Inference: (Wald. Score. Likelihood Ratio Test)

$H_0: \beta = \beta_0 \ (\beta_j = 0)$ versus $H_1: \beta \neq \beta_0 \ (\beta_j \neq 0)$

Test statistic: $z^* = \frac{b_j - 0}{s\{b_j\}}$ $z^* \sim N(0,1)$, $z^{*2} \sim \chi_1^2$

Wald → Horizontal

Confidence Interval: (β): $b_j \pm Z_{1-\alpha/2} \cdot s\{b_j\}$

Monotone: $e^{\beta_0 \pm Z_{1-\alpha/2} s \beta_0}$ for OR.

Score: Test slope = 0. Wald: $T_{SW} = (\hat\beta - \beta_0)^T I(\hat\beta)(\hat\beta - \beta_0)$

$T_{SS} = S(\beta_0)^T \ I^{-1}(\beta_0) \ S(\beta_0)$

Likelihood: $T_{SLR} = 2[l(y, \hat\beta) - l(y, \beta_0)]$.

R: −Residual Deviance + Null Deviance $= G^2 - G_0^2$

Confidence Interval.

$\hat\beta_j \pm Z_{1-\alpha/2}\, se(\hat\beta_j)$. $f(\beta_0, \beta_1) \Rightarrow$ Vcov.

Likelihood Ratio: $l(y, \mu) = y\theta - b(\theta) + c(y)$

$\boxed{\text{Deviance } D}$: $D(y, \hat\mu) = 2[l(y,y) - l(y, \hat\mu)]$ measures discrepancy of 2 fits. $\Rightarrow \chi^2_{n-p}$ Large ⇒ Lack of fit

Normal: $l(y, \mu) = -\frac{n}{2}\log(2\pi) - \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{2}$

$D(y, \hat\mu) = RSS = \sum_{i=1}^{n} (y_i - x_i^T \hat\beta)^2$

Poisson: $\theta_i = \log \mu_i = x_i^T \beta$.

$l(y, \mu) = -\sum_{i=1}^{n} (\log y_i! - y_i \log \mu_i + \mu_i)$

$D(y, \hat\mu) = 2\sum_{i=1}^{n} (y_i \log y_i - y_i - y_i x_i^T\hat\beta + \exp(x_i^T\hat\beta))$

Deviance Analysis [Nested Models]

M_0 (smaller model) M_1 (larger Model)

$D_{M_0} - D_{M_1} \overset{\alpha}{\approx} \chi^2_{p-q}$ under M_0.

Reject smaller models when difference in deviance is large

Dispersion: F-test. → Model Fit $= V(y_i) = b''(\theta)$

$\boxed{\text{Pearson } \chi^2}$: $G = \sum_{i=1}^{n} (y_i - \hat\mu_i)^2 / V(\hat\mu_i)$, $\hat\mu_i = g^{-1}(x_i^T\hat\beta)$.

If the P-parameter model is true, G may have an approximate distribution $\chi^2(n-p)$.

Residual: Pearson Residual.

$r_{P_i} = \frac{y_i - \hat\mu_i}{\sqrt{V(\hat\mu_i)}} = \frac{y_i - \hat\mu_i}{\sqrt{b''(\theta)}}$

$G = \sum_i r_{P_i}^2$

Deviance Residual.

$r_{D_i} = \text{sign}(y_i - \hat\mu_i)\sqrt{d_i}$

$d_i = D(y_i, \hat\mu_i)$

Preferred → $D(y, \hat\mu) = \sum_i r_{D_i}^2$

Contingency Table: Test two variables are correlated.

Poisson Model. $Y_{ij} \sim \text{Poisson}(\mu_{ij})$. <All unknown>

Ho: Row ⫫ col. $\hat\mu_{ij} = y_{ij}$ $(i = 1, \cdots I; \ j = 1, \cdots J)$

$\log(\hat\mu_{ij}) = \mu + \alpha_i + \beta_j$, $\sum_{i=1}^{n} \alpha_i = \sum \beta_j = 0$.

Goodness: $IJ - (I + J - 1)$

$D \text{ (or } G) \sim \chi^2((I-1)(J-1))$

Multinomial Model (n is fixed. Total count)

Saturated. No. of parameter: $IJ - 1$.

MLE $\hat\pi_{i\cdot} = \frac{Y_{i+}}{n}$, $\hat\pi_{\cdot j} = \frac{Y_{+j}}{n}$.

$D \sim \chi^2((I-1)(J-1))$, under Ho.

D is large, Reject Null, and claim there is significant association between 2 factors.

Product Multinomial Model. (With fixed Row Total)

$\hat\pi_{j|i} = \frac{Y_{ij}}{n_{i+}}$

$D \sim \chi^2((I-1)(J-1))$ under Ho

Exponential Family

$f(y) = \exp\left\{ \frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \right\}$.

· $E(y) = b'(\theta) = \mu(\theta) \Leftarrow b'(\theta) = \mu$

· $Var(y) = b''(\theta)\phi$. Variance function in terms of μ.

