03 Theory of linear regression

Leomar Durán

19 September 2023

1 The linear regression model

The linear regression model works by considering each entity with F features and L labels as a set of coordinates in a coordinate system with as many $(F+L)$ dimensions. For each entity indexed $\ell \in [1..L]$, we then estimate one line that crosses through every data point coordinate with features $(x_f | f \in [1..F])$,

$$y_\ell = \beta_0 + \beta_1 x_{\ell 1} + \beta_2 x_{\ell 2} + \cdots + \beta_F x_{\ell F} + \epsilon_\ell, \tag{1}$$

with error term ϵ_ℓ (which we may think of as additive noise) and weights given by $\vec{\beta}$ (to be discussed later). We can use this line to estimate labels given combinations of features that were not part of the original data set.

This is a generalization of finding the equation of a line

$$y = mx + b, \tag{2}$$

given two coordinates and finding other coordinates on the line given either the input or output, however in higher dimensions.

Using the comparison of the slope-intercept equation of a line, we may consider β_0 as analogous to the intercept (or bias), and $\vec{\beta}$ in general as analogous to the slope (which we will see later).

To train the model, we estimate the weights $\vec{\beta} := (\beta_0, \beta_1, \ldots, \beta_F)$. The model is stored as its name "'linear regression" and the weights. From these we can predict any new label given the features.

2 The linear classification model

The linear classification model builds on linear regression by assigning a class label based on the values of the continuous labels. The minimum working example of this is a predicate on a label

$$\mathcal{P}_\theta(y) :\Leftrightarrow (y \geq \theta), \tag{3}$$

that returns true if the value is at least some threshold value θ, or false otherwise, thus assigning the class true (or positive) to labels $y \geq \theta$ and the class false (or negative) to labels $y < \theta$.

For this model, we store a linear regression model, the threshold θ and the corresponding label.

3 Training the model

To train the models, we estimate find the weights, $\vec{\beta} := (\beta_0, \beta_1, \dots, \beta_F)$ given the relation of labels to weights and features

$$y_\ell = \beta_0 + \beta_1 x_{\ell 1} + \beta_2 x_{\ell 2} + \cdots + \beta_F x_{\ell F} + \epsilon_\ell. \tag{4}$$

If we multiply β_0 by 1 treating the 1 as an x-factor, it still remains as β_0. Likewise, if we switch any $\beta_f x_f = x_f \beta_f$ due to commutativity. So there is no change. Further, if we subtract the error term, we have

$$(y_\ell - \epsilon_\ell) = (1)\beta_0 + x_{\ell 1}\beta_1 + x_{\ell 2}\beta_2 + \cdots + x_{\ell F}\beta_F, \tag{5}$$

which is clearly a pattern of multiplication, showing $\vec{\beta}$ as comparative to the slope with an extra x fixed at 1.

Let's explore linear algebra to explain this pattern.

3.1 Relevant concepts in linear algebra

3.1.1 Vectors

A **vector** is a fixed-length linear ordered list of numbers. A common example of a vector is a set of coordinates. We say that a length-N vector $\vec{v}$ over the set of complex numbers $\mathbb{C}$ is s.t. $\vec{v} \in \mathbb{C}^N$.

On the other hand, the real numbers are scalar, meaning that they can be used to define vectors, and we may think of them as length-1 vectors. That is $\mathbb{R} = \mathbb{R}^1$.

Vectors are represented as lowercase variable name ray above ($\vec{x}$) when they are isolated or handwritten, or in bold roman lowercase variable name ($\mathbf{x}$) when used in expressions where the ray would be cumbersome in typing.

As for the value of the vector, one representation may be the same as with coordinates. For example, $\vec{u} := (1, 2, 3, 4) \in \mathbb{R}^4$. Another more common representation is by a column of numbers in square brackets. (Usually they are square brackets, but sometimes they may be parentheses. As always use the standard used in your context.)

For example,

$$\vec{u} := \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} \in \mathbb{R}^4. \tag{6}$$

Elements are referenced as variables with indices (normal font weight, italicized). For example, $u_2 = 2$.

Vectors of the same length form vector spaces, specifically inner product spaces. A vector space defines addition of vectors and scalar multiplication. An inner product space additionally defines the inner product space, another type of multiplication.

