generalized Linear Models and Extensions

Consider a study of childhood obesity, where a sample of n children are followed longitudinally (i.e. over time). The researchers are interested in the time trajectory of Body Mass Index (BMI = weight$_{kg}$ / HT^2_m) as a function of age.

A simple linear regression model might pose that

$$(1) \qquad BMI_{ij} = \beta_0 + \beta_1 AGE_{ij} + \varepsilon_{ij}$$

i = child, j = time point

$i = 1, 2, \cdots, n$

$j = 1, 2, \cdots n_i$

Here we have n_i observations on child i - data are obtained at ages Age_{ij} $j = 1, 2, \cdots, n_i$. In the simplest model we might assume that $\varepsilon_{ij} \sim$ indep $N(0, \sigma^2)$. For simplicity, I am assuming that sex is unimportant, but will consider this later.

A more general model might assume that the intercepts and slopes vary by child

$$(2) \qquad BMI_{ij} = \beta_{0i} + \beta_{1i} Age_{ij} + \varepsilon_{ij}$$

with no other changes in (1).

In standard linear regression we assume the regression effects are <u>fixed effects</u> - fixed unknowns to be estimated

from the data. For model (1) there are 2 regression (β_0, β_1) effects and 1 variance term (σ^2) while model (2) has $2n$ regression effects (β_{0i}, β_{1i}) $i = 1, 2, \cdots, n$ and a variance term σ^2.

One limitation of model (2) is that responses on the same child over time are likely to be correlated, a feature not captured by the model. There are 2 standard extensions of (2) that handle this issue. One approach, common with time series and longitudinal data is to model the residuals, for example with an AR(1) process [auto-regressive process, order 1]

$$\varepsilon_{ij} = \rho\, \varepsilon_{ij-1} + \delta_j \qquad j = 2, 3, \cdots n_i$$

where the δ_j are independent of the $\varepsilon_{i,j-1}$'s. This implies

$$\text{correlation}(\varepsilon_{ij}, \varepsilon_{i(j-k)}) = \rho^k$$

(k time point separation.)

Another approach is to treat the subject specific intercepts β_{0i} and slopes β_{1i} as random variables, with their

own distributions, i.e.

$$\beta_{0i} \overset{ind}{\sim} N(\beta_0, \sigma_0^2) \qquad \beta_{1i} \overset{ind}{\sim} N(\beta_1, \sigma_1^2)$$

and where β_{0i} and β_{1i} might be correlated (or not).

With this approach we can write

$$\beta_{0i} = \beta_0 + \beta_{0i}^* \qquad \beta_{0i}^* \overset{ind}{\sim} N(0, \sigma_0^2)$$

$$\beta_{1i} = \beta_1 + \beta_{1i}^* \qquad \beta_{1i}^* \overset{ind}{\sim} N(0, \sigma_1^2)$$

so that β_{0i}^* and β_{1i}^* measure the difference between the population mean intercept & slopes (β_0, β_1) and the subject specific intercept & slope (β_{0i}, β_{1i}). We usually write the model

$$BMI_{ij} = \beta_{0i} + \beta_{1i} age_{ij} + \varepsilon_{ij}$$

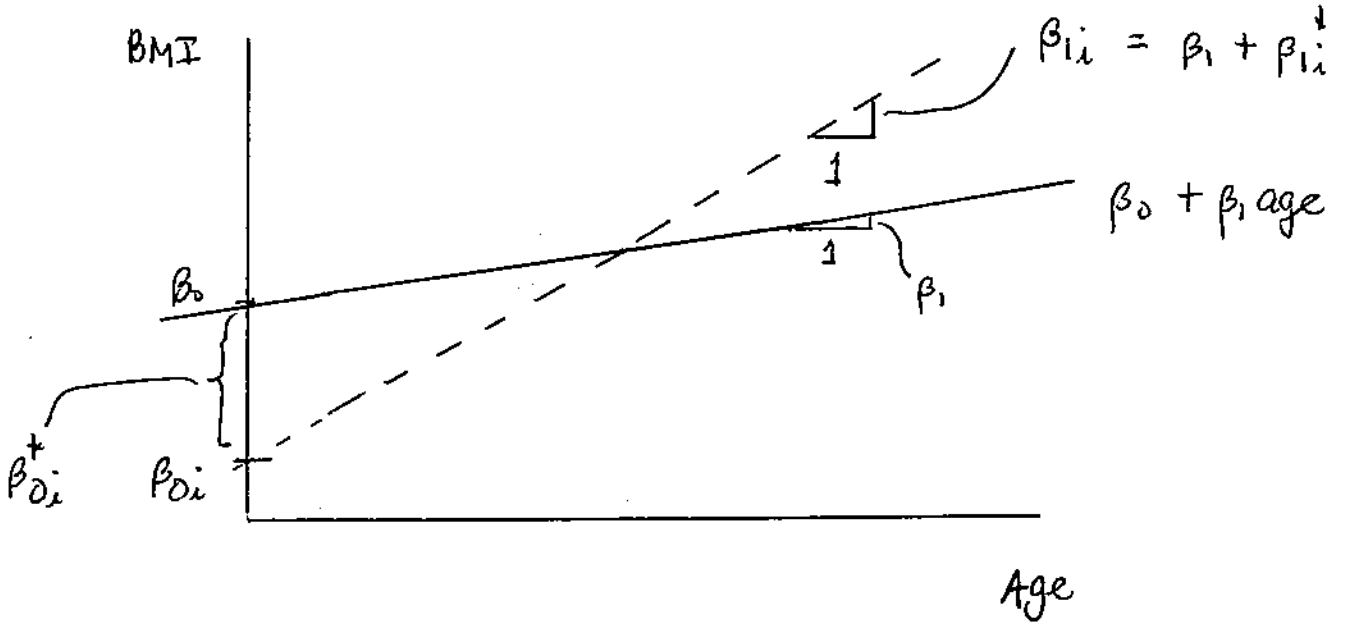
$$\begin{cases} \varepsilon_{ij} \overset{ind}{\sim} N(0, \sigma^2) \\ \beta_{0i} \overset{ind}{\sim} N(\beta_0, \sigma_0^2) \\ \beta_{1i} \overset{ind}{\sim} N(\beta_1, \sigma_1^2) \end{cases}$$

can be indep or not

or

$$(3) \quad BMI_{ij} = (\beta_0 + \beta_{0i}^*) + (\beta_1 + \beta_{1i}^*) age_{ij} + \varepsilon_{ij}$$

$$= \underbrace{\beta_0 + \beta_1 Age_{ij}}_{\text{population average regression line}} + \underbrace{\beta_{0i}^* + \beta_{1i}^* Age_{ij}}_{\text{subject specific deviation from pop average}} + \varepsilon_{ij}$$



It is not entirely clear (yet), but the <u>random intercept & slope</u> model (3) implies that responses on the same individual are correlated because $BM_{i1}, BM_{i2}, \cdots, BM_{ij_i}$ all depend on the same random effects β_{0i}, β_{1i} (or equivalently β_{0i}^* and β_{1i}^*)

Another potential advantage of (3) over (2) is that (2) requires the ~~estimation~~ of $2n$ regression effects. In (3), the subject specific intercepts β_{0i} and slopes β_{1i} are random variables (a <u>random effects</u>) not parameters. This random effect model has only 5 parameters: β_0, β_1 and 3 variance terms σ^2, σ_0^2 and σ_1^2 to be estimated

The rand on effect are predicted based upon the estimated parameters

However, a primary advantage of (3) is not the reduction in parameters to be estimated but rather the important realization that model (2) only applies to those individuals selected for the study. Model (3), through viewing the individuals sampled as representative of an underlying population [via the sampling assumption on the subject specific intercepts and slopes], allows estimation of a population average regression curve $\beta_0 + \beta_1 Age$ plus subject specific inferences on individual regression lines.

Model (3) is a special case of a <u>linear mixed model</u> (LMM) - it includes both <u>fixed</u> (β_0, β_1) and <u>random</u> effect (β_{0i}, β_{1i})

Linear mixed models are widely used to model longitudinal data with normally distributed responses.