GLM Basics: Systematic + Random = link function

$$\theta \text{ — } \overset{y}{\mu} \text{ - - - - } \eta = X^T\beta. \quad \left\{\begin{array}{l} g(\mu) = \eta = \theta \\ b'(\theta) = \mu. \end{array}\right\} \text{Canonical link } g(\mu) = \mu?$$

Normal: $g(\mu) = \mu$.; Binary: $g(\pi) = \log\frac{\pi}{1-\pi}$;

Poisson: log link $g(\lambda) = \log(\lambda)$ Expo: $g(\mu) = -\frac{1}{\mu}$;

Model Diagnostics: (Null vs. Full Model, by itself)

R: Null model: $(n-1)$ degree of f

Residual Deviance: $(n-1-p)$ df. p = No. of predictors

Likelihood: $l(y, \mu) = y\theta - b(\theta) + c(y)$

Deviance: $D(y, \hat\mu) = 2\{l(y,y) - l(y, \hat\mu)\}$.

Score: $\frac{\partial l}{\partial(\lambda)}$ First-order derivative of log-likeli

Information: $E[-\frac{\partial^2 l(y, \pi)}{\partial \pi^2}]$ $\hat\beta_{MLE} = \bar y$.

Invariance: OR of Exposure in case = OR of Disease in E

$$OR = \frac{\text{odds}(E=1 \mid \text{Case})}{\text{odds}(E=1 \mid \text{Control})} = \frac{\text{odds}(D=1 \mid Ex)}{\text{odds}(D=1 \mid \text{Non-E})}$$

RR can Not be used in Retrospective Model.

Vcov: variance-covariance Matrix, Delta Method)

$$Var[f(\hat\beta_0, \hat\beta_1)] = \left(\frac{\partial f}{\partial\beta_0}, \frac{\partial f}{\partial\beta_1}\right) \begin{pmatrix} Var(\hat\beta_0) & Cov(\hat\beta_0, \hat\beta_1) \\ Cov(\hat\beta_0, \hat\beta_1) & Var(\hat\beta_1) \end{pmatrix} \begin{pmatrix} \frac{\partial f}{\partial\beta_0} \\ \frac{\partial f}{\partial\beta_1} \end{pmatrix}$$

① CI for π:

$\hat\eta = \beta_0 + \beta_1 X$ $(X = 20)$ ⇒ 95% CI for η

$\hat\eta \pm 1.96\sqrt{Var(\hat\eta)}$

Variance

$\Rightarrow$ CI for π: $\left[\frac{e^{low}}{1+e^{low}}, \frac{e^{high}}{1+e^{high}}\right]$

② CI for X:

$\log\frac{0.9}{0.1} = \hat\beta_0 + \hat\beta_1 x_0 \Rightarrow x_0 = \frac{\log 9 - \hat\beta_0}{\hat\beta_1}$

$\left(\frac{\partial \hat x_0}{\partial\beta_0}, \frac{\partial \hat x_0}{\partial\beta_1}\right) = \left(-\frac{1}{\hat\beta_1}, -\frac{\log 9 - \hat\beta_0}{\hat\beta_1^2}\right)$

$Var(\hat x_0) \Rightarrow \hat x_0 \pm 1.96\sqrt{Var(\hat x_0)}$

③ Overdispersion. Wald-statistic. / Score /

Wald $= T_{SW} / \phi$ Score $= T_{SS} / \phi$

CI in overdispersion: $Var(\eta^*) = \phi \cdot Var(\eta)$

④

$RR = \pi_2 / \pi_1$

$OR = \frac{\pi_2 / (1-\pi_2)}{\pi_1 / (1-\pi_1)}$

Prospective Model for Retrospective data.

$P(D|S=1) = \frac{\exp(\alpha^* + X\beta)}{1+\exp(\alpha^* + X\beta)}$ Never be interpreted!

Interpretation: $\exp(\beta)$ provided odds ratios of disease corresponding to unit change in different covariates.

Retrospective NOT used for RR.

Multinomial Distribution (No order - Baseline logistic)

$\theta_{ij} = \frac{\log \pi_{ij}}{\log \pi_{i1}} = \alpha_j + \beta_j X$; J-1 logits $\pi_1 + \pi_2 + \cdots + \pi_J = 1$

First

Inter: θ_{ij} is the log odds for Response category j vs 1. β_{jc} is the log odds Ratio with One unit change of X_c.

Exchangable Categories: $\log\frac{\pi_{ij}}{\pi_{ir}} = \log\frac{\pi_{ij}}{\pi_{i1}} - \log\frac{\pi_{ir}}{\pi_{i1}} = X_i^T(\beta_j - \beta_r)$

Goodness: G.D; $\sim \chi^2(n-p)(J-1)$ under Null

Ordinal Response (Proportional Odds Model)

$[\log(\frac{P_j}{1-P_j}) = \log(\frac{Pr(Y \le j|X)}{1-Pr(Y\le j|X)}) = \log\frac{\pi_1+\cdots+\pi_j}{\pi_{j+1}+\cdots+\pi_J}$

all coefficients same, only intercept varies $\Rightarrow J-1+P$

$Pr(Y \le j) = \frac{\exp(\alpha_j + \beta x)}{1+\exp(\alpha_j + \beta X)}$ combined category does not change β_{p-1}

Interpretat: β_K is the log Odds Ratio (of lower Cate vs higher Categories) for a unit change of X_K

Likelihood stati: Residual Deviance(New) - RD(old) | χ^2_{df}

the odds of lower cate vs. Higher cate 1 or C2 combined.

