➢ Approach 2: (profile likelihood-based method)

MLE of θ_2 when θ_1 is fixed. It's a function of θ_1

- Parameters: $\underline{\theta} = (\underline{\theta_1}, \underline{\theta_2})$ of current interest
- Likelihood: $\mathcal{L}(\underline{\theta}) = \mathcal{L}(\underline{\theta_1}, \underline{\theta_2})$

focus only on $\underline{\theta_1}$

- Profile likelihood: $\mathcal{L}^*(\underline{\theta_1}) = \max_{\underline{\theta_2}} \mathcal{L}(\underline{\theta_1}, \underline{\theta_2}) = \mathcal{L}(\underline{\theta_1}, \hat{\underline{\theta}}_2(\underline{\theta_1}))$ fixed first

β: $\hat{\beta}_\lambda = (X^TX)^{-1}X^T[t_\lambda(Y)]$ ⟷ σ^2: $\hat{\sigma}^2_\lambda = RSS_\lambda/(n-p)$

LNp.1-32~33 (recall: the computation of the C.I. for λ in Box-Cox method)

transformation parameter

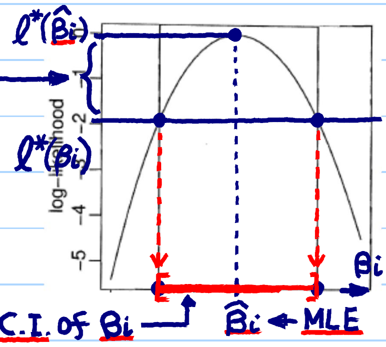
profile log-likelihood

- For a coefficient β_i, its profile likelihood-based C.I. is

$2[\ell^*(\hat{\beta}_i) - \ell^*(\beta_i^*)] < \chi^2_1(1-\alpha)$

MLE under ℓ^*

$$\{\beta_i : l^*(\beta_i) > l^*(\hat{\beta}_i) - (1/2)\chi^2_1(1-\alpha)\}$$



other β_j's, $j \neq i$, set to the MLE when β_i fixed (i.e., MLE as a function of β_i)

compared to Wald test based C.I.

$H_0: \beta_i = \beta_i^*$ vs. $H_1: \beta_i \neq \beta_i^*$ any constant

➢ The profile likelihood method is generally preferable for the same Hauck-Donner reason

Acceptance region of LR test under profile likelihood: $-2\log\left[\frac{\mathcal{L}^*(\beta_i^*)}{\mathcal{L}^*(\hat{\beta}_i)}\right] < \chi^2_1(1-\alpha)$ MLE under profile likelihood

e.g., • Wald: AB • profile: some β_i's

➢ Similar method can be generalized to construct confidence region of several parameters

---

- Some notices about deviance when n_i is small

sparse data

→ asymptotic approaches not work
→ $se(\hat{\beta}_i)$ might be over-estimated (LNp3-12)

➢ Take the extreme case $n_i=1$ (binary data). ← Each covariate class has only one unit.

$\mathcal{L}(\hat{\beta}_{sat}) = \prod_i \hat{p}_{i,sat}^{y_i} \times (1-\hat{p}_{i,sat})^{1-y_i} = \prod_i 0^0 1^1 = 1$

$\hat{p}_{i,sat} = y_i/n_i$ (LNp 3-9)

Then, $y_i = 0$ or 1. For a model S (i.e., $X_S\beta_S$) under logit link, ~ Bernoulli(p_i)

Check ✱ in LNp.3-5

$$D_S = -2\sum_{i=1}^{k}\{\hat{p}_{i,S} \times \text{logit}(\hat{p}_{i,S}) + \log(1-\hat{p}_{i,S})\}$$

$= -2\log\left[\frac{\mathcal{L}(\hat{\beta}_S)}{\mathcal{L}(\hat{\beta}_{sat})}\right] = -2\log[\mathcal{L}(\hat{\beta}_S)] = -2\,\ell(\hat{\beta}_S)$

