Categorical Data Analysis

Chapter 4

Deyuan Li
School of Management
Fudan University

Fall 2015

Outline

4.1 Generalized Linear Models

Classical linear models (CLM):

$$Y = \alpha + \beta x + \varepsilon, \quad \epsilon \sim N(0, \sigma^2),$$

which is equivalent to, for (Y_i, x_i),

$$E[Y_i] = \alpha + \beta x_i, \quad Y_i \sim N(\alpha + \beta x_i, \sigma^2), \quad \text{i.i.d. errors.}$$

How to extend it to generalized linear models (广义线性模型, GLMs)?

4.1.1 Components of generalized linear models

random component; systematic component; link function

1) The *random component* （随机部分） consists of a response variable Y with independent observations $(y_1, \ldots, y_N)$ from a distribution in the natural exponential family, which density or mass function is

The value of the parameter θ_i may vary for $i = 1, \ldots, N$, depending on values of explanatory variables.

The term $Q(\theta)$ is called the *natural parameter*.

4.1.1 Components of generalized linear models

More general formula for the density is

$$f(y;\theta) = a(\theta)\, b(y)\, \exp\left\{ \sum_{i=1}^{k} T_i(y) Q_i(\theta) \right\}$$

Distributions:

Normal, Gamma, Binomial, Multinomial, Possion

4.1.1 Components of generalized linear models

2) The *systematic component*（系统部分） relates a vector $(\eta_1, \ldots, \eta_N)$ to the explanatory variables through a linear model.

Let x_{ij} denote the value of predictor j ($j = 1, 2, \ldots, p$) for subject i. Then

$$\eta_i = \sum_j \beta_j x_{ij}, \quad \text{for } i = 1, \ldots, N.$$

η_i is called the *linear predictor*. Usually, $x_{i1} = 1$ for all i corresponds to an intercept.

4.1.1 Components of generalized linear models

3) The *link function*（连接函数） connects the random and the systematic components.

Let $\mu_i = E(Y_i)$, $i = 1, \ldots, N$. The model links μ_i to η_i by

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

where g is a monotonic, differentiable function.

Thus, g links $E(Y_i)$ to explanatory variables through the formula

4.1.1 Components of generalized linear models

Identical link（恒等连接）: $g(\mu) = \mu$, the link function for ordinary regression with normally distributed Y.

Canonical link（典型连接）: $g(\mu_i) = Q(\theta_i)$, the link function transforming the mean to the natural parameter.
(Note: $\mu_i = \mu_i(\theta_i) \Rightarrow \theta_i = \theta_i(\mu_i)$.)

In summary, a GLM is a linear model for a transformed mean of a response variable that has distribution in the natural exponential family.

4.1.2 Binomial logit models for binary data

Bernoulli distribution: for $y = 0$ and 1,

$$\begin{aligned} f(y;\pi) &= \pi^y (1-\pi)^{1-y} = (1-\pi)[\pi/(1-\pi)]^y \\ &= (1-\pi)\exp\left[y\log\left(\frac{\pi}{1-\pi}\right)\right] \end{aligned}$$

This is in the natural exponential family, with $\theta = \pi$, $a(\pi) = 1-\pi$, $b(y) = 1$ and $Q(\pi) = \log[\pi/(1-\pi)]$.

The natural parameter $Q(\pi) = \log[\pi/(1-\pi)]$ is the log odds of response $y = 1$, i.e., the *logit* of π.

$\Rightarrow$ The canonical link function is the logit link, $\eta = \log[\pi/(1-\pi)]$.

GLMs using the logit link are often called *logit models*（Logit 模型）.

4.1.3 Poisson loglinear models for count data

Poisson distribution: for $y = 0, 1, 2, ...$

$$f(y;\mu) = \frac{e^{-\mu}\,\mu^y}{y!} = \exp(-\mu)\left(\frac{1}{y!}\right)\exp[y\log(\mu)].$$

This is in the natural exponential family, with $\theta = \mu$, $a(\mu) = \exp(-\mu)$, $b(y) = 1/y!$ and $Q(\mu) = \log(\mu)$.

$\Rightarrow$ The canonical link function is the log link, $\eta = \log(\mu)$.

The model using this link

$$\log(\mu_i) = \sum_j \beta_j x_{ij}, \quad \text{for } i = 1, \ldots, N,$$

is called a *Poisson loglinear model*（泊松对数线性模型）.

4.1.4 Generalized linear models for continuous responses

The class of GLMs also includes models for continuous responses. The normal distribution is in a natural exponential family with dispersion parameters.

Its natural parameter is the mean. $\Rightarrow$ The canonical link function is the identical link.

Table 4.1 Types of Generalized Linear Models for Statistical Analysis

Random Component	Link	Systematic Component	Model	Chapters
Normal	Identity	Continuous	Regression	
Normal	Identity	Categorical	Analysis of variance	
Normal	Identity	Mixed	Analysis of covariance	
Binomial	Logit	Mixed	Logistic regression	5 & 6
Poisson	Log	Mixed	Loglinear	8 & 9
Multinomial	Generalized logit	Mixed	Multinomial response	7

4.1.5 Deviance（偏差）

For a particular GLM model with p parameters, let

- $\mathbf{y} = (y_1, \ldots, y_N)$ denote observations,
- $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$ denote means,
- $L(\boldsymbol{\mu}; \mathbf{y})$ denote the log-likelihood function,
- $\hat{\boldsymbol{\mu}}$ denote the ML estimate of $\boldsymbol{\mu}$.
- $L(\hat{\boldsymbol{\mu}}; \mathbf{y})$ denote the maximum value of the log likelihood.

For **all** possible models, the max achievable log likelihood is $L(\mathbf{y}; \mathbf{y})$.

The model having a parameter for each observation and the perfect fit with $\hat{\boldsymbol{\mu}} = \mathbf{y}$, is called *saturated* model（饱和模型）.

The saturated model does not provide data reduction, but can serve as a baseline for comparison with other model fits.

4.1.5 Deviance

The *deviance* of a Poisson or binomial GLM is defined to be

$$-2[L(\hat{\mu}; \mathbf{y}) - L(\mathbf{y}; \mathbf{y})].$$

This is the likelihood-ratio statistic for testing the null hypothesis that the model holds against the saturated model; i.e.,

small P-value $\Rightarrow$ for the saturated model (the reduced model is inadequate);

large P-value $\Rightarrow$ against the saturated model (the reduced model is adequate).

The deviance has a limiting χ^2_{df} with $df = N - p$.

In this book deviance is used for model checking and inferential comparisons of models.

4.1.5 Deviance

Example: Let Y_i be Bin(n_i, π_i), $i = 1, \ldots, N$.

Consider the simple model of homogeneity, $\pi_i = \pi$ for all i.
$\Rightarrow p = 1$ parameter.

The saturated model makes no assumption about $\{\pi_i\}$ and thus has N parameters.

The deviance for the homogeneity model has df$= N - 1$. In fact, it equals the G^2 likelihood-ratio statistic for testing independence in the $N \times 2$ table.

Under independence, it has approximately a chi-squared distribution as the $\{n_i\}$ increase, for fixed N.

Outline

4.2 Generalized Linear Models for Binary Data

Let Y denote a binary response variable with outcome 0 or 1.

Let $\pi(\mathbf{x}) = P(Y = 1|\mathbf{x})$, reflecting the dependence on values $\mathbf{x} = (x_1, \ldots, x_p)$ of predictors.

Then

$$E(Y) = \pi(\mathbf{x}), \quad \text{var}(\mathbf{Y}) = \pi(\mathbf{x})\,[\mathbf{1} - \pi(\mathbf{x})].$$

For simplicity, let $p = 1$ in the following subsections.

4.2.1 Linear probability model

The linear probability model:

$$\pi(x) = \alpha + \beta x.$$

With independent observations it is a GLM with *binomial* random component and *identical link* function.

As a probability, $\pi(x) \in [0, 1]$. But for large or small x values, the model above can yield $\pi(x) < 0$ and $\pi(x) > 1$. This is the major structural defect of the linear probability model.

This model can be valid over a restricted range of x values. When it is plausible, an advantage is its simple interpretation: β is the change in $\pi(x)$ for a one-unit increase in x.

$\text{var}(Y) = \pi(x)[1 - \pi(x)]$ depends on x through $\pi(x)$.

4.2.2 Snoring and heart disease example

Table 4.2 Relation between Snoring and Heart Disease

	Heart Disease		Proportion	Linear	Logit
Snoring	Yes	No	Yes	Fit[a]	Fit[a]
Never	24	1355	0.017	0.017	0.021
Occasionally	35	603	0.055	0.057	0.044
Nearly every night	21	192	0.099	0.096	0.093
Every night	30	224	0.118	0.116	0.132

[a] Model fits refer to proportion of yes responses.

A survey of 2484 subjects to investigate snoring (4 ordinal categories) as a risk factor for heart disease (binary).

The rows (snoring) are treated as independent binomial samples.

The scores (0, 2, 4, 5) are used for the snoring categories.

4.2.2 Snoring and heart disease example

Using SAS PROC GENMOD, the ML estimates are $\hat{\alpha} = 0.0172$ and $\hat{\beta} = 0.0198$ with SE$= 0.0028$ for $\hat{\beta}$.

- For nonsnorers ($x = 0$), the estimated probability of heart disease is $\hat{\alpha} = 0.0172 \approx 0.02$.
- For occasional snorers ($x = 2$), it increases $(2 - 0) \times \hat{\beta} = 2 \times 0.0198 \approx 0.04$.
- For those who snore nearly every night ($x = 4$), it increases further $(4 - 2) \times \hat{\beta} \approx 0.04$.
- For those who always snore ($x = 5$), it increases further $(5 - 4) \times \hat{\beta} \approx 0.02$.

Table 4.2 and Figure 4.1 show the model fits well.

4.2.3 Logistic regression model

In practice, $\pi(x)$ may be nonlinear.

Example.

Let x denote annual family income and $\pi(x)$ denote the probability of buying a new car, instead of a used one.

An increase of \$50,000 in annual income would have less effect when $x = \$1,000,000$ than when $x = \$50,000$.

The S-shaped curves in Figure 4.2 (see below) are typical.

4.2.3 Logistic regression model

The most important S-shaped curve is logistic regression model:

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}.$$

As $x \to \infty$, $\pi(x) \downarrow 0$ when $\beta < 0$ and $\pi(x) \uparrow 1$ when $\beta > 0$.

The odds of the above $\pi(x)$ are

$$\frac{\pi(x)}{1 - \pi(x)} = \exp(\alpha + \beta x).$$

The log odds have the following linear relationship

4.2.3 Logistic regression model

Logistic regression model（Logistic回归模型）is also called logit model, since $\text{logit}(u) = log(u/(1-u))$.

$\pi(x)$ must fall in the $(0, 1)$ range for any x.
$\Rightarrow$ No structural problem as for the linear probability model.

For the snoring data in Table 4.2, the ML fit is

$$\text{logit}[\hat{\pi}(x)] = -3.87 + 0.40\,x.$$

$\hat{\beta} = 0.40 > 0$ reflects the increased incidence of heart disease at higher snoring levels.

4.2.3 Logistic regression model

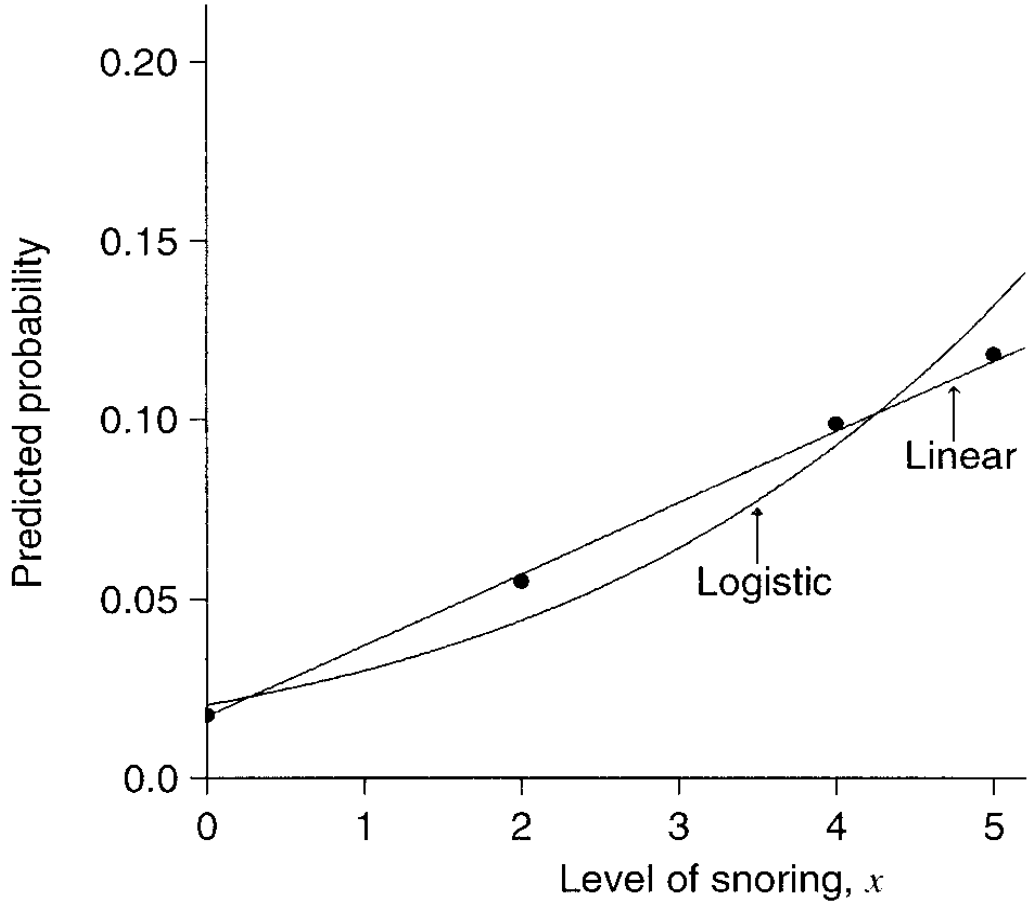


FIGURE 4.1 Predicted probabilities for linear probability and logistic regression models.

