ADVANCED MATHEMATICS

Advanced Modeling and Matrices

G. BURRILL, J. BURRILL, J. LANDWEHR, J. WITMER

DATA-DRIVEN MATHEMATICS

DALE SEYMOUR PUBLICATIONS®

Advanced Modeling and Matrices

DATA-DRIVEN MATHEMATICS

Gail F. Burrill, Jack Burrill, James M. Landwehr, and Jeffrey Witmer

Dale Seymour Publications®
Orangeburg, New York

This material was produced as a part of the American Statistical Association's Project "A Data-Driven Curriculum Strand for High School" with funding through the National Science Foundation, Grant #MDR-9054648. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Managing Editor: Cathy Anderson, Alan MacDonell

Senior Math Editor: Carol Zacny

Project Editor: John Sullivan

Production/Manufacturing Director: Janet Yearian

Senior Production Coordinator: Alan Noyes

Design Manager: Jeff Kelly

Text and Cover Design: Christy Butterfield

Composition: Joan Olson

Art: Carl Yoshihira

Cover photo: David Madison/Bat Hitting Baseball

This book is published by Dale Seymour Publications®, an imprint of Addison Wesley Longman, Inc.

Dale Seymour Publications
125 Greenbush Road South
Orangeburg NY, 10962
Customer Service: 800-872-1100

Printed in the United States of America.

Order number DS21188

ISBN 1-57232-255-1

1 2 3 4 5 6 7 8 9 10-ML-01 00 99 98 97

This Book Is Printed On Recycled Paper

DALE SEYMOUR PUBLICATIONS®

Authors

Gail F. Burrill
Whitnall High School
Greenfield, Wisconsin
University of Wisconsin-Madison
Madison, Wisconsin

Jack Burrill
Whitnall High School
Greenfield, Wisconsin
University of Wisconsin-Madison
Madison, Wisconsin

James M. Landwehr
Bell Laboratories
Lucent Technologies
Murray Hill, New Jersey

Jeffrey Witmer
Oberlin College
Oberlin, Ohio

Consultants

Emily Errthum
Homestead High School
Mequon, Wisconsin

Patrick Hopfensperger
Homestead High School
Mequon, Wisconsin

Henry Kranendonk
Rufus King High School
Milwaukee, Wisconsin

Vince O'Connor
Milwaukee Public Schools
Milwaukee, Wisconsin

Maria Mastromatteo
Brown Middle School
Ravenna, Ohio

Data-Driven Mathematics Leadership Team

Miriam Clifford
Nicolet High School
Glendale, Wisconsin

Kenneth Sherrick
Berlin High School
Berlin, Connecticut

Richard Scheaffer
University of Florida
Gainesville, Florida

James M. Landwehr
Bell Laboratories
Lucent Technologies
Murray Hill, New Jersey

Gail F. Burrill
Whitnall High School
Greenfield, Wisconsin
University of Wisconsin-Madison
Madison, Wisconsin

Acknowledgments

The authors thank the following people for their assistance during the preparation of this module:

- The many teachers who reviewed drafts and participated in field tests of the manuscripts
- The members of the *Data-Driven Mathematics* leadership team, the consultants and writers
- Kathryn Rowe and Wayne Jones for their help in organizing the field-test process and the Leadership Workshops
- Barbara Shannon for many hours of word processing and secretarial services
- Jean Moon for her advice on how to improve the field-test process

Table of Contents

About *Data-Driven Mathematics*

Historically, the purposes of secondary-school mathematics have been to provide students with opportunities to acquire the mathematical knowledge needed for daily life and effective citizenship, to prepare students for the workforce, and to prepare students for postsecondary education. In order to accomplish these purposes today, students must be able to analyze, interpret, and communicate information from data.

Data-Driven Mathematics is a series of modules meant to complement a mathematics curriculum in the process of reform. The modules offer materials that integrate data analysis with high-school mathematics courses. Using these materials will help teachers motivate, develop, and reinforce concepts taught in current texts. The materials incorporate major concepts from data analysis to provide realistic situations for the development of mathematical knowledge and realistic opportunities for practice. The extensive use of real data provides opportunities for students to engage in meaningful mathematics. The use of real-world examples increases student motivation and provides opportunities to apply the mathematics taught in secondary school.

The project, funded by the National Science Foundation, included writing and field testing the modules, and holding conferences for teachers to introduce them to the materials and to seek their input on the form and direction of the modules. The modules are the result of a collaboration between statisticians and teachers who have agreed on statistical concepts most important for students to know and the relationship of these concepts to the secondary mathematics curriculum.

Using This Module

Many problems in the world require more than one variable to predict an outcome. Problems in medicine, science, engineering, and economics involve many variables, and a major task involves describing possible relationships among these variables. Matrices are an important mathematical tool in this process. They enable us to organize information and to efficiently apply different algorithms to data. As a set of mathematical elements, matrices also provide an example for investigating the field properties that are often taken for granted as we work with real numbers: the properties of closure, identity, inverse, associativity, commutativity, and distributivity. Finally, matrices have a very close link to vectors which can be powerful tools in analysis. In each situation presented in the module, the geometry is brought out and its relation to an algebraic approach is explored.

Have you ever wondered how people come up with rankings for the best places to live, the best jobs, or the best schools? In Unit I, you will investigate rating systems and ways to create ranks based on several variables. Here you will use matrices to organize information and to efficiently combine data into ratings.

People often use a regression model to relate the values of one variable to the values of another variable. In Unit II, you will extend the idea of a regression model to the case in which several variables are used to predict another variable. This process makes use of matrices both to organize the information and to produce the mathematical solution that gives the best predictions.

Content

Mathematics content: You will be able to

- Use matrices to organize data.
- Multiply matrices and understand the limitations of matrix multiplication.
- Find the transpose of a matrix.
- Solve a system of equations using matrices.

- Graph equations of the form $ax + by = c$ for given a, b.
- Plot points in three dimensions.
- Graph equations of the form $ax + by + cz = d$ for given a, b, c.
- Identify parallel planes in space.
- Represent ordered n-tuples as vectors.

Statistics content: You will be able to

- Use least squares linear regression to find a model for paired data.
- Find and interpret the residuals for a given model.
- Calculate and interpret the mean and standard deviation for a set of data.
- Find and interpret the sum of squared errors for a model.
- Use multiple linear regression techniques to find a model for multivariate data.

Modeling, Matrices, and Ranks

INTRODUCTORY ACTIVITY

Where Do You Want to Live?

Which city in the United States has the best weather? the best schools? the best jobs?

Which factors are most important when choosing a place to live?

At some point in your life you may have the opportunity to move to a new location. It may be a different place within the state in which you already live, or it may be to a neighboring state, or to one across the continent. Your family may move, you may move to get a better job, or you may move for many other reasons.

OBJECTIVE

Practice using ratings to create ranks.

EXPLORE

Rate and Rank Cities

Assume you can move to any city in the United States. Consider how you might choose the city.

Data Collection and Analysis

1. List five factors about the city you would consider important in helping you make your choice.
2. Are all of these factors equally important?
3. Compare the factors you listed with those listed by others in your class. As a group decide on five factors you all agree are important.
4. Select five cities. Rate them on each of the factors and use your ratings to rank the cities. What was your top ranked city? How did your ranks compare to others in class?

Representing Ratings with Matrices

If you were to buy a new car, which one would you buy?

What information would you want to know before you decided?

OBJECTIVES

Represent information in a matrix.

Multiply matrices and understand the constraints in the process.

Use matrices to apply a formula.

Solve systems of equations using a matrix.

"The best midsize sedan for 1994 was the Honda Accord," according to *Consumer Reports*. "Number one rating for the Intrepid." "Mustang leads the Pack." Minnesota is the healthiest state in Northwestern National Life Insurance Company's state health 1993, ranking. Rochester, Minnesota is one of the most "livable" cities according to *Money* magazine.

Cities, states, cars, and many other things are continually being rated to determine the "best." How is the "best" determined? In particular, what qualities do you think would give a car a top rating?

INVESTIGATE

Car Ratings

The staff of *Car and Driver* magazine selected six station wagons priced around $23,000, loaded them for a trip with fishing, boating and hiking gear and drove them through southern Ohio. The cars and the ratings given by the staff after the trips, as reported in the July 1994 issue of the magazine, are in the matrix below. A *matrix* is an ordered array of information. The dimensions of a matrix give the number of rows and columns in the matrix. The car rating matrix is a 6×10 matrix (read 6 by 10) because it has six rows and ten columns. The vehicles were rated from 1 (low) to 10 (high) in each category; the scores for each category were collected from each staff member and averaged; for example, the average score for the Honda Accord for handling was a 9.

Car Ratings

	Drive Train	Handling	Ride	Comfort (Drive)	Back Seat	Cargo Space	Cargo Utility	Style	Value	Fun to Drive
Honda Accord	8	9	8	9	8	7	7	9	9	8
Mercury Sable	8	7	8	7	7	10	10	8	8	7
Mitsubishi	7	6	8	8	9	10	7	8	6	6
Subaru Legacy	8	9	8	8	7	9	6	7	8	9
Toyota Camry	9	8	9	9	10	9	10	7	8	8
Volkswagen	9	9	8	9	9	8	7	9	8	9

Discussion and Practice

1. Use the ratings in the matrix to determine which wagon is the highest rated. Be ready to justify your method.

2. Chi decided that handling, styling, and value were the most important characteristics and reduced the table to a smaller matrix of just those ratings:

	Handling	Style	Value
Honda Accord	9	9	9
Mercury Sable	7	8	8
Mitsubishi	6	8	6
Subaru Legacy	9	7	8
Toyota Camry	8	7	8
Volkswagen	9	9	8

a. He felt that value (v) was the most important, handling (h) the next most important, and style (s) the least, so he weighted the categories to find the rating (r) as $r = 2h + 1s + 3v$. Find the weighted rating for each car using Chi's formula.

b. Tara felt that styling was more important than handling, so her formula was $r = h + 2s + 3v$. Compare Tara's weighted ratings to Chi's ratings.

A vector can be used to help organize the information. Chi's weighted ratings vector would be [2 1 3]. Written as a column vector, this would be: $\begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}$

To find the weighted ratings for each car, you can use matrices to help organize your work.

Ratings Matrix · Weight Vector = Weighted Ratings

For the Honda Accord, the row vector for the Honda ratings can be multiplied by the column weight vector.

$$
\begin{array}{c} \\ \\ \text{Honda Accord} \\ \\ \\ \end{array}
\begin{array}{c} \text{Handling} \quad \text{Style} \quad \text{Value} \\ \left[\begin{array}{ccc} 9 & 9 & 9 \\ \underline{\qquad} & \underline{\qquad} & \underline{\qquad} \\ \underline{\qquad} & \underline{\qquad} & \underline{\qquad} \\ \underline{\qquad} & \underline{\qquad} & \underline{\qquad} \\ \underline{\qquad} & \underline{\qquad} & \underline{\qquad} \\ \underline{\qquad} & \underline{\qquad} & \underline{\qquad} \end{array}\right] \end{array}
\cdot
\begin{array}{rc} & \text{Chi} \\ \text{Handling} & \left[\begin{array}{c} 2 \\ 1 \\ 3 \end{array}\right] \\ \text{Style} & \\ \text{Value} & \end{array}
= 9\times 2 + 9\times 1 + 9\times 3 =
\left[\begin{array}{c} 54 \\ \underline{\qquad} \\ \underline{\qquad} \\ \underline{\qquad} \\ \underline{\qquad} \\ \underline{\qquad} \end{array}\right]
$$

For the Mercury Sable:

$$
\begin{array}{c} \\ \\ \\ \text{Mercury Sable} \\ \\ \end{array}
\begin{array}{c} \text{Handling} \quad \text{Style} \quad \text{Value} \\ \left[\begin{array}{ccc} \underline{\qquad} & \underline{\qquad} & \underline{\qquad} \\ 7 & 8 & 8 \\ \underline{\qquad} & \underline{\qquad} & \underline{\qquad} \\ \underline{\qquad} & \underline{\qquad} & \underline{\qquad} \\ \underline{\qquad} & \underline{\qquad} & \underline{\qquad} \\ \underline{\qquad} & \underline{\qquad} & \underline{\qquad} \end{array}\right] \end{array}
\cdot
\begin{array}{rc} & \text{Chi} \\ \text{Handling} & \left[\begin{array}{c} 2 \\ 1 \\ 3 \end{array}\right] \\ \text{Style} & \\ \text{Value} & \end{array}
= 7\times 2 + 8\times 1 + 8\times 3 =
\left[\begin{array}{c} \underline{\qquad} \\ 46 \\ \underline{\qquad} \\ \underline{\qquad} \\ \underline{\qquad} \\ \underline{\qquad} \end{array}\right]
$$

For the Mitsubishi:

$$
\begin{array}{c} \\ \\ \\ \text{Mitsubishi} \\ \\ \end{array}
\begin{array}{c} \text{Handling} \quad \text{Style} \quad \text{Value} \\ \left[\begin{array}{ccc} \underline{\qquad} & \underline{\qquad} & \underline{\qquad} \\ \underline{\qquad} & \underline{\qquad} & \underline{\qquad} \\ 6 & 8 & 6 \\ \underline{\qquad} & \underline{\qquad} & \underline{\qquad} \\ \underline{\qquad} & \underline{\qquad} & \underline{\qquad} \\ \underline{\qquad} & \underline{\qquad} & \underline{\qquad} \end{array}\right] \end{array}
\cdot
\begin{array}{rc} & \text{Chi} \\ \text{Handling} & \left[\begin{array}{c} 2 \\ 1 \\ 3 \end{array}\right] \\ \text{Style} & \\ \text{Value} & \end{array}
= 6\times 2 + 8\times 1 + 6\times 3 =
\left[\begin{array}{c} \underline{\qquad} \\ \underline{\qquad} \\ 38 \\ \underline{\qquad} \\ \underline{\qquad} \\ \underline{\qquad} \end{array}\right]
$$

Or, all at once:

$$
\begin{array}{l} \textbf{Ratings Matrix} \\ \begin{array}{l} \\ \text{Honda Accord} \\ \text{Mercury Sable} \\ \text{Mitsubishi} \\ \text{Subaru Legacy} \\ \text{Toyota Camry} \\ \text{Volkswagen} \end{array}
\begin{array}{c} \text{Handling} \quad \text{Style} \quad \text{Value} \\ \left[\begin{array}{ccc} 9 & 9 & 9 \\ 7 & 8 & 8 \\ 6 & 8 & 6 \\ 9 & 7 & 8 \\ 8 & 7 & 8 \\ 9 & 9 & 8 \end{array}\right] \end{array} \end{array}
\cdot
\begin{array}{l} \textbf{Weight Vector} \\ \begin{array}{rc} & \text{Chi Weights} \\ \text{Handling} & \left[\begin{array}{c} 2 \\ 1 \\ 3 \end{array}\right] \\ \text{Style} & \\ \text{Value} & \end{array} \end{array}
=
\begin{array}{l} \textbf{Weighted Ratings} \\ \begin{array}{l} \\ \text{Honda Accord} \\ \text{Mercury Sable} \\ \text{Mitsubishi} \\ \text{Subaru Legacy} \\ \text{Toyota Camry} \\ \text{Volkswagen} \end{array}
\begin{array}{c} \text{Overall Rating} \\ \left[\begin{array}{c} 54 \\ 46 \\ 38 \\ 49 \\ 47 \\ 51 \end{array}\right] \end{array} \end{array}
$$

3. Chi got 47 for the Toyota Camry.

a. Explain how he got that number.

b. Suppose Chi's weight vector was $\begin{bmatrix} 20 \\ 10 \\ 30 \end{bmatrix}$. Find the weighted ratings.

c. Compare the results Chi got using the two different weight vectors.

Paul and Kuo chose different weights for the three categories. The matrix representation below shows the ratings matrix and a matrix of the two weight vectors.

$$\begin{array}{l} \\ \text{Honda Accord} \\ \text{Mercury Sable} \\ \text{Mitsubishi} \\ \text{Subaru Legacy} \\ \text{Toyota Camry} \\ \text{Volkswagen} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Handling} & \text{Style} & \text{Value} \end{array} \\ \begin{bmatrix} 9 & 9 & 9 \\ 7 & 8 & 8 \\ 6 & 8 & 6 \\ 9 & 7 & 8 \\ 8 & 7 & 8 \\ 9 & 9 & 8 \end{bmatrix} \end{array} \qquad \begin{array}{l} \\ \text{Handling} \\ \text{Style} \\ \text{Value} \end{array} \begin{array}{c} \begin{array}{cc} \text{Paul} & \text{Kuo} \end{array} \\ \begin{bmatrix} 5 & 3 \\ 2 & 3 \\ 8 & 5 \end{bmatrix} \end{array}$$

4. To find Paul's weighted ratings for the Honda, multiply the first row, the ratings for each category, by the column with Paul's weight vector. To find Kuo's weighted ratings for the Honda, multiply the first row by Kuo's weight vector in the weighted ratings matrix.

a. Find the weighted ratings for Paul and Kuo.

b. How do the two compare?

5. Refer back to the original matrix on page 5.

a. Explain what each matrix below would represent.

Matrix R

$$\begin{array}{c} \\ H \\ MS \\ MI \\ S \\ T \\ V \end{array} \begin{array}{c} \begin{array}{ccc} S & V & F \end{array} \\ \begin{bmatrix} 9 & 9 & 8 \\ 8 & 8 & 7 \\ 8 & 6 & 6 \\ 7 & 8 & 9 \\ 7 & 8 & 8 \\ 9 & 8 & 9 \end{bmatrix} \end{array}$$

Matrix W

$$\begin{array}{c} \\ S \\ V \\ F \end{array} \begin{array}{c} \begin{array}{cc} \text{Pedro} & \text{Tina} \end{array} \\ \begin{bmatrix} 1 & 3 \\ 1 & 1 \\ 5 & 3 \end{bmatrix} \end{array}$$

b. What formula would describe Pedro's weighted ratings? Tina's weighted ratings?

c. Each entry in a matrix is sometimes called a *cell*. The 7 in the second row and third column (*MS*, *F*) of matrix R is in cell R_{23}. What entry is in cell R_{42}? What does the entry represent?

d. Find the product of the two matrices. Explain what the entries in the result represent.

6. Changing the entries in matrix W can affect the results.

a. Describe the effect of an entry in the second matrix being 0.

b. What would happen to the ratings and the ranking if Tina's weights for each category were exactly double Pedro's weights for those categories?

7. Use the ratings for driver comfort, handling, cargo space, and value with the following weights:

driver comfort 3, handling 5, cargo space 2, value 10.

a. Create the matrices and find the weighted ratings.

b. Explain why the following sets of weighted values don't give the same rankings:

Method 1: driver comfort 5, handling 2, cargo space 1, value 3
Method 2: driver comfort 2, handling 3, cargo space 5, value 3

8. The U.S. Department of Transportation periodically issues information about airlines, and on the basis of this information, airlines make claims about being the "Number 1" airline. The following table contains the information released for January to June 1995 for overbooking and complaints and for the first quarter of 1995 for mishandled baggage and on-time arrival.

Airline Complaints

Airline	Mishandled Baggage per 1,000 Passengers	Complaints per 100,000 Passengers	Percent On-Time Arrival	Overbooked per 10,000 Passengers
Southwest	4.14	0.25	80.5	3.04
US Air	4.81	0.48	80.2	1.67
American	4.99	0.59	75.2	0.44
America West	5.05	0.59	76.7	2.56
Continental	5.19	0.71	76.6	0.89
Delta	5.20	0.77	78.5	0.81
Alaska	5.25	0.85	77.7	1.65
United	5.28	1.09	79.0	0.32
Northwest	6.05	1.38	80.0	0.26
Trans World	6.33	1.51	72.0	0.68

Source: *Air Travel Consumer Report*, July, August, 1995

a. What other factors might you consider important when rating an airline?

b. Based on the data given, how would you rank the airlines? Be ready to justify your method.

c. Basaj ranked the airlines from 1 to 10 in each category, then rated an airline by finding the average of the ranks. What are the advantages and disadvantages of his method?

Transpose of a Matrix

A lumber yard carried wood in different prices and grades. An inventory of the number of board feet of each kind they had is reported in matrix I. Matrix P is the price matrix.

Matrix I

$$\begin{array}{l} \\ \text{Oak bd. ft.} \\ \text{Cherry bd. ft.} \\ \text{Pine bd. ft.} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Grade I} & \text{Grade II} & \text{Grade III} \end{array} \\ \begin{bmatrix} 750 & 526 & 300 \\ 200 & 127 & 300 \\ 250 & 750 & 750 \end{bmatrix} \end{array}$$

Matrix P

$$\begin{array}{l} \\ \text{Oak} \\ \text{Cherry} \\ \text{Pine} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Grade I} & \text{Grade II} & \text{Grade III} \end{array} \\ \begin{bmatrix} \$1.80 & \$1.95 & \$3.00 \\ \$1.42 & \$1.50 & \$2.50 \\ \$0.95 & \$1.10 & \$2.00 \end{bmatrix} \end{array}$$

9. Would it be useful to find the product PI? What would the entry in cell PI_{11} represent?

You can exchange the rows and columns of P and obtain a matrix called P^T, the *transpose* of P, as illustrated below.

Matrix P^T

$$\begin{array}{l} \\ \text{Grade I} \\ \text{Grade II} \\ \text{Grade III} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Oak} & \text{Cherry} & \text{Pine} \end{array} \\ \begin{bmatrix} \$1.80 & \$1.42 & \$0.95 \\ \$1.95 & \$1.50 & \$1.10 \\ \$3.00 & \$2.50 & \$2.00 \end{bmatrix} \end{array}$$

10. Find the product of IP^T and call it C.

a. What does the entry in C_{11} represent?

b. What does the entry in C_{12} represent?

c. How much money does the dealer have tied up in Cherry wood? Which cell in the matrix will answer the question?

d. What is the value of the inventory in oak, cherry, and pine worth? How can you use a matrix to find the answer?

11. Use the same matrices I and P above.

a. Find $I^{T}P$.

b. What information can you obtain from the product?

12. Consider the following matrices:

$$A = \begin{bmatrix} 2 & 4 & 5 \\ 1 & 6 & 7 \end{bmatrix} \quad B = \begin{bmatrix} 3 \\ 2 \\ 8 \end{bmatrix} \quad C = \begin{bmatrix} 5 & 7 \\ 9 & 4 \end{bmatrix}$$

a. What are the dimensions of A^{T}? B^{T}? $(A^{T})A$?

b. Find $(A^{T})A$.

c. Find $(AB)^{T}$, $(A^{T})\ (B^{T})$, and$(B^{T})(A^{T})$. What conjecture can you make?

d. Check your conjecture using A and C.

13. A square matrix can have a multiplicative inverse. Remember that the product of an element and its inverse will yield an identity and that an identity will not change the value of the element when it is used as a factor. In the real numbers, 1 is the multiplicative identity, and $1/a$ is the multiplicative inverse of a, where a is not equal to 0.

a. Verify that $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ is the multiplicative identity matrix for all 2×2 matrices.

b. Why will only a square matrix have a multiplicative inverse?

c. Use the inverse key on your calculator to find the multiplicative inverse of the matrix $\begin{bmatrix} 2 & 4 \\ 5 & 1 \end{bmatrix}$ and verify that it is the inverse.

d. Find the inverse of $\begin{bmatrix} 2 & 4 \\ 1 & 2 \end{bmatrix}$. What conclusion can you make?

Mean and Deviations

Another way to compare the car ratings is to consider the mean rating for each of the cars and how consistent those ratings are. You can use matrices to determine the deviation from the mean for the ratings in each of the categories and to find the sum of the squared deviations, an important statistic that is used in many different ways. Refer to the car rating matrix from the beginning of the lesson.

Car Ratings

	Drive Train	Handling	Ride	Comfort (Drive)	Back Seat	Cargo Space	Cargo Utility	Style	Value	Fun to Drive
Honda Accord	8	9	8	9	8	7	7	9	9	8
Mercury Sable	8	7	8	7	7	10	10	8	8	7
Mitsubishi	7	6	8	8	9	10	7	8	6	6
Subaru Legacy	8	9	8	8	7	9	6	7	8	9
Toyota Camry	9	8	9	9	10	9	10	7	8	8
Volkswagen	9	9	8	9	9	8	7	9	8	9

14. The mean rating over all the categories for the Mercury is 8. Make a 10×1 matrix R for the ratings for the Mercury and another 10×1 matrix M where each entry is the mean, 8.

a. What will $R - M$ give you?

b. Let $R - M = D$. What will you have if you find the sum of the entries in D?

15. Write the transpose of D.

a. What are the dimensions of D^{T}?

b. Find $(D^{T})D$. What does this represent?

c. You can use $(D^{T})D$ to find the standard deviation for the ratings. Divide $(D^{T})D$ by the number of categories and take the square root. What is the standard deviation for the Mercury?

16. Divide the other cars among the members of your group.

a. For your car, find the mean, then use matrices to find the sum of the squared deviations from the mean.

b. Find the standard deviation of the ratings for your car.

c. Compare your results to those of others in your group. Which car has the smallest sum of squared deviations from the mean?

d. For which car are the ratings most consistent?

Solving Systems of Equations

Systems of equations are often used to solve problems involving data, and matrices can be very useful for solving systems of equations. For example, each year *Places Rated Almanac* rates the 343 metropolitan areas in the United States on living costs, job outlook, crime, health, transportation, education, the arts, recreation, and climate. The areas are rated on safety by using a linear combination of their violent crime rate (the sum of the murder, robbery, and aggravated assault rates) and their property crime rate (the sum of the burglary, larceny-theft, and motor vehicle theft rates). The rates are per 100,000 people. The formula for computing the crime rating, r, can be expressed as

$$av + bp = r$$

where v is the violent crime rate, p is the property crime rate, a and b are the coefficients for the formula.

17. The Miami rate for violent crime was 1,721 and for property 10,934. Miami's total crime rating was 2,821. Washington, DC had a 549 violent crime rate and a 4,790 property crime rate for a total rating of 1,028.

a. Write the system of equations you would need to solve to find the coefficients for the formula used to obtain the total crime rating.

The information can be written as a matrix system:

$$C \cdot A = R$$

where C is the matrix of the crime ratings, A is the coefficient matrix, and R is the rating matrix,

or $\begin{bmatrix} 1721 & 10934 \\ 549 & 4790 \end{bmatrix} \times \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 2821 \\ 1028 \end{bmatrix}$.

b. Verify that matrix multiplication will produce the system you wrote for part a.

c. To solve the system for the coefficients, use the inverse of C, C^{-1}. Write the formula to find the total crime rating. Do the coefficients make sense? Why or why not?

d. What is the total crime rating for Ottawa-Hull in Ontario, Canada, if the violent crime rate is 218 and the property crime rate is 5,732?

Summary

A matrix is an ordered array of elements. If the matrix has m rows and n columns, the dimensions of the matrix are $m \times n$. To multiply matrices, multiply each row of the first matrix by each column of the second. If the number of entries in each row of the first matrix does not match the number of entries in each column of the second, you cannot multiply the matrices. The entry in the i, j position of the product matrix is the result of multiplying the entries in the ith row by those in the jth column and summing the products. The transpose of a matrix is the matrix obtained by exchanging the rows and the columns of the original matrix such that the *i*th row becomes the *i*th column.

Practice and Applications

18. The top five metropolitan areas noted in *Places Rated Almanac* for 1993 are given in the matrix below with the ranks assigned by the Almanac to each of the categories. The rankings are based on scores that are a composite of data collected for that category. For example, recreation scores are based on assets such as good restaurants, public golf courses, bowling lanes, zoos, aquariums, NCAA Division I football and basketball teams, and miles of ocean or Great Lakes coastline. The metropolitan areas are then ranked according to those scores, where a 1 is the top ranked or best score for a given category.

Ranking Metropolitan Areas

	Cost of Living	Jobs	Housing	Transp.	Educ.	Health	Crime	Arts	Recreation	Climate
Seattle, WA	302	16	312	15	31	30	240	24	16	15
San Francisco, CA	343	71	543	21	23	8	288	6	5	1
San Diego, CA	326	7	328	27	12	66	285	33	3	2
Pittsburgh, PA	269	109	214	24	24	41	51	36	150	98
Washington, DC	308	5	314	10	6	10	215	3	105	54

Source: *Places Rated Almanac*, 1993

Note that in this problem, low numbers are better than high numbers.

Miguel Hernandez, who is planning to retire, weighted the categories as follows: cost of living 4, crime 2, health care 10, climate 5, and the rest 0.

a. Of the cities given, which have the best ratings for climate?

b. Find the weighted ratings for Miguel. Which town gets the best rating?

c. Chris Strom and Rex McNall rated the categories differently:

Chris: cost of living 8, crime 1, health care 5, climate 5

Rex: cost of living 5, crime 0, health care 8, climate 10.

How do the three different weightings compare? Which weighting scheme do you think is the best? Why?

d. As a group, select three or four of the variables you all consider important. Each of you should independently assign weights to the variables you choose. Use matrix multiplication to compare the effects of the different weighting schemes.

e. Suppose the weights assigned to the categories are changed as given below. Create matrices and describe the effect on the rankings of the metropolitan areas.

Arts 1, health care 2, education 3

Arts 10, health care 20, education 30

Arts 3, health care 2, education 1

Arts 5, health care 25, education 45

19. The rankings for three other metropolitan areas are given in the following matrices, along with two different weighting schemes, I and II.

$$\begin{array}{l} \\ \text{Boston} \\ \text{Anaheim} \\ \text{New York} \end{array} \begin{array}{c} \begin{array}{cccc} \text{Jobs} & \text{Transp} & \text{Educ} & \text{Rec} \end{array} \\ \begin{bmatrix} 29 & 30 & 4 & 57 \\ 1 & 197 & 7 & 49 \\ 71 & 1 & 1 & 15 \end{bmatrix} \end{array} \qquad \begin{array}{l} \\ \text{Jobs} \\ \text{Transp} \\ \text{Educ} \\ \text{Rec} \end{array} \begin{array}{c} \begin{array}{cc} \text{I} & \text{II} \end{array} \\ \begin{bmatrix} 10 & 10 \\ 5 & 1 \\ 10 & 5 \\ 2 & 8 \end{bmatrix} \end{array}$$

a. Find the product P of the two matrices. How do the rating schemes compare?

b. What does the entry in cell P_{32} mean? What calculations created that entry?

20. Consider the matrices below:

$$\begin{bmatrix} 1 & 3 & 5 \\ 2 & 1 & 4 \\ 1 & 2 & 2 \end{bmatrix} = A \qquad \begin{bmatrix} 3 & 7 & 2 \\ 1 & 4 & -2 \\ 3 & 2 & -1 \end{bmatrix} = B \qquad \begin{bmatrix} 4 & -2 \\ 3 & 1 \\ 1 & 2 \end{bmatrix} = C$$

a. Find AB and BA and compare your results. What conclusion can you draw?

b. Find AC and CA. What has to be true about the dimensions of the matrices you want to multiply?

c. If you multiply a matrix that is 3×5 and one that is 5×2, what will the dimensions of the product be?

21. More information about the five wagons reviewed by *Car and Driver* is given in the matrix below.

	0–30 mph	Braking Dist. 70 mph (ft)	City (mpg)	Highway (mpg)	700-Mile trip (mpg)	Price ($)
Honda	3.2	188	23	29	22	21,850
Mercury	3.0	197	19	28	20	21,645
Mitsubishi	3.8	193	18	24	18	26,790
Subaru	2.5	185	18	23	18	23,645
Toyota	2.9	174	18	24	21	23,303
Volkswagen	3.3	178	18	25	20	23,890

a. For which of these six categories are low numbers considered to be good?

b. Using the categories from part a, devise two different weight formulas to rank the cars. Use matrix multiplication to find the weighted ratings. Compare the results in ranking from these two weight formulas.

c. For which of these six categories are high numbers considered to be good?

d. Using the categories from part c, devise two different weight formulas to rank the cars. Use matrix multiplication to find the weighted ratings. Compare the results in ranking from these two weight formulas.

22. The National Association for Stock Car Auto Racing (NASCAR) sponsors the Winston Cup races, covering between 400 and 600 miles each on 17 designated speedways around the country. For each race, drivers are given a certain number of points for winning or for the place in which they finish, for the number of laps in which they lead, and five bonus points for the driver who leads the most laps. In the 1995 Martinsville, Virginia, Goody's 500, Dale Earnhardt won the race and earned 1440 points. He led in 252 of the 500 laps. In the Dover, Delaware, race held in late summer 1995, Jeff Gordon won with a total of 2,180 points and led in 400 of the 500 laps.

a. Set up a system of equations to determine how many points are given for first place and for the number of laps a driver leads.

b. Use matrices to solve the system.

c. How many points would a racer earn in a race if he led in 150 of the 500 laps and won the race?

d. Terry Labonte finished second in the 1995 Richmond race and led in 2 laps. He earned a total of 180 points. How many points are awarded by NASCAR for a second-place finish?

23. After the Wisconsin Badgers won the Rose Bowl in 1994, Badger tee shirts were on sale everywhere in Wisconsin. At one store, they had three colors: red, gray, and white. They came in small, medium, large, and extra large. On the day of the Rose Bowl, the inventory was as given in I, the inventory matrix below.

Matrix *I*

$$\begin{array}{c} \\ \text{White} \\ \text{Red} \\ \text{Gray} \end{array} \begin{array}{c} \begin{array}{cccc} \text{Small} & \text{Medium} & \text{Large} & \text{Extra Large} \end{array} \\ \begin{bmatrix} 120 & 120 & 200 & 250 \\ 100 & 90 & 150 & 150 \\ 150 & 175 & 175 & 200 \end{bmatrix} \end{array}$$

The dealer's cost and selling price for each size of shirt is given below.

$$\begin{array}{c} \\ \text{Dealer cost} \\ \text{Selling Price} \end{array} \begin{array}{c} \begin{array}{cccc} \text{Small} & \text{Medium} & \text{Large} & \text{Extra Large} \end{array} \\ \begin{bmatrix} \$6 & \$7 & \$7 & \$8 \\ \$12 & \$12 & \$15 & \$18 \end{bmatrix} \end{array}$$

a. Create matrices and use matrix multiplication to determine how much money the dealer had tied up in each color of tee shirt.

b. On the day after the Rose Bowl, the store had sold tee shirts as recorded in matrix S.

