Complements on Simple Linear Regression

Terry R. McConnell
Syracuse University

March 16, 2015

Abstract

We present a simple-minded approach to a variant of simple linear regression that seeks to minimize the sum of squared distances to a line rather than the sum of squared vertical distances. We also discuss variants where distances rather than squared distances are summed.

1 Introduction

The Park Service is planning to build a number of shelters at selected scenic locations throughout a large tract of forest. Since much of the labor in this project involves carrying materials and equipment to each shelter site from the nearest road, the Service also plans to construct a straight section of supply road that will minimize the amount of labor involved. How should the road be located?

If we assume labor is proportional to squared distance rather than distance, as is mathematically convenient, then this type of problem is a first cousin of the problem of simple linear regression from elementary statistics. In simple linear regression one is given a data set in the form of a *scatter-plot*, a plot depicting the locations of points in a 2 dimensional Cartesian coordinate system, and one seeks to find the line that best fits the set of points in the sense that it minimizes the sum of squared vertical distances from the points to the line. The type of problem where one seeks to minimize the squared distances, as opposed to squared vertical (or horizontal) distances, will be discussed in this paper. We shall call it *orthogonal regression*.

Regression problems arise in statistics in two quite different contexts, which we review here briefly.

(I) Measurement Error Models

In problems involving measurement error models, one posits that a linear relationship $y = mx + b$ holds between two physical quantities x and y, but the exact relationship is obscured by measurement error when actual pairs (x_i, y_i) are observed. A common mathematical model is to suppose that

$$y_i = mx_i + b + \epsilon_i, i = 1, 2, \ldots, n,$$

where the measurement errors ϵ_i are assumed to be independent and normally distributed with mean zero, and either known or unknown variance σ^2. The problem is to infer the values of the model parameters – m, b, and perhaps σ^2 – from the observed data. The maximum likelihood estimators of m and b are obtained as the slope and y-intercept of the least squares (vertical) regression line described above.

In actual collection of data, the x_i are often selected by the experimenter and then corresponding y_i are observed. In this context it is natural to think of x as the "independent variable" and y as the "dependent variable", but the underlying mathematical model is symmetric in x and y. If we switch to y as the independent variable then the analogous statistical procedure could be termed *horizontal regression*: it would determine the line that minimizes the sum of squared horizontal distances to points in the scatter-plot.

Horizontal and vertical regression generally give different solutions, which might be regarded as a jarring asymmetry in light of the symmetry of the underlying model.

(II) Bivariate Normal Parameter Estimation

Recall that the bivariate normal distributions govern the joint distributions of (x, y) pairs in some models. They depend on 5 parameters: the mean and variance of the marginal distributions of x and y, and the correlation coefficient between x and y. Observed data can again be depicted with scatter-plots, but in this case both coordinates are random. However, if we condition on either the x values or the y values, then the resulting model is mathematically equivalent to a measurement error model, and the maximum likelihood estimators can be found using horizontal and vertical regression.

In this case, the decision of which variable to regard as given is entirely arbitrary.

In the context of both models discussed above, orthogonal regression can be viewed as an alternative method that is symmetric in its treatment of the variables and therefore more mathematically appealing. It is generally regarded as being significantly less tractable than horizontal or vertical regression, but, as we show in the next section, at least in the two-dimensional case it is possible to give a treatment that is no more involved than the usual derivations of least squares lines one sees in calculus and statistics courses.

Orthogonal regression is also known as *Deming regression*, after H.W. Deming who wrote about the problem in the 1940s, but the roots of the problem and its solution date back to the late 19th century. See, for example, [3]. For an exposition similar in spirit to this one, but using somewhat different methods, see [2].

In section 2 we discuss orthogonal regression, and in section 3 we provide some examples to illustrate the results of section 2. In section 4 we discuss absolute value regression, in which distances replace squared distances.

It is a pleasure to thank Vincent Fatica for some interesting conversations on the subject of this paper. In particular, Vince suggested the idea for the Park Service problem stated above, and provided the argument to handle the annoying special case of Theorem 4.1. I also wish to thank Thomas John for pointing out reference [3].

