A student came to his final exam on statistical methods in economics. The professor asked him to compute the linear regression of $y$ versus $x$, and the student successfully computed some $a$ and $b$ of the straight line $y = a + bx$. Then the professor asked the student to compute the linear regression of $x$ versus $y$, and the student immediately rewrote the previous equation into the form $x = (1/b)y - (a/b)$. The professor was expecting that the student would derive the equation of a conjugate regression line and thus evaluated the student’s answer as unsatisfactory. But was the student’s answer really incorrect? That all depends on how the first line was calculated, which in turn depends on the criterion used for determining $a$ and $b$ in the first line.

Least Squares

One could hardly name another method used as frequently as the method known as the least squares method. At the same time, it is difficult to name another method that has been accompanied by such strong and long-lasting controversy. (Details of that story appear in the sidebar, “The Birth of Least Squares”.) It is also difficult to find another method that is both so easy and at the same time so artificial. Figure 1 is a version of the picture that appears frequently in textbooks and slides and on blackboards as a geometric illustration of the least squares method.

Recall how this figure is created. A set of points $(x_k, y_k)$, $k = 1, \ldots, N$, is approximated by a line $y = a + bx$. The classical least squares fitting consists in minimizing the sum of the squares of the vertical deviations of a set of data points

$$E = \sum_{i=1}^{N} [y_i - (a + bx_i)]^2$$

from a chosen linear function $y = a + bx$. Each term in (1) corresponds to a square in Figure 1.

In the opinion of the authors, this picture is ugly: It does not have any sign of mathematical beauty. It could be good for those abstract painters Wassily Kandinsky and Kazimir Malevich, but it is not good for Johann Gauss. The line and the squares are in some visual conflict. This conflict is even more obvious if we assume that the coordinate system is not rectangular. Figure 2 gives an illustration. This conflict is

The Birth of Least Squares

The story of the birth of the least squares method is well covered in the literature and can be summarized as follows. The priority in publication definitely belongs to A. M. Legendre (Nouvelles méthodes pour la détermination des orbites des comètes, 1805), who also gave the method its famous name. C. F. Gauss (Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium, 1809) claimed, however, that he knew and used this method much earlier, about 10 years before Legendre’s publications. In a letter to Gauss about his new book Legendre wrote that claims of priority should not be made without proof in previous publications.

Gauss did not have such a publication, despite which he actively attacked Legendre. We see that his efforts were fruitful enough: In the vast majority of today’s textbooks, the least squares method is attributed to Gauss without further comment. In fact, Gauss’s arguments for his priority were not perfect at all. His diaries with computations claimed to be made by the least squares method were lost. His colleagues did not hurry to acknowledge that he showed him those computations. Indeed, can one imagine that Gauss showed and explained the details of his unpublished computations to his potential competitors?

Only many years later did H. W. M. Olbers (1816) and F. W. Bessel (1832) mention that Gauss showed them something in that sense. But how accurately could they really remember the details of some discussion that happened many years ago? It is also known that H. C. Schumacher suggested repeating Gauss’s lost computations that Gauss claimed to have done by the least squares method.

Gauss totally rejected this idea, claiming that such attempts would only suggest that he could not be trusted. This, however, has been done by Stigler (1981), who could not reproduce Gauss’s results. Later, Celmins (1998) also tried to repeat Gauss’s computations, including the adjustments suggested by Stigler (1981), and arrived at the same conclusion that Gauss’s results cannot be obtained by the least squares method. In other words, it is well known which method Legendre used, and it is not clear at all which method was used by Gauss.

Assuming that it was Gauss who invented the least squares method, it is hard to believe that he did not realize the huge potential of this method and its importance for applications. Knowing Gauss as a prolific mathematician and looking at the present version of the least squares method, one can see a certain contradiction.
absolutely obvious if we would consider, for example, polar or
elliptic coordinates. It is difficult to imagine that C. F. Gauss
would have been happy with such visual interpretation.

