Least Squares

Math 260: Applied Linear Algebra and Matrices
J. Gerlach

Fall 2004, Fall 2005

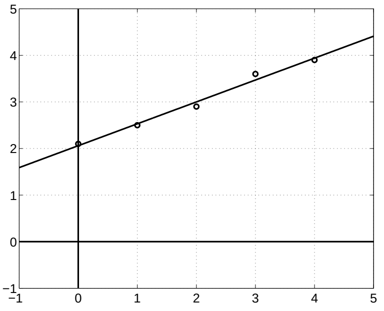
Contents

The goal of this paper is to explain least squares methods which are frequently introduced in Statistics by means of Linear Algebra. In particular, we will derive the formula for the least squares regression line. While the text for the course (Kolman) relies heavily on the QR-factorization in its treatment of least squares, we shall only use norms and dot products in this paper.

1 Problems

Problem 1: Find a line which *best* describes the data

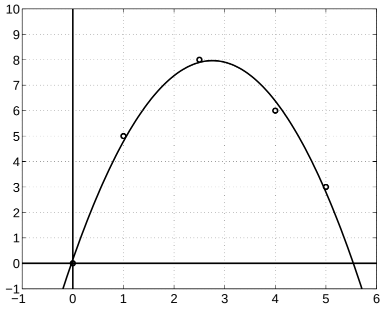
x	0	1	2	3	4
y	2.1	2.5	2.9	3.6	3.9



◇

Problem 2: Find a parabola which *best* describes the data

x	0	1	2.5	4	5
y	0	5	8	6	2



◇

For a given set of data (x_j, y_j), where $j = 1, 2, \ldots, n$ it is natural to define the data vectors $X = [x_1\ x_2 \ldots x_n]^T$ and $Y = [y_1\ y_2 \ldots y_n]^T$. A line of the form $y = mx + b$ which passes through all data points satisfies the system of equations

$$\begin{aligned} x_1 m + b &= y_1 \\ x_2 m + b &= y_2 \\ &\vdots \\ x_n m + b &= y_n \end{aligned}$$

Here m and b are the unknowns. We can write this system in matrix form $Au = y$, with

$$A = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \text{and} \qquad u = \begin{bmatrix} b \\ m \end{bmatrix}$$

Of course, we cannot find a single line which passes though more than two points, unless - by some coincidence - all points are co-linear. Thus we do not expect that the system $Au = y$ has a solution.

If we try to find a parabola $y = ax^2 + bx + c$ passing though the data points we need to satisfy the system

$$\begin{aligned} a x_1^2 + b x_1 + c &= y_1 \\ a x_2^2 + b x_2 + c &= y_2 \\ &\vdots \\ a x_n^2 + b x_n + c &= y_n \end{aligned}$$

for the unknowns a, b, c. Again, we can write this system in the form $Au = y$ with

$$A = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix} \qquad \text{and} \qquad u = \begin{bmatrix} c \\ b \\ a \end{bmatrix}$$

For more than three points it is generally not possible to find a parabola which passes through all points, and we cannot expect to find a solution to our system.

Notice, that the y_j appear on the right hand side of the systems, and the x-coordinates of the data points determine the matrix A. I set up the vector u *backwards*, beginning with c. As a result the matrix A contains the power of x in increasing order 1, x_j, x_j^2. This is just a personal preference, you could also use $u = [a\ b\ c]^T$ and swap the columns of A.