Count Respond (Poisson + offset + NB + zero-inflated)

$P(Y=y) = \frac{e^{-\lambda}\lambda^y}{y!}$, y = 1 to infty, $E(Y) = Var(Y) = \lambda$

$Y_i \sim Poisson(\lambda_i)$, $\log(\lambda_i) = X_i^T\beta$. $\theta_i = \log\lambda_i$

- β_j = difference in log[E(y)] following a unit change in predictor X_j, while others are constant.

Goodness: Pearson $X^2 = \sum_{i=1}^n \frac{(y_i - \hat\lambda_i)^2}{\hat\lambda_i}$ ⇒ different with dispersion Deviance $D = 2\sum_{i=1}^n (y_i \log\frac{y_i}{\hat\lambda_i} - (y_i - \hat\lambda_i))$

Both $\sim \chi^2(n-p)$, Ho: our Model. H1: Saturated (full)

offset: $\begin{cases} Y \sim Poisson(\mu_i) \\ E(Y_i) = \mu_i = n_i\lambda_i = n_i \exp(X_i\beta) \end{cases}$ $\log\mu_i = \underbrace{\log n_i}_{\text{offset}} + X_i^T\beta$.

Inter: β is the log rate ratio with one unit change of X.

Overdispersion: $Var(Y) > E(Y)$.

Source: Correlated Sampling; clustering; Poisson-Gamma Model

$\hat\beta_Q = \hat\beta_{MLE}$, $Cov(\hat\beta_Q) = Cov(\hat\beta_{MLE}) \cdot \phi$

Parameter: $H_0: \beta_2 = 0$ vs $\beta_2 \neq 0$

Test two models without over-dispersion, F-test

$\hat\phi = G/(n - P_1 - P_2)$; $\frac{D_1 - D_2}{P_2\hat\phi} \overset{H_0}{\sim} F(P_2, n-P_1-P_2)$

NB distrib: $Y \sim NB(r,p)$ Handle over-dispersion

mean: $Pr/(1-P)$; variance $= pr/(1-p)^2$

$Y_i \sim NB(\mu_i, \phi)$: $\log\mu_i = X_i\beta$.

ZIP Model. (Two models: 0 and Potentially Non-0)

$P(Z_i = 0) = \pi_i$ Binomial: $\text{logit}(\pi_i) = Z_i r$

$Y_i|(Z_i=0) = 0$ Poisson: $\log(\lambda_i) = X_i\beta$

$Y_i|(Z_i=1) \sim Poisson(\lambda_i)$

E
A B = 0

Interpre: (2 Parts).

Bino: The log odds Ratio of having A is expected to be xx for a vs b, adjusting for other covariates.

Poisson: For subjects in B, the xx for a on average is xx times, adjusting for other covariates.

Binary Response $(Y = 0, 1)$ $E(Y) = \pi = P(Y=1)$

Grouped data - Ungrouped data: They are independen two estimates are the same. Unique value, cannot be grouped. Model diagnostic Different!

$\boxed{\text{Binomial}}$: $P(Y=y) = \binom{n}{y}\pi^y(1-\pi)^{n-y}$, $y = 0, 1, \cdots, n$

1. Log-likelihood f: $L(y;\pi) = \sum_{i=1}^n [Y_i\log\pi_i + (1-Y_i)\log(1-\pi_i)]$ n=1

$L(y_i;\pi_i) = \sum_{i=1}^n [Y_i\theta_i - b(\theta_i)] \Rightarrow \theta = \log\frac{\pi_i}{1-\pi_i}$, $b(\theta_i) = \log(1+\exp(\theta_i))$

GLM Link: ① logit: $g_1(\pi) = \log\frac{\pi}{1-\pi}$; $g_1^{-1}(\eta) = \frac{e^\eta}{1+e^\eta}$

② Probit: $g_2(\pi) = \Phi^{-1}(\pi)$ $g_2^{-1}(\eta) = \Phi(\eta)$

③ C-log-log: $g_3(\pi) = \log(-\log(1-\pi))$ $g_3^{-1}(\eta) = 1 - e^{-e^\eta}$

2 Log-like (m_i, π_i) $L(\beta) = \sum_{i=1}^n [y_i\log(\frac{\pi_i}{1-\pi_i}) + m_i\log(1-\pi_i) + \log\binom{m_i}{y_i}]$

3 Goodness of fit - Group data

Genealized χ^2 statistic

$G(\hat\pi, y) = \sum_{i=1}^n \frac{(y_i - m_i\hat\pi_i)^2}{m_i\hat\pi_i(1-\hat\pi_i)}$

Deviance: $2\sum_{i=1}^n [y_i\log\frac{y_i}{m\hat\pi_i} + (m-y_i)\log\frac{m_i - y_i}{m_i(1-\hat\pi_i)}]$

App $\sim \chi^2(n-p)$ when m_i large

Ungroup data

Deviance & Genealized χ^2 can not be used. χ^2_{HL} $(g \approx 10)$

Hosmer - Lemeshow statistic

$\chi^2_{HL} = \sum_{i=1}^g \frac{(O_i - n_i\bar\pi_i)^2}{n_i\bar\pi_i(1-\bar\pi_i)}$

$\sim \chi^2(g-2)$: O_i: # of 1 in ith group.