Matrix *S*

$$\begin{array}{c} \\ \text{White} \\ \text{Red} \\ \text{Gray} \end{array} \begin{array}{c} \begin{array}{cccc} \text{Small} & \text{Medium} & \text{Large} & \text{Extra Large} \end{array} \\ \begin{bmatrix} 25 & 50 & 75 & 70 \\ 50 & 78 & 123 & 100 \\ 15 & 82 & 105 & 84 \end{bmatrix} \end{array}$$

On which color tee shirt did the dealer make the most profit? Explain how you found your answers.

c. Make an inventory matrix for his remaining stock.

LESSON 2

Ranking and Scatter Plots

What is the best brand of tennis shoes?

Which kind of shampoo works the best?

How do you decide what brand to buy?

It is very important to make a good decision when you are about to make a major purchase such as a VCR or stereo or mountain bike. Many people look through consumer and trade magazines for ratings on the quality of the item they are considering.

In this lesson you will investigate how these comparisons can be better understood through the use of scatter plots, straight lines that are related to weighted values, and matrices that summarize the relationships.

OBJECTIVES

Write a rating equation for two weighted variables.

Understand the relationships between the relative positions of points on a scatter plot.

Determine ranking using a scatter plot.

INVESTIGATE

Selecting a Printer

Suppose you wanted to buy a dot-matrix printer to use with a personal computer. The table that follows contains quality rating scores for seven printers that were reviewed in the *1993 Buying Guide by Consumer Reports.* Each printer is rated in each of two areas: quality when printing graphics and quality when printing text.

Dot-Matrix Printers

Model	Graphics	Text
Star NX-2420 Rainbow	81	77
Epson LQ-510	78	78
Panasonic KX-P1124	84	71
Star NX-2420 Multipoint	82	73
Citizen GSX-140	77	76
NEC P3200	60	82
Panasonic KX-P1123	72	79

A scatter plot of Text score (T) versus Graphics score (G) is in Figure 2.1.

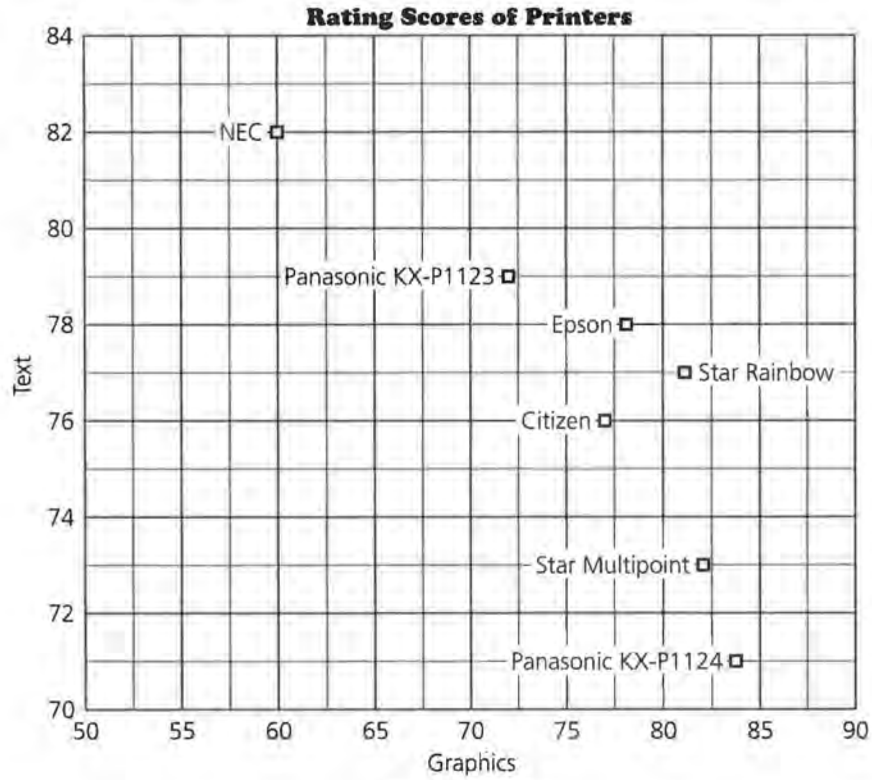


Figure 2.1

Discussion and Practice

1. Use the data in the table on Dot-Matrix Printers.

a. For each variable, which printer has the highest rating?

b. For each variable, which printer has the lowest rating?

To analyze the data more carefully, begin with only two printers, the Epson and the Citizen. The scatter plot in Figure 2.2 shows only the points for these two printers.

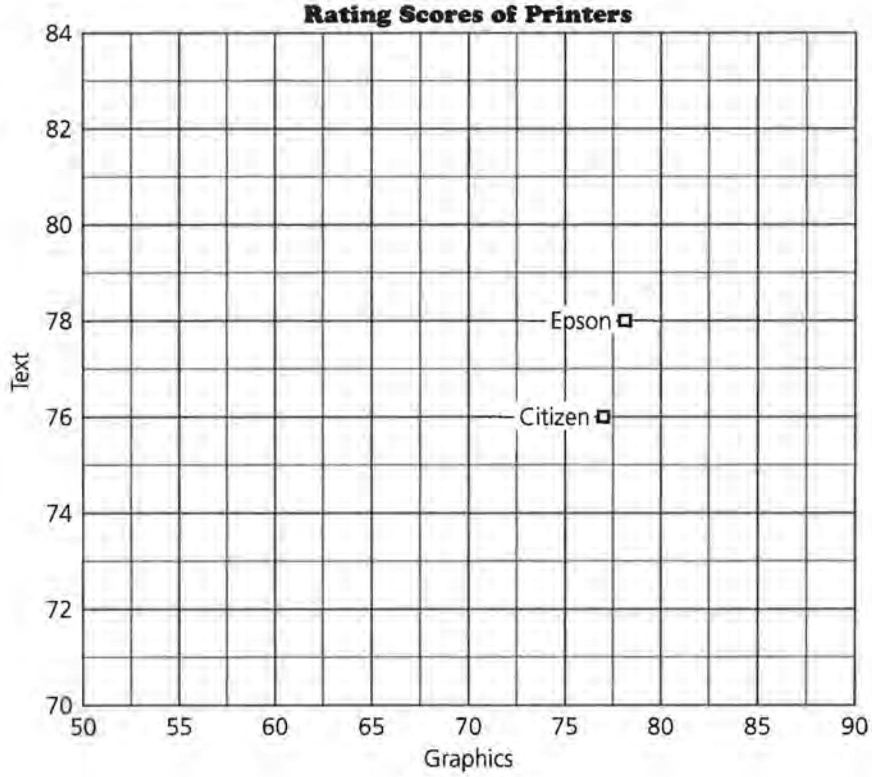


Figure 2.2

2. Use the data to compare the Epson and Citizen printers.

a. If you only cared about printing text, which of these two printers would you choose? Why?

b. If you only cared about printing graphics, which printer would you choose? Why?

c. In terms of the two variables, which printer would you rank higher? Explain why.

In general for a scatter plot in which larger values are better for both variables, you can say that a given point *dominates* all those points having lower values on both variables. That is, if the first printer has larger values on both variables and larger indicates better performance, then the first printer was better than the second printer in terms of these two variables. The indicated point X in the scatter plot in Figure 2.3 dominates any point that is below and to the left, that is any point for another printer that lies in the shaded region.

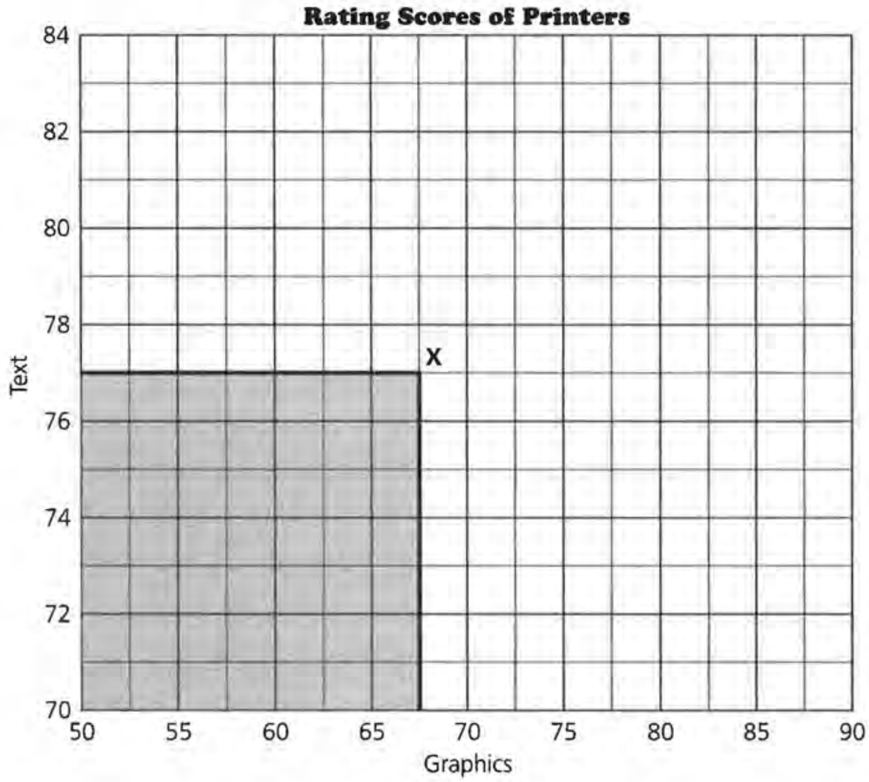


Figure 2.3

3. Suppose point P for another printer dominates the point labeled X in Figure 2.3.

a. Make a sketch of the scatter plot and place point P so that it dominates point X.

b. Sketch the region in the scatter plot that contains all the points that dominate X, that is, all possible locations for P.

4. Now consider the Citizen and Star Multipoint, shown in Figure 2.4.

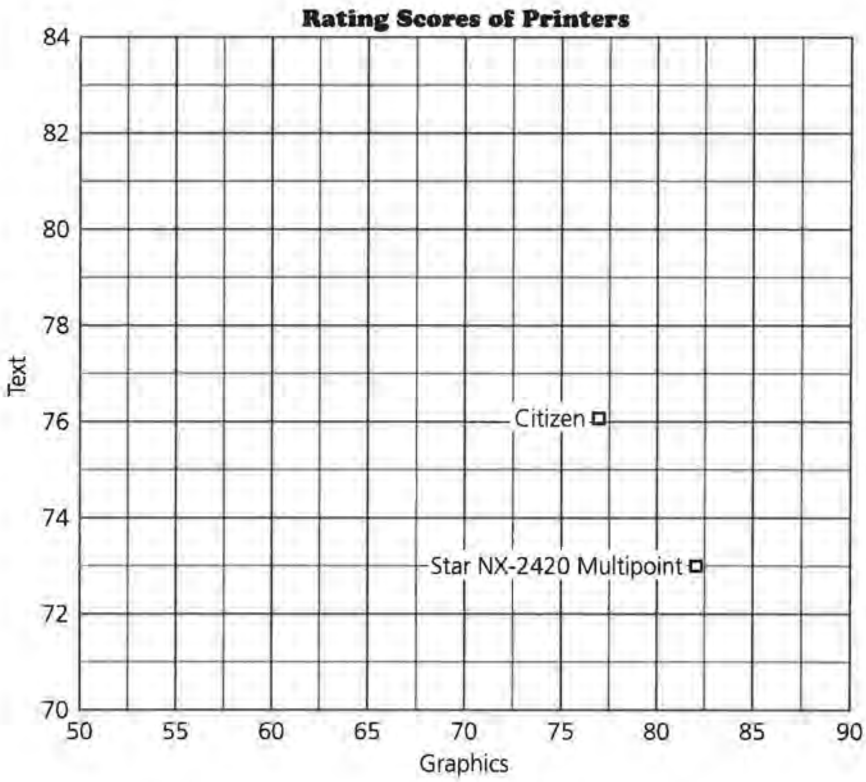


Figure 2.4

a. How do the two printers compare on quality of printing graphics and text?

b. Does one printer dominate the other?

5. Make a sketch of Figure 2.4.

a. Shade the region containing points dominated by the Citizen.

b. Use another color to shade the region containing the points dominated by the Star.

c. Identify the region containing points dominated by both printers.

d. Identify the region containing points dominated by one printer but not the other.

6. Return to the original scatter plot in Figure 2.5 for the printers.

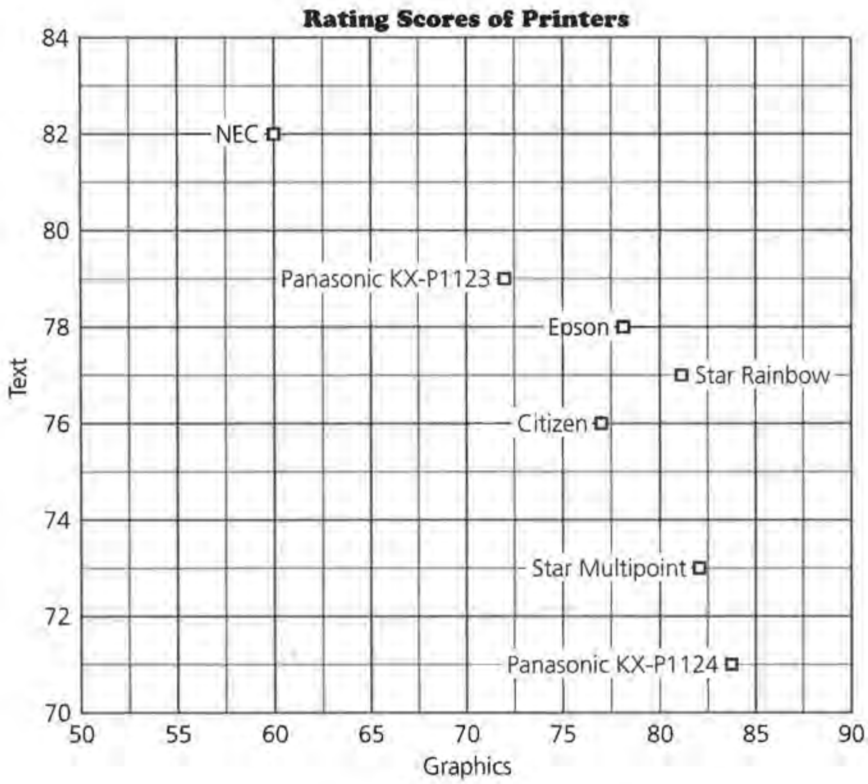


Figure 2.5

a. Jose thinks that graphics is the only thing that should matter when ranking printers. Which printer does he select as best?

b. Gina, on the other hand, thinks that text is all that should matter and that graphics should be ignored. Which printer does she select as best?

7. Which points in Figure 2.5 are not dominated by any other?

a. Explain why an argument could be made for any of these printers as the best-quality printer.

b. Explain why a reasonable argument cannot be made for any of the other printers as the best-quality printer.

In Lesson 1 you used weights to combine information on variables. There are two variables in the printer data, text quality and graphics quality—but when Jose ranks the printers, he gives all of the weight to graphics and no weight to text. His weights for (*graphics, text*) are [1 0]. Gina gives zero weight to graphics and all of the weight to text, so her weights are [0 1].

8. Jeff suggested the following compromise: use equal weights for the two variables.

a. Do you think this is a good compromise? Why or why not?

b. What weights will Jeff use?

c. Explain how the matrices below relate to Jeff's solution.

Data matrix · Weight matrix = Ratings matrix

$$\begin{bmatrix} 81 & 77 \\ 78 & 78 \\ 84 & 71 \\ 82 & 73 \\ 77 & 76 \\ 60 & 82 \\ 72 & 79 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \\ \\ \\ \\ \\ \\ \\ \end{bmatrix}$$

d. Compute the ratings for each printer in the matrix above. Which printer(s) has the highest rating?

When the weights are [1 1], the overall rating is Graphics score + Text score, or $G + T$. In matrix terms, this can be written as $[G\ T]\begin{bmatrix} 1 \\ 1 \end{bmatrix} = [R]$, or as the equation $G \cdot 1 + T \cdot 1 = R$.

To find the printer that is the best choice for Jeff, look at Figure 2.6.

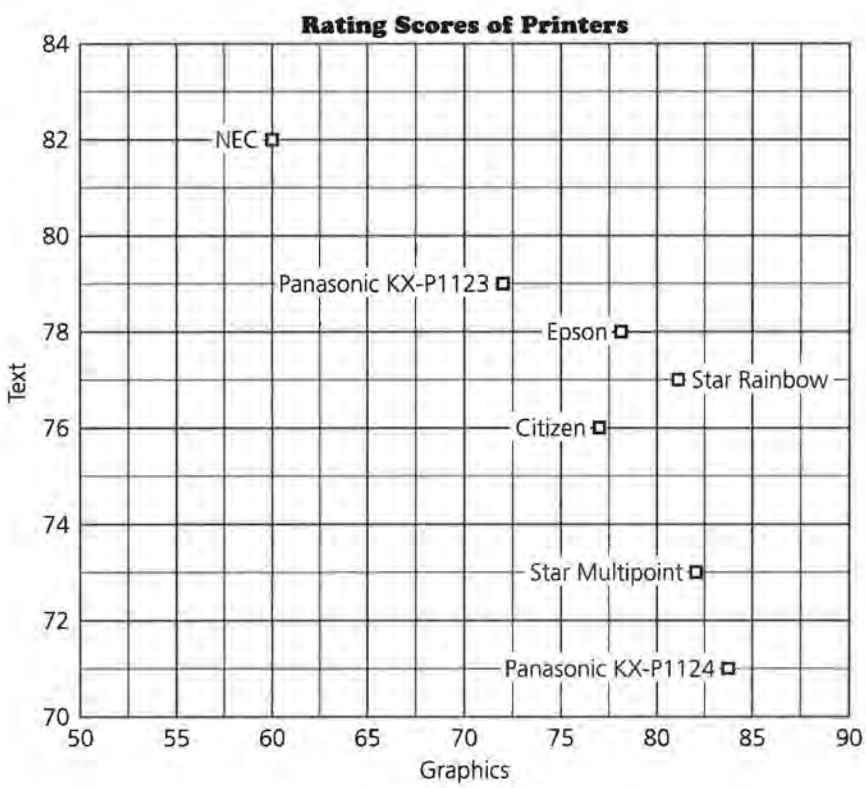


Figure 2.6

For the Citizen, the value of Text score + Graphics score is 153. Any other printer for which Text score + Graphics score is 153 would be just as appealing to Jeff as the Citizen printer—no more appealing and no less appealing. So all printers that fall on the line determined by $G + T = 153$ are equivalent from Jeff's point of view.

9. Graph the line $G + T = 153$ on *Activity Sheet 1*.

  a. What is the slope of this line?

  b. How many printers fall on that line?

10. Draw in the line determined by the rating for the Epson. Use *Activity Sheet 2*.

  a. What is the equation? the slope?

  b. What was the rating for the Star Rainbow?

  c. Where is the point for the Star Rainbow in terms of the line for the Epson?

  d. If a printer has a lower rating than 153, where will the point for that printer lie with respect to the line $G + T = 153$?

11. Figure 2.7 is the scatter plot with several lines added.

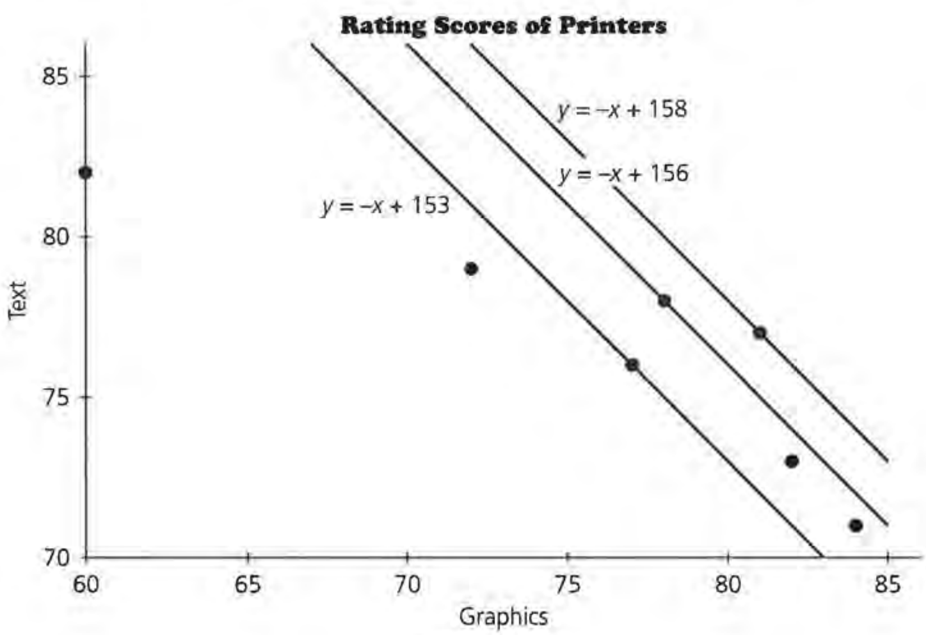


Figure 2.7

  a. Why does each line have the same slope?

  b. Comment on the statement:

Since Jeff wants $G + T$ to be large, he wants the printer that falls on the line having the greatest y-intercept.

c. Which printer will Jeff choose and why?

You could find Jeff's ideal printer by taking a *sweeping line* having slope −1, starting on the top right-hand part of the scatter plot beyond any data points, and "sweeping" down and to the left ("to the southwest"), keeping the slope of the line equal to −1, until the moving line touches one of the points as in Figure 2.8.

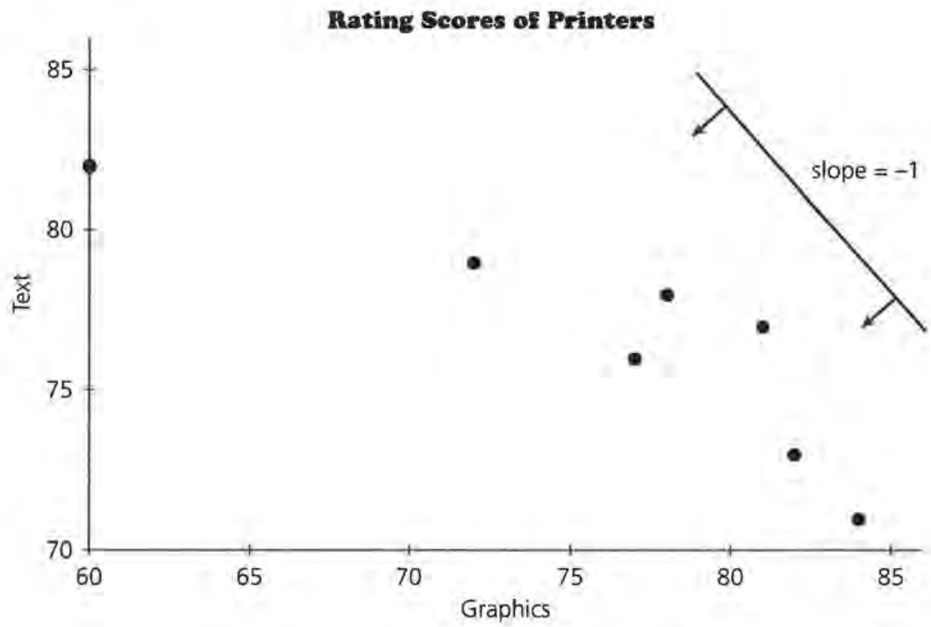


Figure 2.8

12. Why will "sweeping" down using a line with slope −1 find the best printer if the two variables are rated equally?

13. Suppose you had used weights $[\frac{1}{2}\ \ \frac{1}{2}]$rather than [1 1].

a. How would these weights have changed the overall rating for each printer?

b. Would the lines on the graph change? Explain.

c. Which printer would now have the highest rating?

Other Slopes

14. Emily is more concerned with graphics than with text, but text quality does matter to her. She wants to take a weighted average of text score and graphics score in which she gives twice as much weight to graphics as is given to text:

Emily's weighted average for a printer will be:

$R = \frac{2}{3} \cdot \text{graphics score} + \frac{1}{3} \cdot \text{text score.}$

a. Using the data matrix of

$$\begin{bmatrix} 81 & 77 \\ 78 & 78 \\ 84 & 71 \\ 82 & 73 \\ 77 & 76 \\ 60 & 82 \\ 72 & 79 \end{bmatrix}$$ and a weight matrix of $\begin{bmatrix} \frac{2}{3} \\ \frac{1}{3} \end{bmatrix}$, compute the rating for each printer.

b. Which printer(s) has the highest rating? Could you have predicted this before you did the calculations? If so, how?

Emily can take a line having slope –2, start at the upper right-hand part of the scatter plot, and sweep down and to the left ("to the southwest"), keeping the slope of the line equal to –2, until the moving line touches one of the points as in Figure 2.9.

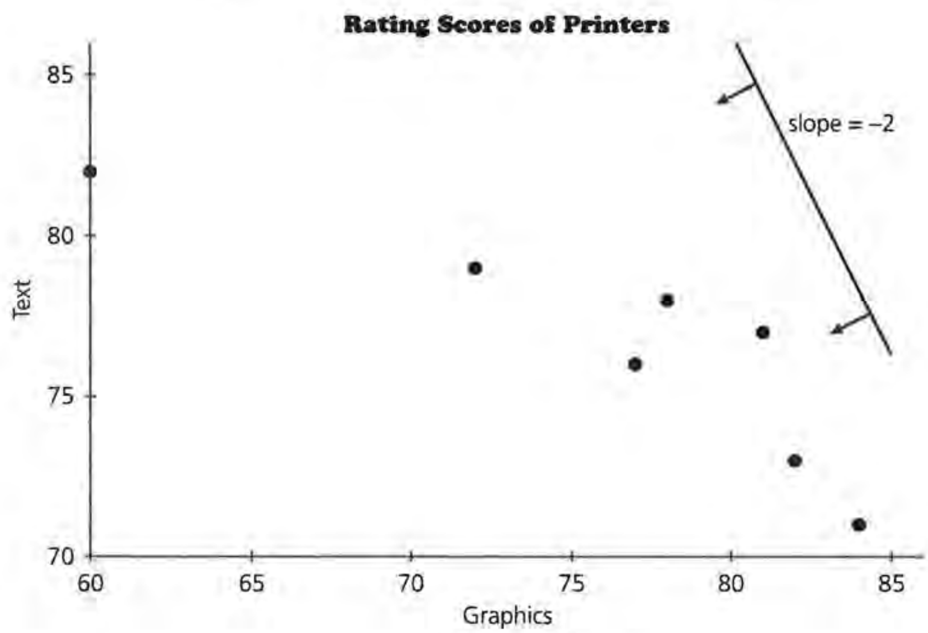


Figure 2.9

c. Verify that the slope of the equation for the sweeping line with these weights is –2.

d. Suppose Emily is considering a printer with ratings of 80 for graphics and 80 for text. What is the drop in text score that Emily would give up in order to get a one-point increase in graphics score:

$\frac{\text{change in } T}{\text{change in } G} = ?$

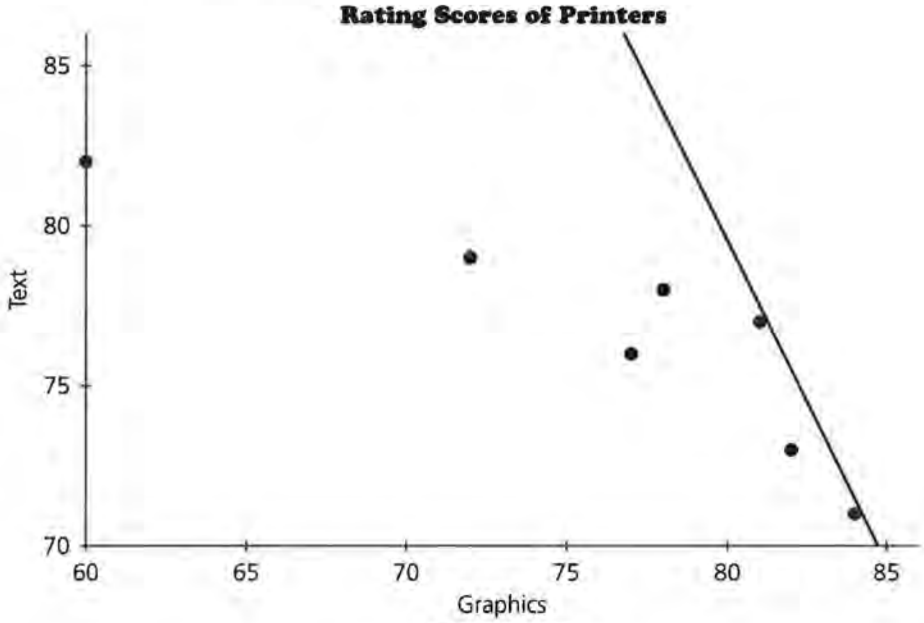


Figure 2.10

15. Consider the sweeping line in Figure 2.10.

a. What point(s) does the line touch first? What does this tell you?

b. What is the equation of this line?

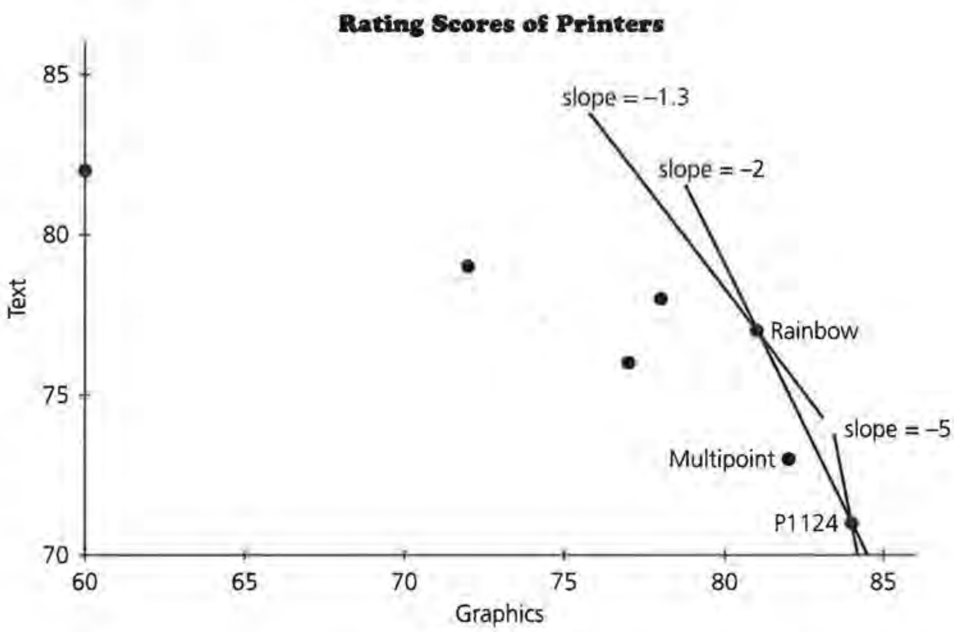


Figure 2.11

16. Figure 2.11 shows lines with several different slopes.

   a. What does it mean if a line has slope –5? –1.3? –0.75?

   b. What does it tell you if the slope of the preference line is zero?

   c. What does a steep "preference line" indicate?

   d. What would a vertical line indicate? What would the slope of the vertical line be?

   e. Is there a slope that will indicate a preference for the Multipoint? How can you tell?

Summary

If the weights on two variables, v_1 and v_2, are equal, the equation that can be used to generate the ratings is $1 \cdot v_1 + 1 \cdot v_2 = R$. For any value R, this equation will determine a straight line in terms of v_1, and v_2 with slope of –1. Using sets of parallel lines, all with slope –1, you can sweep down toward a plot of (v_1, v_2) from the upper right corner of your plot until you reach a point on the scatter plot. That point will have the maximum rating. If you change the weights, the slope of the line will also change, but the technique will still work.

Practice and Applications

17. Suppose you worked for the people who make the Epson printer. Then you might try to find a rating system combining text and graphics that would make the Epson appear to be the best printer. Think about the plot. Is there a pair of weights for text score and graphics score that make Epson the "winner"? How do you know?

Rating Scores of Printers

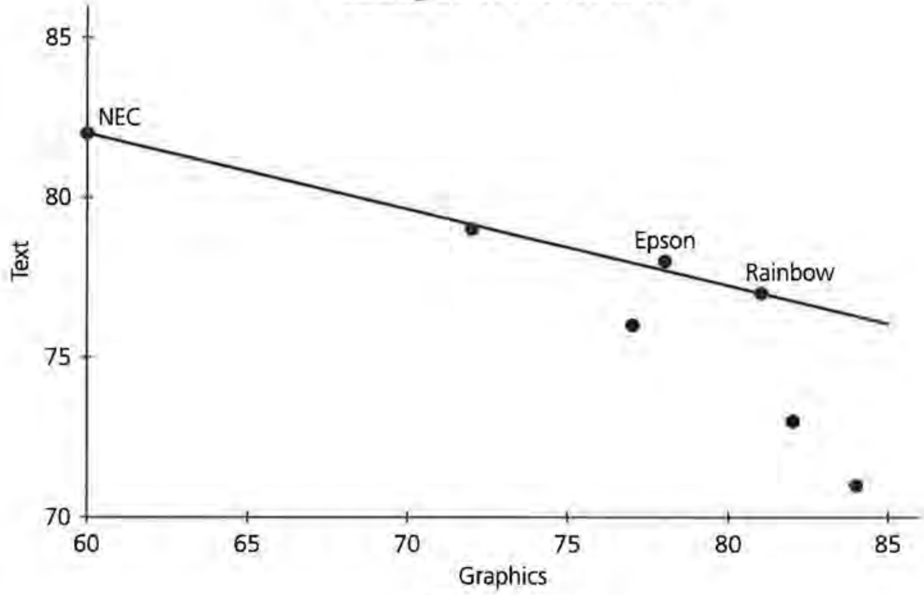


Figure 2.12

18. Look at the line in Figure 2.12 connecting the NEC and Rainbow points.

a. Where does the Epson fall in relation to the line through the NEC and the Rainbow points?

b. What is the slope of the line through the NEC and the Rainbow points?

c. Find a slope for which the Epson is the preferred printer.

d. What pair of weights correspond to the slope you found?