$= -2\sum_i[y_i \cdot \hat{\eta}_{i,S} + n_i \times \log(1-\hat{p}_{i,S})]$

LNp. 3-6

$= \hat{\eta}_S^T Y = (X_S\hat{\beta}_S)^T Y = \hat{\beta}_S^T X_S^T Y \overset{!}{=} \hat{\beta}_S^T X_S^T \hat{\mu}_S = \hat{\eta}_S^T \hat{\mu}_S = \hat{\eta}_S^T(\underline{n} \circ \hat{P}_S) = \hat{\eta}_S^T \hat{P}_S$

∵ normal equation (LNp.3-7)

When $X_S^TY = X_S^T\hat{\mu}_S$, say $X_S^TY =$ $\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_k \end{bmatrix}^T \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}$, ⇒ can arbitrarily change Y, but keep X_S^TY invariant ⇒ same D_S for different Y

cf. LM

This also explains why need to form covariate classes

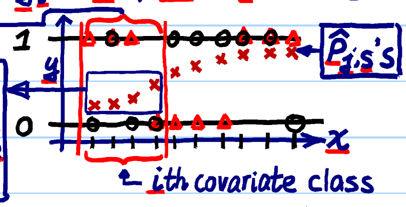
➢ For a D_S to measure goodness-of-fit, it has to compare the fitted values $\hat{y}_{i,S} = n_i\hat{p}_{i,S}$ ($=\hat{\mu}_{i,S}$) to the data y_i (cf., in linear model, we use $y_i - \hat{y}_i$ to compare), but here we have only a function of $\hat{p}_{i,S}$

result in GofF test is suspicious ∵ $D_S|\hat{\beta}_S$ is a degenerate r.v.

Why can it overcome the problem in D_S for sparse data?

n_i^* = # of units in ith class
y_i^* = # of 1's in ith class

➢ Other methods must be used to judge goodness-of-fit for binary data or the case of small n_i's [e.g., Hosmer-Lemeshow test, see Hosmer and Lemeshow (2013, 3rd ed.), *Applied Logistic Regression*]



$P_{i,S}^*$ = average of these $P_{i,S}$'s

based on grouping

required

$y_i^* \leftrightarrow \hat{y}_{i,S}^* = n_i^*\hat{P}_{i,S}^*$, i.e., use n_i^*'s (≥ 1) y_i^*'s, $\hat{P}_{i,S}^*$'s to measure goodness-of-fit

Reading: Faraway (2006, 1st ed.), 2.1, 2.2, 2.3, Appendix A

- a comparison

$\sigma^2 I$ — estimation ($RSS/n-p$); generalization: WLS, GLS, ⋯
$\parallel$
$Var(Y)$

Q: Why no estimation/inference related to $Var(Y)$? $Var(y_i)=\frac{\mu_i(n_i-\mu_i)}{n_i}$, $\mu_i=E(y_i)$

	Linear model	Binomial
	$Y=X\beta+\varepsilon$ / $\boxed{Y\sim N(X\beta,\sigma^2 I)}$	→ GLM
estimation of β	LS = MLE	MLE
Goodness of fit or Lack of fit	RSS, $\hat{\sigma}$ & $R^2=1-RSS/TSS$; $y_i\leftrightarrow\hat{y}_i$ $(y_i-\hat{y}_i)$, $RSS_S\leftrightarrow RSS_{sat}$	Deviance; $y_i\leftrightarrow\hat{y}_i$ $\left(\frac{y_i}{\hat{y}_i}\right)$, $2\log\left(\frac{\mathcal{L}_{sat}}{\mathcal{L}_S}\right)$
H_0: S vs. H_1: $L\backslash S$	(*) — $\frac{(RSS_S-RSS_L)/(df_S-df_L)}{RSS_L/df_L}\sim F$ (H_0) $\boxed{\text{exact}}$	$D_S-D_L \overset{a}{\sim} \chi^2_{df_S-df_L}$ (H_0) —(△); $\boxed{\text{asymptotic } (n_i\text{'s large})}$
H_0: $\beta_i=0$	from (*) ← $\boxed{\text{equivalent}}$; (□) — $\hat{\beta}_i/se(\hat{\beta}_i)\sim t$ (H_0) $\boxed{\text{exact}}$	from (△) ← $\boxed{\text{different}}$; (○) — $\hat{\beta}_i/se(\hat{\beta}_i)\overset{a}{\sim}N(0,1)$ (H_0) $\boxed{\text{asymptotic}}$
Confidence interval or region	from (*) ← $\boxed{\text{equivalent}}$; from (□) ← estimate ± (C.V.) × se(estimate)	profile likelihood method (related to (△)); from (○) ← $\boxed{\text{different}}$

