Quantifying association.

Given random variables X and Y. How to measure the association between X and Y?

① Correlation coefficients.

Pearson correlation coefficient. $\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}}$.

② Regression.

2a. Simple linear regression:

Write $Y = \alpha + \beta X + \varepsilon$, where

$\mathbb{E}[\varepsilon] = 0$, $Cov(\varepsilon, X) = 0$.

Rem: This is not a model. Instead, this is a way of defining β.

Fact: $\beta = \frac{Cov(X,Y)}{Var(X)}$.

Pf. $0 = Cov(\varepsilon, X)$

$= Cov(Y - \alpha - \beta X, X)$.

$= Cov(Y, X) - \beta Cov(X, X)$.

$\Rightarrow \beta \underbrace{Cov(X,X)}_{Var(X)} = Cov(Y, X)$. ∎

We can interpret β as measuring the association

of Y with X.

Fact: $\beta = \rho_{XY} \cdot \sqrt{\frac{Var(Y)}{Var(X)}}$.

3b. Multiple linear regression.

E.g. Y = income, X = years of schooling.

W = county of residence.

$Y = \alpha + \beta X + \gamma W + \varepsilon$,

$\mathbb{E}[\varepsilon] = Cov(\varepsilon, X) = Cov(\varepsilon, W) = 0$.

We can interpret β as a measure of the association of X and Y while "controlling" for W.

(3) Contingency table if X, Y binary.

	$X=0$	$X=1$
$Y=0$	p_{00}	p_{10}
$Y=1$	p_{01}	p_{11}

Risk difference RD $= P(Y=1 \mid X=1) - P(Y=1 \mid X=0)$

$$= \frac{p_{11}}{p_{11}+p_{10}} - \frac{p_{01}}{p_{01}+p_{00}}$$

Risk ratio RR $= \dfrac{P(Y=1 \mid X=1)}{P(Y=1 \mid X=0)} = \dfrac{\frac{p_{11}}{p_{11}+p_{10}}}{\frac{p_{01}}{p_{01}+p_{00}}}$

$$\text{Odds ratio OR} = \frac{P(Y=1|X=1)/P(Y=0|X=1)}{P(Y=1|X=0)/P(Y=0|X=0)}$$

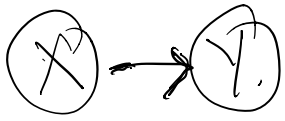
$$= \frac{P_{11} \cdot P_{00}}{P_{01} \cdot P_{10}}$$

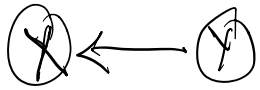
Fact ① If $X \perp Y$, then RD = 0, RR = 1, OR = 1.

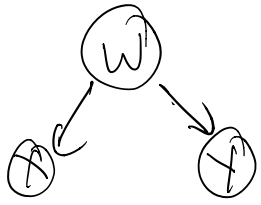
② If $P(Y=1) \ll 1$, then OR $\approx$ RR.

Explaining association.

Multiple causal models for an association between X and Y

① 

② 

③ 

E.g. X = smoking, Y = lung cancer

In ③, W is called a confounder.

Because we don't observe W, we can

marginalizing over W, this leads to an association between X and Y.