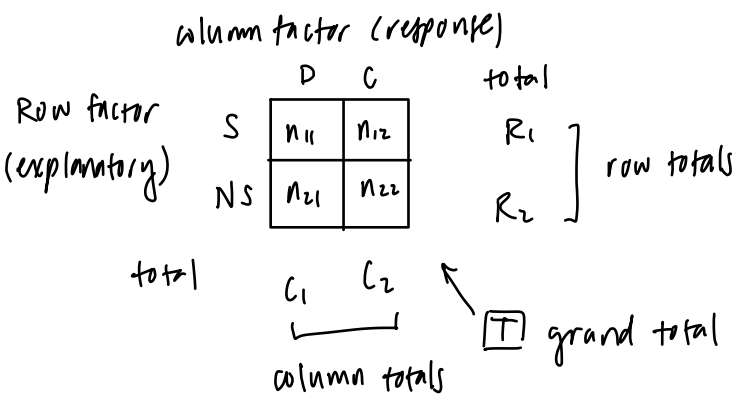
Jan 26

Table of counts



S: smoker
NS: nonsmoker
D: disease
C: control

- Poisson sampling
- Multinomial sampling
- Randomized experiment
- Retrospective product binomial

fix column totals first

- Prospective product binomial

fix row totals first

- Hypothesis of homogeneity : $H_0: \pi_1 = \pi_2$, $\frac{\omega_1}{\omega_2} = 1$
- Hypothesis of independence : $H_0: \pi_1 = \pi_2$,

π_1: P(observe D in first row)

π_2: P(observe D in second row)

odds

$\omega_1 = \frac{\pi_1}{1-\pi_1}$

$\omega_2 = \frac{\pi_2}{1-\pi_2}$

odds ratio

$\frac{\omega_1}{\omega_2}$

Fix marginal totals or grand totals?	Type of sampling	Hypothesis homo	Hypothesis indep	π_1	π_2	$\pi_1 - \pi_2$	ω_1	ω_2	ω_1/ω_2	p (proportion of combined pop. w/ explanatory level = 1)	y (proportion of lung cancer patients in pop.)
None	Poisson	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
grand total	Multinomial	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Row totals	Prospective product binomial	✓	×	✓	✓	✓	✓	✓	✓	×	×*
Column totals	Retrospective product binomial	✓	×	×	×	×	×	×	✓	×*	×
Row totals	Randomized experiment	✓	×	✓	✓	✓	✓	✓	✓	×	×

* can be done when $\frac{\omega_1}{\omega_2} = 1$

b/c otherwise you don't know true proportions of D vs. C in population

$$\frac{\omega_1}{\omega_2} = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}} = \frac{\pi_1}{\pi_2} \cdot \frac{1-\pi_2}{1-\pi_1} = \frac{\frac{\pi_1}{\pi_2}}{\frac{1-\pi_1}{1-\pi_2}}$$

both of these ratios can be correctly estimated using retrospective product binomial

$H_0: \pi_1 = \pi_2$

test proportions:

z-test:

$$\hat{\pi}_1 = \frac{n_{11}}{R_1}, \quad \hat{\pi}_2 = \frac{n_{21}}{R_2}$$

$$\hat{\pi}_c = \frac{n_{11} + n_{21}}{R_1 + R_2}$$ (assume two rows have common π)

$$E(\hat{\pi}_1 - \hat{\pi}_2) = \pi_1 - \pi_2, \quad var(\hat{\pi}_1 - \hat{\pi}_2) = \frac{\pi_1(1-\pi_1)}{R_1} + \frac{\pi_2(1-\pi_2)}{R_2}$$

$$z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\frac{\hat{\pi}_c(1-\hat{\pi}_c)}{R_1} + \frac{\hat{\pi}_c(1-\hat{\pi}_c)}{R_2}}} \sim N$$

chi-squared test:

$$\chi^2 = \sum_{ij} \frac{(Y_{obs} - Y_{exp})^2}{Y_{exp}}$$

$$\sim \chi^2 \quad df = \underbrace{(p-1)(q-1)}_{p \text{ rows, } q \text{ columns}}$$

q

Obs

n_{11}	n_{12}	R_1
n_{21}	n_{22}	R_2
C_1	C_2	T

p

Exp

$\frac{R_1 C_1}{T}$	$\frac{R_1 C_2}{T}$
$\frac{R_2 C_1}{T}$	$\frac{R_2 C_2}{T}$

Fisher's Exact test:

z-test and χ^2 rely on asymptotic distributions. Fisher's does not rely on asymptotic assumptions (can be used for small sample size) (hypergeometric dist.)

$$Pr(n_{11} = k) = \frac{\binom{R_1}{k}\binom{R_2}{C_1 - k}}{\binom{T}{C_1}}$$

can use this to compute a one-sided p-value

If $n_{11} > \frac{R_1 C_1}{T}$ then $p\text{-val} = \sum_{k=n_{11}}^{\min(R_1, C_1)} Pr(n_{11} = k)$

If $n_{11} < \frac{R_1 C_1}{T}$ then $p\text{-val} = \sum_{k=0}^{n_{11}} Pr(n_{11} = k)$

Binary Outcome

$Ber(\pi)$ Binomial data:

$$Y \sim Bin(m, \pi)$$

(binomial denominator)

(i) $Y = Y_1 + \dots + Y_n$

(ii) $Y_1 \sim Pois(m_1)$
$Y_2 \sim Pois(m_2)$

$$Pr(Y_1 = y \mid Y_1 + Y_2 = m) = \binom{m}{y} \pi^y (1-\pi)^{m-y}$$

$$\pi = \frac{m_1}{m_1 + m_2}$$

Moments and cumulants

For Bernoulli:

MGF: $M_Y(\gamma) = E e^{\gamma Y}$
$= (1-\pi) + \pi e^{\gamma}$

cumulant generating function:

$K_Y(\gamma) = \log M_Y(\gamma)$
$= \log((1-\pi) + \pi e^{\gamma})$

1st cumulant = mean
2nd cumulant = 2nd central moment
3rd cumulant = 3rd central moment

no corresponding relationship for 4th and beyond.

For Binomial:

MGF : $M_Y(\gamma) = (1-\pi+\pi e^{\gamma})^m$

cumulant generating function:

$$K_Y(\gamma) = m \log(1-\pi+\pi e^{\gamma})$$

Taylor expansion of CGF:

$$K_Y(\gamma) = K_Y(0) + K_Y'(0)\gamma + \frac{1}{2!}K_Y''(0)\gamma^2 + \dots$$

$$K_Y'(0) = \frac{m\pi e^{\gamma}}{1-\pi+\pi e^{\gamma}}\Big|_{\gamma=0} = m\pi$$

$$K_Y''(0) = m\pi\left(\frac{e^{\gamma}}{1-\pi+\pi e^{\gamma}} - \frac{\pi e^{2\gamma}}{(1-\pi+\pi e^{\gamma})^2}\right)$$

$$= m\pi \frac{(1-\pi)e^{\gamma}}{(1-\pi+\pi e^{\gamma})^2}\Big|_{\gamma=0} = m\pi(1-\pi)$$

$$K_Y'''(0) = m\pi(1-\pi)(1-2\pi)$$

$$K_Y^{(4)}(0) = m\pi(1-\pi)[1-6\pi(1-\pi)]$$

derive as exercise

Jan 31

Cumulant generating function:

$$K_Y(t) = \log M_Y(t)$$

Binomial: $K_Y(t) = m \log\{1 - \pi + \pi e^t\}$

(CGF for Bernoulli)

Normal limit

$$Z = \frac{Y - m\pi}{\sqrt{m\pi(1-\pi)}} \xrightarrow{m \to \infty} N(0,1)$$

Standard Normal CGF: $t^2/2$ (derive as exercise)

$0 \quad 1 \quad 0 \quad \cdots \quad 0$

Scale Y:

original CGF: $E e^{tY}$

$= K_0 + K_1 t + \frac{K_2}{2!} t^2 + \cdots$

after scaling: $\log E e^{tcY}$, $\tilde{Y} = cY$

$= K_0 + K_1 ct + \frac{K_2}{2!}(ct)^2 + \cdots$

$K_0 \quad cK_1 \quad c^2 K_2 \quad \cdots$

$\sqrt{m}$ serves as scaling factor

$N(0,1)$: $\underset{K_1}{0} \quad \underset{K_2}{1} \quad \underset{K_3}{0} \quad \cdots \quad 0$

Z: $0 \quad 1 \quad 0\left(\frac{1}{\sqrt{m}}\right) \quad 0\left(\frac{1}{m}\right) \quad 0\left(\frac{1}{m^{3/2}}\right)$

$\downarrow \quad \downarrow \quad \downarrow$

$0 \quad 0 \quad 0$

slowest converging term

convergence rate $= \sqrt{m}$

Poisson limit:

Poisson CGF: $n(e^t-1)$

all cumulants determined by n (show w/ Taylor expansion)

$Y: m\log(1-\pi+\pi e^t)$

consider the following case:

$m \to \infty$, $\pi \to 0$ s.t. $m\pi \to n$

$m\log(1-\pi+\pi e^t)$

$$\to \frac{n}{\pi}\log(1+\pi(e^t-1))$$

$$= \frac{n}{\pi}\left(\pi(e^t-1) - \frac{1}{2}\pi^2(e^t-1)^2 + \frac{1}{3}\pi^3(e^t-1)^3 \cdots\right)$$

$$= n(e^t-1) - \frac{1}{2}n\pi(e^t-1)^2 + \frac{1}{3}n\pi^2(e^t-1)^3 \cdots$$

poisson CGF; $\frac{n}{m}$; $O(\frac{1}{m})$ → 0; $O(\frac{1}{m^2})$ → 0

$\Rightarrow$ Binomial can be approximated w/ poisson

convergence rate = $1/m$ (faster than $\frac{1}{2}$)

Models for Binary / Binomial Responses

y $\quad$ $Ber(\pi)$ $\quad$ $0 \le \pi \le 1$

$\quad$ $Binomial(m, \pi)$

- need to choose a link function for $X\beta$, π.
  can't use identity b/c $X\beta$ is not bounded.

(1) logit / logistic function

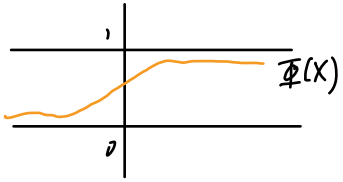
$$g_1(\pi) = \log \frac{\pi}{1-\pi}$$

inverse CDF for logistic distribution:

$$\frac{e^{\eta}}{1-e^{\eta}}$$

(2) probit / inverse Normal function

$$g_2(\pi) = \Phi^{-1}(\pi)$$



(3) complementary log-log function

$$g_3(\pi) = \log\{-\log(1-\pi)\}$$

extreme value distribution CDF:

$$1-e^{-e^{\eta}}$$

(4) log-log link

$$g_4(\pi) = -\log\{-\log(\pi)\}$$

extreme value CDF:

$$e^{-e^{\eta}}$$

- g_1 and g_2 are symmetric around $1/2$.

$$g_1(\pi) = g_1(1-\pi)$$

$$g_2(\pi) = g_2(1-\pi)$$

- g_3 and g_4 are related.

$$g_3(\pi) = -g_4(1-\pi)$$

- g_1 is most commonly used.