2 Results

A convenient parametrization of the set of affine linear functions on $\mathbb{R}^2$ is given by

$$L(x, y) = \cos\theta x + \sin\theta y + \lambda, \theta \in [0, 2\pi), \lambda \in \mathbb{R}.$$

We obtain all lines in the plane as the zero sets of affine functions. In what follows we will abuse notation by identifying lines with the affine functions that vanish on them.

The orthogonal distance d from a given point $\mathbf{x}_0 = (x_0, y_0)$ to L is

$$d^2 = L(x_0, y_0)^2 = |\cos\theta x_0 + \sin\theta y_0 + \lambda|^2.$$

To see this, note that the point $\mathbf{x}_1 = (-\lambda\cos\theta, -\lambda\sin\theta)$ lies on L, and then compute the component of $\mathbf{x}_0 - \mathbf{x}_1$ in the direction of the unit normal – the vector with components $(\cos\theta, \sin\theta)$.

Now suppose we are given a scatter-plot of n points $(x_i, y_i), i = 1, 2, \ldots, n$, and we are to find the values of θ and λ that minimize the *sum of squared orthogonal residuals* given by

$$R_o^2(L) = \sum_{j=1}^{n} (\cos\theta x_j + \sin\theta y_j + \lambda)^2.$$

Setting $\frac{\partial R_o^2}{\partial \lambda} = 0$ we obtain that $L(\bar{x}, \bar{y}) = 0$, i.e., the minimizing line passes through the centroid, as in horizontal and vertical regression. (Here $\bar{x}$ denotes the arithmetic mean of the x-coordinates and $\bar{y}$ denotes the arithmetic mean of the y-coordinates.) It is convenient, then, to put the origin at the centroid and minimize

$$R_o^2(L) = \sum_{j=1}^{n} (\cos\theta x_j + \sin\theta y_j)^2,$$

over $\theta \in [0, 2\pi)$. A little manipulation shows that the normal equation $\frac{\partial R_o^2}{\partial \theta} = 0$ can be rewritten as

$$\frac{1}{2}\tan 2\theta = \frac{SS(xy)}{SS(x) - SS(y)}, \tag{2.1}$$

at least when $SS(x) \neq SS(y)$. Here we use the standard notations

$$SS(x) = \sum_{j=1}^{n}(x_j - \bar{x})^2, \quad SS(y) = \sum_{j=1}^{n}(y_j - \bar{y})^2, \quad SS(xy) = \sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})$$

from Statistics.

It is important to note here that R_o^2 has both a maximum and a minimum value if we restrict to lines that pass through the centroid. This is in contrast to the usual quantities R_v^2 (R_h^2) from vertical (horizontal) regression, where nearly vertical (horizontal) lines lead to unbounded values. In the case of orthogonal regression, it is clear that R_o^2 cannot exceed the square of the diameter of the scatter-plot.

If $S(xy) = 0$ then we have

$$R_o^2(L) = \cos^2\theta SS(x) + \sin^2\theta SS(y).$$

If $SS(x) = SS(y)$ also, then R_o is constant and there is no unique minimum. If $SS(x) < SS(y)$ then the minimum occurs for $\theta = 0$ and π, and the maximum occurs for $\theta = \frac{\pi}{2}$ and $\frac{3\pi}{2}$. If $SS(x) > SS(y)$ then the maximum occurs for $\theta = 0$ and π, and the minimum occurs for $\theta = \frac{\pi}{2}$ and $\frac{3\pi}{2}$.

If $SS(xy) \neq 0$ and $SS(x) = SS(y)$, then we may interpret (2.1) as yielding solutions $\theta = \frac{\pi}{4}, \frac{3\pi}{4}, \frac{5\pi}{4}$, and $\frac{7\pi}{4}$. Thus, in all cases where $SS(xy) \neq 0$, there are exactly 4 solutions in $[0, 2\pi)$ that differ by multiples of $\frac{\pi}{2}$. Pairs that differ by π describe the same line. The minimum line is the one whose slope has the same sign as $SS(xy)$. The maximum line is always perpendicular to the minimum line. (The latter fact is intuitive on geometric grounds, without being completely obvious.)