4.2.4 Binomial GLM for 2 $\times$ 2 contingency tables

A model for a binary response with a single binary explanatory variable X (taking values 0 and 1) is one of the simplest GLMs.

For a GLM with a given link function g, i.e., $g[\pi(x)] = \alpha + \beta x$, we have $g[\pi(0)] = \alpha$ and $g[\pi(1)] = \alpha + \beta$.

4.2.4 Binomial GLM for 2×2 contingency tables

Then the effect of X is described by $\beta = g[\pi(1)] - g[\pi(0)]$.

- For the **identical** (not canonical) link, $\beta = \pi(1) - \pi(0)$ is the **difference** between proportions.
- For the **log** (not canonical) link, $\beta = \log[\pi(1)] - \log[\pi(0)] = \log[\pi(1)/\pi(0)]$ is the **log relative risk**.
- For the **logit** (canonical) link, $\beta = \log \frac{\pi(1)}{1-\pi(1)} - \log \frac{\pi(0)}{1-\pi(0)} = \log \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}$ is the **log odds ratio**.

$\Rightarrow$ Measures of association for 2×2 tables can be obtained from the effect parameter β in GLMs for binary data.

4.2.5 Probit and inverse CDF link functions

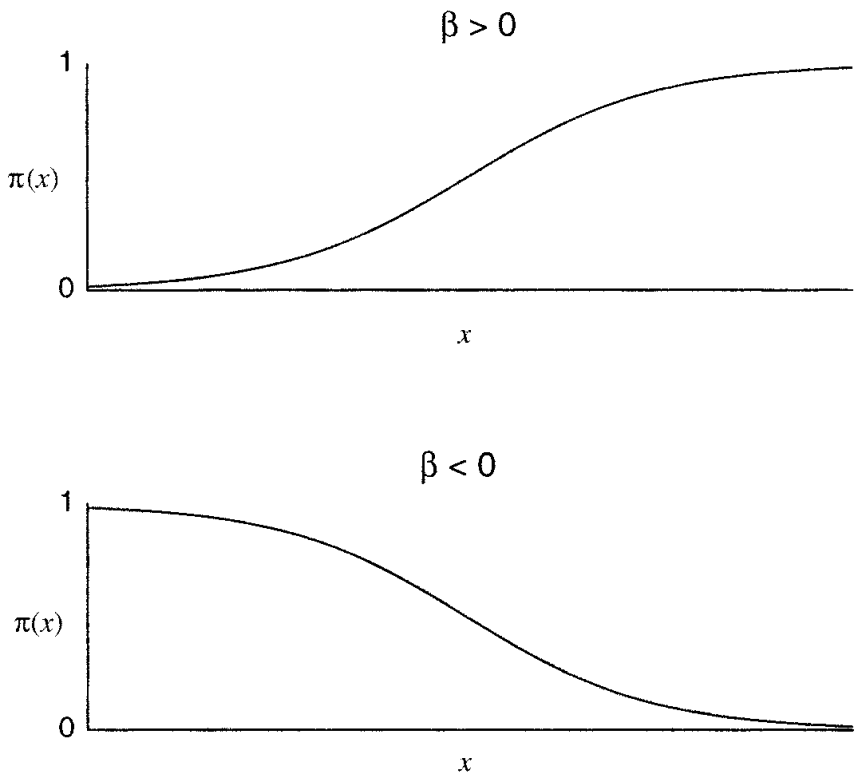


FIGURE 4.2 Logistic regression functions.

A monotone regression curve such as the top one in Figure 4.2 has the shape of a cumulative distribution function (cdf) for a continuous random variable. This suggests a model for a binary response having form $\pi(x) = F(\alpha + \beta x)$ for some cdf F.

:

4.2.5 Probit and inverse CDF link functions

Let $\Phi(\cdot)$ denote the standard cdf of the class, such as the $N(0,1)$ cdf. Writing the model as

$$\pi(x) = \Phi(\alpha + \beta x).$$

When Φ is strictly increasing over the entire real line, its inverse function $\Phi^{-1}(\cdot)$ exists and we can write

$$\Phi^{-1}[\pi(x)] = \alpha + \beta x.$$

For the GLM, the link function is Φ^{-1}.

4.2.5 Probit and inverse CDF link functions

Two common inverse CDF link functions: the inverse standard normal and the inverse standard logistic.

1) Probit model with inverse standard normal link

When $\Phi = \Phi_N =$ the standard normal cdf, the curve $\pi_N(x) = \Phi_N(\alpha + \beta x)$ has the shape of a normal cdf and

$$\Phi_N^{-1}[\pi_N(x)] = \alpha + \beta x$$

is called the *probit* model.

4.2.5 Probit and inverse CDF link functions

2) The logit model with inverse standard logistic link

The cdf of the logistic distribution with mean μ and dispersion parameter $\tau > 0$ is

$$F_L(x) = \frac{\exp[(x-\mu)/\tau]}{1+\exp[(x-\mu)/\tau]}, \quad -\infty < x < \infty.$$

The **standardi** logistic cdf ($\mu = 0$ and $\tau = 1$),

$$\Phi_L(x) = e^x/(1+e^x).$$

The logistic regression model: $\pi_L(x) = \Phi_L(\alpha + \beta x)$.
$\Rightarrow \Phi_L^{-1}[\pi_L(x)] = \alpha + \beta x$.
$\Rightarrow \Phi_L^{-1}[\pi_L(x)] = \log[\pi_L(x)/(1-\pi_L(x)]$.

4.2.5 Probit and inverse CDF link functions

TABLE A.3 SAS Code for Binary GLMs for Snoring Data in Table 4.2

```
data glm;
input snoring disease total @@;
datalines;
0 24 1379    2 35 638    4 21 213    5 30 254
;
proc genmod; model disease/total=snoring/dist=bin link=identity;
proc genmod; model disease/total=snoring/dist=bin link=logit;
proc genmod; model disease/total=snoring/dist=bin link=probit;
```

Outline

4.3.1 Poisson loglinear models

A Poisson loglinear GLM assumes a Poisson distribution for Y and uses the **log** link. For a model with explanatory variable X,

$$\log(\mu) = \alpha + \beta x \quad \Rightarrow \quad \mu = \exp(\alpha + \beta x) = e^{\alpha}(e^{\beta})^{x}.$$

A 1-unit increase in x has a **multiplicative** (乘积) impact of e^{β} on μ, i.e.,

$$\mu_{x+1} = e^{\alpha}(e^{\beta})^{x+1} = [e^{\alpha}(e^{\beta})^{x}](e^{\beta}) = \mu_x e^{\beta}.$$

4.3.2 Horseshoe crab mating example

:

鲎（hòu）、鲎鱼，又称马蹄蟹（horseshoe crab）、王蟹（king crab）。鲎鱼是一种古老的节肢动物，祖先就是生活在大约6亿年前的三叶虫；大约3亿年前鲎鱼出现了，后来三叶虫灭绝，可是鲎鱼却留下来。鲎鱼一直活到现在，保留了原始的蓝色血液。鲎鱼的血液细胞中的蛋白质里所含的却是铜元素，叫血蓝蛋白，它是呈蓝色的。现在，鲎鱼已被列入濒危物种。

4.3.2 Horseshoe crab mating example

4.3.2 Horseshoe crab mating example

TABLE 4.3 Number of Crab Satellites by Female's Characteristics[a]

C	S	W	Wt	Sa	C	S	W	Wt	Sa	C	S	W	Wt	Sa	C	S	W	Wt	Sa
2	3	28.3	3.05	8	3	3	22.5	1.55	0	1	1	26.0	2.30	9	3	3	24.8	2.10	0
3	3	26.0	2.60	4	2	3	23.8	2.10	0	3	2	24.7	1.90	0	2	1	23.7	1.95	0
3	3	25.6	2.15	0	3	3	24.3	2.15	0	2	3	25.8	2.65	0	2	3	28.2	3.05	11
4	2	21.0	1.85	0	2	1	26.0	2.30	14	1	1	27.1	2.95	8	2	3	25.2	2.00	1
2	3	29.0	3.00	1	4	3	24.7	2.20	0	2	3	27.4	2.70	5	2	2	23.2	1.95	4
1	2	25.0	2.30	3	2	1	22.5	1.60	1	3	3	26.7	2.60	2	4	3	25.8	2.00	3
4	3	26.2	1.30	0	2	3	28.7	3.15	3	2	1	26.8	2.70	5	4	3	27.5	2.60	0
2	3	24.9	2.10	0	1	1	29.3	3.20	4	1	3	25.8	2.60	0	2	2	25.7	2.00	0
2	1	25.7	2.00	8	2	1	26.7	2.70	5	4	3	23.7	1.85	0	2	3	26.8	2.65	0
2	3	27.5	3.15	6	4	3	23.4	1.90	0	2	3	27.9	2.80	6	3	3	27.5	3.10	3
1	1	26.1	2.80	5	1	1	27.7	2.50	6	2	1	30.0	3.30	5	3	1	28.5	3.25	9
3	3	28.9	2.80	4	2	3	28.2	2.60	6	2	3	25.0	2.10	4	2	3	28.5	3.00	3
2	1	30.3	3.60	3	4	3	24.7	2.10	5	2	3	27.7	2.90	5	1	1	27.4	2.70	6
2	3	22.9	1.60	4	2	1	25.7	2.00	5	2	3	28.3	3.00	15	2	3	27.2	2.70	3
3	3	26.2	2.30	3	2	1	27.8	2.75	0	4	3	25.5	2.25	0	3	3	27.1	2.55	0
3	3	24.5	2.05	5	3	1	27.0	2.45	3	2	3	26.0	2.15	5	2	3	28.0	2.80	1
2	3	30.0	3.05	8	2	3	29.0	3.20	10	2	3	26.2	2.40	0	2	1	26.5	1.30	0
2	3	26.2	2.40	3	3	3	25.6	2.80	7	3	3	23.0	1.65	1	3	3	23.0	1.80	0
2	3	25.4	2.25	6	3	3	24.2	1.90	0	2	2	22.9	1.60	0	3	2	26.0	2.20	3
2	3	25.4	2.25	4	3	3	25.7	1.20	0	2	3	25.1	2.10	5	3	2	24.5	2.25	0
4	3	27.5	2.90	0	3	3	23.1	1.65	0	3	1	25.9	2.55	4	2	3	25.8	2.30	0
4	3	27.0	2.25	3	2	3	28.5	3.05	0	4	1	25.5	2.75	0	4	3	23.5	1.90	0
2	2	24.0	1.70	0	2	1	29.7	3.85	5	2	1	26.8	2.55	0	4	3	26.7	2.45	0
2	1	28.7	3.20	0	3	3	23.1	1.55	0	2	1	29.0	2.80	1	3	3	25.5	2.25	0
3	3	26.5	1.97	1	3	3	24.5	2.20	1	3	3	28.5	3.00	1	2	3	28.2	2.87	1
2	3	24.5	1.60	1	2	3	27.5	2.55	1	2	2	24.7	2.55	4	2	1	25.2	2.00	1
3	3	27.3	2.90	1	2	3	26.3	2.40	1	2	3	29.0	3.10	1	2	3	25.3	1.90	2
2	3	26.5	2.30	4	2	3	27.8	3.25	3	2	3	27.0	2.50	6	3	3	25.7	2.10	0
2	3	25.0	2.10	2	2	3	31.9	3.33	2	4	3	23.7	1.80	0	4	3	29.3	3.23	12
3	3	22.0	1.40	0	2	3	25.0	2.40	5	3	3	27.0	2.50	6	3	3	23.8	1.80	6
1	1	30.2	3.28	2	3	3	26.2	2.22	0	2	3	24.2	1.65	2	2	3	27.4	2.90	3
2	2	25.4	2.30	0	3	3	28.4	3.20	3	4	3	22.5	1.47	4	2	3	26.2	2.02	2
2	1	24.9	2.30	6	1	2	24.5	1.95	6	2	3	25.1	1.80	0	2	1	28.0	2.90	4
4	3	25.8	2.25	10	2	3	27.9	3.05	7	2	3	24.9	2.20	0	2	1	28.4	3.10	5
3	3	27.2	2.40	5	2	2	25.0	2.25	6	2	3	27.5	2.63	6	2	1	33.5	5.20	7
2	3	30.5	3.32	3	3	3	29.0	2.92	3	2	1	24.3	2.00	0	2	3	25.8	2.40	0
4	3	25.0	2.10	8	2	1	31.7	3.73	4	2	3	29.5	3.02	4	3	3	24.0	1.90	10
2	3	30.0	3.00	9	2	3	27.6	2.85	4	2	3	26.2	2.30	0	2	1	23.1	2.00	0
2	1	22.9	1.60	0	4	3	24.5	1.90	0	2	3	24.7	1.95	4	2	3	28.3	3.20	0
2	3	23.9	1.85	2	3	3	23.8	1.80	0	3	2	29.8	3.50	4	2	3	26.5	2.35	4
2	3	26.0	2.28	3	2	3	28.2	3.05	8	4	3	25.7	2.15	0	2	3	26.5	2.75	7
2	3	25.8	2.20	0	3	3	24.1	1.80	0	3	3	26.2	2.17	2	3	3	26.1	2.75	3
3	3	29.0	3.28	4	1	1	28.0	2.62	0	4	3	27.0	2.63	0	2	2	24.5	2.00	0
1	1	26.5	2.35	0															

[a]C, color (1, light medium; 2, medium; 3, dark medium; 4, dark); S, spine condition (1, both good; 2, one worn or broken; 3, both worn or broken); W, carapace width (cm); Wt, weight (kg); Sa, number of satellites.

