FITTING LINES TO DATA SETS IN THE PLANE

Suppose we have a set of observations (X_j, Y_j) in the plane for $j = 1, ..., n$ with $n \geq 2$ and we want to fit a line as well as possible to the points.

1. Classical y-on-x regression.

This is the oldest and best-known form of regression. Many textbooks present only this form. It is suitable when: X_j are fixed "design points" x_j, or at least have much less random variation in them than the Y_j do, and one wants to predict values of y for other values of x. To fit a line $y = a + bx$ to the data, one minimizes the sum of squared vertical distances from the points to the line,

$$S(a, b) = \sum_{j=1}^{n} (Y_j - a - bx_j)^2. \tag{1}$$

The line is only unique if not all x_j are equal, so let's assume that

$$s_x^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \overline{x})^2 > 0, \tag{2}$$

where $\overline{x} = \sum_{j=1}^{n} x_j/n$. First let b be fixed and consider minimization with respect to a. Clearly $S(a, b) \to +\infty$ as $a \to \pm\infty$, and $S(a, b)$ is quadratic in a, so it's minimized where

$$\partial S(a, b)/\partial a = -2 \sum_{j=1}^{n} Y_j - a - bx_j = 0.$$

Letting $\overline{Y} = \frac{1}{n} \sum_{j=1}^{n} Y_j$, we get $\overline{Y} - a - b\overline{x} = 0$, so the minimizing value of a given b satisfies

$$\widehat{a} = \widehat{a}(b) = \overline{Y} - b\overline{x}. \tag{3}$$

This implies that whatever b is, the line $y = \widehat{a}(b) + bx$ will go through the point $(\overline{x}, \overline{Y})$.

Plugging the value $a = \widehat{a}(b)$ from (3) into (1) gives

$$S(b) = S(\widehat{a}(b), b) = \sum_{j=1}^{n} (Y_j - \overline{Y} - b(x_j - \overline{x}))^2. \tag{4}$$

Date: 18.650, Sept. 21, 2015.

This is a quadratic function of b. The coefficient of b^2 is $(n-1)s_x^2 > 0$ by (2). Thus $S(b) \to +\infty$ as $b \to \pm\infty$, and $S(b)$ is minimized when

$$0 = S'(b) = \sum_{j=1}^{n} -2(Y_j - \overline{Y})(x_j - \overline{x}) + 2b(x_j - \overline{x})^2. \tag{5}$$

Let the sample covariance of x and Y be defined by

$$\mathrm{scov}(x, Y) = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \overline{x})(Y_j - \overline{Y}). \tag{6}$$

In terms of this, the unique solution of (5) for b is

$$b = \widehat{b} = \mathrm{scov}(x, Y)/s_x^2. \tag{7}$$

Then $\widehat{a} = \widehat{a}(\widehat{b})$ gives us a unique value of $a = \widehat{a}$, called the (estimated) "intercept" in the regression, meaning that it's the value of y at the point where the line $y = \widehat{a} + \widehat{b}x$ crosses the y axis. Naturally, $\widehat{b}$ is called the (estimated) slope.

2. Correlation

For any $Y_1, ..., Y_n$ not all equal, and $X_1, ..., X_n$ not all equal, recalling $s_X^2 := \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2$ and $s_X := \sqrt{s_X^2}$, likewise let $s_Y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \overline{Y})^2$ and $s_Y = \sqrt{s_Y^2}$. The *sample correlation* of X and Y is defined by $r_{X,Y} = \mathrm{scov}(X, Y)/(s_X s_Y)$, which is dimensionless. The slope $\widehat{b}$ of the y-on-x regression can be written as $\widehat{b} = r_{X,Y} s_Y / s_X$ (as also stated in Rice, §14.2.3 p. 561).

When X_j are random variables, not necessarily design points x_j, so the prescribed conditions for y-on-x regression may not hold, and Y_j also are random variables, so that x-on-y regression as in the next section may also not satisfy the usual regression models, the sample correlation $r_{X,Y}$ is still well-defined.

3. Regressing x on y

It can happen that data are given such that Y_j are design points or contain little random contribution or error, while X_j are random variables, and we want to predict the value of x given a new value of y. Then we can just reverse the roles of x and y to do x-on-y regression. This will minimize the sum of squares of horizontal deviations of the data points from a line. It will be uniquely defined if $s_Y > 0$. We will get a line $x = c + dy$ which will also pass through $x = \overline{X}$, $y = \overline{Y}$. The estimated value of d will satisfy $\widehat{d} = r_{X,Y} s_X / s_Y$.