Addition over vectors in the space vector space is performed by addition of its elements. That is, for any two length-N vectors over the set of complex numbers $\vec{u}, \vec{v} \in \mathbb{C}^N$, the sum

$$\mathbf{u} + \mathbf{v} := \begin{bmatrix} u_1 + v_1 \\ u_2 + v_2 \\ \vdots \\ u_N + v_N \end{bmatrix} \in \mathbb{C}^N. \tag{7}$$

For example, if you $\vec{u}$ from (6) and also

$$\vec{v} := \begin{bmatrix} 5 \\ 5 \\ 5 \\ 5 \end{bmatrix} \in \mathbb{R}^4. \tag{8}$$

Then the sum

$$\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} + \begin{bmatrix} 5 \\ 5 \\ 5 \\ 5 \end{bmatrix} = \mathbf{u} + \mathbf{v} = \begin{bmatrix} u_1 + v_1 \\ u_2 + v_2 \\ u_3 + v_3 \\ u_4 + v_4 \end{bmatrix} = \begin{bmatrix} 1+5 \\ 2+5 \\ 3+5 \\ 4+5 \end{bmatrix} = \begin{bmatrix} 6 \\ 7 \\ 8 \\ 9 \end{bmatrix} \in \mathbb{R}^4. \tag{9}$$

Scalar multiplication of a vector is another operation defined by vector spaces. This is multiplication of a vector by a scalar such as a real number. For this purpose, a complex number may also be counted as a scalar. When a vector is multiplied by a scalar, each member of the vector is multiplied by the same vector in parallel. Given $\vec{u} \in \mathbb{C}^N$ and $z \in \mathbb{C}$,

$$z\mathbf{u} := \begin{bmatrix} z(u_1) \\ z(u_2) \\ \vdots \\ z(u_N) \end{bmatrix} \in \mathbb{C}^N. \tag{10}$$

For example, suppose we have $z := 1 + i$ and

$$\vec{w} = \begin{bmatrix} 1 - i \\ 2 - i \\ 5 \\ 1 + 4i \end{bmatrix} \in \mathbb{C}^4. \tag{11}$$

Then

$$(1+i)\begin{bmatrix}1-i\\2-i\\5\\1+4i\end{bmatrix}=z\mathbf{w}=\begin{bmatrix}z(w_1)\\z(w_2)\\z(w_3)\\z(w_4)\end{bmatrix}=\begin{bmatrix}(1+i)(1-i)\\(1+i)(2-i)\\(1+i)5\\(1+i)(1+4i)\end{bmatrix}=\begin{bmatrix}2\\3+i\\5+5i\\-3+5i\end{bmatrix}\in\mathbb{C}^4. \quad (12)$$

Multiplication (dot product) However, we are concerned with the inner product, for which, let us consider the set of all complex numbers as an extension of the real numbers
$\mathbb{C} := \left\{x+yi \middle| x,y\in\mathbb{R}, i^2=-1\right\}$.

The inner product of two vectors in the same inner product space is defined as the sum of the product of each pair of parallel elements with the same index after taking the complex conjugate of the right element. So for any two length-N vectors over the set of complex numbers $\vec{u},\vec{v}\in\mathbb{C}^N$, the inner product

$$\langle\mathbf{u},\mathbf{v}\rangle := \sum_{k\in 1}^{N} u_k\overline{v_k} = u_1\overline{v_1}+u_2\overline{v_2}+\cdots+u_N\overline{v_N}\in\mathbb{C}. \quad (13)$$

Now since $0\in\mathbb{R}$, the complex conjugate of any real number x, $\overline{x}=\overline{x+(0)i}=\overline{x+yi}=x-yi=x-(0)i=x$. Thus for real numbers, the multiplication reduces to multiplication between the parallel elements with complex conjugation being equivalent to an identity function.

For example, given $\vec{u}$ defined in (6) and $\vec{v}$ defined in (8), we have the inner product

$$\begin{aligned}&\left\langle\begin{bmatrix}1\\2\\3\\4\end{bmatrix},\begin{bmatrix}5\\5\\5\\5\end{bmatrix}\right\rangle\\ &=\langle\mathbf{u},\mathbf{v}\rangle\\ &=u_1v_1+u_2v_2+u_3v_3+u_4v_4\\ &=1(5)+2(5)+3(5)+4(5)\\ &=5+10+15+20\\ &\in\mathbb{R}.\end{aligned} \quad (14)$$

The inner product of a vector with itself gives us the square of its ℓ^2-norm

$$\|\mathbf{u}\|_2^2 := \langle\mathbf{u},\mathbf{u}\rangle. \quad (15)$$

3.1.2 Matrices

A **matrix** is a fixed-sized rectangular ordered array of numbers. We say that an R-row C-column matrix $\mathbf{M}$ over the set of complex numbers $\mathbb{C}$ is s.t.
$\mathbf{M}\in\mathrm{Mtrx}(\mathbb{C};R\times C)$.

We may think of a vector as a 1-column matrix, sometimes called a column vector.

A 1-row matrix is called a row vector, but it is not really a vector, rather the transpose of one. This is useful for defining vectors as the transpose $\cdot^\dagger$ of a row vector in one line while differentiating from a set of coordinates (but more on transposes later). For example, $\vec{u} := \begin{bmatrix}1 & 2 & 3 & 4\end{bmatrix}^\dagger \in \mathbb{R}^4$.

We can also thing of a matrix as a column vector of row vectors whose transposes are in the same vector space.

The diagonal of a matrix is the oblique line starting from the first element in the top row and every element moving once towards the bottom and opposite end of the matrix. That is, the diagonal of a matrix is every element where the row number and column number are equal.