Logistic Regression

$$\log \frac{\pi}{1-\pi} = X\beta \qquad Y \sim Bin(m, \pi)$$

ex: $X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

1. parameter interpretation:

β_0: log odds of getting disease when $X_1, X_2 = 0$

$$\frac{e^{\beta_0}}{1+e^{\beta_0}} = \pi$$

β_1: hold all other variables constant, change in log odds w/ one unit increase of X_1

$$\log \frac{\pi_1}{1-\pi_1} = \beta_0 + \beta_1(X_1+1) + \beta_2 X_2$$

$$\log \frac{\pi_2}{1-\pi_2} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$\log \frac{\pi_1}{1-\pi_1} - \log \frac{\pi_2}{1-\pi_2} = \beta_1$$

$$= \log \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}}$$

β_1 characterizes a log-odds ratio

$e^{\beta_1} \rightarrow$ odds ratio

- Randomness incorporated into distributional assumption, no error term.

Retrospective trial:

β_1 and β_2 still reflect population behavior.

Assume this true model for the population:

$$\text{logit } P(D|X) = \alpha + \beta^T X \qquad \bar{D} = \text{no disease}, \; D = \text{disease}$$

but data collected is not representative of pop.

$Z=1$ patient sampled

$Z=0$ patient not sampled

$$\pi_1 = P(Z=1|D) \qquad \pi_0 = P(Z=1|\bar{D})$$

fit: $\text{logit } P(D|X, Z=1)$

$$P(D|X,Z=1) = \frac{P(D, Z=1|X)}{P(Z=1|X)}$$

$$= \frac{P(Z=1|D,X)\,P(D|X)}{P(Z=1|D,X)P(D|X) + P(Z=1|\bar{D},X)P(\bar{D}|X)}$$

$$= \frac{\pi_1 P(D|X)}{\pi_1 P(D|X) + \pi_0 P(\bar{D}|X)}$$

$$= \pi_1 \frac{\frac{e^{\alpha+\beta^T x}}{1+e^{\alpha+\beta^T x}}}{\pi_1 \frac{e^{\alpha+\beta^T x}}{1+e^{\alpha+\beta^T x}} + \pi_0 \frac{1}{1+e^{\alpha+\beta^T x}}}$$

$$= \frac{\pi_1 e^{\alpha+\beta^T x}}{\pi_1 e^{\alpha+\beta^T x} + \pi_0}$$

$$= \frac{\frac{\pi_1}{\pi_0} e^{\alpha+\beta^T x}}{\pi_1 e^{\alpha+\beta^T x} + 1}$$

$$P(D|X, z=1) = \frac{e^{\log\frac{\pi_1}{\pi_0} + \alpha + \beta^T x}}{e^{\log\frac{\pi_1}{\pi_0} + \alpha + \beta^T x} + 1}$$

$$\text{logit}\, P(D|X, z=1) = \underbrace{\log\frac{\pi_1}{\pi_0}}_{\text{offset}} + \alpha + \beta^T x$$

without sampling probabilities you can only talk about β's (can't interpret intercept)

$$\alpha^* = \alpha + \log \frac{\pi_1}{\pi_0}$$

log-likelihood:

$$\ell(\pi, y) = \sum_{i=1}^{n} \log \binom{m_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{m_i - y_i}$$

$$= \sum_{i=1}^{n} \left[y_i \log \frac{\pi_i}{1-\pi_i} + m_i \log(1-\pi_i) \right] + \text{const}$$

$\log \frac{\pi_i}{1-\pi_i}$: logit of x_{ij}

const: anything w/o unknown param

$$\ell(\beta, y) = \sum_{i=1}^{n} \sum_{j} y_i x_{ij} \beta_j - \sum_{i=1}^{n} m_i \log\left(1 + e^{\sum_j x_{ij}\beta_j}\right)$$

$$= \sum_{i=1}^{n} y_i x_i^T \beta - \sum_{i=1}^{n} m_i \log\left(1 + e^{x_i^T \beta}\right)$$

$$= y^T X \beta - \sum_{i=1}^{n} m_i \log\left(1 + e^{x_i^T \beta}\right)$$

suff. stat: $X^T y$

MLE: $\ell'(\beta, y) = 0$

$$X^T y - \sum_i m_i \frac{x_i}{1 + e^{x_i^T \beta}} = 0$$

not linear in terms of β

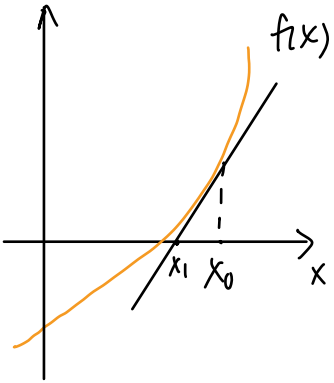
Feb 2 HW: Read chp. 5 of text

logistic regression MLE

$$\ell(\beta, y) = \sum_{i=1}^{n} \left[y_i \log \frac{\pi_i}{1-\pi_i} + m_i \log(1-\pi_i) + c \right]$$

$$= y^T X\beta - \sum_{i=1}^{n} m_i \log\left(1 + e^{x_i^T \beta}\right)$$

$\frac{\partial \ell}{\partial \beta} = 0$, solve using Newton-Raphson



$y - f(x_0) = f'(x_0)(x - x_0)$

$y = f(x_0) + f'(x_0)(x - x_0)$

$= 0$

$x = x_0 - \frac{f(x_0)}{f'(x_0)}$

x_0

$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$

$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$

$\vdots$

$x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)}$

$\vdots$

converges to root of $f(x)$ after iteratively running update

if β is multi-dimensional:

$$\beta_1 = \beta_0 - \left[\mathbb{E}\,\ell''(\beta_0)\right]^{-1}\ell'(\beta_0)$$

(Fisher info: $\mathbb{E}\,\ell''(\beta_0)$; score: $\ell'(\beta_0)$)

↑ usually take expectation b/c X is an RV (Fisher scoring)

$$\vdots$$

$$\beta_{t+1} = \beta_t - \left[\mathbb{E}\,\ell''(\beta_t)\right]^{-1}\ell'(\beta_t)$$

Deriving a concrete form:

$$(1)\quad \frac{\partial \ell}{\partial \beta} = X^T y - \sum_{i=1}^{n} \frac{m_i e^{x_i^T\beta}}{1+e^{x_i^T\beta}} x_i$$

$$= X^T y - (x_1\ x_2\ \dots\ x_n)\begin{pmatrix} m_1\pi_1 \\ m_2\pi_2 \\ \vdots \\ m_n\pi_n \end{pmatrix}$$

$$= X^T(y-\mu)$$

X^Ty is a sufficient statistic for β, $\mathbb{E}(X^Ty) = X^T\mu$

Another way:

$$\frac{\partial \ell}{\partial \beta} = \begin{bmatrix} \frac{\partial \ell}{\partial \beta_1} \\ \vdots \\ \frac{\partial \ell}{\partial \beta_p} \end{bmatrix}$$

$$\frac{\partial \ell}{\partial \beta_r} = \sum_{i=1}^{n} \frac{\partial \ell}{\partial \pi_i} \frac{\partial \pi_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_r}$$

π is a function of linear predictor (link function) (η_i is linear predictor)

$$= \sum_{i=1}^{n} \frac{y_i - m_i \pi_i}{\pi_i (1-\pi_i)} \pi_i (1-\pi_i) x_{ir}$$

$$\pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

$$= \sum_{i=1}^{n} (y_i - m_i \pi_i) x_{ir}$$

$$\frac{d\pi_i}{d\eta_i} = \frac{e^{\eta_i}}{1+e^{\eta_i}} - \frac{e^{2\eta_i}}{(1+e^{\eta_i})^2}$$

$$= \pi_i (1-\pi_i)$$

r^{th} entry of the gradient vector

Second derivative: (only π_i involves β in $\frac{\partial \ell}{\partial \beta_r}$)

$$-\mathbb{E} \frac{\partial^2 \ell}{\partial \beta_r \partial \beta_s} = \mathbb{E} -\sum_{i=1}^{n} -m_i x_{ir} \frac{d\pi_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_s}$$

same result with or without taking expectation

$$= \mathbb{E} \sum_{i=1}^{n} m_i x_{ir} \pi_i (1-\pi_i) x_{is}$$

$$= \sum_{i=1}^{n} m_i \pi_i (1-\pi_i) x_{ir} x_{is}$$

$$-\mathbb{E}\frac{\partial^2 \ell}{\partial \beta^2} = X^T W X$$

where $W = \text{diag}\{m_i \pi_i (1-\pi_i)\}$

$m_i \pi_i (1-\pi_i)$: variance for the ith subject

So iterations for Newton-Raphson look like:

$$\beta_{t+1} = \beta_t + (X^T W X)^{-1} X^T (y - \mu_t)$$

W and μ_t: function of β_t

Another perspective: iterated least squares

$$\beta_1 = \beta_0 + (X^T W X)^{-1} X^T (y - \mu)$$

$$= \beta_0 + (X^T W X)^{-1} X^T W W^{-1} (y - \mu)$$

$$= (X^T W X)^{-1} \{(X^T W X)\beta_0 + X^T W W^{-1}(y - \mu)\}$$

$$= (X^T W X)^{-1} X^T W (X\beta_0 + W^{-1}(y - \mu))$$

$$z = X\beta_0 + W^{-1}(y - \mu) \rightarrow \text{adjusted response}$$

linear predictor

observed - expected error term

→ looks like response variable in linear models

$$\beta_1 = (X^T W X)^{-1} X^T W z$$

weighted least squares estimate for β

$$z = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \alpha W^{-1})$$

alternatively: $W^{1/2} z = W^{1/2} X\beta + W^{1/2}\varepsilon$

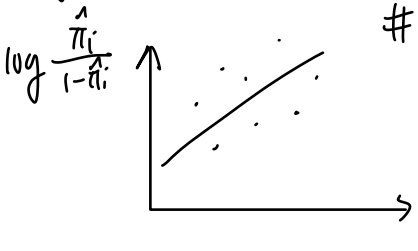
$$\tilde{z} = \tilde{X}\beta + \tilde{\varepsilon} \sim N(0, \alpha I)$$

$$(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{z} = (X^T W^{1/2} W^{1/2} X)^{-1} X^T W^{1/2} W^{1/2} z$$

$$= (X^T W X)^{-1} X^T W z$$

estimating β_0:

have to assume that # obs > # params



Another perspective: link function

higher order terms

$$g(y) = g(\mu) + g'(\mu)(y-\mu) + O(\quad)$$

$$= X\beta + \frac{d\eta}{d\mu}(y-\mu)$$

based on model assumptions

So adjusted response z is using Taylor expansion to approximate link function applied to observed data.

Also called iteratively re-weighted least squares (IRWLS)

Inference:

- for MLEs, $\text{var}\ \hat{\beta}_{mle} \longrightarrow$ inverse of fisher info asymptotically

- when the algorithm converges, we already obtain the fisher info.

$$\text{var}(\hat{\beta}) \cong (X^T W X)^{-1}$$

β_j $\quad logit(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

$$z_j = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\widehat{var}(\hat{\beta}_j)}}$$

would test

/

asymptotic variance, w_{jj}

Deviance

Residual deviance $\triangleq 2(l_{max} - l)$

maximum possible l w/ fully saturated model

(each observation has its own π_i w/o being constrained by shared function for all obs)

Binomial logistic

$$D(y, \hat{\pi}) = 2\{l(\tilde{\pi}, y) - l(\hat{\pi}, y)\}$$

$$= 2\sum_i \left\{ y_i \log \frac{y_i}{m_i \hat{\pi}_i} + (m_i - y_i) \log\left(\frac{1 - \frac{y_i}{m_i}}{1 - \hat{\pi}_i} \right) \right\}$$

small deviance → model sufficient to describe the data.

$$D \sim \chi^2_{n-p} \quad \text{(under certain conditions)}$$

can use χ^2 to determine if D is big or small

mean $\chi^2_{n-p} = n-p$

cannot be used for binary outcomes

Drop-in-Deviance test:

for comparing a null model w/ an alternative model.

