 3) geometry, and 4) discrete mathematics that are taught vertically through the three or four years of the high school curriculum (Hirsch, 1997).

The movement to carry out the recommendations that include statistics and probability into the mathematics curriculum is taking many forms. One called *Data-Driven Mathematics* (DDM) is based on the premise that most mathematics topics in the high school curriculum can be motivated from a data point of view. Not only do data give a meaningful approach to why the mathematics being studied is important, it involves the student. This paper gives an overview of how data can motivate the learning of matrices.

THE USE OF MATRICES TO REPRESENT RATINGS

Studies are often conducted to determine the most desirable city to live in based on a number of variables such as cost of living, quality of education, availability and security of jobs, extent of cultural and recreation opportunities, transportation, health facilities, air quality, safety and crime, and climate. Each city in the study could be given a ranking per variable and the resulting rankings for all the variables weighted somehow to determine the most desirable city in

which to live. Note that if rankings are used per criterion, the advantage is that all variables are on the same scale, but the variability of the original data has been lost.

To solve the difficulty of what to do in the case of the variables of interest not having the same scale of measurement, ratings from 1 (low) to 100, say (high) might be used. Yet other situations may involve variables that are not all on the same scale and it is desired to use neither ratings nor rankings to solve the problem.

As a first example, consider the problem of having to determine a textbook for a course in which the variables of interest are the quality of graphics and text, the accuracy of content material, the topics covered, the quality of problems in the exercises, the quantity of problems, and whether or not there are supplementary materials to aid teacher and student. Teachers at a high school known to the author had to choose a text series for their mathematics curriculum from six available that will be called T1 through T6. There were eight teachers in the department that rated each series from a low score 1 to a high score 10 per each of the seven variables previously listed. The rounded mean of the teachers ratings were taken and are shown in Table 1.

Table 1
Rounded Mean of the Teacher Ratings

	Topics Covered A	Graphics Quality B	Text Quality C	Problems Quality D	Problems Quantity E	Content Accuracy F	Supplements G
T1	7	9	9	7	8	9	7
T2	9	8	9	8	8	9	8
T3	8	8	9	10	6	8	8
T4	8	7	9	8	8	9	8
T5	9	9	8	9	7	8	8
T6	9	10	9	7	8	10	8

For purposes of brevity, suppose that variables B, D, and E are considered the most important and that D is three times as important as B, and E is twice as important as B. A model for the problem is then written as textbook rating R = B + 3D + 2E. To determine that highest rated textbook, students would probably compute R for each Ti by using the weighted formula. In so doing they should easily recognize that each computation involves the (B,D,E) values being multiplied respectively by (1,3,2). They can then be shown that their observation is organized mathematically by the use of matrices and that the definition of multiplication of matrices follows directly from their observation.

Let $\mathbf{S}_{6x3} = \begin{bmatrix} 9 & 7 & 8 \\ 8 & 8 & 8 \\ 8 & 10 & 6 \\ 7 & 8 & 8 \\ 9 & 9 & 7 \\ 10 & 7 & 8 \end{bmatrix}$ $\mathbf{W}_{3x1} = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}$ $\mathbf{R}_{6x1} = \begin{bmatrix} r_{T1} \\ r_{T2} \\ r_{T3} \\ r_{T4} \\ r_{T5} \\ r_{T6} \end{bmatrix}$.

Which textbook is ranked first according to the chosen weighting scheme?

The answer is the ratings vector $\mathbf{R}_{6x1} = \mathbf{S}_{6x3}\mathbf{W}_{3x1} = \begin{bmatrix} 46 \\ 48 \\ 50 \\ 47 \\ 50 \\ 47 \end{bmatrix}$. So, we see that publishers T3 and T5 are tied for being *best* according to the specified weighting scheme. If there were other weighting

schemes, then **W** would be expanded with each new scheme occupying a new column and the size of **R** would change accordingly, one ratings column for each weighting scheme in **W**. The student should also recognize that the weighting scheme $\begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$ yields the same ratings as $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ does except the ratings are one-third as large.

THE USE OF GRAPHS TO REPRESENT RATINGS

A graphical geometric solution is introduced to answer the question as to which text is best if one were interested only in Graphics Quality (B) and Problems Quality (D) (see Figure 1).

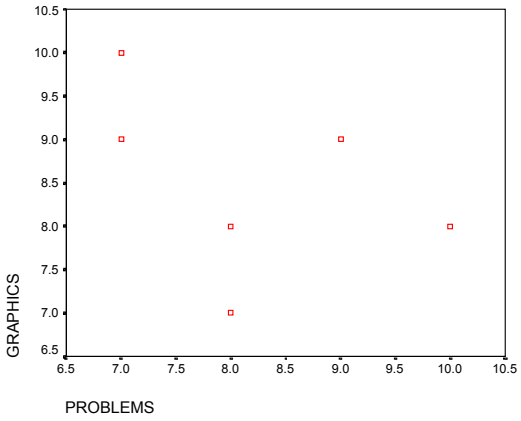

Figure 1. Geometric solution to Graphics/Problem Question.