4.3.2 Horseshoe crab mating example

目的：研究影响雌蟹栖息地附近雄蟹（追随者）数量的因素。

Outcome: Number of satellites of a female horseshoe crab (Y)
Predictors: Female crab's color (C), spine condition (S), weight (Wt) and carapace width (W).

First only use carapace width (W) as a predictor. The mean is 26.3 cm and SE is 2.1 cm.

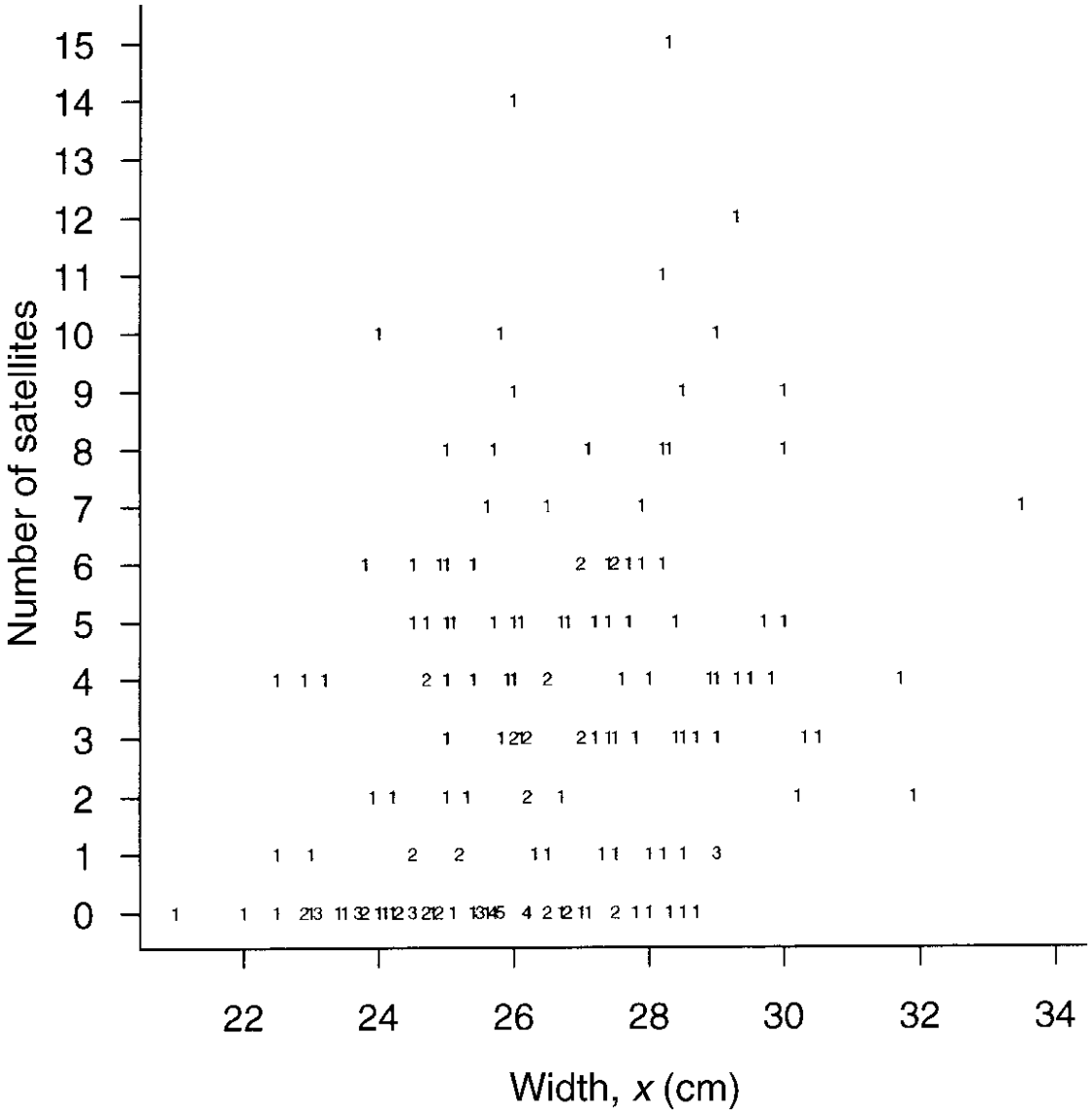
Figure 4.3: No clear trend between W and Y.

Figure 4.4:

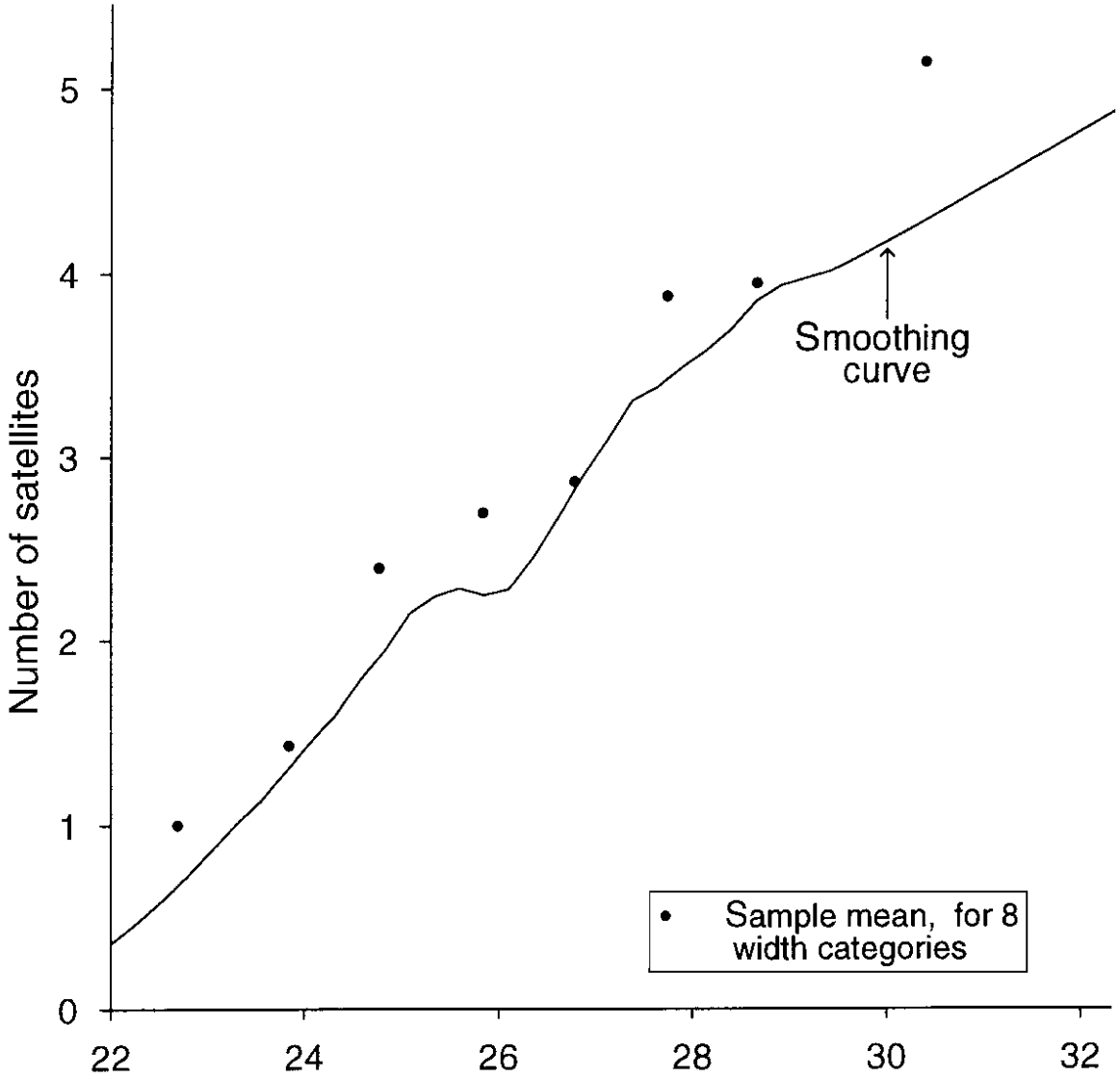
1) Mean numbers of satellites in 8 width categories (Table 4.4).
2) A smoothed curve based on an extension of the GLM in Section 4.8.

$\Rightarrow$ Both show a strong, approximately linear, increasing trend.

4.3.2 Horseshoe crab mating example



4.3.2 Horseshoe crab mating example



4.3.2 Horseshoe crab mating example

Table 4.4 Sample Mean and Variance of Number of Satellites

Width(cm)	Number of Cases	Number of Satellites	Sample Mean	Sample Variance
$<$23.25	14	14	1.00	2.77
23.25–24.25	14	20	1.43	8.88
24.25–25.25	28	67	2.39	6.54
25.25–26.25	39	105	2.69	11.38
26.25–27.25	22	63	2.86	6.88
27.25–28.25	24	93	3.87	8.81
28.25–29.25	18	71	3.94	16.88
$>$29.25	14	72	5.14	8.29

4.3.2 Horseshoe crab mating example

Let $\mu =$ expected number of satellites, and $x =$ width.

Log link

For the ungrouped data, the ML fit of the Poisson loglinear model is

$$\log(\hat{\mu}) = \hat{\alpha} + \hat{\beta}\, x = -3.305 + 0.164\, x.$$

The effect of width is $\hat{\beta} = 0.164 > 0$ with SE$= 0.020$.

$\exp(\hat{\beta}) = \exp(0.164) = 1.18$.
$\Rightarrow$ A 1-cm increase in width (x) yields an 18% increase in the estimated mean number of satellites ($\hat{\mu}$).

4.3.2 Horseshoe crab mating example

Identical link

For the ungrouped data, it has ML fit

$$\hat{\mu} = \hat{\alpha} + \hat{\beta}\, x = -11.53 + 0.55\, x.$$

- As 1-cm increase in x, $\hat{\mu}$ increases $\hat{\beta} = 0.55$.
- On the average, about a 2-cm increase in width is associated with an extra satellite.
- The fitted values are positive at all sampled x.

Figure 4.5: Width vs. fitted values $\hat{\mu}$ for the models with log link and identical link. $\Rightarrow$ The two models provide similar predictions over the width range in which most observations occur.

4.3.2 Horseshoe crab mating example

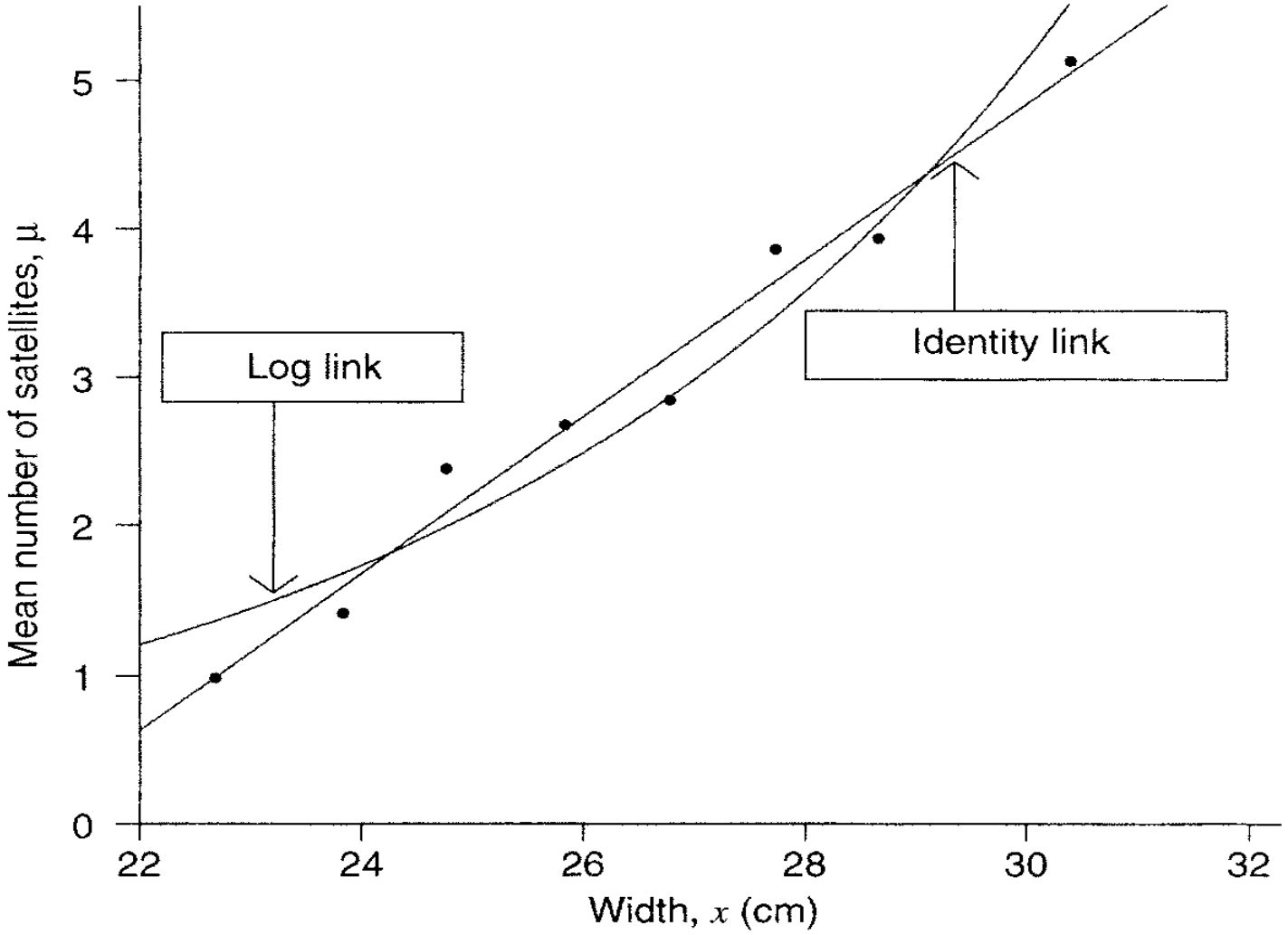


FIGURE 4.5 Estimated mean number of satellites for log and identity links.

