Simple Linear Regression: A PAPIRIS Example

Maher Y. Shawer, Michael J. Bossé, Frederick W. Morgan, John D. Baker
Department of Mathematics, Indiana University of Pennsylvania, Indiana, PA 15701

Introduction

The **PAPIRIS** (Principles and Pedagogy – Internet Resources in Statistics) vision is an easy to use Web site with multiple levels of grade appropriate definitions and concepts, teacher resources, lesson plans, grade-level appropriate activities, and self-help pages for K-14 students. PAPIRIS will help satisfy explanatory and curricular needs for K-14 statistics education within the United States. The following entry represents a sample of the planned entries within PAPIRIS. In this case, entries for linear regressions are provided.

PAPIRIS will be a Web-based statistics reference library for teachers, students, and lifelong learners, consisting of an encyclopedia, annotated links to external Web sites for statistics and statistics education, a collection of lessons plans, and critiques of those lesson plans. It will contain a collection of statistics terms used in K-14 education. In the near future, PAPIRIS will include an archive of lesson plans for teaching the concepts, with accompanying annotations written by PAPIRIS personnel describing the lessons' contents.

PAPIRIS will participate in the Mathematics Education into the 21st Century Project, which is a 20-year international effort dedicated to improving mathematics education worldwide through the publication and dissemination of innovative ideas. PAPIRIS will provide the contents of its repository to the international project's collection of mathematics education resources, and the participants in the international project will contribute statistics education lessons to PAPIRIS.

LINEAR REGRESSION

For the following examples, the accompanying real world data is utilized. The table delineates actual data of the number and weight of chocolate coated, peanut candies within a marked 49.3 gram bag.

# of Peanut Can in bag, X	*Bag Weight, Y*
20	47.0
21	48.8
22	48.7
22	51.0
23	50.3
23	52.0
24	52.2
25	52.9
26	54.8

ELEMENTARY and MIDDLE SCHOOL

Definition: A linear regression is a mathematical model demonstrating the relationship between two variables or sets of data.

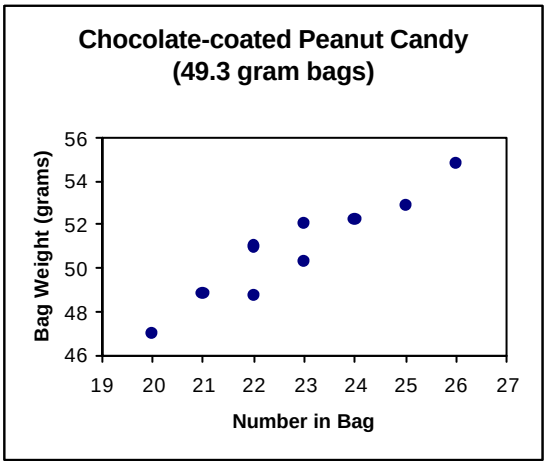
Elementary and Middle School Preparation:
Students should be able to plot data points on an x-y coordinate plane and determine the median and mean of a data set.

Purpose: In elementary and middle school, it is appropriate for school students to examine relationships between such real-life applications as study time and test results, or size of a school and win-loss record of its sports teams (van de Walle, 2001). After relationships are recognized, linear regressions can assist in making predictions regarding the data and applications.

Pedagogy: Interesting and meaningful applications provide motivation for students and build on common-sense understandings. Further, a data-driven approach to learning can foster the basic understandings required for more sophisticated ideas. The following techniques are meant to provide grade appropriate estimations eventually leading to the concept of linear regression through the least squares line in high school.