H_0: simpler model

H_1: more complex model

Drop-in-deviance statistic:

$$D(y_i, \hat{\mu}_0) - D(y_i, \hat{\mu}_1)$$

($\hat{\mu}_0$: MLE from null model; $\hat{\mu}_1$: MLE from alternative model)

$$= 2(\ell_{max} - \ell_0) - 2(\ell_{max} - \ell_1) = 2(\ell_1 - \ell_0)$$

(likelihood ratio)

from likelihood ratio test → $\sim \chi^2_{p_1 - p_0}$

- this is an asymptotic result.

either are ok

$m \to \infty$ p_1, p_0 fixed.

as $n \to \infty$, per-parameter information increases

(different from deviance test)

Feb 9

· models for polytonous data

· polytonous response: has several categories

ex: Blood types (unstructured) $\leftarrow \pi_A \ \pi_B \ \pi_{AB} \ \pi_O$

levels of education (ordered)

$i \quad \underbrace{\pi_1, \ldots, \pi_K}_{\text{response probabilities (not ordered)}}, \quad \sum_{i=1}^{K} \pi_K = 1$

$\gamma_1 = \pi_1, \ \gamma_2 = \pi_1 + \pi_2, \ \gamma_3 = \pi_1 + \pi_2 + \pi_3 \ \ldots$

for ordered responses

(i.e. $\gamma_2 = P(\text{middle} + \text{high school})$

} cumulative response probabilities

· measurement scales

pure scales {

- nominal scales: categories are exchangeable devoid of structure
- ordinal scales: categories are ordered but does not make sense to talk about "distance" between 2 categories
- interval scales: categories are ordered, and with scores, can talk about distance

Compound scales:

ex: bivariate response $\begin{pmatrix} y \\ z \end{pmatrix}$ → ordinal, → binary

binary response { nominal, ordinal

· Modelling multi-category responses

$$y_i \sim x_i$$

$$\begin{pmatrix} \vdots \\ \vdots \end{pmatrix} \quad \pi_i$$

Nominal. $y \sim \text{multinomial}(m, \pi = (\pi_1 \dots \pi_k))$ (m: # trials)

$$\pi \sim X \text{ and } \sum_{i=1}^{k} \pi_i = 1$$

Commonly used model: baseline category logit model (BCLM)

$$\log \frac{\pi_j^{(x)}}{\pi_k^{(x)}} = \theta_j + \beta_j^T x \qquad j = 1, \dots, k-1$$

(θ_j: intercept; β_j: vector containing regression coefs)

similar to binomial logistic regression

will have $K-1$ models. This set of models defines the BCLM.

interpretation:

$$\beta_j = \begin{pmatrix} \beta_{j1} \\ \vdots \\ \beta_{jp} \end{pmatrix} \rightarrow \beta_{jq}$$

$$x \quad x + \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \rightarrow \beta_{jq} = x + e_q \triangleq x'$$

$$\log \frac{\pi_j(x)}{\pi_K(x)} = \theta_j + \beta_j^T x$$

$$\log \frac{\pi_j(x')}{\pi_K(x')} = \theta_j + \beta_j^T K + \beta_j^T e_q$$

$$\log \frac{\pi_j(x')}{\pi_K(x')} - \log \frac{\pi_j(x)}{\pi_K(x)} = \beta^T e_q = \beta_{jq}$$

} difference between log odds ratio between two categories w/ 1-unit increase in q^{th} covariate

· model for ordinal scales

cumulative logit model

$$\text{logit}(\gamma_j) = \theta_j - \beta_j^T x$$

$$\log \frac{\gamma_j}{1-\gamma_K} = \log \frac{P(Y \le j)}{P(Y > j)} = \log \frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_K}, \quad j = 1, \ldots K-1$$

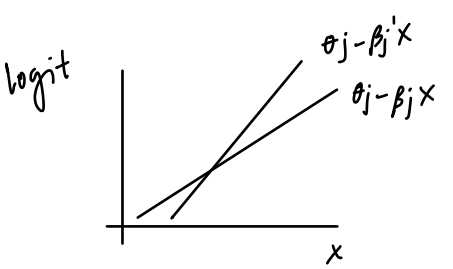
$$\gamma_K = 1$$

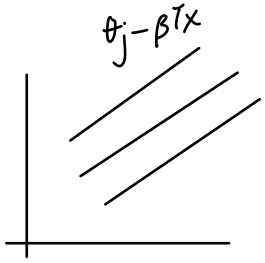
have an intercept and vector of coefs for each model

this and BCLM have the same complexity + same response probabilities after fitting

This model is naive b/c it doesn't consider ordering.

$$\gamma_1 \le \gamma_2 \le \cdots \le \gamma_{K-1} \le \gamma_K$$





higher categories should have higher intercepts

same slope for all categories

$$\text{logit}\, \gamma_j = \theta_j - \beta^T x$$

$$\theta_1 \le \theta_2 \le \cdots \le \theta_{K-1}$$

modified cumulative logit model (proportional odds model)

$$\log \frac{\gamma_j(x_1)}{1-\gamma_j(x_1)} \qquad \theta_j - \beta^T x_1$$

$$\log \frac{\gamma_j(x_2)}{1-\gamma_j(x_2)} \qquad \theta_j - \beta^T x_2$$

difference:

$$\frac{\dfrac{\gamma_j(x_1)}{1-\gamma_j(x_1)}}{\dfrac{\gamma_j(x_2)}{1-\gamma_j(x_2)}} = e^{-\beta^T(x_1-x_2)}$$

education level example:

M H C G

$$\frac{\dfrac{\gamma_H}{1-\gamma_H}}{\dfrac{\gamma_M}{1-\gamma_M}} = \frac{\dfrac{\gamma_C(x_1)}{1-\gamma_C(x_1)}}{\dfrac{\gamma_G(x_1)}{1-\gamma_G(x_1)}}$$

different categories show the same behavior in the odds ratio as a function of the covariates

ex: integrating surveys

$y^{(1)}$: liberal moderate conservative

↑ β

(2)

very liberal liberal moderate conservative very conservative

↑ β

can use same proportional odds model to describe both datasets and then integrate them.

- proportional odds model can also be viewed from a latent variable perspective.

Z: unobserved RV. Assume $Z - \beta^T X \sim$ standard logistic distribution

$$CDF = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}$$

$$\theta_1 \leq \theta_2 \leq \dots \leq \theta_{j-1}$$

(i) Generate Z

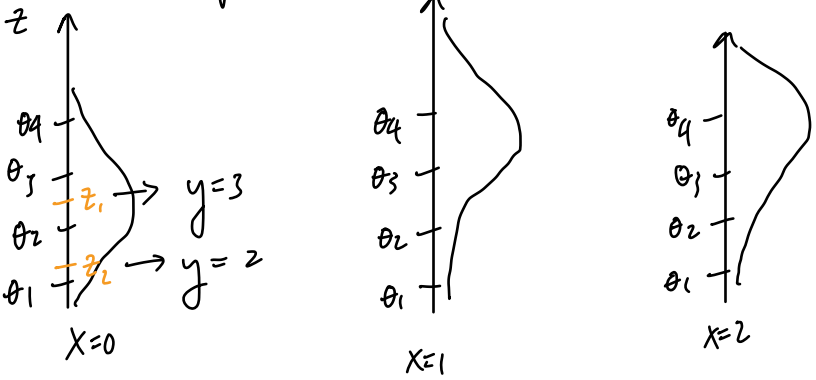
(ii) If $\theta_{j-1} \leq Z \leq \theta_j$, then $y = j$

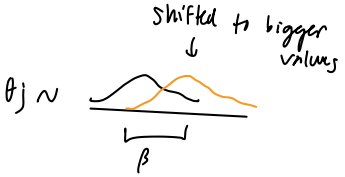
$$P(Y \leq j) = P(Z \leq \theta_j) = P(Z - \beta^T X \leq \theta_j - \beta^T X)$$

$$= \frac{e^{\theta_j - \beta^T X}}{1 + e^{\theta_j - \beta^T X}}$$

$\Rightarrow \text{logit } P(Y \le j) = \theta_j - \beta^T x$

sample z randomly





X increases $\Rightarrow$ probability of seeing data from higher categories increases

If the distribution becomes complementary log-log:

extreme $\Leftrightarrow$ comp. log-log

(i) generate $z \sim$ extreme dist

$$\log(-\log(1-\gamma_j)) = \theta_j - \beta^T x$$] proportional hazards model

normal $\Leftrightarrow$ probit

Feb 14

General Linear Models

- Consists of 3 core components:
  1. Random component
  2. Systematic component
     linear predictor $\eta = X\beta = \sum_{j=1}^{p} x_j \beta_j$
  3. Link — usually mean of y or something that describes the location of the response
     $g(\mu) = \eta$

Linear regression: $\epsilon \sim N(0, \sigma^2 I)$, $y = X\beta + \epsilon \sim N(X\beta, \sigma^2 I)$

No error term in logistic regression b/c you're looking at a parameter

- Exponential family (EF):

$$f(x) = h(x) \exp\{\eta(\theta) T(x) - A(\theta)\}$$

special case: Natural EF (NEF):

$$f(x) = h(x)\exp\{\theta x - A(\theta)\}$$

NEF $\subset$ EF

For EF, if $\eta(\theta) = \theta$, then this exponential family is in canonical form.

Ex: Normal distribution

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2-2x\mu+\mu^2}{2\sigma^2}}$$

(1) when σ^2 is known:

$$h(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}} \qquad \eta(\mu) = \frac{\mu}{\sigma}$$

$$T(x) = x/\sigma \qquad A(\mu) = \frac{\mu^2}{2\sigma^2}$$

suff. stat

(2) When σ^2 is unknown:

$$h(x) = \frac{1}{\sqrt{2\pi}} \qquad \eta = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$$

$$T(x) = (x, x^2)^T \qquad A(\eta) = \frac{\mu^2}{2\sigma^2} + \log|\sigma|$$

- density can always be written in canonical form
- canonical form is not unique.

EX: Beta distribution

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$= \exp\left\{(\alpha-1)\log x + (\beta-1)\log(1-x) + \log\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right\}$$

- EF but not NEF.

