- 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

Q2 in LNp.3-21

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

retrospective

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
- τ_j: $=P(I_j=1)$=prob. j^{th} unit included in the study

depend on the sampling plan

- 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)

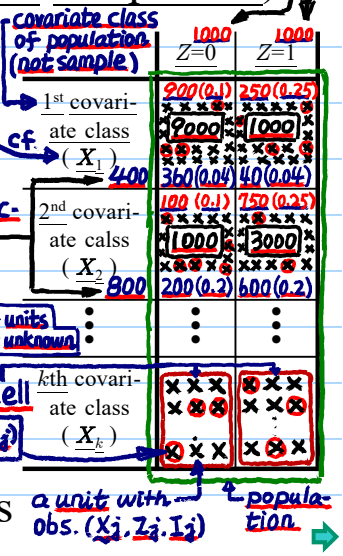
$\tau_j = P(I_j=1 \mid Z=0, X=X_i) = \tau_{i0}$;
$\tau_j = P(I_j=1 \mid Z=1, X=X_i) = \tau_{i1}$, and

eg., $i=1$, In τ_{10}, $j=1, \cdots, 9000$. In τ_{11}, $j=9001, \cdots, 10000$

ρ

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

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



- For fair prospective study, $\tau_{i0}=\tau_{i1}$, $i=1, \ldots, k$. → $\rho = 1$ ← check example in LNp. 3-22
- For retrospective study, often $\tau_{i0} \neq \tau_{i1}$ and τ_{i1} much larger than τ_{i0} → $\rho > 1$
- The probability of interested is

$p_i \equiv P(Z=1 \mid X=X_i)$ ← describe how covariates X affect the response Z

cases (controls, resp.) appear more (less, resp.) often than it should be ↑ prospective

- We can use data (either from prospective or retrospective sampling) to study the probability:

$q_i \equiv P(Z=1 \mid I=1, X=X_i)$ ← in observed data

In the example in LNp 3-21, they are
6/7 (x=0) & 2/3 (x=1) in prospective
3/5 (x=0) & 1/3 (x=1) in retrospective

Q1: $p_i \overset{?}{=} q_i$, check this →

Q2: the way X_i affect p_i $\overset{?}{=}$ the way X_i affect q_i (Recall ANCOVA model)

$$q_i = \frac{P(Z=1, I=1 \mid X=X_i)}{P(I=1 \mid X=X_i)}$$

$$= P(Z=0, I=1 \mid X=X_i) + P(Z=1, I=1 \mid X=X_i)$$
$$= P(I=1 \mid Z=0, X=X_i) \times P(Z=0 \mid X=X_i) + P(I=1 \mid Z=1, X=X_i) \times P(Z=1 \mid X=X_i)$$
$$= \tau_{i0} \times P(Z=0 \mid X=X_i) + \tau_{i1} \times P(Z=1 \mid X=X_i)$$
$$= \tau_{i0} \times (1-p_i) + \tau_{i1} \times p_i$$

Bayes Thm ↓

$$= \frac{\tau_{i1} \times p_i}{\tau_{i1} \times p_i + \tau_{i0} \times (1-p_i)} \quad (\star)$$

representitive data for p_i / not representitive for p_i

$(*) = p_i = \pi_i$ — In prospective $\tau_{i0} = \tau_{i1}$ — $(*)$ in LNp.3-23 — In retrospective $\tau_{i0} \neq \tau_{i1}$ — $(*) \neq p_i$, $= \pi_i$

But, still representitive for $\beta_1, \cdots, \beta_{p-1}$

- Under logit link, we have

$$\text{logit}(p_i) = \eta_i = \beta_0 + \beta_1 h_1(x_i) + \cdots + \beta_{p-1} h_{p-1}(x_i) = h_i \beta$$

What if change logit to other link g?

- Then, for retrospective π_i = prospective π_i

$$\eta_i = \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) \overset{(*)}{=} \log\left(\frac{\tau_{i1} p_i / (\tau_{i1} p_i + \tau_{i0}(1-p_i))}{\tau_{i0}(1-p_i) / (\tau_{i1} p_i + \tau_{i0}(1-p_i))}\right)$$

$$= \log\left(\frac{\tau_{i1}}{\tau_{i0}}\right) + \log\left(\frac{p_i}{1-p_i}\right) \quad \text{(logit}(p_i)\text{)}$$

$$= \log(\rho) + \beta_0 + \beta_1 h_1(x_i) + \cdots + \beta_{p-1} h_{p-1}(x_i) \quad (\Delta)$$

$\beta_0^* = \log(\rho) + \beta_0$ ← different intercept (*); $\beta_1, \cdots, \beta_{p-1}$ same

i.e., still treat z as response (although it's fixed). treat x as covariate (although it's random). an fit a logistic regression

What if $\log(\frac{\tau_{i1}}{\tau_{i0}}) = \log(\rho_i) = f(x_i) \approx \sum_\ell \alpha_\ell h_\ell(x_i)$? ($\alpha_\ell$: parameter)