Orthogonal regression "outperforms" both vertical and horizontal regression in the sense that if L_o, L_v, and L_h denote the respective minimum lines, then

$$R_o^2(L_o) \leq \frac{1}{2}R_v^2(L_v) + \frac{1}{2}R_h^2(L_h). \tag{2.2}$$

To see this, note that neither the horizontal nor the vertical distances from a point to a line can be less than the orthogonal distance. Thus we have that $R_o^2(L) \leq R_v^2(L)$ and $R_o^2(L) \leq R_h^2(L)$ for any test line L. It follows that

$$R_o^2(L_o) \leq R_o^2(L_v) \wedge R_o^2(L_h) \leq \frac{1}{2}R_v^2(L_v) + \frac{1}{2}R_h^2(L_h).$$

It is perhaps worth noting that a much stronger inequality holds for any given test line:

$$R_o^2(L) \leq \frac{1}{4}R_v^2(L) + \frac{1}{4}R_h^2(L). \tag{2.3}$$

This follows from some elementary geometry: the squared distance from vertex to hypotenuse in a right triangle is equal to half the harmonic mean of the squares on the legs. This, in turn, cannot exceed half the arithmetic mean of the squares on the legs, and the desired inequality follows by summing these inequalities over points in the scatter-plot.

Plane-geometric results often assume their most elegant form when expressed in terms of complex numbers. Let w be a non-zero complex number. Then the set of $z \in \mathbb{C}$ such that $\mathcal{R}(\bar{w}z) = 0$ is a line through the origin. We shall call this the line determined by w.

Theorem 2.1. *Let $(x_j, y_j), j = 1, 2, \ldots, n$ be points in the plane having the origin as centroid, and w be a non-zero complex number. Put $z_j = x_j + iy_j$. Then the line determined by w is extremal for orthogonal regression if and only if*

$$\mathcal{I}(\bar{w}^2 \sum_{j=1}^{n} z_j^2) = 0.$$

The proof is a straightforward verification using (2.1). (The centroid is assumed to lie at the origin for convenience only. In general one needs to subtract the arithmetic mean of the z_j from both w and each of the z_j.)

3 Examples

In this section we provide 3 examples to illustrate the foregoing.

Example 1 The collection of 3 points with coordinates (-2,-1), (1,-1), and (1,2) has centroid at the origin. For this arrangement, L_o is the line through the origin of slope 1; and L_v and L_h are, respectively, the lines through the origin with slope 1/2 and with slope 2. We have $R_o^2(L_o) = 3$, while both $R_v^2(L_v)$ and $R_h^2(L_h)$ are equal to 4.5. Also, $R_v^2(L_o)$ and $R_h^2(L_o)$ both equal 6, showing that equality can hold in (2.3).

Example 2 The 4 corners of the unit square also have centroid at the origin. In this case $SS(xy) = 0$ and $SS(x) = SS(y) = 4$. All lines L through the origin have $R_o^2(L) = 4$, so in this case there is no unique minimizing L_o. Also, L_v is horizontal and L_h is vertical, with $R_v^2(L_v) = R_h^2(L_h) = 4$, showing that equality can hold in (2.2).

Example 3 ([1], Example 14.3.9) A survey of agricultural yield against amount of a pest allowed (the latter can be controlled by crop treatment) gave the data in the following table, where x_i is the pest rate and y_i is the crop yield:

i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	8	6	11	22	14	17	18	24	19	23	26	40
y_i	59	58	56	53	50	45	43	42	39	38	30	27

In this example, L_v has slope-intercept equation $y = -1.013x + 64.247$, L_h has equation $y = -1.306x + 69.805$, and L_o has equation $y = -1.172x + 67.264$.

4 Absolute Value (AV) Regression

In absolute value regression, also known as L^1 regression, residuals are obtained by adding distances (horizontal, vertical, or orthogonal) rather than squared distances. This type of regression is more natural for the Park Service problem than are the forms of regression discussed above.

As before, we take as given a set of points $(x_i, y_i), i = 1, 2, \ldots n$, which we shall call *the configuration.* In vertical AV regression we seek to minimize the *residual sum,* given by

$$R_v = \sum_{j=1}^{n} |y_j - mx_j - b|,$$

over all possible lines L having slope-intercept equation $y = mx + b$. In the case of orthogonal AV regression, the summands are replaced by the (orthogonal) distances from the points to the test line. This quantity is denoted R_o.