Tolerance distribution

If fit $Y\sim glm(\beta_0+\beta_1 X)$, how to interpret its link function?

- Consider the following example:

observed data — assume x_i's are different

	X	$Z(1{:}X, 0{:}Y)$	y
1st covariate class	x_1	z_{11} z_{12} ⋯ z_{1n_1}	y_1
	⋮	⋮	⋮
kth covariate class	x_k	z_{k1} z_{k2} ⋯ z_{kn_k}	y_k

y: # of students got X; z columns: from different students — 1st student (T_1), 2nd student (T_2), …, nth student (T_n); not observe

latent continuous r.v.

> Students answers k questions on a test

> The jth student has an aptitude T_j,

F = cdf of $N(0,1)$, $T\sim N(\mu,\sigma^2)$

e.g. T_j's $\overset{i.i.d.}{\sim}$ $F\left(\frac{t-\mu}{\sigma}\right)$ — a (known) cdf; μ: location parameter, population distribution; σ: scale parameter

> The ith question has a fixed difficulty x_i, $i=1,\cdots,k$.

If changed to $(x_{i1},\cdots,x_{im})$

> The jth student will get the ith answer ~~correct~~ wrong only if $T_j > x_i$ [$T_j\le x_i$]

If changed to $T_j \lesseqgtr \sum_\ell \gamma_\ell h_\ell(x_i)$

> The probability that a randomly selected student will get the ith answer wrong is:

GLM for binomial response y_i's and covariate x_i's

response: $y_i\sim B(n_i,p_i)$

$$p_i=P(T_j\le x_i)=F((x_i-\mu)/\sigma)$$

$$\Rightarrow\ F^{-1}(p_i)=(-\mu/\sigma)+(1/\sigma)x_i\equiv\beta_0+\beta_1\times x_i=\eta_i$$

link function $g \Leftrightarrow F=g^{-1}$; linear structure; covariate

Become $\eta_i=\sum_\ell \beta_\ell h_\ell(x_i)$

> The distribution of T_j is called *tolerance distribution*, which arose from toxicity studies where the aptitude would be replaced with the tolerance of the insects.

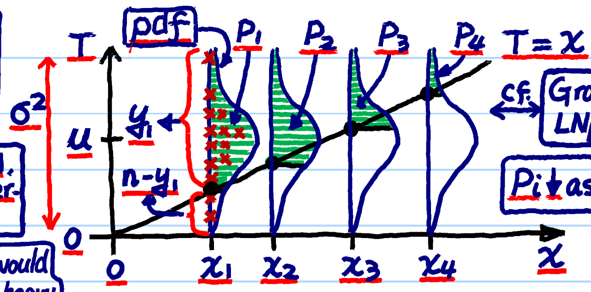
x_i = degree of toxicity

$T_j\lesseqgtr x_i$ — die / survive

- $F = \Phi(\eta) \Rightarrow$ probit link Recall. $g^{-1} = F$: a cdf
- $F = e^{\eta}/(1+e^{\eta}) \Rightarrow$ logit link
- $F = 1 - e^{-e^{\eta}}$ (heavier tail) ← graph in LNp.3-5
  $\Rightarrow$ complementary log-log link

x: not observed. only observe y_i

When would T_i has heavy tail?



cf. Graph in LNp.2-14

$P_i \uparrow$ as $x_i \uparrow$

useful in interpreting β under logit link → **Odds** → Recall.

