Chapter 1

The Noble Eightfold Path to Linear Regression

In this chapter, I show several different ways of solving the linear regression problem. The different approaches are interesting in their own way. Each one reveals something different about the properties of a linear regression fitted to a data set. Given p independent variables $x_1, x_2, \ldots, x_p$ we can perform N independent experiments, for different combinations of values of the p variables, measuring a scalar result y every time. The variables $x_1, x_2, \ldots, x_p$ could represent, for example, the results of p laboratory tests, while y could be a classification ($+1$ for a person who is sick, -1 for a healthy person). The value y could be also a number within a certain range. If we call x_i the vector of p measurements $x_i1, x_i2, \ldots, x_ip$, and y_i the associated result, for $i = 1, \ldots, N$, we call the data set of N pairs (x_i, y_i) the *training set* for a function approximation or for a classifier.
What we would like to do, is to express each y_i as a linear combination of the p independent variables, with a minimum of approximation error. That is, we would like to express each y_i as:

$$\begin{aligned} y_1 &= \phi_0 + \phi_1 x_{11} + \phi_2 x_{11} + \cdots + \phi_p x_{1p} + e_1 \\ y_2 &= \phi_0 + \phi_1 x_{21} + \phi_2 x_{22} + \cdots + \phi_p x_{2p} + e_2 \\ &\vdots \\ y_N &= \phi_0 + \phi_1 x_{N1} + \phi_2 x_{N2} + \cdots + \phi_p x_{Np} + e_N \end{aligned}$$

where the values ϕ_i, for $i = 0, \ldots, p$, are the coefficients of the linear approximation we want to use. The variables e_i represent the approximation error (or residual) for each approximation of y_i.

In matrix terms, what we want to do is find a vector of coefficients Φ such that

$$y = X\Phi + e$$

where $y^{\mathrm{T}} = (y_1, y_2, \ldots, y_N)$ is the vector of the N results and the matrix X contains in each row the components of the x_i vectors, including the constant 1 in the first column:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & & x_{2p} \\ \vdots & & & & \\ 1 & x_{N1} & x_{N2} & & x_{Np} \end{pmatrix}$$

The vector $\Phi^{\mathrm{T}} = (\phi_0, \phi_1, \ldots, \phi_p)$ contains the coefficients for the linear approximation as components, while the vector $e^{\mathrm{T}} = (e_1, e_2, \ldots, e_N)$ contains the individual errors for each linear prediction.
We want to find the best fit in the sense that $e^{\mathrm{T}}e$, the sum of squared residuals, should be minimal. Therefore, we define the error function E as

$$\begin{aligned} E = e^T e &= (Y - X\Phi)^{\mathrm{T}}(y - X\Phi) \\ &= \sum^{N} (y_i - \phi_0 - \phi_1 x_{i1} - \cdots - \phi_p x_{ip})^2 \end{aligned}$$

Our problem now is finding Φ which minimizes the error E.

1.1 Solution by taking partial derivatives

We can find the minimum of E by computing all partial derivatives, relative to $\phi_0, \phi_1, \ldots, \phi_p$, and setting the results to zero. In that case:

$$\begin{aligned} \frac{\partial E}{\partial \phi_0} &= -2 \sum^{N} (y_i - \phi_0 - \phi_1 x_{i1} - \cdots - \phi_p x_{ip}) = 0 \\ \frac{\partial E}{\partial \phi_1} &= -2 \sum (y_i - \phi_0 - \phi_1 x_{i1} - \cdots - \phi_p x_{ip}) x_{i1} = 0 \\ &\vdots \\ \frac{\partial E}{\partial \phi_p} &= -2 \sum (y_i - \phi_0 - \phi_1 x_{i1} - \cdots - \phi_p x_{ip}) x_{ip} = 0 \end{aligned}$$

