- Odds ratio and relative risk

Q1: Which one should be used to characterize the effect of changing from x_1 to x_2?
Q2: How are they related to x_1 & x_2, say under logit link?

➢ Suppose the probability of successes at x_1 (say, in the presence of some condition) is p_1 and p_2 at x_2 (say, in its absence)

always ratio, i.e. division (cf., in LM, use difference $\mu_1-\mu_2$, i.e. subtraction)

- Relative risk $= p_1/p_2$
- Odds ratio $= o_1/o_2$, $\frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \left(\frac{1-p_2}{1-p_1}\right)\left(\frac{p_1}{p_2}\right)$
- Log odds ratio $= \log(o_1/o_2)$

(log odds) difference = log(odds ratio) (for 2 x's)

$= \log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right) = \eta_1 - \eta_2 = x_1^T\beta - x_2^T\beta = (x_1-x_2)^T\beta$

log odds (for 1 x)

$\log\left[\left(\frac{p_1}{1-p_1}\right)\Big/\left(\frac{p_2}{1-p_2}\right)\right] = e^{(x_1-x_2)^T\beta}$

odds (for 1 x) changes in odds ratio (for 2 x's)

changes in linear (additive) / nonlinear (multiplicative)

cf., In LM $= \mu_1-\mu_2$

➢ For rare outcomes, relative risk ≈ odds ratio, but for larger probabilities, they may be substantial differences

i.e., $p_1 \approx 0$, $p_2 \approx 0$

model: $\eta_i = \beta_0 1 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} = x_i^T\beta$ & logit link.

➢ There is some debate over which is the more intuitive way of expressing the effect of changing from x_1 to x_2.

do not mean $p_1 \approx p_2$ ∵ p_1/p_2 could be large or small

check LNp.2-8, Table 3.1 check LNp.2-11, Table 2.5

Prospective and retrospective sampling

issue in survey sampling, not in DOE

In GLM, covariates (fixed) In GLM, response (random)

- Data: $\left(\underline{x_{j1}, x_{j2}, \ldots, x_{jm}}, \underline{z_j}\right)$, $j = 1, 2, \ldots, K$,
  $\left(\underline{x_{i1}, x_{i2}, \ldots, x_{im}}, \underline{y_i}\right)$, $i = 1, 2, \ldots, k$.

$z_j \in \{0, 1\}$, $\sim$ Bernoulli(p_i) true?

$y_i \in \{0, 1, \cdots, n_i\}$, $\sim$ binomial(n_i, p_i) true?

K: # of units; k: # of covariate classes

Q: how is the data collected?

- Sampling methods: choose a sub-population, then draw a sample from it (different) Recall. In DOE control X, then observe Y (cf.)

➢ Prospective sampling: the covariates x are fixed and **then** the response z (or y) is observed, called *cohort* study.

➢ Retrospective sampling: the response z (but not y) is fixed and **then** the covariates x are observed, called *case-control* study.

in data analysis

In data collection, z: like covariate, x: like response

in data analysis

e.g., $z=1$ e.g., $z=0$

z ($z=1$, present; $z=0$, not present)

➢ An infant *respiratory disease* example:

What are the sub-populations?

- Select a sample of newborn *boy/girl* whose parents had chosen a particular *method of feeding*, and **then** monitor whether disease present or not present for their first year. ($z=1$ (random), $z=0$ (random)) (X_1 (fixed), X_2 (fixed))
- Find infants coming to a doctor with a respiratory disease in the first year and **then** record their sex and method of feeding; also obtain a sample of respiratory disease-free infants and record their information (Note. It requires the ratio of inclusion probabilities in the study to be irrelevant to the covariate values) ← check LNp3-22~24 (X_1 (random), X_2 (random))

case $z=1$ (fixed)

control $z=0$ (fixed)

➢ **Q**: which method is better? → consider: time & financial costs; morality issue; if it is a rare disease

- Since the question of interest is how covariates affect response, prospective sampling seems to be required, but retrospective sampling is cheaper, faster, and more efficient.

Q1: Can retrospective sampling obtain this information? (How X affect P_x or γ_x?)

