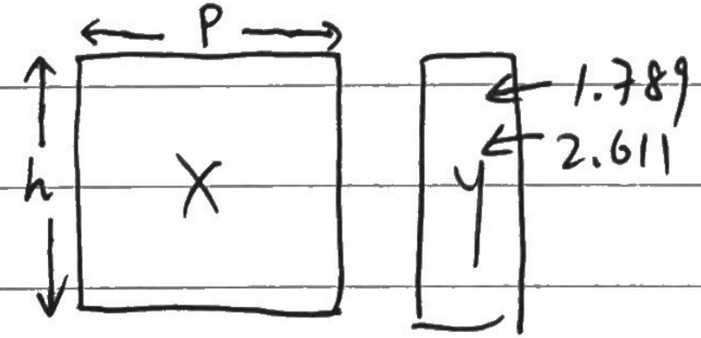
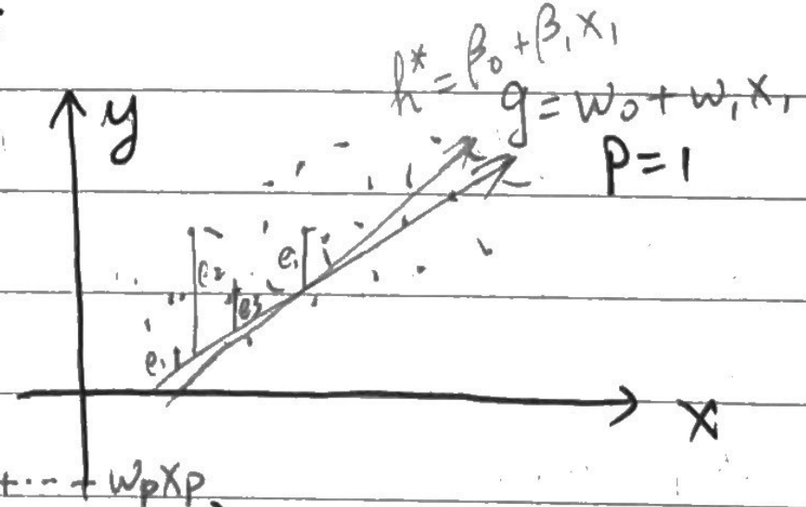
So Far, the response space was $\{0,1\}$, and the models were "binary classification" models. What if $y = \mathbb{R}$ or $y \subset \mathbb{R}$? This means the response is continuous and our predicti ones will be continuous. These models are called "regression" models. The word "regression" is used because of historical circumstance only. (see lab)

What is null model $g_0$? $g_0 = \bar{y}$




$\mathcal{H} = \{\vec{w} \cdot \vec{x} : \vec{w} \in \mathbb{R}^{P+1}\}$ let before, this candidate set, requires, a "1" appended to each of the original P-length x-vectors.

$$h^*(x) = w_0^* + w_1^* x_1 + \dots + w_p^* x_p = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

standard notation for the best / "true" values of the linear coefficients.

$$\Rightarrow y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon$$

We have training data and the candidate set of linear models. We need an algorithm that will compute $w_0$ and $w_1$ for us. We first need an "objective function" or "error function" or "loss function" which gauges the degree of our model mistakes. Let $e_i := y_i - \text{yhat}_i$. Consider the loss function: $\sum_{i=1}^{n} e_i^2 = SSE$ (sum of squared error)

$$= \sum (y_i - \hat{y}_i)^2 = \sum (y_i - w_0 - w_1 x_i)^2$$

Our algorithm will seek to argmin{SSE} over all possible $w_0, w_1$ values. To do this, we take the partial derivative with respect to $w_0$ and set equal to zero and solve for $b_0$ then take the partial derivative wrt. $w_1$ and set equal to zero and solve for $b_1$. We will call $g(x) = b_0 + b_1 X$ the "least squares" regression model or "ordinary least squares" (OLS)

$$\sum (y_i^2 + w_0^2 + w_1^2 x_i^2 - 2 y_i w_1 x_i - 2 y_i w_0 + 2 w_0 w_1 x_i)$$

$$= \sum y_i^2 + n w_0^2 + w_1^2 \sum x_i^2 - 2 w_0 n \bar{y} - 2 w_1 \sum x_i y_i + 2 w_0 w_1 n \bar{x}$$

$$\frac{\partial}{\partial w_0}[SSE] = 2n w_0 - 2n\bar{y} + 2 w_1 n \bar{x} \overset{set}{=} 0$$

$$\Rightarrow b_0 = \frac{n\bar{y} - w_1 n \bar{x}}{n} = \boxed{\bar{y} - b_1 \bar{x}}$$

$$\frac{\partial}{\partial w_1}[SSE] = 2 w_1 \sum x_i^2 - 2 \sum x_i y_i + 2 w_0 n \bar{x} \overset{set}{=} 0$$

$$\Rightarrow w_1 \sum x_i^2 = \sum x_i y_i - w_0 n \bar{x} = \sum x_i y_i - (\bar{y} - b_1 \bar{x}) n \bar{x}$$

$$\Rightarrow b_1 \sum x_i^2 = \sum x_i y_i - n \bar{x} \bar{y} + n \bar{x}^2 b_1$$

$$b_1 \left(\sum x_i^2 - n \bar{x}^2\right) = \sum x_i y_i - n \bar{x} \bar{y}$$

$$b_1 = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2}$$

$\Rightarrow$

$$b_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

$$S_x^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum x_i^2 - 2\bar{x}\underbrace{\sum x_i}_{n\bar{x}} - n\bar{x}^2\right)$$

$$= \frac{1}{n-1}\left(\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right) = \frac{1}{n-1}\left(\sum x_i^2 - n\bar{x}^2\right)$$

$$e := cov[x,y] := \frac{cov[X,Y]}{SE[x]\,SE[y]} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sqrt{Var[Y]\,Var[X]}}$$

