- Normal approximation might be too much a stretch when n_i's not large enough or $p_x \approx 1$ or 0 → Recall. LM data with response having upper/lower bound, & many observations of the response are close/equal to the bounds.
- Some of these problems could be corrected by using transformation and weighting

A suit fitting better for binomial response

eg. $\log(\hat{p}_x/(1-\hat{p}_x)) = X\beta + \varepsilon$, $0 \le$ predicted prob. ≤ 1

→ for { non-constant variance, non-linearity } → for non-constant variance

2/26

Generalized Linear Model for Binomial Data

- Recall: linear model

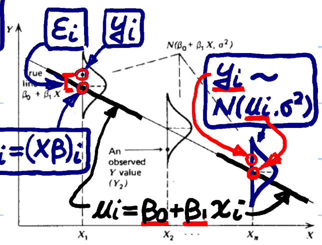
binomial response $\sim B(n_x, p_x)$

$\{0,1,\ldots,n_x\} \ni y_x = X\beta + \varepsilon$ ← cf.

equivalent in LM

➢ model description 1: $Y = X\beta + \varepsilon,\ \varepsilon \sim N(0, \sigma^2 I)$

➢ model description 2: $Y \sim N(X\beta, \sigma^2 I)$



response (隨機) → ⊙ $Y \sim N(\mu_x, \Sigma_x)$, $\mu_x \in (-\infty, \infty)$ — random; parameters (unknown but fixed values)

covariates (規律) → ⊙ $X\beta = \sum_{i=1}^{p} \beta_i \cdot h_i(X_1, \ldots, X_m) \equiv \eta_x(\beta)$ ← a function of x with a linear structure btwn h_i's & β_i's

coefficients (parameters); known functions (base functions); matrix form; functional form; expressed as a function or vector; $\in (-\infty, \infty)$

build the link btwn parameters in Y & $X\beta$: ⊙ $\mu_x = \eta_x$, $\Sigma_x = \sigma^2 I$ — can be assigned a function of x (with parameters) if necessary ★

Q: which description can be generalized to binomial data? — of main interest in regression

---

- 3 components in a generalized linear model (binomial example) y_i ★

response (隨機) ➢ $y_x \sim B(n_x, p_x)$ — random; & indep; fixed but known values; parameters (fixed but unknown) $\in [0,1]$

- $\mu_x = E(y_x) = n_x p_x$
- $\sigma_x^2 = Var(y_x) = n_x p_x(1-p_x) = [\mu_x(n_x - \mu_x)]/n_x$

σ_x^2 = a function of μ_x (cf. LM)

$p_i = \mu_i/n_i$ ↓↑ $\mu_i = n_i p_i$; p_i; g ↓↑ g^{-1}; η_i; $x_i^T\beta$; x_i; interchangeable

covariates (規律) ➢ $X\beta = \sum_{i=1}^{p} \beta_i \cdot h_i(X_1, \ldots, X_m) \equiv \eta_x(\beta) \in (-\infty, \infty)$ — fixed; a linear structure btwn h_i's & β_i's

Build the link btwn the parameters in y_x & $X\beta$: ➢ *link function* g: monotone and differentiable such that — inverse function g^{-1} exists

$\eta_x = g(p_x) \Rightarrow p_x = g^{-1}(\eta_x)$ [for binimial, $g:(0,1) \to (-\infty, \infty)$] — $\ni p_x$, $\ni \eta_x$

reduce the dimension of p_x from β to η; can be regarded as a cdf (LNp3-16); ① $\mu_i/n_i = p_i = g^{-1}(\eta_i) = g^{-1}(x_i^T\beta)$

- Common choices of link function for binomial data — a non-linear function ⇒ relationship btwn μ & x is not linear any more; ② different choice of g results in different interpretation of β

canonical link: ➢ Logit: $\eta_x = \log(p_x/(1-p_x))$ → logistic regression; odds → ; p_x ⇄ η_x (logit; logistic = logit^{-1})

X^TY: suff. stat. of β (check LNp3-17)