If all the points are on a line, then that line will clearly be the best-fitting line either for vertical deviations (y-on-x) or horizontal deviations (x-on-y) because these deviations will be 0 in that case. It may be surprising that these are the only times these two regressions agree:

Theorem 1. *For given observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ in the plane, where $n \geq 2$, $s_X^2 > 0$ and $s_Y^2 > 0$, the lines given by y-on-x and x-on-y regression only agree when all the points (X_i, Y_i) are on a line.*

Proof. Both regression lines pass through the point $(\overline{X}, \overline{Y})$. The slope of the y-on-x line is $r \cdot s_Y/s_X$.

The slope of the x-on-y line, if we take the y axis as horizontal and the x axis as vertical, is then $r \cdot s_X/s_Y$. In the original orientation where the x axis is horizontal and the y axis is vertical, the slope is replaced by its reciprocal, which is $(1/r)s_Y/s_X$. So, the two lines are only the same if $r = 1/r$ so $r^2 = 1$, $r = \pm 1$. Then the points (X_i, Y_i) are all on a line (with positive slope if $r = 1$ or negative slope if $r = -1$). □

About notation: Rice on p. 561 uses definitions of sample covariances and variances with a factor of $\frac{1}{n}$ rather than $\frac{1}{n-1}$. Note that in his next three displays after the definitions, such factors will appear both in the numerator and denominator, so they will divide out. One just needs to be consistent in using one factor or the other. Also note that what Rice calls s_{xx} and s_{yy} are estimators of sample variance (as opposed to standard deviation).

Theorem 1 implies that the two regression lines will in nearly all cases be different (if $n \geq 3$). If the y-on-x regression line has a positive slope, but the correlation $r < 1$, then the x-on-y line always has a larger slope, by a factor of $1/r^2$. In many situations, the assumptions for y-on-x and x-on-y regression may not hold. There is another possible method, as follows.

4. Line-fitting by distance: errors-in-variables regression.

Now suppose X_j and Y_j are both random variables, measured in the same units, so that Euclidean distances in the plane are also in these units. A third way to fit a line to the set of points $(X_1, Y_1), \ldots, (X_n, Y_n)$ is to minimize the sum of squared perpendicular distances of the points to the line. Such distances make good sense when X_j and Y_j are in the same units, not such good sense otherwise. If they are in different units one can just find the sample correlation $r_{X,Y}$ mentioned above.

For any point p and line L in the plane, let $d(p, L)$ be the perpendicular distance from p to L. Given observations (X_j, Y_j), $j = 1, ..., n$, a line L_o will be called a *bfsd line (best-fitting by squared distance line)* if $\sum_{j=1}^{n} d((X_j, Y_j), L)^2$ is minimized at $L = L_o$.

Let $L_{a,b}$ be the line $y = a + bx$ for any real numbers a, b. Let $L_{\infty;c}$ be the vertical line $x \equiv c$, $-\infty < y < \infty$. So every line in the plane is either a line $L_{a,b}$ or a line $L_{\infty;c}$ for some a, b or c. Then bfsd lines are characterized as follows.

Theorem 2. *For any given (X_j, Y_j), $j = 1, ..., n$, there is at least one bfsd line. All such lines go through the point $(\overline{X}, \overline{Y})$. If $s_X = s_Y = 0$, or $s_X = s_Y > 0$ and $r = r_{X,Y} = 0$, then every line through $(\overline{X}, \overline{Y})$ is a bfsd line.*

In all other cases the bfsd line L is unique.

If $s_X > 0 = s_Y$ then $L = L_{\overline{Y},0}$, or if $s_X = 0 < s_Y$ then $L = L_{\infty;\overline{X}}$.

If $s_X > 0$ and $s_Y > 0$ then: if $r = 0$ and $s_X^2 > s_Y^2$ then $L = L_{\overline{Y},0}$, or if $s_X^2 < s_Y^2$ then $L = L_{\infty;\overline{X}}$.

If $s_X > 0$, $s_Y > 0$ and $r \neq 0$ (the general case) then $L_o = L_{a,b}$ has slope $b = \tan\theta$ (which, given θ, uniquely determines the line as $(y - \overline{Y}) = b(x - \overline{X})$) and θ is as follows:

If $s_X > s_Y$ then $\theta = \theta_I$ where

$$\theta_I = \frac{1}{2}\tan^{-1}\left[\frac{2\,\mathrm{scov}(X, Y)}{s_X^2 - s_Y^2}\right]. \tag{8}$$

If $s_X < s_Y$ then $\theta = \theta_{II}$, defined as $\theta_I + \pi/2$.
If $s_X = s_Y$ then since $r \neq 0$, $\mathrm{scov}(X, Y) \neq 0$ and:
if $\mathrm{scov}(X, Y) > 0$, $\theta = \pi/4$, $b = 1$;
if $\mathrm{scov}(X, Y) < 0$, $\theta = -\pi/4$, $b = -1$.