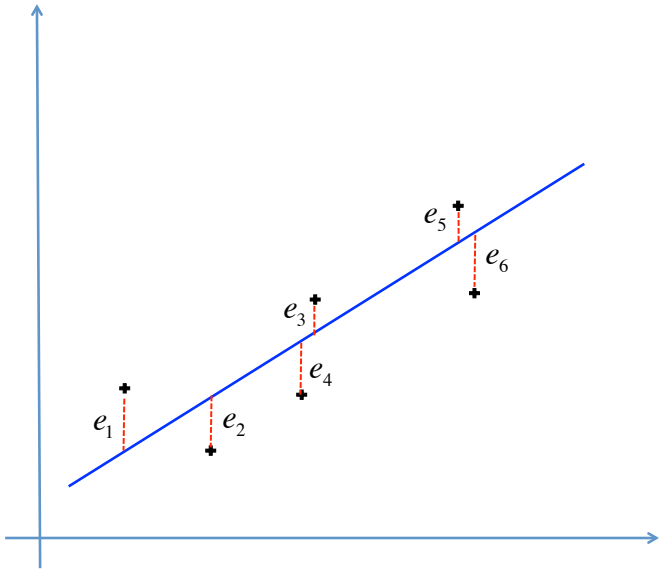

Figure 1.1: Experimental points and the regression error

Simplifying, we obtain a set of $p+1$ linear equations for $p+1$ variables:

$$\begin{aligned}
\sum^{N} y_i &= \phi_0 \sum^{N} 1 + \phi_1 \sum^{N} x_{i1} + \cdots + \phi_p \sum^{N} x_{ip} \\
\sum^{N} y_i x_{i1} &= \phi_0 \sum^{N} x_{i1} + \phi_1 \sum^{N} x_{i1}^2 + \cdots + \phi_p \sum^{N} x_{ip} x_{i1} \\
\sum^{N} y_i x_{ip} &= \phi_0 \sum^{N} x_{ip} + \phi_1 \sum^{N} x_{i1} x_{ip} + \cdots + \phi_p \sum^{N} x_{ip}^2
\end{aligned}$$

This $p+1$ equations correspond to the matrix expression:

$$X^{\mathrm{T}} y = X^{\mathrm{T}} X \Phi \tag{1.1}$$

If $(X^{\mathrm{T}}X)^{-1}$ exists, the solution for Φ is

$$\Phi = (X^{\mathrm{T}}X)^{-1} X^{\mathrm{T}} y$$

The vector Φ minimizes the sum of squared residuals. In the case of just one variable x_1 and two coefficients ϕ and ϕ_1, what we are looking for is the regression line which best approximates the data points as shown in Figure 1.1.

The residuals are measured vertically. The sum of squared distances to the regression line should be a minimum.
Notice that Eq. (1.1) can be rewritten as

$$X^{\mathrm{T}}(y - X\Phi) = X^{\mathrm{T}}e = 0$$

Due to the form of the matrix X, this means that $\sum^N e_i = 0$, $\sum^N e_i x_{i1} = 0, \ldots, \sum^N e_i x_{ip} = 0$, that this the positive residuals always balance the negative residuals, also in the case that the residuals are weighted by the values of the components of the first, second, etc. variables.

1.2 Solution by computing vector derivatives

There is a very convenient shorthand when we want to compute several partial derivatives of the same function. If we are computing the partial derivatives of the function $f(\alpha_1, \alpha_2, \ldots, \alpha_n)$ relative to its vector of variables, which we call α, we write

$$\frac{df}{d\alpha} = \left(\frac{\partial f}{\partial \alpha_1}, \frac{\partial f}{\partial \alpha_2}, \ldots, \frac{\partial f}{\partial \alpha_n}\right)^{\mathrm{T}}$$

The first partial derivative always occupies the first row of the result, the second derivative, the second row, and so on. If we are deriving a vector of functions $f^{\mathrm{T}} = (f_1, f_2, \ldots, f_m)$, the derivative with respect to the vector α is

$$\frac{df^{\mathrm{T}}}{d\alpha} = \left[\frac{\partial f_i}{\partial \alpha_j}\right]_{ij}$$

where the notation in square brackets refers to a matrix, where each component ij has the form shown inside the brackets.
Using this notation, most of the standard rules for the derivation of functions continue to apply, but two special cases are important for us. First

$$\frac{da^{\mathrm{T}}}{da} = \left[\frac{\partial \alpha_i}{\partial \alpha_j}\right]_{ij} = I$$

where a is a n-dimensional column vector of variables and I represents the $n \times n$ identity matrix. Also

$$\frac{d}{dx}(x^{\mathrm{T}}Ax) = Ax + A^{\mathrm{T}}x$$

where x is an n-dimensional vector of variables and A an $n \times n$ constant matrix. If A is symmetrical

$$\frac{d}{dx}(x^{\mathrm{T}}Ax) = 2Ax$$

We now come back to the regression problem. We want to minimize

$$\begin{aligned} E &= (y - X\Phi)^{\mathrm{T}}(y - X\Phi) \\ &= y^{\mathrm{T}}y - y^{\mathrm{T}}X\Phi - \Phi^{\mathrm{T}}X^{\mathrm{T}}y + \Phi^{\mathrm{T}}X^{\mathrm{T}}X\Phi \end{aligned}$$

