Random variables X and Y are said to be dependent if knowing the value of one affects the distribution of the other

$$\mathbb{P}(Y \mid X = x) \neq \mathbb{P}(Y)$$

In data science terminology, if knowing a prediction x allows you to know something about y, then x and y are associated.
Recall covariance:

$$\text{Cov}[X, Y] = \text{E}[(X - \mu_X)(Y - \mu_Y)]$$

which can be estimated by

$$s_{xy} = \frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y}) \in \mathbb{R}$$

Sign of the Covariance: if x increases and y increases, then covariance is positive; if x increases and y decreases, then covariance is negative. Recall:

$$\rho = \text{Corr}[X, Y] = \frac{\text{Cov}[X, Y]}{\text{SE}[X]\text{SE}[Y]} \in [-1, 1]$$

Correlation is estimated by

$$r = \frac{s_{X,Y}}{s_X s_Y} \in [-1, 1]$$

Thus we say correlation is a standardized covariance.

X,Y are positively correlated if $r > 0$ which means if x increases, then y increases. X,Y are negatively correlated if $r < 0$ which means if x increases, then y decreases. X, Y are not correlated if $r = 0$ which means if x increases, then y is unchanged.

Let $\mathbb{Y} \subseteq \mathbb{R}$ where $p = 1$ and $\mathcal{H} = \left\{\vec{w} \cdot \vec{x} = w_0 + w_1 x : w_0 \in \mathbb{R}, w_1 \in \mathbb{R}\right\}$. Now let $p = 2$ and so

$$\mathcal{H} = \left\{w_0 + w_1 x_1 + w_2 x_2 : \vec{w} \in \mathbb{R}^3\right\}$$

For any $\vec{w}$,

$$SSE = \sum_{\langle \vec{x}_i, y_i \rangle \in \mathcal{D}} \overbrace{(y_i - (w_0 + w_1 x_{i1} + w_2 x_{i2}))^2}^{(y_i - \hat{y}_i)^2}$$

To solve for $\vec{w}$, take

$$\frac{\partial SSE}{\partial w_0} \overset{\text{set}}{=} 0, \quad \frac{\partial SSE}{\partial w_1} \overset{\text{set}}{=} 0, \quad \frac{\partial SSE}{\partial w_2} \overset{\text{set}}{=} 0$$

There is a better method to figure out $\vec{w}$.
Let $\mathcal{D} = \langle X, \vec{y} \rangle$ where

$$X = \begin{bmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \end{bmatrix} \in \mathbb{R}^{n \times p} = \mathbb{R}^{n \times 3}$$

Then

$$X\vec{w} = \begin{bmatrix} w_0 + w_1 X_{11} + w_2 X_{12} \\ w_0 + w_1 X_{21} + w_2 X_{22} \\ \vdots \\ w_0 + w_1 X_{n1} + w_2 X_{n2} \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}$$

This means

$$\vec{\hat{y}} = X\vec{w}$$

Recall the following properties from linear algebra:

$$(\vec{a} + \vec{b})^T = \vec{a}^T + \vec{b}^T$$
$$\vec{a} \cdot \vec{b} = \vec{b} \cdot \vec{a}$$
$$(AB)^T = B^T A^T$$
$$\forall \vec{v} \in \mathbb{R}^d, \ \vec{v} \cdot \vec{v} = \sum_{j=1}^{d} v_j^2 = \vec{v}^T \vec{v}$$

Then

$$\begin{aligned} SSE &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= (\vec{y} - \vec{\hat{y}})^T (\vec{y} - \vec{\hat{y}}) = (\vec{y}^T - \vec{\hat{y}}^T)(\vec{y} - \vec{\hat{y}}) \\ &= \vec{y}^T \vec{y} - \vec{\hat{y}}^T \vec{y} - \vec{y}^T \vec{\hat{y}} + \vec{\hat{y}}^T \vec{\hat{y}} \\ &= \vec{y}^t \vec{y} - 2\vec{\hat{y}}^T \vec{y} + \vec{\hat{y}}^T \vec{\hat{y}} \\ &= \vec{y}^T \vec{y} - 2(X\vec{w})^T \vec{y} + (X\vec{w})^T (X\vec{w}) \\ &= \vec{y}^T \vec{y} - 2\vec{w}^t X^T \vec{y} + \vec{w}^T X^T X \vec{w} \end{aligned}$$

Now take the partial derivative with respect to $\vec{w}$ and set it equal to $\vec{0}_{p+1}$ (vector derivative).

$$\frac{\partial SSE}{\partial \vec{w}} = \begin{bmatrix} \frac{\partial}{\partial w_0} SSE \\ \vdots \\ \frac{\partial}{\partial w_p} SSE \end{bmatrix} = \vec{0}_{p+1}$$

Properties:

- For a constant $a \in \mathbb{R}$ and $\vec{c} \in \mathbb{R}^n$,

$$\frac{\partial}{\partial \vec{c}} a = \begin{bmatrix} \frac{\partial}{\partial c_1} a \\ \vdots \\ \frac{\partial}{\partial c_n} \end{bmatrix} = \vec{0}_n$$