4. Confidence interval.

β_j: $\hat\beta_j \pm Z_{1-\alpha/2}\, se(\hat\beta_j)$

$f(\beta)$: $f(\hat\beta) \pm Z_{1-\alpha/2}\, se(f(\hat\beta))$ $\hat\eta_L, \hat\eta_R: X_*^T\hat\beta \pm Z_{\alpha/2}\sqrt{X_*^T I(\hat\beta)^{-1} X_*}$

CI for π_*: $[g^{-1}(\hat\eta_L), g^{-1}(\hat\eta_R)]$

5. Intepretation:

$\log(\frac{\pi}{1-\pi}) = X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ $\begin{cases}\pi: \text{risk} \\ \pi/1-\pi = \text{odds}.\end{cases}$

β_0: log odds for $X_1 = X_2 = 0$ (subpopulation)

β_1: the log odds Ratio per unit change of X_1 holding X_2 fixed (Dichot: male vs. female).

6. Variance: $\hat{x}_0 = f(\beta_0, \beta_1)$: $Var(f(\beta_0, \beta_1)) =$

$Var(\hat{x}_0) = (\frac{\partial X_0}{\partial\beta_0})^2 Var(\hat\beta_0) + (\frac{\partial X_0}{\partial\beta_1})^2 var(\hat\beta_1) + 2(\frac{\partial X_0}{\partial\beta_0})(\frac{\partial X_0}{\partial\beta_1}) Cov(\hat\beta_0, \hat\beta_1)$.

CI: $[\hat{X}_0 - Z_{\alpha/2}\sqrt{Var(\hat{X}_0)}, \hat{X}_0 + Z_{\alpha/2}\sqrt{Var(\hat{X}_0)}]$

7. Over-dispersion in Binomial Ungrouped / Binary ⇒ No overdispersion

$E(Y) = m\pi$, $Var(Y) = m\pi(1-\pi)[1+(m-1)\rho]$

Quasi - $\hat\beta_Q = \hat\beta_{MLE}$ | $Var(\hat\beta_Q) = \phi\, Var(\hat\beta_{MLE})$

Source: ① Intra-class correlation; ② Hierarchical Sampling

Method ③ Not binomial ⇒ Beta-Binomial Distribution

ϕ: $G = \sum_{i=1}^n \frac{(y_i - m_i\hat\pi_i)^2}{m_i\hat\pi_i(1-\hat\pi_i)\phi} \sim \chi^2(n-p) \Rightarrow \hat\phi = G_0/(n-p)$ → from original model

Method 2: $D_0/(n-p) = \hat\phi$

Half-Normal plot: $|r_{(i)}|$ [Order the absolute value of residuals without dispersion]

$\Phi^{-1}(\frac{n+i+0.5}{2n+1.125})$ Reference line | Linear deviation from line indicates constant over-dispersion

Hypothesis Testing: $H_0: \beta = \beta_0$ vs $H_1: \beta \neq \beta_0$.

Wald Test: $T_s = (\hat\beta_Q - \beta_0)^T Var^{-1}(\hat\beta_Q)(\hat\beta_Q - \beta_0) = TS_w/\phi$

Score Test: $T_s = Q(\beta_0)^T Var^{-1}(Q(\beta_0))\, Q(\beta_0) = TS_s/\phi$

Deviance Analysis: $H_0: \beta_2 = 0$ vs $H_1: \beta_2 \neq 0$.

$\frac{(D_1 - D_2)/P_2}{G_0/(n - P_1 - P_2)} \sim F(P_2, n-P_1-P_2)$

Retrospective: Prob of having disease

$Y_1 \sim Bin(m_1, \pi_1)$, $Y_0 \sim Bin(m_0, \pi_0)$ $RR = \pi_1/\pi_0$

$\text{logit}(\pi_i) = \log\frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 X_i$ Exposure $OR = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}$

Example II

In a study of motor vehicle safety, 150 men and 150 women were interviewed to rate how important air conditioning and power steering were to them when they were buying a car.

Table: Importance Rating

It takes 3 values instead of 2. / 3 levels. ~~They have some~~ order. It's ordinal. / Multinormial.

		Response			
Sex	Age	Unimportant	Import	Very Import	Total
Women	18-23	26 (58%)	12 (27%)	7 (16%)	45
	24-40	9 (20%)	21 (47%)	15 (33%)	45
	> 40	5 (8%)	14 (23%)	41 (68%)	60
Men	18-23	40 (62%)	17 (26%)	8 (12%)	65
	24-40	17 (39%)	15 (34%)	12 (27%)	44
	> 40	8 (20%)	15 (37%)	18 (44%)	41
Total		105	94	101	300

Question of interest:

- How are sex and age related to the car preference?

