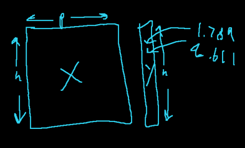
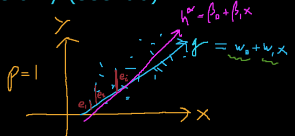
So far, the response space was {0,1} and the models were "binary classification" models. What if $\mathcal{Y} = \mathbb{R}$ or $\mathcal{Y} \subset \mathbb{R}$? This means the response is continuous and our predictions will be continuous. These models are called "regression" models. The word "regression"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

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





$$\mathcal{H} = \left\{ \vec{w} \cdot \vec{x} \;:\; \vec{w} \in \mathbb{R}^{p+1} \right\}$$

$$\vec{w} \cdot \vec{x} = w_0 + w_1 x_1 + \ldots + w_p x_p$$

Like before, this candidate set, requires, a "1" appended to each of the original p-length x-vectors.

$$h^*(\vec{x}) = w_0^* + w_1^* x_1 + \ldots + w_p^* x_p = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

$$\Downarrow$$

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \mathcal{E}$$

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

We have p=1 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 - yhat_i. Consider the loss function:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - w_0 - w_1 x_i)^2$$

(sum of squared error)

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 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).

$$\bar{y} = \frac{1}{n} \sum y_i$$

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

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

$$\frac{\partial}{\partial w_0}[SSE] = \frac{\partial}{\partial w_0}[\;\;] = \cancel{2} n w_0 - \cancel{2} n \bar{y} + \cancel{2} w_1 n \bar{x} \overset{set}{=} 0 \Rightarrow b_0 = \frac{n\bar{y} - w_1 n \bar{x}}{n} = \bar{y} - b_1 \bar{x}$$

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

$$\Rightarrow b_1 \sum x_i^2 = \sum x_i y_i - b_0 n \bar{x} = \sum x_i y_i - (\bar{y} - b_1 \bar{x}) n \bar{x}$$

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

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

this is the answer and now we simplify it using Math 241-like notation:

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

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

$$\rho := \text{Corr}[X,Y] := \frac{\text{Cov}[X,Y]}{SE[X]\,SE[Y]} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{\text{Var}[Y]\,\text{Var}[X]}}$$

$$r := \frac{s_{xy}}{s_x s_y}$$

$$\Downarrow$$

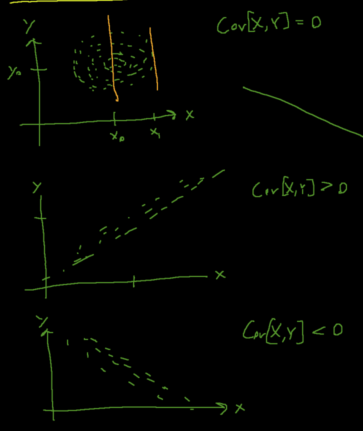
$$s_{xy} = r s_x s_y$$

Covariance is estimated with

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

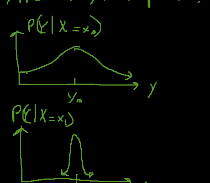
$$= \frac{1}{n-1}\left(\sum x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}\right) = \frac{1}{n-1}\left(\sum x_i y_i - n\bar{x}\bar{y}\right)$$

$$b_1 = \frac{(n-1) s_{xy}}{(n-1) s_x^2} = \frac{s_{xy}}{s_x^2} = \frac{r s_x s_y}{s_x^2} = r \frac{s_y}{s_x} \Rightarrow b_0 = \bar{y} - r \frac{s_y}{s_x} \bar{x}$$



covariance measures change in expected value of the second rv if the first rv changes

Are X,Y independent?



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". Lets 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?

$$\left.\begin{array}{l} g(\text{red}) = \bar{y}_{\text{red}} \\ g(\text{green}) = \bar{y}_{\text{green}} \end{array}\right\} \text{OLS model}$$

$$g_0 = \bar{y}$$

$$\begin{array}{cc} x & y \\ \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ \vdots \end{bmatrix} & \begin{bmatrix} 3.71 \\ 8.43 \\ 6.12 \\ 1.07 \\ 7.11 \\ 7.18 \\ \vdots \end{bmatrix} \end{array}$$