We compute

$$\frac{dE}{d\Phi} = 0 - 2X^{\mathrm{T}}y + 2X^{\mathrm{T}}X\Phi = 0$$

where we took advantage of the fact that $y^{\mathrm{T}}y$ is constant, and $y^{\mathrm{T}}X\Phi = \Phi^{\mathrm{T}}X^{\mathrm{T}}y$, since this is an scalar. Therefore,

$$X^{\mathrm{T}}y = X^{\mathrm{T}}X\Phi$$

and from this, we find, as before, that

$$\Phi = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$$

in case that $(X^{\mathrm{T}}X)^{-1}$ exists. We obtain the same solution faster, if we are acquainted with the derivation rules for vectors.

1.3 The path through the pseudoinverse

Remember: what we want to do is write y as

$$y = X\Phi + e$$

If X could be inverted, we could get an approximate value for Φ. But X is of dimension $N \times (p+1)$. Not being a square matrix, it does not have an inverse. However, any matrix has a pseudoinverse. The pseudoinverse of a matrix X is the unique matrix X^{+} for which these conditions hold:

1.) $XX + X = X$

2.) $X^{+}XX^{+} = X^{+}$

3.) XX^{+} and $X^{+}X$ are symmetric.

If the inverse of a matrix X exists, the pseudoinverse X^+ is just the inverse X^{-1}. If the inverse does not exist, the pseudoinverse is , in some sense, the best approximation. In case that $(X^{\mathrm{T}}X)^{-1}$ exists, it is easy to prove that $X^+ = (X^{\mathrm{T}}X)^{-1}X^T$ fulfills the three conditions above (just test them!).
We want to minimize

$$E = (y - X\Phi)^{\mathrm{T}}(y - X\Phi)$$

We will reduce this expression to now involving the pseudoinverse. We can rewrite E as

$$\begin{aligned} E &= (XX^+y - XX^+y + y - X\Phi)^{\mathrm{T}}(y - X\Phi) \\ &= (XX^+y - X\Phi)^{\mathrm{T}}(y - X\Phi) + y^{\mathrm{T}}(I - XX^+)(y - X\Phi) \\ &= (X^+y - \Phi)^{\mathrm{T}}X^{\mathrm{T}}(y - X\Phi) + y^{\mathrm{T}}(I - XX^+)y \end{aligned}$$

since $y^{\mathrm{T}}(I - XX^+)(-X\Phi) = -y^{\mathrm{T}}(X - XX^+X)\Phi = 0$ because, by property (1) above, $XX^+X = X$. In the last expression obtained above for E, the term $y^{\mathrm{T}}(I - XX^+)y$ is constant. If we want to minimize E, we can therefore just minimize

$$E' = (X^+y - \Phi)X^{\mathrm{T}}(y - X\Phi)$$

and obtain the same result for Φ. Now

$$E' = (X^+y - \Phi)^{\mathrm{T}}\left((XX^+X)^{\mathrm{T}}y - X^{\mathrm{T}}X\Phi\right)$$

since $X = XX^+X$. Therefore

$$E' = (X^+y - \Phi)^{\mathrm{T}}X^{\mathrm{T}}X(X^+y - \Phi)$$

since XX^+ is symmetrical. The above expression for E' corresponds to the product of the same twovectors (one is the transpose of the other). Therefore $E' \geq 0$, and the optimal solution is obtained when

$$\Phi = X^+y$$

If $X^+ = (X^TX)^{-1}X^T$, then $\Phi = (X^TX)^{-1}X^Ty$. Notice that this solution is more general than the previous ones. Even if $(X^TX)^{-1}$ does not exist, the optimal solution to the regression problem is $\Phi = X^+y$, since the pseudoinverse of a matrix X always exists.
Note that the term $y^T(I - XX^+)y$ which we disregard in the algebraic solution (because it is a constant), is equal to

$$y^T(y - XX^+y) = y^T(y - X\Phi) = y^Te$$

But

$$y^T e = (\Phi^T X^T + e^T)e = \Phi^T X^T e + e^T e = e^T e$$

since $X^T e = 0$. Therefore $E = e^T e$, as it should be, when $\Phi = X^+ y$.