The key idea here is that the visualization and, more
important, the resulting approximation is dependent on the
choice of the coordinate system. The vertical distance is not
independent of the coordinate system in which vertical is
defined. But some definitions of distance are invariant to the
coordinate system.

**Shape Recognition and Curve Detection**

Nowadays, the distance between two points in \( k \)-dimensional
space is widely used as an optimal fitting criterion in the
field of image processing for industrial and scientific applica-
tions, especially in problems of shape recognition and curve
detection. An ellipse (a circle) is an ellipse (a circle) in any
coordinate system. A parabola is a parabola in any coordi-
nate system, too. Those objects are not defined by equations
but by their general properties, which include the notion
of distance.

Indeed, everybody knows from the school that a circle is
a set of points in a plane that are at the same distance from a
given point; an ellipse is a set of point for which the sum of
distance from two given points is constant, a parabola is a set
of points which are at the same distance from a given point
and from a given straight line, and so forth. We sometimes
forget that those geometric objects are not related to any par-
ticular coordinate system, although some coordinate systems
are more suitable for describing those objects by equations.
Indeed, one can write simpler equations in a suitable coordi-
nate system.

Figure 3 illustrates the fitting of an ellipse to a set of points in
two-dimensional space. What must one do to fit a set of points
by a circle, or an ellipse, or another geometric shape? One has
to draw a sample curve, measure the distance from each point
of a set under consideration to the curve, and consider the
sum of these distances as a criterion that has to be minimized.
Distance in shape recognition and pattern detection is usually
a function of squared (or absolute) deviations from the point
to the nearest point on the object. For doing this algorithmi-
cally using a computer, it is necessary to set the whole picture
into some coordinate system. The most common coordinate
system we use is the Cartesian rectangular system.

**The Least Squares Method, Revisited**

A. M. Legendre published an idea on how to circumvent the
computational problems that arise in the case of trying to
minimize the sum of orthogonal distances from data points
to a straight line. The idea was to replace the computational
problem with a problem in calculus. Put the whole set of
the objects (data point and a line) in Cartesian coordinates.
Instead of shortest distances from points to a line, consider the
distances from points to the line in a direction that is parallel
to the vertical axis (the vertical offsets). This step would give
a different criterion to be minimized:

\[
E = \sum_i |y_i - f(x_i; \alpha_1, \alpha_2, \ldots, \alpha_n)|,
\]

(2)
from a chosen function \( f \). And the minimization problem could be solved by the standard techniques of the differential calculus.

### The Method of Least Circles

To adjust a viewpoint, let us note that the criterion (3) can be painlessly replaced with

\[
E = \beta \sum_i [y_i - f(x_i, \alpha_1, \alpha_2, \ldots, \alpha_n)]^2. \tag{4}
\]

Indeed, multiplication by a positive number \( \beta \) does not affect the point of minimum. Only the minimum value of the criterion function (\( E \)) will be multiplied by \( \beta \), which itself is not the subject of interest at this stage, since we look for the values of \( \alpha_1, \alpha_2, \ldots, \alpha_n \).

Taking \( \beta = \pi \), we obtain

\[
E = \sum_i \pi [y_i - f(x_i, \alpha_1, \alpha_2, \ldots, \alpha_n)]^2. \tag{5}
\]

Geometrically, the formula (5) means the sum of the areas of the circles shown in Figure 4(a). The radii of the circles in Figure 4(a) are the vertical offsets of \( y \) from the fitting line. Figure 4(a) is just a reformulation of the standard geometric "illustration" of the least squares method (recall Figure 1).

Each of those circles has two points of intersection with the line. It is clear that one cannot consider this picture as elegant. Changing the radii slightly, one can preserve \( n \) pairs of intersection of the circles and the line. That is, the circles can be a little bigger or a little smaller and each one will still intersect the line in two places. Instead, suppose we adopt the perspective of the shortest distance to the line in two-dimensional space. The resulting circles are shown in Figure 4(b). In this case, the fitting line is a tangent line to all circles. The radii of the circles in Fig. 4(b) are equal to distances between the points \((x_i, y_i)\) and the fitting line, and this guarantees the unique picture.