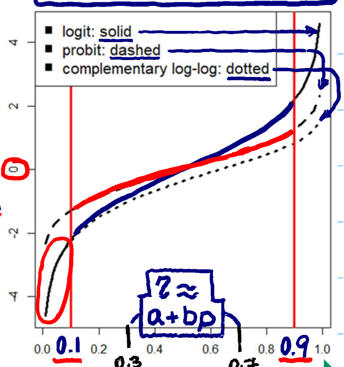
➢ Probit: $\eta_x = \Phi^{-1}(p_x)$, where Φ is the cdf of $N(0,1)$ — the pdfs of Φ & logistic are symmetric about 0; 0 ~ 1

➢ Complementary log-log: $\eta_x = \log(-\log(1-p_x))$ — $-\infty \sim 0$

➢ Note. — g^{-1} corresponds to a cdf with heavy tail (LNp 3-17)

- Logit is close to the complementary log-log when p_x is small — if all $\hat{p}_i$'s < 0.1, results not much difference under the 2 links
- Logit is close to probit when $0.1 < p_x < 0.9$

a linear transformation ← (exercise: compare the 3 functions)



- Recall. Likelihood-based approach for estimation and testing (textbook, Appendix A)
  - widely applicable
  - ➢ Estimation: maximum likelihood estimator (MLE)
  - ➢ Testing: likelihood-ratio (LR) test, Wald test — good statistics, asymptotic properties
- Log-likelihood of the binomial GLM (use logit link as an example):

under logit link:

$$l(\beta) = \sum_{i=1}^{k} \left[y_i \eta_{x_i}(\beta) - n_i \log\left(1 + e^{\eta_{x_i}(\beta)}\right) + \log \binom{n_i}{y_i} \right]$$

($\eta_{x_i}(\beta) = \theta_i$ ← canonical parameter (exponential family); the bracket is $l_i(\beta)$)

$$= \sum_{i=1}^{k} \left[y_i \log\left(\frac{p_i}{1-p_i}\right) - n_i \log\left(\frac{1}{1-p_i}\right) + \log\binom{n_i}{y_i} \right]$$

under any links:

$$= \sum_{i=1}^{k} \left[y_i \log(p_i(\beta)) + (n_i - y_i)\log(1 - p_i(\beta)) + \log\binom{n_i}{y_i} \right]$$

$$= \log\left[\prod_{i=1}^{k} \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i} \right]$$

(∵ independent; LNp.2-17, pmf of $B(n_i, p_i)$; Y_i indep $B(n_i, p_i)$, $i = 1, \cdots, k$)

$\eta = g(p) = \log\left(\frac{p}{1-p}\right)$
- $p = g^{-1}(\eta) = \frac{e^\eta}{1+e^\eta}$ (logistic)
- $1 - p = \frac{1}{1+e^\eta}$

$p : (1-p) = e^\eta : e^0$

$\eta \gg 0 \Rightarrow p \gg 1-p$

$\eta \ll 0 \Rightarrow p \ll 1-p$

$\eta \approx 0 \Rightarrow p, 1-p \approx 0.5$

- Estimation of β
  - ➢ Recall: in linear model, the LS estimator of β is also the MLE → $(X^TX)^{-1}X^TY$ (good projection properties)
  - ➢ For GLM, the concept of LS not appropriate any more (Why? check model description 1 in LNp.3-4)
    - ⇒ still can adopt the method of MLE ← only needs likelihood
    - ⇒ maximizing $l(\beta)$ as a function of β to find MLE (But, no projection explanation)
  - ➢ Obtain MLE by solving $\partial l(\beta)/\partial\beta = \mathbf{0}$: partial derivative of each β_r, $r = 1, \cdots, p$

For one covariate class (x_i, y_i):

$$\frac{\partial l_i}{\partial \beta_r} = \frac{\partial l_i}{\partial \theta_i} \times \frac{\partial \theta_i}{\partial p_i} \times \frac{\partial p_i}{\partial \eta_i} \times \frac{\partial \eta_i}{\partial \beta_r}$$