The conjugate transpose of a matrix $\mathbf{M} \in \mathrm{Mtrx}(\mathbb{C}; R \times C)$, demarcated as $\mathbf{M}^\dagger$ is a matrix s.t. each member of the diagonal of the product $\mathbf{M}^\dagger\mathbf{M}$ is the square of ℓ^2-norm $\|\vec{\cdot}\|_2^2$ of the corresponding column vector of the original matrix M.

To find the conjugate transpose, we first find the complex conjugate of every element in the matrix. Then we swap the row and column numbers. So for example, for matrix $\mathbf{M} \in \mathrm{Mtrx}(\mathbb{C}; R \times C)$, the conjugate transpose

$$\begin{bmatrix} M_{11} & M_{12} & \cdots & M_{1C} \\ M_{21} & M_{22} & \cdots & M_{2C} \\ \vdots & \vdots & \ddots & \vdots \\ M_{R1} & M_{R2} & \cdots & M_{RC} \end{bmatrix}^\dagger = \begin{bmatrix} \overline{M_{11}} & \overline{M_{21}} & \cdots & \overline{M_{R1}} \\ \overline{M_{12}} & \overline{M_{22}} & \cdots & \overline{M_{R2}} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{M_{1C}} & \overline{M_{2C}} & \cdots & \overline{M_{RC}} \end{bmatrix} \tag{16}$$

As noted in paragraph 3.1.1, conjugation of a real number results in the same number. As a result, when the matrix is over the set of real numbers, the conjugate transpose simplifies to the element-wise transpose, denoted as $\mathbf{M}^\intercal$, where the step of complex conjugation is skipped.

Matrix multiplication: base case Matrix multiplication is a recursive problem. In order to understand how to multiply two larger matrices, we must learn how to multiply the base case.

This base case is multiplying a row vector $\mathbf{v}^\dagger$ and a column vector $\mathbf{u}$. As it turns out this is exactly the same finding the inner product. That is $\mathbf{v}^\dagger\mathbf{u} = \langle \mathbf{u}, \mathbf{v} \rangle$ (as a corollary). As noted before, it is then necessary the row vector to have width L corresponding to the column vector's length L. Thus for $\vec{u}, \vec{v} \in \mathbb{C}^L$,

$$\mathbf{v}^\dagger\mathbf{u} = \begin{bmatrix} v_1 & v_2 & \cdots & v_L \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_L \end{bmatrix} := \left\langle \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_L \end{bmatrix}, \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_L \end{bmatrix} \right\rangle = \langle \mathbf{u}, \mathbf{v} \rangle \in \mathbb{C}. \tag{17}$$

Matrix multiplication: recursive case Thus, to multiply two matrices $\mathbf{M}, \mathbf{N}$, first since we are finding the inner product of rows in $\mathbf{M}$ with columns in $\mathbf{N}$, then the number of columns in $\mathbf{M}$ (*i.e.*, the width of the rows in $\mathbf{M}$) must equal the number of rows in $\mathbf{N}$ (*i.e.*, the length of the columns in $\mathbf{N}$). Thus, we have number of rows R, inner vector length L and number of columns $C \in \mathbb{N}$ and two matrices $\mathbf{M} \in \mathrm{Mtrx}(\mathbb{C}; R \times L), \mathbf{N} \in \mathrm{Mtrx}(\mathbb{C}, L \times C)$, which will product $P \in \mathrm{Mtrx}(\mathbb{C}, R \times C)$.

To multiply matrices $\mathbf{M}, \mathbf{N}$, we multiply each row $M_{r:}$ with each column $N_{:c}$ for $r \in [1..R], c \in [1..C]$ which gives each product in product matrix cell P_{rc}.

For example, we have matrices

$$M := \begin{bmatrix} 1 & 2 \\ 4 & 5 \\ 7 & 8 \end{bmatrix} \in \mathrm{Mtrx}(\mathbb{R}, 3 \times 2), \tag{18}$$

$$N := \begin{bmatrix} 10 & 20 & 30 & 40 \\ 60 & 70 & 80 & 90 \end{bmatrix} \in \mathrm{Mtrx}(\mathbb{R}, 2 \times 4). \tag{19}$$

