What Does Linear Algebra Have to do With Linear Regression?

Logan Higginbotham

October 9, 2023

This was a question I was asked multiple times over the Summer at the IAA. I hold a doctorate in mathematics and have taught linear algebra before. I was willing to provide my expert knowledge of the subject to other students of our cohort. Every day during our linear algebra week, I held special sessions to supplement our lectures because I wanted to help all of us succeed. And I know that they would have done the same for me.

Such willingness makes the IAA special: our cohorts comprise unique, talented individuals from many different walks of life and are experts in various fields. Importantly, we are happy to share our talents with others. At the Institute, the collaborative spirit is front and center!

But back to the question. What does linear algebra have to do with linear regression? Regarding math and data science, it can be hard to understand how to apply what we see to what we do. That is why I want to show how the linear algebra we learned can be applied to something we saw very early in our instruction – linear regression!

But let's go over what linear regression is first. When we perform linear regression, we have data points $(x_1, y_1), ..., (x_m, y_m)$. Each point is an observation (e.g., x_4 is the amount of fertilizer given to plant 4, and y_4 is the observed height of the fourth plant). Our goal is to find two coefficients, $\hat{\beta}_0, \hat{\beta}_1$, so that the line $\hat{\beta}_0 + \hat{\beta}_1 x = y$ is as "close" to the the data points as possible.

If we wanted to fit data $(1, 2), (5, 7)$, and $(6, 4)$, then we (ideally) want these points to be on a line for some y-intercept β_0 and slope β_1. In other words, we want a line that contains the three points above. We can create a system of linear equations:

$$\begin{cases} \beta_0 + 1\beta_1 = 2 \\ \beta_0 + 5\beta_1 = 7 \\ \beta_0 + 6\beta_1 = 4 \end{cases}$$

Or in matrix notation

$$\begin{bmatrix} 1 & 1 \\ 1 & 5 \\ 1 & 6 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 2 \\ 7 \\ 4 \end{bmatrix}$$

This is a system of equations $A\vec{x} = \vec{b}$! However, statisticians would instead write $X\vec{\beta} = \vec{y}$. The beta coefficients are the unknown values we want to find, and $\vec{y}$ is our y-variable data put into a vector. Our X matrix (sometimes called a **feature matrix**) has columns that are our features in our modeling. The first column is all 1's (for our y-intercept β_0), while the second column collates our x-variable data.

More generally, if we want to fit data $(x_1, y_1), ..., (x_m, y_m)$, we would get a system $X\vec{\beta} = \vec{y}$ where

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix}, \ \vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \ \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

We don't expect $X\vec{\beta} = \vec{y}$ to have a perfect solution since a straight line *can't* go through all of the data points. When a system of linear equations doesn't have a solution, we can find the "best" choice of $\vec{\beta}$, notated $\hat{\beta}$, that minimizes our sum of squared residuals. By definition, $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$ is a **least squares solution** to $X\vec{\beta} = \vec{y}$.

We learned in our classes that the least squares solution to $X\vec{\beta} = \vec{y}$ is $\hat{\beta} = X^{\dagger}\vec{y}$, where $X^{\dagger} = (X^T X)^{-1} X^T$ is the **pseudo-inverse** of X. In our example above,

$$\begin{bmatrix} 1 & 1 \\ 1 & 5 \\ 1 & 6 \end{bmatrix}^{\dagger} = \begin{bmatrix} 1.19 & .048 & -.238 \\ -.214 & .071 & .143 \end{bmatrix}$$

Thus, the least squares solution (the beta-coefficients for our line of best fit) is $X^{\dagger}\vec{y}$ i.e.

$$\hat{\beta} = \begin{bmatrix} 1.19 & .048 & -.238 \\ -.214 & .071 & .143 \end{bmatrix} \begin{bmatrix} 2 \\ 7 \\ 4 \end{bmatrix} = \begin{bmatrix} 1.76 \\ .64 \end{bmatrix}$$

And our line of best fit is $\hat{y} = 1.76 + .64x$, where $\hat{y}$ is our **predicted** y-value determined by our least squares coefficients $\hat{\beta}$.

Note that $X^{\dagger}$ *only* exists when the columns of X are linearly independent. Analogously, statisticians would say that the columns of X are **uncorrelated**. This is why the assumption of no perfect multi-collinearity is there. Otherwise, there would be no $X^{\dagger}$!

We've talked about simple linear regression, but what about multiple linear regression? Linear algebra makes this easy! If we want to fit data that looks like $(x_1, ..., x_k, y)$ to a higher dimensional "line"

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

We plug each data point into the x's and y to yield a system of linear equations $X\vec{\beta} = \vec{y}$, where the first column of X is a column of 1's, the second column is comprised of the x_1-variable data, the third column is comprised of the x_2-variable data, and so on. $\vec{y}$ is the column of the y-variable data. Lastly,

$$\vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

If the columns of X are linearly independent (i.e., no perfect multi-collinearity), $X^\dagger$ exists, and our least squares solution is $\hat{\beta} = X^\dagger \vec{y}$. Furthermore,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k$$

This demonstration shows how linear algebra can be used in our field. There are other applications, but I'll save that for another time!

Remember, whatever strength you might have at the IAA can be shared and enjoyed by the rest of the cohort (and beyond). Please share your talents so all can benefit. The collaborative mindset is what sets the IAA apart from other institutions!