(chain rule; θ_i: canonical parameter in exponential family; $\frac{\partial l_i}{\partial \theta_i}\times\frac{\partial \theta_i}{\partial p_i} = \partial l_i/\partial p_i$ — binomial pmf; $\frac{\partial p_i}{\partial \eta_i}$ — g^{-1} link; $\frac{\partial \eta_i}{\partial \beta_r}$ — $X\beta$ 規; Check ★ in LNp.3-5)

for any links:

$$\frac{\partial l_i}{\partial p_i} = \frac{\partial}{\partial p_i}\left[y_i \log(p_i) + (n_i - y_i)\log(1-p_i)\right] = \frac{y_i}{p_i} - \frac{n_i - y_i}{1-p_i} = \frac{y_i - n_i p_i}{p_i(1-p_i)}$$

$$\frac{\partial \eta_i}{\partial \beta_r} = \frac{\partial}{\partial \beta_r} \sum_{j=1}^{p} \beta_j h_j(x_i) = h_r(x_i) = x_{ir}$$

for logit link:

$$\frac{\partial p_i}{\partial \eta_i} = \frac{\partial}{\partial \eta}\frac{e^\eta}{1+e^\eta} = \frac{e^\eta(1+e^\eta) - e^\eta \cdot e^\eta}{(1+e^\eta)^2} = \frac{e^\eta}{1+e^\eta} \times \frac{1}{1+e^\eta} = p_i(1-p_i)$$

For all (k) covariate classes

- for any links: for $r = 1, \cdots, p$

$$0 = \sum_{i=1}^{k} \frac{y_i - n_i p_i}{p_i(1-p_i)} \times \frac{\partial p_i}{\partial \eta_i} \times x_{ir}$$

- for logit link: for $r = 1, \cdots, p$

$$0 = \sum_{i=1}^{k} \frac{y_i - n_i p_i}{p_i(1-p_i)} \times p_i(1-p_i) \times x_{ir} = \sum_{i=1}^{k} (y_i - \mu_i) x_{ir} \Rightarrow X^T(Y - \mu) = 0$$

$\Rightarrow X^TY = X^T\mu(\beta)$: normal equation (Suff. stat. of β)

Recall. In LM, $X^TY = X^T\mu = X^TX\beta \Rightarrow \hat\beta = (X^TX)^{-1}X^TY$

But, in binomial GLM,

$\mu = n \circ p = n \circ g^{-1}(\eta) = n \circ g^{-1}(X\beta) \neq X\beta$ (element-wise multiplication)

- ➢ Usually no explicit formula for MLE of β in GLM

  $\hat\beta(\text{MLE}) \to X\hat\beta = \hat\eta \to g^{-1}(\hat\eta) = \hat p \to n \circ \hat p = \hat\mu(\hat Y)$ ← a function of X with unknown parameters (In LM, for non-constant variances modeled by … (check ★ in LNp.3-4))

- ➢ An algorithm (related to IRWLS) to perform the maximization will be discussed in future lecture (Why? Note that $Var(y_x)$ is a function of $E(y_x)$ (LNp.3-5))

  In LM, $cov(\hat\beta) = (X^TX)^{-1}\hat\sigma^2$; cf. $cov(\hat\beta) = (X^TWX)^{-1}\hat\sigma^2$

- ➢ Q: What is the estimated covariance matrix of the MLE?

Inference

Test and Confidence Interval

- RSS (in LM) and deviance (in GLM)
  - Recall: in linear model, RSS play a critical role in inference
  - Q: How to evaluate whether y_i & $\hat{y}_i$ are close enough in GLM?
  - Q: In GLM, what is the concept similar to RSS?

residual $y_i - \hat{y}_i$

① estimate β: $\arg\min_\beta RSS(\beta)$ = LS est'or
② estimate σ^2: $\hat{\sigma}^2 = RSS/(n-p)$
③ do test $\omega \subset \Omega$: use RSS_ω, RSS_Ω
④ evaluate goodness-of-fit: $R^2 = 1 - RSS/TSS$

check LNp.1-13 Test in LM

- consider two models L and S (l, s: # of parameters in $X\beta$)
  - a larger model L: l parameters and likelihood $\mathcal{L}_L$
  - a smaller model S: s ($s < l$) parameters and likelihood $\mathcal{L}_S$
  - S is nested in L ($S \subset L$), e.g., any joint distribution of data in $S \in L$

