Math 390 Lec 5 2/14/19

So far $\mathcal{Y} = \{0, 1\}$. Models built for this output space are called "binary classification".

$\mathcal{H} = \{\mathbb{1}_{\vec{w} \cdot \vec{x} \geq 0} : \vec{w} \in \mathbb{R}^{p+1}\}$ the indicator forced the function to return 0 or 1.

What if $\mathcal{Y} = \mathbb{R}$ or $\mathcal{Y} \subset \mathbb{R}$, a continuous response.

This is called "regression". Why? Historical reasons...

What is null model $g_0 = \bar{y}$ $\mathcal{H} = \mathbb{R}$

Our first regression model:

$\mathcal{H} = \{\vec{w} \cdot \vec{x} : \vec{w} \in \mathbb{R}^{p+1}\}$ where first entry of $\vec{x}$ is a 1.

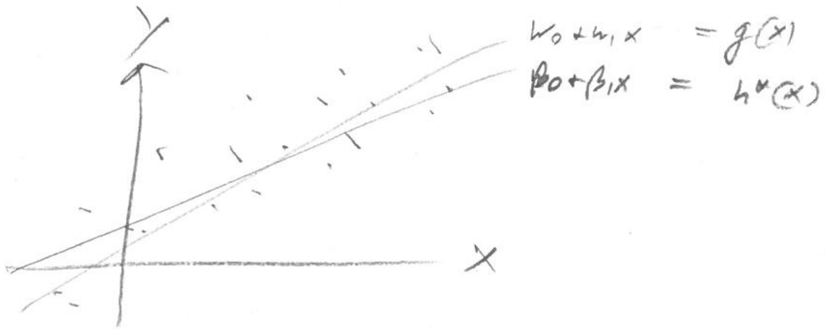
$= \{w_0 + w_1 x_1 + \ldots + w_p x_p : w_0, \ldots, w_p \in \mathbb{R}\}$

$h^*(\vec{x}) = w_0^* + w_1^* x_1 + \ldots + w_p^* x_p$ where the $\vec{w}^*$ is the best candidate model $\in \mathcal{H}$

$= \underbrace{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}_{\text{special notation for linear model}}$

$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$ looks familiar now to ECON 382 people

If $p=1$, we can visualize



$h^*(x) \neq y$ since there are two other errors

How to fit w_0 & w_1? Need an error function / loss function

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

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

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

Choose w_0, w_1 to minimize SSE

$$\frac{\partial}{\partial w_0}[SSE] \overset{set}{=} 0 \Rightarrow 2nw_0 - 2n\bar{y} + 2w_1 n\bar{x} = 0$$

$$\Rightarrow w_0 - \bar{y} + w_1\bar{x} = 0$$

$$\Rightarrow \hat{w}_0 = \bar{y} - w_1\bar{x}$$

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

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

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

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

$$\Rightarrow \hat{w}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$ this's the answer... but we can simplify...

Recall Var(X) is estimated by

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

Math 241

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

Cov(X,Y) is estimated by "sample cov."

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

Corr(X,Y) est by r, "sample corr"

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

$$r = \frac{s_{xy}}{s_x s_y} \cdot \frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2}$$

$$\Rightarrow \hat{w}_1 = \frac{(n-1)\, s_{xy}}{(n-1)\, s_x^2} = \frac{s_{xy}}{s_x^2} = r\frac{s_y}{s_x}$$

$$\hat{w}_0 = \bar{y} - r\frac{s_y}{s_x}\bar{x}$$

Beautiful analytically-derived formulas! That's rare!!

$\hat{w}_0, \hat{w}_1$ has special notation: $\underbrace{b_0, b_1}_{\text{use this}}$ or $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\Rightarrow g(x) = b_0 + b_1 x$$

"Ord. Least Squares" estimates, OLS estimates