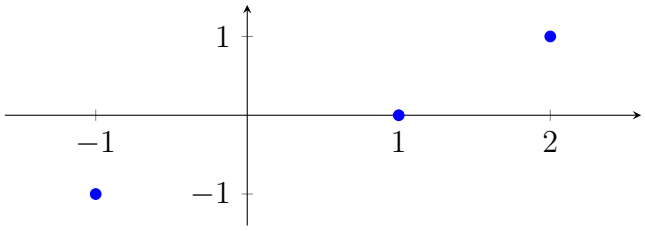
Least Squares Regression **Math 342**

Suppose we have three data points with x and y coordinates given by the following table:

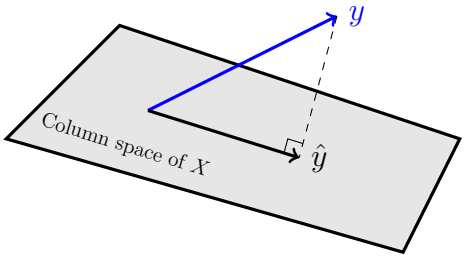
x	y
-1	-1
1	0
2	1



We want to find the linear function $\hat{y} = b_0 + b_1 x$ that is the best fit trendline for the scatterplot. If we **vectorize** this equation, that is treat x and y as vectors containing all of the x and y-values, then $\hat{y} = Xb$ where

$$X = \begin{bmatrix} 1 & x_0 \\ 1 & x_1 \\ 1 & x_2 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}.$$

The goal is to find the value of $\hat{y}$ that is the closest to y. Since $\hat{y}$ is in the column space of X, we should try to find coefficients b_0 and b_1 such that $\hat{y} - y$ is orthogonal to the column space of X.



By the Fundamental Theorem of Linear Algebra, the orthogonal complement of the column space of X is the null space of X^T, so we must have

$$X^T(\hat{y} - y) = 0.$$

By substituting Xb for $\hat{y}$ and rearranging terms, we get the **normal equation** for linear regression:

$$X^T X b = X^T y.$$

1. Calculate $X^T X$ and $X^T y$ for the example above.

2. Use linear algebra to solve the normal equation and find b.

3. Find coefficients b_0, b_1, and b_2 that give the best fit parabola $\hat{y} = b_0 + b_1 x + b_2 x^2$ for the four points $(-2, 3)$, $(-1, 0)$, $(1, -1)$, and $(3, 2)$. To solve this problem, use the normal equations for the matrix

$$X = \begin{bmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \end{bmatrix}.$$

Notice that X is a 4-by-3 Vandermonde matrix. I recommend using Python or Matlab/Octave to solve the normal equations here. In Python, you can enter the Vandermonde matrix using a list comprehension:

```
X = np.array([[x**k for k in range(3)] for x in [-2,-1,1,3]])
```

Then the transpose of X is `X.T` and you can use the `np.linalg.solve()` function to solve the normal equations.