EX: log normal distribution

$$\frac{1}{\sqrt{2\pi}\sigma x} e^{\frac{-(\log x-\mu)^2}{2\sigma^2}}$$

- EF but not NEF

$$f_Y(y;\theta,\phi) = \exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi) \right\}$$

$a(\phi)$ $\quad$ $\underbrace{b(\theta)}$ $\quad$ $c(y,\phi)$

cumulant function

For now, ϕ is assumed to be known

θ is called canonical parameter.

log-likelihood:

$$l(\theta,\phi;y) = \log f_Y(y;\theta\phi)$$

$$= \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)$$

Note:

$$E\frac{\partial l}{\partial \theta} = 0, \quad E\frac{\partial^2 l}{\partial \theta^2} + E\left(\frac{\partial l}{\partial \theta}\right)^2 = 0$$

$$\Rightarrow E\frac{\partial l}{\partial \theta} = E\frac{y - b'(\theta)}{a(\phi)} = 0$$

$$\Rightarrow \underbrace{Ey}_{\mu} = b'(\theta)$$

$$\Rightarrow E\frac{\partial l^2}{\partial \theta^2} + E\left(\frac{\partial l}{\partial \theta}\right)^2 = E\frac{-b''(\theta)}{a(\phi)} + E\left(\frac{y - b'(\theta)}{a\phi}\right)^2$$

$$\frac{d\mu}{d\theta} \quad = \frac{-b''(\theta)}{a(\phi)} + \frac{\mathrm{var}(y)}{[a(\phi)]^2} = 0$$

$$\Rightarrow \text{var}(y) = a(\phi)\, b''(\theta)$$

$$= a(\phi)\, \underbrace{V(\mu)}_{\text{variance function.}}$$

ϕ : dispersion parameter

$a(\phi)$: a priori chosen function, commonly $a(\phi) = \frac{\phi}{w}$

w : prior weight

Examples:

① Normal

$$f_Y(y;\theta,\phi) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$$= \exp\left\{ \frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{1}{2}\left\{ \frac{y^2}{\sigma^2} + \log 2\pi\sigma^2 \right\} \right\}$$

$\theta = \mu \qquad b(\theta) = \theta^2/2$

$\phi = \sigma^2 \qquad a(\phi) = \phi$

$$c(y,\theta) = -\frac{1}{2}\left(\frac{y^2}{\phi^2} + \log 2\pi\sigma^2 \right)$$

$$EY = b'(\theta) = \theta = \mu$$

$$b''(\theta) = 1 = V(\mu)$$

$$\text{var}(Y) = a(\phi)V(\mu) = \sigma^2$$

② Poisson

$$f_Y(y;\theta,\phi) = \frac{\mu^y e^{-\mu}}{y!}$$

$$= \exp\{y\log\mu - \mu - \log y!\}$$

$\theta = \log\mu$ $\quad\quad$ $\phi = 1$

$b(\theta) = \mu = e^\theta$ $\quad\quad$ $a(\phi) = 1$

$b'(\theta) = e^\theta = \mu = E(Y)$

$b''(\theta) = e^\theta = \mu = V(\mu) \Rightarrow \text{var}(y) = a(\phi)\, b''(\theta)$

$= \mu$

$= E(Y)$

Link function

common links: $\quad\quad$ canonical link

① Identity e.g. Normal

② Binomial: $\text{logit} = \log\left(\frac{\pi}{1-\pi}\right) = \eta$

$\text{probit} = \eta = \Phi^{-1}(\pi)$

complementary log-log $= \eta = \log\{-\log(1-\pi)\}$

③ Poisson: $\log n = \log n$

- canonical link: related to the canonical parameter. transforms location param into canonical param.
- using canonical link means $X^T y$ is a sufficient statistic.

$$f_Y(y;\theta,\phi) = \frac{n^y e^{-n}}{y!}$$

$$= \exp\{y \log n - n - \log y!\}$$

$$\log n = X\beta$$

$y\theta = yX\beta$ if we use canonical link

$y h(X\beta)$ otherwise: $y(\theta) = y g^{-1}(X\beta)$

canonical links:

① Normal:

$$\eta = \mu$$

② Poisson:

$$\eta = \log \mu$$

③ Binomial:

$$\eta = \log \frac{\pi}{1-\pi}$$

④ Gamma:

$$\eta = \frac{1}{\mu}$$

⑤ Inverse Gaussian:

$$\eta = \frac{1}{\mu^2}$$

Algorithm for fitting GLM:

(1) iterated reweighted least squares perspective (IWLS)

Basic idea: Fit a regression using a linearized form of link function $g(y)$

$$g(y) = g(\mu) + g'(\mu)(y-\mu)$$
$$= g(\mu) + \frac{d\eta}{d\mu}(y-\mu)$$

linear approx. w/ Taylor expansion

$$z \triangleq X\beta + \frac{d\eta}{d\mu}(y-\mu) \quad \leftarrow \text{adjusted response}$$

$$= X\beta + \underbrace{\epsilon}_{\text{error term, need to find its variance}}$$

$$\text{Var}(\epsilon) = \left(\frac{d\eta}{d\mu}\right)^2 \text{Var}(y)$$

$$\propto \left(\frac{d\eta}{d\mu}\right)^2 V(\mu)$$

(assuming $a(\phi)$ is a known constant)

$$W^{-1} = \left(\frac{d\eta}{d\mu}\right)^2 V(\mu)\, a(\phi)$$ } needs to be estimated

Now can estimate β:

$$\beta = (X^T W X)^{-1} X^T W z$$

Empirically getting β_0 (logistic regression)

$$\log \frac{\pi_i}{1-\pi_i} = X\beta$$

$$(m_i, y_i) \rightarrow \hat{\pi}_i = \frac{y_i}{n_i}$$

$$\tilde{y}_i = \log \frac{\hat{\pi}_i}{1-\hat{\pi}_i}$$

Fit a line and use its slope as β_0 then iterate.

Feb 16

(2) Fisher Scoring

log-likelihood for a single obs:

$$\ell = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi) \qquad \eta = \sum \beta_j x_j$$

$$\Rightarrow \frac{\partial \ell}{\partial \beta_j} = \frac{\partial \ell}{\partial \theta} \cdot \frac{d\theta}{d\mu} \cdot \frac{d\mu}{d\eta} \frac{\partial \eta}{\partial \beta_j}$$

$$= \frac{y - b'(\theta)}{a(\phi)} \cdot \frac{1}{V} \cdot \frac{\partial \mu}{\partial \eta} \cdot x_j$$

$$\mu = b'(\theta) \quad \frac{d\mu}{d\theta} = b''(\theta) = V(\mu)$$

$$W^{-1} = \left(\frac{d\eta}{d\mu}\right)^2 V(\mu) \Rightarrow \frac{d\mu}{d\eta} = W \frac{d\eta}{d\mu} V$$

$$\Rightarrow \frac{\partial \ell}{\partial \beta_j} = \frac{y-\mu}{a(\phi)} \frac{1}{V} W \frac{d\eta}{d\mu} V x_j$$

$$= \frac{W(y-\mu)}{a(\phi)} \frac{d\eta}{d\mu} x_j$$

If $a(\phi) = \phi$:

$$\frac{\partial \ell}{\partial \beta_j} = \frac{W(y-\mu)}{\phi} \frac{d\eta}{d\mu} x_j$$

If $a(\phi) = \phi/w_i$: $y_1, \ldots, y_n$

$$\frac{\partial \ell}{\partial \beta_j} = \frac{w'(y-\mu)}{\phi} \frac{d\eta}{d\mu} x_j$$ where w_i's are absorbed into w'

$$= \frac{\partial \sum \ell_i}{\partial \beta_j} = \sum_i \frac{\partial \ell_i}{\partial \beta_j}$$

$$= \sum_i \frac{w_i (y_i - \mu_i) \frac{d\eta_i}{d\mu_i} x_{ij}}{\phi} = 0$$ } solve to get MLE

take deriv. first, y_i doesn't contain βs

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_s} = \frac{w_i \frac{d\eta_i}{d\mu_i} x_{ij} \frac{\partial (y_i - \mu_i)}{\partial \beta_s} + \frac{\partial w_i \frac{d\eta_i}{d\mu_i}}{\partial \beta_s} x_{ij}(y_i - \mu_i)}{\phi}$$

taking expectation:

$$E \frac{\partial^2 \ell}{\partial \beta_j \partial \beta_s} = \frac{w_i \frac{d\eta_i}{d\mu_i} x_{ij} \frac{-\partial \mu_i}{\partial \beta_s} + 0}{\phi}$$

because $E(y_i - \mu_i) = 0$

$$E \frac{\partial (y_i - \mu_i)}{\partial \beta_s} = -\frac{\partial \mu_i}{\partial \beta_s}$$

$$= \frac{-w_i}{\phi} \frac{d\eta_i}{d\mu_i} x_{ij} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_s} = \frac{-w_i}{\phi} x_{ij} x_{is}$$

second deriv. in matrix form:

$$\frac{(X^T W X)}{\phi}, \quad W = \begin{bmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_n \end{bmatrix}_{n \times n} \qquad X \in \mathbb{R}^{n \times p}$$

$$\Rightarrow \frac{(X^T W X)}{\phi} \in \mathbb{R}^{p \times p}$$

$$X^T W \left[(y - \mu) \frac{d\eta}{d\mu} \right] / \phi$$

Newton Raphson:

$$\beta_1 = \beta_0 + (X^T W X)^{-1} X^T W \left[(y - \mu) \frac{d\eta}{d\mu} \right]$$

ϕ's from these terms cancel out

$$= (X^T W X)^{-1} X^T W \left[X\beta_0 + (y - \mu) \frac{d\eta}{d\mu} \right]$$

Measuring the goodness of fit

1. $\ell(\hat{\mu}, \phi; y)$ - log-likelihood of the current model can be evaluated at the MLE. (maximized over β)

$\ell(y, \phi; \mu)$ - log likelihood of the fully saturated model w/ n parameters

- Derive canonical params from both models:

$$\hat{\theta} = \theta(\hat{\mu}) \quad \text{(current)}$$

$$\tilde{\theta} = \theta(y) \quad \text{(fully saturated)}$$

$$2\left[\ell(\vec{y}, \phi; \vec{y}) - \ell(\hat{\mu}, \phi; \vec{y})\right]$$

$$= \sum_i 2\omega_i \left\{ y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i) \right\} \Big/ \phi$$

$$= \frac{D(y, \hat{\mu})}{\phi} \quad \Big\} \text{ scaled deviance}$$

$$\text{Deviance} \triangleq D(y, \hat{\mu})$$

$$\text{Normal} \quad D = \sum_i (y_i - \hat{\mu}_i)^2 \quad \text{(exercise)}$$

$$\text{Poisson} \quad D = \sum_i \left\{ y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i) \right\}$$

- Generalized Pearson χ^2-statistic:

$$\chi^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

- Residuals:

(1) Pearson's residual:

$$r_p = \frac{y - \mu}{\sqrt{V(\mu)}}$$

(2) Deviance residual:

$$D = \sum_i d_i$$

$$r_D = \underbrace{\text{sign}(y - \mu)\sqrt{d_i}}_{\text{deviance residual}}$$

$\Rightarrow$ deviance is also a sum of squared residuals.

Log-linear model

- Poisson distribution:

$$P(Y=y) = \frac{\mu^y e^{-\mu}}{y!}$$

(a) cumulant generating function:

$$\mu(e^t - 1)$$

$$= \mu\left(t + \frac{t^2}{2!} + \frac{t^3}{3!} + \cdots\right)$$

All cumulants are equal to μ for Poisson

$$\frac{Y-\mu}{\sqrt{\mu}} \xrightarrow{\mu \uparrow} N(0,1)$$

N:	0	1	0	0
Y:	μ	μ	μ	μ
$\frac{Y-\mu}{\sqrt{\mu}}$:	0	1	$O(\frac{\mu}{\mu^{3/2}})$	$O(\frac{\mu}{\mu^{4/2}})$
			$O(\mu^{-1/2})$	$O(\mu^{-1})$

remaining terms converge to 0 ...

So $\frac{Y-\mu}{\sqrt{\mu}}$ cumulants converge to that of Standard Normal.