Proof. In each of the following cases, all the points (X_j, Y_j) are on the given line L, so $\sum_{j=1}^{n}(d((X_j, Y_j), L)^2) = 0$ and L is a bfsd line: $s_X > 0 = s_Y$, so $Y \equiv \overline{Y}$ is constant and $L = L_{\overline{Y},0}$ is horizontal; or $X_j \equiv \overline{X}$ is constant, $s_X = 0 < s_Y$ and $L = L_{\infty;\overline{X}}$, the vertical line $x \equiv \overline{X}$; or all (X_j, Y_j) equal one point $(\overline{X}, \overline{Y})$, i.e. $s_X = s_Y = 0$, and L is any line through $(\overline{X}, \overline{Y})$, either a line $y - \overline{Y} = b(x - \overline{X})$ for any finite slope b, or the vertical line $L_{\infty;\overline{X}}$.

To find the distance $d((X, Y), L)$ from a point (X, Y) to a line L, if $L = L_{\infty;c}$ it's $|X - c|$. If $L = L_{a,0}$ it's $|Y - a|$. So suppose $L = L_{a,b}$ with $b \neq 0$. Here are two ways of evaluating the distance. First, here's a geometric-trigonometric way. The vertical distance from (X, Y) to $L_{a,b}$ is clearly $|Y - a - bX|$. A line M through (X, Y) perpendicular to $L_{a,b}$

forms an angle θ at (X, Y) with a vertical line. Then the perpendicular distance from (X, Y) to $L_{a,b}$ is $|Y - a - bX| \cos\theta$. On the other hand, one can see by drawing a diagram or otherwise that the line $L_{a,b}$ forms the same angle θ with a horizontal line. Thus the slope of $L_{a,b}$, namely b, equals $\tan\theta$. Here two different although symmetric diagrams could be given depending on whether $b > 0$ or $b < 0$. (By the way the French group of mathematical authors with the pseudonym Bourbaki decided that diagrams couldn't be part of proofs and so they have no diagrams in their books.) Anyhow,

$$\sum_{j=1}^{n} (d((X_j, Y_j), L_{a,b})^2) \;=\; \cos^2\theta \sum_{j=1}^{n} ((Y_j - bX_j - a)^2).$$

For fixed b, and so for fixed

$$1 + b^2 = 1 + \tan^2\theta = \sec^2\theta = 1/\cos^2\theta,$$

$\cos^2\theta$ will also be fixed, and the minimization to find a in terms of the other quantities is exactly as in y-on-x regression and gives the same result. Namely, we have a quadratic function of a, which goes to $+\infty$ as $|a|$ does. So it will be minimized at the unique point where the partial derivative with respect to a is 0, which gives $-2(\overline{Y} - b\overline{X}) + 2a = 0$, or $a = \overline{Y} - b\overline{X}$. This says that the point $(\overline{X}, \overline{Y})$ is on the line $L_{a,b}$, again, just as for y-on-x regression, and for each j

$$(9) \quad Y_j - a - bX_j = Y_j - (\overline{Y} - b\overline{X}) - bX_j = (Y_j - \overline{Y}) - b(X_j - \overline{X}).$$

The line $L_{a,b}$ through $(\overline{X}, \overline{Y})$ and the horizontal line $L_{\overline{Y},0}$ form some angles θ. As already indicated, we will take a θ such that the slope b equals $\tan\theta$. Recall that for any real number x, $\tan^{-1} x$ is an angle ϕ such that $\tan\phi = x$ and $-\pi/2 < \phi < \pi/2$. Then $\tan^{-1} x$ is uniquely defined since the tangent function is strictly increasing for $-\pi/2 < \theta < \pi/2$ and takes all real values there. The tangent function is periodic of period π. Thus, all angles ϕ such that $\tan\phi = x$ are of the form $\tan^{-1} x + m\pi$ where m is an integer, positive, negative or 0. On any interval of length π, containing just one of its endpoints, the tangent function takes all real values once each, and also goes to $\pm\infty$ at one point. It's convenient for present purposes to choose θ such that $-\pi/4 \le \theta < 3\pi/4$, which is an interval of length π containing only its lower endpoint. For any real number (slope) b this gives a unique θ such that $\tan\theta = b$.