Their product

$$\begin{aligned}
\mathbf{MN} &= \begin{bmatrix} 1 & 2 \\ 4 & 5 \\ 7 & 8 \end{bmatrix} \begin{bmatrix} 10 & 20 & 30 & 40 \\ 60 & 70 & 80 & 90 \end{bmatrix} \\
&= \begin{bmatrix}
\begin{bmatrix} 1 & 2 \end{bmatrix}\begin{bmatrix} 10 \\ 60 \end{bmatrix} & \begin{bmatrix} 1 & 2 \end{bmatrix}\begin{bmatrix} 20 \\ 70 \end{bmatrix} & \begin{bmatrix} 1 & 2 \end{bmatrix}\begin{bmatrix} 30 \\ 80 \end{bmatrix} & \begin{bmatrix} 1 & 2 \end{bmatrix}\begin{bmatrix} 40 \\ 90 \end{bmatrix} \\
\begin{bmatrix} 4 & 5 \end{bmatrix}\begin{bmatrix} 10 \\ 60 \end{bmatrix} & \begin{bmatrix} 4 & 5 \end{bmatrix}\begin{bmatrix} 20 \\ 70 \end{bmatrix} & \begin{bmatrix} 4 & 5 \end{bmatrix}\begin{bmatrix} 30 \\ 80 \end{bmatrix} & \begin{bmatrix} 4 & 5 \end{bmatrix}\begin{bmatrix} 40 \\ 90 \end{bmatrix} \\
\begin{bmatrix} 7 & 8 \end{bmatrix}\begin{bmatrix} 10 \\ 60 \end{bmatrix} & \begin{bmatrix} 7 & 8 \end{bmatrix}\begin{bmatrix} 20 \\ 70 \end{bmatrix} & \begin{bmatrix} 7 & 8 \end{bmatrix}\begin{bmatrix} 30 \\ 80 \end{bmatrix} & \begin{bmatrix} 7 & 8 \end{bmatrix}\begin{bmatrix} 40 \\ 90 \end{bmatrix}
\end{bmatrix} \\
&= \begin{bmatrix}
\left\langle \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 10 \\ 60 \end{bmatrix} \right\rangle & \left\langle \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 20 \\ 70 \end{bmatrix} \right\rangle & \left\langle \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 30 \\ 80 \end{bmatrix} \right\rangle & \left\langle \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 40 \\ 90 \end{bmatrix} \right\rangle \\
\left\langle \begin{bmatrix} 4 \\ 5 \end{bmatrix}, \begin{bmatrix} 10 \\ 60 \end{bmatrix} \right\rangle & \left\langle \begin{bmatrix} 4 \\ 5 \end{bmatrix}, \begin{bmatrix} 20 \\ 70 \end{bmatrix} \right\rangle & \left\langle \begin{bmatrix} 4 \\ 5 \end{bmatrix}, \begin{bmatrix} 30 \\ 80 \end{bmatrix} \right\rangle & \left\langle \begin{bmatrix} 4 \\ 5 \end{bmatrix}, \begin{bmatrix} 40 \\ 90 \end{bmatrix} \right\rangle \\
\left\langle \begin{bmatrix} 7 \\ 8 \end{bmatrix}, \begin{bmatrix} 10 \\ 60 \end{bmatrix} \right\rangle & \left\langle \begin{bmatrix} 7 \\ 8 \end{bmatrix}, \begin{bmatrix} 20 \\ 70 \end{bmatrix} \right\rangle & \left\langle \begin{bmatrix} 7 \\ 8 \end{bmatrix}, \begin{bmatrix} 30 \\ 80 \end{bmatrix} \right\rangle & \left\langle \begin{bmatrix} 7 \\ 8 \end{bmatrix}, \begin{bmatrix} 40 \\ 90 \end{bmatrix} \right\rangle
\end{bmatrix} \\
&= \begin{bmatrix} 130 & 160 & 190 & 220 \\ 340 & 430 & 520 & 610 \\ 550 & 700 & 850 & 1000 \end{bmatrix} \\
&\in \mathrm{Mtrx}(\mathbb{R}, 3 \times 4).
\end{aligned} \tag{20}$$

Matrices and vectors in representations of systems of linear equations One very important use of matrices and vectors is in the representations of systems of linear equations, which we may remember from algebra.

Note: that this is different from our linear regression problem in that linear regression finds weights given inputs (features) and outputs (labels), whereas the system of linear equations gives inputs and outputs given weights and constants.

For example, we may have the linear system of equations

$$\begin{cases} L_1: & y = x + 1, \\ L_2: & y = 2x - 1. \end{cases} \tag{21}$$

This system may be put into the standard form with all unknowns to the left side of the equal sign and the constants isolated on the right.

$$\begin{cases} L_1: & x - y = -1, \\ L_2: & 2x - y = 1. \end{cases} \tag{22}$$

Further, let's rewrite each term with an unknown so the left side is a sum of the product of each pair of coefficient and unknown.

$$\begin{cases} L_1: & (1)x + (-1)y = -1, \\ L_2: & (2)x + (-1)y = 1. \end{cases} \tag{23}$$

Well now we can rewrite these as one equation with a vector on either side.

$$\begin{bmatrix} (1)x + (-1)y \\ (2)x + (-1)y \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \tag{24}$$

In the vector on the left, we see that both elements of the vector are a sum of some multiplication of x and some multiplication of y. This means we can factor out a vector $\vec{x} := \begin{bmatrix} x & y \end{bmatrix}^\dagger$ leaving a matrix $\mathbf{M}$.

$$\begin{bmatrix} 1 & -1 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \tag{25}$$

Finally, we may augment the matrix on the left with the constants on the left. This form is called an augmented matrix (represented by a matrix with prefixed with an $\mathbf{A}$).

$$\mathbf{AM} = \left[\begin{array}{cc|c} 1 & -1 & -1 \\ 2 & -1 & 1 \end{array}\right]. \tag{26}$$

Row reduction is an operation of transformation of a matrix.