$\sqrt{Y}$ is a variance-stabilizing transform for Poisson and is often used over log.

$$g(Y) = g(\mu) + g'(\mu)(Y-\mu) + O(Y-\mu)$$

$$\text{var}(g(Y)) = [g'(\mu)]^2 \text{var } Y$$

$$= \left[\frac{1}{2}\mu^{-1/2}\right]^2 \mu$$

$$\text{var}(\sqrt{Y}) \approx \frac{1}{4}$$ } no longer depends on μ

Feb 21

log-linear model

assumption: $y_i \sim \text{Pois}(\mu_i)$

$$\log \mu_i = x_i^T \beta \Rightarrow \mu_i = \exp(\eta_i)$$

$$\Rightarrow \frac{\partial \mu_i}{\partial \eta_i} = e^{\eta_i} = \mu_i$$

log-likelihood:

$$l(\mu, y) = \sum_i (y_i \log \mu_i - \mu_i) + \text{const}$$

$$= \sum_i (y_i x_i^T \beta - e^{x_i^T \beta}) + \text{const}$$

$$\frac{\partial l}{\partial \beta_r} = \sum_i \frac{\partial l}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_r}$$

$$= \sum_i \left(\frac{y_i}{\mu_i} - 1\right) \mu_i x_{ir}$$

$$= \sum_i (y_i - \mu_i) x_{ir}$$

$$\frac{\partial^2 l}{\partial \beta_r \partial \beta_s} = \sum_i \frac{-\partial \mu_i}{\partial \beta_s} x_{ir} = \sum_i -\frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_r} x_{ir}$$

$$= -\sum_i \mu_i x_{is} x_{ir}$$

$$E\left[\frac{\partial^2 l}{\partial \beta_r \partial \beta_s}\right] = \sum_i \mu_i x_{is} x_{ir}$$

$\Downarrow$ $X^T W X$, $W = \text{diag}(w_i)$

$$\frac{\partial \ell}{\partial \beta} = X^T (y - \mu)$$

$$-E\left[\frac{\partial^2 \ell}{\partial \beta_r \partial \beta_s}\right] = X^T W X$$

Fisher scoring:

$$\beta_0 \quad \beta_1 = \beta_0 + (X^T W X)^{-1} X^T (y - \mu)$$

Deviance:

$$D(y, \mu) = 2[\ell(y, y) - \ell(\hat{\mu}, y)]$$

$$= 2\sum \left\{ y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i) \right\}$$

Deviance residuals:

$$\text{sign}(y_i - \hat{\mu}_i) \sqrt{2\left\{ y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i) \right\}}$$

Overdispersion:

ex: $Y = \underbrace{Z_1 + Z_2 + \cdots + Z_N}$

indep. Poisson RVs

$Y \sim$ Pois if N fixed

if N is also an RV indep. of Z,

Assume $EZ = \mu$ for all Z and $\text{var}(Z) = \mu$

$$EY = E[E(Y|N)] = E[NEZ]$$
$$= ENEZ = \mu EN$$

$$\text{var}\, Y = E[\text{Var}(Y|N)] + \text{var}[E(Y|N)]$$
$$= E[N\text{var}\, Z] + \text{var}[NEZ]$$
$$= (EN)\text{var}\, Z + \text{var}(N)(EZ)^2$$
$$= (EN)\mu + \text{var}(N)\mu^2$$

If $N \sim$ Pois:

$$\text{var}\, Y = (EN)(\mu + \mu^2)$$

But $\text{var}\, Y > EY$ b/c N is random, additional layer of variability leads to overdispersion

ex: $Y_i|Z_i \sim \text{Pois}(Z_i)$, $Z_i \sim \text{Gamma}(\alpha, \beta)$

$$P(Y_i) = \int P(Y_i, Z_i)\, dZ_i$$

$P(Y_i)$ is not Poisson

should be negative binomial. overdispersion.

$\text{var}(y) = \sigma^2 \mu$, $Ey = \mu$

$\hat{\beta} =$

$D \sim \sigma^2 \chi^2$

$\sigma^2 w_i$

$$\frac{1}{n-p} \sum_i \frac{(y_i - \mu_i)^2}{w_i} = \hat{\sigma}^2$$ estimate of dispersion param

Log linear and multinomial response models

$y_i \sim \text{pois}(\mu_i)$, $\log(\mu_i) = \alpha + x_i^T \beta = \eta_i$

$f(y_1, \dots, y_n) = f(y_+) f(y_1, \dots, y_n | y_+)$

$y_+ = \sum_i y_i \sim \text{pois}(\mu_+)$, $\mu_+ = \sum_i \mu_i$

$$f(y|y_+) = \frac{f(y)}{f(y_+)} = \prod_{i=1}^n \frac{\frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}}{\frac{\mu_+^{y_+} e^{-\mu_+}}{y_+!}}$$

$$= \frac{y_+!}{\prod y_i!} \prod \left(\frac{\mu_i}{\mu_+}\right)^{y_i} \frac{e^{-\sum \mu_i}}{e^{-\mu_+}}$$

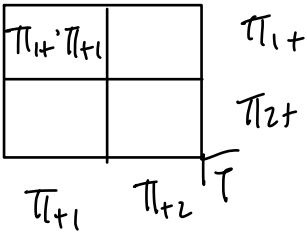
$$\frac{\mu_i}{\mu_+} = \frac{e^{\alpha + x_i^T\beta}}{\sum_i e^{\alpha + x_i^T\beta}} = \frac{e^{x_i^T\beta}}{\sum_i e^{x_i^T\beta}}$$

$$\log f(y|\alpha,\beta) = \underbrace{\log(f(y_+))}_{\text{Poisson}\ \alpha,\beta} + \underbrace{\log(f(y|y_+))}_{\text{multinomial}\ \beta}$$

plug in estimated β from multinomial estimate to get α estimate

first maximize this

y_+ is a sufficient statistic for α



assume T is fixed

Poisson way:

assume each cell's counts follows Poisson

μ_{11}	μ_{12}
μ_{21}	μ_{22}

ex: Independence

multinomial: $\pi_{ij} = \pi_{i+}\pi_{+j}$

log linear: $\mu_{ij} = \exp\{\lambda + \lambda_i^X + \lambda_j^Y\}$

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y \quad (*)$$

$$\pi_{ij} = \frac{\mu_{ij}}{\mu_{++}} = \frac{e^{\lambda + \lambda_i^X + \lambda_j^Y}}{\sum_i \sum_j e^{\lambda + \lambda_i^X + \lambda_j^Y}}$$

$$= \frac{e^{\lambda_i^X} e^{\lambda_j^Y}}{\sum_i e^{\lambda_i^X} \sum_j e^{\lambda_j^Y}}$$

π_{i+} ← $\sum_i e^{\lambda_i^X}$ | $\sum_j e^{\lambda_j^Y}$ → π_{+j}

row factors effects

column factors effects

$$= \pi_{i+}\pi_{+j}$$

we usually constrain the params: $\sum_i \lambda_i^X = 1$

Deviance test tests for independence.

For an $I \times I$ table:

For identifiability, assume

$$\sum_i \lambda_i^X = 0 \qquad \sum_j \lambda_j^Y = 0$$

or

$$\lambda_I^X = 0 \qquad \lambda_I^Y = 0$$

params:

$$\lambda \quad \lambda_i^* \quad \lambda_j^Y$$
$$1 + I-1 + I-1 = 2I-1$$

I^2 data points

df: $I^2 - (2I-1) = (I-1)^2$

ex:

Residence in 1980 \ Residence in 1985	NE	NW	S	W
NE	11607	100	306	124
NW	87	13677	515	302
S	172	255	17819	270
W	63	176	286	10192

under independence:

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y$$

to test for independence:

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \delta_i I(i=j)$$

Quasi-Independence model

$I \subset QI$

independence model

Feb 23

$y_i \sim \text{Pois}_+(\mu_i)$

$\log \mu_i = x_i^T \beta \Rightarrow \mu_i = e^{\alpha + x_i^T \beta}$, $\mu_+ = \sum e^{\alpha + x_i^T \beta}$

$$P(y_1, \dots, y_n) = \prod_{i=1}^{n} \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}$$

$\mu_+ = \sum_i \mu_i$

$$= \frac{y_+ (\mu_+)^{y_+}}{y_+ (\mu_+)^{y_+}} \prod_{i=1}^{n} \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}$$

$$= \underbrace{\frac{(\mu_+)^{y_+} e^{-\mu_+}}{y_+!}}_{\text{Poisson}} \underbrace{\frac{y_+!}{\prod_i y_i!} \prod_i \left(\frac{\mu_i}{\mu_+}\right)^{y_i}}_{\text{multinomial}}$$

$$\ell(\alpha, \beta, y) = y_+ \log \mu_+ - \mu_+ + \ell(\beta | y)$$

$$\frac{\mu_i}{\mu_+} = \frac{e^{x_i^T \beta}}{\sum e^{x_i^T \beta}}$$

reparameterize the log-likelihood

$\ell(\mu_+, \beta; y)$

$\hat{\mu}_+ = y_+$

Independence model:

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y$$

Quasi Independence model:

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \delta_i I(i=j)$$

Symmetry

$$\pi_{ij} = \pi_{ji} \quad \forall\, i,j$$

$$\log \mu_{ij} = \lambda + \lambda_i + \lambda_j + \lambda_{ij}$$

where $\lambda_{ij} = \lambda_{ji}$

$$\log \mu_{ji} = \lambda + \lambda_j + \lambda_i + \lambda_{ji}$$

Don't have X, Y, rows and columns share the same type of effects

symmetry $\Rightarrow$ marginal homogeneity (MH)

↳ marginal row sums = marginal col sums

$$\pi_{+j} = \sum_i \pi_{ij} = \sum_i \pi_{ji} = \pi_{j+}$$

MH $\not\Rightarrow$ symmetry

2X2: MH $\Rightarrow$ S

π_{11}	π_{12}
π_{21}	π_{22}

$\pi_{11} + \pi_{12} = \pi_{11} + \pi_{21}$

$\Rightarrow \pi_{12} = \pi_{21}$

3X3:

x	y	z
z	x	y
y	z	x

MH but not S

Free params:

$$\log m_{ij} = \underbrace{\lambda}_{1} + \underbrace{\lambda_i}_{I-1} + \lambda_j + \underbrace{\lambda_{ij}}_{\frac{I(I-1)}{2}}$$

Only need to consider off diagonals in one half b/c of symmetry

$$df = I^2 - 1 - (I-1) - \frac{I(I-1)}{2}$$

$$= \frac{I(I-1)}{2}$$

Quasi-symmetry: $S \subset QS$

$$\log m_{ij} = \underbrace{\lambda}_{1} + \underbrace{\lambda_i^X}_{I-1} + \underbrace{\lambda_j^Y}_{I-1} + \underbrace{\lambda_{ij}}_{\frac{I(I-1)}{2}}$$

λ: models overall mean

where $\lambda_{ij} = \lambda_{ji}$

params: $1 + I - 1 + I - 1 + \frac{I(I-1)}{2}$

$$df = I^2 - (\#\text{params})$$

$$= \frac{(I-1)(I-2)}{2}$$

$QS + MH \Rightarrow S$:

Consider multinomial probabilities implied by QS.

