Linear regression -- from the vector projection view point.
Instructor: Nam Sun Wang

In linear regression, we wish to express a dependent variable y as a linear combination of independent variables $x^{<0>}$, $x^{<1>}$, ..., $x^{<m>}$.

$$y = a_0 \cdot x^{<0>} + a_1 \cdot x^{<1>} + \ldots + a_m \cdot x^{<m>}$$

The above equation describes how the various individual scalar variables $x^{<0>}$, $x^{<1>}$, ..., $x^{<m>}$ affect the dependent variable y. It also applies to vectors of given data points $x^{<0>}$, $x^{<1>}$, ..., $x^{<m>}$. Due to measurement noise or an error in the functional relationship, the above equation is not exactly satisfied. Thus, there is an error term.

$$y = a_0 \cdot x^{<0>} + a_1 \cdot x^{<1>} + \ldots + a_m \cdot x^{<m>} + \text{error}$$

The following normal equations minimizes the sum of squared error. $\text{Min} \quad sse = \text{error}^T \cdot \text{error}$

$$a = \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot y$$

where X is a matrix whose columns are the independent variables and whose rows are the data points.

$$X = \left(x^{<0>} \quad x^{<1>} \quad \ldots \quad x^{<m>}\right)$$

$$a = \left[\left(x^{<0>} \quad x^{<1>} \quad \ldots \quad x^{<m>}\right)^T \cdot \left(x^{<0>} \quad x^{<1>} \quad \ldots \quad x^{<m>}\right)\right]^{-1} \cdot \left(x^{<0>} \quad x^{<1>} \quad \ldots \quad x^{<m>}\right)^T \cdot y$$

$$= \left[\begin{bmatrix} x^{<0>T} \\ x^{<1>T} \\ \ldots \\ x^{<m>T} \end{bmatrix} \cdot \left(x^{<0>} \quad x^{<1>} \quad \ldots \quad x^{<m>}\right)\right]^{-1} \cdot \begin{bmatrix} x^{<0>T} \\ x^{<1>T} \\ \ldots \\ x^{<m>T} \end{bmatrix} \cdot y$$

$$= \begin{bmatrix} x^{<0>T} \cdot x^{<0>} & x^{<0>T} \cdot x^{<1>} & \ldots & x^{<0>T} \cdot x^{<m>} \\ x^{<1>T} \cdot x^{<0>} & x^{<1>T} \cdot x^{<1>} & \ldots & x^{<1>T} \cdot x^{<m>} \\ \ldots & \ldots & \ldots & \ldots \\ x^{<m>T} \cdot x^{<0>} & x^{<m>T} \cdot x^{<1>} & \ldots & x^{<m>T} \cdot x^{<m>} \end{bmatrix}^{-1} \cdot \begin{bmatrix} x^{<0>T} \\ x^{<1>T} \\ \ldots \\ x^{<m>T} \end{bmatrix} \cdot y$$

If any two columns of x are perfectly correlated, then the inverse of $X^T \cdot X$ does not exist. On the other hand, if the columns of x are completely mutually independent and uncorrelated, then $x^{<j>T} \cdot x^{<k>} = 0$ for $j \neq k$. In other words, if two vectors $x^{<j>}$ and $x^{<k>}$ are orthogonal, $X^T \cdot X$ is diagonal.

$$x^{<0>T} \cdot x^{<1>} = 0 \qquad x^{<0>T} \cdot x^{<2>} = 0 \qquad \ldots \qquad x^{<0>T} \cdot x^{<m>} = 0$$

$$x^{<1>T} \cdot x^{<0>} = 0 \qquad x^{<1>T} \cdot x^{<2>} = 0 \qquad \ldots \qquad x^{<1>T} \cdot x^{<m>} = 0$$

$$\ldots \qquad \ldots \qquad \ldots$$

$$x^{<m>T} \cdot x^{<0>} = 0 \qquad x^{<m>T} \cdot x^{<2>} = 0 \qquad \ldots \qquad x^{<m>T} \cdot x^{<m-1>} = 0$$

