Linear Algebra and Modeling

Joshua Cook

Back to Algebra I

The slope-intercept equation:

$$y = mx + b$$

Here, we are defining a relationship between two variables, x and y.

You should think of this as a **mapping** $x \mapsto y$ or the **function** $f(x) = y$.

We can use R to model a linear function of two variables.

The word "model" here is important. A mathematical function in a computer is by necessity a model. We typically think of mathematical functions as continuous. In between two points on a function curve, we can always find a third point.

Computationally i.e. in a computer, we can not do this. We *must* work with a finite set of values.

A System of Linear Equations

Consider the system of equations:

$$y = 3x - 2$$

$$y = -5x + 6$$

The Solution

The solution to this system of equations is a hyperplane in one less dimension than the two equations. Each equation is a line, a 1-D hyperplane. The hyperplane describing their intersection is a point a 0-D hyperplane.

Find the Solution Analytically

$$
\begin{aligned}
y &= 3x - 2 && (1)\\
y &= -5x + 6 && (2)\\
3x - 2 &= -5x + 6 && \text{(both are equal to } y\text{)}\\
8x &= 8 && (3)\\
x &= 1 && (4)\\
y &= 3 \cdot 1 - 2 = 1 && \text{(plug } x\text{=1 into the first equation)}\\
y &= -5 \cdot 1 + 6 = 1 && \text{(plug } x\text{=1 into the second equation)}
\end{aligned}
$$

Both equations yield $y = 1$ implying that our solution is $(1, 1)$.

The Linear Combination

The basis of linear algebra is the vector and the linear combination.

The Vector

A vector is a multi-dimensional element.

e.g. (1), $(1, 3, 5, 8)$, $(1, 2, \ldots, 50)$

As opposed to a scalar

e.g. $1, \sqrt{2}, e, \pi, 101010, 47$

Scalar Arithmetic

- add
- subtract
- multiply
- divide

Vector Arithmetic

Vector Arithmetic has two defined operations, vector addition and scalar multiplication. That is, provided that we have vectors of equal length, we can add two vectors together, and we can multiply a vector by a scalar value.

Vector Addition

Here, we add two vectors. This is done by adding corresponding elements of the vector. Note the result is a vector of the same shape, signifying that it is a member of the same vector space.

$$(-1, 2) + (3, 5) = (2, 7)$$

Note that all three of these vectors are elements of the two-dimensional real-valued vector space, $\mathbb{R}^2$. We use the symbol $\in$ to mean "an element of" so that we can write

$$(-1, 2), (3, 5), (2, 7) \in \mathbb{R}^2$$

which means that these three vectors are elements of $\mathbb{R}^2$.

Vector Addition in General

Generally, where vector $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^p$, where $\mathbb{R}^p$ is a p-dimensional real-valued vector space

$$\mathbf{u} + \mathbf{v} = (u_1, \ldots, u_p) + (v_1, \ldots, v_p) = (u_1 + v_1, \ldots, u_p + v_p) = (w_1, \ldots, w_p) = \mathbf{w}$$

Scalar Multiplication

Here, we add multiply a vector by a scalar value. This is done by multiplying each element of the vector by the scalar. Note the result is a vector of the same shape, signifying that it is a member of the same vector space.

$$3 \cdot (-1.5, 1.3) = (-4.5, 3.9)$$

$$(-1.5, 1.3), (-4.5, 3.9) \in \mathbb{R}^2$$

Scalar Multiplication in General

Generally, with vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$, where $\mathbb{R}^p$ is a p-dimensional real-valued vector space and scalar $\beta \in \mathbb{R}$

$$\beta\mathbf{u} = \beta(u_1, \ldots, u_p) = (\beta u_1, \ldots, \beta u_p) = (v_1, \ldots, v_p) = \mathbf{v}$$

The Linear Combination

A linear combination is the vector result of a vector addition and scalar multiplication of vectors.