19. Is there a slope that will make the Panasonic KX-P1123 the preferred printer? Explain why or why not.

20. Each year the World Resources Institute releases information about the environment for areas in the United States (as well as the rest of the world). The plot in Figure 2.13 below shows the ratings for 14 of the largest cities in the United States. The ratings are based on studies done by the World Resources Institute. For example, air quality is based on daily readings for five pollutants regulated by the Clean Air Act: sulfur dioxide, nitrogen oxides, particulate matter, carbon monoxide, and ozone. The lower the number the better the rating.

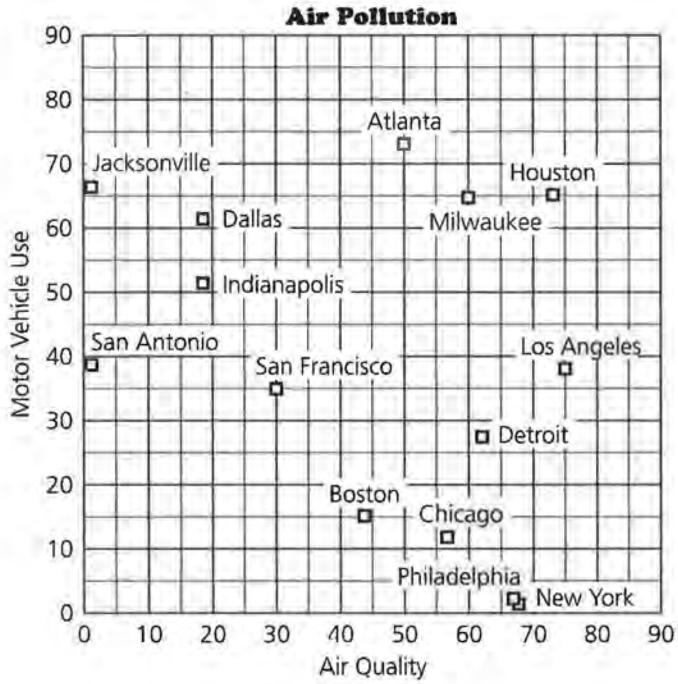


Figure 2.13

Source: *The 1994 Information Please Environment Almanac*

a. Smaller values are now better. How does that affect your interpretation of the plot? Which cities are not dominated by any others with respect to both variables?

b. Looking at the plot, which cities are likely candidates for the "worst" cities environmentally, if you judge only on the two variables in the plot?

c. If you weight air quality and motor vehicle use equally, find the line that will give you the worst (highest) rating. Explain how you found your answer.

d. Explain how you can use the sweeping line to determine the best (lowest) city.

21. The actual data from the environmental study are in the table below.

Environmental Study for Selected Cities

City	Unhealthy Air Quality	Toxic Releases	Energy Use	Motor Vehicle Use
Atlanta	50	31	20	73
Boston	44	17	49	15
Chicago	57	64	68	11
Dallas	19	33	29	61
Detroit	62	47	71	27
Houston	73	71	14	65
Indianapolis	19	46	56	52
Jacksonville	1	51	7	66
Milwaukee	60	52	74	64
Los Angeles	75	32	2	38
New York	68	27	42	1
Philadelphia	67	61	43	2
San Antonio	1	14	19	39
San Francisco	30	11	3	35

Source: *The 1994 Information Please Environment Almanac*

a. Make a scatter plot of (toxic releases, energy use). Which cities seem likely to be the worst cities based on these two variables if they are equally weighted?

b. Weight the toxic releases twice as much as the energy use, and then compute the ratings.

c. What is the slope of the line that will give you the worst city using the weights from part b? Draw in a series of lines with that slope and use them to verify your answer to b.

d. How can you tell from the plot if there are any cities that are tied in the ratings?

22. Examine the scatter plot.

a. Can you find a weighting that will make Houston the worst city? Justify your answer.

b. Can you find a weighting that will make Detroit the worst? Again, justify your answer.

More on Ranking

Who is the best quarterback? the leading tennis player?

What breed of dog is the most intelligent?

How do you combine successful operations with quality of staff and cost to produce a rating for a hospital or the number of touchdowns, interceptions, and completions to rate a quarterback?

OBJECTIVES

Rank variables measured in different units.

Relate ranking with unequal scales to an equation and to a scatter plot.

Sports figures are constantly being ranked as well as hospitals, movies, and schools: best quarterback, leading tennis player, top-ranked golfer. In each case there are many variables that could be considered important in determining ratings. The variables, however, are not always on the same scale. Some of them are percents, some might be in tens and some in thousands. Is there a way to determine weights that can be adjusted for these different scales? In this lesson you will learn some techniques to help you out when you work with two variables that are on different scales. You will also investigate how algebraic calculations using matrices relate to geometric interpretations of the ratings.

INVESTIGATE

Rating Baseball Players

Baseball players are elected to the Hall of Fame for a variety of reasons: pitching, hitting, home run hitting, and good all-around play. Which player in the Baseball Hall of Fame is the best career hitter? How can you rank some of the top hitters? The career batting average and runs batted in for several candidates are given in the table that follows.

Hall of Fame Best Hitters

Player	Batting Average (*BA*)	Runs Batted In (*RBI*)
Hank Aaron	.305	2,297
Rod Carew	.328	1,015
Ty Cobb	.367	1,961
Lou Gehrig	.340	1,990
Rogers Hornsby	.358	1,584
Reggie Jackson	.262	1,702
Mickey Mantle	.298	1,509
Willie Mays	.302	1,903
Stan Musial	.331	1,951
Babe Ruth	.342	2,211
Ted Williams	.344	1,839

Source: *Universal Almanac*, 1994

Discussion and Practice

1. Consider the data in the table above.

- **a.** Which aspects of a hitter's performance are treated as important here? Do you think these aspects are important? Why or why not?
- **b.** What other variables might be considered important?
- **c.** Does the number of seasons in the player's career directly affect batting average? Does it directly affect runs batted in? Explain.
- **d.** Suppose you ranked the players from 1 to 11 in each category and used the ranking to obtain a combined rating. Name an advantage and a disadvantage of finding the ratings this way.

2. Figure 3.1 is a scatter plot of the two variables.

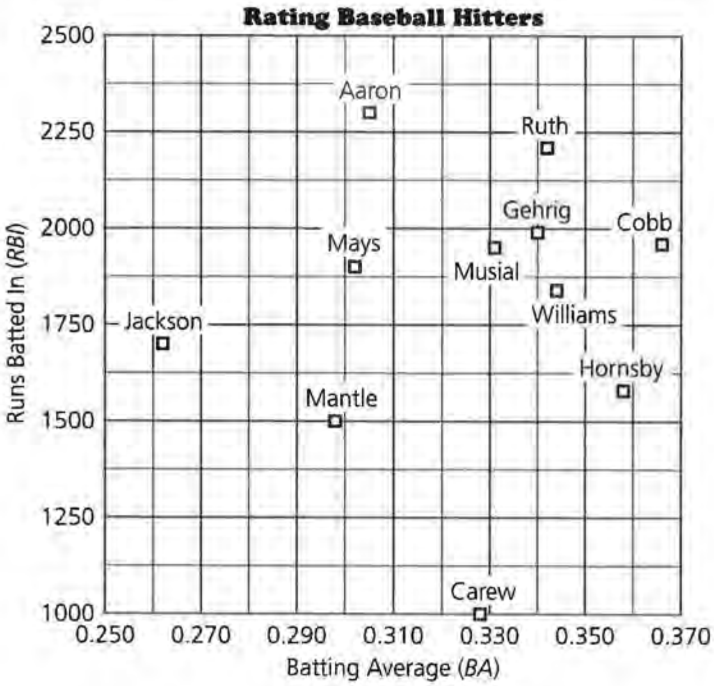


Figure 3.1

a. Is there a way to make Musial come out with the best rating using these variables? Explain.

b. For each of the variables, which of the eleven players seems to be the strongest?

c. For each of the variables, which of the eleven players seems to be the weakest?

d. Is there any player who completely dominates the others? Explain how you can tell.

3. Suppose you decide to use equal weights to ranking the players.

a. Explain how the matrices below relate to your work.

$$\begin{array}{ccccc} \text{Data Matrix} & & \text{Weight Matrix} & & \text{Ratings Matrix} \\ \begin{bmatrix} .305 & 2297 \\ .328 & 1015 \\ .367 & 1961 \\ .340 & 1990 \\ .358 & 1584 \\ .262 & 1702 \\ .298 & 1509 \\ .302 & 1903 \\ .331 & 1951 \\ .342 & 2211 \\ .344 & 1839 \end{bmatrix} & \cdot & \begin{bmatrix} 1 \\ 1 \end{bmatrix} & = & \begin{bmatrix} 2297.305 \\ 1015.328 \\ 1961.367 \\ 1990.340 \\ 1584.358 \\ 1702.262 \\ 1509.298 \\ 1903.302 \\ 1951.331 \\ 2211.342 \\ 1839.344 \end{bmatrix} \end{array}$$

b. Why are the ratings in the Ratings Matrix problematic?

There is a way to use weights to combine information in a problem such as this. You need to find weights that take into consideration the fact that RBIs are large numbers while batting averages are all less than 1.

In general, with a weight w_1 for BA and a weight w_2 for RBIs, a player's overall rating will be

$$BA \cdot w_1 + RBI \cdot w_2 = R.$$

In matrix terms, this can be written as $[BA \quad RBI] \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = [R]$.

Variables and Different Scales

Since BA and RBIs are in different units, here is one way to get a reasonable pair of weights. First, decide what score on one variable you would consider to be equivalent to a given score on the other variable. That is, these two scores have the same contribution to the overall rating. For example, you could use the minimum batting average of .262 as equivalent to the minimum number of runs batted in, 1015, for the 11 players; or you could consider the maximums in both cases to be equivalent, .367 and 2297. Suppose you decide to use the mean of the batting averages for all 121 of the Hall of Fame players listed in the 1995 *Universal Almanac*, .306, and decide it should have the same contribution to the overall rating as the mean of the RBIs earned by all of the Hall of Fame players, 1171. You can achieve this by choosing

$$w_1 = \frac{1}{.306} = 3.268, \text{ and } w_2 = \frac{1}{1171} = 0.000854.$$

Thus, a player with $BA = .306$ and $RBI = 1171$ would receive a rating of

$$\begin{aligned} w_1 \cdot BA + w_2 \cdot RBI &= 3.268 \cdot 0.306 + 0.000854 \cdot 1171 \\ &= 1 + 1 \\ &= 2. \end{aligned}$$

4. Consider the scores of .306 batting average and 1171 runs batted in.

a. Why might these scores be considered equivalent?

b. Think of other choices for batting average and the number runs batted in that you could consider to be equivalent.

5. Consider the equation with weights $w_1 = \frac{1}{.306}$ and $w_2 = \frac{1}{1171}$ that would give a player an overall rating of 2 if the player's batting average was .306 and had 1171 runs batted in.

a. What is the slope of the line represented by your equation?

b. Why will these two values for *BA* and *RBI* give equal contribution to the rating that uses the weights $\frac{1}{.306}$ and $\frac{1}{1171}$?

6. Consider this relationship in terms of Figure 3.2. The plot contains a graph of the equation from Question 5. *X* represents a player with (*BA*, *RBI*) = (.306, 1171).

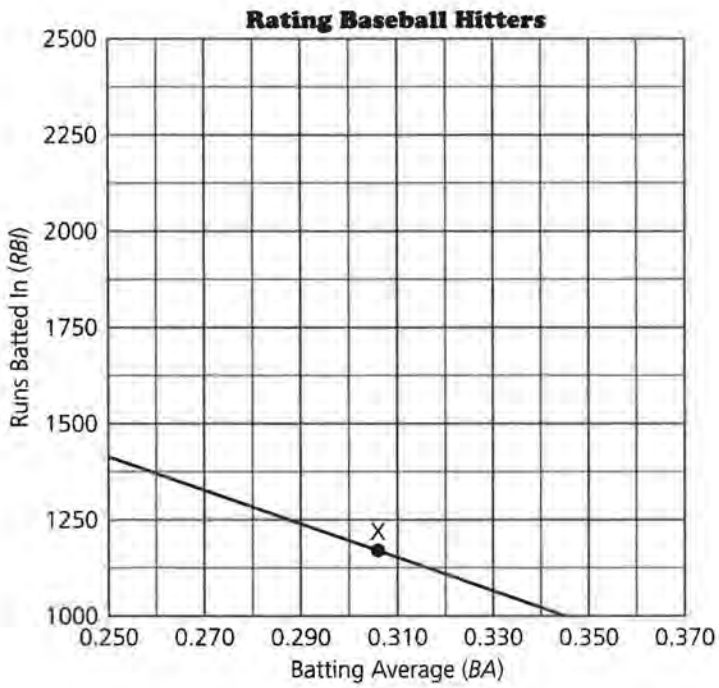


Figure 3.2

a. Find another point on the line and calculate the overall rating that a player would have with that combination of *BA* and *RBIs*.

b. Explain why any player with (*BA*, *RBI*) falling on the line would have the same rating.

c. If a player had a greater rating using the same weights, would the point (*BA*, *RBI*) lie above or below the line? Explain.

7. Find an equation using the weights [3.268 0.000854] that would represent all of the players who have a rating of 3.

   a. Draw that equation in the plot.

   b. How does this equation compare to the equation in problem 6?

8. Return to the original question. How do the 11 baseball players compare, using the weights $[w_1\ w_2] = [3.268\ 0.000854]$?

   a. Using matrices, compute the rating of all 11 players using these weights. That is, calculate

$$\begin{bmatrix} .305 & 2297 \\ .328 & 1015 \\ .367 & 1961 \\ .340 & 1990 \\ .358 & 1584 \\ .262 & 1702 \\ .298 & 1509 \\ .302 & 1903 \\ .331 & 1951 \\ .342 & 2211 \\ .344 & 1839 \end{bmatrix} \cdot \begin{bmatrix} 3.268 \\ 0.000854 \end{bmatrix}.$$

   b. Which player has the highest rating? Does this make sense in terms of the lines your drew in the plot above?

9. Now, consider the original scatter plot redrawn in Figure 3.3, with Henry Aaron displayed as A.

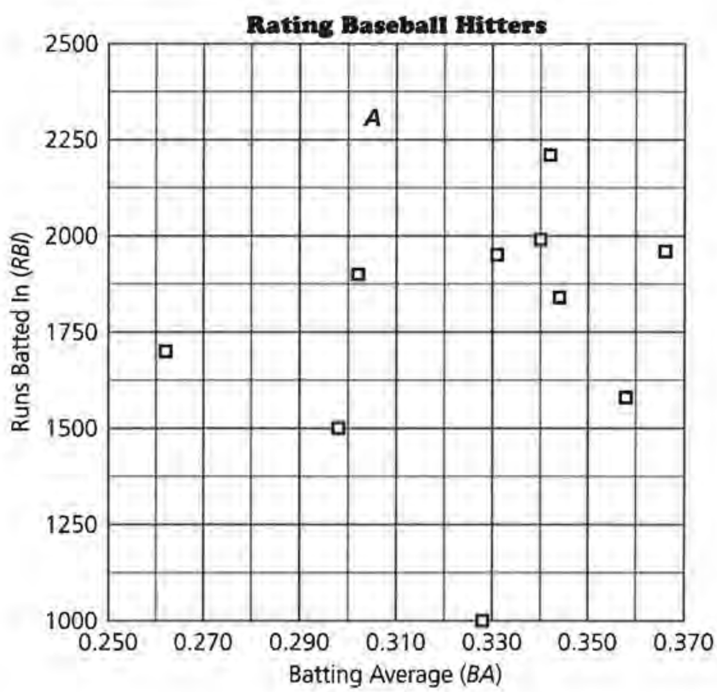


Figure 3.3

a. What is Henry Aaron's rating using these weights?

b. Create a *BA* and *RBI* for a player who would have the same rating as the rating Aaron got in part a but different scores for *BA* and *RBI*.

Find the equation for all *BA* and *RBI* that would tie Aaron. Graph the equation for this (*BA*, *RBI*) on *Activity Sheet 3*.

c. Does any other player have exactly the same rating as Aaron? How can you tell?

10. Repeat Problems 9a–9c for Reggie Jackson.

11. The rating for Willie Mays lies between the ratings for Reggie Jackson and Henry Aaron. Where does the point for Willie Mays lie in relation to the straight lines from Problems 9b and 10b?

12. Consider the straight line representing all points with rating equal to that of Aaron's. What property do all the points *below* this line have? Explain why.

13. Draw a "sweeping line" on *Activity Sheet 3*.

a. Using your scatter plot and a "sweeping line" such as that in Figure 3.4, explain how you could determine which player has the highest rating (using these weights) without actually doing the calculations.

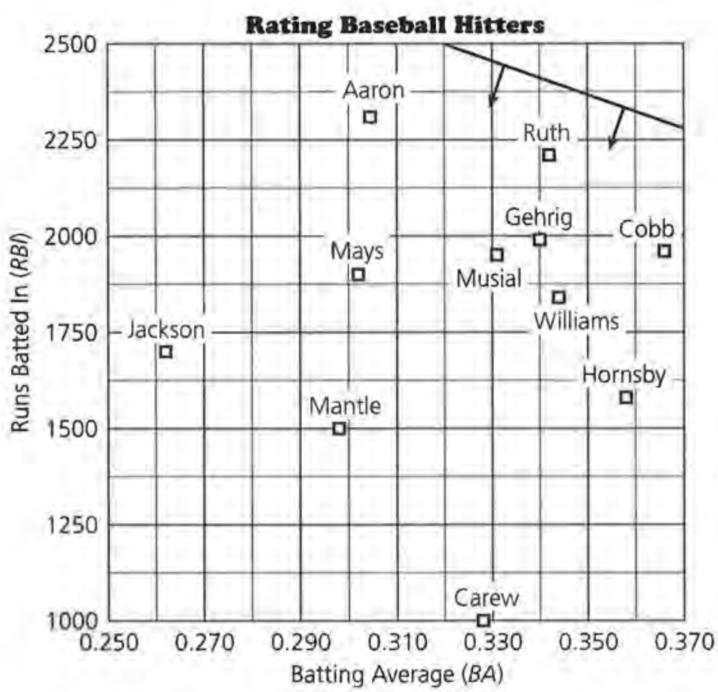


Figure 3.4

b. What do you think will happen in terms of the sweeping line if you change your weights and make the batting average worth twice as much as the number of runs batted in?

14. Is there a way to choose a set of weights that will make Ted Williams the top-ranked hitter? Explain why or why not.

15. Earlier you used the weights $\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 3.268 \\ 0.000854 \end{bmatrix}$ to rate baseball hitters. Suppose you multiply both weights by the same positive number, say 50, so the weights become $\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = 50 \cdot \begin{bmatrix} 3.268 \\ 0.000854 \end{bmatrix} = \begin{bmatrix} 163.400 \\ 0.0427 \end{bmatrix}$.

a. How will the new ratings for these players compare with the ratings you found above in problem 8?

b. Consider the straight lines on a scatter plot that represent points with the same rating using the weights 3.268 and 0.000854. How will these straight lines change when the weights are multiplied by 50?

c. How will the answers to parts a and b change if the multiplier is 187.4 instead of 50?

For the weights $[w_1\ w_2] = [3.268\ 0.000854]$, you found that all (BA, RBI) combinations that result in a rating of 2 are the solutions of the equation

$$2 = 3.268 \cdot BA + 0.000854 \cdot RBI.$$

16. Rewrite the equation so the number of runs batted in (RBI) is a function of the batting average, BA.

a. What is the slope of the line?

b. What is the y-intercept?

Consider the situation in general for any positive weights $[w_1\ w_2]$. The (BA, RBI) pairs that satisfy a given rating R for those weights determine the equation

$$R = w_1 \cdot BA + w_2 \cdot RBI.$$

Thus,

$$\begin{aligned} w_2 \cdot RBI &= R - w_1 \cdot BA \\ RBI &= \frac{R}{w_2} - \frac{w_1}{w_2} \cdot BA. \end{aligned}$$

17. In general, what is the slope of the line for (BA, RBI)?

a. Explain why the line representing a rating of 3.2 will be parallel to the line representing a rating of 2.5.

b. What can you conclude about all lines determined by a weighting of $[w_1\ w_2]$?

18. Suppose you have ratings using $[w_1\ w_2]$ but think that *RBI*s deserve more weight in the overall rating. You decide to rate the hitters using the weights $[w_1\ 3w_2]$.

a. Can the new ratings be determined from the previous ratings, or do you have to use the original (BA, RBI) data?

b. How will the straight line for a constant rating of 2 from the new weights compare with the straight line for a constant rating of 2 from the previous weights?

c. Consider the plot in Figure 3.5 showing two players with equal rating according to the original weights.

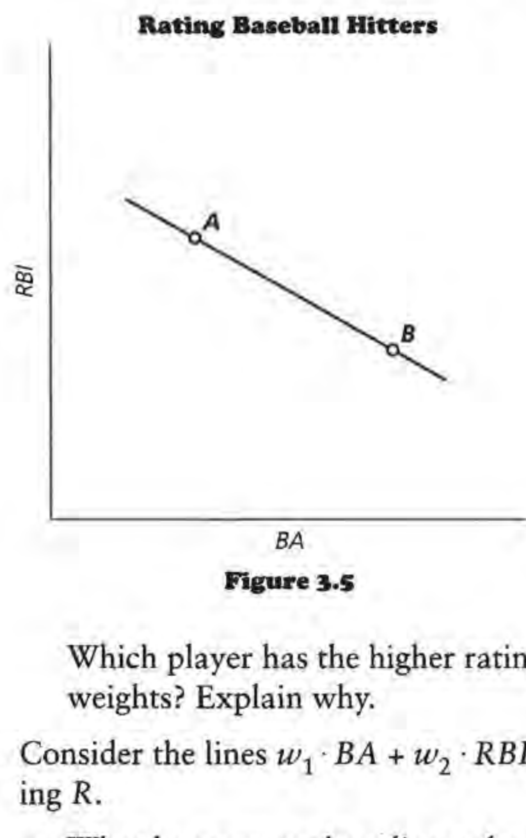


Figure 3.5

Which player has the higher rating using the new weights? Explain why.

19. Consider the lines $w_1 \cdot BA + w_2 \cdot RBI = R$ for a constant rating R.

a. What happens to these lines when $w_2 = 0$, and when $w_1 = 0$?

b. Describe how the players are rated and rank-ordered when $w_2 = 0$, and when $w_1 = 0$.

20. Different weights sometimes result in equivalent rankings.

a. For what c will the weights $[w_1\ w_2]$ yield equivalent rankings to those from the weights $[1\ c]$? Explain your reasoning.

b. For any positive weights w_1 and w_2, what is the only property of these two numbers that has a real effect on the ratings and rank-ordering of the hitters? Why?

You can see by examining the graph that neither Henry Aaron nor Babe Ruth dominate each other's career-hitting performance in terms of *BA* and *RBI*. Given a pair of weights for these variables, you can rate these hitters and rank them. The next question to investigate is the following: Is there some pair of weights that will give Aaron and Ruth equal ratings?

21. Suppose some pair of weights gives Aaron and Ruth equal ratings.

a. If so, what can you say about the line of constant rating that passes through Aaron's point?

b. The (*BA*, *RBI*) data for Aaron is (.305, 2297) and for Ruth is (.342, 2211). What is the slope of the line through these two points?

c. Recall that, for weights $[w_1\ w_2]$, the slopes of the lines of constant rating are all equal to $-\frac{w_1}{w_2}$. Find a pair of weights $[w_1\ w_2]$ that should give Aaron and Ruth equal rating.

d. Verify that the two players have equal ratings using your weights $[w_1\ w_2]$ from part c.

Summary

If you have variables that are not in the same scale, determine what might be typical for each of the scales. Use the reciprocal of each of the typical values to define weights that would have equal value in your rating scheme. In general the equation you work with will be $w_1 \cdot y + w_2 \cdot x = R$.

You can find the ratings using either a system of matrices $[x\ \ y]\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = [R]$ or by generating a sweeping line. If you work with the sweeping line, the slope of the line will be determined by your weights. It is possible to change the weights for your variables and still use either the matrix system or the sweeping line to find the rankings. If the slope of a line is $-\frac{a}{b}$, you can use the weights [a b] to determine your ratings.

Practice and Applications

Cars

22. The table below shows horsepower and gas mileage (EPA highway miles per gallon) for each of seven cars. A scatter plot is given in Figure 3.6.

Cars HP and MPG

Car	HP	MPG
Toyota Camry	130	27
Nissan NX200	140	30
Saturn SC	124	33
Toyota Paseo	108	34
Mazda MX-3	130	28
Honda Accord	140	29
Honda Prelude	160	26

Source: *Consumer Reports*, March 1992, July 1992, January 1993.

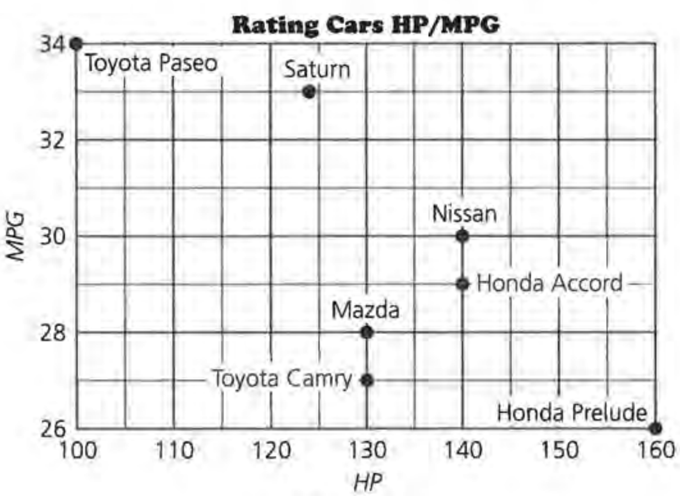


Figure 3.6

a. Describe the trend in the scatter plot of (HP, MPG).

b. Find a set of weights that will give equal value to both the miles per gallon and the horsepower. Calculate the ratings of each of the cars using those weights.

c. Explain how you could use a sweeping line to select the highest rated car.

23. For some of the cars there is a set of weights for which that car has the highest rating.

a. Use the scatter plot to determine which cars those are.

b. Find a set of weights for which the Saturn SC has the greatest weighted average.

24. The average number of inches of rainfall and the average relative humidity in the afternoon (in percent) for selected cities in the United States are contained in the table below and graphed in Figure 3.7.

Average Rainfall/Humidity

City	Yearly Average Number of Precipitation Days	Yearly Average Relative P.M. Humidity (Percent)
Mobile, AL	122	57
Phoenix, AZ	36	23
San Diego, CA	42	62
Boise, ID	90	43
Chicago, IL	126	60
Boston, MA	126	58
Omaha, NB	99	59
Albuquerque, NM	61	59
Columbus, OH	137	62
Seattle, WA	155	29
Salt Lake City, UT	91	43

Source: *American Almanac,* 1994–95

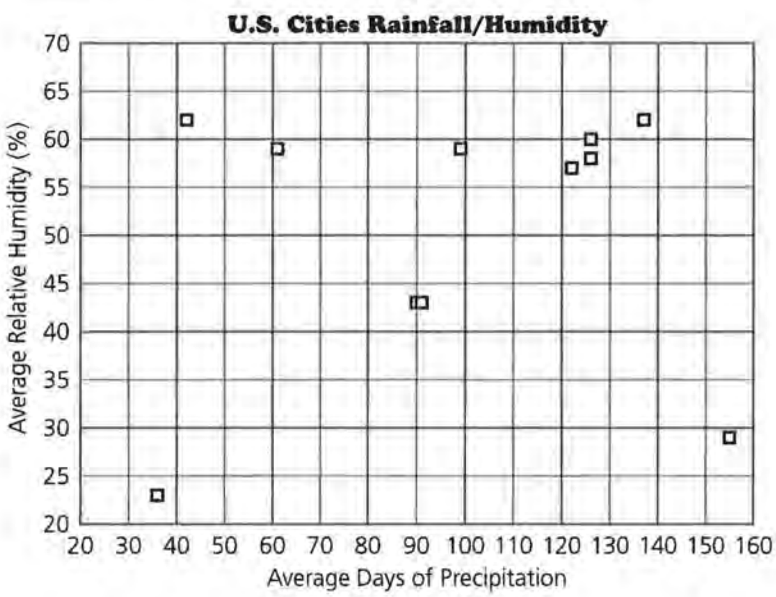


Figure 3.7

a. Which of the cities seems to have the least rain and the lowest humidity? the most?

b. Suppose the slope of the sweeping line was –1.5. Draw in some of the lines using *Activity Sheet 4*. Which of the cities would have the highest rating? What does this rating tell you?

c. Suppose the slope of the sweeping line was –1.5. What does this tell you about the weights for the two variables?

25. Find two weights that you think would make the variables have equal contribution.

a. Find the equation you could use to determine the ratings.

b. Draw in the sweeping lines on the second plot on *Activity Sheet 4* and determine which city would have the highest rating.

c. Calculate the ratings using matrices. How do your results compare to those you found with the sweeping line?

d. Are there any weights that would make the ratings tied for two of the cities? How can you tell?

Extension

26. Return to the top hitters in the Baseball Hall of Fame. Use your weights from problem 21, part c, to answer the following:

a. Suppose a player has a career *BA* of .305. What number of *RBI*s will result in *BA* and *RBI* having equal contribution to this player's overall rating?

b. If a second player has *BA* of .342, what number of *RBI*s will result in *BA* and *RBI* having equal contribution to this player's overall rating?

c. Now recall the weights you found in problem 21, part c. Would you conclude that these weights are rather reasonable, or that *BA* is weighted too heavily , or that *BA* is not weighted heavily enough? Explain your reasoning.

LESSON 4

Ratings with Three or More Variables

Baseball statistics are kept on many variables: batting average, home runs, hits, number of games, runs batted in. How do you use all of the data to rate the players?

Just as cities, hospitals, and airlines are rated on many variables, so are colleges. How can all of the variables be used to find a rating?

When you rate something you usually have many variables to consider. In Lessons 2 and 3 you investigated rating formulas with two variables and the corresponding geometric representation in the two-dimensional coordinate plane. When you use three variables, you can find an algebraic formula much as you did before, but now the geometric representation is in three dimensions. In this lesson you will find weighted rating formulas for three or more variables, learn to plot points in 3-D, write the equation of a plane and of a line in 3-D, and think about the geometry of finding weighted ratings for three variables.

OBJECTIVES

Generalize the approach to the ratings problems using matrices and algebra.

Investigate algebraic representation of a plane.

INVESTIGATE

Baseball Statistics

Consider once again the Hall of Fame baseball players from Lesson 3. Suppose you knew the number of home runs hit by each of the players in addition to the information about batting average and runs batted in.

Batting Average, Runs Batted In, Home Runs

Player	Batting Average (*BA*)	Runs Batted In (*RBI*)	Home Runs (*HR*)
Hank Aaron	.305	2,297	755
Rod Carew	.328	1,015	92
Ty Cobb	.367	1,961	118
Lou Gehrig	.340	1,990	493
Rogers Hornsby	.358	1,584	301
Reggie Jackson	.262	1,702	563
Mickey Mantle	.298	1,509	536
Willie Mays	.302	1,903	660
Stan Musial	.331	1,951	475
Babe Ruth	.342	2,211	714
Ted Williams	.344	1,839	521

Source: *Universal Almanac*, 1994

A plot of the three variables for each player would be a three-dimensional plot such as the one shown below.

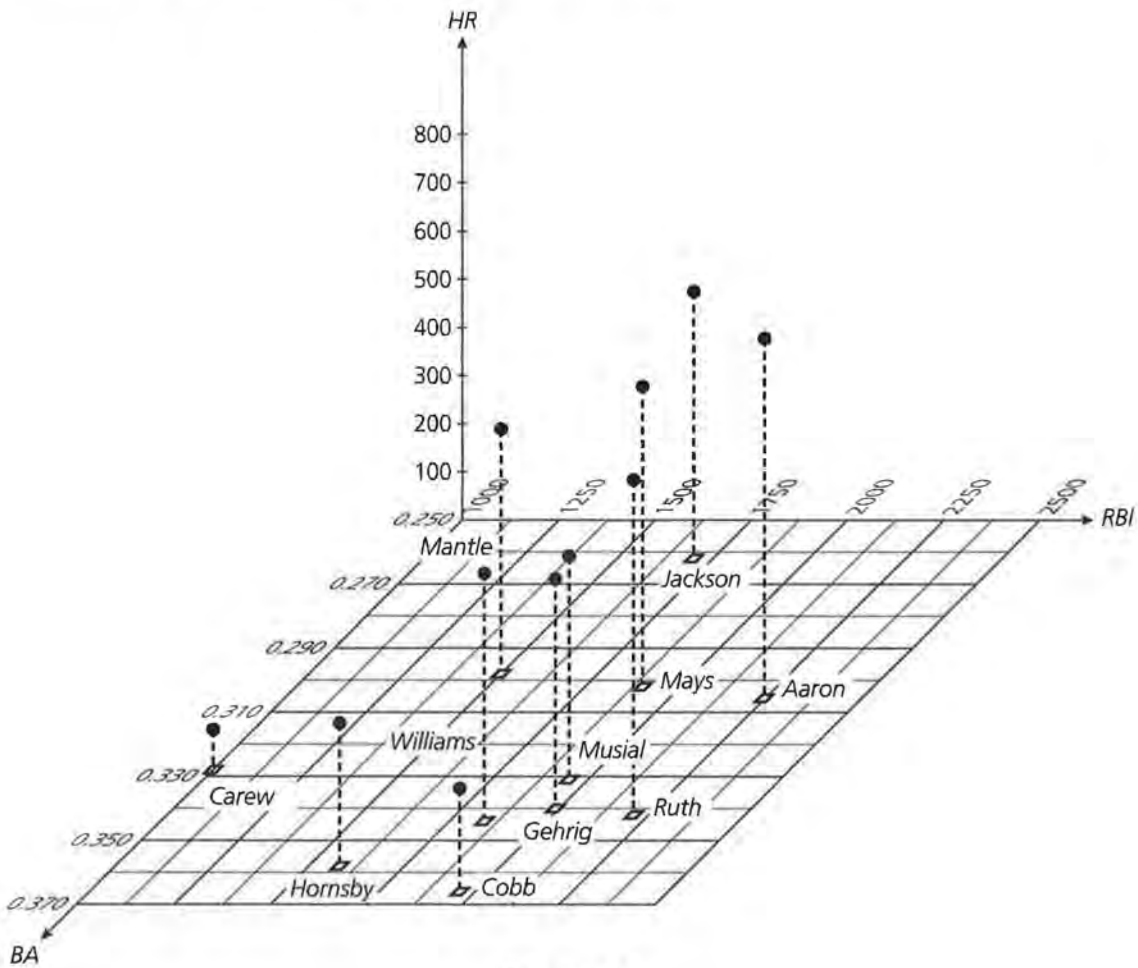


Figure 4.1

Discussion and Practice

1. Study the plot carefully.

   a. Is there one player who dominates with respect to home runs? Where is that player in the plot?

   b. Is there one player who dominates with respect to two variables? If so, where is that player in the plot?