$(\Delta) = \sum_\ell (\alpha_\ell + \beta_\ell) h_\ell$ — aliasing

When would $\frac{\tau_{i1}}{\tau_{i0}}$ depend on x_i? check example in LNp.3-22

- In a retrospective study,
  - $\beta_1, \ldots, \beta_{p-1}$ (effects of X) are estimable
  - β_0 is inestimable when no information about τ_{i1}/τ_{i0} ⇒ cannot estimate p_i's (check example in LNp.3-22)

Q: Why should the ratio of τ_{i1} and τ_{i0} be irrelevant to X? (check (ii) in LNp.3-22)

❖ **Reading**: Faraway (2006, 1st ed.), 2.4, 2.5, 2.6

Recall tolerance distribution (LNp.3-16~17) → logit, probit, CLL

Choice of link function

Note different links, different interpretation of β (cf., different interpretation of β under different dummy variables)

- Usually not possible to make the choice based on the data alone (i.e., based on fits) because (e.g., use deviance of the fit ($y_i \leftrightarrow \hat{y}_i$) of $\hat{y}_i$ or $\hat{p}_i$)
  ① the fits based on the 3 links would be quite similar for moderate p (why?) (different; p small or large → LNp.3-5)
  ② for p_i that is close to 0 or 1, a very large amount of n_i would be necessary to distinguish them (why?)

usually, do not have large n_i data on x_i with $p_i \approx 0$ or $p_i \approx 1$

Q: But, $\text{Var}(\hat{p}_i)$ usually small when $p_i \approx 0$ or $p_i \approx 1$, e.g., if $\hat{p}_i = y_i/n_i$, ≈0 when $p_i \approx 0$ or 1

$\text{Var}(\hat{p}_i) = \frac{1}{n_i} p_i(1-p_i)$ — based on $\hat{p}_i = p_i$

If $\hat{p}_i$ is an accurate est'or, why need large n_i?

Consider $\hat{p}_i/p_i$. $E(\hat{p}_i/p_i) = 1$

$\text{Var}(\hat{p}_i/p_i) = \frac{1}{n_i}\left(\frac{1-p_i}{p_i}\right)$ — large when $p_i \approx 0$

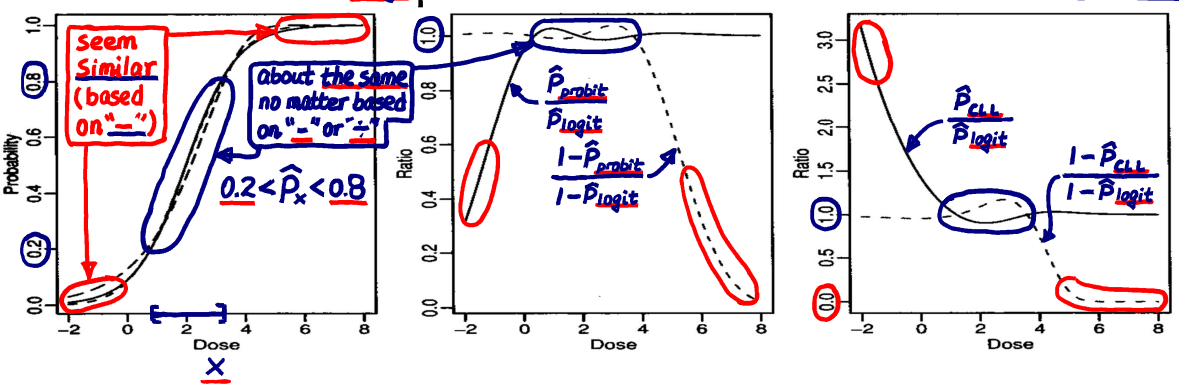
same thing for $(1-\hat{p}_i)/(1-p_i)$ when $p_i \approx 1$

- A lab example

Recall $y_i \leftrightarrow \hat{y}_i$: RSS: $y_i - \hat{y}_i$; deviance: $y_i/\hat{y}_i$ | e.g., 0.01 ↔ 0.0001; diff = 0.0099; ratio = 100

—logit --CLL ---probit

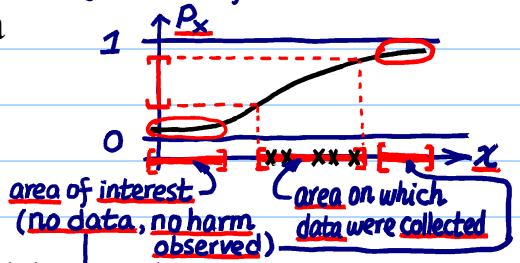
$\hat{p}_x = g^{-1}(\hat{\eta}_x) = g^{-1}(x_x^T \hat{\beta})$



- The choice of link function is usually made based on assumptions derived from physical knowledge or simple convenience
  - e.g., the idea of tolerance distribution (LNp.3-16) can be helpful in some cases.
- Logit link (logistic regression) is popular. Reasons to use logit:
  - theoretically/mathematically simpler due the intractability of Φ (canonical link)
  - easier to interpret – logit is related to log-odds ★ in LNp.3-18 & 19
  - retrospective sampling ★ in LNp.3-21, □ in LNp.3-24

e.g., Normal cdf (probit link) has no explicit form

- In some cases, harmful effects only become apparent at large dosages — more extreme cases (observed data), which make us start to realize something is harmful.
  - In order to estimate the probability of a harmful effect at low dose, it would be necessary to select an appropriate link function. (regular condition, connected with rare event) check graphs in LNp.3-25