1.4 The statistical approach

Assume that we have centered all data, that is $E(y) = 0$, and $E(x_j) = 0$ for all variables $x_1, x_2, \ldots, x_p$. We want to express y as

$$y = X\Phi + e$$

Multiplying by X^T

$$X^T y = X^T X\Phi + X^T e$$

Now, $X^T e$ represents the sum of residuals and weighted residuals. We require $X^T e$ to be a zero vector, because if $X^T e \neq 0$, then one component, for example the k-th, $\sum e_i x_{ik}$ is non-zero. But $\sum e_i x_{ik}$ is the covariance of $\{x_{ie}\}$ and $\{e_i\}$. If this covariance is non-zero, then part of the residuals can be explained by $\{x_{ik}\}$. That is, the vector e has a linear model proportional to the k-th row in the matrix X. Therefore Φ would not be optimal. For an optimal Φ we need $X^{\mathrm{T}} e = 0$. But then

$$X^{\mathrm{T}} y = (X^{\mathrm{T}} X)\Phi \Rightarrow \Phi = (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} y$$

1.5 The normal projection approach

The solution obtained in the last section can be obtained by looking again at the expression

$$y = X\Phi + e$$

If we think of the columns of the matrix X as vectors in N-dimensional space, what we want is to write the vector y as a linear combination of the column vectors of X. We want the error to be minimal. The error can be minimized when the distance to the linear combination defined by $X\Phi$ has minimum length. This is the case for the normal projection onto the subspace spannend by the column vectors of X. The difference between the normal projection $X\Phi$ and y is e. Since e has projection zero onto the subspace spanned by the columns of X, all scalar products with the columns of X must be zero. That is $X^{\mathrm{T}} e = 0$ and therefore $X^{\mathrm{T}}(y - X\Phi) = 0$, and from this we obtain the usual result $\Phi = (X^{\mathrm{T}} x)^{-1} X^{\mathrm{T}} y$.

Notice that this argument is very similar to the argument in the previous section. When we say that if the covariance between the vector e and one of the columns of X is nonzero, that means that e has a nonzero projection onto the subspace spanned by the columns of X. This can be corrected by modifying the parameter for that column of X.
The normal projection approach also helps explains better the concept of pseudoinverse. The pseudoinverse is the matrix that computes the normal projection of a vector y onto the subspace spanned by the columns of X. If the inverse exists, then we can express y as a linear combination of the columns of X and the error is zero. We do not have to project y, it lives in the subspace spanned by the columns of X. Its normal projection is y itself.

1.6 The physics solution

Now we can look at the linear regression problem from the viewpoint of physics. Given a set of points (x_i, y_i), for $i = 1, \ldots, N$, we would like to minimize the distance along the y-direction to a hyperplane. Think of the distance to that hyperplane as a "force" generated by small springs pulling on the hyperplane. If all forces are in equilibrium this should give us the best "fit" to the data set. In the case of springs whose pulling force is proportional to the distance e to the hyperplane, its potential energy is proportional to e^2. Minimizing the sum of squared distances (potential energy) is the same as bringing all forces into equilibrium.
Assume that we connect every point $(y, x_1, x_2, \ldots, x_p)$ in $\mathbb{R}^p$ to the regression hyperplane using springs, as shown in Fig. 1.2.