The algebraic approach for two variables began with identifying some sort of sensible weight to make the variables approximately equivalent or some sort of standard to use as a baseline weight for each variable. For the Hall of Fame players, you used the means of *BA* (.306) and *RBI* (1171) as weights to find a general equation:

$R = w_1\ BA + w_2\ RBI$
$R = \frac{1}{.306}\ BA + \frac{1}{1171}\ RBI$ or
$R = 3.268\ BA + 0.000854\ RBI$

If a third variable, home runs, is included in the analysis, it seems reasonable to use the mean of the home runs for all of the players in the Hall of Fame as the baseline for the weight for *HR*s. Using this mean of 190 and extending the equation for two variables, you get

$R = w_1\ BA + w_2\ RBI + w_3\ HR$
$R = \frac{1}{.306}\ BA + \frac{1}{1171}\ RBI + \frac{1}{190}\ HR$ or
$R = 3.268\ BA + 0.000854\ RBI + 0.00526\ HR.$

2. Is it possible to have more than one person with the same *R* but with different batting averages, number of runs batted in, and number of home runs? Explain your answer and give an example.

3. Consider the ordered triple (*BA*, *RBI*, *HR*) that belongs to Henry Aaron: (.305, 2297, 755). What rating would the formula give for Aaron?

Where on the plot would you find the players who have the same rating as Aaron? To answer, consider a simple equation: $y = 6$. In one dimension, this equation will generate a point. In two dimensions, it will generate a set of points where x is any value, and y is always 6. In three dimensions, it will generate a set of points where x is any value, z is any value, and y is always 6.

4. Consider the equation $y = 6$ in two dimensions and in three dimensions.

   a. What geometric figure is represented in a two-dimensional coordinate plane by $\{(x, y) \mid y = 6\}$?

   b. What figure is represented in a three-dimensional coordinate system by $\{(x, y, 0) \mid y = 6\}$ pictured in the diagram below?

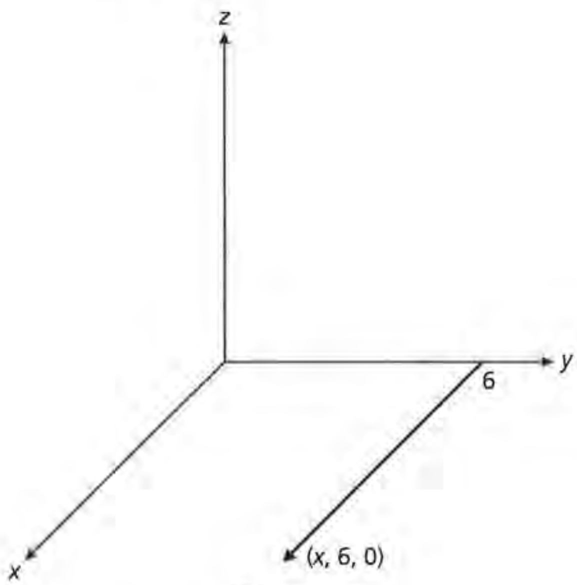


Figure 4.2

   c. What figure is represented in a three-dimensional coordinate system by $\{(x, y, 3) \mid y = 6\}$ pictured in the diagram below?

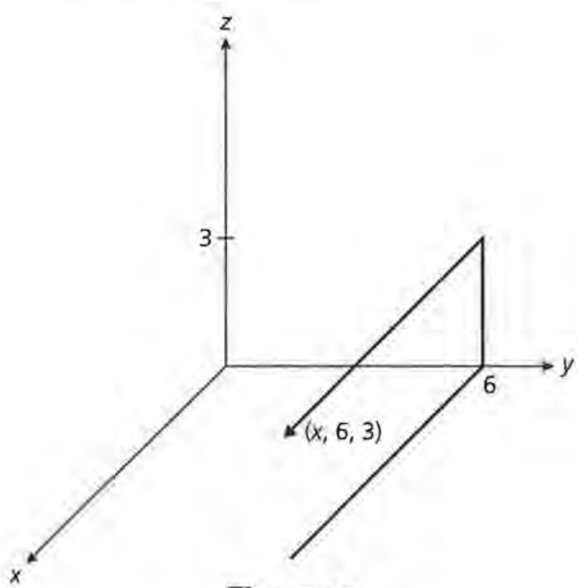


Figure 4.3

d. What figure is represented in a three-dimensional coordinate system by $\{(x, y, z) \mid y = 6\}$?

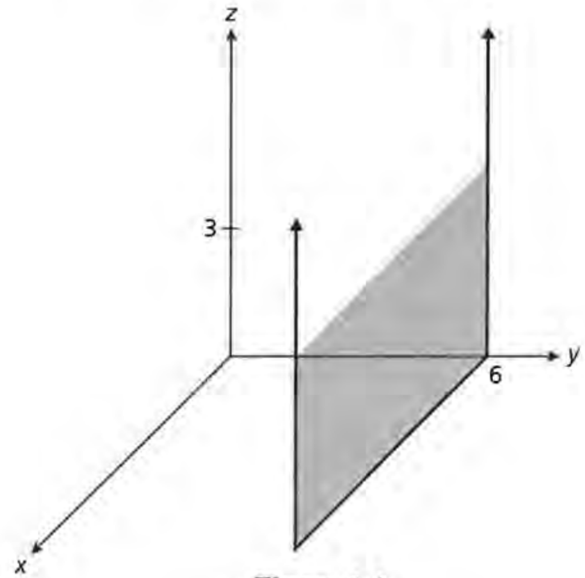


Figure 4.4

5. Consider the statement: For some a and b not both equal to zero, the graph of $ax + by = c$ in three dimensions will be a plane parallel to the z-axis. Use the equation $x + 3y = 6$ to help decide whether the statement is true or false.

6. A trace of a plane is the line in which the plane intersects one of the coordinate planes (the xy-plane, the yz-plane, or the xz-plane). To draw a picture of a plane, you can often draw the traces to show how the plane would look in the first octant.

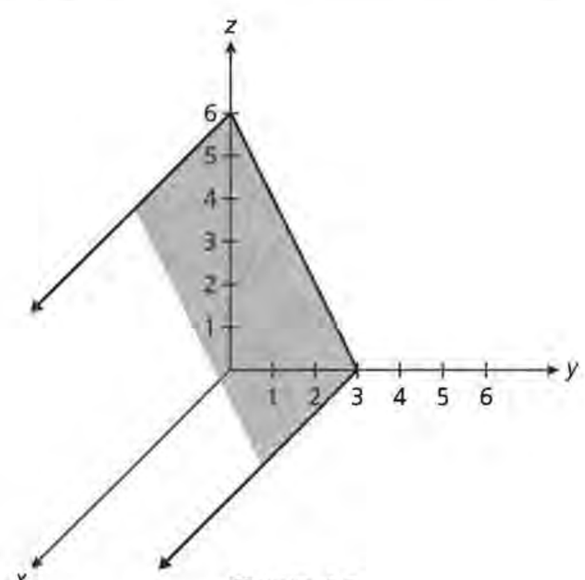


Figure 4.5

a. What is the equation for each of the three traces in the figure drawn above?

Sketch the traces for each of the equations to see how the plane would look in the first octant:

b. $3x + 5y = 10$

c. $x + 4z = 8$

d. $2y + 5z = 10$

7. The traces of the equation $x + 2y + 3z = 6$ are shown in the drawing below. Write an equation for each of the traces and explain their origin.

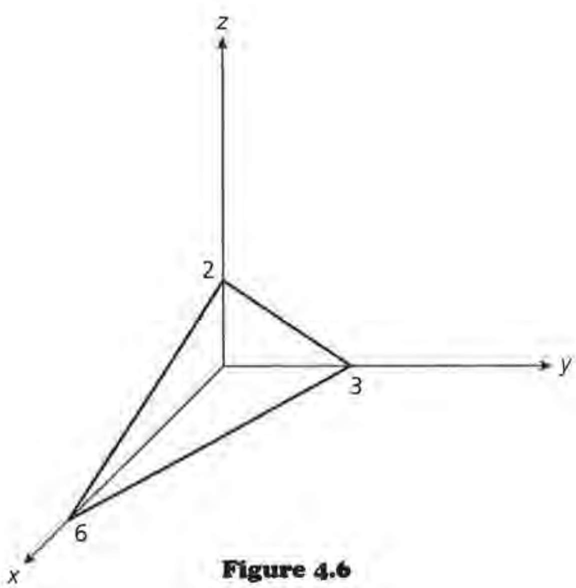


Figure 4.6

To answer the question posed earlier, all of the points that satisfy the equation determined by the weights [3.268 0.000854 0.00526] and Aaron's point (.305, 2297, 755) will be on a *plane* in three-dimensional space. This means that all players who have the same rating of 6.93 for the given weights will lie on that plane. The equation of that plane is

$$3.268\ BA + 0.000854\ RBI + 0.00526\ HR = 6.93.$$

8. Find a possible (BA, RBI, HR) for a player who had the same total rating as Aaron.

a. Explain how you know they have the same rating.

b. Write the equations of the traces of the plane representing all of the players with a rating of 6.93. Make a rough sketch of the plane using the traces.

9. To see how the other players fit into the picture, use the same weights but a new player, say Willie Mays.

   a. Write the equation that will give the rating for Willie Mays and determine the rating.

   b. Where is the set of players who, with these weights, will have the same rating as Willie Mays?

The next question to consider is how the plane containing Willie Mays is related to the plane containing Henry Aaron.

10. Describe the possible geometric relationships between any two planes.

11. Consider the plane $x + 2y + 3z = 6$ and the plane $x + 2y + 3z = 12$. The trace for each of the planes is drawn in Figure 4.7.

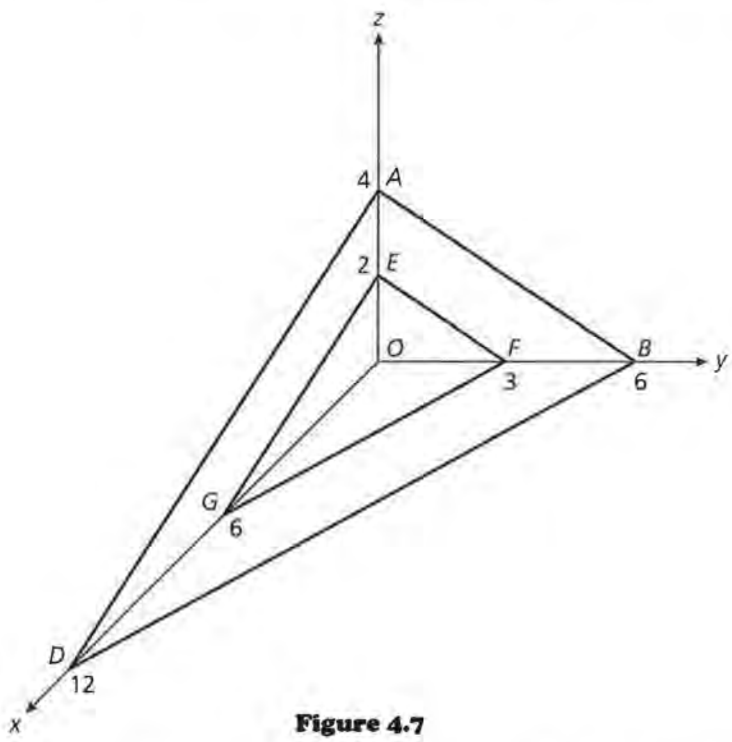


Figure 4.7

   a. Consider the xy-plane. How is the line containing GF related to the line containing DB? Explain how you know.

   b. Consider the yz-plane. How is the line containing EF related to the line containing AB? Explain how you made your conclusion.

   c. What conclusion can you make about the plane ADB and the plane GEF? Explain your reasoning.

d. Suppose you know that $\angle EGO$ is congruent to $\angle ADO$ and that $\angle EFO$ is congruent to $\angle ABO$. What conclusions can you make?

12. Now consider the general plane $w_1 x + w_2 y + w_3 z = R$ for some given R. The intercepts can be written as $(\frac{R}{w_1}, 0, 0)$, $(0, \frac{R}{w_2}, 0)$, $(0, 0, \frac{R}{w_3})$.

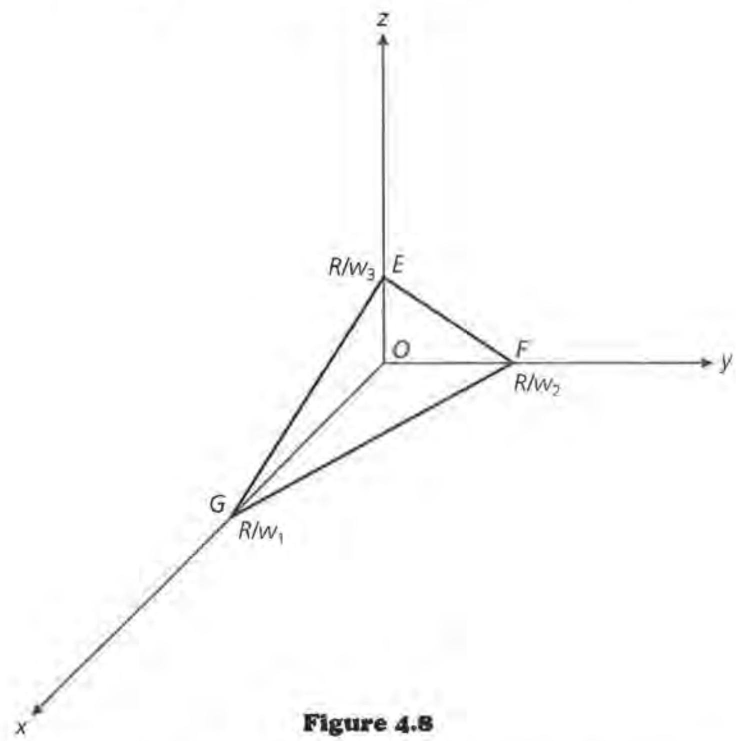


Figure 4.8

a. Find a trigonometric relationship between $\angle EGO$ and the intercepts. Do the same thing for $\angle GEO$, $\angle EFO$, $\angle FEO$, $\angle GFO$, and $\angle FGO$.

b. Find the angles that the traces of the plane make with each of the axes if the plane is given by $x + 5y + z = 8$.

c. Find the angles for a plane given by $2x + 10y + 2z = 8$. What conclusion can you make about this plane and the plane from b?

13. Consider the planes $x + 2y + 4z = 8$ and $2x + y + 4z = 12$.

a. Make a trace of each plane.

b. Find the angle each trace makes with each coordinate axis.

c. Are the two planes parallel or not? Justify your answer.

14. It seems reasonable to make the following conjecture: If the corresponding angles made by the traces of two different planes and the axes are congruent, the planes are parallel. Construct a proof of the conjecture.

15. Return to the planes containing Aaron and Mays. Sketch the three traces for each of the equations on the same axes.

a. How are these two planes related?

b. Describe the angles made by the traces of the planes for Mays and for Aaron with the *BA*, *RBI* coordinate plane.

c. Is the plane containing Mays above or below the plane for Aaron? How can you tell?

16. Suppose one player using the same weights had a rating of 10.

a. Write the equation of the plane.

b. Is the plane parallel to the plane for Aaron and Mays? How can you tell?

c. Is the plane for Aaron above or below this plane? How do you know?

17. Now consider a plane such that the traces form angles to each axis determined by the weights $[w_1 \ \ w_2 \ \ w_3]$. Begin sweeping the plane from afar toward the origin until you touch one of the points in the space.

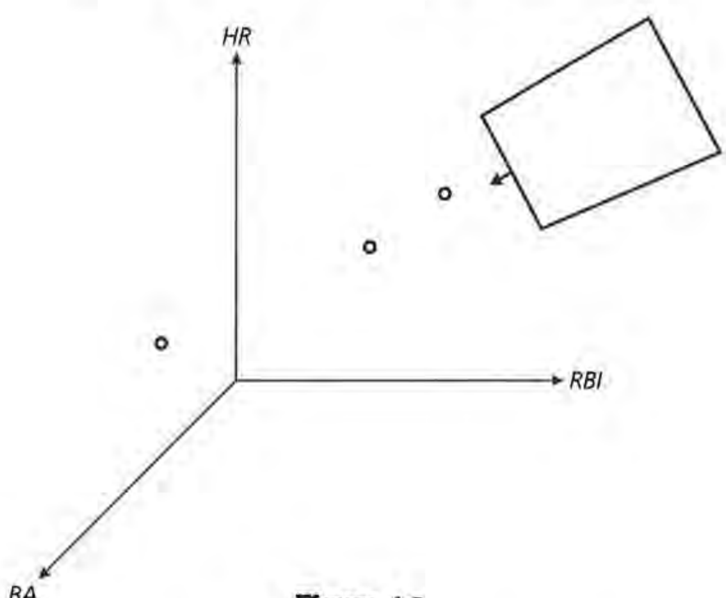


Figure 4.9

a. How will the sweeping plane determined by those weights determine the best player with respect to the three variables?

b. Using the agreed-on weights from problem 1, who is the best player among the eleven listed at the beginning of the lesson?

c. How do the other players rank? Explain what this means in terms of the sweeping plane.

18. Suppose you felt that batting average should be three times as important as the other variables.

a. Write an equation of the possible planes for those weights.

b. What angle will the traces of those planes in the *BA*-*HR*-coordinate plane make with the *BA*-axis?

c. Find the equation of the plane that would contain Henry Aaron and sketch its traces.

d. How will the plane that contains Henry Aaron for these weights compare to the planes that used the original weights?

19. Using the new weights, find the ratings of each player.

a. Determine the ranking based on these ratings.

b. Is it possible to find a set of weights that will enable Willie Mays to be in first place? Explain.

Four or More Variables

A more complete set of data for the Hall of Fame players consists of the following.

Baseball Hall of Fame

Player	Games	At Bats	Hits	Batting Average (*BA*)	Runs Batted In (*RBI*)	Home Runs
Hank Aaron	3,298	12,364	3,771	.305	2,297	755
Rod Carew	2,469	9,315	3,053	.328	1,015	92
Ty Cobb	3,034	11,429	4,191	.367	1,961	118
Lou Gehrig	2,164	8,001	2,721	.340	1,990	493
Rogers Hornsby	2,259	8,173	2,930	.358	1,584	301
Reggie Jackson	2,820	9,864	2,584	.262	1,702	563
Mickey Mantle	2,401	8,102	2,415	.298	1,509	536
Willie Mays	2,992	10,881	3,283	.302	1,903	660
Stan Musial	3,026	10,972	3,630	.331	1,951	475
Babe Ruth	2,503	8,399	2,873	.342	2,211	714
Ted Williams	2,292	7,706	2,654	.344	1,839	521

Source: *Universal Almanac*, 1994

20. The number of hits divided by the number of at bats will give the batting average. The only variable that is not really already involved in the formula is the number of games. Again, you want to have a standard for the number of games. Because all of the other data are based on means, it makes sense to continue to do so. The mean number of games for all the players in the Hall of Fame is 2030. Adding in this variable in the same way that the third variable was included would produce an equation for finding the total ranking of the form

$3.268\,BA + 0.000854\,RBI + 0.00526\,HR + 0.000493G = R.$

a. What total ranking would this produce for Aaron?

b. Use the new equation to find a ranking for each of the players.

c. How does the ranking with four variables compare to that with three?

d. Carl Yastrzemski, elected to the Baseball Hall of Fame in 1989, had the statistics shown below. How does Yastrzemski compare to the other players you have been ranking?

Carl Yastrzemski Stats.

Player	Games	At Bats	Hits	Batting Average (*BA*)	Runs Batted In (*RBI*)	Home Runs
Yastrzemski	3308	11,988	3,419	.285	1,844	452

21. What procedure could you use if you had five variables?

The geometry for four variables is impossible to sketch in a two-dimensional plane. The weights form a vector in four dimensions, and the point for each player is projected onto a scalar multiple of this vector. The player whose projection creates the longest vector is rated the "best" according to the given weights. An extension exercise in the Practice and Applications demonstrates how this works in two dimensions.

Summary

In this lesson you learned that an equation with three variables was the equation of a plane. The line in which a given plane intersects a coordinate plane is called a *trace of the plane*. If you considered equations of the form $w_1x + w_2y + w_3z = R$ for different values for R, you would produce a series of planes, parallel to each other. The intercepts of a plane are the points

$(\frac{r}{w_1}, 0, 0)$, $(0, \frac{r}{w_2}, 0)$, $(0, 0, \frac{r}{w_3})$, and the tangent of the angle a trace makes with an axis can be expressed in terms of the intercepts of the plane:

$$\tan A = \frac{\frac{R}{w_2}}{\frac{R}{w_1}} = \frac{w_1}{w_2}.$$

Two planes are parallel if the angles they make with the axes are congruent. You can use weights determined by some standard to establish a formula to rank something using as many variables as you choose, although you can only see the geometric interpretation in two and three dimensions.

Practice and Applications

22. Are the following planes parallel? Why or why not?

a. $x + y + z = 10$ and $x + y + z = 5$

b. $x + 2y + 4z = 8$ and $3x + 6y + 12z = 24$

c. $6x + 9y + 15z = 8$ and $4x + 6y + 10z = 10$

d. $x + 2z = 8$ and $3x - z = 8$

23. In *Places Rated Almanac*, the data for rating cities is collected for a variety of categories, then put into a formula to determine a score. The scores are then ranked in order from best to worst and the ranks are used to determine the composite rating. The table below contains the scores that cities earned in the Arts, Education, and Recreation. In each case, the higher the number the better the score.

Rating Cities: Arts, Education, Recreation

City	Arts	Education	Recreation
Ann Arbor, MI	944	1461	1480
Austin, TX	560	1730	1611
Jacksonville, FL	512	440	2642
Minneapolis, MN	1660	2212	2273
Portland, OR	597	944	2851
Richmond, VA	690	269	1130
San Jose, CA	1685	1854	1430
Santa Fe, NM	501	60	2294

Source: *Places Rated Almanac*, 1993

a. Make a three-dimensional sketch of the scores. Does any city dominate the others? Explain.

b. Describe a method to make the scores equivalent. Following is some information you might find useful, based on all cities in the *Places Rated Almanac.*

For Arts, the scores ranged from 9681 to 46 with a median score of 300.

For Education, the scores ranged from 6728 to 0 with a median score of 309. For Recreation, the scores ranged from 3940 to 200 with a median score of 1460. You may choose to use some other measure if it seems reasonable.

c. Create a set of weights using the standards you chose in part b and write a formula to calculate a total rating based on those weights.

d. Find the total rating for Minneapolis using the formula and describe the plane this would generate. Will any of the other cities have the same rating as Minneapolis? How can you tell?

e. Find the total ratings for the cities using your weights and rank the eight cities accordingly. Would any of these cities be on the same plane?

24. The scores given in *Places Rated Almanac* for the eight cities on Cost of Living, Jobs, and Housing are in the table below. In the case of Cost of Living and Housing, a low score is better than a high score.

Rating Cities: Cost of Living, Jobs, Housing

City	Cost of Living	Jobs	Housing
Ann Arbor, MI	12,024	3,080	9,404
Austin, TX	10,158	5,448	8,277
Jacksonville, FL	9,598	3,275	7,535
Minneapolis, MN	11,099	6,242	9,316
Portland, OR	11,142	6,321	8,904
Richmond, VA	10,475	3,872	8,612
San Jose, CA	20,493	5,845	29,395
Santa Fe, NM	12,374	2,518	12,987

Source: *Places Rated Almanac,* 1993

a. Do you think these scores will yield a very different ranking from the one you found using the Arts, Education, and Recreation? Why or why not?

b. Describe a method to make the scores equivalent. (Remember that a low score is better for Housing and

Cost of Living.) Following is some information you might find useful.

For the Cost of Living, the scores ranged from 7,325 to 21,932 with a median of 9,514. For Jobs, the scores ranged from 23,028 to 1,623 with a median score of 2,434. For Housing, the scores ranged from 4,102 to 32,211 with a median of 6,944. You may choose to use some other measure if it seems reasonable.

c. Create a set of weights using the standards you choose in part b and write a formula to calculate a total rating based on those weights.

d. Find the total rating for Minneapolis using the formula and describe the plane this would generate.

e. Find the total ratings for the cities using your weights and rank the eight cities accordingly. How did the rankings compare with those using the Arts, Education, and Recreation?

25. Rank each of the cities in problems 23 and 24 from one to eight for each of the six categories, where one is the best rank.

a. Find the total rating for each city for Arts, Education, and Recreation by summing the individual ranks. Use the results to rank the cities for the three categories. How did the results compare to those you found in problem 23?

b. Find the total rating for each city for Cost of Living, Jobs, and Housing by summing the individual ranks. Use the results to rank the cities for the three categories. How did the results compare to those you found in problem 24?

26. Suppose the angle the trace of a plane made in the xz-coordinate plane with the x-axis was 78.69°, and the angle in the yz-coordinate plane with the z-axis was 26.56°.

a. Make a sketch of a plane that satisfies the two conditions.

b. What is an equation for the plane?

c. What is the angle made by the trace of the plane with the y-axis in the xy-coordinate plane?

27. *Money Guide* ranked the "100 best college buys," those schools they feel have the highest quality education for the tuition and fees. Their rankings are based on what the editors consider will provide an excellent education at a much lower price than at schools of similar quality. The table below has some of the factors listed in the *Guide* for the top ten universities according to their ranking. The student academic level is based on class rank, test scores, and high school grades, where 1 represents those with students with the highest academic records.

Best College Buys

	Tuition and Fees	Room and Board	% Students Receiving Aid	% of Need Met	Student-Faculty Ratio	% Who Graduate in 6 yrs	Student Academic Level
California Institute of Technology	17,586	6,620	75	100	3:1	78	1
New College of University of South Florida	7,950	3,847	70	93	10:1	60	1
Northeast Missouri State	3,975	3,330	N.A.	80	22:1	49	4
Rice University	12,034	5,900	84	100	9:1	88	1
State University of New York at Binghamton	8,679	4,654	49	62	19:1	80	1
State University of New York at Albany	8,856	4,836	90	80	18:1	73	2
Spelman College	8,875	5,890	76	38	15:1	74	2
Trenton State College	6,658	5,650	48	90	15:1	78	2
University of Illinois at Urbana/Champaign	9,130	4,408	81	75	13:1	79	2
University of North Carolina-Chapel Hill	10,162	5,350	44	95	10:1	85	1

Source: *Money Guide*, 1996 Edition

In 1994–95, the average tuition at private universities was \$10,333 and at public universities, \$2,730. On average, 69% of the students in the schools surveyed received financial aid for about half of the costs. On the average, 55% of the students in those schools graduated in six years. The typical student-faculty ratio was 34 to 1. Ranking for student academic levels went from 1 to 5, with a 1 as best. According to the U.S. Census, the average cost of room and board at a private college was \$4,793 and at a public college \$3,680 in 1994.

a. Using the information above, rate and then rank the ten universities in terms of percentage of students receiving aid, percentage of need met by aid, and the percentage of students who graduate in 6 years. (You might use the typical 69% as the replacement for the NA category for Northeast Missouri State.)

b. Find a formula to rate and then rank the ten universities using all seven categories given. (You might use the average of the cost and tuition for public and private schools.)

c. Compare the rankings from parts a and b.

d. In their rating scheme, *Money Guide* actually used 16 variables, not all of which are given here. Using these variables, the universities ranked as follows.

Ranking

California Institute of Technology	5
New College of University of South Florida	1
Northeast Missouri State	3
Rice University	2
State University of New York at Binghamton	7
State University of New York at Albany	10
Spelman College	8
Trenton State College	4
University of Illinois at Urbana/Champagne	9
University of North Carolina at Chapel Hill	6

Compare your rankings to those from *Money Guide.*

Extension

28. Consider a pair of weights $[w_1 \ \ w_2]$ for two variables. These weights will determine a vector in a two-dimensional plane.

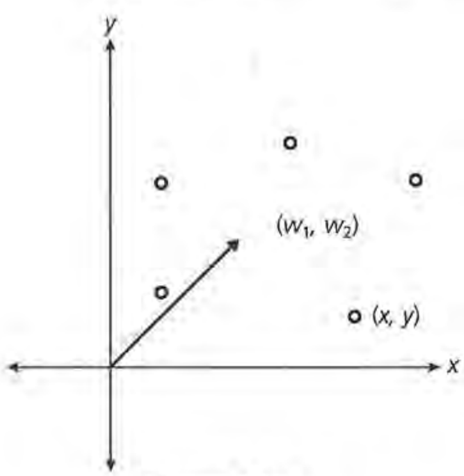


Figure 4.10

Using those weights, each data point determines a rating and the equation for a line that will contain that data point.

a. Each of these lines will be perpendicular to the ray formed by taking all of the positive scalar multiples of $[w_1 \ \ w_2]$. Why is this so? (Recall how the slope of the line is related to the weights.)

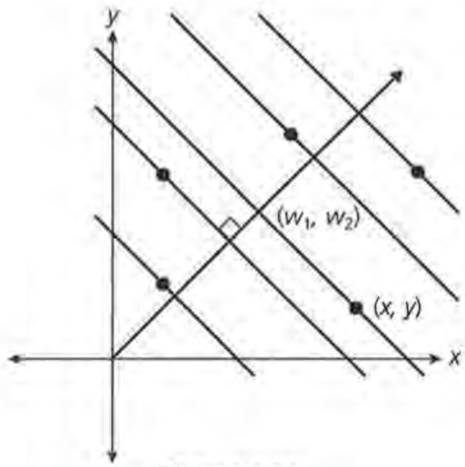


Figure 4.11

This means that each of the data points can be projected orthogonally (at right angles with the line) onto the vector $[aw_1 \ aw_2]$ for some a.

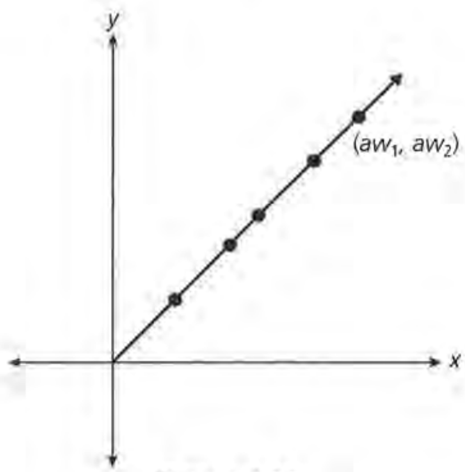


Figure 4.12

Another way to determine the data point that is the "best" according to a given set of weights is to use this projection onto a multiple of the weight vector. The data point for which the projection produces the vector of greatest magnitude will be the data point that has the highest weighted rating and is the "best" according to those weights. To prove that statement, consider the following:

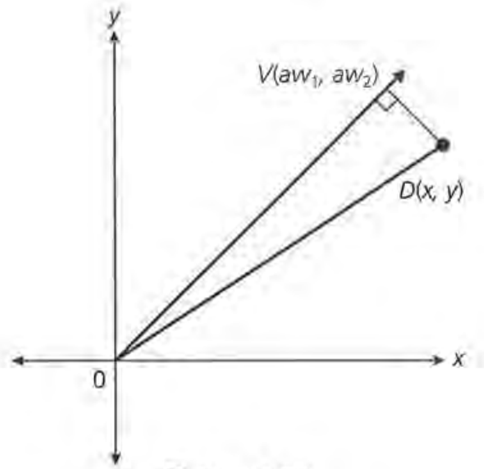


Figure 4.13

b. What is the equation of the line that contains OV? Use the equation of this line to find the distance from D to OV. (The distance from a point (x_1, y_1) to a line $Ax + By + C = 0$ is given by $\frac{|Ax_1 + By_1 + C|}{\sqrt{A^2 + B^2}}$.

c. Find length of OD.

d. Remember that the equation of the line containing DV is $w_1x + w_2y = R$.

Prove that the length of $OV = \frac{R}{\sqrt{w_1^2 + w_2^2}}$.

e. Show how this supports the statement made above: The data point for which the projection produces the vector of greatest magnitude will be the data point that has the highest weighted rating and is the "best" according to those weights.

The method above works for any number of variables. In each case, for n variables, there is an n-dimensional weight vector. Each data point is projected orthogonally onto that vector, and the point that has the highest rating will be the point for which $\frac{R}{\sqrt{w_1^2 + w_2^2 + w_3^2 + \ldots + w_n^2}}$ is the largest.

29. Demonstrate that the statement above works using the games played, home runs, batting average and runs batted in data.

ASSESSMENT

From Best Companies for Women to Cars

Each year *Woman's World* identifies the best 100 companies for working mothers. The following data came from the October 1995 issue.

OBJECTIVE

Use weights, a sweeping line, and matrices to create ratings and ranks.

