p_x = average of all p_j's at $X=x$, ∵ random sampling clusters

LM cf. what if not? cf.

(*) In LM, if assume $Y \sim N(X\beta, I)$
t-test → Z-test
F-test → χ^2-test
if assume $Y \sim N(X\beta, \sigma^2 I)$ cf.

- $E(y) = \sum_i E(s_i) = \sum_i\{E[\underbrace{E(s_i|\pi_i)}_{m\pi_i}]\} = l \times mp = np$ ← cf.
- $Var(y) = \sum_i Var(s_i) = \sum_i\{E[Var(s_i|\pi_i)] + Var[E(s_i|\pi_i)]\}$

$s_1, \ldots, s_\ell$ independent

$= \sum_i\{E[m\pi_i(1-\pi_i)] + Var(m\pi_i)\}$

Var(y) if y ~ binomial

$= l \times \{mp - m[\tau^2 p(1-p) + p^2] + m^2\tau^2 p(1-p)\}$

y not binomial if $\sigma^2 > 1$

$\sigma^2 \equiv$ ✱ $[1+(m-1)\tau^2]\,np(1-p) \geq np(1-p)$

$1 \leq \sigma^2 \leq m \leq n$

- Overdispersion cannot arise when n=1 (sparse case) & m=1

3/26

$Var(y)$
$= Var(\sum_u Z_u)$ ← $B(1,p)$
$= \sum_u Var(Z_u)$ ← $np(1-p)$
$+ 2\sum_{u<v} cov(Z_u, Z_v)$

■ Violation of independence assumption can cause

Check 2 Z's from same cluster in LNp.3-39 ↕ cf. LM

□ over-dispersion, e.g., response has a common cause, say a disease is influenced by genes, the responses will tend to be positively correlated

$E(Z_uZ_v) = E[E(Z_uZ_v|\pi_i)]$
$= E(\pi_i^2) = \tau^2 p(1-p) + p^2$
$cov(Z_u, Z_v)$

assumption: $\sigma^2 = \dfrac{Var(y_x)}{n_x p_x(1-p_x)}$ is a constant over x & n_x (△)

□ under-dispersion, e.g., when food supply is limited, survival probability of an animal may be increased by the death of others, i.e., negatively correlated cf.

$P(Z_u=1|Z_v=0) > P(Z_u=1|Z_v=1)$ survive

- **Q**: how to model overdispersion and do analysis? why? check ✱

in addition to mean structure $\eta = \sum_\ell \beta_\ell x_\ell$

Note. not assign a likelihood

➢ Introduce one additional dispersion parameter σ^2, i.e.,

check (✱)

$$Var(y_\mathbf{x}) = \sigma^2 \times n_\mathbf{x} p_\mathbf{x}(1-p_\mathbf{x}) \Leftarrow \text{notice its similarity to linear model}$$

defined by $g^{-1}(\eta_x) = E(y_x)/n_x$

often include under-dispersion → $\sigma^2 < 1$

(standard binomial case ⇒ σ^2=1; over-dispersion ⇒ σ^2>1) ⇒ $0 < \sigma^2 < \infty$

➢ For a model S, σ^2 can be estimated using $\hat{\sigma}^2_S = X^2_S/(k-p)$

why? check ✱ in LNp.34

$\eta = \sum_{\ell=1}^{p} \beta_\ell x_\ell$ ← $D_S \overset{a}{\approx} X^2_S = \sum_i (r^P_{i,S})^2 \overset{cf.}{\longleftrightarrow}$ RSS in LM

# of covariate classes

# of parameters in η (S)

When n_i's large, using D_S is ok

(using deviance D in place of X^2 is not very recommended as D may be inconsistent for sparse data)

check LNp.3-14 → $D/(k-p)$ might not converge to $\sigma^2=1$ as $k \to \infty$ ← n_i's=1, σ^2=1

MLE

➢ Estimation of β is unaffected since $E(y_\mathbf{x}) = n_x p_x = n_x g^{-1}(\eta_x) = n_x g^{-1}(\sum \beta_\ell x_\ell)$ is not changed (Why? Note that $y_\mathbf{x}$ is not ~ binomial so that likelihood is different) → •quasi-likelihood

rationale
- IRWLS only need mean & var functions
- weights ∝ $[Var(y_x)]^{-1}$ $= 1/[\sigma^2 n_x p_x(1-p_x)]$ ∝ $1/[n_x p_x(1-p_x)]$ ← binomial case

Intuition. In LM, if $\varepsilon \to c\varepsilon$
- $Var(\varepsilon) \to c^2 Var(\varepsilon)$
- $\hat{\varepsilon} \to c\hat{\varepsilon}$
- RSS → c^2RSS
- $E(\hat{\beta})$ not change
- $Var(\hat{\beta}) \to c^2 Var(\hat{\beta})$

➢ But, $Var(\hat{\beta}) \approx \sigma^2(X^T W X)^{-1}$ and $\hat{Var}(\hat{\beta}) = \hat{\sigma}^2(X^T \hat{W} X)^{-1}$

parameter — from IRWLS under binomial | estimate

➢ For S nested in L, difference in their deviances

$$D_S - D_L \overset{a}{\sim} \sigma^2 \chi^2_{df_S - df_L} \quad \text{(under } S\text{)}$$

parameter — cannot use it as a null dist.