If a straightedge were taken vertically, sweeping from the right of the graph toward the Graphics axis, then clearly T3 is the best on Problems Quality. If the straightedge were swept horizontally from the top toward the Problems axis, then T6 is the sole choice for Graphics Quality. Note that T5 is better than T2 with respect to both variables. But overall no text dominates all texts.

The equally weighted model would be R = B + D, or B = -D + R, where B represents Graphics Quality and D represents Problems Quality. If the line B = -D + R were swept from the upper right hand corner across the scatterplot toward the origin, then one would see that T3 and T5 are met first at the same time. This result clearly coincides with the matrix result of ratings, written in its transpose for convenience of space, $\begin{bmatrix} 16 & 16 & 18 & 15 & 18 & 17 \end{bmatrix}^t$. T3 and T5 are tied.

Students could be asked that if they were employees of the publisher of T6, then what model(s) would they be happy with? Note that the slope of the line through T6 and T5 is -1/2. The students should suggest that they would be pleased with any model whose slope is larger than -1/2. Then ask them that for such models how would the customer be viewing the relationship between Graphics Quality and Problems Quality? Suppose the model with slope -5/16, say, were chosen. Then, B + 5/16 D = R. The customer would interpret this model as one in which Problems Quality is considered 5/16 as important as Graphics Quality, or Graphics Quality is 16/5 as important as Problems Quality.

PREDICTION

The statistics topic of prediction or regression analysis is rich in the use of matrices. There is only space enough here to outline the major topics with one example. Manatees are aquatic mammals found, for example, along the Florida (USA) coast. Many of them have been killed or seriously injured by powerboats. Data on the number of powerboat registrations (in thousands) and the number of manatees killed by boats in Florida in the years 1977 to 1990 are presented in Table 2 (see Yates et al., 1999).

Table 2
Relationship between Powerboat Registrations and Boating Deaths

Year	'77	'78	'79	'80	'81	'82	'83	'84	'85	'86	'87	'88	'89	'90
Registrations	447	460	481	498	513	512	526	559	585	614	645	675	711	719
Manatee Kill	13	21	24	16	24	20	15	34	33	33	39	43	50	47

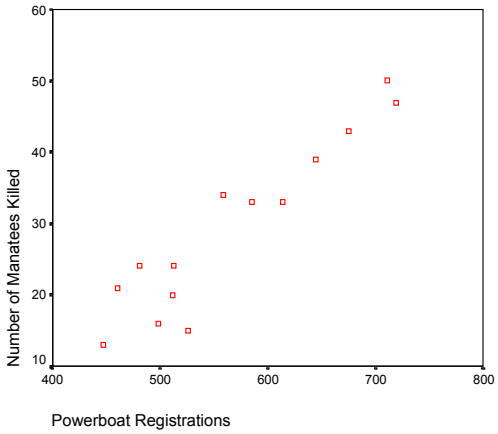

Figure 2. Graphical Representation of the Relationship between Powerboat Registrations and Boating Deaths.

As Figure 2 shows, the data appear to have a linear relationship, hence a linear model $Y = \hat{Y} + r$ would be appropriate, where Y is the actual number of manatees killed, $\hat{Y} = b_0 + b_1 X$ is the predicted number of manatees killed, and r is the residual $Y - \hat{Y}$. The task is to find the *best* possible values for b_0 and b_1 based on the fourteen pieces of paired data. Written in matrix notation, $\mathbf{Y}_{14x1} = \mathbf{X}_{14x2}\,\mathbf{b}_{2x1} + \mathbf{r}_{14x1}$ where **Y** consists of the column of number of manatees killed data, **X** consists of two columns, the first of which is a column of 1's and the second contains the powerboat registrations (in thousands) data, and **r** contains the residuals r_1 through r_{14}, and **b** is the column vector $(b_0 \;\; b_1)^t$. Note that the column of 1's is needed in X because b_0 is multiplied by 1 in the $\hat{Y}$ equation. If the criterion to find b_0 and b_1 is to minimize the sum of the squared residuals, then it can be shown that the vector $\mathbf{b} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$. A measure that is used to judge how good is the model is the root mean squared error, the square root of the sum of the squared residuals divided by n.

CONCLUSION

Matrices and the operations involving matrices can be introduced in a way that engages students. This paper has tried to show that using data is a powerful way to cover mathematics topics, namely matrices, as well as statistics topics. Details for the prediction modeling and also transforming nonlinear data to linearity may be found in *Advanced Modeling and Matrices* (Burrill et al., 1998).

REFERENCES

Burrill, G., Burrill, J., Landwehr, J., & Witmer, J. (1998). *Advanced modeling and matrices*. Dale Seymour Publications.

Hirsch, C. (1997). *Core-Plus Mathematics Project*. Available www.wmich.edu/cpmp/

NCTM (1989). *Curriculum and evaluation standards for school mathematics.* Reston, Virginia: NCTM

NCTM (2000). *Principles and standards for school mathematics.* Reston, Virginia: NCTM.

Steen, L.A. (Ed.) (2001). *Mathematics and democracy: The case for quantitative literacy.* The Woodrow Wilson National Fellowship Foundation.

Yates, D., Moore, D., & McCabe, G. (1999). *The practice of statistics.* W. H. Freeman, p. 773.