To perform row reduction, we must manipulate its rows according to the following rules

1. We may swap rows.
2. We may multiply a row by a nonzero scalar.
3. We may add rows.

These operations are in practice just matrix multiplications by square matrices with no 0 rows or 0 columns. However, for the sake of clarity for this introduction, we will use the notation $R_k \leftrightarrow R_\ell$ for swapping, and $R_k \leftarrow a_k R_k + a_\ell R_\ell$ for $k, \ell \in [1..R] a_k \in \mathbf{R} \backslash \{0\}$ for multiplying by scalars and adding rows, omitting rows that are multiplied by 0.

When solving a system of equals represented in augmented matrix form, the goal is to put the augmented matrix into reduced row echelon form.

Reduced row echelon form is when

1. The leading nonzero element of each row is to the right of that of the previous row.

2. Zero rows (if any) are at the bottom of the matrix.

3. The leading nonzero element of each row is 1.

4. All other elements in each column with a leading nonzero element must be 0.

The reduced row echelon form is unique to every matrix.

Let's look at a more complex example of a system of linear equations, 4 equations, 4 unknowns in 4 dimensions. In standard form,

$$\begin{cases} L_1: & x_1 - x_2 - x_3 + x_4 = 1, \\ L_2: & 2x_1 - x_2 = 1, \\ L_3: & x_1 + 4x_2 + 3x_4 = 35, \\ L_4: & x_1 + 2x_2 + x_3 + x_4 = 20. \end{cases} \tag{27}$$

Thus we have the augmented matrix

$$\mathbf{AM} = \left[\begin{array}{cccc|c} 1 & -1 & -1 & 1 & 1 \\ 2 & -1 & 0 & 0 & 1 \\ 1 & 4 & 0 & 3 & 35 \\ 1 & 2 & 1 & 1 & 20 \end{array}\right]. \tag{28}$$

We perform the row reduction.

$$
\begin{aligned}
& \mathbf{AM} \\
& = \left[\begin{array}{cccc|c} 1 & -1 & -1 & 1 & 1 \\ 2 & -1 & 0 & 0 & 1 \\ 1 & 4 & 0 & 3 & 35 \\ 1 & 2 & 1 & 1 & 20 \end{array}\right] \\
\underbrace{\begin{matrix} R_1 \leftarrow (-1)R_1 + R_2 \\ R_2 \leftarrow R_2 - 2R_1 \\ R_3 \leftrightarrow R_4 \end{matrix}} \quad & \left[\begin{array}{cccc|c} 1 & 0 & 1 & -1 & 0 \\ 0 & 1 & 2 & -2 & -1 \\ 1 & 2 & 1 & 1 & 20 \\ 1 & 4 & 0 & 3 & 35 \end{array}\right] \\
\underbrace{\begin{matrix} R_3 \leftarrow R_3 - R_1 - 2R_2 \\ R_4 \leftarrow R_4 - R_1 - 4R_2 \end{matrix}} \quad & \left[\begin{array}{cccc|c} 1 & 0 & 1 & -1 & 0 \\ 0 & 1 & 2 & -2 & -1 \\ 0 & 0 & -4 & 6 & 22 \\ 0 & 0 & -9 & 12 & 39 \end{array}\right] \\
\underbrace{\begin{matrix} R_3 \leftarrow 2R_3 - R_4 \\ R_4 \leftarrow -4R_4 + 9R_3 \end{matrix}} \quad & \left[\begin{array}{cccc|c} 1 & 0 & 1 & -1 & 0 \\ 0 & 1 & 2 & -2 & -1 \\ 0 & 0 & 1 & 0 & 5 \\ 0 & 0 & 0 & 6 & 42 \end{array}\right] \\
\underbrace{R_4 \leftarrow (1/6)R_4} \quad & \left[\begin{array}{cccc|c} 1 & 0 & 1 & -1 & 0 \\ 0 & 1 & 2 & -2 & -1 \\ 0 & 0 & 1 & 0 & 5 \\ 0 & 0 & 0 & 1 & 7 \end{array}\right] \\
\underbrace{\begin{matrix} R_1 \leftarrow R_1 - R_3 + R_4 \\ R_2 \leftarrow R_2 - 2R_3 + 2R_4 \end{matrix}} \quad & \left[\begin{array}{cccc|c} 1 & 0 & 0 & 0 & 2 \\ 0 & 1 & 0 & 0 & 3 \\ 0 & 0 & 1 & 0 & 5 \\ 0 & 0 & 0 & 1 & 7 \end{array}\right] \\
& = \mathrm{RREF}(\mathbf{AM}).
\end{aligned}
\tag{29}
$$

So we have found that $\vec{x} = \begin{bmatrix} 2 & 3 & 5 & 7 \end{bmatrix}^\dagger$ given that the constants $\mathbf{Mx} = \vec{b} := \begin{bmatrix} 1 & 1 & 35 & 20 \end{bmatrix}^\dagger$.