4.3.3 Overdispersion for Poisson GLMs

For the grouped horseshoe crab data, Table 4.4 shows the sample mean and variance in each width category.
$\Rightarrow$ The variances are much larger than the means.
$\Rightarrow$ Overdispersion.

A common cause of overdispersion is subject heterogeneity.

For instance, suppose that width, weight, color and spine condition are predictors that affect the number of satellites of a female crab.

4.3.3 Overdispersion for Poisson GLMs

Our model uses only width as a predictor.
$\Rightarrow$ Crabs having a certain width are a mixture of crabs of various weights, colors and spine conditions.

Overdispersion is not an issue in ordinary regression with normally distributed Y, because the variance is described by a separate parameter independent of the mean.

However, for binomial and Poisson distributions, the variance is a function of the mean.

4.3.4 Negative binomial GLMs

Negative binomial distribution（负二项分布）: for $y = 0, 1, 2, ...$

$$f(y; k, \mu) = \frac{\Gamma(y+k)}{\Gamma(k)\,\Gamma(y+1)}\left(\frac{k}{\mu+k}\right)^k\left(1-\frac{k}{\mu+k}\right)^y,$$

where k and μ are parameters. It has

$$E(Y) = \mu, \quad \text{var}(Y) = \mu + \mu^2/k.$$

$f(y; k, \mu)$ can be expressed in natural exponential family form.
The index k^{-1} is called a *dispersion parameter*.

As $k^{-1} \to 0$, $\text{var}(Y) \to \mu$.
$\Rightarrow$ The negative binomial distribution converges to the Poisson.

4.3.4 Negative binomial GLMs

For the ungrouped crab data with identical link, $\hat{\mu} = \hat{\alpha} + \hat{\beta}\, x$.

Results of two GLMs:

Random component	$\hat{\alpha}$	$\hat{\beta}$	SE($\hat{\beta}$)	$\hat{k}^{-1}$	v̂ar(Y)
Poisson	-11.53	0.55	0.06	—	$\hat{\mu}$
Negative binomial	-11.15	0.53	0.11	$0.98 \approx 1$	$\approx \hat{\mu} + \hat{\mu}^2$

The fitted values $\hat{\mu}$ are similar (because $\hat{\alpha}$ and $\hat{\beta}$ are similar).

The negative binomial model

1) has greater SE($\hat{\beta}$) and v̂ar(Y) than the Poisson model;
2) reflects the overdispersion uncaptured by the Poisson model.

4.3.6 Poisson GLM of independence in $I \times J$ contingency tables

Consider independent counts $\{Y_{ij}\}$ with $Y_{ij} \sim$ Poisson(μ_{ij}). Suppose

$$\mu_{ij} = \mu\,\alpha_i\,\beta_j, \quad \text{(a multiplicative model)}$$

where $\alpha_i > 0$, $\beta_j > 0$ and $\sum_i \alpha_i = \sum_j \beta_j = 1$.

Using **log link**, then

$$\log(\mu_{ij}) = \log(\mu) + \log(\alpha_i) + \log(\beta_j) = \lambda + \alpha_i^* + \beta_j^*.$$

$\Rightarrow$ This Poisson loglinear model has **additive** main effects of the two classifications but no interaction.

By independence, $n = \sum_i \sum_j Y_{ij} \sim$ Poisson(μ) with $\mu = \sum_i \sum_j \mu_{ij}$.

4.3.6 Poisson GLM of independence in $I \times J$ contingency tables

Conditional on n,

- the cell counts $\{Y_{ij}\}$ have a multinomial distribution with $\{\pi_{ij} = \mu_{ij}/\mu = \alpha_i \beta_j\}$;
- the row totals $\{Y_{i+}\}$ have a multinomial distribution with $\{\pi_{i+} = \sum_j \pi_{ij} = \sum_j \alpha_i \beta_j = \alpha_i \sum_j \beta_j = \alpha_i\}$;
- the column totals $\{Y_{+j}\}$ have a multinomial distribution with $\{\pi_{+j} = \sum_i \pi_{ij} = \sum_i \alpha_i \beta_j = \beta_j \sum_i \alpha_i = \beta_j\}$.

$\Rightarrow$ Conditional on n, the model is a multinomial one that satisfies $\pi_{ij} = \alpha_i \beta_j = \pi_{i+} \pi_{+j}$, i.e., the two classifications are independent.

$\Rightarrow$ In Poisson form, independence is the loglinear model above.

4.3.6 Poisson GLM of independence in $I \times J$ contingency tables

TABLE A.4 SAS Code for Poisson and Negative Binomial GLMs for Horseshoe Crab Data in Table 4.3

```
data crab;
input color spine width satell weight;
datalines;
3  3  28.3  8  3.05
4  3  22.5  0  1.55
...
3  2  24.5  0  2.00
;
proc genmod;
  model satell=width/dist=poi link=log;
proc genmod;
  model satell=width/dist=poi link=identity;
proc genmod;
  model satell=width/dist=negbin link=identity;
```

Outline

4.4 Moments and likelihood for generalized linear models

Assume y_i has the density or mass function

This is called the *exponential dispersion family*(带离散参数的指数型分布族).

ϕ is called the *dispersion parameter* and θ_i^* the *natural parameter*. When ϕ is known, f simplifies to the form in Section 4.1.1 for natural exponential family:

$$f(y_i;\, \theta_i^*, \phi) = \{\exp[y_i\, \theta_i^*/a^*(\phi)]\}\, \{\exp[-b^*(\theta_i^*)/a^*(\phi)]\}\, \{\exp[c(y_i, \phi)]\}$$
$$= \exp[y_i\, Q(\theta_i)]\; a(\theta_i)\; b(y_i).$$

4.4 Moments and likelihood for generalized linear models

Note: in the book, no a^*, b^* and θ_i^*. But we prefer it to make the difference.

The more general formula (exponential dispersion family) is not needed for one-parameter families such as the binomial and Poisson.

Usually, $a^*(\phi) = \phi/w_i$ for a known weight w_i.

When y_i is a mean of n_i independent readings, such as a sample proportion for n_i Bernoulli trials, $w_i = n_i$ (Section 4.4.2).

4.4.1 Mean and variance functions

The contribution of y_i to the log likelihood is

$$L_i = \log f(y_i; \theta_i^*, \phi) = [y_i \theta_i^* - b^*(\theta_i^*)]/a^*(\phi) + c(y_i, \phi).$$

Therefore,

$$\begin{aligned} \partial L_i/\partial \theta_i^* &= [y_i - b^{*'}(\theta_i^*)]/a^*(\phi), \\ \partial^2 L_i/\partial \theta_i^{*2} &= -b^{*''}(\theta_i^*)/a^*(\phi), \end{aligned}$$

where $b^{*'}$ and $b^{*''}$ denote the first two derivatives of $b^*(\cdot)$ evaluated at θ_i^*.

The total log-likelihood is $L = \sum_i L_i$.

4.4.1 Mean and variance functions

Under regularity conditions satisfied by the exponential family, the general likelihood results:

Applied these results to a single observation, we obtain

$$E[(Y_i - b^{*'}(\theta_i^*))/a^*(\phi)] = E[Y_i - b^{*'}(\theta_i^*)]/a^*(\phi) = 0,$$
$$\Rightarrow \mu_i = E(Y_i) = b^{*'}(\theta_i^*);$$

$$\begin{aligned} b^{*''}(\theta_i^*)/a^*(\phi) &= E\{[(Y_i - b^{*'}(\theta_i^*))/a^*(\phi)]^2\} \\ &= E\{(Y_i - b^{*'}(\theta_i^*))^2\}/[a^*(\phi)]^2 \\ &= E\{(Y_i - \mu_i)^2\}/[a^*(\phi)]^2 = \text{var}(Y_i)/[a^*(\phi)]^2, \end{aligned}$$
$$\Rightarrow \text{var}(Y_i) = b^{*''}(\theta_i^*)\, a^*(\phi).$$

In summary, the function $b^*(\cdot)$ determines moments of Y_i.

4.4.1 Mean and variance functions

Explain $E\left(\frac{\partial L}{\partial \theta}\right) = 0$ and $-E\left(\frac{\partial^2 L}{\partial \theta^2}\right) = E\left\{\left(\frac{\partial L}{\partial \theta}\right)^2\right\}$.

4.4.2 Mean and variance functions for Poisson and binomial

When Y_i is Poisson

$$\begin{aligned}
f(y_i;\mu_i) &= (e^{-\mu_i}\,\mu_i^{y_i})/y_i! \\
&= \{\exp(-\mu_i)\}\{\exp[\log(\mu_i^{y_i})]\}\{\exp[\log(y!^{-1})]\} \\
&= \exp[-\mu_i + \log(\mu_i^{y_i}) + \log(y!^{-1})] \\
&= \exp[-\mu_i + y_i\,\log(\mu_i) - \log(y!)] \\
&= \exp\{-\exp[\log(\mu_i)] + y_i\,\log(\mu_i) - \log(y!)\} \\
&= \exp[-\exp(\theta_i^*) + y_i\,\theta_i^* - \log(y!)] \\
&= \exp[y_i\,\theta_i^* - \exp(\theta_i^*) - \log(y!)],
\end{aligned}$$

where $\theta_i^* = \log(\mu_i)$ is the natural parameter. (Note: In Section 4.1.3, $\theta_i = \mu_i$ and $Q(\theta_i) = \log(\mu_i) = \theta_i^*$ is the natural parameter.)

4.4.2 Mean and variance functions for Poisson and binomial

This has exponential dispersion form with $b^*(\theta_i^*) = \exp(\theta_i^*)$, $a^*(\phi) = 1$ and $c(y_i, \phi) = -\log(y_i!)$.

Following the relationships between the function $b^*(\cdot)$ and moments of Y_i, we have

$$\begin{aligned} E(Y_i) &= b^{*'}(\theta_i^*) = \exp(\theta_i^*) = \mu_i, \\ \text{var}(Y_i) &= b^{*''}(\theta_i^*)\, a^*(\phi) = \exp(\theta_i^*) = \mu_i. \end{aligned}$$

4.4.2 Mean and variance functions for Poisson and binomial

When $n_i Y_i$ has a Bin(n_i, π_i) distribution
i.e., here y_i is the sample *proportion* (rather than *number*) of successes, so $E(Y_i)$ is independent of n_i.

Let $\theta_i^* = \log[\pi_i/(1-\pi_i)] = \log(\pi_i) - \log(1-\pi_i)$, i.e., logit, then

$$\begin{aligned} \pi_i &= \exp(\theta_i^*)/[1+\exp(\theta_i^*)], \\ \log(1-\pi_i) &= \log\{1 - \frac{\exp(\theta_i^*)}{1+\exp(\theta_i^*)}\} = -\log[1+\exp(\theta_i^*)], \\ \log(\pi_i) &= \theta_i^* + \log(1-\pi_i). \end{aligned}$$

The binomial density can be expressed as

4.4.2 Mean and variance functions for Poisson and binomial

$$
\begin{aligned}
f(y_i;\,\pi_i, n_i) &= P(n_i Y_i = n_i y_i) = \binom{n_i}{n_i y_i} \pi_i^{n_i y_i} (1-\pi_i)^{n_i - n_i y_i} \\
&= \exp\{\log \binom{n_i}{n_i y_i} + \log(\pi_i^{n_i y_i}) + \log[(1-\pi_i)^{n_i - n_i y_i}]\} \\
&= \exp\{\log \binom{n_i}{n_i y_i} + n_i y_i \log(\pi_i) + (n_i - n_i y_i)\log(1-\pi_i)\} \\
&= \exp\{\log \binom{n_i}{n_i y_i} + n_i y_i\,[\theta_i^* + \log(1-\pi_i)] \\
&\quad + (n_i - n_i y_i)\log(1-\pi_i)\} \\
&= \exp\{\log \binom{n_i}{n_i y_i} + n_i y_i\,\theta_i^* + n_i \log(1-\pi_i)\} \\
&= \exp\{\log \binom{n_i}{n_i y_i} + n_i y_i\,\theta_i^* - n_i \log[1+\exp(\theta_i^*)]\} \\
&= \exp\left\{\log \binom{n_i}{n_i y_i} + \frac{y_i\,\theta_i^* - \log[1+\exp(\theta_i^*)]}{1/n_i}\right\}.
\end{aligned}
$$

4.4.2 Mean and variance functions for Poisson and binomial

This has exponential dispersion form with $b^*(\theta_i^*) = \log[1 + \exp(\theta_i^*)]$, $a^*(\phi) = 1/n_i$ and $c(y_i, \phi) = \log \binom{n_i}{n_i y_i}$.

