check in data analysis

- Some possible explanations for large deviance D_S, say when H_0: S is correct is rejected (**cf.**, possible reasons causing $\hat{\sigma}_S^2$ larger than σ^2 in Normal linear model) lack-of-fit test rejected or RSS too large

individual pattern

➢ Sparse data (i.e., $n_{\mathbf{x}}$'s too small) $\Rightarrow D_S \overset{a}{\sim} \chi^2_{k-p}$ is questionable ← hold when n_i's large; might not close to true null dist.

➢ Presence of some outliers (can be detected in diagnostics for GLM, future lecture) i.e., overall pattern, some assumptions in model need reviewed and revised

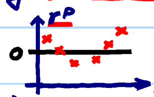
check 2nd graph in LNp.3-37

◉ For a larger number of outliers, we might conclude that they are not exceptional $\Rightarrow$ something amiss with other structures

➢ Wrong linear structure in $\eta_{\mathbf{x}}$ ← deterministic component 規 in LNp 3-5

too simple

cause $\hat{p}_i$ too far away from true p_i

◉ Covariates not transformed/combined in a correct way (diagnostics for GLM can help) → change effect codings R_i's; e.g.  ⇒ add quadratic effect of x.

check example in LNp.3-39

◉ Important covariates are not included ($\Rightarrow$ cause overdispersion if these important predictors are not available) 2 situations: (1) in data (2) not in data

隨 in LNp 3-5

➢ Deficiencies in the random part of GLM (i.e., $y_{\mathbf{x}}$ not binomial) cf.

Consider two scenarios when $y_{\mathbf{x}}$ not binomially distributed: $\{0, 1, 2, \ldots, n_{\mathbf{x}}\} \ni$

Why no likelihood assigned?

usually, same treatment for each scenario rationale: quasi-likelihood

- over-dispersion: $Var(y_{\mathbf{x}}) \gg n_{\mathbf{x}} p_{\mathbf{x}}(1 - p_{\mathbf{x}})$
- under-dispersion: $Var(y_{\mathbf{x}}) \ll n_{\mathbf{x}} p_{\mathbf{x}}(1 - p_{\mathbf{x}})$ (in rare cases)

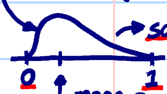
$p_{\mathbf{x}} = E(y_{\mathbf{x}}/n_{\mathbf{x}})$, $g^{-1}(\eta_{\mathbf{x}})$

---

missing important covariates in LNp.3-38

[Note: $y_{\mathbf{x}} \sim B(n_{\mathbf{x}}, p_{\mathbf{x}})$ only when the corresponding $n_{\mathbf{x}}$ $z_{\mathbf{x}}$'s are (1) *independent* and (2) *identically* distributed as $B(1, p_{\mathbf{x}})$]

population dist. of P_i's



sampling; $\pi_1, \ldots, \pi_l$; mean p; Var $\leq p(1-p)$

$0 \leq \tau^2 = \frac{Var}{p(1-p)} \leq 1$

- Violation of the assumption of same $p_{\mathbf{x}}$ in a covariate class

cf. ◉ Example: in shuttle disaster case, position of O-ring on the booster rocket may have effect on the failure probability. Yet, this (important) covariate was not recorded. not in data; depend on position

non-constant P_x

□ **Q**: How can the heterogeneity cause overdispersion? e.g., (i) the sub-population of **X**=**x** can be divided into clusters (families, litters, …); (ii) l clusters sampled, and (for simplicity) m units drawn from each of the l clusters; can be generalized to $m_1, \ldots, m_l$ (iii) total number of units sampled at **X**=**x** is n=l×m.

same covariate class

m units	X=x
1st cluster (1st sampled)	p_1 000111 000111
2nd cluster (not sampled)	p_2 001111 001111
⋮	⋮
uth cluster (lth sampled)	p_u 000001 000001
⋮	⋮

different

- In the i^{th} sampled cluster, i=1, …, l, number of successes = $s_i \sim B(m, \pi_i)$ ← i.e., $S_i | \pi_i$; $s_i = \sum_{i\text{th cluster}} z_{\mathbf{x}}$ where π_i's could vary (say, vary with other covariates)

e.g., a function of other covariates

If they are in data, can assign their effects in η

- Regard π_i's as independent random variables ← ∵ sampling with mean p and variance $\tau^2\, p(1-p)$ ← $\pi_i \sim$ population dist. of P_i's; $\tau^2 \in [0, 1]$
- $y = s_1 + \ldots + s_l$ ← $\sum_{\mathbf{X}=\mathbf{x}} z_{\mathbf{x}}$

p. 3-40

◆ $E(y) = \sum_i E(s_i) = \sum_i \{E[E(s_i|\pi_i)]\} = l \times mp = np$

◆ $Var(y) = \sum_i Var(s_i) = \sum_i \{E[Var(s_i|\pi_i)] + Var[E(s_i|\pi_i)]\}$

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

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

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

◆ Overdispersion cannot arise when $n=1$ (sparse case)

■ Violation of independence assumption can cause

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

□ 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

- **Q**: how to model overdispersion and do analysis?
  - ➢ Introduce one additional dispersion parameter σ^2, i.e.,

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

(standard binomial case $\Rightarrow \sigma^2=1$; over-dispersion $\Rightarrow \sigma^2>1$)

p_x = average of all p_j's at $X=x$, ∵ random sampling clusters
LM cf. what if not?

$m\pi_i$

cf.

$s_1, \cdots, s_l$ independent

Var(y) if y ~ binomial

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

$\sigma^2 =$ ✱

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

(✱) 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.

$Var(y) = Var(\sum_u Z_u)$ 3/26
$= \sum_u Var(Z_u)$ ← $B(1,p)$, $np(1-p)$
$+ 2\sum_{u<v} cov(Z_u, Z_v)$

Check 2 Z's from same cluster in LN p. 3-39 cf. LM

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

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

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

why? check ✱

in addition to mean structure $\eta = \sum_l \beta_l k_l$

check (✱)

Note. not assign a likelihood

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

often include under-dispersion → $\sigma^2 < 1$ ⇒ $0<\sigma^2<\infty$