11.1 Method of Least Squares

Suppose we have a collection of observations $(x_1, y_1), \ldots, (x_n, y_n)$. Our goal is to find the equation of the line $y = \beta_0 + \beta_1 x$ which best "fits" the data.

Define residuals $r_1, \ldots, r_n$ as

$$r_i = y_i - (\beta_0 + \beta_1 x_i)$$

Intuitively, we might suggest that the *best* line is one for which the net residuals are zero

$$\sum_{i=1}^{n} r_i = \sum_{i=1}^{n} y_i - (\beta_0 + \beta_1 x_i) = 0$$

and which minimizes the total distance between the line and the data:

$$\sum_{i=1}^{n} d(y_i, \beta_0 - \beta_1 x_i) = \sum_{i=1}^{n} d(r_i, 0)$$

where $d(x, y)$ is some distance function. Both $d(x, y) = |y - x|$ and $d(x, y) = (y - x)^2$ are popular choices.

The first condition is equivalent to requiring the line to pass through the point $(\bar{x}, \bar{y})$:

$$0 = \sum_{i=1}^{n} y_i - (\beta_0 + \beta_1 x_i) = \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = n \left[\frac{1}{n} \sum_{i=1}^{n} y_i - \beta_0 - \beta_1 \frac{1}{n} \sum_{i=1}^{n} x_i \right] = n \left[\bar{y} - \beta_0 - \beta_1 \bar{x} \right]$$

Dividing both sides by n and solving for $\bar{y}$ gives the result.

Least Absolute Line

Now, consider the optimization problem with $d(x, y) = |y - x|$. We seek β_0, β_1 so that

$$\sum_{i=1}^{n} |y_i - (\beta_0 + \beta_1 x_i)|$$

is minimal. But as the line must pass through $(\bar{x}, \bar{y})$, it suffices to solve the following simpler optimization problem (which corresponds to translating our coordinate system to put the origin at $(\bar{x}, \bar{y})$):

$$e(\beta_1) = \sum_{i=1}^{n} |(y - \bar{y}) - \beta_1 (x_i - \bar{x})|$$

Observe that the function $e(\beta_1)$ is piecewise linear in β_1, with the contribution from each observation y_i decreasing linearly as β_1 increases until $\beta_1 = \dfrac{y_i - \bar{y}}{x_i - \bar{x}}$, and then increases linearly thereafter.

This optimization problem can be meticulously solved by hand for small data sets, but for larger data sets, it is far more computationally challenging, and are an example of linear programming problems. The general algorithm for solving these problems (called the simplex algorithm) was discovered in the 1950s by George Danzig.

Least Squares Line

As an alternative, we can consider the optimization problem with $d(x, y) = (y - x)^2$. To minimize

$$e(\beta_0, \beta_1) = \sum r_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i)^2$$

we differentiate with respect to β_0 and β_1:

$$\frac{\partial e}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i)) \qquad \frac{\partial e}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))x_i$$

Setting the partials equal to 0 gives

$$\beta_0 n + \beta_1 \sum x_i = \sum y_i \qquad \beta_0 \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i$$

which are called the *normal* equations for β_0 and β_1 (Normal in this case means perpendicular, and arises for a reason discussed at the end of these notes). Dividing both sides in the first equation by n shows that

$$\beta_0 + \beta_1 \bar{x} = \bar{y}$$

That is, the line that minimizes the squared residuals must pass through the point $(\bar{x}, \bar{y})$.

Suppose now that $\bar{x}, \bar{y}$ are both 0. Then the second equation is equivalent to

$$\beta_1 = \frac{\sum x_i y_i}{\sum x_i^2}.$$

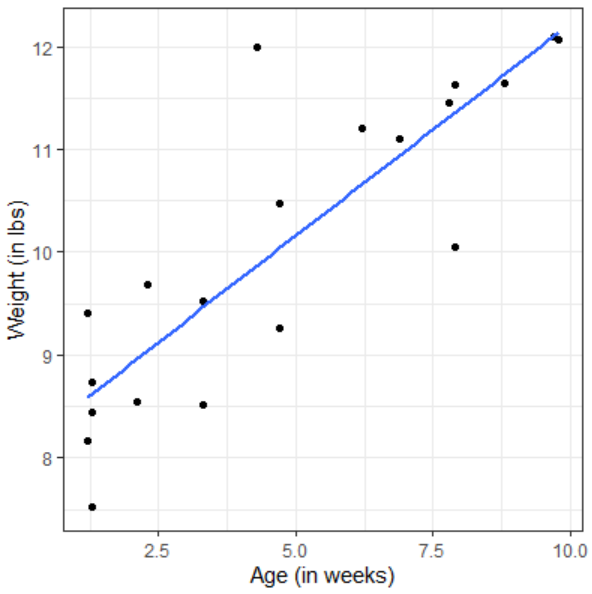
More generally, if $\bar{x}, \bar{y}$ are not 0, consider the collection of points $x_i' = x_i - \bar{x}$ and $y_i' = y_i - \bar{y}$. The line minimizing the squared sum of residuals must have the same slope for both collection of points (since translating a line preserves slope). But since $\bar{x}' = 0$ and $\bar{y}' = 0$, then

$$\beta_1 = \frac{\sum x_i' y_i'}{\sum (x_i')^2} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}.$$