Linear Combination in General

Generally, with vectors $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^p$, where $\mathbb{R}^p$ is a p-dimensional real-valued vector space and scalars $\beta, \gamma \in \mathbb{R}$

$$\beta\mathbf{u}+\gamma\mathbf{v} = \beta(u_1, \ldots, u_p)+\gamma(v_1, \ldots, v_p) = (\beta u_1+\gamma v_1, \ldots, \beta u_p+\gamma v_p) = (w_1, \ldots, w_p) = \mathbf{w}$$

The length of a vector

The length of a vector is the number of elements in that vector.

The magnitude of a vector

The magnitude of a 2-d vector is something you have certainly seen before.

The magnitude of a is

$$||a|| = \sqrt{a_x^2 + a_y^2}$$

which should be all to familiar to you as the Pythagorean Theorem.

Incredibly, this generalizes to p-dimensional vectors.

$$||a_p|| = \sqrt{a_1^2 + a_2^2 \cdots + a_p^2}$$

The ℓ_2-Norm

This computation has a fancy name, the ℓ_2-norm and can be computed using `norm` with the argument `type="2"`.

We will see our friend, the ℓ_2-norm again.

For now, let's consider the first computation.

$$||a|| = \sqrt{a_x^2 + a_y^2}$$

We can rewrite this as

$$||a|| = \sqrt{a_x a_x + a_y a_y}$$

Let's consider just the computation under the radical

$$a_x a_x + a_y a_y$$

This is actually a special compuation, the dot product.

The Dot Product

The dot product, also known as the inner product, is an operation defined over a vector space that yields a scalar.

Given

$$\mathbf{u} = (1, 0, -1)$$

$$\mathbf{v} = (-3, 3, -2)$$

The dot product of $\mathbf{u}$ and $\mathbf{v}$ is $\langle \mathbf{u}, \mathbf{v} \rangle$ where

$$\langle \mathbf{u}, \mathbf{v} \rangle = 1 \cdot (-3) + 0 \cdot 3 + (-1) \cdot (-2) = -1$$

$$\langle \mathbf{u}, \mathbf{v} \rangle = \sum u_i \cdot v_i$$

NOTE:

$$\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$$

for all $\mathbf{u}$, $\mathbf{v}$.

The Magnitude is the Square Root Dot Product of a Vector with Itself

Knowing this, it is easy to see that

$$a_x a_x + a_y a_y$$

is the dot product of $\mathbf{a}$ with itself.

Then

$$||a|| = \sqrt{\langle \mathbf{a}, \mathbf{a} \rangle}$$

Geometric Interpretation of the Dot Product

Geometrically, the dot product is the magnitude of the project of one vector onto another vector. Below $proj_B A$ is the projection of A onto B - it is the part of A that is in the same direction as B.

The projection of a vector onto itself is the vector itself!!!

Thus, the dot product is the magnitude of the vector!

A Note on Writing Vectors

These two forms of vector representation are equivalent

$$(a, b, c, d) = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}$$

This form is known as the column vector form.

A vector written as boldface later e.g. $\mathbf{u}$ is typically considered to be in column vector form.

A vector written as

$$\begin{pmatrix} e & f & g & h \end{pmatrix}$$

is considered to be written in the row vector form.

Changing Vectors Forms

Vectors can be transformed from one form to the other via the transpose operation.

$$\mathbf{m} = (1, 2, 3, 4)$$

then

$$\mathbf{m}^T = \begin{pmatrix} 1 & 2 & 3 & 4 \end{pmatrix}$$

The importance of row and column vectors

This is important because we think of the dot product as a row vector times a column vector

$$\langle \mathbf{u}, \mathbf{v} \rangle = \sum u_i \cdot v_i = \mathbf{u}^T \mathbf{v} = \begin{pmatrix} 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} -3 \\ 3 \\ -2 \end{pmatrix} = 1 \cdot (-3) + 0 \cdot + (-1) \cdot (-2) = 1$$

Solving Systems via Linear Algebra

Back to our system of linear equations.