Mixed Model : Fixed effects
+
Random effects

Suppose instead that the researcher is interested in childhood obesity as a function of age. More specifically, they define

$$OBESE_{ij} = \begin{cases} \text{Yes (1)} & \text{if } BMI_{ij} > c(age_{ij}, sex_i) \\ \text{No (2)} & \end{cases}$$

$OBESE_{ij}$: i = child, j = age

$c(age_{ij}, sex_i)$: some cutoff that may be age & sex specific

Set

$$P_{ij} = Pr(OBESE_{ij} = 1)$$

= probability child i is obese at age j

A possible model for how P_{ij} varies with age and sex is

sex_i: 1 Male, 0 Female

$$(4) \quad \log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \beta_{0i} + \beta_{1i} age_{ij} + \beta_{2i} sex_i$$

This is a <u>logistic regression model</u> with subject specific effects for the "intercept" and age & sex.

Logistic regression is a special case of a <u>binary response model</u> (or more generally a <u>binomial response model</u>)

The response $OBese_{ij}$ is a Bernoulli rv with success probability

$$p_{ij} = Pr(OBese_{ij} = 1)$$

i.e. $OBese_{ij} \sim \text{Bernoulli}(p_{ij})$

Note that

$$\mu_{ij} \equiv E(OBese_{ij}) = p_{ij}$$

If we assume responses are independent over time and across individuals, we can specify the model via

(4) $$OBese_{ij} \sim \text{indep Bernoulli}(\mu_{ij})$$

where

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_{0i} + \beta_{1i}\, age_{ij} + \beta_{2i}\, sex_i$$

Let us contrast this to the normal theory linear regression model

(2) $$BMI_{ij} = \beta_{0i} + \beta_{1i}\, age_{ij} + \varepsilon_{ij} \quad \text{where } \varepsilon_{ij} \overset{\text{indep}}{\sim} N(0, \sigma^2)$$

Since $E(\varepsilon_{ij}) = 0$

$$E(BMI_{ij}) \equiv \mu_{ij} = \beta_{0i} + \beta_{1i}\, age_{ij}$$

Thus, we can alternatively define the model via

$$BMI_{ij} \sim \text{indep } N(\mu_{ij}, \sigma^2)$$

where

$$\mu_{ij} = \beta_{0i} + \beta_{1i}\, age_{ij}$$

Model (2)	Model (4)	
$BMI_{ij} = N(\mu_{ij}, \sigma^2)$	$OBESE_{ij} \sim Bernoulli(\mu_{ij})$	Response Dist
μ_{ij}	μ_{ij}	Mean
$g(\mu_{ij}) = \mu_{ij}$	$g(\mu_{ij}) = \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right)$	Link function

The models have different response distributions, but in both cases the <u>mean response</u> μ_{ij} is linearly related to one or more predictors through a <u>link function</u>:

(2) $g(\mu_{ij}) = \mu_{ij} = \beta_{0i} + \beta_{1i}\, age_{ij}$

(4) $g(\mu_{ij}) = \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_{0i} + \beta_{1i}\, age_{ij} + \beta_{2i}\, sex_i$

Thus, the normal linear regression model and the binary response model (4) are <u>structurally similar</u>

Indeed (2) and (4) are special cases of <u>generalized linear models (GLMs)</u> with fixed effects. GLMs allow responses that have <u>exponential family (EF)</u> distributions such as normal, Poisson, Binomial and gamma, where the mean response is linearly related to fixed effects through a <u>link function</u>

Aside (Conceptually important for GLM specification)

We are more used to seeing the normal linear regression model written in the form

$$BMI_{ij} = \underbrace{\beta_{0i} + \beta_{1i}\, age_{ij}}_{\mu_{ij}} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim \text{indep } N(0, \sigma^2)$$

versus

$$BMI_{ij} \sim \text{indep } N(\mu_{ij}, \sigma^2)$$

where

$$\mu_{ij} = \beta_{0i} + \beta_{1i}\, age_{ij}$$

The earlier specification clearly separates the "signal" μ_{ij} from the "noise" ε_{ij} whereas the latter only implicitly makes that distinction – the normal variation σ^2 about the mean μ_{ij} is the "noise".

If we are more "comfortable" with the first representation why not write the Bernoulli response model in the similarly suggestive form

(5) $$OBESE_{ij} = \mu_{ij} + \varepsilon_{ij}$$

where

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \underbrace{\beta_{0i} + \beta_{1i}\, age_{ij} + \beta_{2i}\, sex_i}_{\eta_{ij}}$$

or equivalently

$$\mu_{ij} = \exp(\eta_{ij}) / \{1 + \exp(\eta_{ij})\}$$

This has the appearance of a non-linear regression model. Although the representation (5) is permissable, one immediate recognition is that the definition of the residual ε_{ij} must be defined as a function of the mean μ_{ij} because the response is discrete:

$$\varepsilon_{ij} = \begin{cases} 1 - \mu_{ij} & \text{if } \text{obese}_{ij} = 1 \\ -\mu_{ij} & = 0 \end{cases} \tag{6}$$

In comparison, a normally distributed response can be specified in terms of a mean plus an independently defined residual because the normal distribution is preserved through linear transformations. To avoid the awkward specification (6) we use (4) to represent the model.

The binary response model (4) shares the limitation of (2) that neither model accounts for dependence among the responses on the same individual. For binary response models and more generally for GLMs, there are 2 standard ways to remedy this deficiency.

In normal theory regression, either modelling the residuals ε_{ij} or adding random effects, or both, leads to a complete specification of the joint distribution of responses $Y_{ij}; j = 1,2,\cdots n_i$ within an individual. Assuming responses across subjects are independent leads to a fully specified probability model for which standard inferential methods (MLE or Bayes) are applicable.

One approach to including dependence in the binary response model (4) is based on the recognition that it is difficult to specify a flexible multivariate model for the joint distribution of the responses $OBESE_{ij}$ $j = 1,2,\cdots n_i$ on a given individual. Instead, the marginal distributions are specified

$$OBESE_{ij} \sim \text{Bernoulli}(\mu_{ij})$$

with

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_{0i} + \beta_{1i}\, age_{ij} + \beta_{2i}\, sex_i \tag{7}$$

and estimates of the regression effects are obtained treating the responses as if they were independent. However, the

standard errors for estimate and tests on regression effects are defined in such a way that the dependence among responses is respected. This is the so-called <u>generalized Estimating Equations</u> (GEE) approach to inference for longitudinal (and clustered) data.

A second approach introduces <u>subject specific random effects</u> into the model

$$OBese_{ij} \mid u_{ij} \sim \text{Bernoulli}(u_{ij})$$

where

$$(*) \quad \log\left(\frac{u_{ij}}{1-u_{ij}}\right) = \beta_{0i} + \beta_{1i}\, age_{ij} + \beta_{2i}\, sex_i$$

and $\beta_{0i} \sim \text{indep } N(\beta_0, \sigma_0^2)$

$\beta_{1i} \sim \text{indep } N(\beta_1, \sigma_1^2)$ or $\beta_{Ki} \overset{\text{indep}}{\sim} N(\beta_K, \sigma_K^2)$ $K = 0,1,2$

$\beta_{2i} \sim \text{indep } N(\beta_2, \sigma_2^2)$

writing $\beta_{Ki} = \beta_K + \beta_{Ki}^*$ where $\beta_{Ki}^* \sim \text{indep } N(0, \sigma_K^2)$

we can write

$$\log\left(\frac{u_{ij}}{1-u_{ij}}\right) = \beta_0 + \beta_1 age_{ij} + \beta_2 sex_i + \beta_{0i} + \beta_{1i}^* age_{ij} + \beta_{2i}^* sex_i$$

There are some fundamental difference between the GEE approach captured by (7) and the random effects approach considered in (8). This point will be elaborated upon later in the semester.

I will note that (8) is a special case of a <u>generalized linear mixed model (GLMM)</u>, which is a GLM with possibly both fixed and random effects.

This course focusses attention on studying the <u>structure</u>, <u>analysis</u> and <u>computations</u> for a variety of these generalizations of normal theory regression. In particular we will examine in some depth issues surrounding

GLMs
LMMs and either GEE for GLMs
or GLMM

Much if not all of attention will be devoted to ML based inference, but Bayesian inference may also be entertained depending on interest (yours and MINE)

Generalized Linear Models

- Fairly standard notation used here - close to writeup for Proc genMOD in SAS - see course webpage

glms extend the standard linear regression model for normal responses to allow both non-normal responses and non-linear relationships. glms focus on modelling a transformation of the mean as a linear function of covariates in contrast to transforming the response to induce linearity.