The criterion to minimize in this case is

\[
E = \sum_i \pi d[(x_i, y_i), f(x, \alpha_1, \alpha_2, \ldots, \alpha_n)]^2, \tag{6}
\]

which is up to a constant multiplier \( \pi \) the formula known under the name of orthogonal regression. It is also known as total least squares or as the errors in variables method. Here \( d((x, y), f) \) denotes the distance between the point \((x, y)\) and the fitting line \( f \).

There are several obvious advantages to using least circles (squared orthogonal distance) fitting:

1. The shortest (orthogonal) distance is the most natural viewpoint on any fitting.
2. The sum of orthogonal distances is invariant with respect to the choice of the system of coordinates (see Figure 5).
3. There are no conifugate regression lines, which appear after swapping \( x \) and \( y \), because in the case of orthogonal regression the fitting \( y = f(x) \) gives exactly the same line as the fitting \( x = f^{-1}(y) \) (So, the student from the story in the beginning of this article could be absolutely right if he used the orthogonal "least circles" method to produce
the first coefficients $a$ and $b$ instead of the classical least squares method)

4. There are no problems with causality. (Normally, determination of what is an independent variable and what is a dependent variable is simply unclear or even impossible; this is always postulated.)

5. Implementation of the orthogonal fitting does not depend on the number of spatial dimensions.

The fourth point above could be a sticking point for some. Often the goal is to predict an outcome. In that case, one dimension ($y$) is of primary interest, and one often considers distance in that dimension of primary importance.

Still, orthogonal regression can be less sensitive to outlying observations and useful as a form of robust regression.

**Orthogonal Distance Linear Regression**

In general, the use of orthogonal distance fitting requires the use of numerical routines for minimization of the criteria. Fortunately for the student in the econometrics course, in the case of orthogonal distance fitting it is possible to obtain simple formulas for evaluating the parameters of a straight line that fits a given set of points in a plane (orthogonal linear regression problem). Indeed, the orthogonal distance between a point $P(x, y)$ and a straight line $y = a + bx$ is illustrated in Figure 6 with values

$$d_i = \frac{|y_i - (a + bx_i)|}{\sqrt{1 + b^2}}$$  \hspace{1cm} (7)

Following Legendre, instead of minimizing the sum of orthogonal distances, minimize the sum of their squares:

$$E^2 = \sum_{i=1}^{n} \left[ y_i - (a + bx_i) \right]^2 = \frac{1}{1 + b^2} \sum_{i=1}^{n} \left[ y_i - (a + bx_i) \right]^2$$  \hspace{1cm} (8)

As usual, take partial derivatives with respect to the parameters $a$ and $b$ equal zero:

$$\frac{\partial E^2}{\partial a} = 0, \quad \frac{\partial E^2}{\partial b} = 0.$$

A system of two equations for determining the values of $a$ and $b$ is obtained. Details can be found on the CHANCE website at www.amstat.org/publications/chance under the supplemental material. Obviously, there are two possible fitting lines, $y = a_1 + b_1x$ and $y = a_2 + b_2x$, which both run though the centroid $(\bar{x}, \bar{y})$ and are mutually orthogonal, since $b_1b_2 = -1$. The proper fitting line (the proper pair of the values of $a$ and $b$) can be determined by a smaller value of the criterion (8).

This has a very simple geometric interpretation. Indeed, the set of points that we try to fit with a straight line is a "cloud" of points in 2D. One of the possible fitting lines coincides with the direction of the main "axis" of that cloud, and the second line corresponds to the direction of the width of that cloud. It is worth mentioning that this is an elementary illustration of the relationship between the orthogonal distance fitting, on one side, and the principal components analysis (PCA) on the other. There are a few mathematical tools which can be used for orthogonal distance fitting. For solving the linear problems in n-dimensional space the PCA method is appropriate.