2 Least Squares Solutions

Let us study this problem in a more general setting. In both examples we were facing a linear system with more equations than unknowns. In the general matrix setting we attempt to solve $Ax = b$ where A is an $m \times n$ matrix with $m > n$. Since the system is over-determined we cannot expect that a solution exists. Therefore we try to cut our losses, and select x such that the residual vector is as small as possible. Let's be more specific. Denote the residual by r, then

$$r \;=\; b - Ax$$

We see that r depends on x, and our goal is to find x such that the square error $||r||^2$ is minimized. Spelling this out, we attempt to minimize the expression

$$E \;= ||r||^2 \;=\; r \bullet r \;=\; r^T r \;=\; r_1^2 + r_2^2 + r_3^2 + \cdots + r_m^2$$

Hence the terminology *least squares.*

The least squares error $||r||^2 = E$ depends on x, and we shall write $E(x)$ to indicate this dependence. A brief computation, using the laws of matrix algebra, shows

$$\begin{aligned} E(x) &= (b - Ax)^T\,(b - Ax) \;=\; b^T b - b^T Ax - x^T A^T b + x^T A^T Ax \\ &= x^T A^T Ax - 2x^T A^T b + b^T b \end{aligned}$$

Here we used $b^T Ax = x^T A^T b$, since 1×1 matrices are always symmetric. A similar computation yields (h is a vector)

$$\begin{aligned} E(x+h) &= (b - A(x+h)) \bullet (b - A(x+h)) \\ &= (x+h)^T A^T A(x+h) - 2(x+h)^T A^T b + b^T b \\ &= x^T A^T Ax + 2h^T A^T Ax + h^T A^T Ah \\ &\qquad -2x^T A^T b - 2h^T A^T b + b^T b \\ &= E(x) + ||Ah||^2 + 2h^T(A^T Ax - A^T b) \end{aligned}$$

Here we used $h^T A^T Ah = (Ah)^T(Ah) = ||Ah||^2$. Notice, that $A^T A$ is a symmetric, $n \times n$ square matrix, and if x is a solution of $A^T Ax = A^T b$, then we get

$$E(x+h) \;=\; E(x) + ||Ah||^2 \;\geq\; E(x)$$

for any choice of h. Hence, x is desired optimal solution of the minimization problem. Let us summarize the results:

Theorem: Given are an $m \times n$ matrix A and an m-vector b. We define the residual vector $r = r(x) = b - Ax$. If x solves the system $A^T Ax = A^T b$, then

$$||r(x+h)||^2 \geq ||r(x)||^2 \qquad \text{for all} \;\; h \in R^n \;,$$

i.e. x minimizes the residual in the least squares sense. ◇

The system of equations

$$A^T Ax = A^T b$$

is called the *normal equations*. A solution yields the optimal choice for the least squares problem. What is so normal about these equations? We may rewrite this system as $0 = A^T(b - Ax) = A^T r$. Hence, the residual vector r is perpendicular (normal) to the columns of A.

Problem 3: Find the least squares solution to the system

$$\begin{aligned} 2x + y &= 8 \\ x + y &= 4 \\ y &= 2 \end{aligned}$$

First let us experiment with this problem. $y = 2$ in the last equation, implies $x = 2$ from the second equation, which leads to $6 = 8$ in the first equation. Clearly, we have no solution. The residual for the choice $[x\ y]^T = [2\ 2]^T$ is $r = [2\ 0\ 0]^T$, with $||r||^2 = 4$. If we combine the first two equation alone, we get $[x\ y]^T = [4\ 0]$, and $y = 0$ clearly contradicts the last equation. The residual is $r = [0\ 0\ 2]^T$, again with $||r||^2 = 4$. Since the two residual vectors have the same norm, the two alternatives $x = 2$, $y = 2$ and $x = 4$, $y = 0$ are equally good, or equally bad. The point in question is whether we can find a better solution, namely one with a smaller residual vector.

Now let us solve the normal equations for this problem. The matrices $A^T A$ and $A^T b$ become

$$A^T A \;=\; \begin{bmatrix} 5 & 3 \\ 3 & 3 \end{bmatrix} \qquad \text{and} \qquad A^T b \;=\; \begin{bmatrix} 20 \\ 14 \end{bmatrix}$$