Exponential Families (EF)

glms assume a scalar rv Y is a member of the EF of distributions which includes several well known discrete (Bernoulli, Binomial, Poisson) and continuous (normal, gamma, exponential) distributions.

A scalar rv Y has an EF distribution if it's probability function (discrete case) or density function (continuous case) has the form

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

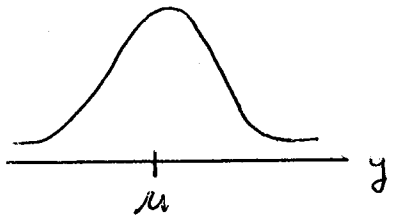
for functions $b(\theta)$, $c(y, \phi)$ and $a(\phi)$, where θ and ϕ are parameters.

In many cases ϕ is known, and it is also usually assumed that

$$a(\phi) = \frac{\phi}{\omega}$$

for some known weight ω. It is assumed $\phi > 0$ and $\omega > 0$

ex: $Y \sim N(\mu, \sigma^2)$ then



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

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

$$= \exp\left\{\frac{y\mu - .5\mu^2}{\sigma^2} + \left[\frac{-y^2}{2\sigma^2} - \log(\sqrt{2\pi}\,\sigma)\right]\right\}$$

If we define: $\theta = \mu$, $\phi = \sigma^2$, $\omega = 1$

$$b(\theta) = .5\theta^2 = .5\mu^2$$

$$c(y, \phi) = \frac{-y^2}{2\phi} - \log(\sqrt{2\pi\phi}) = \frac{-y^2}{2\sigma^2} - \log(\sqrt{2\pi}\,\sigma^2)$$

we have density in form of an EF

ex: If $Y \sim \text{Poisson}(\mu)$ with

$$f(y) = \Pr(Y = y) = e^{-\mu}\mu^y / y! \qquad y = 0, 1, 2, \cdots$$

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

This is an EF with

$$\theta = \log \mu \qquad \phi = 1 \qquad \omega = 1$$

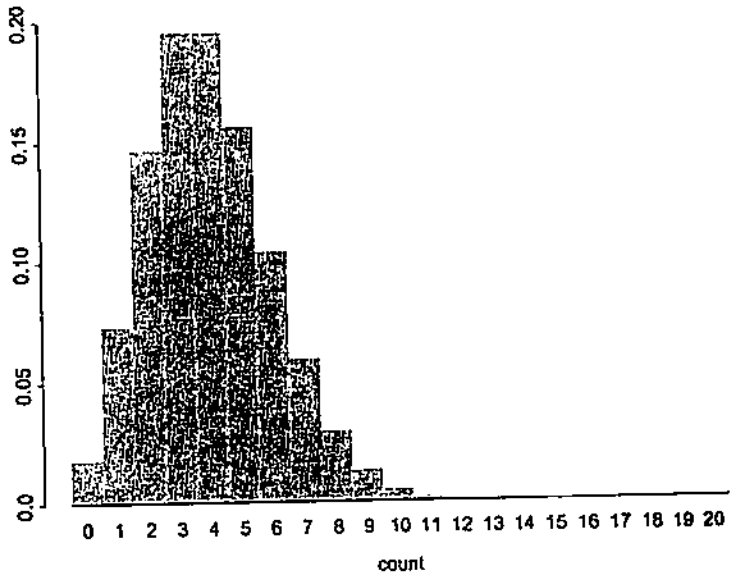
$$b(\theta) = \mu = \exp(\theta) \qquad c(y, \phi) = -\log y!$$

For a Poisson(μ) rv the variance is a function of the mean:

$$\text{Var}(Y) = E(Y) = \mu$$

in contrast to the standard normal model where mean & variance

Figure 1: *Poisson probabilities with mean = 4.*



are distinct parameters, unrelated.

ex: Suppose $Y \sim \text{Binomial}(n, p)$ with

$$f(y) = \Pr(Y = y) = \binom{n}{y} p^y (1-p)^{n-y} \qquad y = 0, 1, \ldots, n$$

$$= \exp\left\{ y \log p + (n-y) \log(1-p) + \log\binom{n}{y} \right\}$$

$$= \exp\left\{ y \log\left(\frac{p}{1-p}\right) + n \log(1-p) + \log\binom{n}{y} \right\}$$

This is in the form of an EF. However, it is customary with GLMs to rewrite this in terms of a RV

$$y^* = y/n = \text{proportion of successes}$$

Then

$$f(y^*) = \Pr(y^* = y^*) = \binom{n}{n y^*} p^{n y^*} (1-p)^{n(1-y^*)} \qquad y^* = 0, \frac{1}{n}, \frac{2}{n}, \cdots, 1$$

$$= \exp\left\{ n y^* \log\left(\frac{p}{1-p}\right) + n \log(1-p) + \log\binom{n}{n y^*} \right\}$$

$$= \exp\left\{ \underbrace{n}_{\frac{1}{a(\phi)}} \left[\underbrace{y^* \log\left(\frac{p}{1-p}\right)}_{\theta} + \underbrace{\log(1-p)}_{-b(\theta)} \right] + \underbrace{\log\binom{n}{n y^*}}_{c(y^*, \theta)} \right\}$$

Here: $\theta = \log\left(\frac{p}{1-p}\right) \Rightarrow p = \frac{e^\theta}{1+e^\theta} \Rightarrow 1-p = \frac{1}{1+e^\theta}$

$$\Rightarrow \log(1-p) = -\log(1+e^\theta)$$

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

$$\left.\begin{array}{rcl} \phi &=& 1 \\ \omega &=& n \end{array}\right\} \quad a(\phi) = \frac{\phi}{\omega} = \frac{1}{n}$$

$$c(y^*, \phi) = \log\binom{\omega}{\omega y^*} \quad \langle \omega \text{ is constant} \rangle$$

Note that if $Y^* \sim Bin(n,p)/n$ i.e. $Y^* = Y/n$ where $Y \sim Bin(n,p)$ then

$$E(Y^*) = \frac{1}{n} E(Y) = \frac{1}{n} np = p$$

$$Var(Y^*) = \frac{1}{n^2} Var(Y) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

i.e. variance is a function of the mean (and the weight)

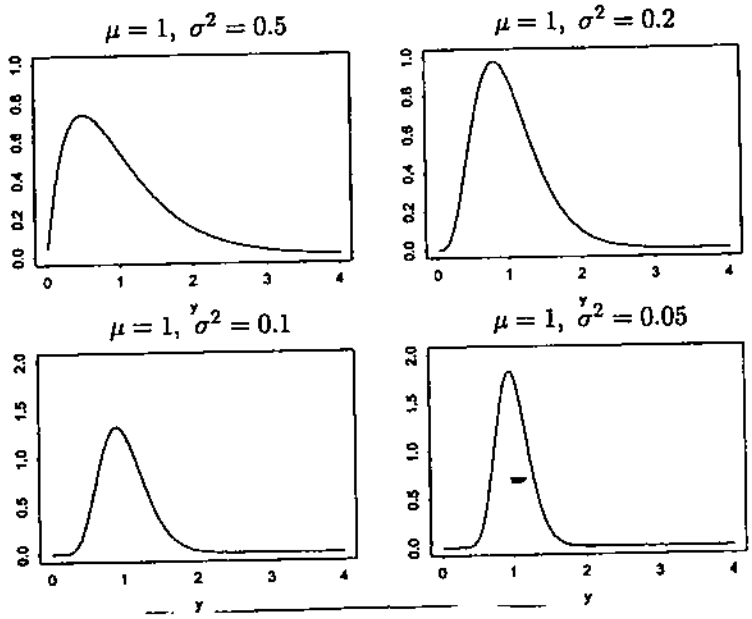
ex: Suppose $Y \sim gamma(\mu, \sigma^2)$ with density for $y > 0$

$$f(y) = \frac{1}{y\,\Gamma(1/\sigma^2)} \left(\frac{y}{\mu\sigma^2}\right)^{1/\sigma^2} \exp\left(-\frac{y}{\mu\sigma^2}\right) \qquad \mu > 0,\ \sigma^2 > 0$$

where $\Gamma(\cdot)$ is the gamma function (for integer $k > 0$: $\Gamma(k+1) = k!$)

This can be written in EF form. (will be homework problem!)