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}$$

$$= \log m_{++} + \log \frac{m_{ij}}{m_{++}}$$

$$= \log m_{++} + \log \pi_{ij}$$

$$\pi_{ij} \propto e^{\lambda_i^x} e^{\lambda_j^y} e^{\lambda_{ij}}$$

$$\pi_{ab} \propto \alpha_a \beta_b \gamma_{ab}$$

α_a: a^{th} row; β_b: b^{th} column; γ_{ab}: interaction effects

$$= \frac{\alpha_a}{\beta_a} \beta_a \beta_b \gamma_{ab}$$

$$= \ell_a \delta_{ab} \quad \delta_{ab} = \delta_{ba} \text{ (symmetric quantity)}$$

$$\pi_{ab} = \ell_a \delta_{ab}$$

$$\pi_{ab} \stackrel{?}{=} \pi_{ba} \quad \text{we know: } \pi_{j+} = \pi_{+j} \text{ (MH)}$$

$$\pi_{j+} = \sum_i \pi_{ji} = \sum_i \ell_j \delta_{ji} = \ell_j \sum_i \delta_{ji} = \ell_j \sum_i \delta_{ij}$$

$$\pi_{+j} = \sum_i \pi_{ij} = \ell_i \sum_i \delta_{ij} = \sum_i \ell_i \delta_{ij}$$

$$\ell_j = \sum_i \ell_i \frac{\partial_{ij}}{\sum_{i'} \partial_{i'j}}$$

$$= \sum_i \ell_i w_i$$

$$w_i \triangleq \frac{\partial_{ij}}{\sum_{i'} \partial_{i'j}}$$

$\ell_1 \leq \ell_2 \leq \cdots \leq \ell_I$ and $\ell_1 < \ell_I$ if they are not all the same.

$$\ell_1 = \sum_i \ell_i w_i > \ell_1 \sum_i^{I-1} w_i + \ell_I w_I$$

$$> \ell_1 \sum_{i=1}^{I} w_i$$

$$= \ell_1$$

which is a contradiction. So

$$\ell_1 = \ell_2 = \cdots = \ell_I$$

$$\Rightarrow \pi_{ab} = \ell_a \delta_{ab} = \ell_b \delta_{ba} = \pi_{ba}$$

Quasi-likelihood

$$E(y) \sim X\beta$$

$$\text{var}(y) \sim V(\mu)$$

(1) Suppose response variable y has

mean: μ

variance: $\sigma^2 V(\mu)$

$$U = \frac{y - \mu^{(\beta)}}{\sigma^2 V(\mu)} = 0$$

behaves like deriv. of log-likelihood. Solve for estimate of β.

$$E_y\, l'(\theta; y) = 0$$

$$-E\, l''(\theta; y) = \text{var}(l'(\theta; y))$$

$$= E[l'(\theta; y)]^2$$

$$\cong \frac{1}{\sigma^2 V(\mu)}$$

asymptotically

$$E(U) = 0 \quad \text{(only } y \text{ is random)}$$

$$\text{Var}(U) = \frac{\sigma^2 V(\mu)}{(\sigma^2 V(\mu))^2} = \frac{1}{\sigma^2 V(\mu)}$$

$$-E\frac{\partial U}{\partial \mu} = E\left[\frac{-1}{\sigma^2 V(\mu)} + (y-\mu)\left(\quad\right)\right]$$

$$= \frac{1}{\sigma^2 V(\mu)}$$

$$\uparrow$$

$$\text{var}(y)^{-1}$$

So U behaves like deriv. of log-likelihood.

March 2

Quasi-Likelihood

μ , $V(\mu)$

$$\frac{Y-\mu}{\sigma^2 V(\mu)}$$

Ex 2:

$V(\mu) = \mu$, $\sigma^2 = 1$

$$Q(\mu, y) = \int_y^{\mu} \frac{y-t}{t}\, dt$$

$$= (y \log t - t) \Big|_y^{\mu}$$

$$= (y \log \mu - \mu) - (y \log y - y)$$

Maximising the quasi-likelihood:

- Differentiate $Q(\mu, y)$ wrt β

$$U(\beta)_{p\times 1} = \frac{\partial \mu^T}{\partial \beta} \frac{V^{-1}(y-\mu)}{\sigma^2}$$

$$= D^T_{p\times n} \frac{V^{-1}(y-\mu)}{\sigma^2} \overset{\text{set}}{=} 0$$

where $D = (D_{ir})_{n\times p} \triangleq \left(\frac{\partial \mu_i}{\partial \beta_r}\right)_{n\times p}$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \sim \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} \begin{bmatrix} x_1 & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$\begin{bmatrix} V(\mu_1) & & 0 \\ & V(\mu_2) & \\ 0 & & \ddots \\ & & V(\mu_n) \end{bmatrix}_{n \times n} \begin{bmatrix} y_1 - \mu_1 \\ \vdots \\ y_n - \mu_n \end{bmatrix}$$

$$-E \frac{\partial U(\beta)}{\partial \beta}$$

0 in expectation

$$= -E \left[-\frac{D^T V^{-1} D}{\sigma^2} + \frac{\partial D^T V^{-1}}{\partial \beta} \frac{(y-\mu)}{\sigma^2} \right]$$

$$= \frac{D^T V^{-1} D}{\sigma^2} \triangleq i_\beta$$

2nd deriv of Quasi likelihood

Newton - Raphson:

dispersion param σ^2 gets cancelled

$$\beta_1 = \beta_0 + \left(\frac{D_0^T V_0^{-1} D_0}{\sigma^2} \right)^{-1} \frac{D_0^T V_0^{-1} (y - \mu_0)}{\sigma^2}$$

$$= \beta_0 + (D_0^T V_0^{-1} D_0)^{-1} D_0^T V_0^{-1} (y - \mu_0)$$

$$\text{var}(\hat{\beta}) \simeq \sigma^2 (D^T V^{-1} D)^{-1}$$

$$\tilde{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V(\mu_i)} = \frac{X^2}{n-p}$$

Optimal estimating functions

Quasi-score $U(\beta) = D^T V^{-1}(Y-\mu)/\sigma^2$ is a special case of estimating function.

(1) Estimating function $g(y;\theta)$: a function of data y and parameter θ with zero mean for all parameter values $E_y g(y;\theta) = 0 \quad \forall \theta$

$H^T_{p\times n}(y-\mu)_{n\times 1}$, $H(\beta)_{n\times p}$

$p\times 1$. Get p estimating functions, set to 0, get p estimating eqns.

$U(\beta) = D^T V^{-1}(y-\mu)/\sigma^2$ is the optimal combination of the elementary est function within the class of linear estimating functions.

$(y-\mu) \to h(y;\beta) \triangleq H^T(y-\mu) = 0 \Rightarrow \tilde{\beta} \Rightarrow \hat{\beta}$

Asymptotically, all linear functions of $\hat{\beta}$ have variances at least as big as the variance of $\hat{\beta}$.

i.e. $\mathrm{var}(\tilde{\beta}) \geq \mathrm{var}(\hat{\beta})$.

$$h(y;\tilde{\beta}) = h(y;\beta) + h'(y;\beta)(\tilde{\beta}-\beta) + o(\tilde{\beta}-\beta)$$

↑ can replace w/ its expectation

$$\text{then } \tilde{\beta}-\beta \simeq -\left[E\,h'(y;\beta)\right]^{-1} h(y;\beta)$$

$$= \left[-E\left(-H^T D + \frac{\partial H}{\partial \beta}(y-m)\right)\right]^{-1} h(y;\beta)$$

$$= (H^T D)^{-1} h(y;\beta)$$

$$= (H^T D)^{-1} H^T (y-m)$$

$$\text{var}(\tilde{\beta}) \simeq \text{var}\left((H^T D)^{-1} H^T (y-m)\right)$$

$$= (H^T D)^{-1} H^T V H (D^T H)^{-1} \sigma^2$$

$$\text{var}(\hat{\beta}) = \sigma^2 (D^T V^{-1} D)^{-1}$$

$$\text{var}(\tilde{\beta}) - \text{var}(\hat{\beta}) \geq 0$$

non-negative definite matrix

Need to show $(H^T D)^{-1} H^T V H (D^T H)^{-1} - (D^T V^{-1} D)^{-1} \geq 0$

$\Updownarrow$ b/c $A - B \geq 0 \iff B^{-1} - A^{-1} \geq 0$

$D^T V^{-1} D - (D^T H)(H^T V H)^{-1} H^T D \geq 0$

$\Updownarrow$

$D^T [V^{-1} - H(H^T V H)^{-1} H^T] D \geq 0$

$\Updownarrow$

$V^{-1} - H(H^T V H)^{-1} H^T \geq 0$

$V^{1/2} [V^{-1} - H(H^T V H)^{-1} H^T] V^{1/2} \geq 0$

$\Updownarrow$

$I - V^{1/2} H (H^T V H)^{-1} H^T V^{1/2} \geq 0$

$\tilde{X} \triangleq V^{1/2} H$

$I - \tilde{X}(\tilde{X}^T \tilde{X})^{-1} \tilde{X} \geq 0$

March 7

Optimal estimating eqns

$$D' V^{-1} (Y - \mu)$$

$$\begin{bmatrix} 0 & & \\ & \ddots & \\ & & 0 \end{bmatrix}^{-1} \begin{matrix} y_1 - \mu_1 \\ y_2 - \mu_2 \\ \vdots \\ y_n - \mu_n \end{matrix}$$

$$E\{ g_i(Y, \theta) \mid A_i \} = 0$$

Time ordered sequence:

$$Y_1, Y_2, \ldots, Y_n$$

$$Y_{(t)} = (Y_1, \ldots, Y_t)$$

$$E\{ g_t(Y_t, \theta) \mid Y_{t-1} \} = 0$$

$$V \triangleq \text{diag}\{ \text{var}(g_i \mid A_i) \}$$

$$D_{ir} \triangleq -E\left\{ \frac{\partial g_i(Y, \theta)}{\partial \theta_r} \mid A_i \right\}$$

$$U(\theta, y) = D' V^{-1} g$$

Ex: Y's are generated by an autoregressive process.

$$Y_t = \theta Y_{t-1} + \varepsilon_t$$

$$Y_0 = \varepsilon_0$$

where $\varepsilon_t \overset{iid}{\sim} N(0,1)$

Want to estimate θ:

Need a function that is 0 in conditional expectation.

$$E\{ y_t - \theta y_{t-1} \mid y_{t-1} \} = 0$$

$$g_t = y_t - \theta y_{t-1}$$

$$g_t^* = \frac{y_t}{\theta} - y_{t-1} \quad (\text{if } \theta \neq 0)$$

$$g^* = Bg, \quad U(\theta, y) = D' V^{-1} g$$

$$\text{Var}(g^*) = B \,\text{Var}(g)\, B' = V^*$$

$$= BVB'$$

$$\frac{\partial g^*}{\partial \theta} = B \frac{\partial g}{\partial \theta} = BD = D^*$$

$$\hookrightarrow D^{*\prime} V^{*-1} g^*$$

$$= (BD)' (BVB')^{-1} Bg$$

$$= D' B' (B')^{-1} V^{-1} B^{-1} Bg$$

$$= D' V^{-1} g$$

$$\text{Dir} = -E\left(\frac{\partial g_i}{\partial \theta} \mid A_i \right)$$

$$D = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \end{bmatrix}$$

$$\underbrace{\begin{bmatrix} y_0 & y_1 & \cdots \end{bmatrix}}_{D^\tau} \underbrace{\begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix}^{-1}}_{V^{-1}} \underbrace{\begin{bmatrix} y_1 - \theta y_0 \\ y_2 - \theta y_1 \\ \vdots \\ y_n - \theta y_{n-1} \end{bmatrix}}_{g} = \sum_t y_{t-1} g_t$$

$$= \sum_t y_{t-1} (y_t - \theta y_{t-1})$$

$$= 0$$

$$\hat{\theta} = \frac{\sum_t y_{t-1} y_t}{\sum_t y_{t-1}^2}$$

g^*: $U(\theta, y) = \sum_t E\left(\frac{y_t}{\theta^2} \mid y_{t-1}\right) \theta^2 \left(\frac{y_t}{\theta} - y_{t-1}\right)$

$$= \sum_t \frac{\theta y_{t-1}}{\theta^2} \theta^2 \left(\frac{y_t}{\theta} - y_{t-1}\right)$$

$$= \sum_t y_{t-1}\left(y_t - \theta y_{t-1}\right)$$

MLE:

$$\prod_t \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(y_t - \theta y_{t-1})^2}{2}\right\}$$

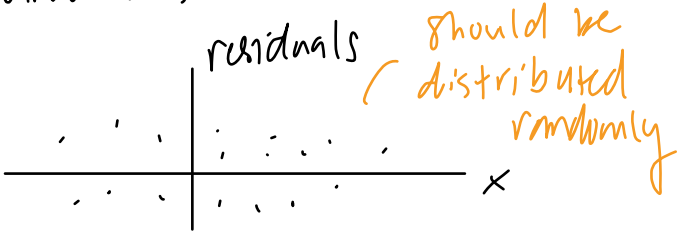
$$l = \sum_t \frac{-(y_t - \theta y_{t-1})^2}{2} + c$$

$$\frac{\partial l}{\partial \theta} = \sum_t \frac{2(y_t - \theta y_{t-1}) y_{t-1}}{2} = 0$$

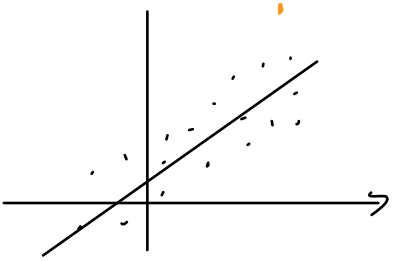
$\Rightarrow$ optimal estimating eqns and MLE return the same estimate.