The springs pull the regression line towards them. The force is proportional to the residual e_i for the i-th point. The energy of such a system is proportional to $\sum e_i^2$ because the energy stored in a spring is proportional to the square of its stretch. In equilibrium the regression hyperplane does not move. The hyperplane has then reached its state of minimal energy (minimal sum of squared residuals, i.e., the minimal $e^T e$). But in equilibrium we should have

$$\sum e_i = 0$$

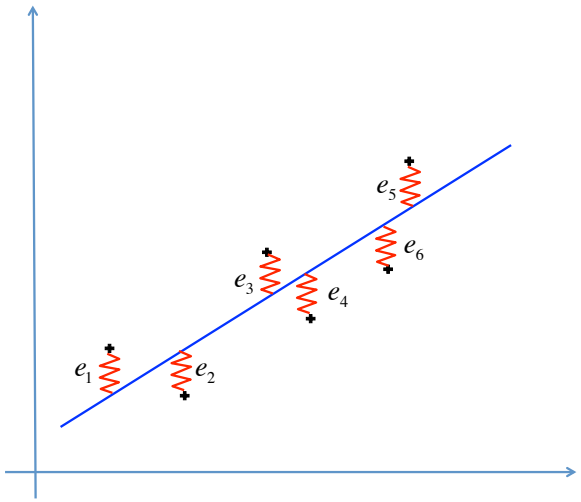

Figure 1.2: The regression error as a force acting on the regression line

since the sum of the translational forces must be zero. And for every variable j (that is for every spatial axis), we should have

$$\sum e_i x_{ij} = 0 \qquad \text{for} \quad j = 1, 2, \ldots, p,$$

since $\sum e_i x_{ij}$ represents the torque around the origin caused by the forces e_i acting on the j-th axis. Since the regression lines (hyperplane) are in equilibrium the total torque must be zero. Total torque and total force equal to zero means that

$$X^{\mathrm{T}} e = 0$$

but since $e = y - X\Phi$, then

$$\begin{aligned} & X^T y - X^T X \Phi = 0 \\ \Rightarrow \quad & \Phi = (X^T X)^{-1} X^T y \end{aligned}$$

This is probably the most intuitive derivation of the regression result, since we start directly from the equation $X^T e = 0$, which in other cases, for example, when taking partial derivatives, must first be proved.

1.7 Completing the square

The regression error function E is quadratic. We first expand it:

$$\begin{aligned} E &= (y - X\Phi)^{\mathrm{T}}(y - X\Phi) \\ &= \Phi^{\mathrm{T}}X^{\mathrm{T}}X\Phi - (y^{\mathrm{T}}X\Phi + \Phi^{\mathrm{T}}X^{\mathrm{T}}y) + y^{\mathrm{T}}y \end{aligned}$$

If we could rewrite E in the form $(\Phi - s)^{\mathrm{T}}X^{\mathrm{T}}X(\Phi - s)$, plus some constants, then, if $\Phi = s$, this would make the complete quadratic term zero and would minimize E. That is s would be the value of Φ we are looking for. Expanding $(\Phi - s)^{\mathrm{T}}X^{\mathrm{T}}X(\Phi - s)$, we obtain:

$$E = \Phi^{\mathrm{T}}X^{\mathrm{T}}X\Phi - (\underbrace{s^{\mathrm{T}}X^{\mathrm{T}}X\Phi}_{y^{\mathrm{T}}X\Phi} + \underbrace{\Phi^{\mathrm{T}}X^{\mathrm{T}}Xs}_{\Phi^{\mathrm{T}}X^{\mathrm{T}}y}) + s^{\mathrm{T}}X^{\mathrm{T}}Xs + \underbrace{c}_{y^{\mathrm{T}}y - s^{\mathrm{T}}X^{\mathrm{T}}Xs},$$

where c is a constant vector. This looks very similar to the first expansion of E computed above. The underbraces show the kind of term to term equalization that we have to make in order to have equivalent expressions for E.
If we identify the term $s^{\mathrm{T}}X^{\mathrm{T}}X\Phi$ with the term $y^{\mathrm{T}}X\Phi$, we see that we need to have $s^{\mathrm{T}}X^{\mathrm{T}}X = y^{\mathrm{T}}X$, and therefore we need $s^{\mathrm{T}} = y^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}$ in order to make both terms equal (assuming, as usual, that the inverse of $X^{\mathrm{T}}X$ exists). The third term in both equations is also made identical by this selection of the value of s. This means that we can rewrite E as

$$E = (\Phi - s)^{\mathrm{T}}X^{\mathrm{T}}X(\Phi - s) + y^{\mathrm{T}}y - s^{\mathrm{T}}X^{\mathrm{T}}Xs.$$

However, since the last two terms in the right-hand side are constant, this means that the minimum of the quadratic function E corresponds to $\Phi = s$, that is $\Phi = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$, as we have obtained through the previous six paths.
Notice that this derivation of the solution to the regression problem is even more general than just taking partial derivatives of E and setting the result to zero. In that case we still have to prove that we obtained a minimum (it is somewhat trivial, but it has to be done). Here we see immediately from the form of the quadratic function that we have a minimum. This corresponds to the usual approach when solving quadratic equations, where we "complete the square".