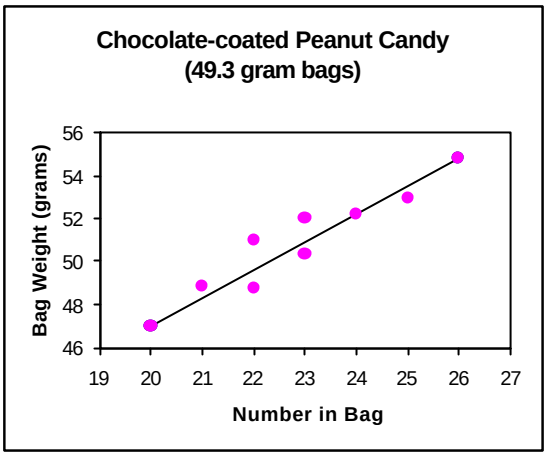
Elementary Education Techniques:

Scatter Plots: Once data is collected, it is reasonable to ask how relationships can best be understood, and later, how predictions can be made from collected data. An appropriate starting point is to make a scatter plot where relationships can often be seen visually, and then, described verbally. With some instruction on plotting points, elementary school students can begin their conceptual understandings with simple activities such as creating scatter plots.



Trend Line: As a next step for young students, an informal trend line (Bennett & Nelson, 1998) can be drawn. A trend line serves to reinforce the relationship that can be seen visually, and also may provide a springboard for discussion. Simple

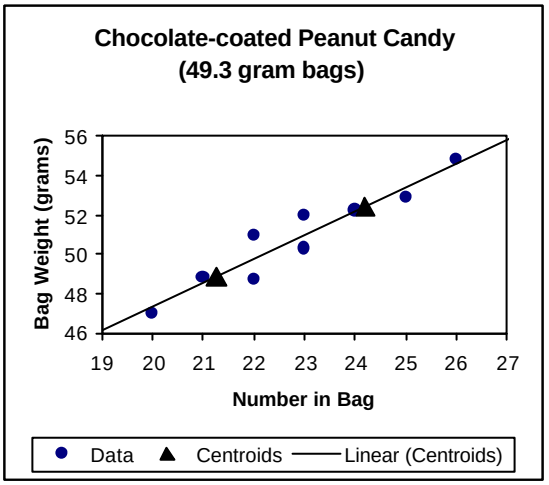
classroom activities could be developed: (a) on a scatter plot, students can explore trend lines with a piece of spaghetti (van de Walle, 2001), (b) a trend line can be drawn with a straightedge, and (c) two values or "choice points" that seem representative of the trend can be selected and a line drawn through them.



From a simplified point of view, the line of regression is an attempt to make sense of general data trends. The closer data is to the trend line, the more confidence one has in predictions. Students may find that making predictions from a trend line that only rarely coincides with actual data values or that varies widely from a trend line is not to be trusted. Teachers may need to reinforce the idea that predictions are estimates. The idea of mathematical models not exactly fitting real-life data is a concept for teachers to explore with their students. Closely associated with concepts of data relationships and trend lines is the notion of "best fit" lines. The coarse lines described above are not intended to show best fit, but rather general trends.

Middle Grade Techniques:

Median-Median Line: A more advanced treatment of trend is the median-median line (van de Walle, 2001). More likely to be available in middle school classrooms, the TI-73 and TI-83 calculators have a built-in function for this purpose.



To determine the median line, the number of data elements is divided from least to greatest such that they form roughly equal sections of elements. [Sections will have elements numbering: n, n, n; n-1, n, n-1; or n+1, n, n+1.] For each section, the median x value and median y value is determined to form three ordered pairs, one for each section. An informal line then connects the two outside median points. Remaining parallel to this line, the line is then shifted 1/3 of the way toward the middle point, and a new line is drawn.

Double Centroid Line: The double centroid line divides the data, from least to greatest, into two sections of equal numbers of data elements. The mean x value and y value is determined for each section, creating two ordered pairs, the lower and upper centroid [$(\bar{x}_l, \bar{y}_l)$ and $(\bar{x}_u, \bar{y}_u)$] . A line then connects both centroids.

SECONDARY SCHOOL

Definition: A linear regression is a linear equation which best fits the data points associated with a given relationship of two variables or data sets. For points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ the best fit line is determined by the least squares line $y = a + bx$, where

$$b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}, \quad a = \bar{y} - b\bar{x}, \quad \bar{x} = \frac{\sum x_i}{n}, \text{ and } \bar{y} = \frac{\sum y_i}{n}.$$

Secondary School Preparation:

Students should understand the x-y coordinate plane, the mean, quadratic functions and their graphs, and summation notation and have an understanding of linear functions.

