Linear Regression and Multilinear Regression

1 Linear Regression

For this section, assume that we are given data points in the form (x_i, y_i) for $i = 1, \ldots, N$ and that we desire to fit a line to these data points. So, we need to find the slope and y–intercept that make a line fit these data "best." The approach used in linear regression is to minimize the sum of the squares of the differences between the data and a line; that is, find the values of a and b that minimize the sum

$$R^2 = \sum_{i=1}^{N} (a x_i + b - y_i)^2. \tag{1}$$

Minimizing this sum is called *least squares minimization.*

The minimum of the sum above will occur at a critical point. By the first derivative test from calculus, we can find critical points by solving the system

$$\begin{cases} \dfrac{\partial R^2}{\partial a} = 0 \\[2ex] \dfrac{\partial R^2}{\partial b} = 0 \end{cases} \tag{2}$$

Specifically, finding the partial derivatives and rewriting gives

$$\begin{cases} a \sum x_i^2 + b \sum x_i = \sum x_i y_i \\[2ex] a \sum x_i + N b = \sum y_i \end{cases} \tag{3}$$

Since the values $\sum x_i^2, \sum x_i, \sum x_i y_i, \sum y_i$ can be computed, the above is a system of 2 equations in 2 unknowns. These equations are known as the *normal equations.*

For example, consider performing linear regression on the data points

$$(1, 10.1), (2, 10.4), (3, 10.9), (4, 10.8), (5, 11.0)$$

Then, the normal equations in Eq. (3) become

$$\begin{cases} 55\,a + 15\,b = 161.8 \\[2ex] 15\,a + 5\,b = 53.2 \end{cases}$$

Solving this system gives $a = 0.22$, $b = 9.98$. So, the regression line (the line of "best fit") for the above data is $y = 0.22x + 9.98$.

The normal equations in Eq. (3) can be solved to obtain general expressions for a and b. These are the nasty formulas that are often taught in traditional, elementary statistics courses. The formulas are:

$$
\begin{aligned}
a &= \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{N \sum x_i^2 - \left(\sum x_i\right)^2} \\
b &= \frac{N \sum x_i y_i - \sum x_i \sum y_i}{N \sum x_i^2 - \left(\sum x_i\right)^2}
\end{aligned}
$$

Using matrices makes dealing with the normal equations a little easier. For example, we can write the normal equations in Eq. (3) as

$$
\begin{pmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & N \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum x_i y_i \\ \sum y_i \end{pmatrix}
$$

To solve this system, we must invert the 2-by-2 coefficient matrix and multiply it on both sides to get

$$
\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & N \end{pmatrix}^{-1} \begin{pmatrix} \sum x_i y_i \\ \sum y_i \end{pmatrix}
$$

So, the real work in solving this system is in inverting (if possible) the coefficient matrix.

Most of the applications of linear regression involve a large number of data points, and hence, a simple way of computing the coefficient matrix is useful. The standard approach is to create the coefficient matrix from another matrix which is defined using the data as follows.

$$
\mathrm{A} = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{pmatrix}
$$

Then, we can write

$$
\mathrm{A}^{\mathrm{t}}\mathrm{A} = \begin{pmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & N \end{pmatrix}, \qquad \mathrm{A}^{\mathrm{t}}\,\mathbf{Y} = \begin{pmatrix} \sum x_i y_i \\ \sum y_i \end{pmatrix}.
$$

(Be sure to check by hand that this is true.) Thus, the normal equations can be written

$$
\mathrm{A}^{\mathrm{t}}\mathrm{A}\,\mathbf{W} = \mathrm{A}^{\mathrm{t}}\,\mathbf{Y} \tag{4}
$$

with

$$
\mathbf{W} = \begin{pmatrix} a \\ b \end{pmatrix}, \qquad \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{pmatrix}.
$$

The solution to this matrix equation is then

$$\mathbf{W} = (\mathrm{A}^{\mathrm{t}}\mathrm{A})^{-1}\mathrm{A}^{\mathrm{t}}\,\mathbf{Y}, \tag{5}$$

if $(\mathrm{A}^{\mathrm{t}}\mathrm{A})^{-1}$ exists.

Although the definition for A appears to be arbitrarily chosen, there is a good reason behind it. Remember that for linear regression, we want to find a, b so that we can reproduce y_i from x_i for each data point. Specifically, we want

$$\begin{aligned} ax_1 + b &= y_1 \\ ax_2 + b &= y_2 \\ ax_3 + b &= y_3 \\ &\vdots \\ ax_N + b &= y_N \end{aligned}$$

This system of equations can also be written using our definition of A, $\mathbf{W}$, and $\mathbf{Y}$ as

$$\mathrm{A}\mathbf{W} = \mathbf{Y}.$$

To solve this system, we would like to multiply by A^{-1} on both sides. BUT, A is not a *square* matrix and cannot be inverted. So, we need to rewrite the system so that we have a square matrix involved. To do this, multiply both sides of the matrix equation by A^{t} to obtain the matrix version of the normal equations

$$\mathrm{A}^{\mathrm{t}}\mathrm{A}\,\mathbf{W} = \mathrm{A}^{\mathrm{t}}\mathbf{Y}.$$

So, the definition of A "makes sense" and works!