1.8 Fisher discriminant and linear regression

We said before that linear regression can be used to separate two classes, when the y values are bipolar (that is $+1$ or -1). The sign of the regression function

can then be used to classify a new data point as belonging to the positive or to the negative class. The Fisher linear discriminant is a direct attempt to find the best separation direction for two classes, and it has an interesting relationship with linear regression.

Assume that we want to separate two classes in a multidimensional space (with respective means μ_1 and μ_2, and covariance matrices Σ_1 and Σ_2) as well as possible, through a projection in one dimension. We are looking for a line such that the sum of the variances of two distributions is as low as possible, while the distance between the means is as high as possible. The expression for S given below tries to achieve this: it grows when the distance between the class means is large along the line u and when the sum of the variance of the projections of the two classes is low:

$$S(u) = \frac{|\mu_1 \cdot u - \mu_2 \cdot u|^2}{u^{\mathrm{T}}\Sigma_1 u + u^{\mathrm{T}}\Sigma_2 u}$$

We want to maximize S for the direction u. This is called the Fisher criterion. To maximize, we move the right-hand side denominator to the left-hand side of the expression. We obtain:

$$(u^{\mathrm{T}}\Sigma_1 u + u^{\mathrm{T}}\Sigma_2 u)S(u) = u^{\mathrm{T}}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^{\mathrm{T}}u$$

Differentiating

$$(2(\Sigma_1 + \Sigma_2)u)S(u) + (u^{\mathrm{T}}\Sigma_1 u + u^{\mathrm{T}}\Sigma_2 u)\frac{dS}{du} = 2(\mu_1 - \mu_2)(\mu_1 - \mu_2)^{\mathrm{T}}u$$

where we have used the fact the Σ_1 and Σ_2 are symmetric matrices. Since we are maximizing, we set dS/du equal to zero, and we obtain:

$$((\Sigma_1 + \Sigma_2)u)S(u_0) = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^{\mathrm{T}}u$$

For given u, $(\mu_1 - \mu_2)^{\mathrm{T}}u$ is an scalar γ, and therefore

$$((\Sigma_1 + \Sigma_2)u)S(u_0) = \gamma(\mu_1 - \mu_2)$$

and finally

$$u = \frac{\gamma}{S(u_0)}(\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2)$$

This direction of maximum separation (according to the Fisher criterion) is the celebrated Fisher linear discriminant. This approach allows us to project two multidimensional classes on a line and proceed finding a classifier for the simpler

one-dimensional problem. Projecting multidimensional data sets on lines is a usual approach when trying to reduce the dimensionality of a problem, as is done, for example, in random trees. The Fisher linear discriminant gives us a heuristic for a fast computation of a good projection direction.
However, notice that for two classes, if we collect all the data in one single matrix X, and if the number of points N in the positive class is the same as the number of points N in the negative class, then

$$\frac{1}{N} X^{\mathrm{T}} y = \mu_1 - \mu_2,$$

where y is a vector containing $+1$ for every vector in the positive class with mean μ_1 and -1 for every vector in the negative class with mean μ_2. The vector u, the Fisher discriminant direction, can then be written as:

$$u = \alpha(\Sigma_1 + \Sigma_2)^{-1} X^{\mathrm{T}} y.$$

where α is a proportionality constant.

1.9 Summary

In this chapter, I have shown that there are many equivalent ways of deriving the least-squares solution for the vector of regression coefficients. We can take partial derivatives (one by one, or using the shorthand notation of vector derivatives). We can solve the problem algebraically, using the pseudoinverse, or a statistical argument about the correlation of X and e. I showed that a physical model of the regression problem allows us to jump directly to the main algebraic expression which we need, namely $X^T e = 0$, which represents a condition over the normal projection onto the subspace spanned by the columns of X. The normal projection of a vector is provided by computing the matrix-vector product with the pseudoinverse of X. Or we can rewrite the error function completing a square, and the solution to the regression problem immediately stands out. Finally, there is subtle relationship between linear regression and the Fisher linear discriminant.