Thm: Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a collection of n points. The line that minimizes the sum of squared residuals is given by $y = \beta_0 + \beta_1 x$ with

$$\beta_1 = \frac{\text{cov}(x, y)}{\text{var}(x)} \quad \beta_0 = \bar{y} - \beta_1 \bar{x}.$$

Ex 1: Consider the following scatterplot showing the age x and weight y for 20 newborn babies:



	age	weight
1	4.7	9.26
2	7.9	11.63
3	7.9	10.05
4	7.8	11.45
5	6.2	11.21
6	1.3	8.74
7	9.8	12.07
8	1.2	8.17
9	3.3	8.51
10	8.8	11.64
11	2.3	9.69
12	3.3	9.53
13	6.9	11.11
14	1.3	7.52
15	4.3	12
16	1.2	9.41
17	1.3	8.44
18	9.7	12.1
19	4.7	10.48
20	2.1	8.55

The least squares line is obtained by computing the means of x and y, along with the variance of x, and the covariance of x and y.

$$\text{mean}(x) = 4.8 \quad \text{mean}(y) = 10$$
$$\text{var}(x) = 9.5 \quad \text{cov}(x, y) = 3.9$$

This gives the equation

$$\beta_1 = \frac{\text{cov}(x, y)}{\text{var}(x)} = \frac{3.9}{9.5} = 0.41$$
$$\beta_0 = \bar{y} - \beta_1 \bar{x} = 10 - 0.41 \cdot 4.8 = 8$$

Least squares in multiple variables

Suppose now that we have collection of points $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$ where $\mathbf{x}_i \in \mathbb{R}^{k+1}$, where the 0th entry of each $\mathbf{x}_i$ is 1. Let x_{ij} denote the ith coordinate of $\mathbf{x}_j$.

For example, suppose we collect measurements on k different attributes for n different people in a population. We can record this information in the $n \times (k+1)$ matrix $\mathbf{X}$ given by

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ & \vdots & & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

The matrix $\mathbf{X}$ is called the **design matrix**. The columns of $\mathbf{X}$ represent observations of a single variable, while the rows represent the observations of a single individual. We wish to find linear surface in $\mathbb{R}^{k+1}$ of best fit. *Note that when $k = 1$, this is identical to the problem of finding the least squares line.* This linear function will have the form

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \quad \text{for } \beta_i \in \mathbb{R}.$$

Let $\boldsymbol{\beta}^T = \begin{pmatrix} \beta_0 & \beta_1 & \cdots & \beta_k \end{pmatrix}$. Note that in matrix notation, we can represent the output of the linear equation at inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n$ as

$$\mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ & \vdots & & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} \end{pmatrix}$$

As before, we proceed by finding coefficients β_i which minimize the sum of squared residuals:

$$\sum r_i^2 = \sum \Big(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) \Big)^2 = \sum (y_i - (\mathbf{X}\boldsymbol{\beta})_i)^2$$

To do so, we now differentiate with respect to $\beta_0, \beta_1, \ldots, \beta_k$, set equal to 0, and solve:

$$\begin{aligned} \beta_0 n + \beta_1 \sum_{i=1}^n x_{i1} + \cdots + \beta_k \sum_{i=1}^n x_{ik} &= \sum_{i=1}^n y_i \\ \beta_0 \sum_{i=1}^n x_{i1} + \beta_1 \sum_{i=1}^n x_{i1}x_{i1} + \cdots + \beta_k \sum_{i=1}^n x_{i1}x_{ik} &= \sum_{i=1}^n x_{i1} y_i \\ &\vdots \\ \beta_0 \sum_{i=1}^n x_{ik} + \beta_1 \sum_{i=1}^n x_{ik}x_{i1} + \cdots + \beta_k \sum_{i=1}^n x_{ik}x_{ik} &= \sum_{i=1}^n x_{ik} y_i \end{aligned}$$

Or in matrix notation:

$$(\mathbf{X}\mathbf{X}^T)\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}$$

Solving for $\boldsymbol{\beta}$ gives

$$\boldsymbol{\beta} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}^T\mathbf{y}$$

Since we make no assumptions about the relationships among the variables, the above framework allows us to find the least squares polynomial for data, by using variables

$$x_0 = 1 \quad x_1 = x \quad x_2 = x^2 \quad \cdots \quad x_k = x^k$$

In this case, we find a polynomial of the form

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \ldots \beta_k x^k$$

for which the sum of squared residuals are minimal.

