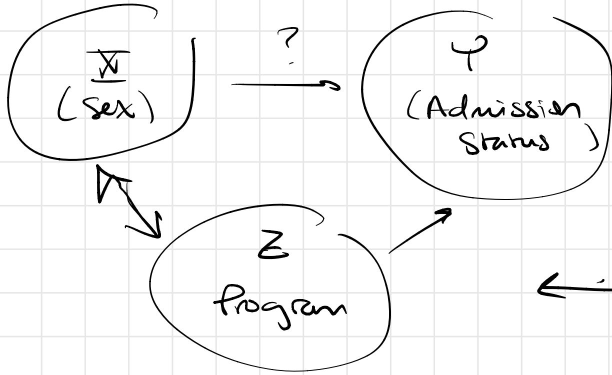
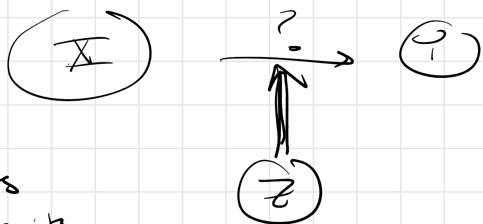
Simpson's Paradox

2/15/22

Example: UC Berkeley Admissions



← Confounder



Z interacts with X on its effect on Y

← effect modification (interaction)

Marginal Independence between X & Y:

$$\longrightarrow P(X = x \mid Y = y) = P(X = x) \quad \forall x, y$$

$$P(X = x \wedge Y = y) = P(X = x)\, P(Y = y)$$

Independent conditional on W

$$P(X = x \mid Y = y, W = w) = P(X = x \mid W = w)$$

$$\forall x, y, w$$

PPT Ex: X & Y indep given $W = 1$?

$$P(Y = 1 \mid X = 1, W = 1) \stackrel{?}{=} P(Y = 1 \mid W = 1)$$

or Odds ratio / RR / diff. prop.

$\theta = 1$ $\quad$ $\theta = 1$ $\quad$ $\theta = 0$

Odds ratios: (conditional)

$$\hat{\theta}_{XY \mid W=1} = \frac{4(9)}{6(6)} = 1$$

$$\hat{\theta}_{XY \mid W=2} = \frac{16(1)}{4(4)} = 1$$

$\rightarrow X \perp\!\!\!\perp Y \mid W$ "independent"

Marginal independence: Sum over W:

	$Y=1$	$Y=2$
$X=1$	20	10
$X=2$	10	10

$\hat{\theta}_{XY} = 2$

$X \not\perp\!\!\!\perp Y$

W an effect modifier? W interact w/ X on Y

$\Rightarrow$ Association between X - Y changes w/ values of W

$$\hat{\theta}_{XY|W=1} = \hat{\theta}_{XY|W=2} = 1$$

$\rightarrow$ **No**

- Homogeneous association between X - Y given W

W a confounder?

- W associated w/ X ?
- W associated w/ Y ?

	W=1	W=2
X=1		
X=2		

	W=1	W=2
Y=1		
Y=2		

$\hat{\theta}_{XW} = 1/6$ $\hat{\theta}_{YW} = 1/6$

$\rightarrow$ **Yes**

X	Y: 1	0
1		
2		
3		

ORs

$$\frac{\text{odds } Y=1 \mid X=1}{\text{odds } Y=1 \mid X=2}$$

$$\text{''} \quad \frac{X=1}{X=3}$$

$$\text{''} \quad \frac{X=2}{X=3}$$

Ch. 3 : Generalized Linear Models

Plan:

1. Review linear model
2. What assumptions of linear models are not met w/ categorical response.
3. "Fix up" linear model to satisfy assumptions.

Linear Model: Y = response variable

$X_1, X_2, \dots, X_k$ = predictor variables

Model $Y \mid X_1 = x_1, \dots, X_k = x_k$.

$i = 1, \dots, n$

$$Y_i = \underbrace{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}}_{\text{fixed}} + \underbrace{\varepsilon_i}_{\text{random}}$$

Assumptions on ε:

1. Normality
2. $E(\varepsilon_i) = 0$
3. $\ast$ $Var(\varepsilon_i) = \sigma^2$ → constant
4. $\ast$ Independence

Mathematically:

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

(independent & identically distributed)

Alternate expression:

$$Y_i \overset{indep.}{\sim} N(\mu_i, \sigma^2)$$

$$\mu_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$$

Generalized linear models – Response variable has some other distribution besides normal.

e.g. Binomial – Linear model assumptions violated → 1 & 3

Binomial: $Y_i \overset{\text{indep.}}{\sim} \text{Bin}(n_i, \pi_i) \quad i = 1, \ldots, n$

$\rightarrow$ Model π_i using $x_{1i} \rightarrow x_{ki}$

$$Y_i \neq \mu_i + \varepsilon_i \qquad E(Y_i) = n_i \pi_i$$

$$\cancel{n_i \pi_i + \varepsilon_i}$$

GLM: 3 components:

① Systematic component (linear predictor):

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

② Link function: $E(Y) = \mu \qquad g(\mu) = \eta$

– how do we model $E(Y)$ as a function of $x_1, x_2, \ldots, x_k$

③ random component – probability distributional assumption on Y

Normal linear model:

$$Y \sim N(\mu, \sigma^2)$$

$$E(Y) = \mu$$

$$\mu = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

Link: $g(\mu) = \mu$

(identity)

Generalized LM:

$$Y \sim ??$$

$$E(Y) = \mu$$

$$g(\mu) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

Examples: ① Bernoulli:

$$Y \sim Bin(1, \pi)$$

$$E(Y) = \pi$$

$$g(\pi) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

Logistic regression $\longrightarrow$ $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$

$$\Rightarrow \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

Common error: $\cancel{\log\left(\frac{Y}{1-Y}\right) = \underbrace{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}_{\text{fixed}}}$

$Y \longrightarrow$ random

$$\cancel{\log\left(\frac{Y}{1-Y}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon}$$

* Don't use an error term!

② Poisson: $Y_i \sim \text{Pois}(\mu)$ $\rightarrow$ counts $y = 0, 1, 2, 3, \ldots$

Most common: "log-linear"

$$\underbrace{\log \mu}_{g(\mu)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Other common link functions:

① Identity: $g(\mu) = \mu$ — Normal

② Logit: $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$ — Binomial

③ Log: $g(\mu) = \log(\mu)$ — Poisson

④ Inverse: $g(\mu) = \frac{1}{\mu}$ — Gamma

⑤ Probit: $g(\mu) = \underbrace{\Phi^{-1}}_{\text{Inverse cdf of Std. normal}}(\mu)$ — Binomial

$\Phi(z) = P(Z \leq z)$

⑥ Complementary log-log link: $g(\mu) = \log(\log(1-\mu))$ (Binomial)