Well the constant relates to the x_4-intercept, 4 being the number of columns in matrix $\mathbf{M}$, in that $\frac{b}{x_4}$ is the x_4-intercept. So the x_4 intercepts are $\frac{1}{1}$ for L_1, undefined for L_2, $\frac{35}{3}$ for L_3 and $\frac{20}{1}$ for L_4.

Let's suppose that we have the same weights for the unknowns $\vec{x}$, but different x_4-intercepts, or different constants, say

$$\vec{b} := \begin{bmatrix} 1 & -3 & 63 & 32 \end{bmatrix}^\dagger . \tag{30}$$

Well, the process is the same because goal the was to change the values in the first 4 nonzero leading columns (because there are 4 equations) to the reduced row echelon form. However, repeating this process for every set of constants would be wasteful.

3.1.3 Square matrices

Square matrices have special properties and operations that other matrices do not.

The Gramian matrix of any matrix $\mathbf{M} \in \mathrm{Mtrx}(\mathbb{C}, R \times C)$ is defined as $G_{\mathbf{M}} := \mathbf{M}^{\dagger}\mathbf{M} \in \mathrm{Mtrx}\left(\mathbb{C}, C^2\right)$.

Diagonal matrices are square matrices where all nonzero elements are along the diagonal of the matrix and all other values are zero.

With $N \in \mathbb{N}$, we can use the function

$$\mathrm{diag} := \vec{v} \mapsto \begin{bmatrix} v_1 & 0 & \dots & 0 \\ 0 & v_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & v_3 \end{bmatrix} : \mathbb{C}^N \to \mathrm{Mtrx}\left(\mathbb{C}, N^2\right) \tag{31}$$

to transform a length N vector into a N^2 matrix.

For our purposes, this is just a useful shorthand and we will not deal with diagonal matrices for other uses.

The identity matrix is a special square matrix that only has 1 values along its diagonal and only 0 values outside of its diagonal, it is identified by $\mathbf{1}_N \in \mathrm{Mtrx}\left(\{0,1\}, N^2\right)$, where $N \in \mathbb{N}$ is the number of length of its equal sides. For example

$$\mathbf{1}_4 := \mathrm{diag}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \in \mathrm{Mtrx}\left(\{0,1\}, 4 \times 4\right). \tag{32}$$

The identity matrix is useful because given sizes $R, N, C \in \mathbb{N}$, matrices $\mathbf{M} \in \mathrm{Mtrx}(\mathbb{C}, R \times N), \mathbf{N} \in \mathrm{Mtrx}(\mathbb{C}, N \times C)$ and identity matrix $\mathbf{1}_N$, it is a property that $\mathbf{M}\mathbf{1}_N = \mathbf{M}$ and $\mathbf{1}_N\mathbf{N} = \mathbf{N}$.

An identity matrix of unspecified size may be referenced as $\mathbf{1}$.

Matrix inverse One property that a square matrix may have is invertibility, although not all square matrices are invertible. An invertible matrix will help us estimate the weights for the regression model.

A matrix $\mathbf{M}$ is invertible if there exists $\mathbf{M}^{-1}$ s.t. $\mathbf{M}\mathbf{M}^{-1} = \mathbf{M}^{-1}\mathbf{M} = \mathbf{1}$.

So for the first matrix multiplication, $\mathbf{M}^{-1}$ must have as many rows as $\mathbf{M}$ has columns, and for the second matrix multiplication $\mathbf{M}^{-1}$ must have as many columns as $\mathbf{M}$ has rows, and for both of these products to equal the same identity matrix, then $\mathbf{M}^{-1}$ must have the same number of rows and columns as $\mathbf{M}$. Thus, $\mathbf{M}$ and $\mathbf{M}^{-1}$ are both square matrices.

Now as we saw earlier row reduction can be used to put a matrix into a form where the greatest leftmost square is an identity matrix if the matrix can be put in that form. On the other hand, we can think of it that the identity matrix can be used to record the effects of any matrix multiplication because for any matrix $\mathbf{M}$, $\mathbf{M1} = \mathbf{M}$. So what we are going to do is augment the matrix of coefficients by the identity matrix and perform row reduction, thereby recording all matrix multiplications.