The natural parameter is $\theta_i^* = \log[\pi_i/(1 - \pi_i)]$.

The moments of Y_i are

$$
\begin{aligned}
E(Y_i) &= b^{*'}(\theta_i^*) = \exp(\theta_i^*)/[1 + \exp(\theta_i^*)] = \pi_i, \\
\text{var}(Y_i) &= b^{*''}(\theta_i^*)\, a^*(\phi) = \exp(\theta_i^*)/\{[1 + \exp(\theta_i^*)]^2\, n_i\} \\
&= \left[\frac{\exp(\theta_i^*)}{1 + \exp(\theta_i^*)}\right]\left[1 - \frac{\exp(\theta_i^*)}{1 + \exp(\theta_i^*)}\right]/n_i = \pi_i(1 - \pi_i)/n_i.
\end{aligned}
$$

4.4.3 Systematic component and link function

Let $(x_{i1}, \ldots, x_{ip})$ be explanatory variables for observation i.

The systematic component of a GLM: $\eta_i = \sum_j \beta_j x_{ij}$, $i = 1, \ldots, N$.

In matrix form: $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$,

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_N)'$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ are column vectors of model parameters, and $\mathbf{X}$ is the $N \times p$ matrix of values of the explanatory variables for the N subjects.

In ordinary linear models, $\mathbf{X}$ is called the *design matrix*. In GLMs, $\mathbf{X}$ is called the *model matrix*.

The GLM links function g: $\eta_i = g(\mu_i) = \sum_j \beta_j x_{ij}, \quad i = 1, \ldots, N$.

4.4.3 Systematic component and link function

The link function for which $g(\mu_i) = \theta_i^*$ (the natural parameter) is the *canonical link*. Then $\theta_i^* = \sum_j \beta_j x_{ij}$.

Since $\mu_i = b^{*'}(\theta_i^*)$, the natural parameter is a function of the mean, i.e., $\theta_i^* = (b^{*'})^{-1}(\mu_i)$, where $(b^{*'})^{-1}(\cdot)$ denotes the inverse function of $b^{*'}(\cdot)$.
$\Rightarrow$ The canonical link is the inverse of $b^{*'}(\cdot)$.

In Poisson case, the canonical link is the log link.

Overall remark: From the $b^*(\cdot)$ in the exponential dispersion form, we can derive

- the mean and variance of Y,
- the canonical link.

4.4.4 Likelihood equations for a GLM

For N independent observations, the log likelihood is

$$L(\beta) = \sum_i L_i = \sum_i \log f(y_i; \theta_i^*, \phi) = \sum_i \frac{y_i \theta_i^* - b^*(\theta_i^*)}{a^*(\phi)} + \sum_i c(y_i, \phi).$$

The likelihood equations are

$$\partial L(\beta)/\partial \beta_j = \sum_i \partial L_i/\partial \beta_j = 0 \quad \text{for all } j.$$

To differentiate the log likelihood, we use the chain rule:

4.4.4 Likelihood equations for a GLM

For N independent observations, the log likelihood is

$$L(\beta) = \sum_i L_i = \sum_i \log f(y_i; \theta_i^*, \phi) = \sum_i \frac{y_i \theta_i^* - b^*(\theta_i^*)}{a^*(\phi)} + \sum_i c(y_i, \phi).$$

The likelihood equations are

$$\partial L(\beta)/\partial \beta_j = \sum_i \partial L_i/\partial \beta_j = 0 \quad \text{for all } j.$$

To differentiate the log likelihood, we use the chain rule:

4.4.4 Likelihood equations for a GLM

Since

$$\frac{\partial L_i}{\partial \theta_i^*} = \frac{y_i - b^{*'}(\theta_i^*)}{a^*(\phi)} = \frac{y_i - \mu_i}{a^*(\phi)} \quad \text{because } \mu_i = b^{*'}(\theta_i^*),$$

$$\frac{\partial \theta_i^*}{\partial \mu_i} = \left[\frac{\partial \mu_i}{\partial \theta_i^*}\right]^{-1} = [b^{*''}(\theta_i^*)]^{-1} = \frac{a^*(\phi)}{\text{var}(Y_i)}$$

$$\text{because } \text{var}(Y_i) = b^{*''}(\theta_i^*)\, a^*(\phi),$$

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial g(\mu_i)} \quad \text{(depending on the link function)},$$

$$\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \sum_j \beta_j x_{ij}}{\partial \beta_j} = x_{ij},$$

we obtain

$$\frac{\partial L_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a^*(\phi)} \frac{a^*(\phi)}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} = \frac{(y_i - \mu_i)\, x_{ij}}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}.$$

4.4.4 Likelihood equations for a GLM

So the likelihood equations are

The β are incorporated in $\mu_i = g^{-1}(\sum_j \beta_j x_{ij})$.

The likelihood equations depend on the distribution of Y_i only through μ_i and var$(Y_i) = \nu(\mu_i)$ for some function ν:

- Poisson: $\nu(\mu_i) = \mu_i$.
- Bernoulli: $\nu(\mu_i) = \mu_i(1 - \mu_i)$.
- Binomial: $\nu(\mu_i) = \mu_i(1 - \mu_i)/n_i$.
- Normal: $\nu(\mu_i) = \sigma^2$ (i.e., a constant, independent of μ_i).

$\Rightarrow$ When Y_i is in the natural exponential family, the relationship between μ_i and var(Y_i) characterizes the distribution.

4.4.5 Likelihood equations for binomial GLMs

Suppose $n_i Y_i \sim \text{Bin}(n_i, \pi_i)$.

For a single predictor, $\pi_i = \Phi(\alpha + \beta x_i)$, where Φ is a cdf.

For several predictors $(\mathbf{x}_j, \ldots, \mathbf{x}_p)$, the GLM is

$$\pi_i = \Phi\Big(\sum_j \beta_j x_{ij}\Big) = \Phi(\eta_i).$$

Since $\pi_i = \mu_i$, we have

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \pi_i}{\partial \eta_i} = \frac{\partial \Phi(\eta_i)}{\partial \eta_i} = \phi(\eta_i) = \phi\Big(\sum_j \beta_j x_{ij}\Big),$$

where $\phi(u) = \partial \Phi(u)/\partial u$ is the density function.

4.4.5 Likelihood equations for binomial GLMs

Since $\text{var}(Y_i) = \pi_i(1-\pi_i)/n_i$, the likelihood equations simplify to

$$\sum_i \frac{n_i\,(y_i - \pi_i)\,x_{ij}}{\pi_i(1-\pi_i)}\,\phi\Big(\sum_j \beta_j\,x_{ij}\Big)$$

$$= \sum_i \frac{n_i\,[y_i - \Phi(\sum_j \beta_j\,x_{ij})]\,x_{ij}}{\Phi(\sum_j \beta_j\,x_{ij})[1-\Phi(\sum_j \beta_j\,x_{ij})]}\,\phi\Big(\sum_j \beta_j\,x_{ij}\Big) = 0.$$

Let $\Phi(x) = e^t/(1+e^t)$. Then $\phi(t) = \Phi(t)(1-\Phi(t))$, and the link function is logit link, i.e. $\eta_i = \log[\pi_i/(1-\pi_i)]$.

$\Rightarrow \partial\eta_i/\partial\pi_i = 1/[\pi_i(1-\pi_i)]$

$\Rightarrow \partial\mu_i/\partial\eta_i = \partial\pi_i/\partial\eta_i = (\partial\eta_i/\partial\pi_i)^{-1} = \pi_i(1-\pi_i) \quad (= \phi(\eta_i))$.

Then the likelihood equations simplify to

$$\sum_i n_i\,(y_i - \pi_i)\,x_{ij} = \sum_i n_i\,[y_i - \Phi(\sum_j \beta_j\,x_{ij})]\,x_{ij} = 0.$$

4.4.6 Asymptotic covariance matrix of model parameter estimators

Recall that $\text{cov}(\hat{\boldsymbol{\beta}})$ is the inverse of the information matrix $\mathcal{J}$, which has elements $E[-\partial^2 L(\boldsymbol{\beta})/\partial\beta_h\,\partial\beta_j]$. Since

$$E\Big(\frac{\partial^2 L_i}{\partial\beta_h\,\partial\beta_j}\Big) = -E\Big[\Big(\frac{\partial L_i}{\partial\beta_h}\Big)\Big(\frac{\partial L_i}{\partial\beta_j}\Big)\Big],$$

for distributions in the exponential family we have

$$\begin{aligned} E\Big(\frac{\partial^2 L_i}{\partial\beta_h\,\partial\beta_j}\Big) &= -E\Big[\Big(\frac{(Y_i-\mu_i)\,x_{ih}}{\text{var}(Y_i)}\,\frac{\partial\mu_i}{\partial\eta_i}\Big)\Big(\frac{(Y_i-\mu_i)\,x_{ij}}{\text{var}(Y_i)}\,\frac{\partial\mu_i}{\partial\eta_i}\Big)\Big] \\ &= \frac{-x_{ih}\,x_{ij}}{\text{var}(Y_i)}\Big(\frac{\partial\mu_i}{\partial\eta_i}\Big)^2. \end{aligned}$$

Then, the (h,j)th element of the information matrix is

$$E\Big(-\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_h\,\partial\beta_j}\Big) = E\Big(-\frac{\partial^2\sum_i L_i}{\partial\beta_h\,\partial\beta_j}\Big) = \sum_{i=1}^{N} E\Big(-\frac{\partial^2 L_i}{\partial\beta_h\,\partial\beta_j}\Big) = \sum_{i=1}^{N}\frac{x_{ih}\,x_{ij}}{\text{var}(Y_i)}\Big(\frac{\partial\mu_i}{\partial\eta_i}\Big)^2.$$

4.4.6 Asymptotic covariance matrix of model parameter estimators

Explain

$$E\Big(\frac{\partial^2 L}{\partial\beta_h\,\partial\beta_j}\Big) = -E\Big[\Big(\frac{\partial L}{\partial\beta_h}\Big)\Big(\frac{\partial L}{\partial\beta_j}\Big)\Big].$$

4.4.6 Asymptotic covariance matrix of model parameter estimators

Hence, the information matrix has the form

$$\mathcal{J} = \mathbf{X}'\mathbf{W}\mathbf{X},$$

where $\mathbf{W}$ is the diagonal matrix with main-diagonal elements

$$w_i = (\partial\mu_i/\partial\eta_i)^2/\text{var}(Y_i).$$

The asymptotic covariance matrix of $\hat{\beta}$ is estimated by

where $\hat{\mathbf{W}}$ is $\mathbf{W}$ evaluated at $\hat{\beta}$. The $\mathbf{W}$ depends on the link function.

4.4.7 Likelihood equations and covariance matrix for Poisson loglinear model

The Poisson loglinear model (Section 4.1.3) has the form

$$\log \mu_i = \sum_j \beta_j x_{ij}, \quad i = 1, \ldots, N \quad \Rightarrow \log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}.$$

For the log link, $\eta_i = \log(\mu_i)$, $\Rightarrow \mu_i = \exp(\eta_i)$, $\partial\mu_i/\partial\eta_i = \mu_i$.

Since $\text{var}(Y_i) = \mu_i$, the likelihood equations simply to

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i)\, x_{ij}}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = \sum_{i=1}^{N} \frac{(y_i - \mu_i)\, x_{ij}}{\mu_i} \mu_i = \sum_{i=1}^{N} (y_i - \mu_i)\, x_{ij} = 0.$$

$\Rightarrow \sum_i y_i x_{ij} = \sum_i \mu_i x_{ij}$.

Since $w_i = (\partial\mu_i/\partial\eta_i)^2/\text{var}(Y_i) = \mu_i^2/\mu_i = \mu_i$, the $\hat{\mathbf{W}}$ in the $\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})$ has elements of $\hat{\boldsymbol{\mu}}$ on the main diagonal.

Outline

4.5.1 Deviance and goodness of fit

Let $\tilde{\theta}^*$ denote the estimate of θ^* for the saturated model, with estimated means $\tilde{\mu}_i = y_i$ for all i.

Let $\hat{\theta}^*$ denote the estimate of θ^* for the unsaturated model, with estimated means $\hat{\mu}_i$.

The lack of fit can be described by

$$\begin{aligned} & -2 \log \frac{\text{max. likelihood for the unsaturated model}}{\text{max. likelihood for the saturated model}} \\ = \quad & -2[L(\hat{\boldsymbol{\mu}}; \mathbf{y}) - L(\mathbf{y}; \mathbf{y})]. \end{aligned}$$

It is the likelihood-ratio statistic for testing the null hypothesis that the unsaturated model holds against the alternative that a more general model holds.