Model Diagnosis

1) check systematic discrepancies between model and data



2) check isolated discrepancies



{ Formal: tests
{ Informal: plots

asymptotically equivalent

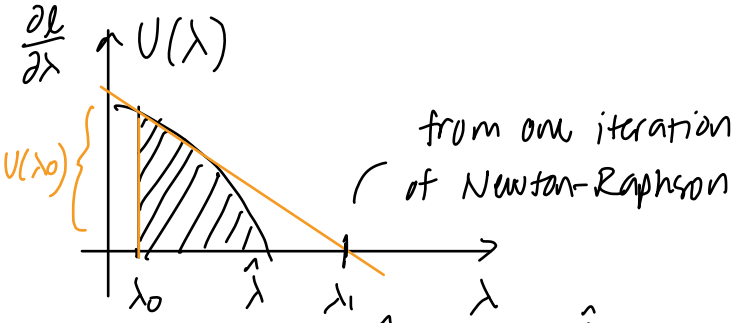
(1) Drop-in-deviance (likelihood ratio) test
- M_0: p params
- M_1: p+k params

$D = 2(\log L_1 - \log L_0)$

$= 2(l_1 - l_0)$

(2) score test (inv. of Fisher eval. at MLE)

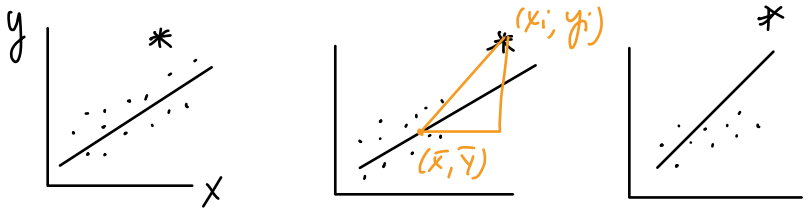
$S(\lambda_0) = U(\lambda_0)^T i^{-1}(\lambda_0) U(\lambda_0)$



$$\ell(\hat{\lambda}) - \ell(\lambda_0) = \int_{\lambda_0}^{\hat{\lambda}} \frac{\partial \ell}{\partial \lambda} \, d\lambda = \int_{\lambda_0}^{\hat{\lambda}} U(\lambda) \, d\lambda$$

$$\frac{1}{2} S(\lambda_0) = \frac{1}{2} U(\lambda_0)(\lambda_1 - \lambda_0)$$

Isolated discrepancies:



$$y - \bar{y} = \gamma \frac{S_y}{S_x}(x - \bar{x})$$

$$= \frac{S_{xy}}{S_x^2}(x - \bar{x})$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} (x_i - \bar{x})$$

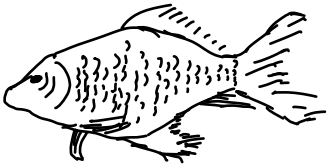
$$\beta = \frac{\sum (x_i - \bar{x})^2 \left(\frac{y_i - \bar{y}}{x_i - \bar{x}} \right)}{\sum_i (x_i - \bar{x})^2}$$

$$= \sum w_i \left(\frac{y_i - \bar{y}}{x_i - \bar{x}} \right)$$

linear combination of slope contributed by each datapoint

if x is very different from $\bar{x}$, it will have a bigger weight.

cabbage



$$\hat{y} = X\hat{\beta}$$

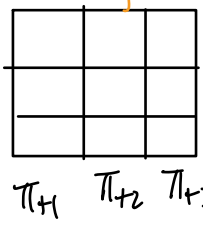
$$H = X(X'X)^{-1}X'$$

$$\hat{y} = Hy$$

$$H = \begin{bmatrix} \ddots & & \\ & h_{ii} & \\ & & \ddots \end{bmatrix}$$

March 7 OH

π_{ij} ← alternative model



I categories

assume total counts are fixed

$I-1$ independent

$I-1 + I-1$ (null model) } multinomial

I^2-1 (alternative model)

difference:

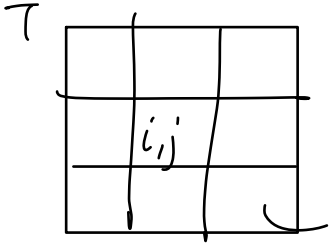
$$I^2-1-2(I-1)$$

$$= I^2-2I+1$$

$$= (I-1)^2$$

χ^2 will have this many df

alternative model:



pick this to be baseline

introduce an indicator — indicators

$$\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{II} \end{bmatrix} = \begin{bmatrix} 1 & \partial_{11} & \partial_{12} & \cdots & \partial_{II} \\ & & & & \\ 1 & \partial_{I1} & \cdots & & \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_{11} \\ \beta_{12} \\ \vdots \\ \beta_{II} \end{bmatrix} \Big\} I^2 - 1$$

exercise: converting models to matrix form

Poisson log-linear is the same:

$$\log m_{ij} = \lambda_0 + \lambda_i^x + \lambda_j^y$$

1 | $I-1$ | $I-1$ indicator of obs. in row i or not | column indicators

$$\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{II} \end{bmatrix} \rightarrow \begin{bmatrix} \log m_{11} \\ \log m_{12} \\ \vdots \\ \log m_{II} \end{bmatrix} = \left[\begin{array}{ccc|cc} 1 & r_{11} & r_{12} & c_{11} & c_{12} \\ & r_{21} & r_{22} & & \\ \vdots & & & & \\ 1 & & & & \end{array}\right] \begin{bmatrix} \lambda_0 \\ \lambda_1^x \\ \lambda_2^x \\ \vdots \\ \lambda_I^x \\ \lambda_1^y \\ \vdots \\ \lambda_I^y \end{bmatrix}$$

$2I - 1$ params

↳ same as independent multinomial model.

alternative model

if it's the alternative (saturated) model in order to capture cell-specific effects but only added to one cell b/c 2x2 table can only have 4 params

---

$$z_{ij} = \underbrace{y_i}_{\sim N(\mu, \sigma^2)} + \varepsilon_{ij} \sim N(0, \tau^2)$$

$$\text{cov}(z_{ij}, z_{i'j'})$$

$$= \text{cov}(y_i + \varepsilon_{ij},\ y_{i'} + \varepsilon_{i'j'}) \quad i \neq i'$$

$$= 0$$

$$\text{cov}(z_{ij}, z_{ik}) \quad j \neq k$$

$$= \text{cov}(y_i + \varepsilon_{ij},\ y_i + \varepsilon_{ik})$$

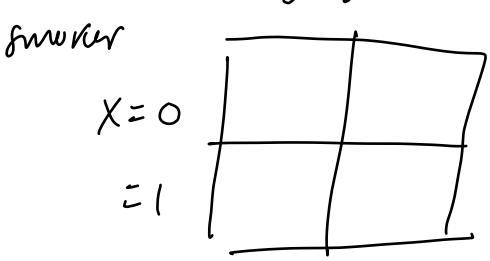
$$= \underbrace{\text{cov}(y_i, y_i)}_{\text{marginal covariance}} = \text{var}(y_i) = \sigma^2$$

$\Downarrow$ non-zero correlation

$$z_{ij} \mid y_i \overset{iid}{\sim} N(y_i, \tau^2),\ y_i \overset{iid}{\sim} N(\mu, \sigma^2)$$

$$\underbrace{Cov(z_{ij}, z_{ik} \mid y_i)}_{0} \neq Cov(z_{ij}, z_{ik})$$

$$\text{logit}\, P(y=1 \mid X) = \beta_0 + \beta_1 X$$



prospective:

don't a priori control cancer status

Conditional on being a smoker, what's the prob. of getting cancer? → same as population prob.

March 9

$$r = y - \hat{y} = y - Hy$$

$$= (I - H)y$$

$$\text{var}\, r = (I - H)\sigma^2 I (I - H)$$

$$= (I - H)\sigma^2 I$$

$$\text{var} \begin{bmatrix} y_1 - \hat{y}_1 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix} \rightarrow \underbrace{\begin{bmatrix} \circ & & \\ & \ddots & \\ & & \circ \end{bmatrix}}_{I - H} (1 - h_{ii})\sigma^2$$

Raw materials for model checking

(1) Linear models

$\hat{u}$: fitted value

$r = y - \hat{u}$: residuals (raw)

S^2 : residual variance

$H = X(X'X)^{-1}X'$: hat matrix

$h_i \triangleq h_{ii}$: leverage $= x_i'(X'X)^{-1}x_i$

$[\cdots] \quad \begin{bmatrix} \vdots \end{bmatrix}$

other types of residuals:

standardized:

$$\frac{y_i - \hat{\mu}_i}{\sqrt{1-h_i}}$$

studentized standardized:

$$r_i' = \frac{y_i - \hat{\mu}_i}{s\sqrt{1-h_i}}$$

fitted w/ problematic point

deleted:

$$r_i^* = \frac{y_i - \hat{\mu}_{(i)}}{s_{(i)}\sqrt{1+h_{(i)}}}$$

i^{th} data point removed

$$\hat{\mu}_{(i)} = x_i' \hat{\beta}_{(i)}$$

obtained w/o i^{th} data point

$$h_{(i)} = x_i' \left(X_{(i)}' X_{(i)}\right)^{-1} x_i$$

- y_i and $\hat{\mu}_{(i)}$ are independent, but y_i and $\hat{\mu}_i$ are not.

that's why variance in deleted residual is additive.

$$r_i^* = \frac{y_i - \hat{\mu}_{(i)}}{s_{(i)}\sqrt{1+h_{(i)}}} = \frac{y_i - \hat{\mu}_i}{s_{(i)}\sqrt{1-h_i}}$$

Pf:/ WLOG, consider obs. 1

$$\text{var}\left(y_1 - \hat{\mu}_{(1)}\right) = \text{var}\left(y_1 - x_1' \left(X_{(1)}' X_{(1)}\right)^{-1} X_{(1)}' y_{(1)}\right)$$