The response of each subject is the preference level, which is categorical and ordinal. One plausible model would be

Probability of choosing unimportant.

$$Y_i \sim multinomial(\pi_{i1}, \pi_{i2}, \pi_{i3}),$$

Much more detailed compared with binomial logistic model.

and

$$\log\left(\frac{\pi_{i1}}{\pi_{i2}+\pi_{i3}}\right) = \mathbf{X}_i\beta_1;\ \log\left(\frac{\pi_{i1}+\pi_{i2}}{\pi_{i3}}\right) = \mathbf{X}_i\beta_2.$$

$\pi_{i1} + \pi_{i2} + \pi_{i3} = 1$

We will discuss in detail when we introduce the multinomial logistic regression.

From General Normal Distribution to a Generalized Context.

Example I

The Kyphosis data consist of measurements on 81 children following corrective spinal surgery. The binary response variable, Kyphosis, indicates the presence or absence of a postoperative deforming. The three covariates are, Age of the child in month, Number of the vertebrae involved in the operation, and the Start of the range of the vertebrae involved. The first five observations are shown below.

(binary Outcome)

	Kyphosis	Age	Number	Start
1	absent	71	3	5
2	absent	158	3	14
3	present	128	4	5
4	absent	2	5	1
5	absent	1	4	15

- Use 3 variables to predict kyphosis.
- The Response Variable is Not continuous, does not meet Assumption.
- But It's still a regression.
- std. linear Regression does not work.

Questions of interest are:

- How do the three explanatory variables relate to the response?
- Can they be used to screen the patients prior to the operation?

Due to the binary nature of the response, it's not reasonable to model it as a linear function of the covariates. A more appropriate model would be the logistic regression:

$$Y_i \sim Bernoulli(\pi_i),$$

→ Fits the distribution.

0 = absent

1 = present.

and

Latent Relation

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{X}_i\beta.$$

π_i: probability of Y_i being 1.

logistic Regression;

Logit link.

Part I: Generalized Linear Model

Outline

- Motivating examples
- Exponential family distributions
- Generalized linear model basics
- Logistic regression
- Multinomial regression
- Poisson regression
- Contingency table

Example

- Normal distribution $N(\theta, 1)$

$$f(x) = \exp\left[\frac{x\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{x^2}{2\sigma^2} - \log \sigma \right]$$

- Binary distribution $\text{Bin}(p)$

$$f(x) = \exp\left[x \cdot \theta - \log(1 + e^{\theta}) \right]$$

$$\theta = \log \frac{p}{1-p}$$

$$b(\theta) = \log(1 + e^{\theta})$$

$$b'(\theta) = \frac{e^{\theta}}{1 + e^{\theta}} = p \quad \text{confirms it's mean.}$$

$$b''(\theta) = \frac{e^{\theta}(1+e^{\theta}) - e^{2\theta}}{(1+e^{\theta})^2} = \frac{e^{\theta}}{(1+e^{\theta})^2} = \frac{e^{\theta}}{1+e^{\theta}} \cdot \frac{1}{1+e^{\theta}} = p \cdot (1-p)$$

Additional Examples

Final : $f(x) = \exp[x\theta - \log(1+e^\theta)]$

$b(\theta) = \log(1+e^\theta)$

pdf ⇒ Exponential. / Natural Parameters

- Bernoulli(p): $p^x(1-p)^{1-x}$; $f(x) = p^x(1-p)^{1-x} = \exp[\log[p^x(1-p)^{1-x}]]$
- Binomial(n, p); $= \exp[x \cdot \log p + (1-x)\log(1-p)]$
- Poisson(λ); $= \exp(x[\log p - \log(1-p)] + \log(1-p))$
- $\chi^2(k)$;
- Gamma(a,b); $= \exp[x \cdot \log\frac{p}{1-p} + \log(1-p)]$ (θ: $\log\frac{p}{1-p}$; $h(\theta)$: $\log(1-p)$)
- Beta(a,b);
- Negative Binomial(p,m).

set $\theta \triangleq \log\frac{p}{1-p}$

$\triangleq$ | := set to be equal to.

$e^\theta = \frac{p}{1-p} = \frac{1-p+2p}{1-p} = 1 + 2\frac{p}{1-p}$

$e^\theta = p/(1-p)$

$= \cancel{\text{}}$

$e^\theta = \frac{1}{1/p - 1}$

$\frac{1}{e^\theta} = \frac{1}{p} - 1$

$\Rightarrow \frac{1}{e^\theta} + 1 = \frac{1}{p}$

$\Rightarrow p = \frac{1}{\frac{1}{e^\theta}+1} = \frac{e^\theta}{1+e^\theta}$

Properties of Exponential Family

Consider the exponential family density function,

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

It has to be Negative <sign>.

Neat & Nice → Directly tell.
- Expectation: $\mathbb{E}(Y) = \mu = b'(\theta)$
- Variance: $\text{var}(Y) = \phi b''(\theta)$ Expection & Variance → $b(\theta)$
- $b(\theta)$ is always a convex function (Sec Der It's All positive) ① $E(\frac{\partial \log f}{\partial \theta}) = 0$. ② $E(\frac{\partial^2 \log f}{\partial \theta^2}) = -E[(\frac{\partial \log f}{\partial \theta}$

  sec Deri : ++ −− They're Universal.
- If we express var(Y) as a function of μ, we get the so-called variance function $V(\mu)$

(Y → variance 用 μ 表示)

In this course, we primarily focus on the exponential family with a single canonical parameter. For a random variable $y \sim EF(\theta)$, the density function is

exponential family.

$$f(y,\theta) = h(y)\exp\left(y\theta - b(\theta)\right)$$

functional.