The resulting linear system is

$$\begin{aligned} 5x + 3y &= 20 \\ 3x + 3y &= 14 \end{aligned}$$

It has the solution $x = 3$ and $y = \frac{5}{3}$. The residual vector is

$$r = \begin{bmatrix} 8 \\ 4 \\ 2 \end{bmatrix} - \begin{bmatrix} 2 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ \frac{5}{3} \end{bmatrix} = \begin{bmatrix} 8 \\ 4 \\ 2 \end{bmatrix} - \begin{bmatrix} 7\frac{2}{3} \\ 4\frac{2}{3} \\ \frac{5}{3} \end{bmatrix} = \begin{bmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ \frac{1}{3} \end{bmatrix}$$

We see that none of the equations is satisfied exactly, but $||r||^2 = \frac{2}{3}$, which is well below 4, the value we found by solving two of the equation exactly. By inspection, we also see that the residual vector r is normal to the columns of A. ◇

Problem 2.1: Solution of Problem 2 using the normal equations. The matrix A takes the form

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2.5 & 6.25 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \end{bmatrix}$$

and we obtain

$$A^T A = \begin{bmatrix} 5 & 12.5 & 48.25 \\ 12.5 & 48.25 & 205.625 \\ 48.25 & 205.625 & 921.0625 \end{bmatrix} \quad \text{and} \quad A^T y = \begin{bmatrix} 21 \\ 59 \\ 201 \end{bmatrix}.$$

The solution of the normal equations becomes (after some work) $c = 0.0478$, $b = 5.9899$ and $a = -1.1215$, and the resulting approximating quadratic function is $q(x) = 0.0478 + 5.9899x - 1.1215x^2$. The residual vector is

$$r = \begin{bmatrix} -0.0478 \\ 0.0838 \\ -0.0131 \\ -0.0633 \\ 0.0404 \end{bmatrix}$$

with $||r||^2 = 0.0151$. The computations were done in matlab, the interested reader may want to verify that r is orthogonal to the columns of A. ◇

3 Regression Lines

3.1 Direct Computation

Now let's look at the case of finding the equation for the least squares regression line. This is a standard topic of statistical analysis. If the line is given in the form $y = mx + b$, the usual formulas are

$$m \; = \; \frac{n \displaystyle\sum_{k=1}^{n} \; x_k y_k - \left(\sum_{k=1}^{n} \; x_k \right) \left(\sum_{k=1}^{n} \; y_k \right)}{n \displaystyle\sum_{k=1}^{n} \; x_k^2 \; - \; \left(\sum_{k=1}^{n} \; x_k \right)^2}$$

$$b \; = \; \overline{y} - m \; \overline{x} \; ,$$

where $\overline{x} = \dfrac{1}{n} \displaystyle\sum_{k=1}^{n} \; x_k$ and $\overline{y} = \dfrac{1}{n} \displaystyle\sum_{k=1}^{n} \; y_k$ represent the means. The trained eye of the linear algebraist recognizes $\displaystyle\sum_{k=1}^{n} \; x_k y_k \; = \; x^T y \; = \; x \bullet y$ as a dot product, and $\displaystyle\sum_{i=k}^{n} \; x_k^2 = x^T x = ||x||^2$ as the square of a norm. Thus,

$$m \; = \; \frac{n \; x^T y - (n \; \overline{x}) \; (n \; \overline{y})}{n \; x^T x - n^2 \; \overline{x}^2} \; = \; \frac{x^T y - n \; \overline{x} \; \overline{y}}{x^T x - n \; \overline{x}^2}$$

Now we shall derive these formulas as solutions of the the normal equations. Recall that for the computation of the least squares regression line the matrix A takes the form

$$A \; = \; \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Thus,

$$A^T A \; = \; \begin{bmatrix} n & \displaystyle\sum_{k=1}^{n} \; x_k \\ \displaystyle\sum_{k=1}^{n} \; x_k & \displaystyle\sum_{k=1}^{n} \; x_k^2 \end{bmatrix} \; = \; \begin{bmatrix} n & n \; \overline{x} \\ n \; \overline{x} & x^T x \end{bmatrix}$$

and

$$A^T y \;=\; \begin{bmatrix} \sum\limits_{k=1}^{n} y_k \\ \sum\limits_{k=1}^{n} x_k y_k \end{bmatrix} \;=\; \begin{bmatrix} n\,\overline{y} \\ x^T y \end{bmatrix}$$

