Properties of Normal random variables (ch 4.6)

$X \sim \text{Normal}(\mu, \sigma^2)$ is a normal r.v. and has density

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

Shorthand notation: $X \sim N(\mu, \sigma^2)$

Combining Normal r.v.s (theorem 4.6.1)

Let $X_i \sim \text{Normal}(\mu_i, \sigma_i^2)$ be independent

then

$$Y = \sum_{i=1}^{N} X_i \sim \text{Normal}\left(\sum_{i=1}^{N} \mu_i, \sum_{i=1}^{N} \sigma_i^2\right)$$

- might talk about theorem 4.6.2 later
- will talk about chi-squared in week 4

Linear Regression Model (Ch 10.)

related variables (Def 10.1.1) = not independent

linear regression w/ one predictor (Ex 10.1.1)

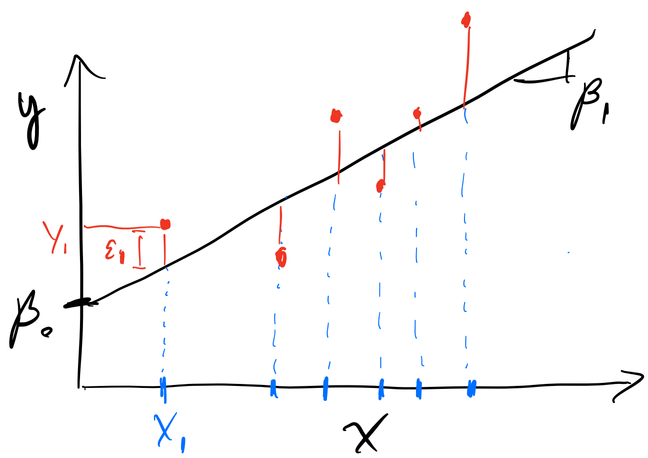
Let X be any r.v. and

$$Y \mid X \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

Can also write

"error" or "residual" term

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$



Very important model because simple linear relationship + Normal error $\Rightarrow$ we can easily estimate $\beta_0, \beta_1, \sigma^2$

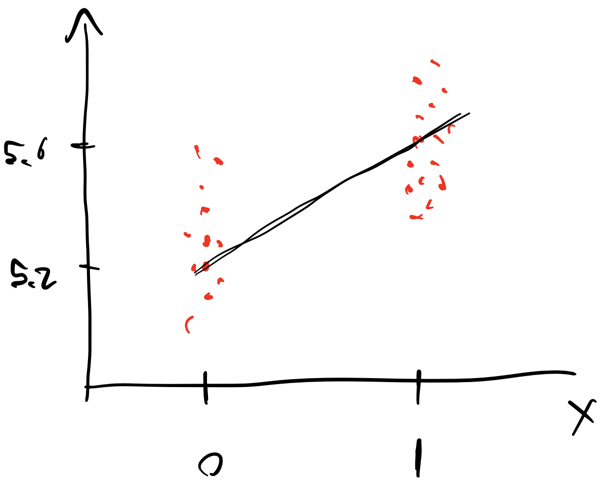
X is called predictor

Y is called response variable

Example (model of hieght)

$X \sim \text{Bernoulli}(1/2)$ (1 = male, 0 = female)

$Y \mid X \sim N(5.2 + 0.4X, 2.5^2)$ (hight in inches)



Note: $\mathbb{E}[Y \mid X=0] = 5.2 = \beta_0$

$\mathbb{E}[Y \mid X=1] = 5.6 = \beta_1$

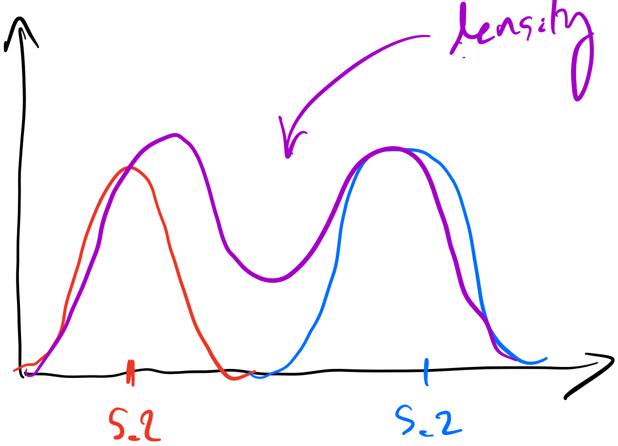
$\text{var}(Y \mid X=1) = 2.6$

$\Rightarrow$ in order to estimate β_0, β_1, σ we merely need to compute mean and variance within each group

what about marginal distribution of Y?