$g(y)$

where θ is the natural parameter.

Sometimes, we need to consider a more general form with a scale (dispersion) parameter ϕ Adding scale parameter ϕ.

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

$$e^\theta = \frac{\pi}{1-\pi} \Rightarrow e^\theta = \frac{1}{\frac{1}{\pi}-1}$$

$$-[-n_i \log(1-\pi)]$$

$$n\log(1-\pi) = -[(-n)\log(1-\pi)] = -[\log(1-\pi)^{-n}]$$

$$\frac{1}{\pi} - 1 = \frac{1}{e^\theta}$$

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

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

$$1 - \frac{e^\theta}{1+e^\theta} = \frac{1}{1+e^\theta}$$

Gaussian Example

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \left(\frac{\mu^2}{2\sigma^2} + \log\sigma\right)\right]$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(\frac{x\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{x^2}{2\sigma^2} - \log\sigma\right)$$

$$\theta = \mu, \quad b(\theta) = \frac{\mu^2}{2} = \frac{\theta^2}{2}$$

Consider $x \sim N(\mu, \sigma^2)$ where σ^2 is known.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

forms a single-parameter exponential family distribution with the canonical parameter $\theta = \mu$ and the dispersion parameter $\phi = \sigma^2$.

Using Distribution for fitting Response ⇒ GLM.

Exponential Family

- Exponential family is a large family of distributions
- Many well known distributions are in the exponential family (e.g., normal, exponential, Poisson, Bernoulli, binomial, gamma, beta, etc)
- The density function satisfies certain form

$y \sim dist(\theta)$

$$f(y, \boldsymbol{\theta}) = h(y) \exp\left[\sum_{i=1}^{s} \eta_i(\boldsymbol{\theta}) T_i(y) - b(\boldsymbol{\theta})\right].$$

- After reparameterization (set $\theta_i = \eta_i(\boldsymbol{\theta})$), the canonical form is

$$f(\mathbf{y}, \boldsymbol{\theta}) = h(\mathbf{y}) \exp\left[\boldsymbol{\theta}^T T(\mathbf{y}) - b(\boldsymbol{\theta})\right],$$

where $\boldsymbol{\theta}$ is the (canonical) natural parameter.

① Exponential : (Normal Distribution)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} - \log \sigma\right)$$

Gaussian Example

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \log\sigma\right)$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2\sigma^2} + \frac{\mu}{\sigma^2}x - \left(\frac{\mu^2}{2\sigma^2} + \log\sigma\right)\right]$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left[\underbrace{-\frac{1}{2\sigma^2}}_{\theta_1} x^2 + \underbrace{\frac{\mu}{\sigma^2}}_{\theta_2} x - \underbrace{\left(\frac{\mu^2}{2\sigma^2} + \log\sigma\right)}_{b(\theta_1, \theta_2)}\right]$$

Consider $x \sim N(\mu, \sigma^2)$.

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

Remember all pdf : (gass + poisson + gamma)

which forms a two-parameter exponential family with

$$\theta_1 = \frac{\mu}{\sigma^2},\ \theta_2 = -\frac{1}{2\sigma^2},\ (\theta_1, \theta_2) \in \mathbb{R} \times (-\infty, 0).$$

Example III

The example concerns a type of damage caused by waves to the forward section of certain cargo-carrying vessels. For the purpose of setting standards for hull construction we need to know the risk of damage associated with the three classifying factors shown below:

- Ship Type: A-E;
- Year of Construction: 1960-64, 65-69, 70-74, 75-79;
- Period of operation: 1960-74, 75-79.

Two other variables are Aggregated months of service and Number of damage accidents.

ship	year	period	month	damage
A	60-64	60-74	127	0
A	60-64	75-79	63	0
A	65-69	60-74	1095	3
A	65-69	75-79	1095	4
A	70-74	60-74	1512	6

Positive Integers

Using GLM instead of LM.

Model counts data.

Question of interest:

- How does the number of damage accidents depend on the other variables?

The response is count-valued, and may follow a Poisson distribution. A more appropriate model would be

$$Y_i \sim Poisson(\lambda_i),$$

and

$$\log \lambda_i = \log m_i + \mathbf{X}_i \beta$$

special feature for poisson process (offset)

where m_i is the number of months of service.
This is a Poisson regression model with offset.

Generalized Linear Model

Canonical
$f(y,\theta) = h(y)\exp(y\theta - b(\theta))$
$E(y) = b'(\theta)$
$Var(y) = \phi\, b''(\theta)$
Can be a function y. Normally, single 'y'.

A linear model $Y = X\beta + \varepsilon$ can be equivalently expressed as

Equivalently as:

$$\Leftrightarrow \begin{cases} Y|X \sim N(\mu, \sigma^2), & \text{Distribution} \\ \mu = X\beta, \quad \beta \in (-\infty, \infty). & \leftarrow \text{identity function.} \end{cases}$$

The model specifies

- the response variable is continuous and *normally* distributed.
- some function of the mean μ (*the identity function here*) can be written as a linear combination of the covariates. (linear predictor)

A generalized linear model (GLM) generalizes normal linear regression models to address a broader class of data structures.