The squared distance from $L_{a,b}$ to (X_j, Y_j) is, using (9),

$$[Y_j - \overline{Y} - (\tan\theta)(X_j - \overline{X})]^2 \cos^2\theta,$$

which since $\tan\theta = (\sin\theta)/\cos\theta$ equals

$$[(Y_j - \overline{Y})\cos\theta - (X_j - \overline{X})\sin\theta]^2.$$

We want to find θ to minimize the sum of these squares, which is

$$f(\theta) \;\equiv\; (n-1)\left[s_Y^2\cos^2\theta - 2\,\mathrm{scov}(X,Y)\sin\theta\cos\theta + s_X^2\sin^2\theta\right].$$

Since f is smooth and periodic of period 2π (actually, of period π because of the product and squaring), setting $f'(\theta) = 0$ we can expect to find at least one minimum and at least one maximum. They will turn out to be in perpendicular directions. We get

$$0 = f'(\theta)/(n-1) = 2\sin\theta\cos\theta(s_X^2 - s_Y^2) - 2\cos(2\theta)\,\mathrm{scov}(X,Y).$$

If $s_X^2 \neq s_Y^2$ this gives $\tan(2\theta) = 2\,\mathrm{scov}(X,Y)/(s_X^2 - s_Y^2)$. There are two solutions for θ, namely θ_I given by (8) and $\theta_{II} = \theta_I + (\pi/2)$, since $\tan(\phi+\pi) = \tan\phi$ for any ϕ. Then $-\pi/4 < \theta_I < \pi/4 < \theta_{II} < 3\pi/4$, so both θ_I and θ_{II} are in the chosen interval for θ. A point where $f'(\theta) = 0$ will be a relative minimum if $f''(\theta) > 0$. We have

$$f''(\theta)/(n-1) = 2\cos(2\theta)(s_X^2 - s_Y^2) + 2\sin(2\theta)\cdot 2\,\mathrm{scov}(X,Y).$$

At a point where $f'(\theta) = 0$ this becomes $f''(\theta)/(n-1) = 2(s_X^2 - s_Y^2)/\cos(2\theta)$. If $s_X^2 > s_Y^2$ we want $\theta = \theta_I$ since $\cos(2\theta_I) > 0$ for a minimum, and θ_{II} will give a maximum. If $s_X^2 < s_Y^2$ we want $\theta = \theta_{II}$ so that $\pi/2 < 2\theta < 3\pi/2$ and $\cos(2\theta_{II}) < 0$ for a minimum, while θ_I then gives a maximum.

Now, what if $s_X^2 = s_Y^2$? In that case $f'(\theta) = -2\cos(2\theta)\,\mathrm{scov}(X,Y)$. If $\mathrm{scov}(X,Y) = 0$, f is a constant and all θ, in other words all lines through $(\overline{X},\overline{Y})$, are equally good.

If $\mathrm{scov}(X,Y) \neq 0$ then we need $\cos(2\theta) = 0$, so we can take $\theta = \pm\pi/4$. Again we need to consider the second derivative, which is $f''(\theta) = 4\sin(2\theta)\,\mathrm{scov}(X,Y)$. To have $f''(\theta) > 0$ for a minimum of f, if $\mathrm{scov}(X,Y) > 0$ we want $\theta = \pi/4$, giving a line with slope 1. If $\mathrm{scov}(X,Y) < 0$ we want $\theta = -\pi/4$, giving a line with slope -1. This completes the proof of Theorem 2. □

If $1/(n-1)$ is replaced by $1/n$ in both sample variances and the sample covariance (as Rice does), the result is the same since these factors cancel out, appearing both in the numerator and denominator of (8), as long as it's done consistently.

5. Outliers

For any of the three kinds of regression mentioned, outlying values of x_j or Y_j can have a bad effect on the regression. Specifically, in y-on-x regression, if the smallest design point x_1 is far below the rest of the

x_j, or the largest one x_n is far above, then the data point (x_1, Y_1) or (x_n, Y_n) can have undue influence on the estimated slope, because the outlying point has excess leverage on it. A reference on such problems is the book by D. Belsley, the late E. Kuh, and R. E. Welsch (1980) (all three authors were at MIT at the time).

6. Historical notes

Reportedly, Gauss discovered y-on-x regression in 1794, but he did not publish it until 1809. Legendre first published it in 1805. There was a publication on it by an American, Robert Adrain, in 1808, in a journal he had just founded and which lasted only through 1814. Both Gauss and Legendre used regression in predicting the future positions of astronomical bodies. Adrain was concerned with surveying (land measurement).

Acknowledgment. Daniel Kane suggested the trigonometric formulation and result for best fitting by squared perpendicular distance in February, 2005.

REFERENCES

Adrain, Robert (1808), "Research concerning the probabilities of the errors which happen in making observations, &c," *The Analyst, or Mathematical Museum*, vol. 1, pp. 93-109.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New York.

Gauss, C. F. (1809), *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum.*

Legendre, A. M. (1805), *Nouvelles méthodes pour la détermination des orbites des comètes*; Appendix, "Sur la méthode des moindres carrés."