So again.

$$
\begin{aligned}
&\left[\mathbf{M} \mid \mathbf{1}\right] \\
= &\left[\begin{array}{cccc|cccc} 1 & -1 & -1 & 1 & 1 & 0 & 0 & 0 \\ 2 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 4 & 0 & 3 & 0 & 0 & 1 & 0 \\ 1 & 2 & 1 & 1 & 0 & 0 & 0 & 1 \end{array}\right] \\
\underbrace{\begin{matrix} R_1 \leftarrow (-1)R_1 + R_2 \\ R_2 \leftarrow R_2 - 2R_1 \\ R_3 \leftrightarrow R_4 \end{matrix}} \quad &\left[\begin{array}{cccc|cccc} 1 & 0 & 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 1 & 2 & -2 & -2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 4 & 0 & 3 & 0 & 0 & 1 & 0 \end{array}\right] \\
\underbrace{\begin{matrix} R_3 \leftarrow R_3 - R_1 - 2R_2 \\ R_4 \leftarrow R_4 - R_1 - 4R_2 \end{matrix}} \quad &\left[\begin{array}{cccc|cccc} 1 & 0 & 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 1 & 2 & -2 & -2 & 1 & 0 & 0 \\ 0 & 0 & -4 & 6 & 5 & -3 & 0 & 1 \\ 0 & 0 & -9 & 12 & 9 & -5 & 1 & 0 \end{array}\right] \\
\underbrace{\begin{matrix} R_3 \leftarrow 2R_3 - R_4 \\ R_4 \leftarrow -4R_4 + 9R_3 \end{matrix}} \quad &\left[\begin{array}{cccc|cccc} 1 & 0 & 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 1 & 2 & -2 & -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & -1 & -1 & 2 \\ 0 & 0 & 0 & 6 & 9 & -7 & -4 & 9 \end{array}\right] \\
\underbrace{R_4 \leftarrow (1/6)\, R_4} \quad &\left[\begin{array}{cccc|cccc} 1 & 0 & 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 1 & 2 & -2 & -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & -1 & -1 & 2 \\ 0 & 0 & 0 & 1 & 9/6 & -7/6 & -4/6 & 9/6 \end{array}\right] \\
\underbrace{\begin{matrix} R_1 \leftarrow (6/6)\, R_1 - (6/6)\, R_3 + R_4 \\ R_2 \leftarrow (6/6)\, R_2 - (12/6)\, R_3 + 2R_4 \end{matrix}} \quad &\left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & -3/6 & 5/6 & 2/6 & -3/6 \\ 0 & 1 & 0 & 0 & -6/6 & 4/6 & 4/6 & -6/6 \\ 0 & 0 & 1 & 0 & 1 & -1 & -1 & 2 \\ 0 & 0 & 0 & 1 & 9/6 & -7/6 & -4/6 & 9/6 \end{array}\right] \\
= &\left[\mathbf{1} \mid \mathbf{M}^{-1}\right].
\end{aligned}
\tag{33}
$$

We can factor out $1/6$ from R_1, $1/6$ from R_2 and $1/6$ from R_4 into a diagonal matrix. This time we produced the inverse matrix

$$\mathbf{M}^{-1} = \left(\operatorname{diag}\begin{bmatrix} 1/6 \\ 1/6 \\ 1 \\ 1/6 \end{bmatrix}\right) \begin{bmatrix} -3 & 5 & 2 & -3 \\ -6 & 4 & 4 & -6 \\ 1 & -1 & -1 & 2 \\ 9 & -7 & -4 & 9 \end{bmatrix}, \tag{34}$$

which encodes the row reduction of matrix $\mathbf{M}$.

Now just as if we had the scalar equation $mx = b$, we may multiply both sides by m^{-1} to find $m^{-1}mx = x = m^{-1}b$, we will multiply both sides of $\mathbf{Mx} = \mathbf{b}$ by $\mathbf{M}^{-1}$ as an analogy to find $\mathbf{M}^{-1}\mathbf{Mx} = \mathbf{x} = \mathbf{M}^{-1}\mathbf{b}$. For example, we may solve (27) for $\vec{x}$ by performing the matrix multiplication

$$\vec{x} = \mathbf{M}^{-1}\mathbf{b} = \left(\operatorname{diag}\begin{bmatrix} 1/6 \\ 1/6 \\ 1 \\ 1/6 \end{bmatrix}\right) \begin{bmatrix} -3 & 5 & 2 & -3 \\ -6 & 4 & 4 & -6 \\ 1 & -1 & -1 & 2 \\ 9 & -7 & -4 & 9 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 35 \\ 20 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 5 \\ 7 \end{bmatrix}. \tag{35}$$

Likewise, we can find the values of $\vec{x}$ given the proposed constants $\vec{b}$ in (30).

$$\vec{x} = \mathbf{M}^{-1}\mathbf{b} = \left(\operatorname{diag}\begin{bmatrix} 1/6 \\ 1/6 \\ 1 \\ 1/6 \end{bmatrix}\right) \begin{bmatrix} -3 & 5 & 2 & -3 \\ -6 & 4 & 4 & -6 \\ 1 & -1 & -1 & 2 \\ 9 & -7 & -4 & 9 \end{bmatrix} \begin{bmatrix} 1 \\ -3 \\ 63 \\ 32 \end{bmatrix} = \begin{bmatrix} 2 \\ 7 \\ 5 \\ 11 \end{bmatrix}. \tag{36}$$

3.2 Back to training the linear regression model

Note that although it's possible to perform linear regression with complex number, from hereon we will stay within the domain of real numbers for the sake of simplicity.