A System of Linear Equations

$$y = 3x - 2$$

$$y = -5x + 6$$

Our system in standard form, where we change our to two variables to x_1 and x_2:

$$-3x_1 + x_2 = -2$$
$$5x_1 + x_2 = 6$$

Can be rewritten as

$$(-3, 1)^T (x_1, x_2) = -2$$
$$(5, 1)^T (x_1, x_2) = 6$$

$$\begin{pmatrix} -3 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = -2$$

$$\begin{pmatrix} 5 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 6$$

$$\begin{pmatrix} -3 & 1 \\ 5 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -2 \\ 6 \end{pmatrix}$$

Let

$$A = \begin{pmatrix} -3 & 1 \\ 5 & 1 \end{pmatrix}$$

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$\mathbf{b} = \begin{pmatrix} -2 \\ 6 \end{pmatrix}$$

Now

$$A\mathbf{x} = \mathbf{b}$$

represents

$$y = 3x - 2$$
$$y = -5x + 6$$

Where the vector $\mathbf{x} = (x_1, x_2) = (x, y)$.

If we had some simple problem like

$$a \cdot x = b$$

we could solve this simply by dividing both sides by a. Another way to say this is that we multiply both sides by the multiplicative inverse of a, which is $\frac{1}{a}$.

$$\frac{1}{a} a \cdot x = \frac{1}{a} b$$

then

$$x = \frac{b}{a}$$

WE CAN NOT DIVIDE VECTORS

In some cases, we can take the inverse.

$$A^{-1}A = I$$

We have

$$A\mathbf{x} = \mathbf{b}$$

If the inverse of A exists

then

$$\begin{aligned} A\mathbf{x} &= \mathbf{b} && (5)\\ A^{-1}A\mathbf{x} &= A^{-1}\mathbf{b} && (6)\\ I\mathbf{x} &= A^{-1}\mathbf{b} && \text{(By definition of the inverse of } A)\\ \mathbf{x} &= A^{-1}\mathbf{b} && \text{(By the property of the inverse)} \end{aligned}$$

Inverting the Problem/Inverting the Matrix

The Inverse

To do this, we will use the linear algebraic inverse operation. This may be completely new to you.

In Algebra I, there were two inverse operations.

The Additive Inverse

Given a number x, the additive inverse is $-x$.

If I add x to a number, transforming the number, I can add $-x$ to the result to transform the number back.

Given: 3, we add 2, which yields 5.

We can transform back to our original number by adding the inverse -2.

The Multiplicative Inverse

Given a number x, the multplicative inverse is $\frac{1}{x}$.

If I multiply a number by x, transforming the number, I can multiply the result by $\frac{1}{x}$ to transform the number back.

Given: 3, we multiply by 2, which yields 6.

$$2 \cdot 3 = 6$$

$$\frac{1}{2} \cdot 6 = 3$$

We can transform back to our original number by multiply by the inverse $\frac{1}{2}$.

Inverting the Problem

Previously, we found a point that was the intersection between two lines. Now, let's find a line that connects two points.

What if we know some points and not the function?

e.g. the points $(-3, 4)$ and $(2, -3)$

We can use linear algebra to find a function to fit these points!

We know that the equation looks like this

$$\beta_0 + \beta_1 x_i = y_i$$

$$1 \cdot \beta_0 + \beta_1 x_i = y_i$$

That we can rewrite as

$$(1, x_i)^T (\beta_0, \beta_1) = y_i$$

or

$$\begin{pmatrix} 1 & x_i \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = y_i$$

We then have two equations

$$\beta_0 + \beta_1 \cdot (-3) = 4 \tag{7}$$
$$\beta_0 + \beta_1 \cdot (2) = -3 \tag{8}$$
$$\tag{9}$$

or

$$(1, -3)^T(\beta_0, \beta_1) = 4 \tag{10}$$
$$(1, 2)^T(\beta_0, \beta_1) = -3 \tag{11}$$
$$\tag{12}$$

We can enter our data using the matrix form

$$\beta_0 \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} -3 \\ 2 \end{pmatrix} = \begin{pmatrix} 4 \\ -3 \end{pmatrix}$$

$$\begin{pmatrix} 1 & -3 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 4 \\ -3 \end{pmatrix}$$

Which we can think of as

$$X\beta = \mathbf{y}$$

The Inverse Matrix

For some matrices, an inverse operation exists. We will come to think of a matrix multiplication as a transformation.