- Instead of being normal, the response variable Y could have any distribution from the **exponential family** distributions.
- The mean $\mu = \mathbb{E}(Y|X)$ may be a more general function of $X\beta$, rather than an identify function.

how mean is related to Response, through some function.

$$g(\mu) = X\beta \Leftarrow \text{This function: link function}$$

Examples

Disease Occurring Rate:

- In the early stages of a disease epidemic, the rate at which new cases occur can often increase exponentially through time.
- We are interested in predicting the number of new cases y_i on day x_i.
- Since y_i is count-valued, we may use the Poisson distribution to model it.

$\mu_i = E(y_i)$

- Let μ_i be the expected number of new cases on day x_i. Based on the description, the following model seems reasonable.

$$\mu_i = \beta_0 \exp(\beta_1 x_i)$$

μ_i be expressed in exponential form of y_i.

$$\log \mu_i = \log \beta_0 + \beta_1 X_i$$
$$= \beta_0^* + \beta_1 X_i$$

Kyphosis Data:

- Children are followed up after corrective spinal surgeries. We are interested in the relationship between clinical covariates and postoperative deforming.
- Binary response: presence or absence of a postoperative deforming (denoted by a binary variable y_i) only takes value : 0 / 1.

$$y_i \sim Bernoulli(\pi_i)$$

[0, 1]

probability of having the postoperative deforming

- Assume log odds of deforming is associated with the linear predictor:

值域 mismatch

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

$\pi_i = X_i\beta$.

$(-\infty, +\infty)$

Problematic

$$\frac{\pi_i}{1-\pi_i} = -\frac{\pi_i - 1}{\pi_i - 1} - \frac{1}{\pi_i - 1} \quad (-\infty, +\infty)$$
$$= \frac{1}{1-\pi_i} - 1$$

Example (Logistic Regression)

Suppose $y_i \sim Bin(1, p_i)$, $i = 1, \ldots, n$, are independent 0/1 indicator responses, and $\boldsymbol{x}_i$ denote a $p \times 1$ vector of predictors for individual i. The log likelihood is as follows

$P(Y = y_i \mid p_i) \Rightarrow L(p_i \mid y_i)$ β maximizes likelihood

likelihood given condition of y.

$$
\begin{aligned}
l(\boldsymbol{y}|\boldsymbol{\beta}) &= \sum_{i=1}^{n} \log\left[p_i^{y_i}(1-p_i)^{(1-y_i)}\right] \\
L(\beta|y) &= \sum_{i=1}^{n} y_i \log\left(\frac{p_i}{1-p_i}\right) - \log\left(\frac{1}{1-p_i}\right) \\
&= \sum_{i=1}^{n} (y_i\theta_i - \log(1+e^{\theta_i})).
\end{aligned}
$$

Choosing the canonical link, the logit link in this case,

$$\eta_i = \theta_i = \log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\beta},$$

which leads to

$$l(\boldsymbol{y}|\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \boldsymbol{x}_i^T \boldsymbol{\beta} - \log(1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}})\}.$$

No closed-form MLE!

Geometric $\Rightarrow$ Special case of Negative Binomial $n = 1$

Examples

Normal: ① $f(y) = \frac{1}{\sqrt{2\pi}} e^{\mu y - \frac{\mu^2}{2} - \frac{y^2}{2}}$

$\theta = \mu$, $b(\theta) = \frac{\theta^2}{2}$, $b'(\theta) = \theta$ $b'^{-1}(\mu) = \mu$, $g(\mu) = \mu$.

② Poisson $f(y) = \frac{1}{y!} e^{-\lambda} \cdot \lambda^y = \frac{1}{y!} e^{-\lambda + \log\lambda \cdot y} = \frac{1}{y!} e^{y\theta - e^\theta}$ $\theta = \log\lambda$ $\lambda = e^\theta$

Only one / Unique for link.

$b(\theta) = e^\theta$ $b'(\theta) = e^\theta$ $b'^{-1}(\mu) = \log\mu$.

- Normal: Identity link $g(\mu) = \mu$
- Poisson: Log link $g(\lambda) = \log(\lambda)$
- Binary: Logit link $g(\pi) = \log \frac{\pi}{1-\pi}$
- $\frac{1}{\lambda} \to$ Mean Exponential: Negative reciprocal link $g(\mu) = -1/\mu$

③ Binary

$f(y) = \pi^y (1-\pi)^{1-y} = e^{y\log\pi + (1-y)\cdot\log(1-\pi)} = e^{y\log\frac{\pi}{1-\pi} + \log(1-\pi)} = e^{y\theta - \log(1+e^\theta)}$

$\theta = \log\frac{\pi}{1-\pi}$ $\Rightarrow \pi = \frac{e^\theta}{1+e^\theta}$ $b(\theta) = \log(1+e^\theta)$

$b'(\theta) = \frac{e^\theta}{1+e^\theta}$ $b'^{-1}(\mu) = \log\frac{\mu}{1-\mu}$