Figure 2: *Gamma probability density functions.*



It is also common to write the density using $v = 1/\sigma^2$, as in SAS:

$$f(y) = \frac{1}{y\,\Gamma(v)} \left(\frac{y v}{\mu} \right)^{v} \exp\left\{ -\frac{y v}{\mu} \right\} \qquad y > 0$$

With our earlier representation (where we use a "moments parameterization")

$$E(Y) = \mu \qquad \text{and} \qquad \text{var}(Y) = \mu^2 \sigma^2$$

i.e. the variance depends on the model. Further, the gamma distribution has constant coefficient of variation

$$CV(Y) = \frac{SD(Y)}{E(Y)} = \frac{\sqrt{\sigma^2 \mu^2}}{\mu} = \sigma$$

i.e. If we have $Y_i \sim$ ind gamma(μ_i, σ^2) $i = 1, 2, \cdots n$ where the "inverse scale" parameter σ^2 is identical across observations, then this sequence of rv's has a constant CV.

Remark: Wikipedia is a useful online reference for probability distributions.

The inverse gamma is another EF distribution, as is the exponential (μ), with density

$$f(y) = \frac{1}{\mu} \exp(-y/\mu) \qquad y > 0$$

which is a gamma$(\mu, 1)$ distribution.

Moments

For an EF distribution

$$E(Y) \equiv \mu = b'(\theta) \qquad b'(\theta) = \frac{\partial}{\partial \theta} b(\theta)$$

$$\text{var}(Y) = \frac{b''(\theta)\phi}{\omega} \qquad b''(\theta) = \frac{\partial^2 b(\theta)}{\partial \theta^2}$$

Since μ is used to identify $E(Y)$, we see that μ is a function of θ (and vice-versa). We often define

$$V(\mu) \equiv b''(\theta)$$

and call $V(\cdot)$ the "variance function" as

$$\text{var}(Y) = V(\mu)\frac{\phi}{\omega}$$

These results come from properties of the <u>score function</u>. Let us see where these result originate for continuous EF distributions

Noting that $f(y) \equiv f(y;\theta,\phi)$ is a function of θ and ϕ

$$1 = \int f(y)\,dy$$

so

$$\frac{\partial}{\partial \theta}(1) = \frac{\partial}{\partial \theta}\int f(y)\,dy$$

$$= \int \left\{\frac{\partial}{\partial \theta} f(y)\right\} dy$$

derivative can be passed under integral sign with EF distributions

Now

$$\frac{\partial}{\partial\theta}\log f(y) = \frac{1}{f(y)}\frac{\partial}{\partial\theta}f(y)$$

$$\Rightarrow \frac{\partial}{\partial\theta}f(y) = f(y)\frac{\partial}{\partial\theta}\log f(y)$$

so

$$0 = \frac{\partial}{\partial\theta}(1) = \int \left\{\frac{\partial}{\partial\theta}f(y)\right\} dy$$

$$= \int f(y)\frac{\partial}{\partial\theta}\log f(y)\, dy$$

$$= E_Y\left\{\underbrace{\frac{\partial}{\partial\theta}\log f(Y)}_{\text{score function}}\right\}$$

Fn an EF

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

$$\frac{\partial \log f(y)}{\partial\theta} = \frac{1}{a(\phi)}\{y - b'(\theta)\}$$

so

$$0 = E_Y\left\{\frac{\partial}{\partial\theta}\log f(Y)\right\} = \frac{1}{a(\phi)}E_Y\{Y - b'(\theta)\}$$

$$\Rightarrow E(Y) = b'(\theta)$$

Similarly, we can show in general

$$E_Y\left\{\frac{\partial^2}{\partial\theta^2}\log f(Y)\right\} = -E_Y\left\{\frac{\partial}{\partial\theta}\log f(Y)\right\}^2$$

For an EF

$$\frac{\partial^2}{\partial\theta^2}\log f(y) = -\frac{b''(\theta)}{a(\phi)}$$

and

$$-\left\{\frac{\partial}{\partial\theta}\log f(y)\right\}^2 = -\left(\frac{y-b'(\theta)}{a(\phi)}\right)^2$$

so

$$E_Y\left(\frac{\partial^2 \log f(Y)}{\partial\theta^2}\right) = -\frac{b''(\theta)}{a(\phi)} = -E_Y\left\{\frac{\partial}{\partial\theta}\log f(Y)\right\}^2$$

$$= -\frac{1}{a(\phi)^2}E(Y-b'(\theta))^2$$

or

$$\frac{b''(\theta)}{a(\phi)} = \frac{\mathrm{var}(Y)}{a^2(\phi)} \Rightarrow \mathrm{var}(Y) = a(\phi)\,b''(\theta)$$

$$= \frac{\phi}{\omega}b''(\theta) = V(\mu)\frac{\phi}{\omega}$$

ex: $Y \sim N(\mu, \sigma^2)$. We know $E(Y) = \mu$ and $\mathrm{var}(Y) = \sigma^2$

As an EF: $\theta = \mu$ $b(\theta) = .5\mu^2$ $\phi = \sigma^2$ $b(\theta) = .5\theta^2$ (p15)

$E(Y) = b'(\theta) = \theta = \mu$ ✓

$\mathrm{var}(Y) = \frac{b''(\theta)\,\phi}{\omega} = \frac{1\cdot\sigma^2}{1} = \sigma^2$ Here $V(\mu) = 1$

ex: $Y \sim \text{Poisson}(\mu)$. we know

$$E(Y) = \text{var}(Y) = \mu$$

As an EF (p16) with

$$\theta = \log \mu \qquad b(\theta) = \exp(\theta) = \mu \qquad \phi = 1 \quad \omega = 1$$

we get

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

$$\text{var}(Y) = b''(\theta)\frac{\phi}{\omega} = \frac{b(\theta)\,\phi}{1} = \mu \qquad \text{here } V(\mu) = \mu$$

ex: $Y^* \sim \text{Bin}(n, \mu)/n$. we know

$$E(Y^*) = \mu \quad \text{and} \quad \text{var}(Y^*) = \frac{\mu(1-\mu)}{n}$$

here I am using "μ" instead of "p" so that $E(Y^*) = \mu$ as per notation.

As an EF (p17)

$$\theta = \log\left(\frac{\mu}{1-\mu}\right) \qquad \mu = \frac{e^\theta}{1+e^\theta} \qquad \phi = 1$$

$$b(\theta) = \log(1+e^\theta) = -\log(1-\mu) \qquad \omega = n$$

$$E(Y^*) = b'(\theta) = \frac{1}{1+e^\theta}\frac{\partial}{\partial\theta}(1+e^\theta) = \frac{e^\theta}{1+e^\theta} = \mu$$

$$\text{var}(Y^*) = \frac{b''(\theta)\phi}{\omega} = \frac{1}{n}\frac{\partial}{\partial\theta}\left\{e^\theta(1+e^\theta)^{-1}\right\}$$

$$= \frac{1}{n}\left\{e^\theta(1+e^\theta)^{-1} - e^\theta(1+e^\theta)^{-2}e^\theta\right\}$$

$$= \frac{1}{n}\left\{\frac{e^\theta}{1+e^\theta} - \left(\frac{e^\theta}{1+e^\theta}\right)^2\right\}$$

$$= \frac{1}{n}\left\{\mu - \mu^2\right\} = \frac{1}{n}\mu(1-\mu) \qquad \text{so } V(\mu) = \mu(1-\mu)$$

Remarks

1) The mean μ for an EF depends on θ but not ϕ nor ω. The variance may depend on θ (or μ), ϕ and ω.

2) ϕ is often a "scale parameter"

3) The Binomial & Poisson do not have a scale parameter while the normal and gamma do, with $\phi = \sigma^2$.

4) Distributions for which the variance depends on the mean have a non-constant variance function $V(\mu)$. Such models include the gamma, Poisson & Binomial but not the normal

5) In the EF definition, it is implicit that the domain of Y does not depend on either θ or ϕ. Thus, a Uniform $(0,\theta)$ distribution is not a member of the EF. The Cauchy is also not an EF member.

6) Properties of EF can be derived for both discrete & continuous cases simultaneously by defining distributions or densities with respect to a dominating measure (typically lebesgue measure or a counting measure). We don't need this level of sophistication here.

