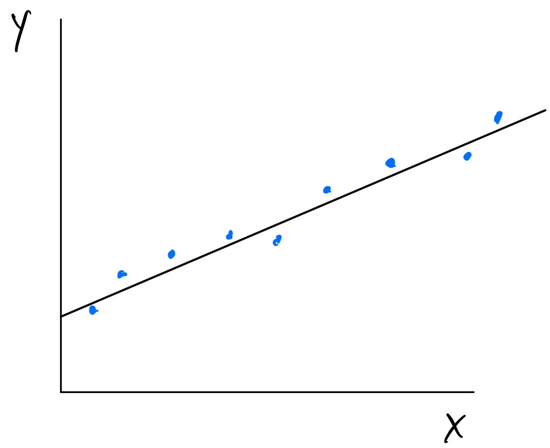
Coefficient of determination

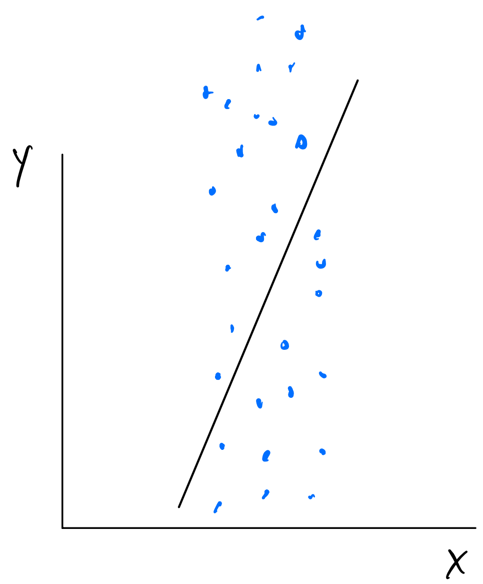
$$Y \mid X \sim \text{Normal}(\beta_0 + \beta_1 X, \sigma^2)$$

Q: How "well" does does X predict Y?

↳ need a way of quantify association which does not depend on units because could have a <u>huge</u> slope but if σ_X^2 is small and σ^2 is large.

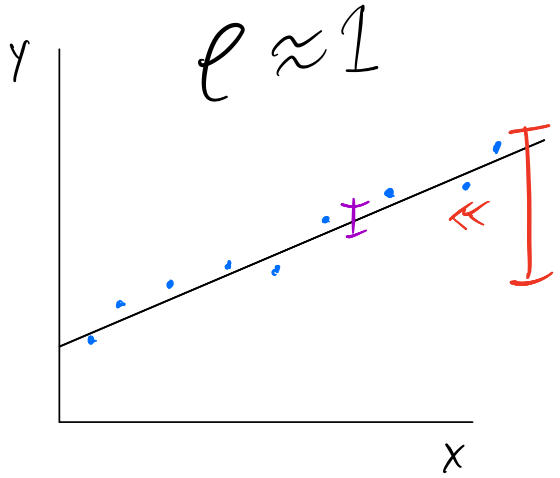


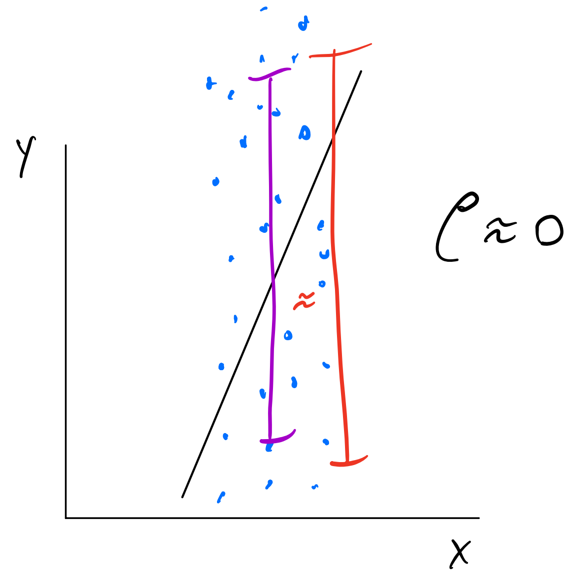
vs.



<u>Idea</u>: look at relation between overall Y variation and conditional Y variation:

$$\rho^2 = 1 - \frac{\sigma^2}{\sigma_Y^2} = 1 - \frac{\sigma^2}{\sigma^2 + \beta_1^2 \sigma_X^2} = \frac{\cancel{\sigma^2} + \beta_1^2 \sigma_X^2 - \cancel{\sigma^2}}{\sigma_Y^2} = \frac{\beta_1^2 \sigma_X^2}{\sigma_Y^2}$$





<u>remember:</u> $\mathrm{Cov}(Y, X) = \beta_1 \sigma_X^2$

$$\Longrightarrow \rho^2 = \frac{\mathrm{Cov}(Y, X)^2}{\sigma_X^2 \, \sigma_Y^2}$$

Estimating this gives "R-squared"

Statistical interecne (Ch4 in regression and other stories. See referenaces in ER for more details)

So far: We talked about how to estimate mean, variance etc.