  - However, the data for high dosages will be of little help in selecting an appropriate link — extrapolation: it becomes more difficult (compared to LM) because of the lower (0) & upper (1) bounds in p. other example: highly reliable products

❖ **Reading**: Faraway (1st ed.), 2.7

Prediction and Effective Doses

an important objective in regression-type analysis

- Recall: At covariate values $\mathbf{x}_0 = (x_{01}, x_{02}, \ldots, x_{0m})^T$, denoted by $\hat{O}_{\mathbf{x}_0}$ & $[\underline{O}_{\mathbf{x}_0}, \bar{O}_{\mathbf{x}_0}]$

$$p_{\mathbf{x}_0} \xrightarrow{g} \eta_{\mathbf{x}_0} = \mathbf{h}_0{}^T \beta, \text{ where } \mathbf{h}_0 = (1, h_1(\mathbf{x}_0), \ldots, h_{p-1}(\mathbf{x}_0))^T$$

(g^{-1}; e.g., g = logit $\Rightarrow \log(O_{\mathbf{x}_0}) = \eta_{\mathbf{x}_0}$)

- Prediction of the probability $p_{\mathbf{x}_0}$ (or odds $o_{\mathbf{x}_0}$, or $\eta_{\mathbf{x}_0}$) of success for given covariate values $\mathbf{x}_0$ (for parameters in model) — similar to $\hat{P}_{\mathbf{x}_0}$ & $[\underline{P}_{\mathbf{x}_0}, \bar{P}_{\mathbf{x}_0}]$
  - For the MLE $\hat{\beta}$, we have $\hat{\beta} \overset{a}{\sim} N(\beta, \Sigma)$. Denote the estimate of Σ by $\hat{\Sigma}$ (consistent). ($= (X^T \hat{W} X)^{-1}$ from IRWLS; $\hat{W}$ change it)
  - Predict $\eta_{\mathbf{x}_0}$ by $\mathbf{h}_0{}^T \hat{\beta}$, denoted by $\hat{\eta}_{\mathbf{x}_0}$. (lik $\hat{u}_{x_0}$ in LM)
  - Predict $p_{\mathbf{x}_0}$ by $g^{-1}(\hat{\eta}_{\mathbf{x}_0})$, denoted by $\hat{p}_{\mathbf{x}_0}$. ($E\left(\frac{y_{\mathbf{x}_0}}{n_{\mathbf{x}_0}}\right)$; cf. mean response)
  - 100(1−α)% confidence interval for $\eta_{\mathbf{x}_0}$:

$$\hat{\eta}_{\mathbf{x}_0} \pm z(\alpha/2) \times \sqrt{\mathbf{h}_0{}^T \hat{\Sigma} \mathbf{h}_0} \equiv [\underline{\eta}_{\mathbf{x}_0}, \bar{\eta}_{\mathbf{x}_0}] \Leftarrow P\left(\left|\frac{\hat{\eta}_{\mathbf{x}_0} - \eta_{\mathbf{x}_0}}{\sqrt{\mathbf{h}_0^T \Sigma \mathbf{h}_0}}\right| < z\left(\frac{\alpha}{2}\right)\right) = 1-\alpha$$

($\mathrm{se}(\hat{\eta}_{x_0})$)

because $\hat{\eta}_{\mathbf{x}_0} = \mathbf{h}_0{}^T \hat{\beta} \overset{a}{\sim} N(\mathbf{h}_0{}^T \beta = \eta_{\mathbf{x}_0}, \mathbf{h}_0{}^T \Sigma \mathbf{h}_0)$.

  - 100(1−α)% confidence interval for $p_{\mathbf{x}_0}$:

$$[g^{-1}(\underline{\eta}_{\mathbf{x}_0}), g^{-1}(\bar{\eta}_{\mathbf{x}_0})] \equiv [\underline{p}_{\mathbf{x}_0}, \bar{p}_{\mathbf{x}_0}]$$

$P(\eta_{x_0} \in [\underline{\eta}_{x_0}, \bar{\eta}_{x_0}]) = P(p_{x_0} = g^{-1}(\eta_{x_0}) \in [g^{-1}(\underline{\eta}_{x_0}), g^{-1}(\bar{\eta}_{x_0})])$ (random; g^{-1} monotone)

Note. In Normal LM, wider C.I. for μ_{x_0} when x_0 is away from center $\bar{x}$ (i.e., has larger leverage), same for binomial GLM?

Ans. Yes for η_{x_0}, No for p_{x_0}. But, $\log\left(\frac{\bar{O}_{x_0}}{\underline{O}_{x_0}}\right)$ would get larger as $\eta_{x_0} - \bar{\eta}_{x_0}$ can be regarded as a log-odds-ratio