If we have two matrices, say A and B and multiply them, we think of A as transforming B into a new matrix. Let's call the new matrix C.

$$AB = C$$

If we want to change C back into the original matrix, B, and provided that A has an inverse, we can transform C by the inverse matrix A^{-1}.

$$A^{-1}C = B$$

Matrix Multiplication is not Commutative

$$AB \neq BA$$

We must place the transforming matrix first!

$$\frac{1}{2}2 = 1$$

$$A^{-1}A = I$$

$$A^{-1}A = AA^{-1} = I$$

Solving the System using the Inverse

Our system is

$$X\beta = \mathbf{y}$$

Here we see that X is transforming β into $\mathbf{y}$. Keep in mind that we know X and we know $\mathbf{y}$. What we want to know is β.

What we want to do is transform $\mathbf{y}$ back into β. We can do this with the inverse matrix, X^{-1} (provided that it exists).

$$\beta = X^{-1}\mathbf{y}$$

An Over-determined System

What if I have more points than variables?

e.g. $(-3.1, 4.2)$, $(-2.1, 2.4)$, $(1.8, -2.5)$, $(0.5, -1.3)$, $(-1.1, 1.9)$

Consider that this is often the case in data science. We generally refer to the number of instances or points as n and the number of features as p. Here we have $n > p$. A system with more n than p is known as an **over-determined system**.

There is no line that will work!!!

What we want is the "best" line

Our problem could be defined as such

Given $(-3.1, 4.2)$, $(-2.1, 2.4)$, $(1.8, -2.5)$, $(0.5, -1.3)$, $(-1.1, 1.9)$, we seek β_0 and β_1 so that

$$\begin{pmatrix} 1 & -3.1 \\ 1 & -2.1 \\ 1 & 1.8 \\ 1 & 0.5 \\ 1 & -1.1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 4.2 \\ 2.4 \\ -2.5 \\ -1.3 \\ 1.9 \end{pmatrix}$$

$$X\beta = \mathbf{y}$$

This implies that

$$\beta = X^{-1}\mathbf{y}$$

X has no inverse!!!

While the precise mathematical meaning of this may elude you, intuitively, you should be able to understand why. We framed the problem as such

Given $(-3.1, 4.2)$, $(-2.1, 2.4)$, $(1.8, -2.5)$, $(0.5, -1.3)$, $(-1.1, 1.9)$, we seek β_0 and β_1 so that

$$\begin{pmatrix} 1 & -3.1 \\ 1 & -2.1 \\ 1 & 1.8 \\ 1 & 0.5 \\ 1 & -1.1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 4.2 \\ 2.4 \\ -2.5 \\ -1.3 \\ 1.9 \end{pmatrix}$$

$$X\beta = \mathbf{y}$$

But there can be no solution to this system of equations. There is no vector $\beta = (\beta_0, \beta_1)$ that will work for all five equations.

If we are going to solve this, we will need to reframe the problem.

Can't we find a non-linear solution

Non-linear functions are linear in Coefficient!!

Why is this important? Because we want to develop our linearly intuition **even if we will use non-linearity later**.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

$$\begin{pmatrix} 1 & x & x^2 & x^3 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = y$$

When is there no Inverse?

When the matrix is singular. This is deep, deep concept. But basically, it signifies that via elimination a row and/or a column of the matrix can be reduced to a column of zeros.

This is obviously true for a rectangular matrix ($n > p$ and $p > n$).

$$\begin{pmatrix} 1 & -3.1 \\ 1 & -2.1 \\ 1 & 1.8 \\ 1 & 0.5 \\ 1 & -1.1 \end{pmatrix} = \begin{pmatrix} 1 & -3.1 & 0 & 0 & 0 \\ 1 & -2.1 & 0 & 0 & 0 \\ 1 & 1.8 & 0 & 0 & 0 \\ 1 & 0.5 & 0 & 0 & 0 \\ 1 & -1.1 & 0 & 0 & 0 \end{pmatrix}$$

When we use this matrix as a transformer **it will lose information that is unrecoverable**. Thus, when we try to invert, we will have incomplete information and will not be able to.