So now we have covered enough linear algebra to explain the linear regression model.

From (5), we have

$$(y_\ell - \epsilon_\ell) = (1)\beta_0 + x_{\ell 1}\beta_1 + x_{\ell 2}\beta_2 + \cdots + x_{\ell F}\beta_F, \tag{37}$$

If we consider every line ℓ, this should now begin to look like linear algebra. We have a difference of vectors on the left, and a multiplication of a vector by a 1-augmented matrix on the right.

Let us first define the define the 1-augmented feature matrix, using $\vec{1}_L$, which is just a length L vector with all 1 elements, for multiplying the bias β_0.

$$\mathbf{AX} := \left[\vec{1}_L \;\middle|\; \mathbf{X}\right]. \tag{38}$$

Then (5) becomes

$$\mathbf{y} - \boldsymbol{\epsilon} = (\mathbf{AX})\boldsymbol{\beta}. \tag{39}$$

Solving this model will differ from the system of linear equations because here we are solving for weights given features and label, whereas in the system of linear equations we were solving for inputs and outputs given weights and constants. However, we can use the same techniques.

The first technique that we used was the row reduction in (29). However, we do not want to do row reduction to find a linear regression model. Our training data had 16,512 entities × 12 features (plus the augmented 1 column).

Luckily, we have your favorite spreadsheet application to handle the calculations for us. However, note that it does not perform row reduction. So this leaves our second method, which is multiplication by the matrix inverse such as (35). Algebra systems and spreadsheet applications generally support the matrix inverse operation. So maybe we can multiply by $(\mathbf{AX})^{-1}$. The only issue is that only square matrices may be inverted. This is where The Gramian matrix comes into play.

If we think of a vector as we do in physics (as a magnitude and direction from the origin), and the matrix as a transformation of that vector, then it makes sense that the matrix-matrix-vector product $\mathbf{MNx}$ is a transformation of $\mathbf{x}$ by $\mathbf{N}$ followed by a transformation of $\mathbf{M}$. In that case, taking the inverse of the transformations $(\mathbf{MN})^{-1} = \mathbf{N}^{-1}\mathbf{M}^{-1}$, that is the inverse operations happen in reverse, as intuition may imply.

Thus, the inverse of the Gramian matrix,

$$G_{\mathbf{AX}}^{-1} = \left(\mathbf{AX}^{\intercal}\right)\left(\mathbf{AX}\right) = \left(\mathbf{AX}\right)^{-1}\left(\mathbf{AX}^{\intercal}\right)^{-1}. \tag{40}$$

Let's assume that the feature matrix $\mathbf{AX}$ is a square matrix. We may find the inverse of $\mathbf{AX}$ as follows.

$$\begin{aligned}
&\left(\mathbf{AX}\right)^{-1} \\
\overset{\text{by identity}}{=}\; &\left(\mathbf{AX}\right)^{-1}\mathbf{1} \\
\overset{\text{by def'n of } [\cdot]^{-1}}{=}\; &\left(\mathbf{AX}\right)^{-1}\Big(\left(\mathbf{AX}^{\intercal}\right)^{-1}\left(\mathbf{AX}^{\intercal}\right)\Big) \\
\overset{\text{by association}}{=}\; &\Big(\left(\mathbf{AX}\right)^{-1}\left(\mathbf{AX}^{\intercal}\right)^{-1}\Big)\left(\mathbf{AX}^{\intercal}\right) \\
\overset{\text{by } G^{-1}}{=}\; &G_{\mathbf{AX}}^{-1}\left(\mathbf{AX}^{\intercal}\right).
\end{aligned} \tag{41}$$

Thus, we can use $G_{\mathbf{AX}}^{-1}\left(\mathbf{AX}^{\intercal}\right)$ as a generalization of the inverse of a matrix because neither of these operations requires a square matrix.

Finally, the number of entities provided by our data is $16{,}512 \gg 12$, which is the number of features. This means that the linear regression model provides more equations than there are unknowns. To fit the entities to the weights, we use the least squares approach. This approach minimizes the square of the error when comparing the expected "ground truth" value of the labels to the predicted value.

So to estimate the labels, we assume that the error term $\boldsymbol{\epsilon} = \vec{0}_L$ and solve for the estimator of the weights $\boldsymbol{\beta}$.

$$\hat{\boldsymbol{\beta}} = \left(G_{\mathbf{AX}}^{-1}\left(\mathbf{AX}^{\intercal}\right)\right)(\mathbf{AX})\hat{\boldsymbol{\beta}} = G_{\mathbf{AX}}^{-1}\left(\mathbf{AX}^{\intercal}\right)\left(\mathbf{y} - \vec{0}_L\right) = G_{\mathbf{AX}}^{-1}\left(\mathbf{AX}^{\intercal}\right)\mathbf{y}. \tag{42}$$