4.5.1 Deviance and goodness of fit

Following Section 4.4.4,

$$
\begin{aligned}
-2[L(\hat{\boldsymbol{\mu}};\mathbf{y}) - L(\mathbf{y};\mathbf{y})] &= 2[L(\mathbf{y};\mathbf{y}) - L(\hat{\boldsymbol{\mu}};\mathbf{y})] \\
&= 2\Big[\Big(\sum_i \frac{y_i\tilde{\theta}_i^* - b^*(\tilde{\theta}_i^*)}{a^*(\phi)} + \sum_i c(y_i,\phi)\Big) \\
&\quad -\Big(\sum_i \frac{y_i\hat{\theta}_i^* - b^*(\hat{\theta}_i^*)}{a^*(\phi)} + \sum_i c(y_i,\phi)\Big)\Big] \\
&= 2\sum_i [y_i\tilde{\theta}_i^* - b^*(\tilde{\theta}_i^*) - y_i\hat{\theta}_i^* + b^*(\hat{\theta}_i^*)]/a^*(\phi) \\
&= 2\sum_i [y_i(\tilde{\theta}_i^* - \hat{\theta}_i^*) - b^*(\tilde{\theta}_i^*) + b^*(\hat{\theta}_i^*)]/a^*(\phi).
\end{aligned}
$$

Usually $a^*(\phi) = \phi/w_i$, then the likelihood-ratio statistic equals

$$
2\sum_i w_i[y_i(\tilde{\theta}_i^* - \hat{\theta}_i^*) - b^*(\tilde{\theta}_i^*) + b^*(\hat{\theta}_i^*)]/\phi = D(\mathbf{y};\hat{\boldsymbol{\mu}})/\phi.
$$

4.5.2 Deviance for Poisson models

$D(\mathbf{y}; \hat{\boldsymbol{\mu}})/\phi$: *scaled deviance*; $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$: *deviance*. For some GLMs the scaled deviance has an approximate chi-square distribution.

For Poisson GLMs, $\theta_i^* = \log(\mu_i)$, $b^*(\theta_i^*) = \exp(\theta_i^*)$ and $a^*(\phi) = \phi/w_i = 1$ for all i (Section 4.4.2).

- For the unsaturated model, $\hat{\theta}_i^* = \log(\hat{\mu}_i)$ and $b^*(\hat{\theta}_i^*) = \exp(\hat{\theta}_i^*) = \hat{\mu}_i$;
- For the saturated model, $\tilde{\theta}_i^* = \log(\tilde{\mu}_i) = \log(y_i)$ and $b^*(\tilde{\theta}_i^*) = \exp(\tilde{\theta}_i^*) = \tilde{\mu}_i = y_i$;
- The deviance and the scaled deviance equal to

$$\begin{aligned} D(\mathbf{y}; \hat{\boldsymbol{\mu}}) &= 2\sum_i \{y_i[\log(y_i) - \log(\hat{\mu}_i)] - y_i + \hat{\mu}_i\} \\ &= 2\sum_i \{y_i \log(y_i/\hat{\mu}_i) - y_i + \hat{\mu}_i\}. \end{aligned}$$

4.5.2 Deviance for Poisson models

When a model with log link containing an intercept term, e.g., β_1 is the intercept with $x_{i1} = 1$ for all i, the likelihood equation (see Section 4.4.7) implied by that parameter (e.g., β_1) is

$$\sum_i (y_i - \mu_i)\, x_{i1} = \sum_i (y_i - \mu_i) = 0 \quad \Rightarrow \quad \sum_i y_i = \sum_i \hat{\mu}_i.$$

Then the deviance simplifies to $D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2\sum_i \{y_i \log(y_i/\hat{\mu}_i)\}$.

For two-way contingency tables, this reduces to the G^2 statistic in Section 3.2.1, substituting cell count n_{ij} for y_i and the independence fitted values $\hat{\mu}_{ij}$ for $\hat{\mu}_i$.

4.5.3 Deviance for binomial models: grouped and ungrouped data

Consider binomial GLMs with sample proportions $\{y_i\}$ based on $\{n_i\}$ trials. Following Section 4.4.2,

- unsaturated model:

$$\begin{aligned}
\hat{\theta}_i^* &= \log[\hat{\pi}_i/(1-\hat{\pi}_i)], \\
b^*(\hat{\theta}_i^*) &= \log[1+\exp(\hat{\theta}_i^*)] = \log[1+\exp\{\log[\hat{\pi}_i/(1-\hat{\pi}_i)]\}] \\
&= \log[1+\hat{\pi}_i/(1-\hat{\pi}_i)] = \log[(1-\hat{\pi}_i+\hat{\pi}_i)/(1-\hat{\pi}_i)] \\
&= \log[1/(1-\hat{\pi}_i)] = -\log(1-\hat{\pi}_i);
\end{aligned}$$

- saturated model:

$$\tilde{\theta}_i^* = \log[y_i/(1-y_i)] \quad \text{and} \quad b^*(\tilde{\theta}_i^*) = -\log(1-y_i).$$

Also, $a^*(\phi) = 1/n_i$, so $\phi = 1$ and $w_i = n_i$.

4.5.3 Deviance for binomial models: grouped and ungrouped data

The (scaled) deviance equals

$$
\begin{aligned}
& 2\sum_i n_i\left\{y_i\left[\log\left(\frac{y_i}{1-y_i}\right)-\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right)\right]+\log(1-y_i)-\log(1-\hat{\pi}_i)\right\} \\
= \;& 2\sum_i n_i\left\{y_i\log\left(\frac{y_i}{1-y_i}\right)-y_i\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right)+\log\left(\frac{1-y_i}{1-\hat{\pi}_i}\right)\right\} \\
= \;& 2\sum_i\left\{n_i y_i\log\left(\frac{n_i y_i}{n_i-n_i y_i}\right)-n_i y_i\log\left(\frac{n_i\hat{\pi}_i}{n_i-n_i\hat{\pi}_i}\right)\right. \\
& \left.+n_i\log\left(\frac{n_i-n_i y_i}{n_i-n_i\hat{\pi}_i}\right)\right\} \\
= \;& 2\sum_i\{n_i y_i\log(n_i y_i)-n_i y_i\log(n_i-n_i y_i)-n_i y_i\log(n_i\hat{\pi}_i) \\
& +n_i y_i\log(n_i-n_i\hat{\pi}_i)+n_i\log(n_i-n_i y_i)-n_i\log(n_i-n_i\hat{\pi}_i)\} \\
= \;& 2\sum_i n_i y_i\log\left(\frac{n_i y_i}{n_i\hat{\pi}_i}\right)+2\sum_i(n_i-n_i y_i)\log\left(\frac{n_i-n_i y_i}{n_i-n_i\hat{\pi}_i}\right).
\end{aligned}
$$

4.5.3 Deviance for binomial models: grouped and ungrouped data

At setting i, $n_i\, y_i$ is the number of successes and $(n_i - n_i\, y_i)$ is the number of failures.

Thus, the deviance is a sum over the $2N$ cells of successes and failures and has the same form

$$D(\mathbf{y}; \hat{\mu}) = 2 \sum \text{observed} \times \log(\text{observed/fitted}),$$

as the deviance for Poisson loglinear models with intercept term.

4.5.3 Deviance for binomial models: grouped and ungrouped data

With binomial responses, it is possible to construct the data file in two ways:

1) With counts of successes and failures at each setting for the predictors. $\Rightarrow$ *Grouped data*.

$\Rightarrow$ The saturated model has a parameter at each **setting** for the the predictors.

$\Rightarrow$ The approximate chi-square distribution for the deviance occu

2) With the individual Bernoulli 0-1 observations at the subject level. $\Rightarrow$ *Ungrouped data*.

$\Rightarrow$ The saturated model has a parameter for each **subject**.

$\Rightarrow$ The approximate chi-square distribution for the deviance does not occur.

4.5.4 Likelihood-ratio model comparison using the deviance

For a Poisson or binomial model M, $\phi = 1$; so

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = -2[L(\hat{\boldsymbol{\mu}}; \mathbf{y}) - L(\mathbf{y}; \mathbf{y})].$$

Consider two models: M_0 with fitted values $\hat{\boldsymbol{\mu}}_0$ and M_1 with fitted values $\hat{\boldsymbol{\mu}}_1$, with M_0 a special case of M_1.

Since M_0 is simpler than M_1, a smaller set of parameter values satisfies M_0 than satisfies M_1.
$\Rightarrow$ Maximizing the log likelihood over a smaller space cannot yield a larger maximum value.
$\Rightarrow L(\hat{\boldsymbol{\mu}}_0; \mathbf{y}) \leq L(\hat{\boldsymbol{\mu}}_1; \mathbf{y})$.
$\Rightarrow D(\mathbf{y}; \hat{\boldsymbol{\mu}}_0) \geq D(\mathbf{y}; \hat{\boldsymbol{\mu}}_1)$, i.e., simpler models have larger deviances.

4.5.4 Likelihood-ratio model comparison using the deviance

Assuming that model M_1 holds, the likelihood-ratio test of the hypothesis that M_0 holds uses the test statistic

$$\begin{aligned}
&-2[L(\hat{\boldsymbol{\mu}}_0;\mathbf{y}) - L(\hat{\boldsymbol{\mu}}_1;\mathbf{y})] \\
&= -2[L(\hat{\boldsymbol{\mu}}_0;\mathbf{y}) - L(\mathbf{y};\mathbf{y})] - \{-2[L(\hat{\boldsymbol{\mu}}_1;\mathbf{y}) - L(\mathbf{y};\mathbf{y})]\} \\
&= D(\mathbf{y};\hat{\boldsymbol{\mu}}_0) - D(\mathbf{y};\hat{\boldsymbol{\mu}}_1) \\
&= 2\sum_i w_i[y_i(\tilde{\theta}_i^* - \hat{\theta}_{0i}^*) - b^*(\tilde{\theta}_i^*) + b^*(\hat{\theta}_{0i}^*)] \\
&\quad -2\sum_i w_i[y_i(\tilde{\theta}_i^* - \hat{\theta}_{1i}^*) - b^*(\tilde{\theta}_i^*) + b^*(\hat{\theta}_{1i}^*)] \\
&= 2\sum_i w_i[y_i(\hat{\theta}_{1i}^* - \hat{\theta}_{0i}^*) - b^*(\hat{\theta}_{1i}^*) + b^*(\hat{\theta}_{0i}^*)].
\end{aligned}$$

4.5.5 Residuals for GLMs

First check overall goodness-of-fit of a GLM. If it fits poorly, then check residuals to find out where the fit is poor.

1) Deviance residual: $\sqrt{d_i} \times \text{sign}(y_i - \hat{\mu}_i)$,
where $d_i = 2\, w_i[y_i(\tilde{\theta}_i^* - \hat{\theta}_i^*) - b^*(\tilde{\theta}_i^*) + b^*(\hat{\theta}_i^*)]$. $\Rightarrow \sum_i d_i = D(\mathbf{y}; \hat{\boldsymbol{\mu}})$.
For two-way contingency tables, this is the same as the likelihood-ratio statistic for testing independence (Section 3.2.1):
$\Rightarrow \sum_i \sum_j (\text{deviance residual}_{ij})^2 = \sum_i \sum_j d_{ij} = G^2$.

2) Pearson residual: $e_i = (y_i - \hat{\mu}_i)/[\widehat{\text{var}}(Y_i)]^{1/2}$.
For a Poisson GLM, $\text{var}(Y_i) = \mu_i$, then $e_i = (y_i - \hat{\mu}_i)/\sqrt{\hat{\mu}_i}$.

For two-way contingency tables, this is the same as the Pearson residual defined in Section 3.3.1: $e_{ij} = (n_{ij} - \hat{\mu}_{ij})/\sqrt{\hat{\mu}_{ij}}$, with $\sum_i \sum_j e_{ij}^2 = X^2$, the Pearson X^2 statistic.

4.5.5 Residuals for GLMs

In linear models with the hat matrix **Hat**, the data are projected through **Hat** $\times$ **y** to the fitted values $\hat{\mu}$.

For GLMs, applying the estimated hat matrix to a linearized approximation for $g(\mathbf{y})$ yields $\hat{\boldsymbol{\eta}} = g(\hat{\boldsymbol{\mu}})$.

For GLMs the asymptotic covariance matrix of the vector of the raw residuals $\{y_i - \hat{\mu}_i\}$ is

$$\text{cov}(\mathbf{Y} - \hat{\boldsymbol{\mu}}) = \text{cov}(\mathbf{Y})[\mathbf{I} - \mathbf{Hat}];$$

where $\mathbf{I}$ is the identity matrix and

$$\mathbf{Hat} = \mathbf{W}^{1/2}\,\mathbf{X}\,(\mathbf{X}'\,\mathbf{W}\,\mathbf{X})^{-1}\,\mathbf{X}'\,\mathbf{W}^{1/2},$$

where $\mathbf{W}$ is as defined in Section 4.4.6.

4.5.5 Residuals for GLMs

Let $\hat{h}_i$ be the estimated i-th diagonal element of **Hat**. Then, the standardized Pearson residual is

For Poisson GLMs, $r_i = (y_i - \hat{\mu}_i)/\sqrt{\hat{\mu}_i(1-\hat{h}_i)}$.

As in ordinary regression, $h_i \in [0, 1]$ for all i and $\sum_i h_i = p$ (number of model parameters).

Outline

4.6.1 Newton-Raphson method

The likelihood equations are usually nonlinear in $\hat{\boldsymbol{\beta}}$.

The *Newton-Raphson method* is an iterative method for solving nonlinear equations.