An example Q2: When can? When cannot?

- response z: disease present/not present – D/D^c (2 categories)
- covariate x: risk factor present/not present – R/R^c (reference: use 0-1 coding for R)

$\gamma=\beta_0+\beta_1 R \leftarrow R(x)$		D^c ($=0$)		D ($=1$)	
1000	R^c ($=0$)	1000/7	π_{00} (0.1) 500 1/2	6000/7	π_{01} (0.6) 750 3/4
1000	R ($=1$)	1000/3	π_{10} (0.1) 500 1/2	2000/3	π_{11} (0.2) 250 1/4

$\gamma_1-\gamma_0=\log\left(\frac{P_1}{1-P_1}/\frac{P_0}{1-P_0}\right)=\log\left(\frac{2/3}{1/3}/\frac{6/7}{1/7}\right)=\log(1/3)=\log\left(\frac{1/3}{3/3}/\frac{3/5}{3/5}\right)$

$P_0=\frac{\pi_{01}}{\pi_{00}+\pi_{01}}=\frac{750}{500+750}=\frac{3}{5}$ ≠

$P_1=\frac{\pi_{11}}{\pi_{10}+\pi_{11}}=\frac{250}{500+250}=\frac{1}{3}$ ≠

Under logit link:

- Given R^c, log-odds for disease is $\log(\pi_{01}/\pi_{00})=\log\left(\frac{P_0}{1-P_0}\right)=\gamma_0$
- Given R, log-odds for disease is $\log(\pi_{11}/\pi_{10})=\log\left(\frac{P_1}{1-P_1}\right)=\gamma_1$
- The difference between the two log-odds is of interest (why?):

$$\Delta = \log(\pi_{11}/\pi_{10}) - \log(\pi_{01}/\pi_{00}) = \log(\pi_{11}/\pi_{01}) - \log(\pi_{10}/\pi_{00})$$

$=\gamma_1-\gamma_0=(\beta_1-\beta_0)^T\beta=\beta_1$

$\log\left(\frac{\pi_{00}\times\pi_{11}}{\pi_{10}\times\pi_{01}}\right)$ odds ratio

log odds ratio; estimable

- The two ratios π_{11}/π_{01} and π_{10}/π_{00} can be estimated in a retrospective manner.

$\pi_{11}/\pi_{01}=\frac{250}{750}=\frac{1}{3}$, $\pi_{10}/\pi_{00}=\frac{500}{500}=1$

But, can γ_0 & γ_1 be estimated? Why cannot?

- A retrospective sampling is as effective as a prospective one for estimating Δ (provided (1) the probabilities of inclusion in D and in D^c are homogeneous or their ratio is irrelevant to covariates, and (2) data is reliable, e.g., no problems such as inaccurate or incomplete historical records; or unreliable memory of the subject)

check ★ in LNp. 3-21

$\tau_{i0}=\tau_{20}=\cdots=\tau_{k0}$, $\tau_{i1}=\tau_{21}=\cdots=\tau_{k1}$

- This manipulation is not possible for other links ever mentioned.

3/12

Q: When can retrospective sampling work (under logit link)? retrospective

Q2 in LNp.3-21

Consider a scenario: a study with response Z and covariates X and

a random variable before sampling

- z_j: binary response of j^{th} unit (e.g., disease present/not present)
- x_j: covariate values of j^{th} unit in the population
- I_j: $=1$ if j^{th} unit is included in the study, 0 if not

depend on the sampling plan

- τ_j: $=P(I_j=1)$=prob. j^{th} unit included in the study
- assume that (i) for j^{th} units in the cell of i^{th} covariate class with $Z=0$ or $Z=1$, respectively,

irrelevant to j (i.e., each unit in a cell has equal prob. to be drawn)

$\tau_j=P(I_j=1\mid Z=0, X=X_i)=\tau_{i0}$;

$\tau_j=P(I_j=1\mid Z=1, X=X_i)=\tau_{i1}$, and

(ii) τ_{i1}/τ_{i0}'s irrelevant to (i.e., constant over) X_i's

$I_j=1, (x_j, z_j)$ observed