Purpose: In secondary school, students examine relationships within real-world situations. Linear regressions

can assist to analyze and make predictions regarding the data derived from the investigated situation.

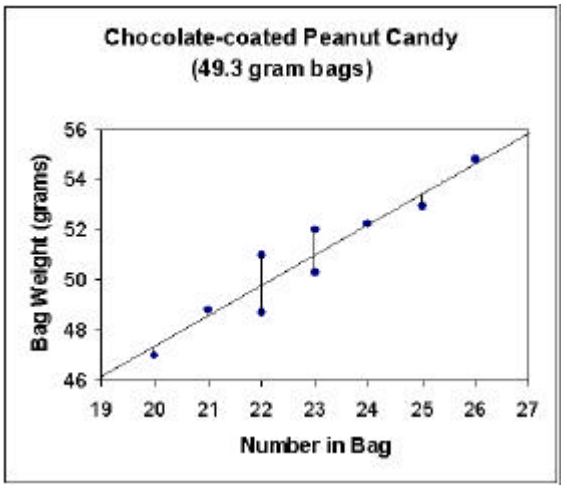
Pedagogy: Interesting and meaningful applications provide motivation for students and build on common-sense understandings. At the secondary school level, data analysis can be connected to other mathematical concepts previously studied. Following informal investigations resulting from scatter plots, trend lines, median lines, and double centroid lines, linear regression formally connects concepts from data acquisition/analysis with concepts from algebra and linear equations.

Techniques:

Least Squares Line: For slightly easier calculations, the least squares line, $y' = ax + b$, can be determined by $a = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2}$ and $b = \frac{\sum y_i - a(\sum x_i)}{n}$.

This provides the best fit line to the data points (x_1, y_1), (x_2, y_2), … , (x_n, y_n). The least squares line minimizes the sum of the squares of the differences of each data point from a set line. The least squares line can be calculated utilizing the accompanying table.

From the given data observed through the relationship, the least squares line allows one to predict additional values for the relationship.



Example:

The candy data provided above would produce the following table of data. These are then substituted into the equations to determine values for a and b.

x_i	y_i	$x_i y_i$	x_i^2
x_1	y_1	$x_1 y_1$	x_1^2
x_2	y_2	$x_2 y_2$	x_2^2
*	*	*	*
x_n	y_n	$y_n y_n$	x_n^2
$\sum x_i$	$\sum y_i$	$\sum x_i y_i$	$\sum x_i^2$

Parabolic Method to the Least Squares Line: The least squares line technique previously depicted leads to the following analysis.

$$y = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2} x + \frac{\sum y_i - a(\sum x_i)}{n}$$

$$y = ax + (\bar{y} - a\bar{x})$$

$$y = \bar{y} + ax - a\bar{x}$$

$$y - \bar{y} = a(x - \bar{x})$$

Therefore, the point $(\bar{x}, \bar{y})$ is on the least squares line.

The accompanying steps can be followed to determine the slope of the least squares line (which passes through the point $(\bar{x}, \bar{y})$). Relatively few data sets of n ordered pairs are actually needed as slopes fall along a parabola.

x_i	y_i	$x_i y_i$	x_i^2
20	47.0	940	400
21	48.8	1024.4	441
22	48.7	1071.4	484
22	51.0	1122	484
23	50.3	1156.9	529
23	52.0	1196	529
24	52.2	1252.8	576
25	52.9	1322.5	576
26	54.8	1424.8	676
206	**457.7**	**.0510.8**	**4695**

Find the slope of the line between at least three pairs of given points. (Notably, a quadratic can be determined through three points. However, most high school students will need to select at least five points to graph a quadratic. Therefore, students should select approximately five pairs of points for which to determine the slope.) Arrange slopes in ascending order from least to greatest, labeling the slopes as $x_1, x_2, \ldots, x_i$.