When we set up the normal equations $A^T A u = A^T b$ with $u = [b\ m]^T$ as our unknown, we arrive at

$$\begin{array}{rcrcl} n\,b & + & n\,\overline{x}\,m & = & n\,\overline{y} \\ n\,\overline{x}\,b & + & x^T x\,m & = & x^T y \end{array}$$

Application of the usual techniques for linear systems (details omitted) yields the solutions

$$\begin{aligned} m &= \frac{x^T y - n\,\overline{x}\,\overline{y}}{x^T x - n\,\overline{x}^2} \\ b &= \overline{y} - m\overline{x} \end{aligned}$$

which are identical to the formulas above. The regression line becomes

$$y \;=\; mx + b \;=\; mx + \overline{y} - m\overline{x} \;=\; \overline{y} + m(x - \overline{x})$$

Hence the regression line will always contain the point $(\overline{x}, \overline{y})$.

Problem 1.1: Determine the regression line for Example 1. We have the data

x	0	1	2	3	4
y	2.1	2.5	2.9	3.6	3.9

and a straight forward computations show

$$\sum_{k=1}^{5} x_k \;=\; 10 \qquad \sum_{k=1}^{5} y_k \;=\; 15 \qquad \sum_{k=1}^{5} x_k^2 \;=\; 30 \qquad \sum_{k=1}^{5} x_k y_k \;=\; 34.7$$

It follows that $\overline{x} = 2$ and $\overline{y} = 3$. The normal equations take the form

$$\begin{array}{rcrcl} 5b & + & 10m & = & 15 \\ 10b & + & 30m & = & 34.7 \end{array}$$

with solution $m = 0.47$ and $b = 2.06$. The regression line has equation

$$y = 0.47x + 2.06\ .$$

We see that for $x = \overline{x} = 2$ we get $y = 0.94 + 2.06 = 3 = \overline{y}$, and the regression line passes through $(\overline{x}, \overline{y})$. ◇

3.2 Data Preprocessing

Least squares analysis goes further. It is a known fact that the computations become more efficient if the data are pre-processed. We shall center the coordinates so that the respective means become zero, and we shall reset the scales for both axes so that both data vectors become unit vectors. Details: Let

$$s_x^2 = \sum_{k=1}^{n} (x_k - \overline{x})^2 \qquad \text{and} \qquad s_y^2 = \sum_{k=1}^{n} (y_k - \overline{y})^2$$

and define

$$u_k = \frac{x_k - \overline{x}}{s_x} \qquad \text{and} \qquad v_k = \frac{y_k - \overline{y}}{s_y}$$

The trained statistician recognizes standard deviations (up to a factor) and Z-scores. This is just a change of variables, $u = \frac{x-\overline{x}}{s_x}$ and $v = \frac{y-\overline{y}}{s_y}$. Once the least squares line has been found in the uv-setting, we can convert back to x and y using $x = \overline{x} + s_x u$ and $y = \overline{y} + s_y v$.

By construction we have

$$\begin{aligned} \overline{u} &= \frac{1}{n}\sum_{k=1}^{n} u_k = \frac{1}{n}\sum_{k=1}^{n} \frac{x_k - \overline{x}}{s_x} \\ &= \frac{1}{ns_x}\left(\sum_{k=1}^{n} x_k - \sum_{k=1}^{n} \overline{x}\right) = \frac{1}{ns_x}(n\overline{x} - n\overline{x}) = 0 \\ ||u||^2 &= u^T u = \sum_{k=1}^{n} u_k^2 = \sum_{k=1}^{n} \frac{(x_k - \overline{x})^2}{s_x^2} = \frac{1}{s_x^2}\, s_x^2 = 1\,, \end{aligned}$$