Covariance is estimated with

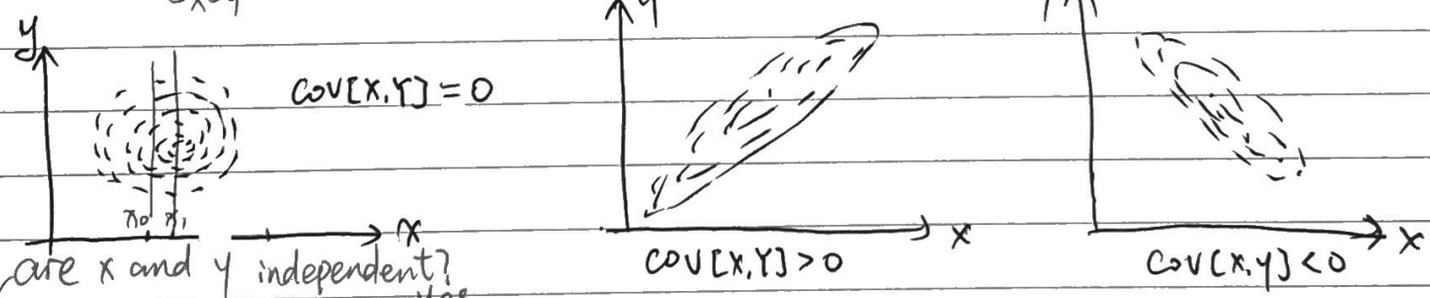
$$S_{x,y} = \frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum x_i y_i - \bar{y}\sum x_i - \bar{x}\sum y_i + n\bar{x}\bar{y}\right)$$

$$= \frac{1}{n-1}\left(\sum x_i y_i - \bar{y}n\bar{x} - \bar{x}n\bar{y} + n\bar{x}\bar{y}\right)$$

$$= \frac{1}{n-1}\left(\sum x_i y_i - n\bar{x}\bar{y}\right)$$

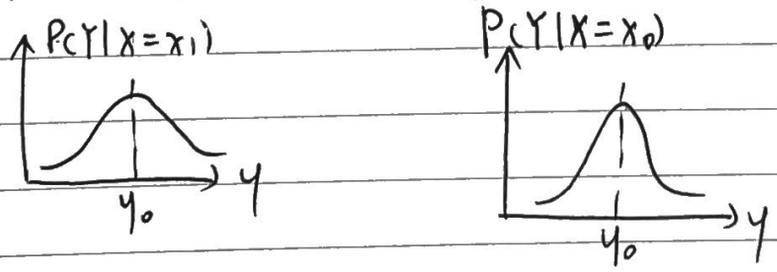
$$\Rightarrow b_1 = \frac{(n-1)\,S_{x,y}}{(n-1)\,S_x^2} = \frac{S_{x,y}}{S_x^2} \Rightarrow \frac{rS_xS_y}{S_x^2} = \boxed{\frac{rS_y}{S_x}} \Rightarrow b_0 = \boxed{\bar{y} - r\frac{S_y}{S_x}\bar{x}}$$

$$r := \frac{S_{x,y}}{S_xS_y} \Rightarrow S_{xy} = r\,S_x\,S_y$$




are x and y independent? yes

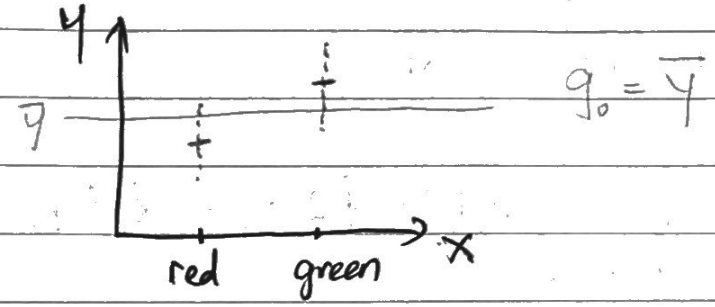
Covariance measures change in expected value of the second r.v if the first r.v changes.




The word "association" just means "dependence". Correlation means linear dependence (and covariance means linear dependence). Correlation is a type of association (it is linear association).

Let's examine a special case of OLS where $P=1$. Let the only feature be a binary feature e.g. $x_i$ is either red or green. Let's create a new $x$ which is a dummy / binary variable which is 0 if red and 1 if green. What is a good model for prediction?

$g(\text{red}) = \bar{y}_{\text{red}}$
$g(\text{green}) = \bar{y}_{\text{green}}$
} OLS model




ex.

| x | y |
|---|---|
| 0 | 3.71 |
| 1 | 8.43 |
| 0 | 6.72 |
| 1 | 1.87 |
| ⋮ | 7.11 |
| | 7.98 |
| | ⋮ |