Now we want to quantify uncertinty in these estimates

Let $Y \sim Normal(\mu, \sigma^2)$

Given data $Y_1, Y_2, \ldots Y_N$ we can estimate

$$\hat{\mu} = \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$ (see Example 5.5.6 in ER)

We define the sample distribution of an estimator as the distribution over many replications of our dataset (in this case $Y_1, \ldots Y_N$)

In the example above: $\hat{\mu}$ is normal with

$$\mathbb{E}[\hat{\mu}] = \frac{1}{N} \cdot N \mathbb{E}[Y_i] = \mu_Y$$

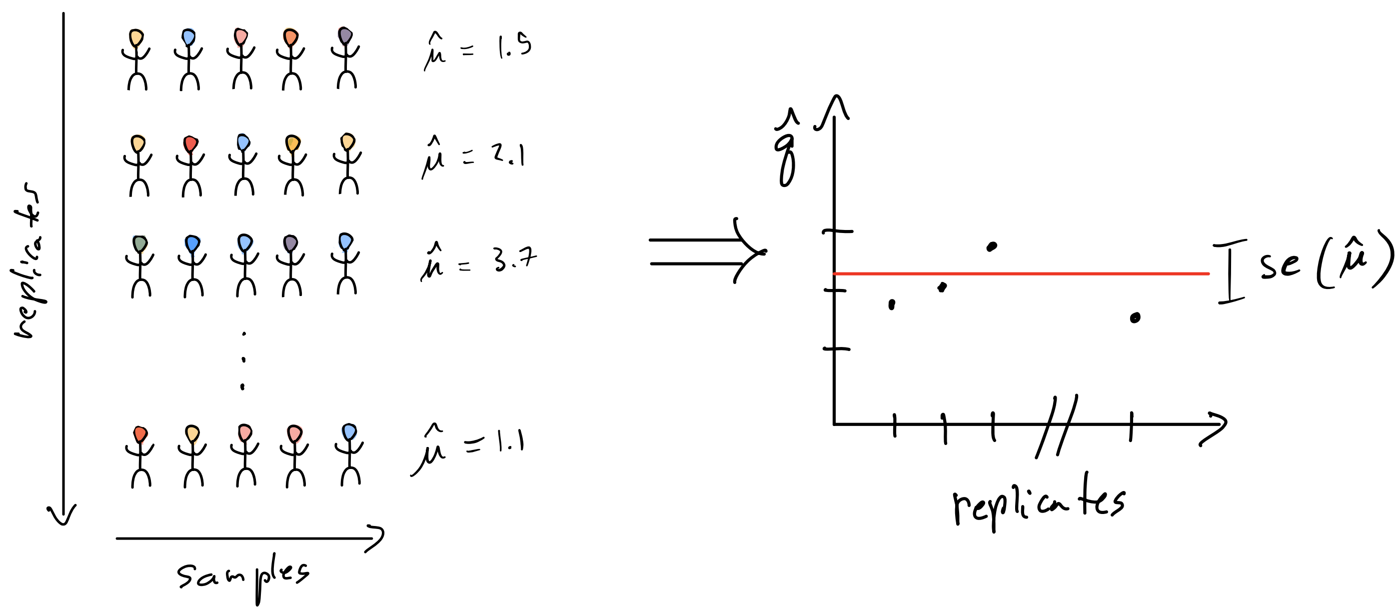
$$Var(\hat{\mu}) = N\, var(Y_i) = \sigma_Y^2/N$$

Hence the sample distribution is

$$\hat{\mu} \sim Normal(\mu_Y, \sigma_Y^2/N)$$

The standard error (se) is the standard deviation of the sample distribution. In the example above

$$se(\hat{\mu}) = \sigma_Y^2 / \sqrt{N}$$ (See section 6.3.1 in EK)



Example $X \sim \text{Bernoulli}(q)$

$$\hat{q} = \bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

by CLT $\hat{q} \sim \text{Normal}\left(q, \frac{q(1-q)}{N}\right)$

Properties of Estimators

Let $\hat{\Theta}$ be an estimator of a parameter Θ using N samples

<u>Unbiased</u> if $\mathbb{E}[\hat{\Theta}] = \Theta$ (Def 6.3.2 in EK)

<u>Consistent</u> if $se(\hat{\Theta}) \longrightarrow 0$ as $N \to \infty$ (Def 6.3.3 in EK)

Example

	Unbiased	Consistent
$\hat{\mu} = \overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$	✓	✓
$\hat{\mu}_1 = \frac{1}{N}\sum_{i=1}^{N} Y_i + \frac{1}{N}$	✗	✓
$\hat{\mu}_2 = \frac{1}{2}(Y_1 + Y_2)$	✓	✗

Estimating variance

$$Y \sim \text{Normal}(0, \sigma^2)$$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} Y_i^2$$

$$\mathbb{E}[\hat{\sigma}^2] = \frac{1}{N} \cdot N\, \mathbb{E}[Y^2] = \text{Var}(Y) = \sigma^2$$

Now suppose Y has (unknown) mean $\mu_Y > 0$

$$\hat{\sigma}_1^2 = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \bar{Y})^2$$

$$\mathbb{E}[\hat{\sigma}_1^2] = \mathbb{E}[(Y - \bar{Y})^2] \neq \mathbb{E}[(Y - \mu_Y)^2] = \text{Var}(Y)$$

Biased ... but

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2 \quad \text{is unbiased}$$