Theorem 4.1. *If $n \geq 2$ then in all 3 cases of AV regression (vertical, horizontal, and orthogonal) there is an extremal line passing through two distinct points of the configuration.*

This result provides an effective procedure for finding a best-fit line: simply try all possible lines through pairs of points in the configuration and select the one that gives the smallest residual sum. Given the speed of today's computers, this procedure is even practical for problems of modest size.

Theorem 4.1 is a well-known "folklore" result, at least in the case of vertical regression, but some proof sketches we have seen in the literature miss (or at least fail to mention) one annoying detail. Accordingly, it seems desirable to give a proof here.

Consider first the case of vertical AV regression. Let L be a line that avoids all points in the configuration. We shall argue that L either does not minimize R_v, or can be replaced by another line with the same value of R_v that does go through some point in the configuration.

Since no point lies on L, all points can be classified as belonging to one of two groups: group (I) - points lying above L, i.e. (x_j, y_j) for which $y_j > mx_j + b$; and group (II) - points lying below L. If more than half of the points belong to group I, then translating the line vertically, i.e., in the direction of the positive y-axis, decreases R_v until a point of the configuration is encountered. Similarly, if more than half of the points belong to group II, then a better line can be obtained by translating L downward. Finally, if the number of points in the two groups is equal, L can be translated either up or down without changing the value of R_v until some point of the configuration is encountered.

In any case, L can be replaced by a line going through at least one point of the configuration and that has a value of R_v that is no larger than the value L had itself. Accordingly, we may assume that L passes through some point of the configuration. Without loss of generality, that point is (x_1, y_1).

Next, we repeat the argument, but with rotation about (x_1, y_1) in place of vertical translation. Assume that L does not pass through any other point of the configuration. Let θ be the angle L makes at (x_1, y_1) with the horizontal.

A straightforward, but somewhat tedious, calculation shows that

$$R_v = \sum_{j=1}^{n} |y_j - y_1| - (S_I - S_{II}) \tan\theta, \tag{4.1}$$

where S_I is the sum of $|x_j - x_1|$ over points in group I, and S_{II} is the sum over points in group II. A differential change $d\theta$ in the inclination of L that does not cause the line to encounter any other point of the configuration produces a differential change in R_v of $-(S_I - S_{II}) \sec^2\theta \, d\theta$. Thus, if $S_I \neq S_{II}$, the sign of $d\theta$ can be chosen so as to lead to a smaller value of R_v for the rotated line. On the other hand, if $S_I = S_{II}$, then the value of R_v is independent of θ, and L can be rotated until some other point of the configuration is encountered without changing the value of R_v. See (4.1).

Thus, in any case, we can replace L by a line - produced by rotation around (x_1, y_1) - that goes through a second point of the configuration, and that has a value of R_v that is no larger than the value L had itself.

The case of horizontal AV regression is similar.

Finally, consider a best-fit line L for orthogonal AV regression, and let R be the value of R_o for this line. Rotate the coordinate system about the origin until the line L becomes horizontal. Since rotations preserve distance, the resulting horizontal line is necessarily a best-fit line for the rotated configuration under orthogonal AV regression. It also necessarily a best fit line for the rotated configuration under vertical AV regression, since the minimum value for R_v cannot be less than R. It may be that this line does not go through two points of the configuration, but if so, it can be replaced by another having the same value of $R_v = R$, and that does go through two points. But for this line we must have $R_o = R_v = R$, since $R_o \leq R_v$ for any line. (Indeed, the latter line must also be horizontal, in addition to passing through two points of the configuration.) Applying the inverse rotation transforms this line into one that has a value of R_o that is equal to R, and that goes through two points of the original configuration.

References

[1] E. Dudewicz and S. Mishra, *Modern Mathematical Statistics,* Wiley, New York, 1988.

[2] P. Glaister, Least Squares Revisited, *The Mathematical Gazette* **85**(2001),104-107.

[3] K. Pearson, On Lines and Planes of Closest Fit to Systems of Points in Space, *Philosophical Magazine* **2**(1901), 559-572.