Find the equations of the lines with slope m_i through the point $(\bar{x}, \bar{y})$. $y_i = \bar{y} + m_i(x - \bar{x})$

Determine the sum of the squares of the residuals, r_i, for each line, y_i.

For m_i, the slope between two data points,
and $r_i = \sum (residuals)^2$

m_1	m_2	***	m_n
r_1	r_2	***	r_n

Plot each point (m_i, r_i) and sketch the graph of the respective parabola.

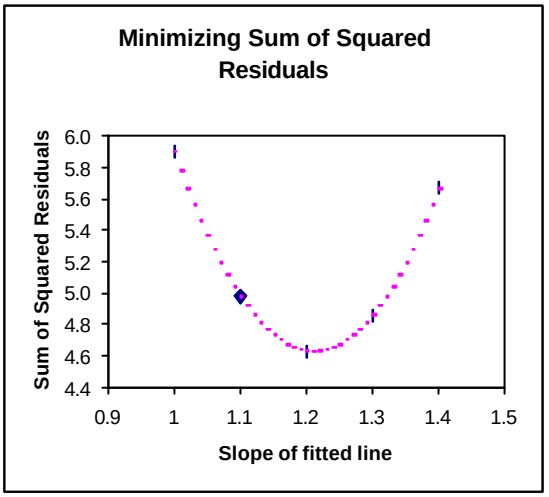
The x- coordinate of the minimum point (vertex of the parabola will represent the slope of the line that has the least $(residuals)^2$.

Example:

Utilizing this technique produces the following table which in turn produces the following graph.

				$y_p=1.0x+27$	$y_p=1.1x+25.6$	$y_p=1.2x+23.3$	$y_p=1.3x+21.1$	$y_p=1.4x+18.8$
x	**y**	**y_p**	**$y-y_p$**	**$(y-y_p)^2$**	**$(y-y_p)^2$**	**$(y-y_p)^2$**	**$(y-y_p)^2$**	**$(y-y_p)^2$**
20	47.0	47.97	-0.97	0.941	0.462	0.152	0.010	0.036
21	48.8	48.97	-0.17	0.029	0.000	0.044	0.160	0.348
22	48.7	49.97	-1.27	1.613	1.392	1.188	1.000	0.828
22	51.0	49.97	1.03	1.061	1.254	1.464	1.690	1.932
23	50.3	50.97	-0.67	0.449	0.462	0.476	0.490	0.504
23	52.0	50.97	1.03	1.061	1.040	1.020	1.000	0.980
24	52.2	51.97	0.23	0.053	0.014	0.000	0.010	0.044
25	52.9	52.97	-0.07	0.005	0.078	0.240	0.490	0.828
26	54.8	53.97	0.83	0.689	0.270	0.044	0.010	0.168
Sum of Squared Residu				**5.900**	**4.976**	**4.629**	**4.860**	**5.669**

This example demonstrates that the slope of the line with the least sum of squared residuals is approximately 1.2.



COLLEGE

Definition:

Regression: A model that describes the dependence of the value of one random variable on one or more other variables. If the mean value of a random variable Y for a fixed x is written as $E(Y/x)$ and called the mean of Y conditional upon x, then when x varies, $E(Y/x)$ is a function called the regression of Y on x.

The simplest case is simple linear regression or straight-line regression:

$$E(Y/x) = a + bx$$

The parameters a, b are called the regression coefficients and are usually unknown. Regression Analysis is concerned with estimating the parameters a and b, and finding the best-fit line.

College Preparation:

Students should understand the concepts: linear equation, the minimizing and maximizing using differentiation, the expected value and conditional probability.

Purpose: In college, students use regression line to summarize the relationship between dependent variables and independent variables, predict the dependent variable from one or more independent variables, measure the size of the effect to any independent variable through its regression coefficient, and discover mathematics or empirical laws.