7) The primary interest with a GLM will be to model the mean function μ, or equivalently θ, assuming the responses follow an EF distribution

8) θ is called the <u>natural parameter</u> of an EF distribution

If $Y \sim N(\mu, \sigma^2)$ the natural parameter is $\theta = \mu$

If $Y \sim \text{Poisson}(\mu)$ $\theta = \log(\mu)$

If $Y^+ \sim \text{Bin}(n, \mu)/n$ $\theta = \log\left(\frac{\mu}{1-\mu}\right)$

This has important consequences for EF theory, where "optimal" tests and CI can be constructed for natural parameters.

In GLM, using a link function corresponding to the natural parameter does have certain (minor) implications that we will explore.

Logistic Regression - Some background

Logistic regression & other binomial response models are extensively used in medical research. Given their prominence, I will provide some additional discussion on these models below

For a simple logistic regression model we have

$$Y_j \sim \text{indep } Bin(n_j, \mu_j)$$

or in EF format $Y_j^* = Y_j/n_j \sim \text{indep } Bin(n_j, \mu_j)/n_j$. The response probabilities satisfies

$$\log\left(\frac{\mu_j}{1-\mu_j}\right) = \beta_0 + \beta_1 x_j \qquad j = 1, 2, \cdots$$

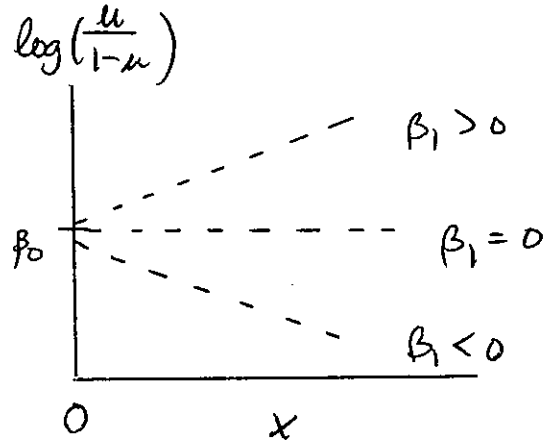
If we write

$$\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 x$$

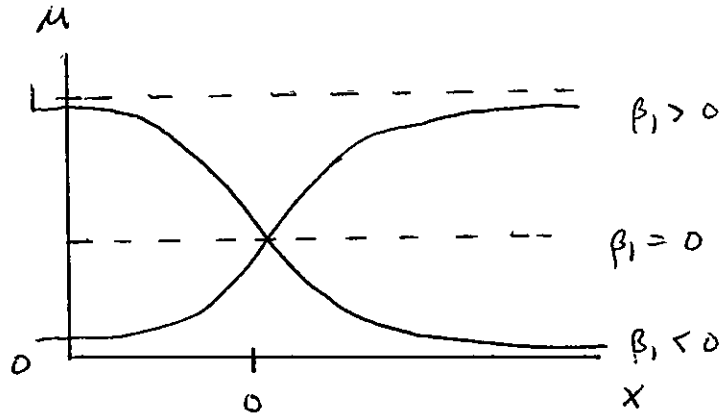
to define the general form of the relationship, then the "logit" transformed probability is linearly related to x, with intercept β_0 and slope β_1. On the probability scale

$$\mu = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

Plots of μ as a function of x are given below. Curve is Sigmoidal (S) shaped.



logit scale



probability scale

Odds and Odds Ratio

Suppose individuals can be classified according to whether they have been exposed to a risk factor and ultimately whether they developed a specific disease

$$Y = \begin{cases} 1 & \text{if develop disease} \\ 0 & \text{if not} \end{cases} \qquad E = \begin{cases} 0 & \text{not exposed} \\ 1 & \text{exposed to risk factor} \end{cases}$$

Let

$$p_0 = Pr(Y=1 \mid E=0) \qquad p_1 = Pr(Y=1 \mid E=1)$$

p_0: prob develop disease given not exposed

p_1: prob develop disease given exposed

The odds of developing disease given exposed are

$$p_1/(1-p_1)$$

The odds of developing disease given not exposed

$$p_0/(1-p_0)$$

Note that if

$$p = Pr(\text{Success})$$

for an arbitrary binary event, Then odds of success

$$\frac{p}{1-p} \begin{matrix} < \\ = \\ > \end{matrix} 1 \iff p \begin{matrix} < \\ = \\ > \end{matrix} .5$$

Furthermore, the odds is an increasing function of p.

The <u>odds ratio</u> for developing disease for exposed relative to unexposed individual is

$$OR = \frac{\text{odds disease given exposed}}{\text{odds disease given not exposed}}$$

$$= \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$$

Can show

$$OR \begin{matrix} > \\ = \\ < \end{matrix} 1 \iff p_1 \begin{matrix} > \\ = \\ < \end{matrix} p_0$$

The odds ratio is a <u>measure of association</u> for binary variables.

i.e. $OR > 1 \Rightarrow$ more likely to develop disease given exposed versus not exposed

$< 1 \Rightarrow$ less "

$= 1 \Rightarrow$ just as likely

Now consider the simple logistic regression model

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 E_i$$

where $E_i = 1$ if individual exposed and 0 if not. Here

$Y_i \sim$ indep Bernoulli (p_i)

where

$$p_i = pr(Y_i = 1 \mid E_i)$$

$$= pr(\text{develop disease} \mid \text{exposure status})$$

Note that if $E_i = 1$ then

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \equiv \text{log odds of disease given exposed}$$

whereas if $E_i = 0$ then

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 \equiv \text{log odds of disease given not exposed}$$

Thus

$$\beta_1 = \text{log odds disease given exposed} - \text{log odds disease given not exposed}$$

or equivalently

$$\exp(\beta_1) = OR \qquad [\text{or } \beta_1 = \log(OR)]$$

as defined on previous page.

A consequence here is that inferences on the OR can be based on formulating a simple logistic regression model with a binary predictor.

Now let us consider a more complicated model that also includes a continuous covariate X :

$$\log\left(\frac{P_i}{1-p_i}\right) = \beta_0 + \beta_1 E_i + \beta_2 X_i$$

on the logit scale we have 2 parallel lines, one for each Exposure level : If $E_i = 1$

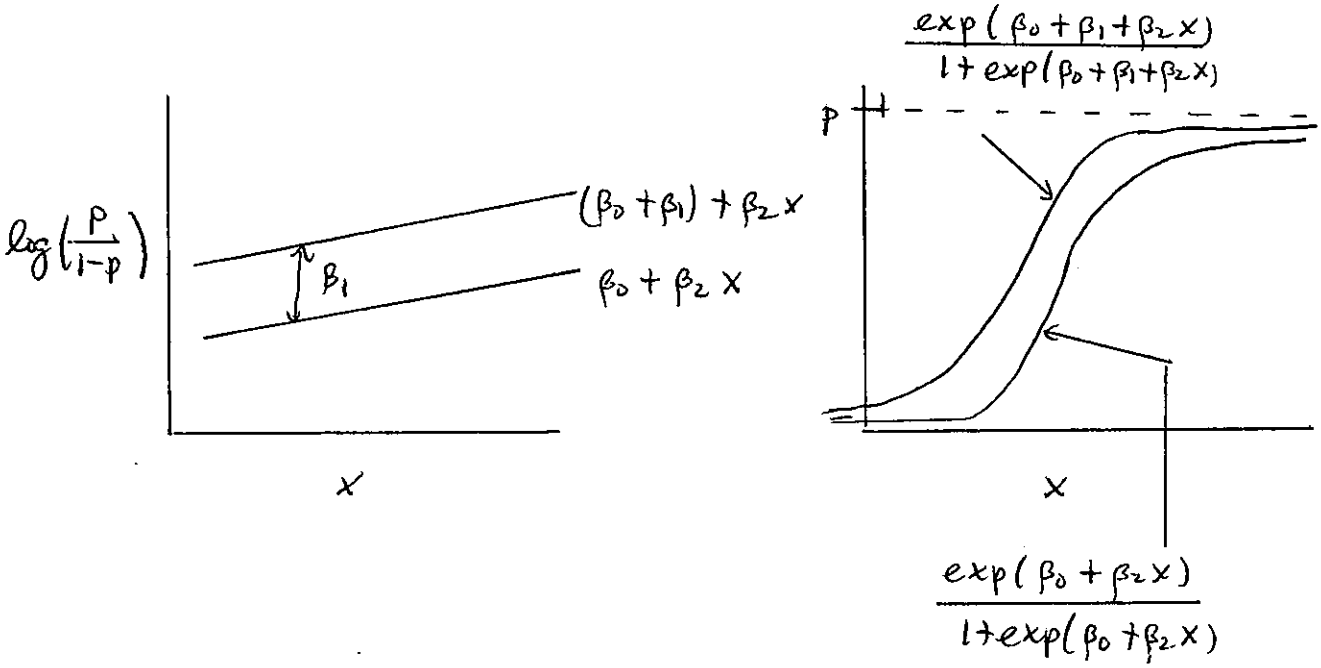
$$\log\left(\frac{P_i}{1-p_i}\right) = \beta_0 + \beta_1(1) + \beta_2 X_i = \underbrace{(\beta_0 + \beta_1)}_{\text{intercept}} + \underbrace{\beta_2 X_i}_{\text{slope}}$$

whereas if $E_i = 0$

$$\log\left(\frac{P_i}{1-p_i}\right) = \beta_0 + \beta_1(0) + \beta_2 X_i = \beta_0 + \beta_2 X_i$$

The coefficient β_1 for Exposure measure the change in intercepts between Exposed ($E=1$) and non-Exposed individuals ($E=0$), whereas β_0 is the intercept for non-exposed individuals - the "baseline group" to which other groups are compared. The slopes are identical.