Alternate Proof for Normal Equations using Linear Algebra

Let $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{y} = (y_1, \ldots, y_n)$ and $\mathbf{1} = (1, \ldots, 1)$ be vectors in $\mathbb{R}^n$. Let $V = \text{span}\{\mathbf{1}, \mathbf{x}\}$, and in particular, note that any element $\mathbf{v} \in V$ can be written as

$$\mathbf{v} = \beta_0\mathbf{1} + \beta_1\mathbf{x} = (\beta_0 + \beta_1 x_1, \ldots, \beta_0 + \beta_1 x_n)$$

In particular, using $\{\mathbf{1}, \mathbf{x}\}$ as a basis, we can identity vectors in V with lines in $\mathbb{R}^2$, where β_0 and β_1 are the coefficients on the line $y = \beta_0 + \beta_1 x$.

Now, consider the problem of finding the point $\hat{\mathbf{y}} = (\beta_0, \beta_1)$ in V that is closest to another point $\mathbf{y}$. That is, we want to find $\hat{\mathbf{y}}$ minimizing $\|\mathbf{y} - \hat{\mathbf{y}}\|^2$. But note that the displacement vector can be written as

$$\mathbf{y} - \hat{\mathbf{y}} = (y_1 - (\beta_0 + \beta_1 x_1), \ldots, y_n - (\beta_0 + \beta_1 x_n)) \quad \text{with} \quad \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x))^2$$

So $\hat{\mathbf{y}}$ is exactly the solution to the previous optimization problem.

But to find $\hat{\mathbf{y}}$ we simply need to project $\mathbf{y}$ onto V. To do so, we first need an orthogonal basis for V. Observe that the projection of $\mathbf{x}$ onto $\text{span}\{\mathbf{1}\}$ is

$$\frac{\langle \mathbf{x}, \mathbf{1} \rangle}{\|\mathbf{1}\|^2}\mathbf{1} = \frac{1}{n}\sum x_i \mathbf{1} = \bar{x}\mathbf{1}$$

which means $\mathbf{1}$ and $\mathbf{x} - \bar{x}\mathbf{1}$ are orthogonal.

The projection of $\mathbf{y}$ onto V is then

$$\mathbf{y} = \frac{\langle \mathbf{y}, \mathbf{x} - \bar{x}\mathbf{1} \rangle}{\|\mathbf{x} - \bar{x}\mathbf{1}\|^2}(\mathbf{x} - \bar{x}\mathbf{1}) + \frac{\langle \mathbf{y}, \mathbf{1} \rangle}{\|\mathbf{1}\|^2}\mathbf{1}.$$

To simplify, observe that $\mathbf{1}$ and $\mathbf{x} - \bar{x}\mathbf{1}$ are orthogonal, and so $\langle \bar{y}\mathbf{1}, \mathbf{x} - \bar{x}\mathbf{1} \rangle = 0$. Additionally, note that $\|\mathbf{1}\|^2 = n$ and $\langle \mathbf{y}, \mathbf{1} \rangle = \sum y_i$. Hence

$$\mathbf{y} = \frac{\langle \mathbf{y} - \bar{y}\mathbf{1}, \mathbf{x} - \bar{x}\mathbf{1} \rangle}{\|\mathbf{x} - \bar{x}\mathbf{1}\|^2}(\mathbf{x} - \bar{x}\mathbf{1}) + \bar{y}\mathbf{1}$$

Letting

$$\beta_1 = \frac{\langle \mathbf{y} - \bar{y}\mathbf{1}, \mathbf{x} - \bar{x}\mathbf{1} \rangle}{\|\mathbf{x} - \bar{x}\mathbf{1}\|^2} \quad \text{and } \beta_0 = \bar{y} - \beta_1 \bar{x}$$

Then

$$\mathbf{y} = \beta_1\mathbf{x} - \beta_1\bar{x}\mathbf{1} + \bar{y}\mathbf{1} = \beta_1\mathbf{x} + \beta_0\mathbf{1}$$

Finally, note that

$$\beta_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}.$$

Least Squares in Multiple Variables

Let $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k) \in \mathbb{R}^{k+1}$ and let $\mathbf{X}$ denote the $n \times (k+1)$ matrix whose columns are $\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_k$. Note that the squared sum of residuals can then be expressed as

$$\sum r_i^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

Minimizing the squared sum of residuals corresponds to projecting $\mathbf{y}$ onto the span of the columns of $\mathbf{X}$.

We won't go through the details right now (they can be found in your MAT 215 text), but the coordinates of the projection are given by

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

As a quick exercise, you should verify that this formula is consistent with the one previously derived in the case when $k = 1$.