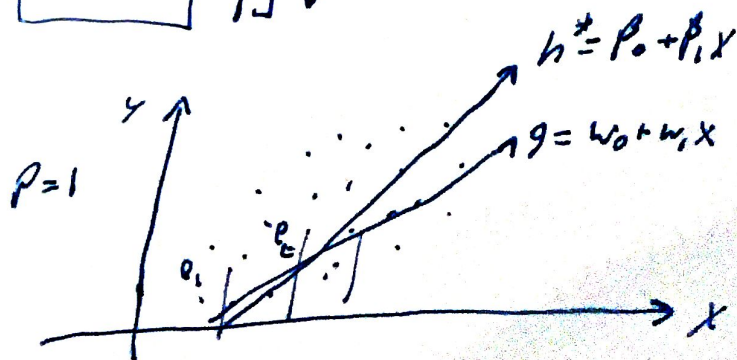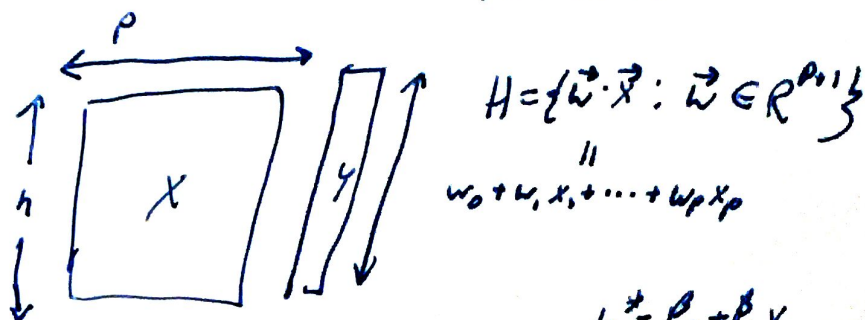# Lecture 06

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 predictions will be continuous. These models are called "regression" models. The word "regression" is used because of historical circumstances

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

$$H = \{\vec{w} \cdot \vec{x} : \vec{w} \in \mathbb{R}^{p+1}\}$$
$$\|$$
$$w_0 + w_1 x_1 + \cdots + w_p x_p$$

$$h^* = \beta_0 + \beta_1 x$$

$$g = w_0 + w_1 x$$

$p = 1$

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

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

Standard notation for the best / true value of the linear coefficient

we have $p=1$ training data and the candidate set of linear models. we need an algorithm that will compute w-o 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 $e_i := y_i - \hat{y}_i$. consider the loss function:

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

$\underbrace{\qquad}$ 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 solve for b-0 then take the partial derivative w.r.t w-1 and set equal to zero and solve for b-1. we will call $g(t) = b_0 + b_1 x$ the "least squares" regression model or "ordinary least squares" (OLS)

$$\{ y_i^2 + b_0^2 + w_1^2 + {}^2 - 2 y_i w_0 - 2 y_i w_1 x_i + 2 w_0 - x_i \}$$

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

$$\frac{\partial}{\partial w_0} [SSE] = \frac{\partial}{\partial w_0} [\quad] = 2n w_0 - 2n \bar{y} + 2 w_1 n \bar{x} \overset{set}{=} 0 = 7 \, b_0$$

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

$$\frac{\partial}{\partial b_1} [SSE] = \frac{2}{\partial b_1} [\quad] = 2 w_1 \sum x_i^2 - 2 \sum x_i y_i + 2 b_0 n \bar{y} \overset{x^*}{=} 0$$

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

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

$$\Rightarrow b_1 = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2}$$ this can be simplified using

Math 241

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

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

$$e = corr[x,y] = \frac{cov[x,y]}{SE[x]\,SE[y]} = \frac{E[(x-\mu_x)(y-\mu_y)]}{\sqrt{var[y]\,var[x]}}$$

Covariance is estimated with

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

$$= \frac{1}{n-1} \left(\sum x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y} \right) = \frac{1}{n-1} \left(\sum x_i y_i - n\bar{x}\bar{y} \right)$$
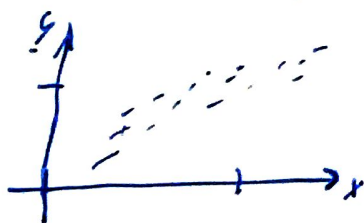
$$b_1 = \frac{(n-1)\,S_{xy}}{(n-1)\,S_x^2} = \frac{S_{xy}}{S_x^2} = \frac{r\,S_x\,S_y}{S_x^2} = r\frac{S_y}{S_x} \implies b_0 = \bar{y} - r\frac{S_y}{S_x}\bar{x}$$

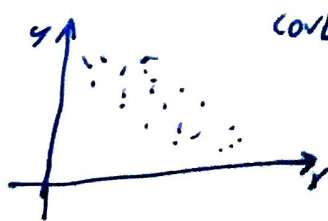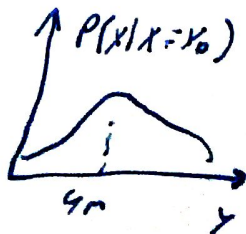Covariance measures change in expected value of the second rv if the first rv changes
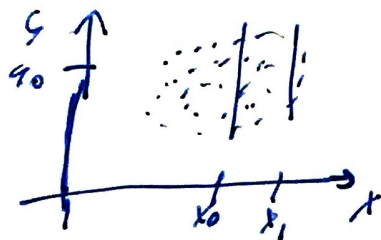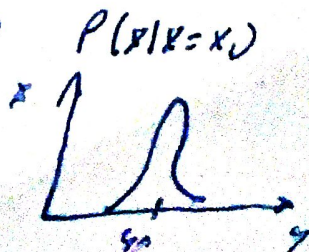

$$Cov[x,y] = 0$$

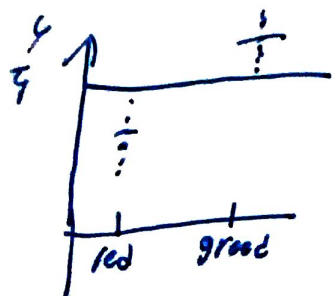Are x,y independent

$$Cov[x,y] > 0$$

$P(y|x=y_0)$

$$Cov[x,y] < 0$$

$P(y|x=x_i)$

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 model for prediction:

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

$$g_0 = \bar{y}$$

$$X \quad \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ \vdots \end{bmatrix} \quad \overset{y}{\begin{bmatrix} 3.71 \\ 6.47 \\ 6.72 \\ 1.07 \\ 7.11 \\ 7.22 \\ \cdots \end{bmatrix}}$$