If we assume $\beta_1 > 0$ and $\beta_2 > 0$ then plots on the logit and probability scale might look like



The curves are not parallel on the probability scale, but "order" is preserved.

Note that the model is a logistic regression analog of ANCOVA model

We will see later that adding an <u>interaction</u> or product term

$$\log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 E_i + \beta_2 X_i + \beta_3 (E_i * X_i)$$

implies a structure where each Exposure group has its own intercept and slope.

In the ANCOVA model (dropping subscripts)

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 E + \beta_2 X$$

both β_1 & β_2 can be interpreted as <u>adjusted log OR</u>. In particular, consider fixing X and varying E from 0 to 1

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 + \beta_2 X \quad \text{if } E=1$$

$$= \beta_0 + \beta_2 X \quad = 0$$

Then β_1 is the difference in log odds of developing disease between exposed & unexposed individuals with the same value of X. This is called the adjusted log OR. Exponentiating β_1 gives the adjusted OR for exposed vs non-exposed individuals, holding X fixed.

Similarly if we fix E and vary x to $x+1$

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 E + \beta_2 (x+1)$$

$$= \underbrace{\beta_0 + \beta_1 E + \beta_2 x}_{\text{log odds when } X=x} + \underbrace{\beta_2}_{\uparrow}$$

log odds when $X=x+1$

log odds when $X=x$

increase in log odds when $X=x \to X=x+1$ holding E fixed

Thus β_2 is the increase in log-odds of developing the disease for each increase in 1 unit of X, holding E fixed. This is adjusted log OR for X, and $\exp(\beta_2)$ is the corresponding adjusted OR.

Remark: In the model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 E$$

the odds-ratio $\exp(\beta_1)$ is said to be the unadjusted OR - it measures association between exposure & disease without consideration of other factors.

Adjusted OR's are ORs obtained from multi-variable models, which adjust effects relative to other factors included in the model. To be clear we need to always specify what other effect are included in model

i.e. in ANCOVA model $\exp(\beta_1)$ is adjusted OR for exposure, adjusting for X (only).

Similarly, regression coefficient in more complex models

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

Can be interpreted as adjusted OR's provided the coefficient under consideration, say β_h, coresponds to an effect X_h that is not part of another effect - i.e. can't fix levels of all other effect and at the same time increase X_h by 1.

Bioassay

Binomial response models, and in particular probit regression more so than logistic regression were developed extensively for bioassay problems in which groups of bugs, animals etc were exposed to different dosages of a drug and the proportion of "responders" was recorded.

Suppose for simplicity a randomly selected rat is given dose X

Let

$$Y = \begin{cases} 1 & \text{if rat dies} \\ 0 & \text{else} \end{cases}$$

and define

$$p = Pr(Y=1 \mid X)$$

A probit model relating p to X has the form

$$\Phi^{-1}(p) = \beta_0 + \beta_1 x$$

or equivalently

$$p = \Phi(\beta_0 + \beta_1 x)$$

where

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp(-.5 z^2)\, dz$$

is the $N(0,1)$ cdf.

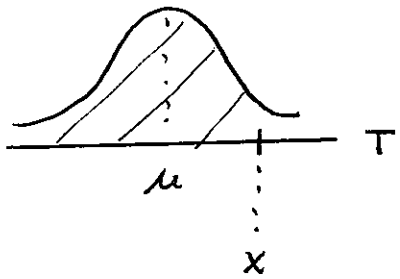
This model can be motivated through the concept of a tolerance distribution. Suppose there exists a tolerance T which is the maximum drug dose a rat can take before dying.

That is if

$T = t$ then $x > t \Rightarrow$ rat dies $y = 1$

$x \leq t \Rightarrow$ rat live $y = 0$

If the distribution of tolerances is $N(\mu, \sigma^2)$ for example



Then

$$p = Pr(y=1|x) = Pr(\text{rat dies given dose } x)$$

$$= Pr(\text{rat's tolerance } T < x)$$

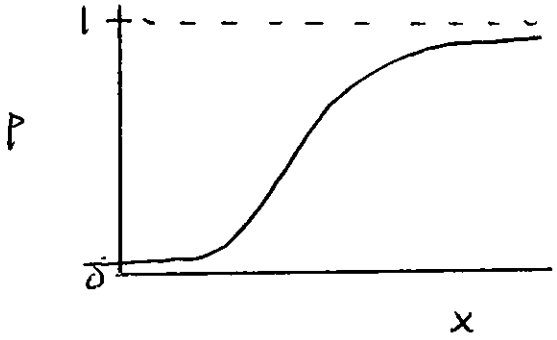
$$= \Pr\left(\frac{T-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right)$$

$$= \Phi\left(\frac{x-\mu}{\sigma}\right)$$

$$= \Phi(\beta_0 + \beta_1 x)$$

where $\beta_0 = -\mu/\sigma$ and $\beta_1 = 1/\sigma$. This is in the form of a probit regression model with intercept $-\mu/\sigma$ and slope parameter $1/\sigma$.

Note that on the probability scale, a graph of p vs x is sigmoidal



$$p = \Phi(\beta_0 + \beta_1 x)$$

The tolerance-based derivation of probit regression requires $\beta_1 > 0$ but in general $\beta_1 < 0$ is possible within this probit model.

Similarly, a logistic regression model arises if we assume T has a logistic distribution

$$T \sim \text{Logistic}(\mu, \sigma)$$

with density

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

Mean $= \mu$

Variance $= \frac{\pi^2}{3}\sigma^2$

and cdf

$$F_T(x) = \frac{\exp\left(+\left(\frac{x-\mu}{\sigma}\right)\right)}{1+\exp\left(\frac{x-\mu}{\sigma}\right)}$$

The logistic is symmetric, and very similar to the normal except with slightly heavier (?) tails. Consequently logistic & probit fits to the same data are often similar, unless the fitted probabilities are extreme (near 0 or 1)

To see the connection of the logistic tolerance distribution to logistic regression, note as with probit model

$$p = \Pr(Y=1 \mid x) = \Pr(T \le x)$$

$$= F_T(x)$$

$$= \frac{\exp\left(\frac{x-\mu}{\sigma}\right)}{1+\exp\left(\frac{x-\mu}{\sigma}\right)}$$

$\beta_0 = -\mu/\sigma$

$\beta_1 = 1/\sigma$

$$= \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$$

or equivalently

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

More generally, within the tolerance distribution approach, if we assume T has cdf $F_T(x)$ then as written on previous page

$$p = Pr(Y=1 \mid x) = F_T(x)$$

If T has a location-scale family distribution (center $= \mu$, scale $= \sigma$) then

$$p = F_T(x) = Pr(T \le x)$$

$$= Pr\left(\frac{T-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right)$$

cdf for standardized T

$$= F_{T_0}\left(\frac{x-\mu}{\sigma}\right) = F_{T_0}(\beta_0+\beta_1 x)$$

Specification of a GLM

- $Y_1, Y_2, \cdots Y_n$ are independent
- $Y_i \sim f(y; \theta_i, \phi_i, w_i)$ an EF distribution that depends on θ_i, ϕ_i and w_i. For simplicity we assume $\phi_i = \phi$ i.e. constant across observations. Note $\phi > 0$ & $w_i > 0$
- $E(Y_i) = \mu_i = b'(\theta_i)$ satisfies

$$g(\mu_i) = \eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

$$= [x_{i1}\ x_{i2} \cdots x_{ip}] \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$= x_i' \beta$$

for some <u>known</u> link function $g(\cdot)$ and known covariate vector x_i'. Note that distributions are <u>conditional</u> on x_i (fixed covariates)

There is no constraint on $g(\cdot)$, except an assumption that it is increasing and differentiable. Standard choices depend on the distribution.