Best Companies for Working Mothers

Companies	Pay Rating	Opportunity to Advance	Child Care Support	Family-Friendly Benefits	Median Salary	% of Women Professionals	% of Highest Paid Women
Calvert Group	4	4	4	4	39,000	53	43
Corning Incorporated	4	2	5	3	30,200	28	12
The Dow Chemical Company	4	2	3	4	44,400	25	7
Dupont Company	5	2	4	4	42,840	21	14
Fannie Mae	4	3	2	4	34,950	49	43
Fel-Pro Incorporated	4	2	5	5	37,999	40	19
General Motors	4	2	3	4	39,570	25	10
Johnson & Johnson	5	3	5	4	31,000	45	26
Mattel, Inc.	4	3	3	4	36,376	52	29
Merck & Co., Inc.	5	3	5	4	37,000	47	24
Motorola	3	2	4	3	38,000	23	6
Quad/Graphics, Inc.	3	2	5	3	30,000	50	9
Security Benefit Group of Companies	3	2	4	4	25,194	59	30

Source: *Woman's World*, October 1995

Woman's World staff rated each company in the first four categories from 1 to 5 based on certain criteria, where 5 is the best rating. A typical salary, the percent of women professionals employed in the company, and the percent of the highest paid employees who were women constituted part of the information given about each company and used to create the rankings.

The companies in the table were those whose salary was given for engineers or accountants.

1. Use the set of data given to answer the following:
   a. How would you find the total ratings and ranks for the companies based on the first four numerical ratings, and how could you use matrices to do so?
   b. How would you use two variables that are in different units to obtain a ranking of the companies? Explain in this context how to use a sweeping line.
   c. How would you use three variables that are in different units to obtain a ranking of the companies? Explain in this context how to use a **sweeping plane.**
   d. Based on your work for a to c and anything else that seems appropriate, how would you rank these thirteen companies in terms of the best companies for working mothers? Explain your reasoning.
2. The data in the table below are customer satisfaction ratings. The data are based on the number of problems per 100 cars during the first three months of ownership according to JD Power and Associates Initial Quality Survey and the amount of damage caused by four crashes (front and rear into flat barrier, rear into a pole, front into barrier at angle) at a speed of 5 mph. Note that high customer satisfaction is good as is low damage cost.

Car Problems

Car	Customer Satisfaction Rating	Total Cost of Damage ($)
Nissan Maxima	138	3,605
Honda Accord	149	1,433
Saab 900S	132	1,734
Subaru Legacy	155	1,966
Volvo 850	155	2,131
Toyota Camry	148	2,328
Ford Taurus	128	2,814
Lumina	125	2,629

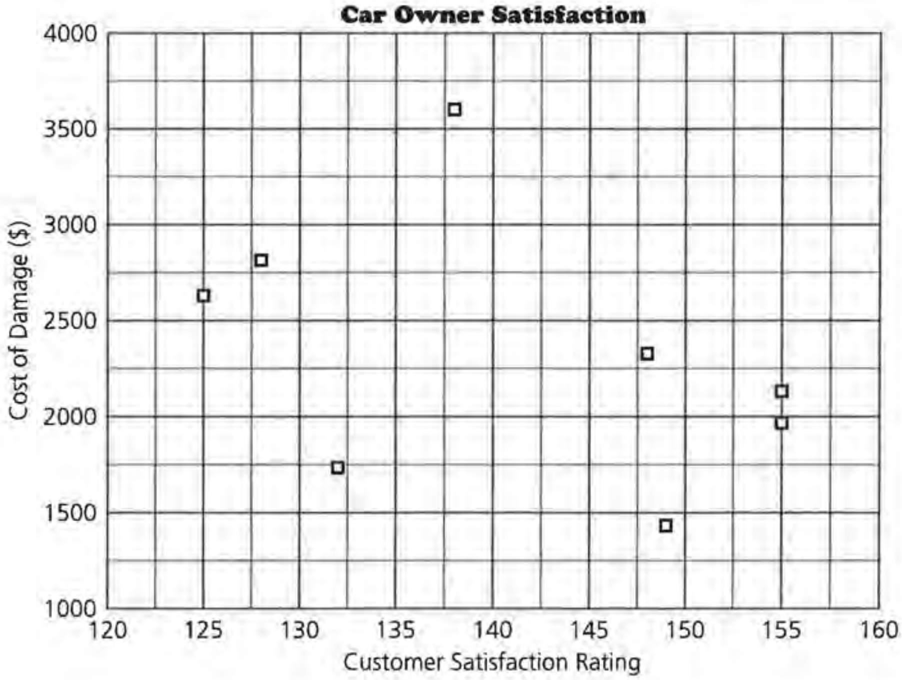


Figure A1.1

a. The average customer satisfaction rating for all cars was 100. Find the average cost for damage in the crashes for all of the cars tested. Use this information to find weights such that both variables contribute about equally to the rating.

b. What rating will the Volvo have using those weights?

c. Graph the equation for the line that shows cars with the same rating as the Volvo. How will the equation containing the rating for the other cars compare to the equation for the Volvo? Sketch at least one other equation on the graph.

d. Can a sweeping line help you identify the "best" car according to those weights for the variables? Explain why or why not.

e. Use your equation to find a total rating for each of the cars. Rank each of the cars according to the ratings.

f. How reliable do you think the rating you produced in part e will be? Write a paragraph describing your conclusions.

Modeling, Matrices, and Multiple Regression

INTRODUCTORY ACTIVITY

What Affects Your Walking Speed?

Do people who have long strides walk faster than those who do not?

Does the length of your stride affect how fast you can walk?

We are often interested in the relationship between two or more variables. We can use linear relationships to make predictions for one variable when we know the value of the other variable. In this unit you will explore ways to use several variables to predict the values of another variable.

OBJECTIVE

Compare the strengths of different relationships.

EXPLORE

Take a Walk

1. Design an experiment to determine the length of each person's stride and the time it takes the person to rapidly walk a given course. Have each person in class walk the course as fast as possible and record both the person's stride length and time.

   a. Plot (length of stride, time). Describe the plot.

   b. Does there appear to be a relationship between the variables?

2. Does your height affect your speed? Do taller people walk faster?

   a. Collect data on height from the class. Plot (height, time).

   b. Describe the plot. Does there appear to be a relationship between the variables?

c. Which appears to have the stronger relationship: (height, time) or (length of stride, time)? How did you decide?

3. How could you use both length of stride and height to determine time? Would a plot help? Explain your answer.

4. What other factors might affect your walking speed? How could these factors be incorporated into the process for predicting your walking speed?

LESSON 5

Matrices and Linear Regression

How can you find a model for the relationship between height and weight? between the number of calories and the grams of fat in fast food? between time and the change in temperature in a physics lab?

How can matrices be useful in the process?

In earlier work you studied different methods to represent relationships: linear models and nonlinear models such as exponential, logarithmic, or power models. You considered questions relative to the ideas that are important in the modeling process. How do you know if a model is appropriate?

You can use matrices to help you in the modeling process. In the first section you studied how to use matrices to solve equations, to evaluate formulas, and to determine weighted values for what is "best." Matrices can be used in a variety of situations to help represent and manipulate information. An important area in mathematics and science is to find an algebraic model to describe the relationship in a set of data. In this lesson you will learn how to use matrices to help determine how good your model is for making predictions.

OBJECTIVES

Review the process of modeling a relationship between variables by fitting a line to the data.

Use matrices to express the least squares regression equation.

Review correlation, root mean squared error, and residuals.

INVESTIGATE

Airline Operating Costs

Is there any relationship between the number of seats of an aircraft and the operation costs? The matrix on the next page lists the operating costs and the number of seats on different airplanes.

Airplanes: Seats, Cost

Airplane	No. of Seats	Operating Cost per Hour
B747-100	405	$6132
L-1011-100/200	296	3885
DC-10-10	288	4236
A300 B4	258	3526
A310-300	240	3484
B767-300	230	3334
B767-200	193	2887
B757-200	188	2301
B727-200	148	2247
MD-80	142	1861
B737-300	131	1826
DC-9-50	122	1830
B727-100	115	2031
B737-100/200	112	1772
F-100	103	1456
DC-9-30	102	1778
DC-9-10	78	1588

Source: *World Almanac*, 1992

Discussion and Practice

1. Figure 5.1 is a scatter plot of (*number of seats, operating costs*).

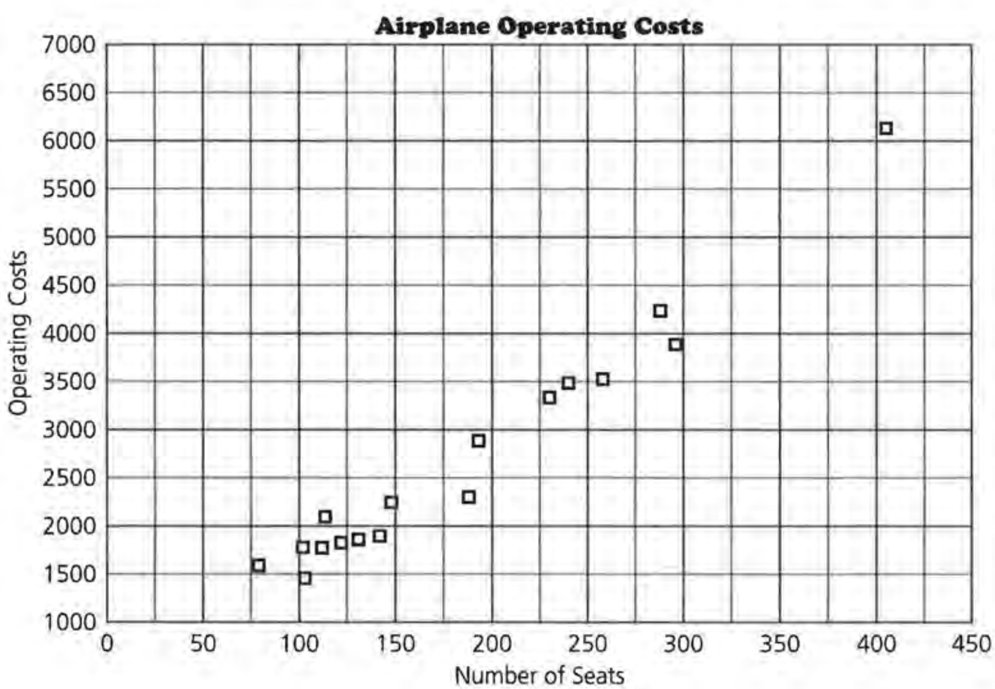


Figure 5.1

a. Describe the relationship between the variables in the scatter plot.

b. Draw a line that you think captures the trend in the relationship. Determine the rate of change for the line and explain what it represents in terms of the data.

c. Write the equation of the line you drew.

d. Using the equation of the line you created, predict the cost to run a plane that has 240 seats. Compare your value with the actual value given in the original data.

The process of fitting a model to data is called *regression*. One of the ways to determine whether you have a good model is to analyze the way the model behaves in relation to the data used to create the model. Finding the predicted values for a given model can be represented using matrices. Suppose the line you created for (number of seats, operating cost) has the equation: $C = 425 + 13.3S$. To find the predicted C for each value of S you will have to substitute for S and do the calculation. Just as you did in the earlier lessons, you can write this as a matrix problem.

$425 + 13.3S = C$ can be written

$1 \cdot 425 + S \cdot 13.53 = C$ and in matrices for some value S,

$$[1\ S] \cdot \begin{bmatrix} 425 \\ 13.3 \end{bmatrix} = [C] \text{ or } X\,B = \hat{Y}$$

where X is the matrix that contains the data for the number of seats along with a 1; B is the slope and intercept matrix, and $\hat{Y}$ is the predicted value matrix.

For $S = 405$ seats, the system would give you $[1\ 405] \cdot \begin{bmatrix} 425 \\ 13.3 \end{bmatrix} = [5811.5]$,

or an airplane with 405 seats will cost about \$5811.50 per hour to operate.

Rather than doing just one x-value at a time, rewrite matrix X with a column vector of 1s as the first column and a column vector of all possible values for the number of seats as the second column.

$$X \cdot B = \begin{bmatrix} 1 & 405 \\ 1 & 296 \\ 1 & 288 \\ 1 & 258 \\ 1 & 240 \\ 1 & 230 \\ 1 & 193 \\ 1 & 188 \\ 1 & 148 \\ 1 & 142 \\ 1 & 131 \\ 1 & 122 \\ 1 & 115 \\ 1 & 112 \\ 1 & 103 \\ 1 & 102 \\ 1 & 78 \end{bmatrix} \cdot \begin{bmatrix} 425 \\ 13.3 \end{bmatrix}$$

2. Multiplying X and B will give you matrix $\hat{Y}$, the column vector of all predicted y-values.

a. Find XB and justify the dimensions of B.

b. Find the predicted operating cost for 296 seats using matrix multiplication.

c. $1,781 is one of the entries in XB. Explain how it was obtained.

Residuals

A *residual* is the difference between the predicted operating cost and the actual operating cost for a given number of seats. A careful study of the residuals will help you determine whether a model is a good fit for the data. If there is a pattern to the residuals, the error is predictable and another model might be better. The best models will have residuals that are not predictable and that seem to be randomly scattered above and below the line where the residuals equal 0.

The residuals are indicated on the plot in Figure 5.2.

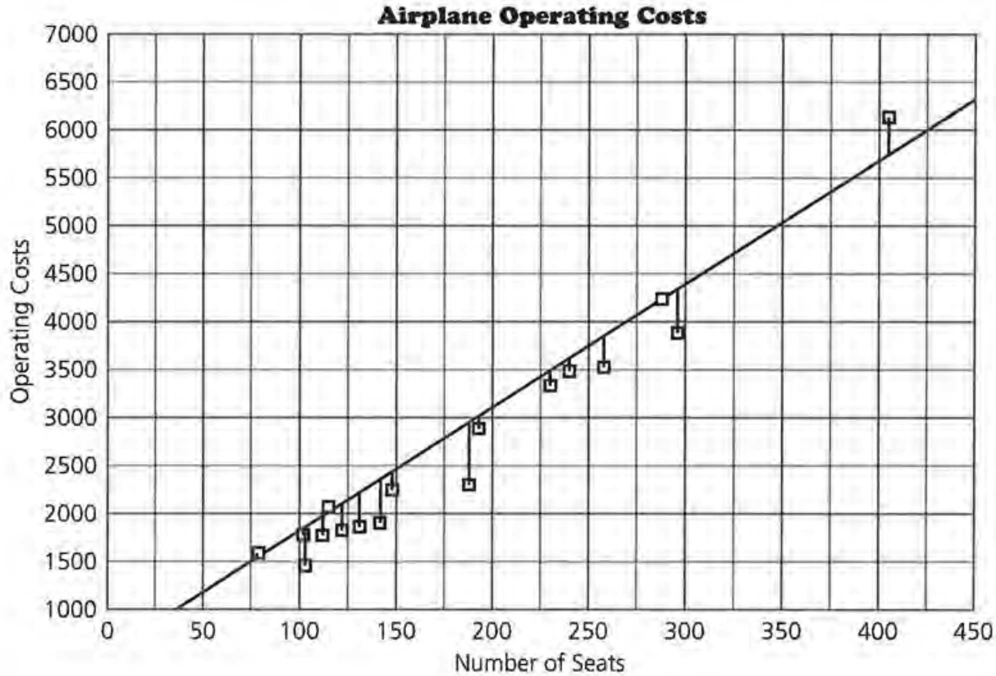


Figure 5.2

The residuals vector (or matrix) can be found by taking the difference between the matrix of the vector for the actual cost, Y, and the matrix of the vector for the predicted cost, $\hat{Y}$.

$Y - \hat{Y} = R$ where R is the matrix of residuals.

$$\begin{bmatrix} 6132 \\ 3885 \\ 4236 \\ 3526 \\ 3484 \\ 3334 \\ 2887 \\ 2301 \\ 2247 \\ 1861 \\ 1826 \\ 1830 \\ 2031 \\ 1772 \\ 1456 \\ 1778 \\ 1588 \end{bmatrix} - \begin{bmatrix} 5811.5 \\ 4361.8 \\ 4255.4 \\ 3856.4 \\ 3617.0 \\ 3484.0 \\ 2991.9 \\ 2925.4 \\ 2393.4 \\ 2313.6 \\ 2167.3 \\ 2047.6 \\ 1954.5 \\ 1914.6 \\ 1794.9 \\ 1781.6 \\ 1462.4 \end{bmatrix} = \begin{bmatrix} 320.5 \\ -476.8 \\ -19.4 \\ -330.4 \\ -133.0 \\ -150.0 \\ -104.9 \\ -624.4 \\ -146.4 \\ -452.6 \\ -341.3 \\ -217.6 \\ 76.5 \\ -142.6 \\ -338.9 \\ -3.6 \\ -125.6 \end{bmatrix}$$

Root mean squared error

In earlier work on fitting lines to data, you probably studied different ways to determine how well a line fits data. One of the measures was the sum of the squared residuals, $\Sigma(y_i - \hat{y})^2$. The *root mean squared error* is the square root of the average of this sum, $\sqrt{\frac{(\Sigma(y_i - \hat{y}_i)^2)}{n}}$. This tells you how large in magnitude a typical residual is. To find the root mean squared error, take the residuals, square them, find the sum of the squares, divide by the number of residuals, and then take the square root. The smaller the root mean squared error (or, equivalently, the smaller sum of the squared residuals), the better the line will yield predicted y-values that are close to the actual y-values.

3. In order to get the sum of squared residuals, you have to square each individual residual.

a. Could you square the residual matrix to accomplish this? Why or why not?

b. Remember that the transpose of a matrix, A^{T}, is formed when the rows and columns of the matrix are interchanged. What are the dimensions of the transpose of the residual matrix R?

c. In the product AB, A is said to premultiply B. Describe the result if the transpose of the residual matrix premultiplies the residual matrix itself.

$$\begin{array}{cc} R^{\mathrm{T}} & \cdot \; R \end{array}$$

$$[320.5\ -476.8\ -19.4\ -330.4\ -133.0\ -150.0\ -104.9\ -624.4\ -146.4\ -452.6\ -341.3\ -217.6\ 76.5\ -142.6\ -338.9\ -3.6\ -125.6] \cdot \begin{bmatrix} 320.5 \\ -476.8 \\ -19.4 \\ -330.4 \\ -133.0 \\ -150.0 \\ -104.9 \\ -624.4 \\ -146.4 \\ -452.6 \\ -341.3 \\ -217.6 \\ 76.5 \\ -142.6 \\ -338.9 \\ -3.6 \\ -125.6 \end{bmatrix}$$

d. Which of the following is equivalent to $(R^{\mathrm{T}})(R)$?

i. R^2 ii. $\Sigma(y_i - \hat{y}_i)^2$ iii. $(R)(R^{\mathrm{T}})$ iv. $\Sigma(y_i - \overline{y})^2$

4. Calculate the product $(R^{\mathrm{T}})(R)$.

a. Use it to find the root mean squared error.

b. Use the line $C = 425 + 13.3S$ to predict the operating costs for a plane that has 400 seats. What does the root mean squared error indicate about your prediction?

c. In general, what conclusions can you make about the prediction model?

5. Inspecting a plot of the residuals can help you determine whether there is a pattern in them.

a. How does a residual relate to the graph of the data and to the equation? How can you tell from the plot when the fit is relatively good?

b. Figure 5.3 contains a plot of (*number of seats, residuals*). This is called a *residual plot.* Describe the plot.

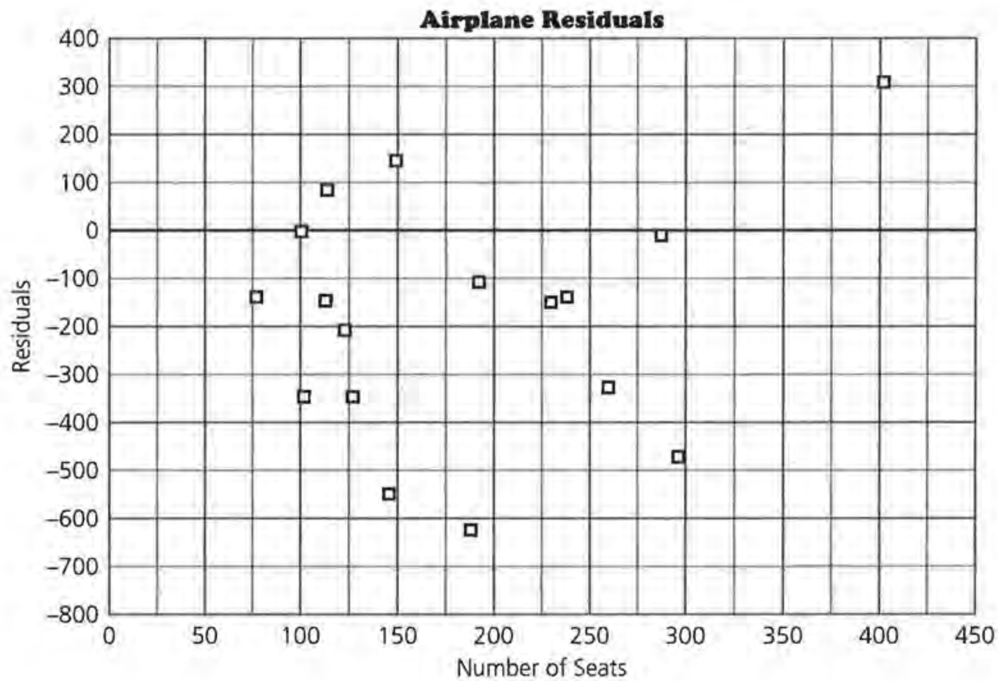


Figure 5.3

c. How does the plot of the residuals in Figure 3 compare to the residuals you observed in the plot in Figure 2?

d. Write an argument explaining why or why not the line $C = 425 + 13.3S$ is a *good* line to represent the relationship between the number of seats and the operating costs of the planes.

Linear regression

One of the lines you studied earlier is the *least squares linear regression* line. You may remember from your earlier work that the least squares linear regression equation has the smallest sum of squared errors of all lines that can be fit to a data set.

Most graphing calculators or computers can easily calculate the least squares linear regression line after you enter the data.

6. Calculate the least squares linear regression line for (*number of seats, operating costs*).

   a. Use it to write a matrix representation for the prediction model.

   b. Find $\hat{Y}$, the predicted y-values matrix, and R, the residuals matrix.

   c. Use the matrix model to find the sum of the squared residuals. How does the sum compare to the one you found earlier?

7. Find the root mean squared error for the predicted operating costs based on the number of seats using the equation for the least squares regression line.

   a. Use the least squares regression line to predict the operating costs for a plane with 240 seats. What does the root mean squared error indicate about your prediction?

   b. How do you think outliers will affect the root mean squared error?

8. Plot (*number of seats, residuals*) using the least squares regression line for (*number of seats, operating costs*).

   a. Describe the plot.

   b. How well does the least squares linear regression model fit the data?

Correlation

Another tool in the modeling process is correlation. *Correlation* is a measure of the degree of linear association between two variables. The correlation coefficient, r, indicates how tightly packed the data are around a line; r^2 indicates the percent of change in the y-values that can be attributed to changes in the x-values. If the correlation is weak (close to zero), a linear model is not appropriate, although another model may be. It is important to look at a plot of the data before drawing any conclusions about the model. A strong correlation (close to 1 or –1) does not necessarily mean the model you are investigating is the best model. Correlation is very sensitive to outliers and clusters of data points.

9. Using a calculator or computer, find the correlation for (*number of seats, operating costs*).

   a. What does it tell you about the relation between the two variables?

   b. What is the value of r^2? What does it tell you about the relationship?

   c. Are there any outliers? If there are, how do they affect the correlation?

10. Summarize the work you have done in problems 6–9 to find a model for the relationship between the number of seats and the operating costs for the set of planes.

There are many other variables related to the operating costs of airplanes. Will a least squares linear regression model help you describe any of these relationships? Additional information about airplanes is included in the table below.

More Airplane Data

Aircraft	No. of Seats	Air Speed (mph)	Operating Cost/Hour ($)
B747-100	405	519	6132
L-1011-100/200	296	498	3885
DC-10-10	288	484	4236
A300 B4	258	460	3526
A310-300	240	473	3484
B767-300	230	478	3334
B767-200	193	475	2887
B757-200	188	449	2301
B727-200	148	427	2247
MD-80	142	415	1861
B737-300	131	413	1826
DC-9-50	122	378	1830
B727-100	115	422	2031
B737-100/200	112	388	1772
F-100	103	360	1456
DC-9-30	102	377	1778
DC-9-10	78	376	1588

Source: *World Almanac*, 1992

11. Consider the speed at which a plane flies and its relation to operating costs. Plot (*speed, operating costs*).

   a. Describe the association you can see in the plot.

b. Find the correlation for (*speed, operating costs*). What does this indicate about the data?

c. Find the least squares linear regression line. How well do you think it describes the relationship?

12. Create a matrix system to use the least squares linear regression line to predict operating costs using speed.

a. Use the matrix system to find the residuals.

b. What is the root mean squared error? Estimate the error if you use your linear model to predict the operating costs for a plane that travels 400 miles per hour.

c. Plot (*speed, residuals*). What does this plot indicate about using the model?

Transforming Data

When a linear regression model is not appropriate, but the data do seem to form a pattern, you can try transforming the data to determine a better model. Because logarithms use a different scale, they may help to find another model. In this case, to get a better fit it might help to try a transformation that would lower the points with large operating costs on the plot.

13. Divide the work among group members so that each member has a strategy to transform the speed and operating costs. Select your strategy from the list below.

(*speed, ln*(*operating costs*))

(*speed, multiplicative inverse of operating costs*)

(*ln(speed), operating costs*)

(*ln(speed), ln(operating costs)*)

a. Plot the paired data after each transformation. Which of the transformations seems to produce the most linear relationship?

b. Find the least squares linear regression line for the transformed data you chose in part a.

14. Use the least squares regression line you found above to determine how well the line will fit the data by answering the following questions.

a. Write a matrix representation of the linear equation. Use it to create a set of predicted values.

b. Use matrices to find the residuals and the root mean squared error for your model. What do these indicate about using your model for the relation between speed and operating costs?

c. Make a residual plot. What does this indicate about the relationship?

d. Consider using the model you found to be the best from working the first part of this problem. If you were to predict the operating costs for a speed of 550 miles per hour, how reliable do you think your prediction will be? Explain how you found your answer.

15. A model that seems to be a good fit for (age of plane, cost) is $y = 10x^{0.52}$ where y is in hundreds of dollars.

a. Does the matrix system below represent the model? Explain why or why not.

$$[1\ x] \cdot \begin{bmatrix} 10 \\ 0.52 \end{bmatrix} = [y]$$

b. Take the natural logarithm of both sides of the equation.

$$\ln y = \ln 10x^{0.52}$$

Use the properties of logarithms to express the relationship in the form of a sum.

c. Use both the equation $y = 10x^{0.52}$ and your answer for part b to predict the cost of operating a seven-year-old airplane. How do the two compare?

Summary

Matrices can be a useful tool in the modeling process. If you see a pattern in a set of data, you can search for a model to describe the pattern. If the pattern is linear, you can use the least squares linear regression line. If the data are not linear, you can try to straighten the data by transforming them, using what you know about the situation and about the plot. In either case, you can find the predicted values using your model by writing a matrix expression: $XB = \hat{Y}$, where X is the independent variable, Matrix B is the coefficient matrix for the model, and $\hat{Y}$ is the matrix of predicted values. The residuals are the difference between $\hat{Y}$ and Y, $Y - \hat{Y} = R$. If the residuals are small and seem to be randomly distributed, the model is a good model for the data. Correlation and root mean squared error are two other tools that can be an aid in the modeling process. The sum of the squared errors is the product $R^T \cdot R$ where R is the residual matrix.

Practice and Applications

16. The least squares linear regression line for (*fat grams, calories*) for a variety of fast foods can be expressed in the following matrix system:

$$[1\ x] \cdot \begin{bmatrix} 182 \\ 12.5 \end{bmatrix} = [y]$$

where x = number of fat grams, y = calories.

a. Use the system to find the predicted number of calories for the following food items:

Fast Foods Calories/Fat

Place	Food	Calories	Fat (gms)
McDonald's	McLean Deluxe	320	10
	Chicken salad	162	4.5
	Quarter Pounder	510	28
Burger King	Broiler	267	8
	Chunky chicken salad	172	5
	Whopper with cheese	706	44
Pizza Hut	Cheese pizza	492	18
	Pan supreme pizza	589	30
Taco Bell	Bean burrito	447	14
	Chicken burrito	334	12
	Taco salad	905	61
	Nachos-Bell Grande	649	35

Source: *Eating on the Run*, Human Kinetics, Tribole, Evelyn. *USA Today*, Jan. 7, 1992

b. Find the residuals using matrices and make a residual plot. What does this tell you about your model?

17. An economics student claims the relationship between the per capita gross national product in 1992 (G) and the percent of students enrolled in secondary schools in the following countries (E) can be given by the equation $G = -8152 + 223E$.

a. Write the matrix system for the equation. Use the system to predict the gross national product for the given countries.

	% Enrolled Secondary School	Predicted GNP (in US $)
China	44	
Hungary	76	
Iran	53	
Peru	67	
Mexico	63	
Kenya	23	
Thailand	28	
Germany	97	

Source: *Universal Almanac*, 1995

b. If the actual GNPs were as follows, use matrices to find the residuals and the sum of squared residuals.

	% Enrolled Secondary School	Predicted GNP (in US $)	Actual GNP	Residuals
China	44		380	
Hungary	76		3010	
Iran	53		2190	
Peru	67		1020	
Mexico	63		3470	
Kenya	23		340	
Thailand	28		1840	
Germany	97		23030	

c. What is the root mean squared error and what does it tell you about your model?

d. Comment on the economic student's attempt to find a linear model to relate secondary education to gross national product.

18. A plot of (*speed, length of flights*) is given in Figure 5.4 with a regression line drawn through the data.

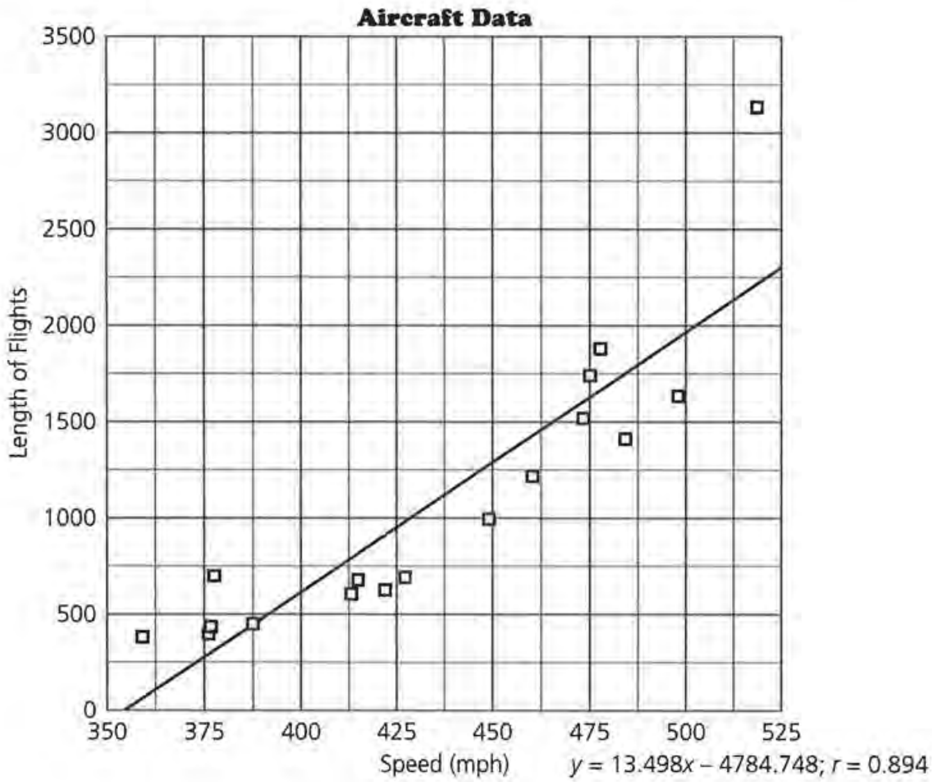


Figure 5.4

a. How well does the line seem to represent the relationship in the data?

b. The correlation for the line is 0.894. If the outlier around 520 mph were removed from the data, how do you think the correlation would change? Why?

c. Figure 5.5 contains a plot of the residuals. Describe what the residual plot indicates about using the line to predict the length of flights from the speed.

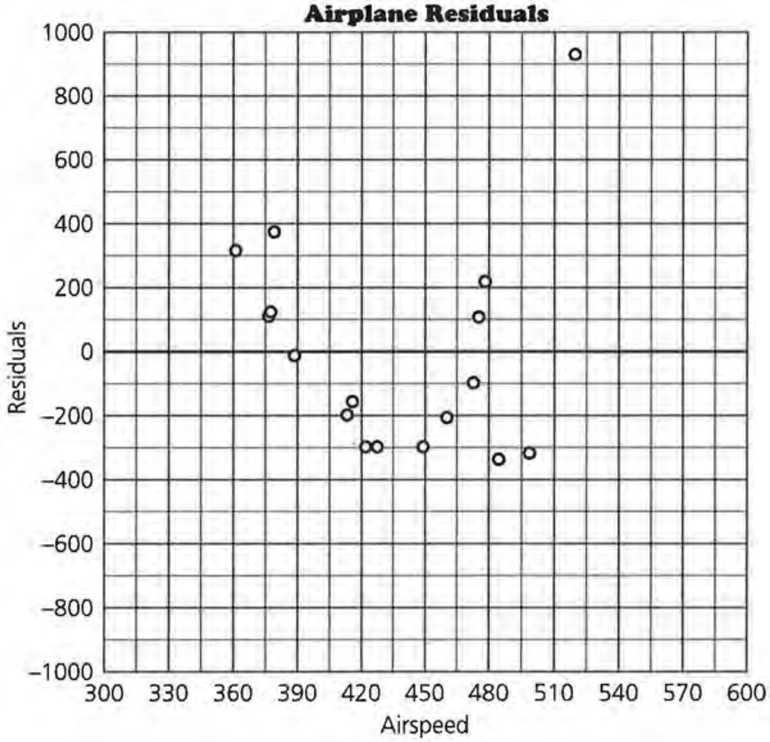


Figure 5.5

19. Figure 5.6 is a residual plot for the least squares linear regression line on (*flight length, fuel consumption*) for the set of planes.