2 Multilinear Regression

With multilinear regression, we assume that the dependent data, y_i, depends linearly on several independent variables, $x_1, x_2, \ldots, x_k$. For the purposes of this discussion, assume that the given data depends only on two independent variables. So, data points are of the form

$$(x_{11}, x_{21}, y_1), (x_{12}, x_{22}, y_2), (x_{13}, x_{23}, y_3), \ldots, (x_{1N}, x_{2N}, y_N)$$

The goal is to minimize the sum

$$R^2 = \sum_{i=1}^{N} (a_1 x_{1i} + a_2 x_{2i} + b - y_i)^2. \tag{6}$$

Ideally, we want to find a_0, a_1, a_2 so that

$$\begin{aligned} a_1 x_{11} + a_2 x_{21} + b &= y_1 \\ a_1 x_{12} + a_2 x_{22} + b &= y_2 \\ a_1 x_{13} + a_2 x_{23} + b &= y_3 \\ &\vdots \\ a_1 x_{1N} + a_2 x_{2N} + b &= y_N \end{aligned}$$

Rewrite this system as

$$\mathrm{A}\mathbf{W} = \mathbf{Y},$$

where

$$\mathrm{A} = \begin{pmatrix} x_{11} & x_{21} & 1 \\ x_{12} & x_{22} & 1 \\ x_{13} & x_{23} & 1 \\ \vdots & \vdots & \vdots \\ x_{1N} & x_{2N} & 1 \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} a_1 \\ a_2 \\ b \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{pmatrix}.$$

To solve this system, we would like to multiply by A^{-1} on both sides. BUT, A is not a *square* matrix and cannot be inverted. So, we need to rewrite the system so that we have a square matrix involved. To do this, multiply both sides of the matrix equation by A^{t} to obtain

$$\mathrm{A}^{\mathrm{t}}\mathrm{A}\,\mathbf{W} = \mathrm{A}^{\mathrm{t}}\mathbf{Y}.$$

Note that this equation is the same as Eq. (4). The difference lies only in how the coefficient matrix A is created. Indeed, if you take the partial derivatives $\partial R^2/\partial a_1$, $\partial R^2/\partial a_2$, and $\partial R^2/\partial b$ as we did in linear regression, you will find that the equation above represents the normal equations in multilinear regression.

Now, $\mathrm{A}^{\mathrm{t}}\mathrm{A}$ is a square matrix. We can multiply by its inverse on both sides of our system, if it exists. Thus, the solution for multilinear regression is

$$\mathbf{W} = (\mathrm{A}^{\mathrm{t}}\mathrm{A})^{-1}\mathrm{A}^{\mathrm{t}}\,\mathbf{Y}, \tag{7}$$

if $(\mathrm{A}^{\mathrm{t}}\mathrm{A})^{-1}$ exists. The fact that the matrix equations for linear and multilinear regression appear the same make this matrix approach very appealing. Also, there are lots of numerical methods for finding $(\mathrm{A}^{\mathrm{t}}\mathrm{A})^{-1}$ accurately.

Finally, consider the data set below as an example.

$$\begin{gathered} (0, 0.30, 10.14),(0.69, 0.60, 11.93), (1.10, 0.90, 13.57) \\ (1.39, 1.20, 14.17),(1.61, 1.50, 15.25), (1.79, 1.80, 16.15) \end{gathered}$$

Then,

$$\mathrm{A} = \begin{pmatrix} 0 & 0.30 & 1 \\ 0.69 & 0.60 & 1 \\ 1.10 & 0.90 & 1 \\ 1.39 & 1.20 & 1 \\ 1.61 & 1.50 & 1 \\ 1.79 & 1.80 & 1 \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} 10.14 \\ 11.93 \\ 13.57 \\ 14.17 \\ 15.25 \\ 16.15 \end{pmatrix}$$

and so

$$\mathrm{A}^{\mathrm{t}}\mathrm{A} = \begin{pmatrix} 9.41 & 8.71 & 6.58 \\ 8.71 & 8.19 & 6.30 \\ 6.58 & 6.30 & 6.00 \end{pmatrix}.$$

Then,

$$\mathbf{W} = \begin{pmatrix} a_1 \\ a_2 \\ b \end{pmatrix} = (\mathrm{A}^{\mathrm{t}}\mathrm{A})^{-1}\mathrm{A}^{\mathrm{t}}\,\mathbf{Y} = \begin{pmatrix} 2.09 \\ 1.50 \\ 9.69 \end{pmatrix}.$$

Thus, the line that best fits the data is $y = 2.09x_1 + 1.50x_2 + 9.69$.