$$a = \begin{bmatrix} x^{\langle 0\rangle T} \cdot x^{\langle 0\rangle} & 0 & \dots & 0 \\ 0 & x^{\langle 1\rangle T} \cdot x^{\langle 1\rangle} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & x^{\langle m\rangle T} \cdot x^{\langle m\rangle} \end{bmatrix}^{-1} \cdot \begin{bmatrix} x^{\langle 0\rangle T} \\ x^{\langle 1\rangle T} \\ \dots \\ x^{\langle m\rangle T} \end{bmatrix} \cdot y$$

$$\begin{bmatrix} a_0 \\ a_1 \\ \dots \\ a_m \end{bmatrix} = \begin{bmatrix} \dfrac{x^{\langle 0\rangle T} \cdot y}{x^{\langle 0\rangle T} \cdot x^{\langle 0\rangle}} \\ \dfrac{x^{\langle 1\rangle T} \cdot y}{x^{\langle 1\rangle T} \cdot x^{\langle 1\rangle}} \\ \dots \\ \dfrac{x^{\langle m\rangle T} \cdot y}{x^{\langle m\rangle T} \cdot x^{\langle m\rangle}} \end{bmatrix} \qquad a_j = \frac{x^{\langle j\rangle T} \cdot y}{x^{\langle j\rangle T} \cdot x^{\langle j\rangle}}$$

Projection of y into x is the scalar product of (y,x) in the direction of x.

$$\text{projection}(y, x) = \frac{(y, x)}{(x, x)} \cdot x$$

Best representation of y in a space spanned by $x^{\langle 0\rangle}$, $x^{\langle 1\rangle}$, ..., $x^{\langle m\rangle}$ is the projection of y into this space.

$$y = \frac{\left(x^{\langle 0\rangle}, y\right)}{\left(x^{\langle 0\rangle}, x^{\langle 0\rangle}\right)} \cdot x^{\langle 0\rangle} + \frac{\left(x^{\langle 1\rangle}, y\right)}{\left(x^{\langle 1\rangle}, x^{\langle 1\rangle}\right)} \cdot x^{\langle 1\rangle} + \dots + \frac{\left(x^{\langle m\rangle}, y\right)}{\left(x^{\langle m\rangle}, x^{\langle m\rangle}\right)} \cdot x^{\langle m\rangle}$$

$$. = \frac{x^{\langle 0\rangle T} \cdot y}{x^{\langle 0\rangle T} \cdot x^{\langle 0\rangle}} \cdot x^{\langle 0\rangle} + \frac{x^{\langle 1\rangle T} \cdot y}{x^{\langle 1\rangle T} \cdot x^{\langle 1\rangle}} \cdot x^{\langle 1\rangle} + \dots + \frac{x^{\langle m\rangle T} \cdot y}{x^{\langle m\rangle T} \cdot x^{\langle m\rangle}} \cdot x^{\langle m\rangle}$$

$$. = a_0 \cdot x^{\langle 0\rangle} + a_1 \cdot x^{\langle 1\rangle} + \dots + a_m \cdot x^{\langle m\rangle} = \sum_j a_j \cdot x^{\langle j\rangle} \quad \text{where} \quad a_j = \frac{x^{\langle j\rangle T} \cdot y}{x^{\langle j\rangle T} \cdot x^{\langle j\rangle}}$$

By inspection, we see that vector projection yields identical expression as the normal equation from linear regression. Thus, the graphical interpretation of least squares linear regression is that it captures y by projecting y into x. Furthermore, when the independent vectors $x^{\langle j\rangle}$ are normalized, the regression coefficients are simply the scalar product of y and $x^{\langle j\rangle}$, which is an indication of how much y has in common with x. Of course, the regression coefficient is 0 when y has nothing in common with $x^{\langle j\rangle}$, when y contains no component in the $x^{\langle j\rangle}$ direction, or, when y is orthogonal to $x^{\langle j\rangle}$. (These are all equivalent statements.)

$a_j = 0$ when $y \perp x^{\langle j\rangle}$: $\left(x^{\langle j\rangle}, y\right) = x^{\langle j\rangle T} \cdot y = 0$