Generally, for solving the linear and nonlinear problem the singular value decomposition (SVD) method and QR decomposition method are suitable. SVD is widely used in statistics where it is related to the PCA method. Now PCA is mostly used as a tool in exploratory data analysis and for making predictive models, but the applicability of the PCA is limited by several assumptions (linearity, statistical importance of mean and covariance, etc.).

**Least Circles, Least Spheres, Least Hyperspheres!**

Now let us consider the 3D case. Suppose we have a set of data points, which look to be close to a straight or curved line in 3D, and we want to obtain the equation of the optimal fitting line. First, moving from 2D to 3D makes the idea of "least squares" absolutely useless. The only natural criterion is the minimum sum of distances from the data points to the fitting line, and this criterion, in Legendre’s manner, can be replaced with the
volumes of spheres with the radii equal to the distances from the data points to the fitting line,

\[ E_\perp^3 = \sum_i \frac{4\pi}{3} \left[ d\left((x_i, y_i, z_i), F_x, y, z; \alpha_1, \alpha_2, \ldots, \alpha_n\right) \right]^3 \tag{9} \]

where \( F_{x, y, z; \alpha_1, \alpha_2, \ldots, \alpha_n} \) denotes a line in 3D described by implicit or explicit equations containing \( n \) parameters \( \alpha_1, \alpha_2, \ldots, \alpha_n \).

The idea is illustrated in Figure 7, where \( F \) is a straight line.

Obviously, this approach can be used in \( n \)-dimensional space too, where we have to minimize the sum of hyper-volumes of the hyperspheres with radii that are equal to the distances from the data points to the fitting line. Besides fitting data by lines, we can consider fitting data by geometric shapes in \( n \)-dimensional space (recall the example of fitting data points by an ellipse in previous section, or imagine fitting a set of 3D data by the surface of an ellipsoid), which is a part of image processing theory.

Such an approach can have many unexpected applications: for example, the description of national economies in state space, where the 3D data describing the behavior of national economies have been fitted by planes. In other words, there is a uniform approach to fitting lines (either straight or curved) and shapes in \( n \)-dimensional space by minimizing the volumes of \( n \)-dimensional spheres with radii equal to the orthogonal distances from the data points to the fitting line or shape. Maybe Gauss's method, which has not been successfully reproduced until today, was close to such a viewpoint?

**Further Reading**


Least Squares or Least Circles?

Online Supplement:

Partial derivative with respect to the parameter $a$ equal zero is

$$\frac{\partial E_a^2}{\partial a} = \frac{-2}{(1 + b^2)} \sum_{i=1}^{n} (y_i - (a + bx_i)) = 0$$

which implies

$$\sum_{i=1}^{n} (y_i - (a + bx_i)) = 0$$

$$\sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i,$$

where

$$\bar{y} = a + b\bar{x} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

Partial derivative with respect to the parameter $b$ equal zero is

$$\frac{\partial E_b^2}{\partial b} = \frac{(1 + b^2)[-2 \sum_{i=1}^{n} (y_i - (a + bx_i)) - 2b \sum_{i=1}^{n} (y_i - (a + bx_i))^2]}{(1 + b^2)^2}$$

$$= -\frac{2b \sum_{i=1}^{n} (y_i - (a + bx_i))^2}{(1 + b^2)} = 0$$

Since

$$\sum_{i=1}^{n} (y_i - (a + bx_i)) = 0$$

so

$$0 = \sum_{i=1}^{n} (y_i - (a + bx_i))^2 = \sum_{i=1}^{n} (y_i - (\bar{y} - b\bar{x} + bx_i))^2$$

$$= \sum_{i=1}^{n} [(y_i - \bar{y}) - b(x_i - \bar{x})]^2$$

Quadratic equation of $b$ is:

$$b^2 S_{xx} - 2b S_x S_y + S_{yy} = 0$$

where

$$b_{1,2} = \frac{-S_x S_y \pm \sqrt{(S_x S_y)^2 - 4 S_x S_y}}{S_{xx}}$$
Dear Readers,

This issue of CHANCE begins with an article by Jana Asher on collecting data in challenging settings. In particular, Asher describes her experiences conducting in-person survey interviews in East Timor. She gives us personal anecdotes, practical statistical advice, and an interesting story.