Intuitively, think of holding a color image in a matrix. What if we apply a matrix transformation to this to transform the image to be black and white. There is no inverse operation to transform the image back into color. **We lost information!!**.

Statistical Learning

> Statistical learning theory deals with the problem of finding a predictive function based on data. wikipedia

We observe:

- a set of **input** predictors ($\mathbf{X} = X_1, X_2, \ldots, X_P$)
- **output** response (Y).

We assume that there is some relationship between Y and $\mathbf{X}$ that is, that there is some function f such that f maps $\mathbf{X}$ to Y i.e.

$$f : \mathbf{X} \mapsto Y$$

Then, this relationship can be written in some general form

$$Y = f(\mathbf{X}) + \varepsilon$$

Here:

- $\mathbf{X}$ is our input
- Y is our output

- f is a fixed but unknown function
- ε is an error term

Statistical Learning

In essence, statistical learning refers to a set of approaches for estimating f, that is the function defining the distribution of our data.

Minimizing the Loss

We have thus far framed the problem as

Given $(-3.1, 4.2)$, $(-2.1, 2.4)$, $(1.8, -2.5)$, $(0.5, -1.3)$, $(-1.1, 1.9)$, we seek $\beta = (\beta_0, \beta_1)$ so that

$$\begin{pmatrix} 1 & -3.1 \\ 1 & -2.1 \\ 1 & 1.8 \\ 1 & 0.5 \\ 1 & -1.1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 4.2 \\ 2.4 \\ -2.5 \\ -1.3 \\ 1.9 \end{pmatrix}$$

$$X\beta = \mathbf{y}$$

If we think of this in terms of statistical learning, we can add in some element of error and assume that the relationship between our inputs and outputs can be represented by

$$Y = f(\mathbf{X}) + \varepsilon = X\beta + \varepsilon$$

Let us reformulate this as

$$\varepsilon = X\beta - Y$$

Note that we don't care what the sign of the error is, we just want it to be small!

Here, ε is a vector of errors, one element for each instance in our dataset.

The Magnitude of the Error Vector

We recall that the magnitude of the error vector is given by the dot product.

$$||\varepsilon|| = \sqrt{\sum_{i=1}^{n} \epsilon_i^2} = \sqrt{\varepsilon^T \varepsilon}$$

For convenience, we will let the square of the magnitude stand in and simply consider $\varepsilon^T \varepsilon$. We will define this as our **loss function**, a function of our β vector,

$$\mathcal{L}(\beta) = \varepsilon^T \varepsilon$$

Infinite Possibilities

We have established a form for this function. We have decided that it has the form

$$\widehat{Y} = \widehat{f}(\mathbf{X}) = \beta_0 + \beta_1 x$$

Consider how many different $\widehat{f}$ functions there are. β_0 could be anything from $-\infty$ to ∞, as could β_1. In order to determine the "best" $\widehat{f}$ we will seek the β vector that minimizes the loss function $\mathcal{L}$.

Reframing in terms of "best" fit

Given $(-3.1, 4.2)$, $(-2.1, 2.4)$, $(1.8, -2.5)$, $(0.5, -1.3)$, $(-1.1, 1.9)$, we seek $\beta = (\beta_0, \beta_1)$ so that for

$$\begin{pmatrix} 1 & -3.1 \\ 1 & -2.1 \\ 1 & 1.8 \\ 1 & 0.5 \\ 1 & -1.1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \approx \begin{pmatrix} 4.2 \\ 2.4 \\ -2.5 \\ -1.3 \\ 1.9 \end{pmatrix}$$

$$X\beta \approx \mathbf{y}$$

and with error

$$\varepsilon = X\beta - \mathbf{y}$$

the loss function

$$\mathcal{L}(\beta) = \varepsilon^T \varepsilon$$

is at a minimum.