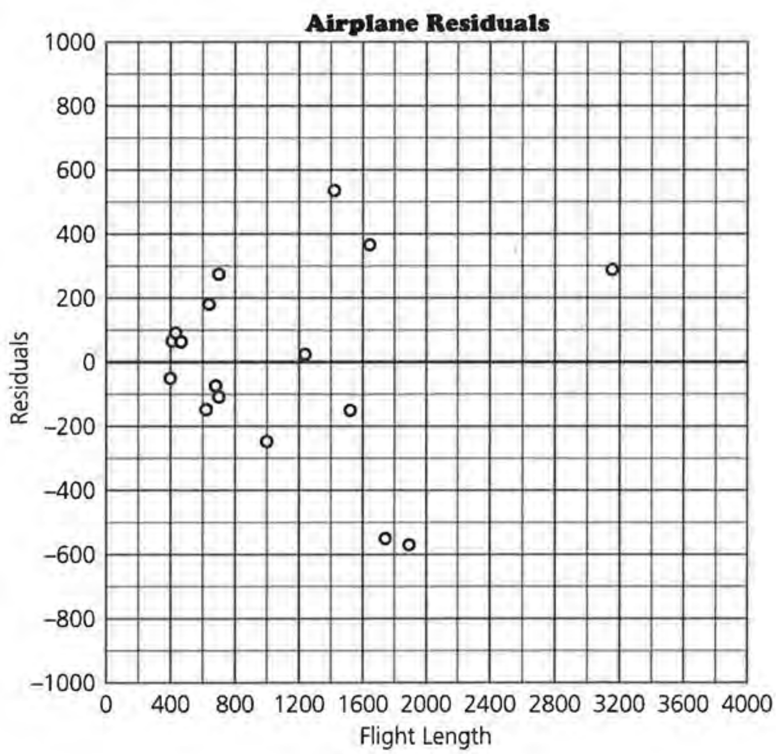


Figure 5.6

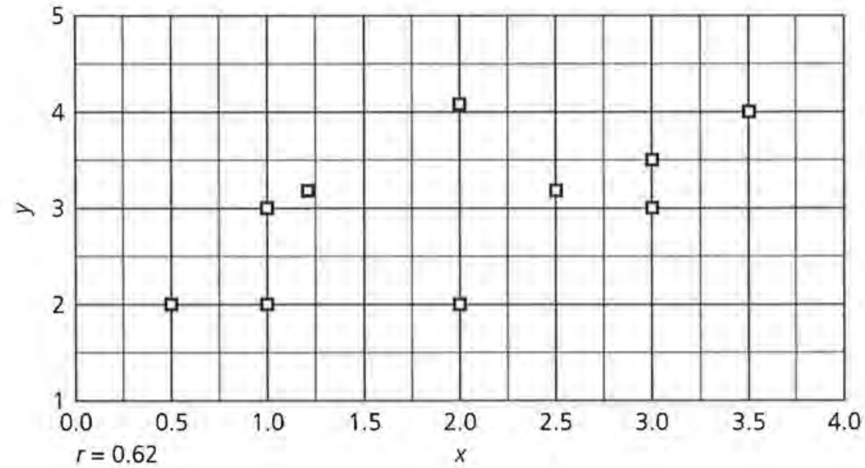
a. Describe any pattern you see in the residuals.

b. What do the residuals indicate about using the model that generated the residuals to predict the fuel consumption from the length of the flight?

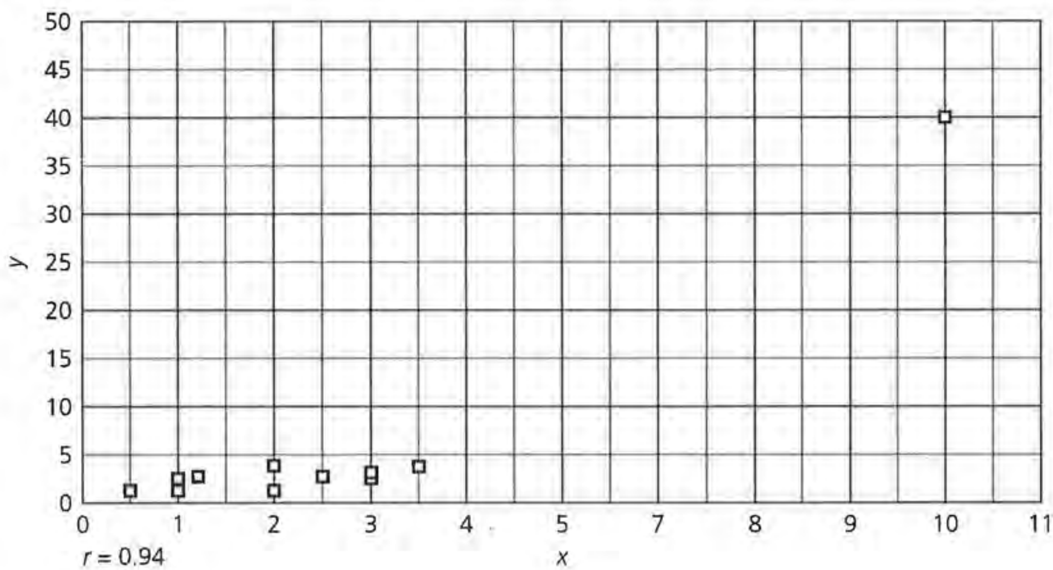
20. The correlation coefficient is given for each of the plots i, ii, iii, and iv.

a. Comment on the relation between the plot and the numerical value of the correlation.

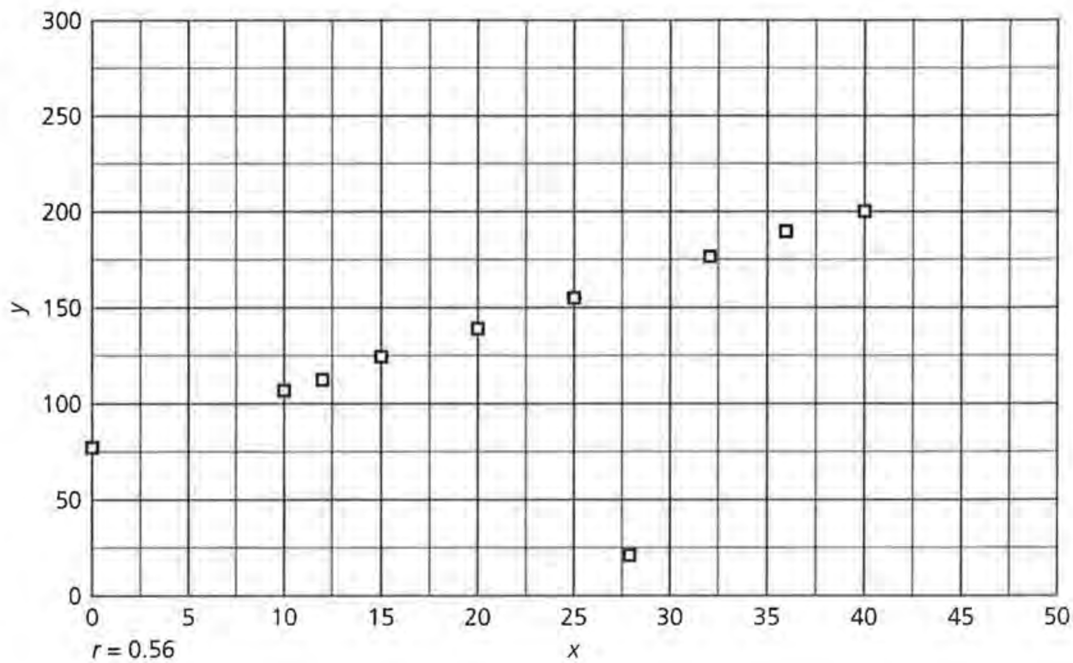
i.



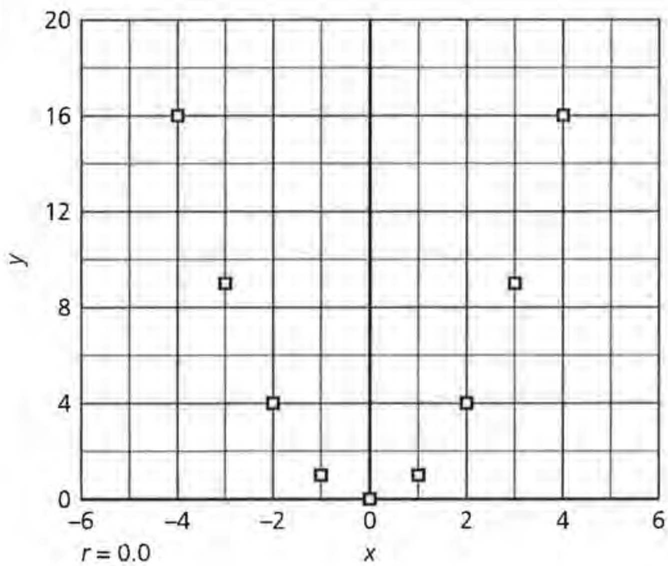
ii.



iii.



iv.



b. Plots i and ii are of the same data set with one exception. The point (10, 40) was added in plot ii. Comment on the effect of this one point on the correlation.

c. What observations can you make about using only correlation as a tool to help you find a good model for a data set?

21. Find the prices of 12 used cars of the same size or type. Plot (*age of car, price*) and look for a model that describes the price of used cars as a function of the age of the car. Explain how you chose your model and which of the tools you used to help make your choice.

LESSON 6

Multiple Variables and Modeling

What factors might have an impact on your success in college?

Can you predict the grade point average a student will have in college?

What variables might be useful to help make a good prediction: high school grades, the courses the student took in high school, SAT scores, or ACT scores?

OBJECTIVES

Examine how two different factors can be used to predict an outcome.

Use recursion to generate a procedure.

In earlier lessons you learned how to find models to make predictions when you had only one independent variable. In this lesson you will learn how to use two independent variables to make a prediction. Students with high SAT verbal (*SATV*) and SAT math (*SATM*) scores tend to get better grades in college than do students who do not score well on the SAT exam. Suppose you only had those two variables available. The investigation in this lesson will show that it is possible to model the way in which college grade point average, *GPA*, depends on both *SATV* and *SATM*.

INVESTIGATE

College Entrance Data

The table that follows contains the data on *GPA*, *SATV*, and *SATM* for each of 15 college students.

College Grade Point Average

Student Number	*GPA*	*SATV*	*SATM*
1(*X*)	3.58	670	710
2	3.17	630	610
3	2.31	490	510
4	3.16	760	580
5	3.39	450	510
6	3.85	600	720
7	2.55	490	560
8	2.69	570	620
9	3.19	620	640
10	3.50	640	660
11	2.92	730	780
12	3.85	800	630
13	3.11	640	730
14	2.99	680	630
15	3.08	510	610

Source: Oberlin College, 1993

The scatter plot in Figure 6.1 shows the relationship between *GPA* and *SATV*. The fitted least squares linear regression line has been added to the plot. The first student on the list is shown with an *X* in the plot and on the table so you can keep track of this student as you progress through the lesson. ($\hat{GPA}$ is a notation for the predicted grade point average using the regression model.)

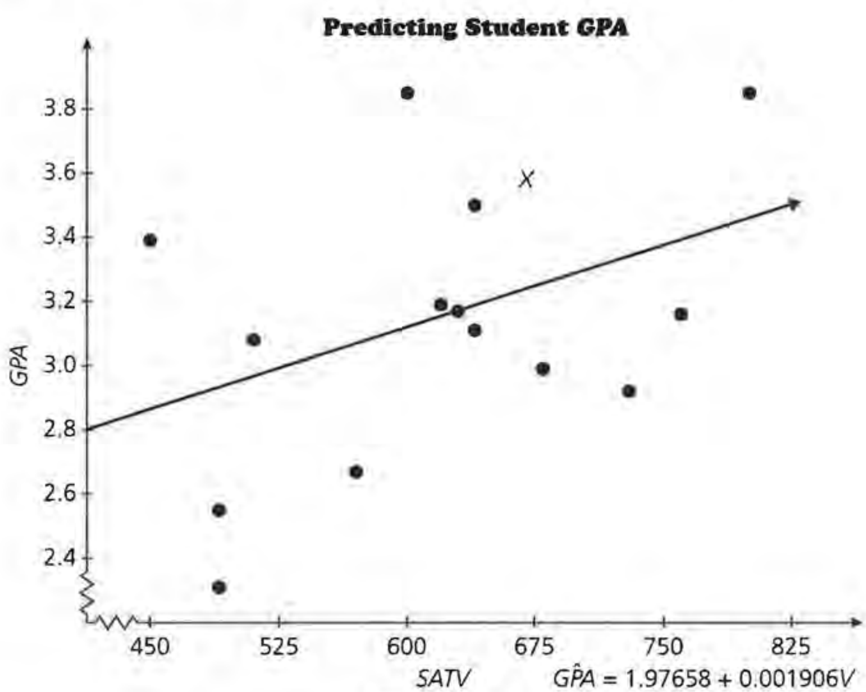


Figure 6.1

Figure 6.2 shows the relation between *SATM* and *GPA*.

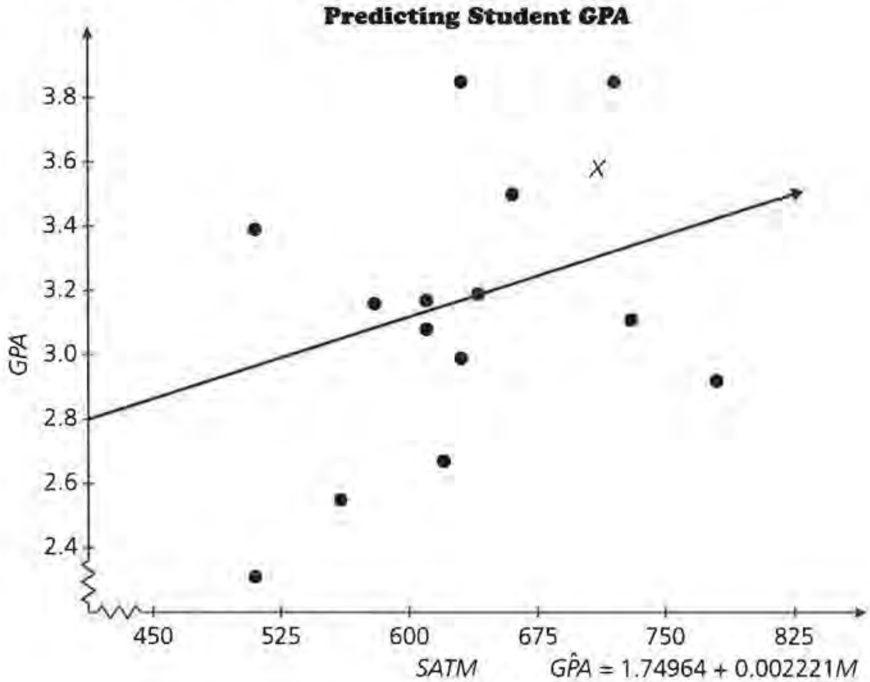


Figure 6.2

Clearly, you could use either *SATV* or *SATM* to predict *GPA*. However, it would seem even better to use both *SATV* **and** *SATM* as predictors.

SATV and the Regression Line

Consider again the plot of *GPA* versus *SATV*. The equation of the fitted linear regression model is $\hat{GPA} = 1.97658 + 0.001906V$. (Note that the coefficients are given to several decimal places; in doing the calculations on a computer or a calculator, carry as many digits of accuracy as possible.)

Discussion and Practice

1. Enter the data into your calculator or computer. Verify the equation for the least squares linear regression line.

- **a.** What is the slope of the line and what does it tell you about the relation between *SATV* and *GPA*?
- **b.** If *SATV* score goes up by 100, by how much on average will the *GPA* change?
- **c.** If one student has an *SAT* verbal score that is 100 points higher than that of a second student, what can you say about their *GPA*s?

d. What is the intercept on the vertical axis (*GPA*-intercept) and what could it mean in terms of the data?

As you may know, the lowest possible *SATV* score is 200 (and the highest possible score is 800), so zero is not a possibility. Because of this, the *GPA*-intercept only tells you where the regression line crosses the *GPA*-axis, but it does not give the predicted *GPA* for any student.

The table below contains a listing of the 15 students in the sample, the *GPA* and *SATV* for each, and the predicted *GPA* for the first student based on the equation:

$\hat{GPA}_V = 1.97658 + 0.001906\ V.$

We will call the predicted *GPA*s from this equation $\hat{GPA}_V$, since this model uses the SAT verbal score.

SAT Verbal, GPA

Student Number	*GPA*	*SATV*	$\hat{GPA}_V$
1(*X*)	3.58	670	3.254
2	3.17	630	
3	2.31	490	
4	3.16	760	
5	3.39	450	
6	3.85	600	
7	2.55	490	
8	2.69	570	
9	3.19	620	
10	3.50	640	
11	2.92	730	
12	3.85	800	
13	3.11	640	
14	2.99	680	
15	3.08	510	

2. Consider student number 1, marked with an *X*.

a. Explain the numbers in the first row for student 1.

b. Fill in the column on the table for the predicted *GPA* for the rest of the students.

Residuals

3. Look carefully at the plot of the data with the regression line in Figure 6.3 and at the values you now have in the table on the previous page. Remember that a residual is the difference of the predicted and the actual *GPA*.

Predicting Student GPA

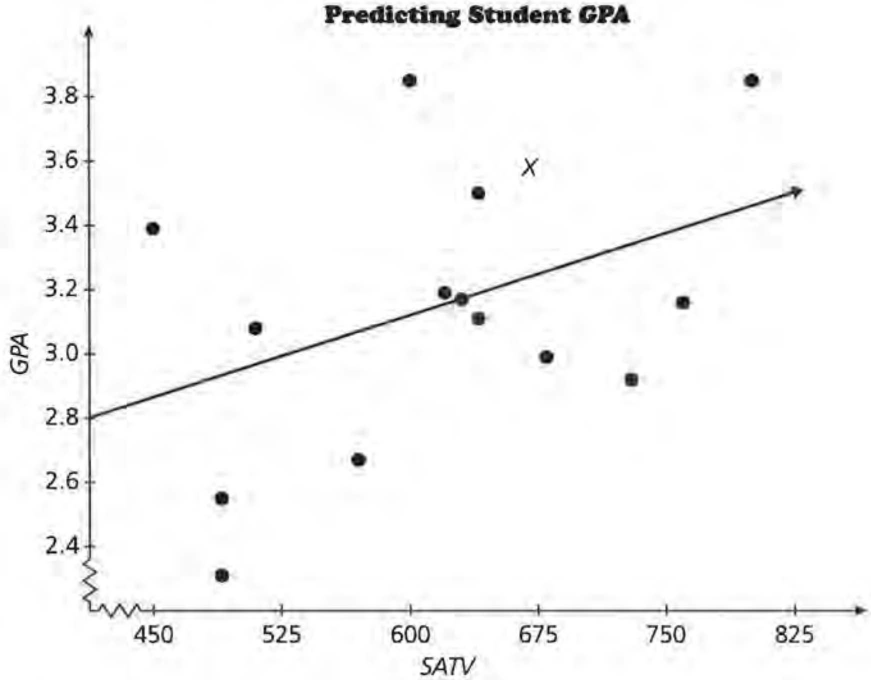


Figure 6.3

a. For which students will the prediction be the closest?

b. Let r_V represent the residuals. What is r_V for the first student? What does that tell you?

Residuals SAT Verbal, GPA

Student Number	GPA	$\hat{GPA}_V$	r_V
1	3.58	3.254	
2	3.17	3.178	
3	2.31	2.911	
4	3.16	3.425	
5	3.39	2.834	
6	3.85	3.120	
7	2.55	2.911	
8	2.69	3.063	
9	3.19	3.159	
10	3.50	3.197	
11	2.92	3.368	
12	3.85	3.502	
13	3.11	3.197	
14	2.99	3.273	
15	3.08	2.949	

c. Complete the table above by finding the residuals for all of the students. The table contains values of

- the actual *GPA*;
- the predicted values found by $\hat{GPA}_V = 1.97658 + 0.001906V$;
- and r_V, the residuals obtained using the equation for $\hat{GPA}_V$ for each of the 15 students and calculated by $r_V = GPA - \hat{GPA}_V$.

d. What does a negative sign in a residual indicate? Use an actual data point in your explanation.

4. Figure 6.4 contains a residual plot, the plot of $(SATV, r_V)$ obtained from predicting *GPA* using *SATV*. What does the residual plot tell you about how good this model is?

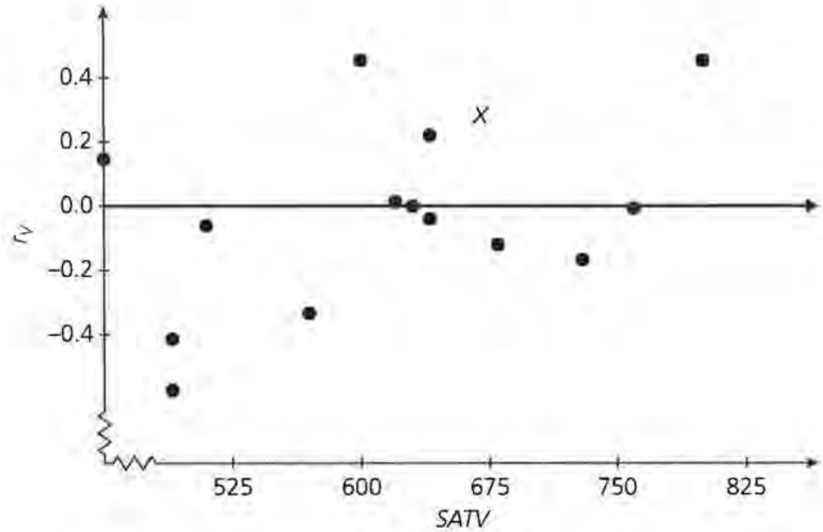


Figure 6.4

5. Remember from earlier work that you would like to have the smallest possible sum of squared residuals.

a. Why do you want the smallest sum of squared residuals?

b. If you did not have the *SATV* or *SATM* data and were trying to predict the *GPA* of one of the students, the best you could do would be to use the average, $\bar{y}$, as the predicted value. Find the mean *GPA*.

c. Verify that using the mean $\bar{y}$ as the predicted *GPA* for each of the 15 students yields a sum of squared residuals of $\Sigma(y_i - \bar{y})^2 = 2.709$.

d. What is the sum of squared residuals from the model that uses *SATV*? What conclusions can you make about the model?

The fitted model is $G\hat{P}A_V = 1.97658 + 0.001906V$. It is quite clear, however, that the model using *SATV* is not as good as it could be. There is an "error" in using the model. A more complete description of the relationship between *GPA* and *SATV* is $GPA = 1.97658 + 0.001906V + r_V = G\hat{P}A_V + r_V$. The residual term, r_V, represents the part of *GPA* that is not explained by *V*, the *SAT* verbal score, in the regression model.

Adding SAT Mathematics Score

SATV is not the only thing that determines the *GPA* of a student. As you saw earlier, one of the other factors is *SATM*. You would like the residuals to be as small as possible, so possibly *SATM* can help explain r_V.

6. Figure 5 is a scatter plot of r_V versus *SATM*.

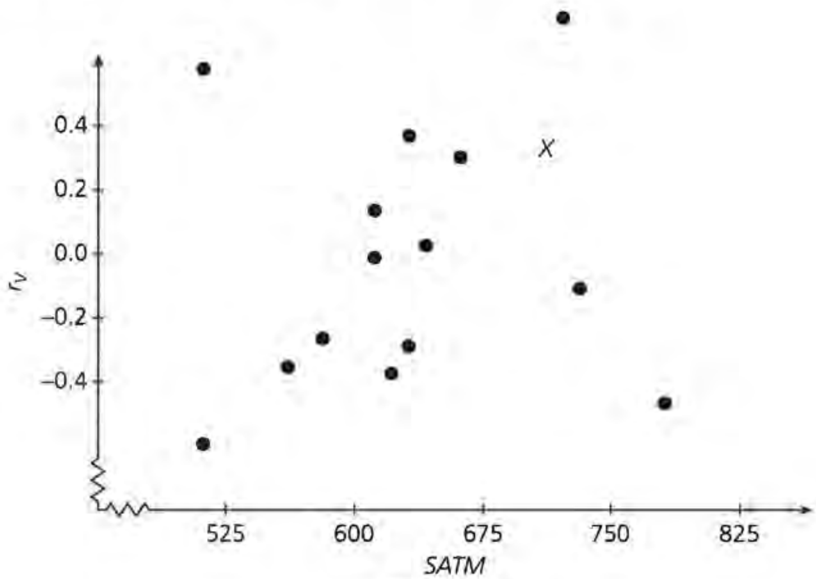


Figure 6.5

a. Explain in your own words what the plot represents.

b. What trend can you see in the plot?

7. It seems reasonable that *GPA* would depend on both *SATV* and *SATM*, probably with some error; in other words, $GPA = f(V) + h(M) + e$ for some functions f and h. So far you have $f(V) = 1.97658 + .001906V$ which predicts *GPA* with error r_V. Think of $h(M) + e$ as r_V. This indicates that r_V can be explained by M or $r_V = h(M) + e$. Thus, to find r_V, fit a regression line to $(SATM, r_V)$, and you will get the equation $\hat{r}V = -.52534 + 0.00083\ M$. This is the equation of the line shown in Figure 6.6.

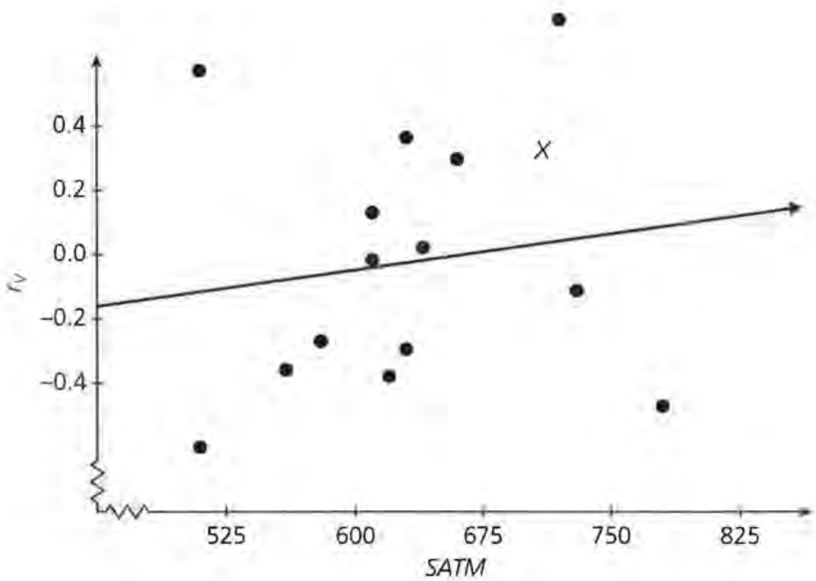


Figure 6.6

a. Explain what the equation above represents.

b. Use your own copy of the data to find the least squares regression equation for (M, r_V). How does it compare to the equation for the line in Figure 6.6?

Now, the equation for using only V to predict GPA is $G\hat{P}A_V = 1.97658 + 0.001906V$. Another statement of this relationship is that GPA depends on V and a residual amount r_V: $GPA = 1.97658 + 0.001906V + r_V$. If you replace r_V by the predicted value of r_V, from the equation $\hat{r}_V = -.52534 + 0.00083M$, you get a prediction equation for GPA in terms of V and M:

$$G\hat{P}A = 1.97658 + 0.001906V + (-.52534 + 0.00083M)$$

or

$$G\hat{P}A = 1.45124 + 0.001906V + 0.00083M$$

You found this equation by first using $SATV$ to predict GPA and then adding $SATM$ to the process, so the predicted values from this equation could be labeled $G\hat{P}A_{VM}$. The notation r_{VM} will denote the residuals from this model ($r_{VM} = GPA - G\hat{P}A_{VM}$).

What have you done so far? You began by finding the regression equation that uses $SATV$ to predict GPA. To obtain a better model, you modified the original prediction, $G\hat{P}A_V$, by refining the residual term. That is, $GPA = G\hat{P}A_V + r_V$ and $r_V = \hat{r}_V + r_{VM}$. Your new equation is now $GPA = G\hat{P}A_V + r_{VM}$. College GPA is a function of SAT verbal, SAT math, and still another "error."

8. The table below summarizes what you should have so far. $G\hat{P}A_{VM} = 1.45124 + 0.001906V + 0.00083M$ represents the prediction model using both $SATM$ and $SATV$, and r_{VM} represents the residuals found using this model.

Residuals *SAT* Verbal, Math, GPA

Student Number	*GPA*	*SATV*	*SATM*	$G\hat{P}A_{VM}$	Residual (r_{VM})
1x	3.58	670	710	3.318	0.262
2	3.17	630	610		
3	2.31	490	510		
4	3.16	760	580		
5	3.39	450	510		
6	3.85	600	720		
7	2.55	490	560		
8	2.69	570	620		
9	3.19	620	640		
10	3.50	640	660		
11	2.92	730	780		
12	3.85	800	630		
13	3.11	640	730		
14	2.99	680	630		
15	3.08	510	610		

a. Explain what the values in the first row represent for student X.

b. Complete the table.

c. Find the sum of the squared residuals r_{VM}. How does it compare to the sum of the squared residuals r_V?

Can you do even better than this? That is, can you improve the model

$$G\hat{P}A_{VM} = 1.45124 + 0.001906V + 0.00083M$$

to get even better predictions?

Iterating the Process

You created a model, which will be called model VM, by first regressing GPA on $SATV$. You then computed residuals, which you labeled r_V, and regressed the residuals, r_V, on *SATM*. You combined the two equations

$$GPA = 1.97658 + 0.001906V + r_V$$

and

$$\hat{r}_V = -.52534 + 0.00083M$$

to get

$$\hat{GPA}_{VM} = 1.45124 + 0.001906V + 0.00083M.$$

Another way to express the relationship is to say that *GPA* depends on *V* and *M* plus a residual, labeled r_{VM}: $GPA = 1.45124 + 0.001906V + 0.00083M + r_{VM}$. You would like to improve the model and get smaller residuals, but *V* and *M* are the only predictor variables available.

The last step was to add *SATM* to the model. How are the residuals r_{VM} related to *SATV*? Figure 6.7 is a scatter plot of $(SATV, r_{VM})$.

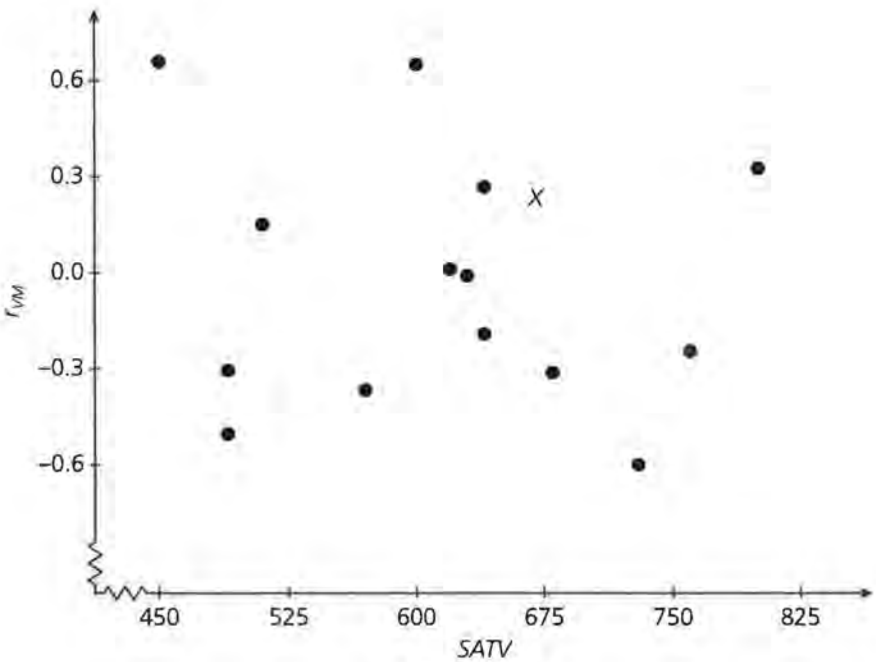


Figure 6.7

If *SATV* contributed a great deal to the determination of r_{VM}, then the pattern in this plot would be closer to linear. Apparently there is not a very strong relationship here, but you can capture the trend that exists by fitting a regression line (Figure 6.8).

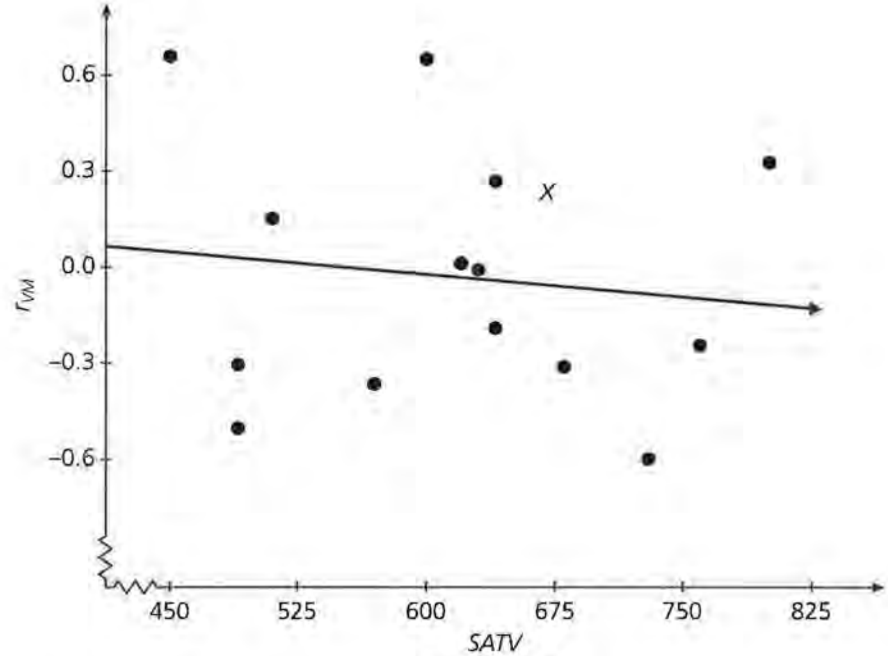


Figure 6.8

9. Find the prediction equation for this regression line, $\hat{r}_{VM}$. (Note: you are plotting *SATV* and the residuals from the previous model.)

   a. The equation to predict *GPA* can be written

   $GPA = 1.45124 + 0.001906V + 0.00083M + \hat{r}_{VM}$.