$= g(p_i)$, meaningful in terms of p_i?

linear → $\eta_i = \beta_0 + \beta_1 h_1(x_i) + \cdots + \beta_p h_p(x_i) = h_i^T \beta$

nonlinear → $\mu_i / n_i = p_i = g^{-1}(\eta_i) = g^{-1}(h_i^T \beta)$

- Definition of odds o_x (賠率)
  - ➢ Sometimes, a better scale than probability to present chance
  - ➢ Express the payoffs for bets $O(A) = \frac{P(A)}{P(A^c)}$ (an event A)

In LM, use difference. cf.

If it's a fair bet

- even bet: pay \$1 for every \$1 bet, odds=1 $\Rightarrow p$=0.5 $\boxed{1:1}$
- 3-1 against bet: pay \$3 for every \$1 bet, odds=1/3 $\Rightarrow p$=0.25 $\boxed{1/3:1}$ winning prob.
- 1-3 on bet: pay \$1 for every \$3 bet, odds=3 $\Rightarrow p$=0.75 $\boxed{3:1}$

compare 2 probabilities of 2 categories in terms of ratio

➢ $o_x = p_x/(1-p_x) \Rightarrow p_x = o_x/(1+o_x)$ $p_x \overset{1}{\underset{0}{\uparrow}} \Leftrightarrow o_x \overset{\infty}{\underset{0}{\uparrow}}$

$0 \le p_x \le 1$
$0 \le o_x < \infty$
$-\infty < \log(o_x) < \infty$

➢ A mathematical advantage of odds: unbounded above and unbounded above & below after taking log. Recall. problem with upper/lower bounds (LNp 3-12)

- odds and logit link (logistic regression)
  - ➢ Consider the model — an advantage of using logit link

its value can be intepreted as "log odds" at x

$$\underbrace{\log(o_{\mathbf{x}})}_{\text{log odds}} \overset{*}{=} \log\left(\frac{p_{\mathbf{x}}}{1-p_{\mathbf{x}}}\right) = \text{logit}(p_{\mathbf{x}}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 = \eta_{\mathbf{x}}$$

$$\log(o_{\mathbf{x}'}) = \text{logit}(p_{\mathbf{x}'}) = \beta_0 + \beta_1(x_1+1) + \beta_2 x_2 = \eta_{\mathbf{x}'} = \eta_{\mathbf{x}} + \beta_1$$

When changing from $\mathbf{x}$ to $\mathbf{x}'$,

$\log(O_{x'}) = \log(O_x) + \beta_1$ $\Rightarrow$ $\beta_1 = \log(O_{x'}) - \log(O_x) = \log\left(\frac{O_{x'}}{O_x}\right)$ ← log odds ratio of x' relative to x

taking exp: $O_{x'} = O_x \times e^{\beta_1}$ $\qquad e^{\beta_1} = \frac{O_{x'}}{O_x}$ ← odd ratio

at two covariate values $x=(x_1, x_2)$ and $x'=(x_1+1, x_2)$. they are already ratios

➢ The coefficient β_1 can be interpreted as: compare 2 odds in terms of ratio

A unit increase in x_1 with x_2 held fixed

Probit: $\eta = \Phi^{-1}(p)$ CLL: $\eta = \log[-\log(1-p)]$ meaning in terms of p? especially in interpreting β

- increases the log-odds of successes by β_1, or
- increases the odds of successes by a factor of $\exp(\beta_1)$

check in LNp. 3-5

➢ The usually interpretational difficulties regarding causation apply as in standard regression ← check LNp.1-19~21

➢ No such simple interpretation exists for other links, e.g.,