the Normal Equations

Given $(-3.1, 4.2)$, $(-2.1, 2.4)$, $(1.8, -2.5)$, $(0.5, -1.3)$, $(-1.1, 1.9)$, we seek $\beta = (\beta_0, \beta_1)$ so that for

$$\begin{pmatrix} 1 & -3.1 \\ 1 & -2.1 \\ 1 & 1.8 \\ 1 & 0.5 \\ 1 & -1.1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \approx \begin{pmatrix} 4.2 \\ 2.4 \\ -2.5 \\ -1.3 \\ 1.9 \end{pmatrix}$$

$$X\beta \approx \mathbf{y}$$

and with error

$$\varepsilon = X\beta - \mathbf{y}$$

the loss function

$$\mathcal{L}(\beta) = \varepsilon^T \varepsilon$$

is at a minimum.

How do we find a minimum?

Calculus and Linear Algebra!!

Special Transpose Property

We are going to need this fact: The transpose has a special property so that

$$(UV)^T = V^T U^T$$

Minimizing the Loss Function

$$\mathcal{L}(\beta) = \varepsilon^T \varepsilon \tag{13}$$
$$= (X\beta - \mathbf{y})^T (X\beta - \mathbf{y}) \tag{14}$$
$$\tag{15}$$

Expansion

Let's expand this using FOIL.

$$\mathcal{L}(\beta) = (X\beta)^T X\beta - \mathbf{y}^T X\beta - (X\beta)^T \mathbf{y} + \mathbf{y}^T \mathbf{y} \tag{16}$$

Application of Transpose Property

We apply the special property of the transpose to the first term

$$\mathcal{L}(\beta) = \beta^T X^T X\beta - \mathbf{y}^T X\beta - (X\beta)^T \mathbf{y} + \mathbf{y}^T \mathbf{y} \tag{17}$$

Combine middle terms using Properties of Dot Product

Recall that the dot product is order **independent**. Therefore

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

Therefore, we can combine the middle terms which gives us

$$\mathcal{L}(\beta) = \beta^T X^T X\beta - 2(X\beta)^T \mathbf{y} + \mathbf{y}^T \mathbf{y} \tag{18}$$

Application of the Transpose Property

We next apply the transpose property again

$$\mathcal{L}(\beta) = \beta^T X^T X\beta - 2\beta^T X^T \mathbf{y} + \mathbf{y}^T \mathbf{y} \tag{19}$$

Minimize by Taking the Derivative

To find the minimum of the loss function for a vector β, we take the derivative with respect to β.

$$\frac{d}{d\beta}\mathcal{L}(\beta) = \frac{d}{d\beta}\left(\beta^T X^T X\beta - 2\beta^T X^T \mathbf{y} + \mathbf{y}^T \mathbf{y}\right) \tag{20}$$

And set it equal to zero.

$$\frac{d}{dx}x^b = bx^{b-1}$$

$$\frac{d}{dx}x^2 = \frac{d}{dx}x \cdot x = 2x$$

$$\frac{d}{dx}3x = 3$$

$$\frac{d}{dx}5 = 0$$

$$\frac{d}{d\beta}\mathcal{L}(\beta) = 2X^T X\beta - 2X^T \mathbf{y} = 0 \tag{21}$$

$$\tag{22}$$

This leaves us with

$$X^T X\beta = X^T \mathbf{y}$$

Recall the Original Data

$$\begin{pmatrix} 1 & -3.1 \\ 1 & -2.1 \\ 1 & 1.8 \\ 1 & 0.5 \\ 1 & -1.1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 4.2 \\ 2.4 \\ -2.5 \\ -1.3 \\ 1.9 \end{pmatrix}$$

this result shows that we have a minimum loss when

$$X^T X\beta = X^T \mathbf{y}$$

which we can easily solve by

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

Normal Equations in Matrix Form

This equation is known as the matrix form of the normal equations

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

|

"Best" Fit

It can be shown by the Gauss-Markov theorem that the β vector we found defines the "best" fit, that is it defines the line with the Best Linear Unbiased Estimator.