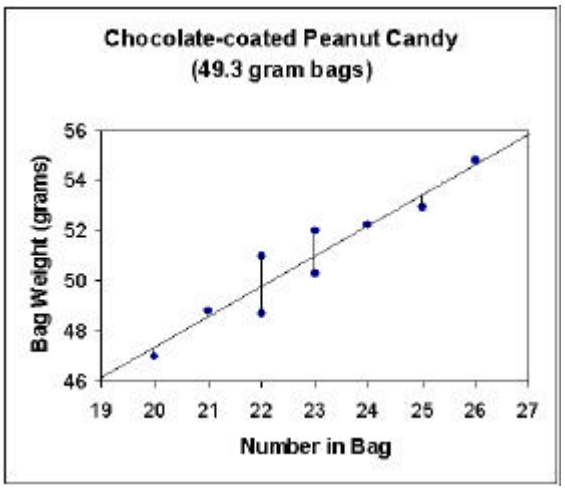
Pedagogy: Using calculus to develop the regression model will give the students understanding about the importance of calculus in developing models, and using the regression model introduces the students to a model that incorporates a random element and has application in many fields of the real world.

Techniques:

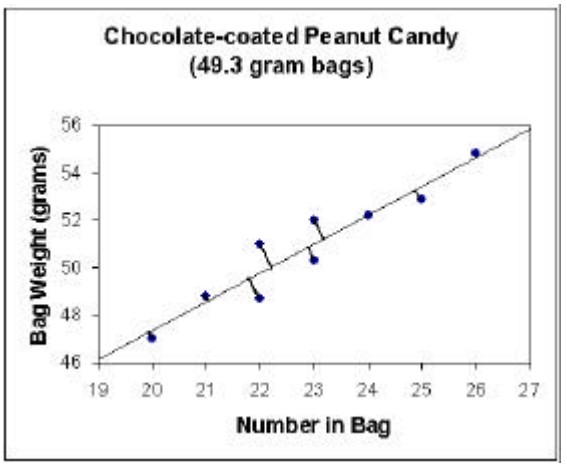
If Y is assumed to be normally distributed with constant and usually unknown variance, which is independent of X, then given a sample of independent pairs of observation (x_i, y_i); $i = 1,...,n$, the estimators a , b of α and β can be found by a procedure called the least square. This procedure is minimizing $\sum (y_i - a - bx_i)^2$ where $(y_i - a - bx_i)$ is called the residual or the vertical offset. By using this procedure, the maximum likelihood estimators b and a for β and α are:

$$b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}, \quad a = \bar{y} - b\bar{x}, \quad \bar{x} = \frac{\sum x_i}{n}, \text{ and } \bar{y} = \frac{\sum y_i}{n}.$$

GRAPH OF VERTICAL OFFSET



GRAPH OF PERPENDICULAR OFFSET



Other methods to find a and b are by using the sum of the perpendicular distances d_i of the points (x_i , y_i) from the line

$$R = \sum d_i = \frac{\sum |y_i - (a + bx|}{\sqrt{1 + b^2}}.$$

But because the absolute value function does not have continuous derivatives, minimizing R is not amenable to analytic solution. However, by minimizing the square of the sum of the perpendicular distance, $R^2 = \frac{|y_i - (a + bx)|^2}{1 + b^2}$, then: $b = -b \pm \sqrt{b^2 + 1}$,

$$b = \frac{1}{2}\left[\frac{(\sum y^2 - n\bar{y}^2) - (\sum x^2 - n\bar{x}^2)}{n\bar{x}\bar{y} - \sum xy}\right] \text{ and } a = \bar{y} - b\bar{x}.$$

Example:

i) Vertical offset method:

By minimizing the sum square of the vertical residuals: $b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{34.955}{28.889} = 1.21$,

$a = 50.68 - 1.2(22.89) = 23.16.$

The regression equation is: $\hat{y} = 23.16 + 1.21x$, and the sum of the square of the residuals,

$\sum (\hat{y} - a - bx)^2 = 4.626$.

ii) Perpendicular offset method:

By minimizing the sum square of the perpendicular residuals:

$B = \frac{1}{2}\left[\frac{46.922 - 28.889}{-34.956}\right] = -0.258$, $b = 0.258 + \sqrt{1 + (0.258)^2} = 1.29$,

$a = 50.86 - 1.28(22.89) = 21.33$

The regression equation is: $\hat{y} = 21.33 + 1.29x$, and the sum of the square of the residuals, $\frac{\sum (y - a - bx)^2}{1 + b^2} = 4.811$.