and in similar manner we obtain $\overline{v} = 0$ and $||v||^2 = 1$. In uv-coordinates the matrix $A^T A$ becomes

$$A^T A = \begin{bmatrix} n & \sum_{k=1}^{n} u_k \\ \sum_{k=1}^{n} u_k & \sum_{k=1}^{n} u_k^2 \end{bmatrix} = \begin{bmatrix} n & 0 \\ 0 & 1 \end{bmatrix}$$

and on the left hand side we have

$$A^T v = \begin{bmatrix} \sum_{k=1}^{n} v_k \\ \sum_{k=1}^{n} u_k v_k \end{bmatrix} = \begin{bmatrix} 0 \\ u^T v \end{bmatrix}$$

Hence, the normal equations reduce to (this is the promised simplification of the problem)

$$\begin{aligned} nb &= 0 \\ m &= u^T v \end{aligned}$$

Thus the regression line is $v = (u^T v)\, u$, and in the original variables we have the least squares line

$$y \;=\; \overline{y} + s_y v = \overline{y} + s_y(u^T v)u \;=\; \overline{y} + \frac{s_y(u^T v)}{s_x}\,(x - \overline{x})$$

Problem 1.2: Revisit the data from Problem 1, and confirm previous results using the *preprocessing* technique. We summarize the computations in a table.

x_k	$x_k - \overline{x}$	$(...)^2$	u_k	y_k	$y_k - \overline{y}$	$(...)^2$	v_k	$u_k v_k$
0	-2	4	-0.6325	2.1	-0.9	0.81	-0.6013	0.3803
1	-1	1	-0.3162	2.5	-0.5	0.25	-0.3341	0.1056
2	0	0	0	2.9	-0.1	0.01	-0.0668	0
3	1	1	0.3162	3.6	0.6	0.36	0.4009	0.1268
4	2	4	0.6325	3.9	0.9	0.81	0.6013	0.3803
10	0	10	0	15	0	2.24	0	0.9931

The last line contains the sums of the respective columns. Hence, $s_x^2 = \sum_{k=1}^{5}(x_k - 2)^2 = 10$, $s_y^2 = 2.24$ and $m = 0.9931 = \frac{4.7}{\sqrt{10}\sqrt{2.24}}$. In *uv*-coordinates we get $v = \frac{4.7}{\sqrt{10}\sqrt{2.24}}\, u$, which becomes

$$y \;=\; 3 + \frac{\sqrt{2.24}\;\frac{4.7}{\sqrt{10}\sqrt{2.24}}}{\sqrt{10}}\,(x-2) \;=\; 3 + 0.47(x-2) \;=\; 0.47x + 2.06$$

in xy-coordinates, as before. ◇

4 Correlation Coefficients

Not all data are suited for an approximation by a linear function. In an ideal match, when all points belong to a common line, all the equations $y_k = mx_k + b$ will be satisfied, which in uv-coordinates becomes $v_k = mu_k$. This means that the vector v is a multiple of the vector u. But we have normalized our vectors u and v to length one, hence, in the perfect match scenario we get $v = \pm u$, that is either $m = 1$ or $m = -1$. In this case the angle between u and v is either 0^o or 180^o. We choose the cosine of the angle between u and v as an indicator of the appropriateness of a linear model. If the cosine is close to ± 1, a linear model is appropriate, else it is not. The cosine of the angle between u and v is computed as

$$\cos\theta \;=\; \frac{u \bullet v}{||u||\;||v||} \;=\; u \bullet v \;=\; u^T v$$

since u and v are unit vectors. This cosine is usually denoted by ρ, and it is called the Pearson correlation coefficient. In the original variables it becomes

$$\rho \;=\; \sum_{k=1}^{n} u_k v_k \;=\; \sum_{k=1}^{n} \frac{x_k - \overline{x}}{s_x} \frac{y_k - \overline{y}}{s_y} \;=\; \frac{\displaystyle\sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y})}{\sqrt{\displaystyle\sum_{k=1}^{n} (x_k - \overline{x})^2}\;\sqrt{\displaystyle\sum_{k=1}^{n} (y_k - \overline{y})^2}}$$

Examples:

1. The data from Problem 1 provide a strong linear relationship. We already computed $u \bullet v = 0.9931$ which is extremely close to 1.