- η_i is called the "linear predictor" - it is linearly related to β, while the mean is linked to the linear predictor through the link function.

Remark: Some texts define the model via

$$\mu_i = f(x_i'\beta) = f(\eta_i)$$

so necessarily $f(\cdot)$ is the "inverse" link function.

example: The standard linear regression model

$$Y_j \sim \text{indep } N(\mu_j, \sigma^2)$$

$$\mu_j = x_j'\beta$$

$\phi = \sigma^2$

is a GLM with identity link $g(\mu_j) = \mu_j = \eta_j$ $\quad w_j = 1$

Alternatively, if we assume

$$\log(\mu_j) = x_j'\beta$$

then we have a GLM with a log-link: $g(\mu_j) = \log(\mu_j)$

example: For Poisson data

$\phi = 1$

$$Y_j \sim \text{Poisson}(\mu_j) \qquad w_j = 1$$

a common GLM uses the log link

$$\log(\mu_j) = x_j'\beta$$

which is equivalent to

$$\mu_j = \exp(x_j'\beta)$$

<u>example</u> Fu Binomial data ($\phi = 1$ $w_j = n_j$ = sample size)

$$Y_j \sim Bin(n_j, \mu_j) \quad \text{or} \quad Y_j^* = Y_j/n_j \sim Bin(n_j, \mu_j)/n_j$$

a glm has the form

$$g(\mu_j) = x_j'\beta$$

for some link function. Common choices are

$$g(\mu_j) = \begin{cases} \log\left(\frac{\mu_j}{1-\mu_j}\right) & \text{logit link: logistic regression} \\ \Phi^{-1}(\mu_j) & \text{probit link: Probit regression} \\ \log\{-\log(1-\mu_j)\} & \text{complementary log-log link} \end{cases}$$

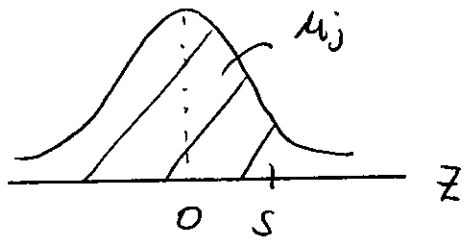
Here

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} \exp(-.5z^2)\, dz$$

is the standard normal distribution function, so $\Phi^{-1}(\mu_j)$ is the "percentile function"

i.e. if $s = \Phi^{-1}(\mu_j)$ then

$$\mu_j = \Phi(s)$$



Fu these 3 choices of $g(\cdot)$, the transformation allows

all possible values of

$$-\infty < g(\mu_j) < +\infty \quad \text{for} \quad 0 < \mu_j < 1$$

This is a natural requirement for transforming a probability in a regression model, otherwise we would need to constrain the values of β in

$$g(\mu_j) = x_j'\beta$$

I'll elaborate in words - but a simple way to see this is to suppose $g(\mu_j) = \mu_j$.

Examples from Handouts

- O-ring failures (SAS logistic regression II)

  Space Shuttle has 6 O-rings

  data on 23 pre-Challenger flight

$$Y_j = \begin{cases} 1 & \text{if } \geq 1 \text{ O-ring fails during flight } (j) \\ 0 & \text{else} \end{cases}$$

covariate information : $Temp_j$ = temperature at lift off for flight j

$press_j$ = pressure on O-ring joints during " "

Initial model

$$Y_j \sim \text{indep Bernoulli}(\mu_j)$$

with

$$\log\left(\frac{\mu_j}{1-\mu_j}\right) = \beta_0 + \beta_1 \text{Temp}_j + \beta_2 \text{Pressure}_j$$

$$= \begin{bmatrix} 1 & \text{Temp}_j & \text{pressure}_j \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = x_j'\beta$$

- UNM trauma data (SAS logistic regression II)
- Budworm experiment (SAS logistic regression I)

Batches of 20 moths (either all Male or all female) exposed to dose of cypermethrin. The number of 20 that were "knocked down" or dead 3 days after treatment was recorded.

Initial model

$$Y_j = \#\text{respond in batch } j \sim \text{Bin}(20, \mu_j)$$

(n_j = 20)

where

$$\log\left(\frac{\mu_j}{1-\mu_j}\right) = \beta_0 + \beta_1 \text{sex}_j + \beta_2 \text{LD}_j + \beta_3 \text{sex}_j * \text{LD}_j$$

(product)

- sex_j: sex in batch j
- LD_j: log-dose in batch j
- $\text{sex}_j * \text{LD}_j$: sex-by-logdose interaction

Here

$$sex_j = \begin{cases} 1 & \text{Females } (F) \\ 0 & \text{Males } (M) \end{cases}$$

The model implies that M & F have separate logistic regression models.

To see this note that if $sex_j = 0$ (Males) then

$$\log\left(\frac{\mu_j}{1-\mu_j}\right) = \beta_0 + \beta_1(0) + \beta_2 LD_j + \beta_3(0 * LD_j) = \beta_0 + \beta_2 LD_j$$

whereas if $sex_j = 1$ (Females), then

$$\log\left(\frac{\mu_j}{1-\mu_j}\right) = \beta_0 + \beta_1(1) + \beta_2 LD_j + \beta_3 \overbrace{(1 * LD_j)}^{LD_j}$$

$$= (\beta_0 + \beta_1) + (\beta_2 + \beta_3) LD_j$$

Thus, β_0 & β_2 are the intercept & slope for Males and

$(\beta_0 + \beta_1)$ & $(\beta_2 + \beta_3)$ are the corresponding intercepts & slopes for Females. In essence, the coefficient β_1 for sex effect is the <u>difference in intercepts</u> between the baseline sex Males (sex = 0) and Females and the coefficient β_3 for the sex-by-LD interaction is the difference in slopes between Males & Females.

I'll draw some pictures (soon) to make this clearer.

° Suicides (SAS Poisson Regression Handout)

Y_{ij} = # Suicides of type i in year j in Great Britain

$i = 1, 2, \cdots, 7$ $\quad$ $j = 1, 2, \cdots 8$

Initial model

$$Y_{ij} \overset{\text{ind}}{\sim} \text{Poisson}(\mu_{ij})$$

with

$$(A) \quad \log \mu_{ij} = \mu + \alpha_i + \beta_j$$

α_i: effect of Type i $\quad$ β_j: effect of Year j

i.e. additive two-way ANOVA type model for log means, a Poisson GLM with log link.

Remark: ANOVA notation for models typically "hides" the connection to a regression structure because not all parameters appear directly in (A). That is because the implied predictors are binary and many effects are zeroed out. In particular, (A) is equivalent

to

$$\log \mu_{ij} = \begin{bmatrix} 1 & \vdots & 0\,0\cdots 0 & 1 & 0\cdots 0 & \vdots & 0\cdots 0 & 1 & 0\cdots 0 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_7 \\ \beta_1 \\ \vdots \\ \beta_8 \end{bmatrix}$$

in i^{th} spot, corresponds to α_i

corresponds to μ

j^{th} spot, corresponds to β_j

$$= x_{ij}' \beta$$

This model has 1 effect for μ

7 binary effects for $\alpha_1 \cdots \alpha_7$ (1 for each leve)

8 binary effects for $\beta_1 \cdots \beta_8$ (ditto)

as such the parameters are not identifiable without adding 1 constraint on the α_i's and 1 on the β_j's - usually set one of each to zero, leading to baseline categories (as in Bud worm experiment - more later)

Canonical Link

In the EF framework, the mean is written as a function of θ_i

$$E(Y_i) = \mu_i = b'(\theta_i)$$

Whereas in the glm framework the mean is modelled in terms

of predictors

$$g(\mu_i) = \eta_i = x_i'\beta$$

Necessarily, if we write the model in terms of θ_i, then θ_i is a function of μ_i, or η_i or β.

The special case where

$$\theta_i = g(\mu_i)$$

corresponds to a GLM with a "canonical link function".