cf. check LNp.3-10

Let L be the saturated model. Then, under S $D_S, X^2_S \overset{a}{\sim} \sigma^2 \chi^2_{df_S}$

➢ When comparing models, e.g., testing H_0:S vs. H_1:$L\backslash S$, can use

0, D_{sat}, D_L, D_S

$$F = \frac{(D_S - D_L)/(df_S - df_L)}{\hat{\sigma}^2_L} \overset{a}{\sim} F_{df_S - df_L,\, df_L} \quad \text{(under } S\text{)}$$

X^2_L/df_L

LM cf.

i.e., adding a σ^2 multiplier on binomial variance

check graph ✱ in LNp.3-37

∵ no information about true σ^2 and $X^2_L = 0$ if L is saturated

more possible that $Var(y_x)/[n_x p_x(1-p_x)]$ is a constant (check ✱ in LNp.3-40)

➢ No goodness-of-fit test is possible

➢ This dispersion parameter method is more appropriate when the covariate classes are roughly equal in size (i.e., $n_1 \approx n_2 \approx \ldots \approx n_k$)

p. 3-42

➢ Alternative approaches to over-dispersion

more flexible than dispersion parameter method

likelihood is known (cf. dispersion parameter method)

- beta-binomial method (Williams, 1982; Crowder, 1978)
- quasi-likelihood: specify only how the mean and variance of the response are connected to covariates.

1st moment 2nd moment

$P_x \sim beta(\alpha_x, \beta_x)$

$y_x | P_x \sim B(n_x, P_x)$

$y_x \sim$ beta-binomial

$$E(y_x) = n_x\underbrace{\left(\frac{\alpha_x}{\alpha_x+\beta_x}\right)}_{P_x^*}$$

a linear function of n_x (cf. (a) in LNp.3-40)

$$Var(y_x) = n_x\underbrace{\left(\frac{\alpha_x}{\alpha_x+\beta_x}\right)}_{P_x^*}\underbrace{\left(\frac{\beta_x}{\alpha_x+\beta_x}\right)}_{1-P_x^*}\underbrace{\left(\frac{\alpha_x+\beta_x+n_x}{\alpha_x+\beta_x+1}\right)}_{\geq 1}$$

But, not the whole dist. family (i.e., likelihood unknown)

use them to define a function working as likelihood for estimation & testing ⇒ same estimator & test stat. for all dist. with same mean & var. functions

❖ **Reading**: Faraway (2006, 1st ed.), 2.11

LNp 3-19 ~24

Matched Case-Control Studies

Recall 1. blocking in DOE

Recall 2. paired t (within block comparison) vs. 2-sample t

- **Q**: In a case-control study, how should we choose the controls if there exist some *confounding variables* W, say age and sex, that may affect the outcome in addition to the *risk factors* X?

block factor in DOE, i.e., covariates of no interest, but should be considered in design and analysis.

covariates of main interest

deal with W in data analysis

➢ Approach 1: record and include confounding variables as covariates in GLM analysis (however, we may not be interested in the effects of the confounding variables)

ANCOVA cf.

Why include covariates of no interest in model? Recall. In LM, true: $Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$, fitted: $Y = X_1\beta_1 + \varepsilon$

may cause a large number of covariate classes (β) & possibly sparse data

deal with W in data collection

➢ Approach 2: confounding variables are explicitly adjusted for in the design

blocked design cf.

Recall. In paired t comparison block effect removed.

p. 3-43

- *Matched case-control design (MCCD)*: match each case with one or more controls that have the same or similar values of some set of potential confounding variables. A group of a case and its corresponding controls is called a *matched set*, e.g.,

having same value of W

allows within block comparison in this contingency table

units in a matched set are regarded as homogeneous (∵ they have same values of W, check ☆ in LNp.3-39, W = other covariates)

controls

values of risk factor

1st case

age=20; sex=male

	D^c	D
X^c	y_{100}	1
X	y_{110}	0

⊕=1

⊕=M

a block

2st case

age=20; sex=female

	D^c	D
X^c	y_{200}	0
X	y_{210}	1

a block

cf. a matched set

• • •

nst case

age=70; sex=female

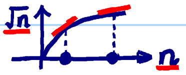
	D^c	D
X^c	y_{n00}	1
X	y_{n10}	0

➢ 1:M MCCD: M controls for each case

case

controls

Recall. s.e.$(\bar{y}) = \sigma/\sqrt{n}$ ↓ as n ↑. But,



- M typically small, can vary in size in every matched set
- Each additional control yields a diminished return in terms of increased efficiency in estimating the effects of risk factors
- It is usually not worth exceeding M=5

more controls higher efficiency

$Var(\bar{y}_1 - \bar{y}_2) = \sigma_1^2/n_1 + \sigma_2^2/n_2$

if $\sigma_1^2/n_1 \ll \sigma_2^2/n_2$, σ_2^2/n_2 dominates

Then, it's of not much use to increase n_1

- Some disadvantages of MCCD

➢ Lose the possibility of discovering the effects of the confounding variable W

cf. Approach 1 in LNp 3-42 can estimate effects of W

cannot estimate true block effects of W, e.g., in each matched set, estimate of $P(Z=1|W=w_j)$ is $1/(M+1)$ due to 1:M setting ⇒ irrelevant to w_j

➢ The data will likely be far from a random sample of the population of interest

Check example in LNp.3-21

might not able to gain some information of the population, e.g.