this is not normal:

density of Y has two bumps $\Rightarrow$ not normal



Suppose X has mean and var μ_X, σ_X^2

$$Y \mid X \sim \text{Normal}(\beta_0 + \beta_1 X, \sigma_X^2)$$

Q: what is $E[XY]$?

will give solution in class

we find $\mathbb{E}[XY] = \beta_1 \sigma_X^2 + \beta_1 \mu_X^2 + \beta_0 \mu_X$

$$\mathbb{E}[X]\mathbb{E}[Y] = \mu_X \mu_Y$$
$$= \beta_1 \mu_X^2 + \beta_0 \mu_X$$

$$\Longrightarrow \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \beta_1 \sigma_X^2$$

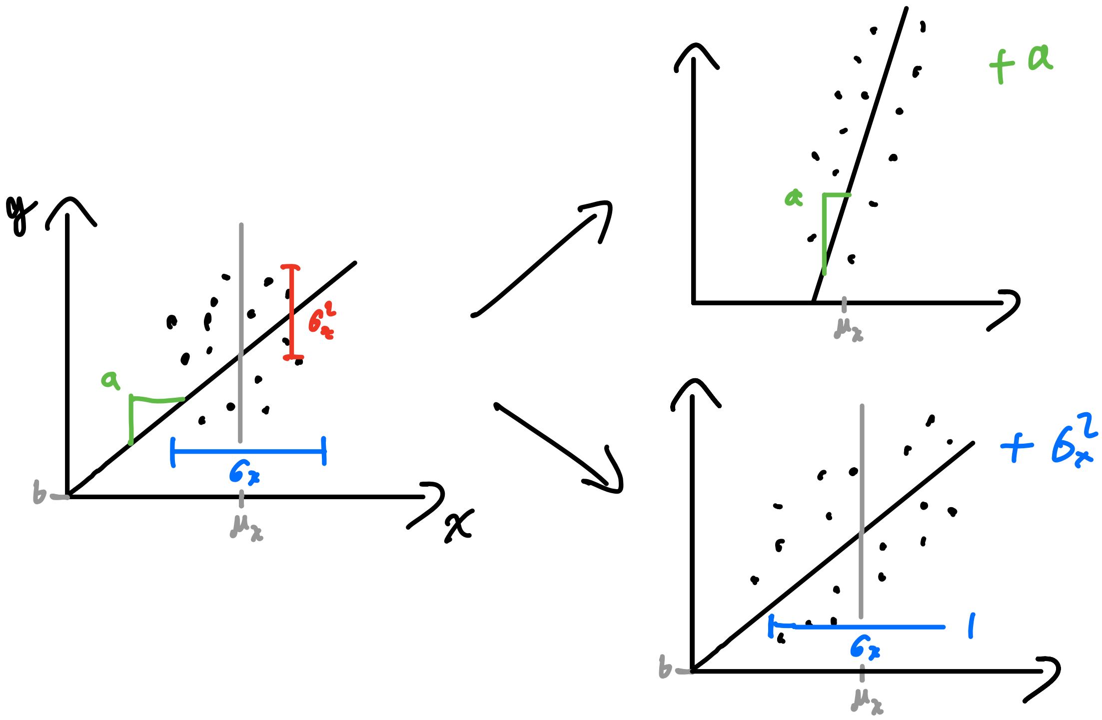
Covariance (Def 3.3.3)

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

theorem 3.3.3

We just saw $\mathrm{Cov}(Y, X) = \beta_1 \sigma_X^2$ (★)

Cov is measure of association between X, Y



(★) suggests way to estimate slope from samples $(X_1, Y_1) \ldots (X_N, Y_N)$:

$$\beta_1 = \frac{\mathrm{Cov}(Y, X)}{\sigma_X^2} \approx \frac{\overline{(X-\mu_X)(Y-\mu_Y)}}{\overline{(X-\mu_X)^2}}$$

$$= \underbrace{\frac{\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{(X_i - \bar{X})^2}}_{\hat{\beta}_1}$$

hat "^" means estimate of, like "—" when we write sample avg.

then we can estimate β_0 via

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1$$

these are formulas in theorem 10.3.1 derived differently

I will also talk about R^2, correlation coefficient.