$\text{var}(y_1 - \hat{\mu}_{(1)})$ $\quad \text{var}(Ay) = A\,\text{var}(y)\,A'$

$$= \text{var}\left(\begin{bmatrix} 1 & -x_1'(X_{(1)}'X_{(1)})^{-1}X_{(1)}' \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \right)$$

$$= \begin{bmatrix} 1 & -x_1'(X_{(1)}'X_{(1)})^{-1}X_{(1)}' \end{bmatrix} \sigma^2 I \begin{bmatrix} 1 \\ -X_{(1)}(X_{(1)}'X_{(1)})^{-1}x_1 \end{bmatrix}$$

$$= \left[1 + x_1'(X_{(1)}'X_{(1)})^{-1}X_{(1)}'X_{(1)}(X_{(1)}'X_{(1)})^{-1}x_1 \right]\sigma^2$$

$$= \left[1 + x_1'(X_{(1)}'X_{(1)})^{-1}x_1 \right]\sigma^2$$

$$= \left[1 + h_{(1)} \right]\sigma^2$$

so $r_i^* = \dfrac{y_i - \hat{\mu}_{(i)}}{s_{(i)}\sqrt{1+h_{(i)}}}$

Next, show $h_{(i)} = \dfrac{h_i}{1-h_i}$. WLOG, consider h_1:

$$X = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix} = \begin{pmatrix} x_1' \\ X_{(1)} \end{pmatrix}$$

$$H = X(X'X)^{-1}X' = \begin{pmatrix} x_1' \\ X_{(1)} \end{pmatrix} \left[(x_1 \; X_{(1)}') \begin{pmatrix} x_1' \\ X_{(1)} \end{pmatrix} \right]^{-1} (x_1 \; X_{(1)}')$$

$$= \begin{pmatrix} X_1' \\ X_{(1)} \end{pmatrix} \left[X_1 X_1' + X_{(1)}' X_{(1)} \right]^{-1} \left(X_1 \; X_{(1)}' \right)$$

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}$$

$A : X_{(1)}' X_{(1)}$, $B : X_1$, $C : X_1'$, $D = I$

$$\left[X_1 X_1' + X_{(1)}' X_{(1)} \right]^{-1}$$

$$= \left[X_{(1)}' X_{(1)} \right]^{-1} + \left[X_{(1)}' X_{(1)} \right]^{-1} X_1 \left(-1 - X_1' (X_{(1)}' X_{(1)})^{-1} X_1 \right)^{-1} X_1' (X_{(1)}' X_{(1)})^{-1}$$

$$= \left[X_{(1)}' X_{(1)} \right]^{-1} - \left[X_{(1)}' X_{(1)} \right]^{-1} X_1 (1 + h_{(1)})^{-1} X_1' (X_{(1)}' X_{(1)})^{-1}$$

$$h_1 = X_1' \left[X_1 X_1' + X_{(1)}' X_{(1)} \right]^{-1} X_1$$

$$= X_1' \left[X_{(1)}' X_{(1)} \right]^{-1} X_1 - \frac{X_1' \left[X_{(1)}' X_{(1)} \right]^{-1} X_1 \cdot X_1' \left[X_{(1)}' X_{(1)} \right]^{-1} X_1}{1 + h_{(1)}}$$

$$h_1 = h_{(1)} - \frac{h_{(1)}^2}{1 + h_{(1)}} = \frac{h_{(1)}}{1 + h_{(1)}}$$

$$h_{(1)} = \frac{h_1}{1 - h_1}$$

$$y_1 - \hat{m}_1 = y_1 - x_1'(X'X)^{-1}X'y$$

$$= y_1 - x_1'\left[(x_1\ x_{(1)}')\begin{pmatrix} x_1 \\ x_{(1)}' \end{pmatrix}\right]^{-1}(x_1\ x_{(1)}')\begin{pmatrix} y_1 \\ y_{(1)} \end{pmatrix}$$

$$= (1-h_1)(y_1 - \hat{m}_{(1)})$$

$$\Rightarrow h_{(1)} = \frac{h_1}{1-h_1}$$

show these steps at home.

$$\frac{y_1 - \hat{m}_1}{\sqrt{1-h_1}} = \frac{y_1 - \hat{m}_{(1)}}{\sqrt{1+h_{(1)}}}$$

$$r_i^* = r_i' \frac{S_i}{S_{(i)}}$$

GLMs:

$y \to z$ (adjusted response)

$\hat{m} \to \hat{\eta}$

$S^2 \to \hat{\phi}$

$H \to W^{1/2}X(X'WX)X'W^{1/2}$

$W = V^{-1}\left(\frac{\partial m}{\partial \eta}\right)^2$

$$r_i' = \frac{y_i - \hat{m}_i}{\sqrt{\hat{\phi} V(\hat{m}_i)(1-h_i)}}$$

$$g(y) \approx g(m) + g'(m)(y-m)$$

$$z = X\beta + \frac{\partial \eta}{\partial m}(y-m)$$

$$= X\beta + \varepsilon \rightarrow var(\varepsilon) = \left(\frac{\partial \eta}{\partial m}\right)^2 \phi V(m)$$

$$w^{-1} = \phi V(m)\left(\frac{\partial \eta}{\partial m}\right)^2$$

$$W^{1/2}z = W^{1/2}X\beta + W^{1/2}\varepsilon$$

$$\tilde{z} = \tilde{X}\beta + \tilde{\varepsilon}$$ ← reduced to linear regression

So

$$\underbrace{\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'}_{\text{hat matrix}} = W^{1/2}X(X'WX)^{-1}X'W^{1/2}$$

$$var(\tilde{\varepsilon}) = W^{1/2}\phi V W^{1/2} I$$

$$= \phi I$$

$$\frac{\tilde{z}_i - \tilde{x}_i'\beta}{\sqrt{(1-h_i)\phi}}$$

$$= \frac{W^{1/2}(z - x_i'\beta)}{\sqrt{(1-h_i)\phi}}$$

$$= \frac{W_i^{1/2}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)(y_i - \hat{\mu}_i)}{\sqrt{(1-h_i)\phi}}$$

$$W = \left(\frac{\partial \mu}{\partial \eta_i}\right)^2 V^{-1}$$

$$W^{1/2} = \left(\frac{\partial \mu}{\partial \eta}\right) V^{-1/2}$$

$$= \frac{\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\left(\frac{\partial \eta_i}{\partial \mu_i}\right) V_i^{-1/2}(y_i - \hat{\mu}_i)}{\sqrt{(1-h_i)\phi}}$$

$$= \frac{(y_i - \hat{\mu}_i)}{\sqrt{\phi V(\mu_i)(1-h_i)}}$$

$$\sum_i h_i = p = tr(H)$$

$$= tr\left(X(X'X)^{-1}X'\right)$$

$$= tr\left((X'X)^{-1}X'X\right)$$

$$= tr(I_{p\times p}) = p$$

so each datapoint should contribute p/n

if $h_i \geq \frac{2p}{n}$, it's high leverage

in linear models, large h_i = covariate is different. But not necessarily in GLM.

Because $\tilde{X}$ contains a weight. A very different data point can have a small weight and thus exert less influence.

Cook's distance:

Linear:

$$\hat{\beta}_{(i)} - \hat{\beta}$$

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'(X'X)(\hat{\beta}_{(i)} - \hat{\beta})}{pS^2}$$

$$= \frac{r_i^2}{p} \cdot \frac{h_i}{1 - h_i}$$

monotone function of leverage

$\Rightarrow$ large leverage $\rightarrow$ large D_i

GLM:

add weight matrix from linear approx.

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})(X'WX)(\hat{\beta}_{(i)} - \hat{\beta})}{p\,\phi}$$

↑ # params

March 14

Matched case-control design

Y: outcome { 1 disease; 0 control }

X: primary risk factor: smoking status

↖ secondary risk factors: alcohol

Z: potential confounders: age, education levels, etc.

sample the data so that they match on potential confounders.

	1	...	m_i
$i=1$	x	x	x x
.			
.			
.			
N			

N cases, for each case collect m_i controls matched w/ potential confounders.

Fit a logistic regression model:

$$\text{logit}\,\pi = \beta_0 + X\beta_X + Z\beta_Z$$

Assuming linear dependence of logit on confounders (not guaranteed to be true in the data).

how to get rid of this assumption?

$\text{logit}\ \pi_i = \alpha_i + x_{i\beta x}^T$ for the i^{th} matching group

each matching group has its own intercept

primary risk factor

- b/c all datapoints in the group have the same confounder values.
- No longer making assumptions about confounders.
- Issue: Many parameters to fit.
  - N cases, Nk datapoints (k-constant)
  - O(N) params

    ⇒ Not enough datapoints to estimate in a consistent way

  - H^{-1} will be expensive to compute in Newton-Raphson
- Can't directly use this model

conditional likelihood approach

- used to handle above issues.

$$P\left(Y_{i0}=1, Y_{i1}=0, Y_{i2}=0, \ldots Y_{im_i}=0 \,\middle|\, \sum_j Y_{ij}=1\right)$$

(show this doesn't depend on α_i

- only one obs. is 1, rest are controls for ith matching group.

- $Y_{i0}=1$ — case
- $Y_{ij}=0,\ j=1,\ldots,m_i$ — controls
- $x_{ij1},\ldots,x_{ijp}$ — primary risk factors
- $z_{i1},\ldots,z_{iq}$ — matching covariates (same value across different people)

$$\text{logit}\ \pi_{ij} = \alpha_i + x_{ij}^T\beta$$

$$P\left(Y_{i0}=1, Y_{i1}=0, \ldots, Y_{im_i}=0 \,\middle|\, \sum_j Y_{ij}=1\right)$$

First evaluate the joint probability:

$$P(y_{i0}=1, y_{i1}=0, \ldots, y_{im_i}=0)$$

$$= \frac{\exp(\alpha_i + x_{i0}^T\beta)}{1+\exp(\alpha_i + x_{i0}^T\beta)} \prod_{j=1}^{m_i} \frac{1}{1+\exp(\alpha_i + x_{ij}^T\beta)}$$

$$= \frac{\exp(\alpha_i + x_{i0}^T\beta)}{\prod_{j=0}^{m_i}\{1+\exp(\alpha_i + x_{ij}^T\beta)\}}$$

can derive probabilities for $y_{i0}=0, y_{i1}=1, y_{i2}=0 \cdots$ and all other configurations.

$$P\left(\sum_j y_{ij}=1\right) = \sum_{j=1}^{m_i} \frac{\exp(\alpha_i + x_{ij}^T\beta)}{\prod_{k=0}^{m_i}\{1+\exp(\alpha_i + x_{ik}^T\beta)\}}$$

$$P\left(y_{i0}=1, y_{i1}=0, \ldots, y_{im_i}=0 \mid \sum_{j=1}^{m_i} y_{ij}=1\right)$$

$$= \frac{\exp(\alpha_i + x_{i0}^T\beta)}{\sum_{j=0}^{m_i}\exp(\alpha_i + x_{ij}^T\beta)} = \frac{\exp(x_{i0}^T\beta)}{\sum_{j=0}^{m_i}\exp(x_{ij}^T\beta)}$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- After taking log, can estimate β using Newton-Raphson / Fisher scoring
- variance is the inverse of the last iteration of NR

since a_i disappears, we can't say anything about the population.

- cannot draw conclusions about potential confounders
- cannot durive absolute risk (i.e. prob someone gets lung cancer)
- can only evaluate relative risk for cancer between smoker and non-smoker, adjusting for confounders.

GLMs

- random
- systematic
- link

difference between NR and Fisher scoring

generalized Pearson → overdispersion

Poisson log-linear and multinomial (table of counts)

reformulation

$\frac{D}{n-p} \sim \chi^2_{n-p} \approx \phi$ (not necessarily consistent)

Pearson X^2 can be used to estimate dispersion param

↓ this one is consistent

drive algs for QL fitting / $\hat{\beta}$ estimate

- how to draw inference?
- asymptotic variance?

QL is an optimal estimating function within cases of given functions.

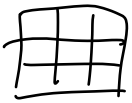
model checking

- residual vs. fitted
- score test ≈ drop in deviance asymptotically
- isolated data points (x_i, y_i)

. influence, cook's distance

.

Quasi-symmetric vs. symmetric



$\log m_{ij} = \lambda_0 + \lambda_i + \lambda_j + \lambda_{ij}$

$\log m_{ii} = \lambda_0 + 2\lambda_i$

↑ no λ_{ij}

Quasi-symmetric: symmetric after getting rid of marginal (row/column) effects