2. The data in Problem 2 are not linearly related. We get

x_k	$x_k - 2.5$	$(x_k - 2.5)^2$	y_k	$y_k - 4.2$	$(y_k - 4.2)^2$	s_{xy}
0	-2.5	6.25	0	-4.2	17.64	10.5
1	-1.5	2.25	5	0.8	0.64	-1.2
2.5	0.0	0.0	8	3.8	14.44	0.0
4	1.5	2.25	6	1.8	3.24	2.7
5	2.5	6.25	2	-2.2	4.84	-5.5
12.5	0	17	21	0	40.80	6.5

where the last column contains the products $s_{xy} = (x_k - 2.5)(y_k - 4.2)$. Hence,

$$\rho \;=\; \frac{6.5}{\sqrt{17 \times 40.8}} \;=\; 0.2468$$

There is no evidence supporting a linear relationship, since ρ is not close to ± 1. ◇

5 MATLAB

5.1 Backslash

In matlab we have two options to solve least squares problems. For one, the backslash operator automatically applies least squares techniques when the system is over-determined.

Problem 3.1: Solve Problem 3 in matlab with the backslash operator.

```
>> A = [2 1; 1 1; 0 1]; b = [8 4 2]';
>> x = A\b
x =
    3.0000
    1.6667
>>
```

5.2 Polyfit

A second option is the `polyfit` command, which applies, as the name indicates, when data are to be approximated by polynomials. The input requires the data set x and y, as well as the degree of the polynomial, the output are the coefficients of the approximating polynomial.

Problem 1.3: Solve problem 1 with the backslash operator and with `polyfit`.

```
>> x = (0:4)'; y = [2.1 2.5 2.9 3.6 3.9]';
>> A = [ones(size(x)) x];
>> A\y
ans =
    2.0600
    0.4700
>>
>> polyfit(x,y,1)
ans =
    0.4700    2.0600
>>
```

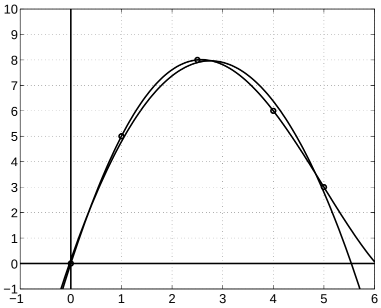
Problem 2.2: Solve problem 2 with the backslash operator and with `polyfit` using polynomials of degree 1, 2 and 4.

```
>> x = [0 1 2.5 4 5]’; y = [0 5 8 6 2]’;
>> A = [ones(size(x)) x x.^2]
A =
    1.0000         0         0
    1.0000    1.0000    1.0000
    1.0000    2.5000    6.2500
    1.0000    4.0000   16.0000
    1.0000    5.0000   25.0000
>> A\y
ans =
    0.0478
    5.9899
   -1.1215
>> % degree 1:
>> polyfit(x,y,1)
ans =
    0.3824    3.2441
>> % degree 2:
>> polyfit(x,y,2)
ans =
   -1.1215    5.9899    0.0478
>> % degree 4:
>> polyfit(x,y,4)
ans =
   -0.0022    0.0389   -1.3144    6.2778    0.0000
```

We can always fit five datapoints by a polynomial of the degree four. The last matlab result indicates that the interpolating polynomial is

$$y = -0.0022x^4 + 0.0389x^3 - 1.3144x^2 + 6.2778x + 0.0000$$

and by the size of the coefficients we can conjecture that a quadratic approximation is appropriate. We conclude with a plot of the approximating polynomial (degree 2) and the interpolating polynomial in a common graph.


10
9
8
7
6
5
4
3
2
1
0
−1
−1
0
1
2
3
4
5
6

◇