$b'(\theta) = \mu$, $g(\mu) = \eta$, $\theta = \eta$

$[b'(\theta)]^{-1} = \frac{1}{\frac{1}{e^\theta} + 1} = \log(\frac{\mu}{1-\mu})$

④ Tips: $\theta = \underbrace{\log\frac{\pi}{1-\pi}}_{\text{Canonical}} = \eta$ (Shortcut) parameter = Mean

Exponential

$f(y) = \lambda \cdot e^{-\lambda x} = e^{-\lambda x + \log\lambda}$

$\theta = -\lambda =$ ~~$e^{x\theta}$~~

$f(y) = e^{x\theta + \log(-\theta)} = e^{x\theta - [-\log(-\theta)]}$

$b(\theta) = -\log(-\theta)$ $b'(\theta) = -\frac{1}{-\theta}$ $b'(\mu) = -\frac{1}{\mu}$

$b'(\lambda) = +\frac{1}{\lambda}$

$b'(\lambda) = \frac{1}{\lambda} = \mu$

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

GLM Model Fitting

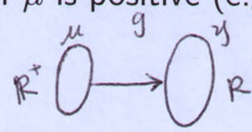
- In GLM of $(\boldsymbol{y}, \boldsymbol{X})$ with a given link function, we can write out the likelihood function as a function of β
- To estimate β, we use maximum likelihood (ML) approach
- However, unlike LM, no closed-form MLE for β
- Need to maximize the log likelihood function numerically
  - Newton-Raphson method
  - Fisher-Scoring method
  - Iteratively reweighted least squares (IRLS) algorithm

Examples of link functions:

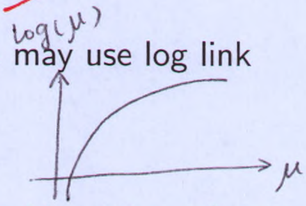
- If μ is unbounded (e.g., Normal distribution), may use identity link

$$g(\mu) = \mu = \eta$$

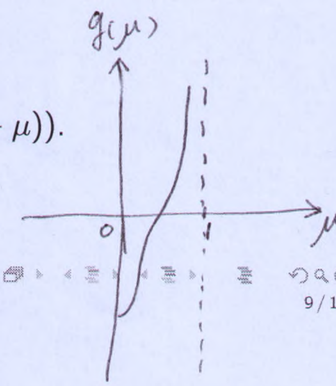
- If μ is positive (e.g., Poisson, Exponential), may use log link



$$g(\mu) = \log(\mu)$$

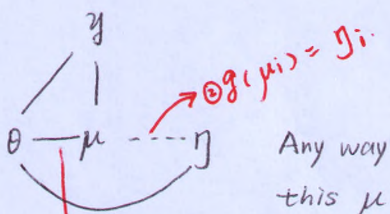


- If μ is bounded (e.g., Binomial, Bernoulli), without loss of generality, consider $0 < \mu < 1$:
  - logit link: $g(\mu) = logit(\mu) = \log \frac{\mu}{1-\mu}$;
  - probit link: $g(\mu) = \Phi^{-1}(\mu)$;
  - complimentary log-log link: $g(\mu) = \log(-\log(1-\mu))$.



Canonical Link Functions

Goal: $g(\mu)$? = $g[E(y)]$.



Any way of bypassing? this μ.

Parameter relations:

$$\theta \longleftarrow \mu = \mathbb{E}(y) \longrightarrow \eta$$

① $b'(\theta_i) = \mu_i$

Can we connect the natural parameter θ with the linear predictor?

- Canonical Link: the special link function g which makes $\theta = \eta$. ☆ Set $\times\beta$
- $g(\mu) = \eta = \boxed{\theta = b'^{-1}(\mu)}$, namely 易混 → 求倒 inverse

~~$[b^{-1}(\mu)]'$~~ prime → reverse.

Canonical link func: $g = (b')^{-1}$

- We know b' is strictly increasing and differentiable, so its inverse is a valid link function.

$\theta = b'^{-1}(\mu) = g(\mu)$

$b'(\theta) = \mu$

$g(\mu) = \eta$

μ

$\theta = \eta$

Link function

Reparametrization

y

$\theta - \mu - - \eta = x\beta$

$b'(\theta) = \mu$

$\eta = g(\mu)$

$g[b'(\theta)] = \eta$

Suppose y has a density from an exponential family:

$$f(y;\theta,\phi) = e^{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)}.$$

For n observations, $(y_i, x_{i1}, \cdots, x_{ip})$, $i = 1, \cdots, n$

- $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$ is the linear predictor.
- $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$ is the parameter of interest, and needs to appear somehow in the likelihood function.
- A link function g relates the linear predictor η_i to the mean parameter μ_i:

Connection :

$$\eta_i = g(\mu_i)$$

- With a little abuse of notation, sometimes we write $\boldsymbol{\eta} = g(\boldsymbol{\mu})$ to represent entry-wise mapping

"1 - 1" (invertible)

- g is required to be *monotone increasing* and *differentiable*.

$$\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}) = g^{-1}(\boldsymbol{X}\beta).$$

- It's generally preferred that the image of g is $\mathbb{R}$. (The domain depends on the exponential family.)

Determine the linear predictor .

① Through the empircal evidence / hypothesis

② Range of domain for 'Mean'

③ Canonical link.