   This prediction equation will be labeled $\hat{GPA}_{VMV}$. Explain why.

   b. Write the equation for $\hat{GPA}_{VMV}$ substituting for $\hat{r}_{VM}$.

   c. Find the sum of squared residuals for model *VMV*. Compare model *VMV* to model *VM*.

You can continue the process of adding *V*, followed by *M*, followed by *V*, and so on. The best model you have so far is model *VMV*, which yields predicted values $\hat{GPA}_{VMV}$ from an expression of the form $b_0 + b_1V + b_2M$. Another statement of the situation is that *GPA* depends on *V* and *M* plus a residual, which can be labeled r_{VMV}:

$GPA = b_0 + b_1V + b_2M + r_{VMV}$

Following the pattern of adding V, then M, then V, consider how r_{VMV} is related to M. Figure 6.9 is a scatter plot of r_{VMV} versus $SATM$.

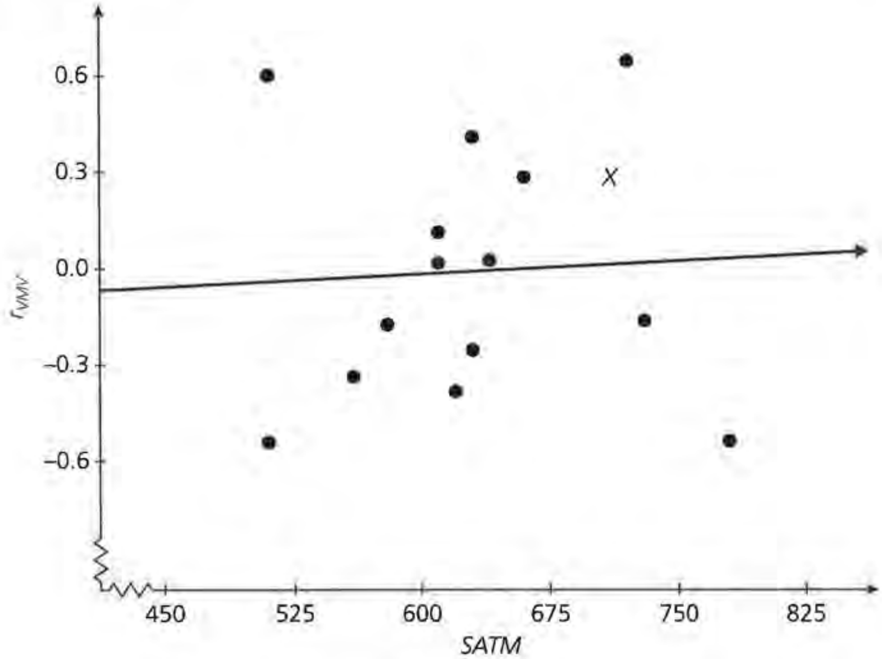


Figure 6.9

10. There is only a very weak association between r_{VMV} and $SATM$. The equation of the regression line is $\hat{r}_{VMV} = -0.15916 + 0.000252M$.

a. Write the equation for $G\hat{P}A_{VMVM}$.

b. The sum of squared residuals from this model is 2.086. Compare model $VMVM$ to model VMV.

11. Reflect back over the lesson so far.

a. Fill in the table with the corresponding equations and squared sum of residuals.

Sum of Squared Residuals Verbal, Math

	Equation	Sum of Squared Residuals
GPA , SATV	$G\hat{P}A_V =$	
GPA, SATV, SATM	$G\hat{P}A_{VM} =$	
GPA, SATV, SATM, SATV	$G\hat{P}A_{VMV} =$	
GPA, SATV, SATM, SATV, SATM	$G\hat{P}A_{VMVM} =$	

b. If you wrote down the equations for $G\hat{P}A_V$, $G\hat{P}A_{VM}$, $G\hat{P}A_{VMV}$, and $G\hat{P}A_{VMVM}$ and then continued the iterative process to add $G\hat{P}A_{VMVMV}$, $G\hat{P}A_{VMVMVM}$, and so on, what do you think you would see?

Summary

One process of finding a model when two variables, x_A and x_B, are used to predict a third, y, is an iterative one. You begin by finding a regression line for (x_A, y). The residuals from this model can be written as a function of the second variable, x_B. The regression line for the residuals from the plot of (x_B, r_A) can be used to adjust the original model. The residuals from (x_B, r_A) can be explained by using (x_A, r_{AB}) and this can be used to obtain a new model. The process continues, alternating between the variables and the residuals, and each time you adjust the model you had before.

Practice and Applications

Using *M* followed by *V*

12. Rather than using *SATV* first as a predictor of *GPA* and then adding *SATM*, you could use these two variables in the reverse order. First, plot (*M*, *GPA*) and fit a regression line (Figure 6.10). (Because this model uses *M*, and only *M*, to predict *GPA*, the notation $G\hat{P}A_M$ will denote the predicted values. In the same way, the residuals can be called r_M.)

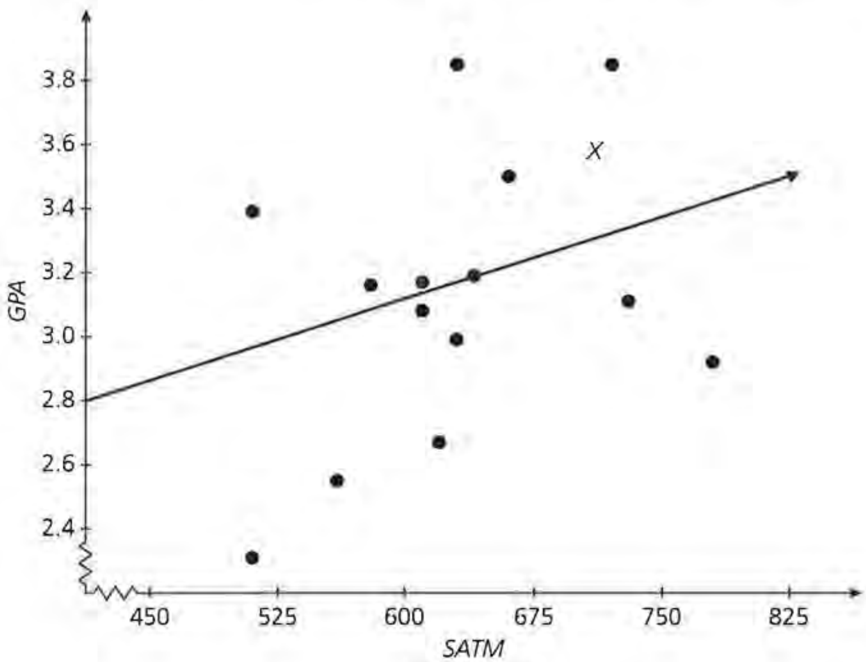


Figure 6.10

a. What is the equation of the least squares regression line?

b. Use your equation to complete the table below.

Residuals *SAT* Math, *GPA*

Student Number	*GPA*	*SATM*	Predicted *GPA* $\hat{GPA}_M$	Residuals r_M
1	3.58	710		
2	3.17	610		
3	2.31	510		
4	3.16	580		
5	3.39	510		
6	3.85	720		
7	2.55	560		
8	2.69	620		
9	3.19	640		
10	3.50	660		
11	2.92	780		
12	3.85	630		
13	3.11	730		
14	2.99	630		
15	3.08	610		

A plot of r_M versus $SATM$ shows no alarming pattern (Figure 6.11).

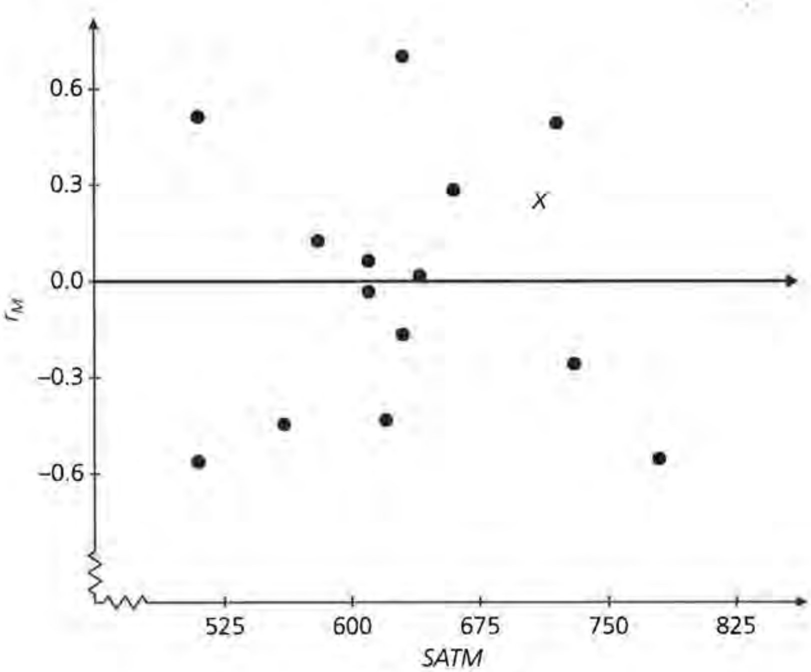


Figure 6.11

c. Find the sum of squared residuals from the model that uses *SATM*. How does it compare to the sum of squared residuals for the model that uses *SATV*?

13. You can improve the model that uses only *SATM* by adding *SATV* to it. A scatter plot of r_M versus *SATV* is in Figure 6.12.

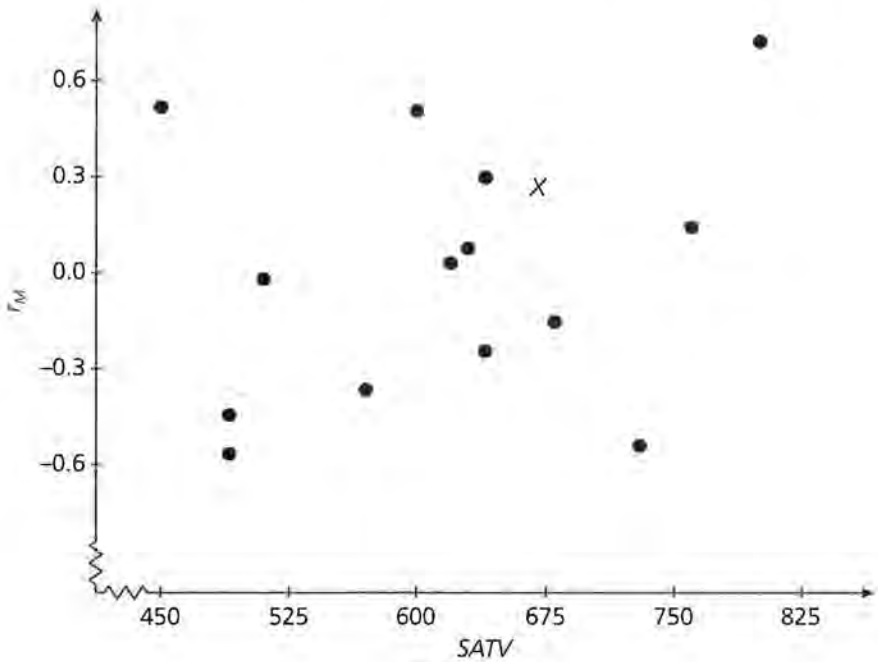


Figure 6.12

a. Describe the association in the plot.

b. The least squares regression line has been plotted in Figure 6.13. Find the equation.

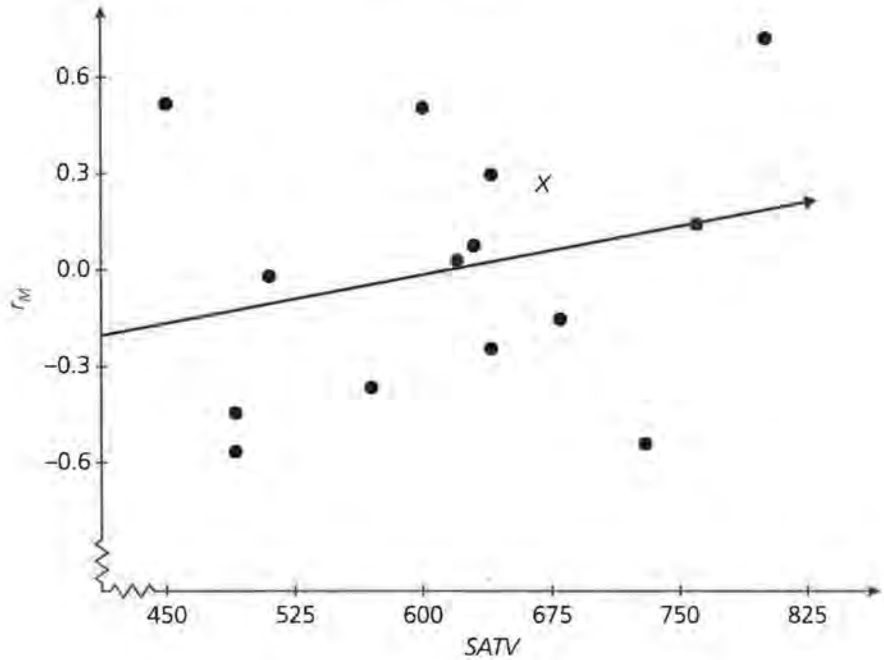


Figure 6.13

c. Another statement of this relationship is that *GPA* depends on *M* and a residual amount r_M:

$GPA = a_0 + a_1M + r_M$.

If you replace r_M by the predicted value from the equation $\hat{r}_M$, you get a prediction equation for *GPA* in terms of both *V* and *M*:

$$G\hat{P}A = a_0 + a_1M + (c_1 + a_2V)$$

or

$$G\hat{P}A = a_3 + a_1M + a_2V$$

Write the new prediction equation for $G\hat{P}A$ and use your equation to fill in the $G\hat{P}A_{MV}$ values and the corresponding residuals for each of the 15 students in the table that follows. (You found this equation by first using *M* to predict *GPA* and then adding *V* to the process, so call the predicted values from this equation $G\hat{P}A_{MV}$ and the resulting residuals r_{MV}.)

Residuals SAT Math, Verbal, GPA

Student Number	*GPA*	*SATV*	*SATM*	$\hat{GPA}_{MV}$	Residual (r_{MV})
1	3.58	670	710		
2	3.17	630	610		
3	2.31	490	510		
4	3.16	760	580		
5	3.39	450	510		
6	3.85	600	720		
7	2.55	490	560		
8	2.69	570	620		
9	3.19	620	640		
10	3.50	640	660		
11	2.92	730	780		
12	3.85	800	630		
13	3.11	640	730		
14	2.99	680	630		
15	3.08	510	610		

d. What is the sum of the squared residuals for r_{MV}? How does it compare to the sum of squared residuals for r_{VM}?

14. You used M and then added V to the model. Now use M again by plotting r_{MV} against M and fitting a regression line.

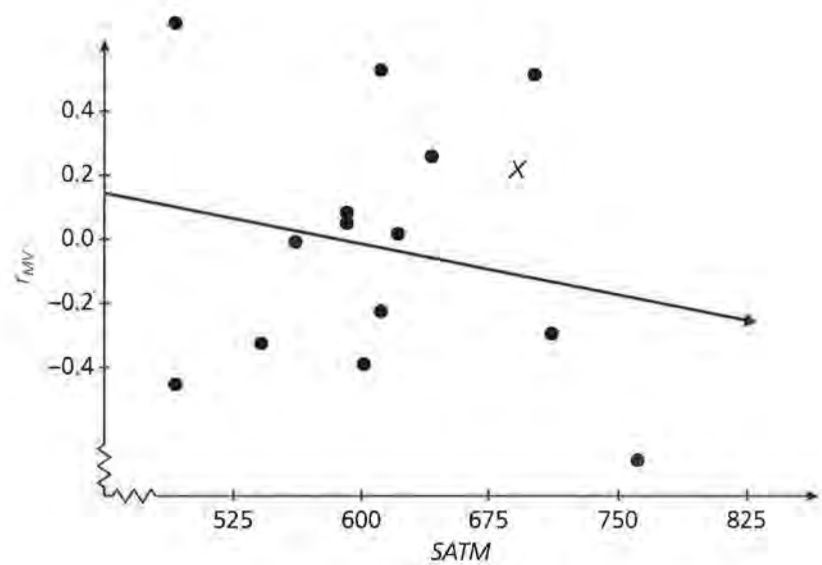


Figure 6.14

a. The equation of the regression line is $\hat{r}_{MV} = 0.453533 - 0.000716M$. Find the refined prediction equation for GPA by substituting $0.453533 - 0.000716M$ for r_{MV} in the equation

$$GPA = 1.142410328 + 0.0009815276V + 0.0022205651M + r_{MV}.$$

b. Why does it make sense to call the predicted values from this equation $\hat{GPA}_{MVM}$?

c. The sum of squared residuals for model MVM is 2.103. How do these compare to the sum of squared residuals for model MV?

15. Fill in the chart using the results of problems 12, 13, and 14.

Sum of Squared Residuals Math Verbal

	Equation	Sum of Squared Residuals
GPA, SATM	$\hat{GPA}_M =$	
GPA, SATM, SATV	$\hat{GPA}_{MV} =$	
GPA, SATM, SATV, SATM	$\hat{GPA}_{MVM} =$	

What observations can you make?

16. The plots below are of the residuals for each successive iteration. What observation can you make? Does the development of the model support your observation? If so, how?

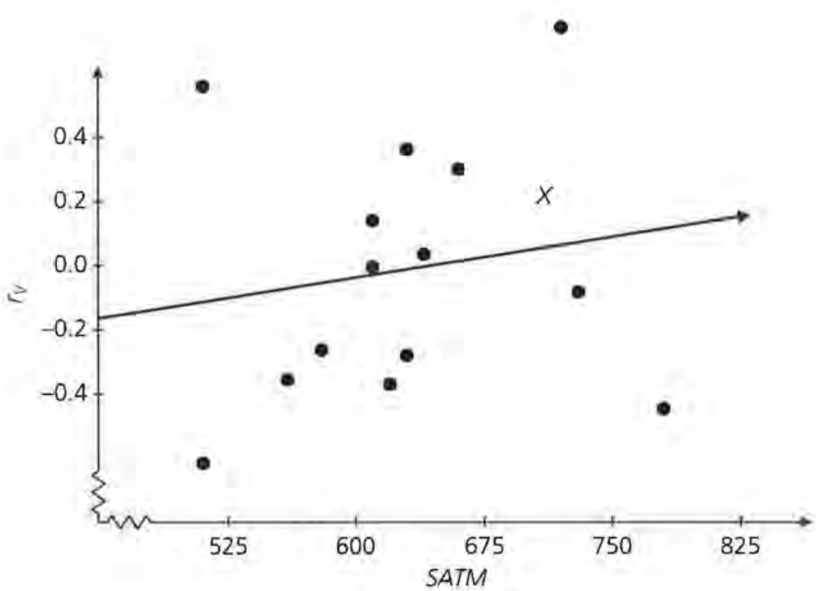


Figure 6.15

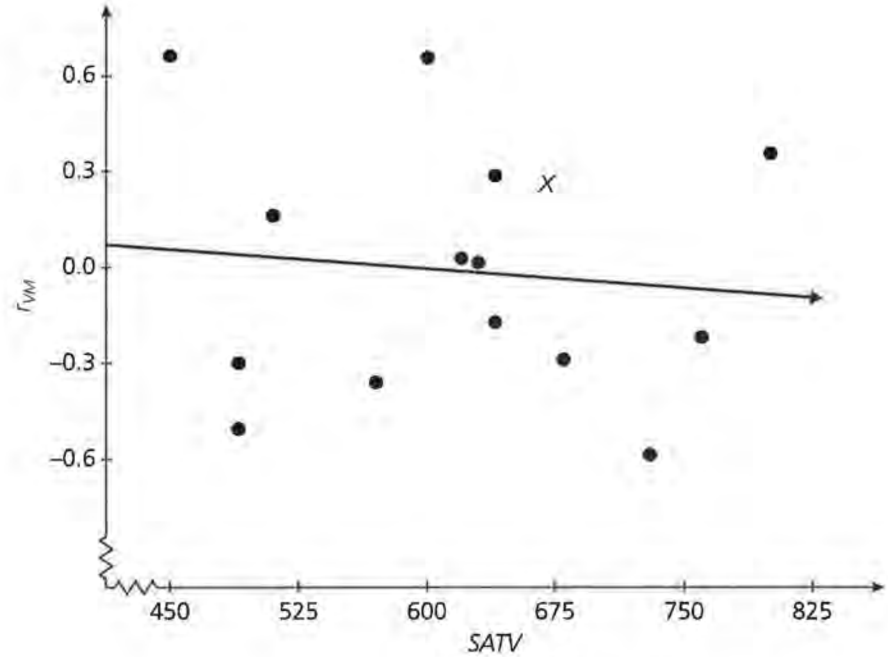


Figure 6.16

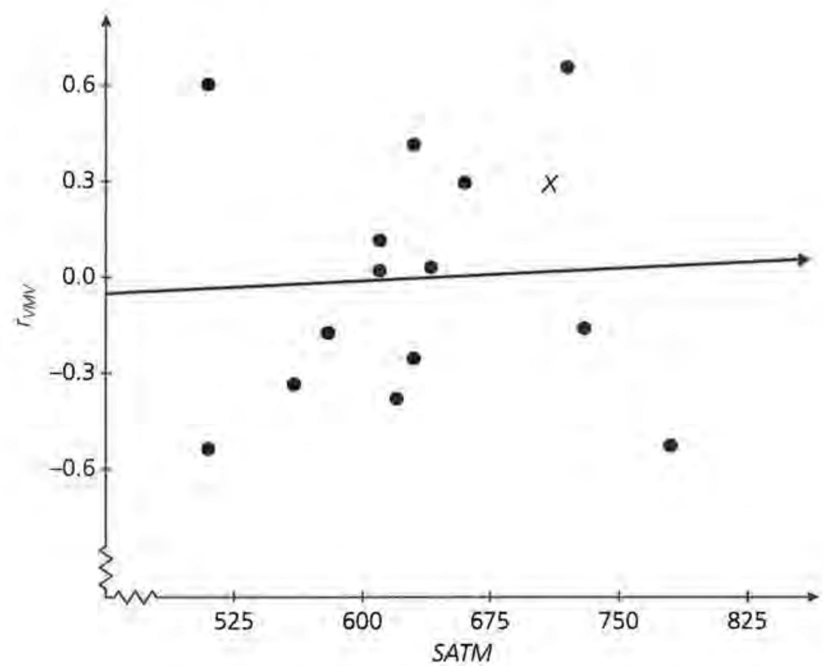


Figure 6.17

17. Write a summary of the process that can be used to find a model if two variables are involved in making a prediction.

LESSON 7

Comparing Order in Regression

Does it make a difference if you start finding a model using either *SATM* or *SATV*?

Which variable should you use first, *SATM* or *SATV*? How would you choose?

What would you do if you had more variables to use in predicting *GPA*?

OBJECTIVES

Investigate the impact of order in the regression process.

Recognize what will occur as the process continues.

You have used *SATV* and *SATM* to make predictions of *GPA* in two different ways. One approach was to start with *SATV* as a predictor and then to take *SATM* into consideration; this led to the list of predictions labeled $G\hat{P}A_V$, $G\hat{P}A_{VM}$, and so on. The second approach was to reverse the order, first using *SATM* as a predictor and then adding *SATV* to the model; you labeled the predicted values from this model $G\hat{P}A_M$, $G\hat{P}A_{MV}$, and so on. In this lesson you will compare the two different approaches by looking at the coefficients, the graphs, and the sums of the squared residuals for each iteration.

INVESTIGATE

Comparing Two Models

The equations to predict *GPA* if you began with *SATV* and if you began with *SATM* were:

$G\hat{P}A_V = 1.97658 + 0.001906V$ and
$G\hat{P}A_M = 1.74964 + 0.002221M.$

The coefficients in the two models are different, but how different are the results the models produce? One way to compare the two is to consider the predictions they give for a single student, such as student number 1. Student number 1 was plotted with an X in all of the scatter plots. Notice that in both of the scatter plots in Figure 7.1, student number 1 is above most of the other points. In particular, student number 1 has a positive residual in each case.

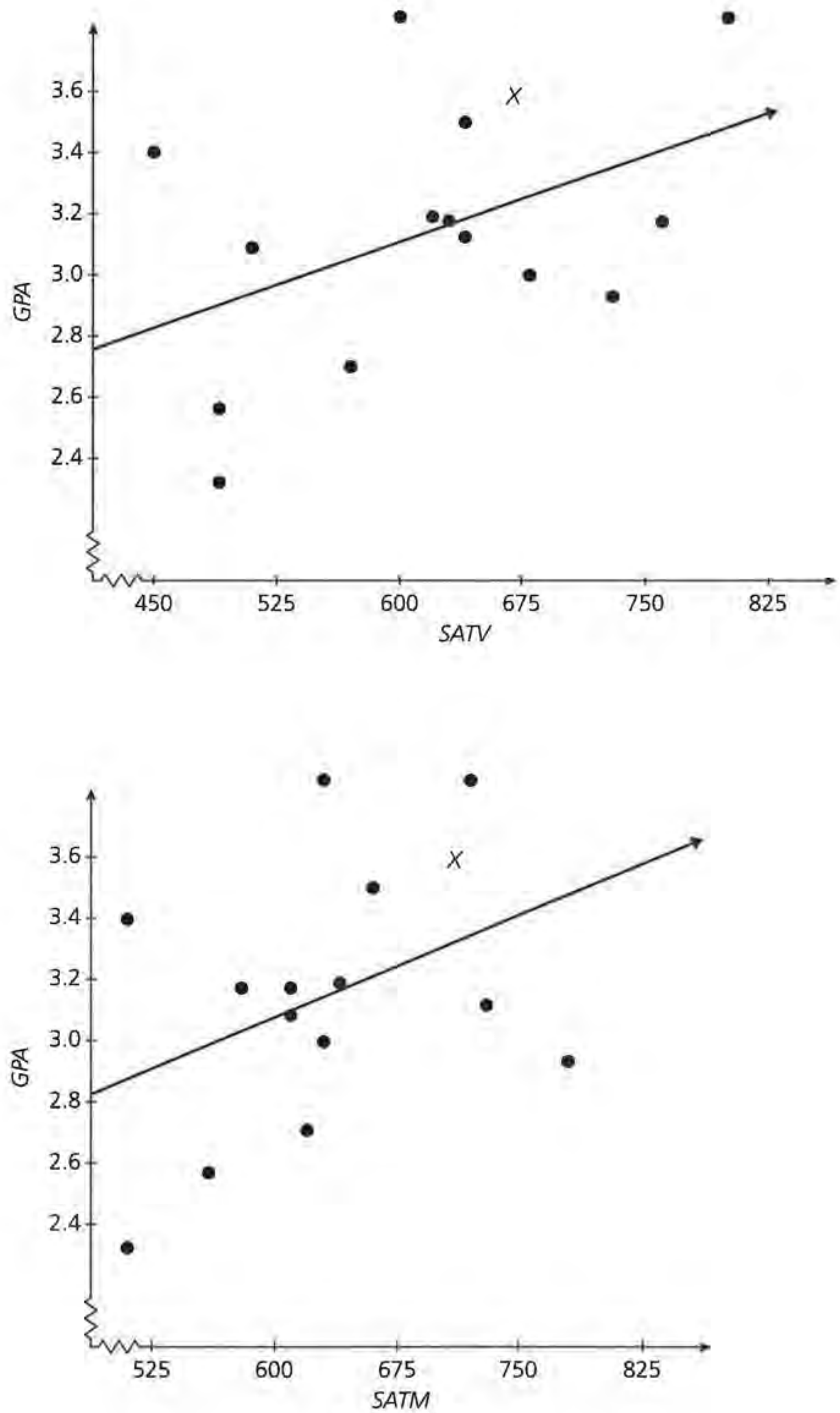


Figure 7.1

Discussion and Practice

1. What are the predicted *GPA* values from the two models for student number 1?

a. Complete the table with the predicted values for each student.

GPA Verbal, GPA Math

Student Number	*GPA*	*SATV*	*SATM*	$\hat{GPA}_V$	$\hat{GPA}_M$
1*x*	3.58	670	710		
2	3.17	630	610		
3	2.31	490	510		
4	3.16	760	580		
5	3.39	450	510		
6	3.85	600	720		
7	2.55	490	560		
8	2.69	570	620		
9	3.19	620	640		
10	3.50	640	660		
11	2.92	730	780		
12	3.85	800	630		
13	3.11	640	730		
14	2.99	680	630		
15	3.08	510	610		

b. If GPA_V and GPA_M were equally good at predicting *GPA*, describe the plot of $(\hat{GPA}_V, \hat{GPA}_M)$.

c. Make a plot of $(\hat{GPA}_V, \hat{GPA}_M)$. How are the predictions from the two models related to each other?

Consider models *MV* and *VM*. You constructed model *VM* by first using *SATV* to predict *GPA*, computing the residuals r_V, and then regressing r_V on *M*. You constructed model *MV* by first using *SATM* to predict *GPA*, computing the residuals, r_M, and then regressing r_M on *V*. The two equations that resulted were

$$\hat{GPA}_{VM} = 1.45124 + 0.001906V + 0.00083M$$

and

$$\hat{GPA}_{MV} = 1.14241 + 0.0009815276V + 0.0022205651M.$$

The order in which you add *V* and *M* to the model affects the coefficients in the model. That is, the coefficients in the model

VM are not exactly the same as the coefficients in the model MV.

2. What are the predicted *GPA*s for student 1 using models MV and VM?

a. How do these results compare to those for model V and model M?

b. If the models were equivalent, what do you know about the two predictions?

c. You learned in the first part of this module that the equations for $\hat{GPA}_{VM}$ and $\hat{GPA}_{MV}$ would represent planes. Decide whether the two planes are parallel or intersect and justify your decision.

Graphs

It is not sufficient to compare the models on only one student. Just as you did in problem 1, consider the ordered pair $(\hat{GPA}_{VM}, \hat{GPA}_{MV})$ for each of the 15 students.

3. Figure 7.2 is a plot of those pairs of predictions, with student number 1 again plotted using an X.

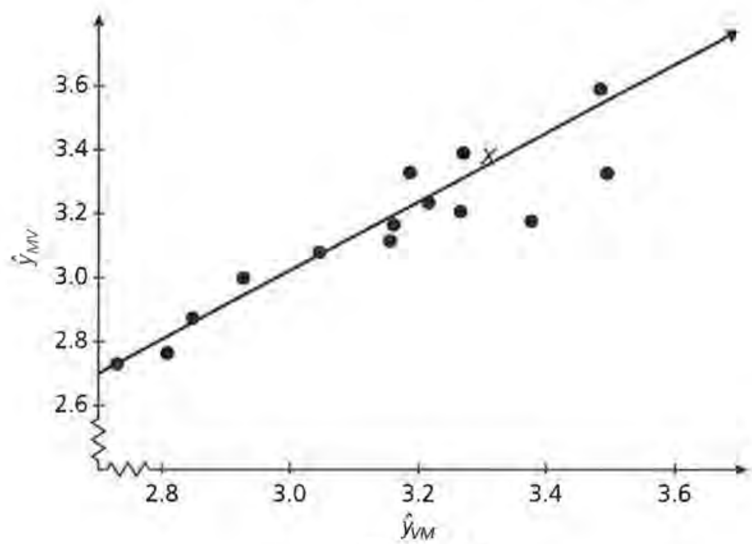


Figure 7.2

a. What does the line in Figure 7.2 represent?

b. Describe the trend in the plot. What does this tell you about the relationship between the *GPA*s predicted by the two models?

c. If the two models made exactly the same prediction for each of the 15 *GPA*s, describe what the plot would look like.

4. The prediction equations for $G\hat{P}A_{VMV}$ and $G\hat{P}A_{MVM}$ are similar but have different coefficients:

$G\hat{P}A_{VMV} = 1.664753 + 0.0015707V + 0.00083M$

and

$G\hat{P}A_{MVM} = 1.59594333 + 0.0009815276V + 0.0015045651M$

a. Compare the predictions that the models *VMV* and *MVM* give for the *GPA* of student number 1.

b. How do these predictions compare to those using model *VM* and model *MV*?

c. Figure 3 is a scatter plot of $(G\hat{P}A_{VMV}, G\hat{P}A_{MVM})$ for each of the 15 students. Describe the trend in the plot.

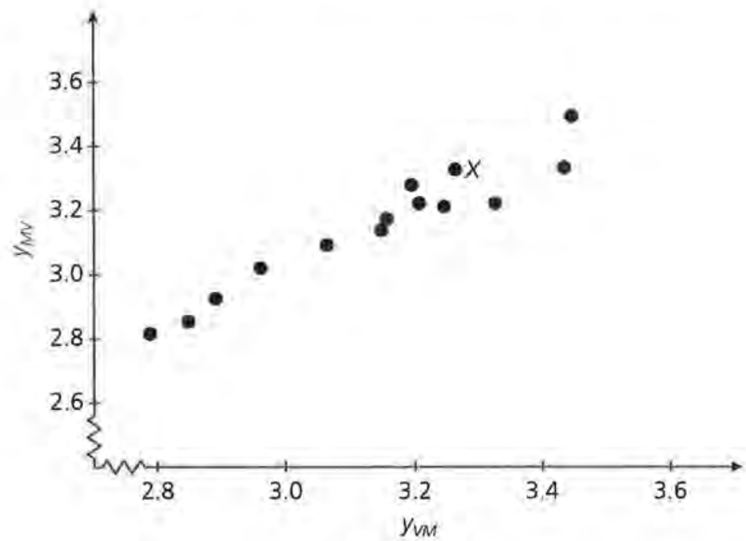


Figure 7.3

d. Look carefully at the plot you made for problem 1 and at Figure 7.2 and Figure 7.3. What observation can you make?