- For $\vec{c} \in \mathbb{R}^n$,

$$\begin{aligned}
\frac{\partial}{\partial \vec{c}}(af(\vec{c}) + g(\vec{c})) &= \begin{bmatrix} \frac{\partial}{\partial c_1}(af(\vec{c}) + g(\vec{c})) \\ \vdots \\ \frac{\partial}{\partial c_n}(af(\vec{c}) + g(\vec{c})) \end{bmatrix} \\
&= \begin{bmatrix} a\frac{\partial}{\partial c_1}f(\vec{c}) + \frac{\partial}{\partial c_1}g(\vec{c}) \\ \vdots \\ a\frac{\partial}{\partial c_n}f(\vec{c}) + \frac{\partial}{\partial c_n}g(\vec{c}) \end{bmatrix} \\
&= a\frac{\partial}{\partial \vec{c}}f(\vec{c}) + \frac{\partial}{\partial \vec{c}}g(\vec{c})
\end{aligned}$$

- For $\vec{c} \in \mathbb{R}^n$ and $\vec{b} \in \mathbb{R}^n$,

$$\frac{\partial}{\partial \vec{c}}\ \vec{c}^T\vec{b} = \frac{\partial}{\partial \vec{c}}(c_1b_1 + c_2b_2 + \cdots + c_nb_n) = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} = \vec{b}$$

- For $A \in \mathbb{R}^{n\times n}$ and $\vec{c} \in \mathbb{R}^n$ and A symmetric, note first that

$$A\vec{c} = \begin{bmatrix} a_{11}c_1 + a_{12}c_2 + \cdots + a_{1n}c_n \\ a_{21}c_1 + a_{22}c_2 + \cdots + a_{2n}c_n \\ \vdots \\ a_{n1}c_1 + a_{n2} + \cdots + a_{nn}c_n \end{bmatrix}$$

  and that

$$\vec{c}^T(A\vec{c}) = c_1(a_{11}c_1+a_{12}c_2+\cdots+a_{1n}c_n)+\cdots+c_n(a_{n1}c_1+a_{n2}c_2+\cdots+a_{nn}c_n) = \sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij}c_ic_j$$

  Then

$$\frac{\partial}{\partial \vec{c}}\ \vec{c}^T(A\vec{c}) = \begin{bmatrix} \frac{\partial}{\partial c_1}A\vec{c} \\ \vdots \\ \frac{\partial}{\partial c_n}A\vec{c} \end{bmatrix} = \begin{bmatrix} 2\vec{a}_{1\cdot}^T\vec{c} \\ \vdots \\ 2\vec{a}_{n\cdot}^T\vec{c} \end{bmatrix} = 2\begin{bmatrix} \vec{a}_{1\cdot} \\ \vdots \\ \vec{a}_{n\cdot} \end{bmatrix}\vec{c} = 2A\vec{c}$$

Now let's apply this to SSE.

$$\begin{aligned}
\frac{\partial}{\partial \vec{w}}SSE &= \frac{\partial}{\partial \vec{w}}(\vec{y}^T\vec{y} - 2\vec{w}^TX^T\vec{y} + \vec{w}^TX^TX\vec{w}) \\
&= \frac{\partial}{\partial \vec{w}}(\vec{y}^T\vec{y}) - 2\frac{\partial}{\partial \vec{w}}(\vec{w}^TX^T\vec{y}) + \frac{\partial}{\partial \vec{w}}(\vec{w}^TX^TX\vec{w}) \\
&= \vec{0} - 2X^T\vec{y} + 2X^TX\vec{w} \overset{\text{set}}{=} 0 \\
X^T\vec{y} &= X^TX\vec{w} \\
\vec{w} = \vec{b} &= (X^TX)^{-1}X^T\vec{y}
\end{aligned}$$

Note that X^TX is of dimension $\mathbb{R}^{p+1\times p+1}$. It is invertible only when X^TX is of full rank $p+1$ (linearly independent), or $\text{rank}(X) = p+1$.

Proof by Contradiction: Assume rank$(X^TX) = p+1$ and rank$(X) < p+1$. Then there is a non-trivial null space, meaning a vector $\vec{u} \neq \vec{0}$ and in $\mathbb{R}^{p+1}$ that can be mapped to $\vec{0}$. This means $X\vec{u} = \vec{0}_n$. But then if we use X^TX to map vector $\vec{u}$,

$$(X^TX)\vec{u} = X^T(X\vec{u}) = X^T\vec{0}_n = \vec{0}_{p+1}$$

This means $\vec{u}$ is in the null space of X^TX. Therefore the dimension of the null space of X^TX is greater than 0. Then X^TX is not full rank. Contradiction.

When we say rank$(X) = p+1$, we mean that each column is not linearly dependent on other columns. Therefore no predictive information is duplicated.

Henceforth,

$$\vec{\hat{y}} = X\vec{b} = X(X^TX)^{-1}X^T\vec{y} = \vec{H}\vec{y}$$

where $\vec{H} = X(X^TX)^{-1}X^T$ is a hat matrix.