Qi Zheng explains the origins of the Luria-Delbrück distribution and its role in studying evolutionary change in E. coli. The statistical reasoning underlying the phenomenon has a connection to the distribution of slot machine returns.

Holmes Finch’s article, “Using Item Response Theory to Understand Gender Differences in Opinions on Women in Politics,” compares and contrasts item response models and how they describe a data set. The models are explained using formulas, pictures, and examples.

In Volume 22, Number 4, Jürgen Symanzik proposed a puzzle based on 10 data points and a set of seven instructions. Contest winner Stephanie Kovalchik, a graduate student at UCLA, provided a solution in the form of an amusing letter and an illustrative graphic. The 10 data values were flight times in seconds recorded on the log 10 scale of the Space Shuttle Challenger. Brad Thiessen earned honorable mention for his graph that included temperature and historical facts.

Bernard Dillard asks, “Who turned out the lights?” We are all concerned with energy demand and production. Bernard uses a discrete wavelet transformation to analyze electricity consumption data measured on a frequent time scale. The fit of the model is used in multiscale statistical process control. The ultimate goal is to be able accurately predict points of extreme energy demand and respond appropriately.

Students in virtually all statistics courses learn something of least squares estimation when studying prediction of an outcome from an explanatory variable. Ivo Petras and Igor Podlubny ask whether there is a reasonable alternative to the default criterion. “Least circles” is presented for your consideration.

To introduce students to concepts of design of experiments, instructors sometimes have students conduct taste tests of various food items, such as gummy bears (see Vol. 23, No. 1). John Bohannon, Robin Goldstein, and Alexis Herschkowitsch compared dog food and pâté. Really, they did. Read about their design and the results in this issue.

Ronald Smeltzer shows us an early time-line bar graph by Philippe Buache depicting the water level of the Seine River in Paris from 1760 to 1766. The picture creatively and effectively depicts data in print before the advent of the modern printing techniques that we enjoy today.

Howard Wainer, in his Visual Revelations column, writes about the graphics in the 2008 National Healthcare Quality Report and State Snapshots. Usefully and accurately displaying information graphically is important and challenging. Wainer makes suggestions for improving some of the displays.

Continuing a series of articles on postage stamps, Peter Loly and George P. H. Styan discuss stamps issued in sheets with 5x5 Latin square designs. Color versions of the stamps, as well as previous articles on stamps, are available online at www.amstat.org/publications/chance.

Jonathan Berkowitz’s puzzle celebrates the 2010 Winter Olympics, which was held in his home city of Vancouver, British Columbia. The puzzle, titled “Employs Magic,” is actually five smaller puzzles, each a cryptic five-square of 10 words.

Mark Glickman’s Here’s to Your Health column will appear in the next issue.

In other news, the Executive Committee of the ASA met recently and made decisions that impact CHANCE. First, the committee voted to continue CHANCE for another three years in both print and online versions. The next executive editor will serve 2011–2013. I’ll enjoy reading CHANCE in the years to come. Second, the Executive Committee voted to make the online version of CHANCE free to the ASA’s certified student members. This is a great development, because students are potential long-term subscribers and future authors. They also can be inspired by the significant role that probability and statistics can play in major studies and activities. I hope that other professionals will be motivated to submit articles to CHANCE to entertain and influence this group.

I look forward to your suggestions and submissions.

Enjoy the issue!

Mike Larsen