Residuals

Another way to compare the two models is to consider the residuals from each.

5. This table has the predicted *GPA* for student 1 for each of the models.

Residuals Student 1

Student	GPA	$\hat{GPA}_V$	$\hat{GPA}_M$	$\hat{GPA}_{VM}$	$\hat{GPA}_{MV}$	$\hat{GPA}_{VMV}$	$\hat{GPA}_{MVM}$
1	3.58	3.2536	3.3262	3.3176	3.3766	3.3064	3.3218

a. Find and compare the residuals for each case:

r_V, r_M
r_{VM}, r_{MV}
r_{VMV}, r_{MVM}

b. What conclusions can you draw?

c. Why is investigating the sum of squared residuals a better way to compare the models than using an individual student?

6. The sum of squared residuals for each of the models, except for the model *MVMV*, is given in the following table.

Comparing Sum of Squared Residuals

Model	Sum of Squared Residuals for Model	Model	Sum of Squared Residuals for Model
V	2.168	*M*	2.290
VM	2.110	*MV*	2.147
VMV	2.092	*MVM*	2.103
VMVM	2.086	*MVMV*	

a. Back in Lesson 6, you learned that if you had no information at all about SAT scores and were trying to predict GPA using only the mean GPA, the sum of the squared residuals for the mean would be 2.709. What does this say about using either of the two models to begin with?

b. Describe the trend in the sum of squared residuals for each group of models in the table.

c. Use the information above to comment on the statement: The predictions from the model that uses only *M* are about as accurate, overall, as the predictions generated by the model that uses only *V*. The equations for the two models are rather different, but overall, each model makes predictions that are reasonable.

d. Compare the difference between the sum of squared residuals for model *V* and model *M;* model *VM* and model *MV;* model *VMV* and model *MVM.*

e. What do you think the sum of squared residuals is for model *MVMV?*

Conclusion

If you continue to repeat the process outlined in Lesson 6 using either order, you end up with the following prediction equation:

$$\hat{GPA} = 1.52885 + 0.00141V + 0.001192M$$

This is the *least-squares multiple regression model* for predicting *GPA* using *SATV* and *SATM.*

7. Use the multiple regression model to predict the *GPA* for each of the students.

a. What is the sum of squared residuals for this model? How does it compare to the sum of squared residuals from the earlier models?

b. If you were a college entrance administrator at the school from which these data came and a student applied for entrance with an *SATV* of 450 and an *SATM* of 480, what would you tell that student?

8. Think back to the original problem. How can you use both *SATV* and *SATM* to predict college *GPA?* You might have wondered what would happen if you had added the two values and worked with the total *SAT* score rather than going through this entire process. That is, why not add *SATM* to *SATV* to get *SAT TOTAL* and use *SAT TOTAL* to predict *GPA?*

a. Do you think that this will give better predictions then you found with the models above? Use the information in the table that follows to find the linear regression model for (*SAT TOTAL*, *GPA*) and calculate the sum of squared residuals to help you answer the question.

b. Think about different variables that could be used to predict *GPA*. Would it always make sense to add two predictor variables together?

Sum of Math, Verbal

Student Number	*GPA*	*SATV*	*SATM*	*SAT TOTAL*	Predicted *GPA*	Residuals
1	3.58	670	710			
2	3.17	630	610			
3	2.31	490	510			
4	3.16	760	580			
5	3.39	450	510			
6	3.85	600	720			
7	2.55	490	560			
8	2.69	570	620			
9	3.19	620	640			
10	3.50	640	660			
11	2.92	730	780			
12	3.85	800	630			
13	3.11	640	730			
14	2.99	680	630			
15	3.08	510	610			

c. Why is it better to use *SATM* and *SATV* as two separate predictors rather than combining them into *SAT TOTAL?*

Summary

You can find a prediction equation in two variables by beginning with one variable, taking the residuals from the model, regressing them on the other variable, and continuing to repeat the process. The equations are slightly different, depending on which variable you choose to begin the process. As you compare the *GPA* predictions for each step, however, the differences begin to lessen. The plots of predicted values from the iterations using different orders of the variables $G\hat{P}A_{VM}$ and $G\hat{P}A_{MV}$ converge to the straight line $y = x$. The sums of the squared residuals for each case converge toward a common sum. The equation you approach is called the *least squares multiple regression model.*

Practice and Applications

9. Work with a partner on the following data about colleges. One of you should begin with *room and board cost* (R) and use it to predict *tuition and fees*. The other should begin with *% of need met* (N) and use it to predict *tuition and fees*. Use the process of iteration to generate $\hat{T}_{RNRN}$ and $\hat{T}_{NRNR}$ and compare your results. (Note that *% of need met* is a measure of financial aid at the college.)

Best College Buys

School	Tuition and Fees ($)	Room and Board ($)	% of Need Met
California Inst. of Tech.	17,586	6,620	100
New College of U. of South Florida	7,950	3,847	93
Northwest Missouri State	3,975	3,330	80
Rice University	12,034	5,900	100
State U. of NY at Binghamton	8,679	4,654	62
State U. of NY at Albany	8,856	4,836	80
Spelman College	8,875	5,890	38
Trenton State College	6,658	5,650	90
U. of Illinois at Urbana/ Champaign	9,130	4,408	75
U. of North Carolina at Chapel Hill	10,162	5,350	95

Source: *Money Guide*, 1996 Edition

10. The following table gives the width, length, and perimeter of each of seven rectangles. Suppose you did not know that $P = 2W + 2L$, and you wanted to determine the relationship between P, W, and L by using regression. Use the iterative process to find $\hat{P}_{WLWL}$, etc. How many iterations does it take until the sum of squared residuals gets close to zero and the fitted model gets close to $P = 2W + 2L$?

W = width	L = length	P = perimeter
2	2	8
3	5	16
4	4	16
1	1	4
5	5	20
2	3	10
6	2	16

11. What are some underlying assumptions that are necessary if a least squares multiple regression model is to be a good model? Why do you think the model is called the least squares multiple regression model?

12. Explain the role of residuals in the process of finding the least squares multiple regression model.

13. Make a conjecture about how you might find a model if you had three variables available to determine a prediction equation.

LESSON 8

Matrices and Multiple Regression

How can you find a model more quickly? Is there a method that would be more efficient?

Is there any way that another mathematical approach for finding a model can help?

The iterative process outlined in Lessons 6 and 7 seems to be a long, involved process. Is there an easier way to find a good model using both variables? Is there a mathematical technique that can help? What happens if there are more than two variables that can be used to find a prediction equation? In this lesson you will investigate a shortcut formula using matrices that will enable you to find a model with any number of variables.

OBJECTIVES

Use matrices to represent the process of regression.

Justify the process mathematically.

Generalize the formula to many dimensions.

INVESTIGATE

Using Matrices in Regression

In Lesson 5 you learned to write a system of matrices to represent a regression rule. Equations of the form $aw + bx = c$ were written as

$$[w\ x] \cdot \begin{bmatrix} a \\ b \end{bmatrix} = [c]$$

When used with a least squares regression equation, the matrix $[w\ x]$ became an $n \times 2$ matrix that contained a column vector of 1s and a column of the x-values; [c] became an $n \times 1$ column vector with the predicted y-values. $\begin{bmatrix} a \\ b \end{bmatrix}$ was the coefficient matrix determined by the least squares regression equation. So the matrix equation could be written:

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{bmatrix} \cdot \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 1 & y_1 \\ 1 & y_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & y_n \end{bmatrix}$$

You have been considering prediction equations of the form

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2.$$

For example, the model labeled *VM* is

$$\hat{GPA}_{VM} = 1.45124 + 0.001906V + 0.00083M.$$

In this model, $b_0 = 1.45124$, $b_1 = 0.001906$, and $b_2 = 0.00083$.

You can rewrite the model as

$$b_0(1) + b_1 V + b_2 M = \hat{GPA}.$$

to make explicit the fact that b_1 is multiplied by *V*, b_2 is multiplied by *M*, and b_0 is multiplied by 1. In matrix form, this is

$$[1\ V\ M] \cdot \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = [\hat{GPA}].$$

A matrix formulation of this model for predicting *GPA*s is $\mathrm{Xb} = \hat{\mathrm{Y}}$.

$$\mathrm{X} = \begin{bmatrix} 1 & 670 & 710 \\ 1 & 630 & 610 \\ 1 & 490 & 510 \\ 1 & 760 & 580 \\ 1 & 450 & 510 \\ 1 & 600 & 720 \\ 1 & 490 & 560 \\ 1 & 570 & 620 \\ 1 & 620 & 640 \\ 1 & 640 & 660 \\ 1 & 730 & 780 \\ 1 & 800 & 630 \\ 1 & 640 & 730 \\ 1 & 680 & 630 \\ 1 & 510 & 610 \end{bmatrix}$$

The first column in X is a column of 1s, the second column in X holds the values of *SATV*, and the third column holds the values of *SATM*.

The matrix b is a 3×1 matrix that holds the parameters b_0, b_1, and b_2 for the equation $\hat{GPA} = b_0 + b_1 V + b_2 M$. For model *VM*,

$$\mathrm{b} = \begin{bmatrix} 1.45124 \\ 0.001906 \\ 0.00083 \end{bmatrix}.$$

When you multiply X by b you have

$$\begin{bmatrix} 1 & 670 & 710 \\ 1 & 630 & 610 \\ 1 & 490 & 510 \\ 1 & 760 & 580 \\ 1 & 450 & 510 \\ 1 & 600 & 720 \\ 1 & 490 & 560 \\ 1 & 570 & 620 \\ 1 & 620 & 640 \\ 1 & 640 & 660 \\ 1 & 730 & 780 \\ 1 & 800 & 630 \\ 1 & 640 & 730 \\ 1 & 680 & 630 \\ 1 & 510 & 610 \end{bmatrix} \cdot \begin{bmatrix} 1.45124 \\ 0.001906 \\ 0.00083 \end{bmatrix}.$$

Discussion and Practice

1. Enter the data for X and b into two matrices on your calculator.

a. Find the predicted *GPA*s, Xb.

$$\text{Let } [\mathbf{Y}] = \begin{bmatrix} 3.58 \\ \cdot \\ \cdot \\ 3.08 \end{bmatrix}.$$

Enter the data for Y into a matrix on your calculator.

b. Find the residuals using matrices.

c. What is the sum of squared residuals? Explain how to find this using matrices.

In the example above, you used the coefficients from the model *VM*. In general, how can you find good choices for b_0, b_1, and b_2? You want to find a matrix b such that Xb is a good approximation of Y, where Y is an $n \times 1$ matrix that contains the known values of the variable y. To find the matrix formula for the least-squares solution (to find the best values of b_0, b_1, and b_2) you have to find the best b to use in the equation Y = Xb. You determined a way to find b by the iterative process you used in Lessons 6 and 7.

2. You might also think about using the matrix method we used in Lesson 1 to find b_0, b_1, and b_2 that in general will give the best prediction for *GPA* based on *SATV* and *SATM*. If Xb = Y, then b = X^{-1}Y. Will this method work? Why or why not?

A Matrix Solution

Fortunately, there is another method, and it uses matrices! If you premultiply both sides of the equation Xb = Y by the transpose of X, you get $X^TXb = X^TY$. X^TX has an inverse as long as no column in X is a linear combination of any other column. You can find the inverse of X^TX and use that to find b.

$$Xb = Y$$
$$X^TXb = X^TY$$
$$((X^TX)^{-1})(X^TXb) = (X^TX)^{-1}(X^TY)$$
$$\text{or } ((X^TX)^{-1})(X^TX)b = (X^TX)^{-1}X^TY$$
$$\text{so } b = (X^TX)^{-1}X^TY$$

Therefore, if you premultiply Xb = Y by $(X^TX)^{-1}X^T$, you get $b = (X^TX)^{-1}X^TY$. Note that b is not the exact solution to the equation Xb = Y because you did not directly "undo" multiplication by X. Instead, you found a way to create the possibility of an inverse, so what you have found is the "best" approximation you can for the solution. In the last part of this section, you will see why it is the "best."

3. Consider the points (2, 5) and (4, 7) and the graph of the line below.

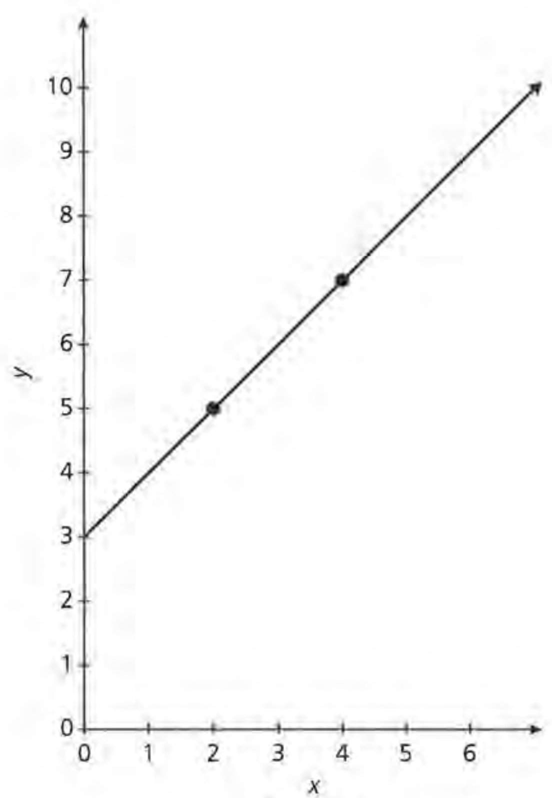


Figure 8.1

a. Write the equation of the line by using the two points. What is the sum of the squared residuals?

b. Write the equation of the least squares linear regression line that would fit two points.

c. Write the equation of the line you would find by using the matrix formula to determine the coefficients as described above.

All three of these methods give the same answer, that is, the same values for the slope and the intercept of the least squares regression line when there are only two points.

4. Hector claims the matrix method using the formula works in the general case in which there are three or more data points. Is he correct? Justify your reasoning. Use the points (2, 5), (3, 8), and (4, 7) shown in Figure 8.2.

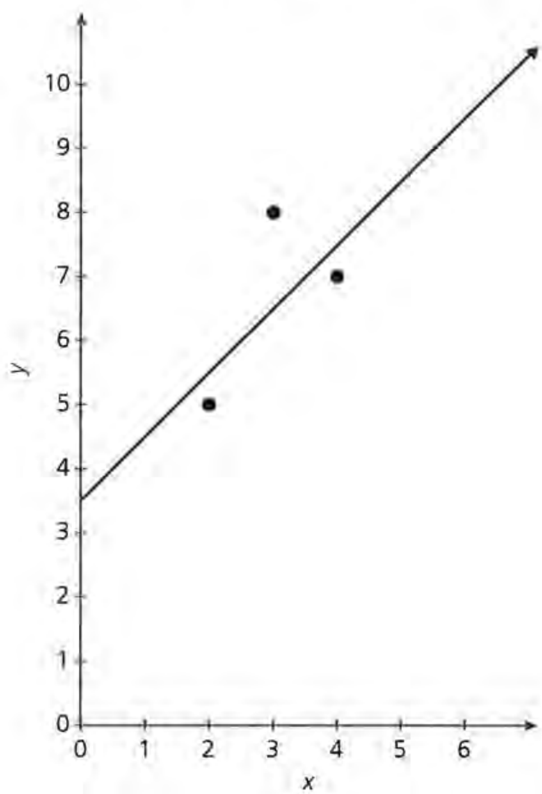


Figure 8.2

5. It is impossible to use either of the two methods described in problems 3a and b above to find the equation for the SAT verbal and mathematics scores to predict college grade point.

a. Explain why.

b. Apply the formula b = $(X^TX)^{-1}X^TY$ to the original SAT and GPA example. How does your solution compare to the one you found using the iterative process?

6. Hardwood trees are used to make furniture. The amount of lumber that can be realized from a tree is critical to determine how much the tree is worth when it is sold. To measure volume before the tree is cut, two measurements are taken: the diameter of the tree 4.5 feet from the ground and the height of the tree measured with sighting instruments. After the tree is cut, the actual volume can be measured. The better the estimate, the more likely the seller will get the correct amount for a crop of trees. Therefore, a good regression model will be useful. The data in the table below are on black cherry trees from the Allegheny National Forest, Pennsylvania.

Cherry Wood Trees

Diameter (ft)	Height (ft)	Volume (ft³)	Diameter (ft)	Height (ft)	Volume (ft³)
0.691	70	10.3	1.075	74	19.1
0.716	65	10.3	1.075	74	22.2
0.733	63	10.2	1.108	85	33.8
0.875	72	16.4	1.142	86	27.4
0.892	81	18.8	1.150	71	25.7
0.900	83	19.7	1.183	78	34.5
0.917	66	15.6	1.208	80	31.7
0.917	75	18.2	1.333	74	36.3
0.925	80	22.6	1.358	72	38.3
0.933	75	19.9	1.442	77	42.6
0.942	79	24.2	1.458	81	55.4
0.950	76	21.0	1.492	82	55.7
0.950	76	21.4	1.500	80	58.3
0.975	69	21.3	1.500	80	51.5
1.00	75	19.1	1.717	87	77.0

Source: B. F. Ryan, B. L. Joiner, and T. A. Ryan, *Minitab Handbook* 2d ed., PWS-Kent, Boston 1986

a. How do you think the foresters measure the height of a tree? Use an example in your answer.

b. Make a plot of (*diameter, volume*). Find the least squares linear regression model for the data by using the matrix formula and by using the linear regression on your calculator. How do the two models compare?

c. Find the sum of squared residuals for the least squares model.

d. Make a plot of (*height, volume*). Find the least squares linear regression model for the plot. Find the sum of squared residuals.

e. Find a regression model to predict volume from both diameter and height. Compare the three models and describe how well each fits the data.

7. Think carefully about the procedure for finding the coefficients for a model that can serve as a predictor equation. How would the procedure have to be modified if you had three independent variables? Consider using height, shoulder width, and age to predict the weight of a person.

a. Write the general form of the appropriate model and set up a matrix system using the three variables.

b. Suppose you have the measurements for 12 people. Describe the system you would be using.

c. Explain how you could use the procedure described above to find the coefficients of the model that will give you the least sum of squared residuals.

8. Obtain from your teacher the grades for a quiz, homework or a project, a test, and a final or semester exam for at least ten students. Find a model that could be used to predict the final exam grade based on the set of other grades.

Why the Formula Works (Optional)

The formula $b = (X^TX)^{-1}X^TY$ will always give you the least squares regression model for any number of variables. You must be careful, however, because, as in any curve-fitting process, a variable may have no relation to the response variable and so will not improve the model. The idea behind least squares regression is to find values of the parameters b_0, b_1, and b_2 that make the sum of squared residuals as small as possible. (Hence the term *"least squares."*) A residual is the difference between an actual value of y and the corresponding predicted value $\hat{y}$, where you subtract $\hat{y}$ from y: $r = y - \hat{y}$.

In matrix terms, the predicted values (i.e., the $\hat{y}$s) are given by $Xb = \hat{Y}$. Thus, you can use matrices to find the residuals by subtracting Xb from the matrix Y. This gives an $n \times 1$ matrix of residuals: $Y - \hat{Y}$. If you want to calculate the sum of squared residuals, take this $n \times 1$ matrix and premultiply it by its transpose (which is a $1 \times n$ matrix). The product is a 1×1 matrix (i.e., a scalar) that equals the first residual squared + the second residual squared + ...

$$\text{Sum of squared residuals} = SSR = [Y - \hat{Y}]^T[Y - \hat{Y}] = [Y - Xb]^T[Y - Xb]$$

To find the least squares values for the parameters b_0, b_1, and b_2, you want to minimize $[Y - Xb]^T[Y - Xb]$. First, multiply out this expression, using the rules identified in Lesson 1. This will involve four terms:

$$\begin{aligned} SSR &= [Y - Xb]^T[Y - Xb] \\ &= [Y^T - (Xb)^T]\,[Y - Xb] \\ &= Y^TY - Y^T(Xb) - (Xb)^TY + (Xb)^T(Xb) \\ &= Y^TY - Y^TXb - b^TX^TY + b^TX^TXb \text{ since } (Xb)^T = b^TX^T \end{aligned}$$

Note that Y^TXb is the product of a $1 \times n$ matrix (Y^T) with an $n \times 1$ matrix (Xb). Thus, it is a 1×1 matrix. Likewise, b^TX^TY is a 1×1 matrix. Since the transpose of a 1×1 matrix (i.e., a scalar) is equal to itself and since the transpose of b^TX^TY is Y^TXb, it follows that b^TX^TY equals Y^TXb. Thus,

$$SSR = Y^TY - 2Y^TXb + b^TX^TXb.$$

You wish to minimize this expression by finding good values for the parameters b_0, b_1, and b_2. That is, you want to find a good matrix b. The first term in *SSR*, Y^TY, does not involve b, so you can ignore it and concentrate on minimizing $-2Y^TXb + b^TX^TXb$.

Earlier you found that a good choice of b is $b = (X^TX)^{-1}X^TY$. You will see that this is the best choice of b. You will see that if you change $(X^TX)^{-1}X^TY$ by adding a vector to it, you make matters worse, unless the vector is the zero vector (that is, a vector whose elements are all equal to zero).

Suppose $b = (X^TX)^{-1}X^TY + v$, where v is an $n \times 1$ vector. Then the sum of squared residuals

$$SSR = Y^TY - 2Y^TXb + b^TX^TXb$$

becomes

$$= Y^TY - 2Y^TX[(X^TX)^{-1}X^TY + v] + [(X^TX)^{-1}X^TY + v]^TX^TX[(X^TX)^{-1}X^TY + v].$$

If you expand the product of these terms you get

$$\begin{aligned} SSR = {} & Y^TY - 2Y^TX[(X^TX)^{-1}X^TY - 2Y^TX[v] \\ & + [(X^TX)^{-1}X^TY]^TX^TX[(X^TX)^{-1}X^TY] \\ & + [(X^TX)^{-1}X^TY]^TX^TX[v] \\ & + [v]^TX^TX[(X^TX)^{-1}X^TY] \\ & + [v]^TX^TX[v]. \end{aligned}$$

This expression contains seven terms, but the first, second, and fourth terms do not involve v. Combine those terms and call them c (for "constant"). Then you have

$$\begin{aligned} SSR = {} & c - 2Y^TX[v] + [(X^TX)^{-1}X^TY]^TX^TX[v] \\ & + [v]^TX^TX[(X^TX)^{-1}X^TY] \\ & + [v]^TX^TX[v]. \end{aligned}$$

The next thing to notice is that the second to last term, $[v]^TX^TX[(X^TX)^{-1}X^TY]$, can be simplified. The X^TX part cancels with the $(X^TX)^{-1}$ part, leaving $[v]^T[X^TY]$. Moreover, v is a $p \times 1$ vector (matrix), so $[v]^T$ has dimension $1 \times p$. X^TY is the product of a $p \times n$ matrix with an $n \times 1$ matrix, so it has dimension $p \times 1$. Thus, $[v]^T[X^TY]$ has dimension 1×1, which means that $[v]^T[X^TY]$ is equal to its transpose $Y^TX[v]$.

Likewise, the term $[(X^TX)^{-1}X^TY]^TX^TX[v]$ can be simplified. First, take the transpose of the leading part, $[(X^TX)^{-1}X^TY]$, to get $Y^TX[(X^TX)^{-T}]$, so that

$$[(X^TX)^{-1}X^TY]^TX^TX[v] = Y^TX[(X^TX)^{-T}]X^TX[v].$$

Now, (X^TX) is a symmetric matrix, which means that $(X^TX)^{-1}$ is also symmetric. Thus, $(X^TX)^{-T} = (X^TX)^{-1}$. Hence, $(X^TX)^{-T}$ cancels with (X^TX), giving us the result that $Y^TX[(X^TX)^{-T}]X^TX[v] = Y^TX[v]$.

Thus, you have simplified the expression for the sum of squared residuals to the form

$$\begin{aligned} SSR = {} & c - 2Y^TX[v] + Y^TX[v] \\ & + Y^TX[v] \\ & + [v]^TX^TX[v]. \end{aligned}$$

Notice that the middle part of this expression cancels, which means that

$$SSR = c + [v]^TX^TX[v].$$

Write the last term a bit differently to get

$$SSR = c + v^TX^TXv = c + [Xv]^T[Xv].$$

In summary, the sum of squared residuals equals a constant plus $[Xv]^T[Xv]$.

Now, Xv is the product of an $n \times p$ matrix and a $p \times 1$ matrix, so it has dimension $n \times 1$. That is, Xv is an $n \times 1$ vector. The term $[Xv]^T[Xv]$ is the sum of the squared values of the elements of Xv. That is, $[Xv]^T[Xv]$ equals the first element squared + the second element squared + ... You wish to minimize *SSR*, which means that you want $[Xv]^T[Xv]$ to be as small as possible. Clearly, the smallest possible value for the sum of squared elements is 0, which is what you get if *each* of the elements equals 0.

You can make each of the elements of Xv equal to 0 by making v the zero vector. Thus, you have shown that the sum of squared residuals is minimized when v is the zero vector, which means that the least squares solution is

$$b = (X^TX)^{-1}X^TY + v = (X^TX)^{-1}X^TY.$$

This is what you wanted to show.

Summary

When you have more than one variable to use to predict an outcome, it is possible to find a good model by using an iterative regression process. There is a matrix formula, however, that will produce the model much faster. It can be shown that using this formula produces the model with the least sum of squared residuals. If you write the problem as a matrix system, you can find the elements in the coefficient matrix b by using the formula $b = (X^TX)^{-1}X^TY$.

Practice and Applications

9. What cereal do you eat for breakfast? How high is your cereal in calories? What variables affect the number of calories? The table shown on the next page contains the amount of sugar, the amount of sodium, and the number of calories for a one-ounce serving of 16 different cereals.

Cereal

Cereal	Sugar (gms)	Sodium (mg)	Calories
Cap'n Crunch	12	220	120
Apple Jacks	14	125	110
Corn Pops	12	90	110
Frosted Flakes	11	200	110
Frosted Mini Wheats	7	0	100
Life	6	150	100
Nut & Honey Crunch	9	190	120
Raisin Nut Bran	8	140	100
Raisin Squares	6	0	90
Rice Chex	2	240	110
Shredded Wheat	0	0	80
Wheaties	3	200	100
Golden Graham	9	280	110
Nutri Grain Wheat	2	170	90
Cocoa Puffs	13	180	110
Grape Nuts	3	170	110

Source: American Statistical Association, 1993 statistical graphics exposition

a. Plot the data for each of the following. Find a regression model and decide how well the model seems to fit the data.

i. (*sugar, calories*)

ii. (*sodium, calories*)

b. Find a multiple regression model using both sugar and sodium to predict the number of calories. How does this model compare to either of the other two?

c. Use your model from part b to predict the number of calories in a serving of Wheat Chex that has 5 grams of sugar and 390 milligrams of sodium.

d. Look carefully at your model from part b. Which variable has the greater effect on the prediction? Explain how you can tell.

e. Do you think other variables might also have an influence on the number of calories? If so, what might they be? Describe how would you adjust your model to accommodate them.

10. In Lesson 5 you studied the relation between two of the variables that might be factors in finding the operating cost of an airplane. Other variables are given in the table on the next page.

Airplane Data

Aircraft	No. of Seats	Air Speed (mph)	Flight Length (miles)	Fuel Consumption (gal/hr)	Operating Cost/hour ($)
B747-100	405	519	3149	3529	6132
L-1011-100/200	296	498	1631	2215	3885
DC-10-10	288	484	1410	2174	4236
A300 B4	258	460	1221	1482	3526
A310-300	240	473	1512	1574	3484
B767-300	230	478	1886	1503	3334
B767-200	193	475	1736	1377	2887
B757-200	188	449	984	985	2301
B727-200	148	427	688	1249	2247
MD-80	142	415	667	882	1861
B737-300	131	413	605	732	1826
DC-9-50	122	378	685	848	1830
B727-100	115	422	626	1104	2031
B737-100/200	112	388	440	806	1772
F-100	103	360	384	631	1456
DC-9-30	102	377	421	804	1778
DC-9-10	78	376	394	764	1588

Source: *The World Almanac and Book of Facts,* 1993

a. Find a model for using speed and flight length to predict operating costs.

b. What is the sum of squared residuals?

c. Use your model to predict the operating cost for an airplane that travels at 500 mph and usually makes flights of around 400 miles.

d. Is your model more effective for short trips or for long ones or is there no difference? How can you tell?

11. Use all of the factors to determine a model to predict the cost of operating an airplane. Compare your model using all of the factors to the one you found using only speed and flight length.

12. The matrix formula can be used to generate a prediction equation in one variable that is quadratic rather than linear. Suppose you had just speed to predict fuel consumption.

a. Plot (fuel consumption, speed). Describe the trend you see in the plot.

b. Find a least squares linear regression line to describe the plot. How well does your line fit the data?

c. Use the matrix formula to find an equation of the form

$$b_0 + b_1x + b_2x^2 = y$$

where $[1 \ \ x \ \ x^2]$ is the matrix just as $[1 \ \ V \ \ M]$ was in the lesson. Explain why the matrix formula can be used to find the model.

d. Graph your equation with the data. How does the new model compare to the linear model?

13. The cost of stamps has consistently risen over the years.

Cost of Stamps

Year	Cost of Stamps (Cents)
1885	2
1917	3
1919	2
1932	3
1958	4
1963	5
1968	6
1971	8
1974	10
1975	13
1978	15
1981	18
1981	20
1985	22
1988	25
1991	29
1995	32

Source: *The Milwaukee Journal,* December 1994

a. Plot (*year, cost*). Find a model to predict the cost of stamps.

b. What feature of the data has a strong impact on the model?

c. Predict the cost of stamps in the year 2000. How reliable do you think your prediction will be?

14. The table below shows the all-time leading passers in the National Football League (NFL) as of 1993, along with the variables used by the NFL to rate its quarterbacks on their passing ability.

NFL Quarterback Rating

Player	Att	Comp	TD	Int	Yards	Rating
Ken Anderson	4475	2654	197	160	32838	81.9
Len Dawson	3741	2136	239	183	28711	82.6
Sonny Jurgensen	4262	2433	255	189	32224	82.6
Jim Kelly	3494	2112	179	126	26413	86.0
David Kreig	4178	2431	217	163	30485	82.0
Neil Lomax	3153	1817	136	90	22771	82.7
Dan Marino	5434	3219	298	168	40720	88.1
Joe Montana	4898	3110	257	58	37268	93.1
Roger Staubach	2958	1685	153	109	22700	83.4
Steve Young	1968	1222	105	58	15900	93.0

Source: *The Universal Almanac*, 1995

a. Using the variables $\frac{C}{A}$, $\frac{T}{A}$, $\frac{I}{A}$, and $\frac{Y}{A}$, where

R = the rating
A = attempts
C = number of completions
T = number of touchdowns
I = number of interceptions
Y = yards gained by passing,

find a model to predict the ratings in general.

b. The formula used by the NFL is

$$R = \frac{50 + 2000\left(\frac{C}{A}\right) + 8000\left(\frac{T}{A}\right) - 10000\left(\frac{I}{A}\right) + 100\left(\frac{Y}{A}\right)}{24}$$

How does your formula compare?

15. Choose a situation that has at least two variables involved in the outcome. Collect data and find a regression model. In a brief report, describe each of the following:

- the problem
- the data source
- the model and how you selected it
- an example of how the model works
- your opinion about how effective your model will be to make predictions

ASSESSMENT

Rating Universities

The following data were taken from some of the information given in the *U.S. News & World Report*, September 18, 1995 issue on America's best colleges. The universities were the top fourteen in *U.S. News & World Report* list, given here alphabetically.

OBJECTIVE

Create ratings and ranks when high scores are good for one variable but bad for another variable.

Data on Top Universities

University	SAT/ACT 25th Percentile	Acceptance Rate %	% Attend After Accepted	Cost per Student ($)
Brown University (RI)	1210	22	49	22,704
California Institute of Technology	1350	25	46	63,575
Cornell University (NY)	1180	33	47	21,864
Dartmouth (NH)	1250	23	48	32,162
Duke University (NC)	1220	30	40	31,585
Harvard University (MA)	1320	14	75	39,525
Johns Hopkins University (MD)	1210	44	28	58,691
Massachusetts Institute of Technology	1290	30	51	34,870
Northwestern University (IL)	1160	39	36	28,052
Princeton University (NJ)	1280	14	56	30,220
Stanford University (CA)	1270	20	54	36,450
University of Chicago (IL)	1180	50	29	38,380
University of Pennsylvania	1190	36	47	27,553
Yale University (CT)	1290	19	53	43,514

1. The percentage of students who actually attend a university after having been offered an invitation to attend is given in the third column.

a. Which factors do you think might influence the attendance rate for the universities?

b. Analyze the data given above, looking for a relationship between two of the variables. Describe what you find and indicate how the elements of the curve-fitting process are part of the discussion.

c. Look for two variables that may have an impact on the attendance rate. Find a multivariate regression equation that would predict the attendance rate.

d. Explain how using matrices simplifies the process.

2. The final rating for each of the universities as given by *U.S. News and World Report* is shown below.

Rating Top Universities

University	Overall Scores
Brown University (RI)	95.3
California Institute of Technology	95.5
Cornell University (NY)	94.0
Dartmouth (NH)	95.5
Duke University (NC)	96.8
Harvard University (MA)	100.0
Johns Hopkins University (MD)	94.6
Massachusetts Institute of Technology	98.0
Northwestern University (IL)	94.0
Princeton University (NJ)	98.8
Stanford University (CA)	98.1
University of Chicago (IL)	94.4
University of Pennsylvania	94.4
Yale University (CT)	98.8

a. Use appropriate variables and find a multivariable regression equation to predict the final rating based on those variables.

b. Show how you can verify that multiple regression produces a better model than any using just one independent variable.

Dale Seymour Publications® is a leading publisher of K–12 educational materials in mathematics, thinking skills, science, language arts, and art education.

DALE
SEYMOUR
PUBLICATIONS®

90000

9 781572 322554

ISBN 1-57232-255-1
21188