Let

$$
\begin{aligned}
\mathbf{u}' &= (\partial L(\boldsymbol{\beta})/\partial\beta_1,\ \partial L(\boldsymbol{\beta})/\partial\beta_2,\ \ldots); \\
\mathbf{H} &= \{h_{ab}\} = \{\partial^2 L(\boldsymbol{\beta})/\partial\beta_a\,\partial\beta_b\}, \quad \text{i.e., the } \textit{Hessian matrix}; \\
\boldsymbol{\beta}^{(t)} &= \text{the guess for } \hat{\boldsymbol{\beta}} \text{ at iteration } t \quad (t = 0, 1, 2, \ldots); \\
\mathbf{u}^{(t)} &= \mathbf{u} \text{ evaluated at } \boldsymbol{\beta}^{(t)}; \\
\mathbf{H}^{(t)} &= \mathbf{H} \text{ evaluated at } \boldsymbol{\beta}^{(t)}.
\end{aligned}
$$

4.6.1 Newton-Raphson method

The Newton-Raphson method involves the following steps:

1) Give an initial guess ($\boldsymbol{\beta}^{(0)}$) for the solution.

2) For each cycle with $t = 0, 1, 2, \ldots$, approximate $L(\boldsymbol{\beta})$ near $\boldsymbol{\beta}^{(t)}$ by Taylor expansion,

$$L(\boldsymbol{\beta}) \approx L(\boldsymbol{\beta}^{(t)}) + \mathbf{u}^{(t)\prime}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) + (\frac{1}{2})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})' \mathbf{H}^{(t)} (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}).$$

3) Solve $\partial L(\boldsymbol{\beta})/\partial\boldsymbol{\beta} \approx \mathbf{u}^{(t)} + \mathbf{H}^{(t)} (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) = \mathbf{0}$ for $\boldsymbol{\beta}$ to obtain

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - (\mathbf{H}^{(t)})^{-1} \mathbf{u}^{(t)},$$

assuming that $\mathbf{H}^{(t)}$ is nonsingular.

4) Repeat 2) and 3) until changes in $L(\boldsymbol{\beta}^{(t)})$ between successive cycles are sufficiently small.

4.6.1 Newton-Raphson method

The convergence of $\boldsymbol{\beta}^{(t)}$ to $\hat{\boldsymbol{\beta}}$ is usually fast when the function is suitable and/or the initial guess is good.

For large t, the convergence satisfies, for each j,

$$|\beta_j^{(t+1)} - \hat{\beta}_j| \leq c|\beta_j^{(t)} - \hat{\beta}_j|^2 \quad \text{for some } c > 0$$

and is referred to as *second-order*.

$\Rightarrow$ The number of correct decimals in the approximation roughly doubles after sufficiently many iterations.

4.6.1 Newton-Raphson method

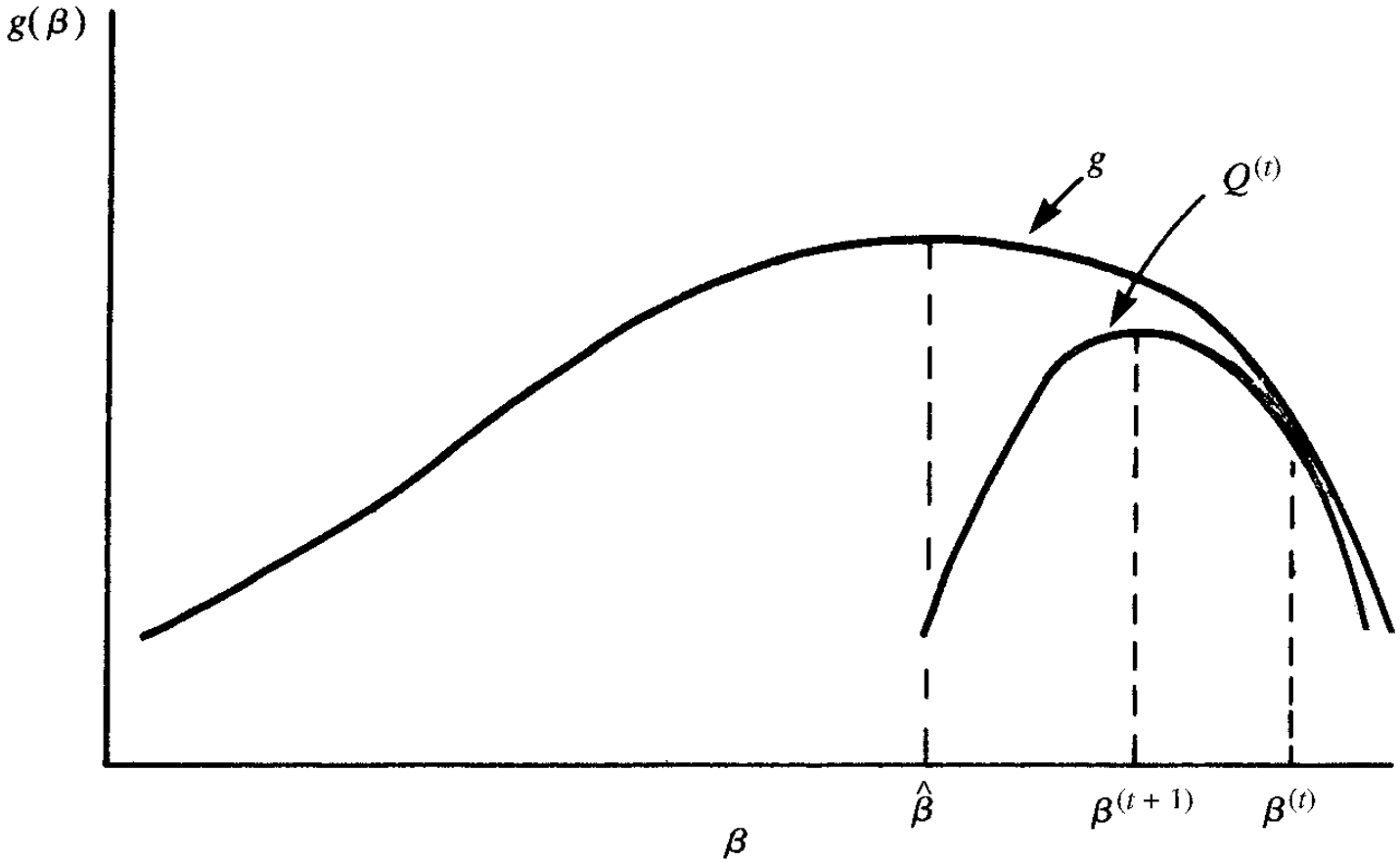


FIGURE 4.6 Cycle of Newton−Raphson method.

4.6.1 Newton-Raphson method

Illustration: One observation y from Bin(n, π). From Section 1.3.2, $L(\pi) = y\log(\pi) + (n-y)\log(1-\pi)$.

The first two derivatives of $L(\pi)$ are

$$\left(\frac{dL}{d\pi} = \right) u = \frac{y}{\pi} - \frac{n-y}{1-\pi} = \frac{y(1-\pi) - \pi(n-y)}{\pi(1-\pi)} = \frac{y - n\pi}{\pi(1-\pi)},$$

and

4.6.1 Newton-Raphson method

$$
\begin{aligned}
H &= \frac{du}{d\pi} = \frac{d(y-n\pi)/d\pi}{\pi(1-\pi)} - \frac{(y-n\pi)\,[d(\pi(1-\pi))/d\pi]}{[\pi(1-\pi)]^2} \\
&= \frac{-n}{\pi(1-\pi)} - \frac{(y-n\pi)(1-2\pi)}{[\pi(1-\pi)]^2} \\
&= -\frac{n\pi(1-\pi)+(y-n\pi)(1-2\pi)}{[\pi(1-\pi)]^2} \\
&= -\frac{n\pi - n\pi^2 + y - n\pi - 2y\pi + 2n\pi^2}{[\pi(1-\pi)]^2} = -\frac{y-2y\pi+n\pi^2}{[\pi(1-\pi)]^2} \\
&= -\frac{y-2y\pi+y\pi^2-y\pi^2+n\pi^2}{[\pi(1-\pi)]^2} = -\frac{y(1-\pi)^2+(n-y)\pi^2}{[\pi(1-\pi)]^2} \\
&= -\left[\frac{y}{\pi^2}+\frac{n-y}{(1-\pi)^2}\right].
\end{aligned}
$$

4.6.1 Newton-Raphson method

Each Newton-Raphson step has the form

$$\pi^{(t+1)} = \pi^{(t)} + \left[\frac{y}{(\pi^{(t)})^2} + \frac{n-y}{(1-\pi^{(t)})^2}\right]^{-1} \frac{y - n\pi^{(t)}}{\pi^{(t)}(1-\pi^{(t)})}.$$

The $\pi^{(t)}$ is adjusted

- upwards if $y - n\pi^{(t)} > 0$ or equivalently $y/n > \pi^{(t)}$;
- downwards if $y - n\pi^{(t)} < 0$ or equivalently $y/n < \pi^{(t)}$.

When $\pi^{(t)} = y/n$, no adjustment occurs and $\pi^{(t+1)} = \pi^{(t)} = y/n$, which is the correct answer for $\hat{\pi}$.

For instance, when $\pi^{(0)} = 1/2$, we obtain $\pi^{(1)} = y/n$.

For starting values other than $1/2$, adequate convergence usually takes four or five iterations.

4.6.2 Fisher scoring method

Fisher scoring is an alternative iterative method for solving likelihood equations.

It resembles the Newton-Raphson method, the distinction being with the Hessian matrix:

- Fisher scoring uses the *expected value* of the Hessian matrix
  $\Rightarrow$ *expected information*;
- Newton-Raphson uses the observed matrix itself
  $\Rightarrow$ *observed information*.

4.6.2 Fisher scoring method

Let $\mathcal{J}^{(t)}$ denote the approximation at iteration t for the ML estimate of the expected information matrix, i.e., $\mathcal{J}^{(t)}$ has elements $E[-\partial^2 L(\boldsymbol{\beta})/\partial\beta_a\,\partial\beta_b]$, evaluated at $\boldsymbol{\beta}^{(t)}$.

The formula for Fisher scoring is

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + (\mathcal{J}^{(t)})^{-1}\,\mathbf{u}^{(t)}$$

or

$$\mathcal{J}^{(t)}\,\boldsymbol{\beta}^{(t+1)} = \mathcal{J}^{(t)}\,\boldsymbol{\beta}^{(t)} + \mathbf{u}^{(t)}.$$

For estimating a binomial parameter, the information is $n/[\pi(1-\pi)]$ (Section 1.3.2).

4.6.2 Fisher scoring method

A step of Fisher scoring gives

$$\begin{aligned}\pi^{(t+1)} &= \pi^{(t)} + \left[\frac{n}{\pi^{(t)}(1-\pi^{(t)})}\right]^{-1} \frac{y - n\pi^{(t)}}{\pi^{(t)}(1-\pi^{(t)})} \\ &= \pi^{(t)} + \frac{y - n\pi^{(t)}}{n} = \frac{y}{n}.\end{aligned}$$

In Section 4.4.6 we saw $\mathcal{J} = \mathbf{X}'\mathbf{W}\mathbf{X}$. Similarly, here $\mathcal{J}^{(t)} = \mathbf{X}'\mathbf{W}^{(t)}\mathbf{X}$, where $\mathbf{W}^{(t)}$ is $\mathbf{W}$ evaluated at $\beta^{(t)}$.

With Fisher scoring, the estimated asymptotic covariance matrix of $\hat{\beta}$ is a by-product as

$$\widehat{\mathcal{J}^{-1}} = (\mathcal{J}^{(t)})^{-1} \quad \text{for } t \text{ at which convergence is adequate.}$$

For GLMs with a canonical link, the observed and expected information are the same (see Section 4.6.4 below).
$\Rightarrow$ Fisher scoring and Newton-Raphson are the same.

4.6.2 Fisher scoring method

For noncanonical link models,

- Fisher scoring has the following advantages:
  1. It estimates asymptotic covariance matrix as a by-product.
  2. The expected information is necessarily nonnegative definite $\Rightarrow (\mathcal{J}^{(t)})^{-1}$ exists.
  3. It is closely related to weighted least squares methods for ordinary linear models (see Section 4.6.3 below).
- Newton-Raphson has the following advantages:
  1. It has second-order convergence.
  2. It is easier to calculate the observed information for complex models.
  3. The variance estimates of the observed information better approximate a relevant conditional variance.
  4. It is closer to the data.

Outline

4.7 Quasi-likelihood and generalized linear models

Recall Section 4.4.4, for natural exponential family:

$$u_j = \frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^{N} \frac{(y_i - \mu_i)\, x_{ij}}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = \sum_{i=1}^{N} \frac{(y_i - \mu_i)\, x_{ij}}{\nu(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i}.$$

- The likelihood equations depend on the distribution of Y_i only through μ_i and $\text{var}(Y_i) = \nu(\mu_i)$ for some function ν.
- The relationship between μ_i and $\text{var}(Y_i)$ characterizes the distribution
  $\Rightarrow$ the choice of distribution determines the mean-variance relationship $\nu(\mu_i)$.

4.7.1 Mean-variance relationship determines quasi-likelihood estimates

The approach of *quasi-likelihood estimation* (拟似然估计)

- assumes only a mean-variance relationship $\text{var}(Y_i) = \nu(\mu_i)$,
- without assumption of a specific distribution for Y_i.

Compared with usual ML approach for GLM, it has the same

- link function, linear predictor,
- equations as the likelihood equations.