In practice, the method of the vertical offsets from the line (the least square) is used instead of the perpendicular offsets. This allows uncertainties of the data points along the x and y axes to be incorporated, and also provide a much simpler analytic form for finding the parameter, than would be obtained by using the perpendicular distances. In addition, the vertical distance methods can easily be generalized from a best-fit line to the best-fit polynomial which is not the case using the perpendicular distance. However, the difference between vertical and perpendicular methods of estimating the parameter is very small as shown from the example above.

Type of Regression:

Regression Analysis may be applied to data gathered in two ways:

1) The amount of y may be measured for certain value of factor x,
2) The point (x, y) may be chosen at random from a two dimensional normal distribution of points in x-y plane.

In either case y value observed is considered to be a random observation from a normal population of a possible y for each particular x used. The assumption of normality and common variance are not necessary for fitting an equation, but are necessary for significance tests and confidence intervals.

Bibliography:

Bennett, A. B. & Nelson L. T. (1998). *Mathematics for elementary teachers: A conceptual approach, 4th Ed.* WCB/McGraw-Hill: Boston, MA.

Bowerman B L, O'Connell R T. *Linear Statistical Models: An Applied Approach (Second Edition).* Duxbury Press, 1990.

Burrill G F, Burrill J C, Hopfensperger P W, Landwehr J M. *Exploring Regression (Teacher's Edition).* Dale Seymour Publications, 1999.

Crow E L, Davis F A, Maxfield M W. *Statistics Manual.* Dover Publications, Inc., 1960.

James G, James R C (Eds.). *Mathematics Dictionary (Multilingual Edition).* D. Van Nostrand Company, Inc., 1959.

Mosteller F, Tukey J W. *Data Analysis and Regression: A Second Course in Statistics.* Addison-Wesley Publishing Company, Inc., 1977.

Nelson D (Ed.). *The Penguin Dictionary of Mathematics (Second Edition).* Penguin Books Ltd., 1989.

Stout W F, Travers K J, Marden J. *Statistics: Making Sense of Data (Second Edition).* M¬bius Communications, Ltd., 1999.

Van de Walle, J. (2001). Elementary and middle school mathematics: Teaching developmentally, 4th Ed. Addison, Wesley, Longman: New York, NY.

http://mathworld.wolfram.com/LeastSquaresFitting.html
http://cne.gmu.edu/modules/dau/stat/regression/linregsn/nreg_1_frm.html
http://whatis.techtarget.com/definition/0,,sid9_gci212884,00.html
http://qed.econ.queensu.ca/walras/custom/300/351B/notes/mul_02.htm#B
http://www.wpi.edu/Academics/Depts/IGSD/IQPHbook/ch13.html
http://cne.gmu.edu/modules/dau/stat/regression/multregsn/mreg_1_frm.html
http://qed.econ.queensu.ca/walras/custom/300/351B/notes/reg_02.htm#B
http://ebook.stat.ucla.edu/textbook/singles/single_structured/regression/
http://www.ics.uci.edu/~eppstein/280/regress.html
http://cmap.coginst.uwf.edu/cmaps/MDM2/The%20Multiple%20Regres.html
http://stat-www.berkeley.edu/users/stark/SticiGui/Text/ch6.htm
http://www.itl.nist.gov/div898/handbook/pmd/section1/pmd142.htm

http://demography.anu.edu.au/Courses/DEMO8014/sec06.htm
http://www.amstat.org/publications/jse/v2n2/laviolette.html
http://www.uvm.edu/~dhowell/gradstat/psych341/lectures/Logistic%20Regression/LogisticReg1.html
http://www.phoenix5.org/glossary/regression.html