L: $X_L\beta_L = \eta_L = g(\mu_L)$ ← $H_0 \cup H_1$, $A\beta_L \neq c$
↓ & H_0: $A\beta_L = c$
S: $X_S\beta_S = \eta_S = g(\mu_S)$ ← H_0

dim = s, $span(X_S) \subset span(X_L)$, dim = l

likelihood ratio $-2\log\left(\frac{\mathcal{L}(\hat{\beta}_S)}{\mathcal{L}(\hat{\beta}_L)}\right)$

- To test H_0: S (say, $A\beta = c$) vs. H_1: $L \setminus S$, likelihood methods suggests the likelihood ratio statistics:

$$2[l(\hat{\beta}_L) - l(\hat{\beta}_S)] = 2\log\left[\mathcal{L}(\hat{\eta}_L = X_L\hat{\beta}_L)/\mathcal{L}(\hat{\eta}_S = X_S\hat{\beta}_S)\right] \overset{a}{\simeq} \chi^2_{df_{L\setminus S}}$$

approximately or asymptotically (n_i's large)

log-likelihood (LNp.3-6); MLE under L; MLE under S

where $df_{L\setminus S} = \dim(L) - \dim(S) = l - s$.

- Suppose that the larger model L is saturated.

an identifiable model with maximum # of parameters

- For R covariate classes with distinct x_i's, a saturated L has R parameters in β.
- Say, $X = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}_{R\times R}$, $\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_R \end{bmatrix}$ ⇒ $\mu_i = n_i p_i = n_i g^{-1}(\eta_i) = n_i g^{-1}(\beta_i)$, $i = 1, \ldots, R$

can freely change in estimation; $(X\beta)_i$

MLE $\hat{\beta}_i$ would make $\hat{p}_i = y_i/n_i$, $\hat{\mu}_i = y_i$

$\hat{p}_x = y_x/n_x$; saturated model S; x

For other saturated X^*, $span(X^*) = span(X)$

We have $\hat{y}_{i,L} = y_i$ under L, and the LR test statistic becomes:

under any links; cf. saturated LM, $\hat{y}_x = \bar{y}_x$ (note 1 y in 1 covariate class)

$$D_S = 2\sum_{i=1}^{k}\left\{y_i \log(y_i/\hat{y}_{i,S}) + (n_i - y_i)\log\left[(n_i - y_i)/(n_i - \hat{y}_{i,S})\right]\right\}$$

$2[l(\hat{\beta}_L) - l(\hat{\beta}_S)]$; check LNp.3-6

# of sucess; $y_i \leftrightarrow \hat{y}_{i,S}$; ≈ 1, if $y_i \approx \hat{y}_{i,S}$; # of failure; $n_i - y_i \leftrightarrow n_i - \hat{y}_{i,S}$; ≈ 1, if $y_i \approx \hat{y}_{i,S}$

where $\hat{y}_{i,S}$ is the fitted values from the smaller model S. ($n_i \times \hat{p}_{i,S}$)

- D_S is called *deviance* of S, which plays a role similar to RSS

check ① – ④ in LNp.3-8

small D_S better fit; $\hat{y}_i \leftrightarrow y_i$

cf. RSS = $\sum_i (y_i - \hat{y}_i)^2$

- Since the saturated model fits as well as any model can fit, the deviance D_S measures how close the (smaller) model S comes to the perfection (i.e., D=0 under saturated model).

prefer S with small D_S & fewer parameters

In LM, can use (i) $\hat{\sigma}^2 = RSS/(n-p)$ (ii) $R^2 = 1 - \frac{RSS}{TSS}$

(similar)

- Deviance can be treated as a measure of goodness-of-fit

can do test !!!

Compare S with saturated model (most complicated) ↔ (cf.) compare S with intercept-only model (simplest)