4.7.1 Mean-variance relationship determines quasi-likelihood estimates

Illustration: Suppose $\{Y_i\}$ are independent with $\nu(\mu_i) = \mu_i$.

The quasi-likelihood (QL) estimates are the solution to

$$u_j = \sum_{i=1}^{N} \frac{(y_i - \mu_i)\, x_{ij}}{\nu(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} = \sum_{i=1}^{N} \frac{(y_i - \mu_i)\, x_{ij}}{\mu_i} \frac{\partial \mu_i}{\partial \eta_i} = 0.$$

In this case, the QL estimates are also ML estimates when the random component has a Poisson distribution (in the exponential dispersion family).

The QL estimates have asymptotic covariance matrix of the same form as in GLMs, namely $\text{cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}$, $w_i = (\partial \mu_i / \partial \eta_i)^2 / \text{var}(Y_i)$.

4.7.2 Overdispersion for Poisson GLMs and quasi-likelihood

The Poisson GLM assumes var$(Y_i) = \nu(\mu_i) = \mu_i$, which is unrealistic because of overdispersion, i.e., var$(Y_i) > \mu_i$. One cause of overdispersion is heterogeneity among subjects.

An alternative mean-variance relationship has the form

$$\nu(\mu_i) = \phi\,\mu_i \quad \text{for some constant } \phi.$$

$\phi > 1$ represents overdispersion for the Poisson model.

Estimate β: The QL estimating equations become

$$u_j = \sum_{i=1}^{N} \frac{(y_i - \mu_i)\,x_{ij}}{\phi\,\mu_i}\,\frac{\partial \mu_i}{\partial \eta_i} = 0 \quad \Rightarrow \quad \sum_{i=1}^{N} \frac{(y_i - \mu_i)\,x_{ij}}{\mu_i}\,\frac{\partial \mu_i}{\partial \eta_i} = 0.$$

The ϕ drops out.

4.7.2 Overdispersion for Poisson GLMs and quasi-likelihood

Thus, the equations are identical to likelihood equations for Poisson models, and model parameter estimates are also identical. However, $w_i = (\partial\mu_i/\partial\eta_i)^2/(\phi\,\mu_i)$.
$\Rightarrow$ The estimated $\text{cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}$ is ϕ times that for the Poisson model.

Estimate ϕ:
The ϕ cannot be estimated from the estimating equation. Consider a variance function with the form $\nu(\mu_i) = \phi\,\nu^*(\mu_i)$, where $\nu^*(\mu_i)$ is the variance assumed by the distribution of random component.
$\Rightarrow \phi = \nu(\mu_i)/\nu^*(\mu_i)$ for all μ_i.

4.7.2 Overdispersion for Poisson GLMs and quasi-likelihood

Let $X^2 = \sum_i (y_i - \hat{\mu}_i)^2 / \nu^*(\hat{\mu}_i)$, a Pearson-type statistic for the simpler model with $\phi = 1$. Then

$$\frac{X^2}{\phi} = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\phi\, \nu^*(\hat{\mu}_i)} = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\nu(\hat{\mu}_i)}.$$

Since X^2/ϕ has a limiting χ^2_{df} with $df = N - p$, we have $E(X^2/\phi) \approx N - p$.

Hence, ϕ can be estimated by $\hat{\phi} = X^2/(N - p)$.

4.7.2 Overdispersion for Poisson GLMs and quasi-likelihood

Summary of QL approach for count data

- Fit the ordinary Poisson model and use its p parameter estimates.
- Multiply the ordinary standard error estimates by $\sqrt{X^2/(N-p)}$.

Illustration: The horseshoe crab data in Section 4.3.2.

1) Poisson GLM (Section 4.3.2)

Number of obs: 173 crabs

Outcome Y_i: Number of satellites of crab i, $i = 1, \ldots, 173$

Predictor for α: $X_{i1} = 1$ for all i $(j = 1)$

Predictor for β: $X_{i2} = X_i =$ width of crab i $(j = 2)$

Estimated model: $\log \hat{\mu}_i = -3.305 + 0.164\, x_i$,
with SE$= 0.020$ for $\hat{\beta} = 0.164$.

4.7.2 Overdispersion for Poisson GLMs and quasi-likelihood

2) QL approach

To improve the adequacy of using a chi-squared statistic to summarize fit, we use the satellite totals and fit for all female crabs at a given width.

Number of obs:	66 distinct width levels
Outcome Y_k^*:	Total number of satellites of all crabs of width level k, $k = 1, \ldots, 66$
Predictor for α:	$X_{k1}^* = 1$ for all k $(j = 1)$
Predictor for β:	$X_{k2}^* = X_k^* =$ width of width level k $(j = 2)$

That is, we re-index each crab i $(i = 1, \ldots, 173)$ by two indexes: the width level index k $(k = 1, \ldots, 66)$ and the within-level index l $(l = 1, \ldots, n_k)$, with $\sum_{k=1}^{66} n_k = 173$.

4.7.2 Overdispersion for Poisson GLMs and quasi-likelihood

Hence,

$$Y_k^* = \sum_{l=1}^{n_k} Y_{(kl)} = \sum_{\{\text{all } i \text{ in width level } k\}} Y_i,$$

$$\mu_k^* = \sum_{l=1}^{n_k} \mu_{(kl)} = \sum_{\{\text{all } i \text{ in width level } k\}} \mu_i,$$

and $X_{(kl)} = X_k^*$ or equivalently $X_i = X_k^*$ for all i in width level k.

As specified in Section 4.4.7, $\partial \mu_i / \partial \eta_i = \mu_i$ for Poisson model with log link, so the GLM likelihood equations are

$$\sum_{i=1}^{173} y_i \, x_{ij} = \sum_{i=1}^{173} \mu_i \, x_{ij} \quad \text{for } j = 1, 2,$$

$$\text{replace } i \text{ by } (kl) \quad \Rightarrow \quad \sum_{k=1}^{66} \sum_{l=1}^{n_k} y_{(kl)} \, x_{(kl)j} = \sum_{k=1}^{66} \sum_{l=1}^{n_k} \mu_{(kl)} \, x_{(kl)j}.$$

4.7.2 Overdispersion for Poisson GLMs and quasi-likelihood

$$\text{For } j=1 \quad \Rightarrow \quad \sum_{k=1}^{66}\sum_{l=1}^{n_k} y_{(kl)} = \sum_{k=1}^{66}\sum_{l=1}^{n_k} \mu_{(kl)} \quad \Rightarrow \sum_{k=1}^{66} y_k^* = \sum_{k=1}^{66} \mu_k^* ;$$

$$\text{For } j=2 \quad \Rightarrow \quad \sum_{k=1}^{66} x_k^* \sum_{l=1}^{n_k} y_{(kl)} = \sum_{k=1}^{66} x_k^* \sum_{l=1}^{n_k} \mu_{(kl)}$$

$$\Rightarrow \quad \sum_{k=1}^{66} x_k^* \, y_k^* = \sum_{k=1}^{66} x_k^* \, \mu_k^* \, .$$

The model gives $\log \hat{\mu}_k^* = \hat{\alpha} + \hat{\beta}\, x_k^*$, with the same $\hat{\alpha}$ and $\hat{\beta}$ as in the Poisson GLM above,

$\Rightarrow \hat{\mu}_k^* = \exp(\hat{\alpha} + \hat{\beta}\, x_k^*)$.

$X^2 = \sum_{k=1}^{66} (y_k^* - \hat{\mu}_k^*)^2 / \hat{\mu}_k^* = 174.3$.

$\Rightarrow \hat{\phi}^{1/2} = \sqrt{174.3/(66-2)} = 1.65$ (two parameters α and β),

$\Rightarrow$ SE $= 1.65 \times 0.020 = 0.033$ for $\hat{\beta} = 0.164$.

4.7.3 Overdispersion for binomial GLMs and quasi-likelihood

When y_i is the sample mean of n_i independent binary observations with parameter π_i, $i = 1, \ldots, N$, then $E(Y_i) = \pi_i$ and $\text{var}(Y_i) = \pi_i(1 - \pi_i)/n_i$.

A simple quasi-likelihood approach uses the alternative variance function

$$\nu(\pi_i) = \phi\, \pi_i\, (1 - \pi_i)/n_i.$$

Overdispersion occurs when $\phi > 1$.

4.7.3 Overdispersion for binomial GLMs and quasi-likelihood

As in the overdispersed Poisson case, ϕ drops out of the estimating equations and enters the denominator of w_i.

1. The QL estimates are the same as the ML estimates for the binomial model.
2. The asymptotic covariance matrix multiplied by ϕ and SE multiplied by $\sqrt{\phi}$.

Using the X^2 fit statistic for the ordinary binomial model, $\hat{\phi} = X^2/(N-p)$ (Finney 1947).

4.7.4 Teratology overdispersion example

Table 4.5 Response Counts of Litter Size, Number Dead for 58 Litters of Rats in Low-Iron Teratology Study

Group 1: Untreated (low iron).
(10, 1) (11, 4) (12, 9) (4, 4) (10, 10) (11, 9) (9, 9) (11, 11) (10, 10)
(10, 7) (12, 12) (10, 9) (8, 8) (11, 9) (6, 4) (9, 7) (14, 14) (12, 7) (11, 9)
(13, 8) (14, 5) (10, 10) (12, 10) (13, 8) (10, 10) (14, 3) (13, 13)
(4, 3) (8, 8) (13, 5) (12, 12)

Group 2: Injections days 7 and 10
(10, 1) (3, 1) (13, 1) (12, 0) (14, 4) (9, 2) (13, 2) (16, 1)
(11, 0) (4, 0) (1, 0) (12, 0)

Group 3: Injections days 0 and 7
(8, 0) (11, 1) (14, 0) (14, 1) (11, 0)

Group 4: Injections weekly
(3, 0) (13, 0) (9, 2) (17, 2) (15, 0) (2, 0) (14, 1) (8, 0) (6, 0) (17, 0)

4.7.4 Teratology overdispersion example

Number of obs: 58 litters
Outcome: Number of dead fetuses in each litter
Predictor: Treatment group

Due to unmeasured covariates and genetic variability, the probability of death may vary from litter to litter within a particular treatment group.

Let

$n_{i(g)} =$ number of fetuses in litter i in treatment group g.
$x_{i(g)} =$ number of dead fetuses in litter i in treatment group g.
$y_{i(g)} = x_{i(g)}/n_{i(g)} =$ proportion of dead fetuses out of the $n_{i(g)}$.
$\pi_{i(g)} =$ probability of death for a fetus in litter i in group g.

Consider the model with $n_{i(g)}\, y_{i(g)}$ a bin$(n_{i(g)}, \pi_{i(g)})$ variate, where $\pi_{i(g)} = \pi_g$ for $g = 1, 2, 3, 4$, i.e., all litters in a group have the same probability of death π_g.

4.7.4 Teratology overdispersion example

The ML estimate of π_g is $\hat{\pi}_g = \sum_i x_{i(g)} / \sum_i n_{i(g)}$, with SE$= \sqrt{\hat{\pi}_g(1-\hat{\pi}_g)/\sum_i n_{i(g)}}$.

g	1	2	3	4
$\hat{\pi}_g$	0.758	0.102	0.034	0.048
SE$_g$	0.024	0.028	0.024	0.021

The estimated probability of death is considerably higher for group 1 (placebo). For litter i in group g,

$n_{i(g)}\,\hat{\pi}_g$ = fitted number of dead fetuses,

$n_{i(g)}\,(1-\hat{\pi}_g)$ = fitted number of live fetuses.

Comparing these fitted values with the observed counts of dead and live fetuses, the Pearson statistic is

$$X^2 = \sum_g \sum_i \left\{ \frac{[x_{i(g)} - n_{i(g)}\,\hat{\pi}_g]^2}{n_{i(g)}\,\hat{\pi}_g} + \frac{[(n_{i(g)} - x_{i(g)}) - n_{i(g)}\,(1-\hat{\pi}_g)]^2}{n_{i(g)}\,(1-\hat{\pi}_g)} \right\} = 154$$

4.7.4 Teratology overdispersion example

With total 58 litters and 4 parameters (π_g), the df$= 58 - 4 = 54$.

Since the X^2 is quite large (i.e., lack of fit), there is considerable evidence of overdispersion.

Using the QL approach, $\hat{\pi}_g$ are the same as the ML estimates, with $\hat{\phi} = X^2/(N-p) = 154.7/(58-4) = 2.86$ and $\hat{\phi}^{1/2} = 1.69$.

Even with this adjustment for overdispersion, there is still strong evidence that the probability of death is substantially higher for the placebo group.

For instance, a 95% CI for $\pi_1 - \pi_2$ is

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm 1.96[\hat{\phi} \times \text{SE}_1^2 + \hat{\phi} \times \text{SE}_2^2]^{1/2} = (0.54,\ 0.78).$$

This is wider than the CI without adjustment for overdispersion, (0.59, 0.73).

4.7.4 Teratology overdispersion example

TABLE A.5 SAS Code for Overdispersion Modeling of Teratology Data in Table 4.5

```
data moore;
  input litter group n y @@;
datalines;
 1 1 10 1     2 1 11 4     3 1 12 9     4 1 4 4     5 1 10 10
...
55 4 14 1    56 4  8 0    58 4 17 0
;
proc genmod;  class group;
  model y/n=group/dist=bin link=identity noint;
estimate 'pi1-pi2' group 1 -1 0 0;
proc genmod;  class group;
  model y/n=group/dist=bin link=identity noint scale=pearson;
```