For example

Normal response	θ_i	$= \mu_i$
Poisson response	θ_i	$= \log(\mu_i)$
Binomial response	θ_i	$= \log\left(\frac{\mu_i}{1-\mu_i}\right)$

Thus, the canonical link for normal responses is the <u>identity function</u>

Poisson	log
Binomial	logit

Likelihood function

We will use Maximum Likelihood (ML) for GLM parameter estimates.

I'll assume $Y_1, Y_2, \dots Y_n$ are independent with EF type distributions

$$f(y_j; \theta_j, \phi, w_j) = \exp\left\{ \frac{y_j \theta_j - b(\theta_j)}{a_j(\phi)} - c_j(y_j, \phi) \right\}$$

- ϕ: same ϕ
- $a_j(\phi)$: ϕ / w_j
- $c_j(y_j, \phi)$: allow to depend on w_j

The log-likelihood function

$$L(\underset{\sim}{y}; \underset{\sim}{\theta}, \phi) = \log \prod_{j=1}^{n} f(y_j; \theta_j, \phi, w_j) = \sum_{j=1}^{n} \log f(y_j; \theta_j, \phi, w_j)$$

$(y_1 \dots y_n)$ $(\theta_1 \dots \theta_n)$

$$\equiv \sum_{j=1}^{n} \ell_j$$

where

$$\ell_j = \frac{y_j \theta_j - b(\theta_j)}{a_j(\phi)} - c_j(y_j, \phi)$$

Remark: following convention, I will leave w_j and ϕ in definitions even though some distributions have either $w_j = 1$ for all j and/or $\phi = 1$. The reason will be explained later.

I will leave L in this general form. However, for certain summaries it is more useful to write the ℓ_j's in terms of the usual parameters, such as means - will do later when needed.

Likelihood Equations

For the moment we will assume ϕ is fixed and known. Then the sole parameter in the model is β. Since each θ_i is linked to β via $\mu_j = b'(\theta_j)$ where $g(\mu_j) = \eta_j = x_j'\beta$, we can write the log-likelihood as a function of β

$$L(\beta) = \sum_{j=1}^{n} \ell_j(\beta) = \sum_{j=1}^{n} \left[\frac{y_j \theta_j(\beta) - b(\theta_j(\beta))}{a_j(\phi)} - c_j(y_j, \phi) \right]$$

suppress dependence on y.

$$= \sum_{j=1}^{n} \frac{y_j \theta_j(\beta) - b(\theta_j(\beta))}{a_j(\phi)} + \text{constant}$$

constant: independent of β

To find the MLE of β, we solve the score equation

$$\frac{\partial L(\beta)}{\partial \beta} = 0_{p \times 1}$$

$$\frac{\partial L(\beta)}{\partial \beta} = \begin{bmatrix} \partial L/\partial \beta_1 \\ \partial L/\partial \beta_2 \\ \vdots \\ \partial L/\partial \beta_p \end{bmatrix}$$

To get the score equation, let us do some preliminary calculations.

First, suppose we have 2 functions

$$y = f(x) \quad \text{and} \quad x = h(y)$$

that are inverses of each other, i.e. $h = f^{-1}$.

Then, using standard notation

$$y = f(x) \quad \Rightarrow \quad \frac{\partial y}{\partial x} = f'(x)$$

$$x = h(y) \quad \Rightarrow \quad \frac{\partial x}{\partial y} = h'(y)$$

But also (chain rule)

$$y = f(h(y)) \Rightarrow \frac{\partial}{\partial y}(y) = 1$$

$$= \frac{\partial}{\partial y} f(h(y)) = f'(h(y))\, h'(y)$$

$$\Rightarrow \quad h'(y) = \frac{1}{f'(h(y))}$$

This is the well-know formula for the derivative of the inverse of a function. Put another way

$$h'(y) = 1/f'(h(y)) = 1/f'(x)$$

or

$$\frac{\partial x}{\partial y} = \frac{1}{\partial y/\partial x}$$

or $$\frac{\partial x}{\partial y}\frac{\partial y}{\partial x} = 1 \qquad \text{(not surprising!)}$$

We will use this result a number of times.

For an EF model we know (suppressing subscripts on μ_i, θ_i, n_i etc)

- $\mu = b'(\theta) \Rightarrow \frac{\partial \mu}{\partial \theta} = b''(\theta) = V(\mu)$ — variance function

$$\Rightarrow \boxed{\frac{\partial \theta}{\partial \mu} = \frac{1}{V(\mu)}}$$

For a GLM

- $g(\mu) = \eta \quad \Rightarrow \quad \frac{\partial \eta}{\partial \mu} = g'(\mu)$

$$\Rightarrow \boxed{\frac{\partial \mu}{\partial \eta} = \frac{1}{g'(\mu)}} \qquad [\, g(\,) \text{ is strictly } \uparrow\,]$$

- $\eta = x'\beta = x_1\beta_1 + \cdots + x_p\beta_p \quad \Rightarrow \quad \frac{\partial \eta}{\partial \beta_j} = x_j$

(x_j: a scalar here (j^{th} predictor))

Putting these last 3 results together gives us

$$\frac{\partial \theta(\beta)}{\partial \beta_j} = \frac{\partial \theta}{\partial \beta_j} = \frac{\partial \theta}{\partial \mu}\frac{\partial \mu}{\partial \eta}\frac{\partial \eta}{\partial \beta_j} \qquad \text{(chain rule)}$$

($\frac{\partial \theta}{\partial \beta_j}$: shorthand)

$$= \frac{1}{V(\mu)} \cdot \frac{1}{g'(\mu)} \cdot x_j \qquad (B)$$

This is the fundamental calculation needed for us.

Now with (using i as subscript for observation)

$$L(\beta) = \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \{ y_i \theta_i(\beta) - b(\theta_i(\beta)) \} + \text{constant}$$

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \left\{ y_i \frac{\partial \theta_i}{\partial \beta_j} - b'(\theta_i)\frac{\partial \theta_i}{\partial \beta_j} \right\} \qquad \text{chain rule}$$

$$= \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \left(y_i - b'(\theta_i) \right) \frac{\partial \theta_i}{\partial \beta_j}$$

$$= \sum_{i=1}^{n} \frac{1}{a_i(\phi)} (y_i - \mu_i) \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$

$$= \sum_{i=1}^{n} \frac{1}{a_i(\phi)} (y_i - \mu_i) \frac{1}{V(\mu_i)} \frac{1}{g'(\mu_i)} x_{ij}$$

x_{ij}: j^{th} predictor on i^{th} observation

Matrix notation helps here. Let

$$\underset{n\times 1}{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \underset{n\times 1}{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} \qquad \underset{n\times p}{X} = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix}$$

i^{th} row is x_i'

and define W and Δ to be $n \times n$ diagonal matrices with diagonal elements

$$w_{ii} = \frac{1}{a_i(\phi)} \cdot \frac{1}{V(\mu_i)} \frac{1}{\{g'(\mu_i)\}^2} \qquad \text{(for } W\text{)}$$

$$\delta_{ii} = g'(\mu_i) \qquad \text{(for } \Delta\text{)}$$

Then it is easy to see

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij} \underbrace{v_{ii} w_{ii} (y_i - \mu_i)}$$

i^{th} element of $n \times 1$ vector $W \Delta (y - \mu)$

x_{ij}: i^{th} element of j^{th} column of X or j^{th} row of X'

$$= [j^{th} \text{ row of } X'] \, W \Delta (y - \mu)$$

Thus

$$\frac{\partial L}{\partial \beta} = \begin{bmatrix} \partial L/\partial \beta_1 \\ \partial L/\partial \beta_2 \\ \vdots \\ \partial L/\partial \beta_p \end{bmatrix} = X' W \Delta (y - \mu)$$

(impt) useful form only depends on data, means and variances, not distributions or θ's

The score or likelihood equation is

$$\boxed{\frac{\partial L}{\partial \beta} = 0_{p\times 1} \iff X' W \Delta (y - \mu) = 0} \qquad (c)$$

In this equation W, Δ and μ are functions of β! This result allows for weights w_i and a scale ϕ, even if they are not usually present!

Standard iterative methods (Fisher Scoring, Newton Raphson, Iteratively Reweighted LS) can be used to solve the likelihood equations to get the MLE

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}$$

Large sample theory for GLMs can be used to establish that (under suitable regularity conditions)

$$\hat{\beta} \dot{\sim} N_p(\beta, I^{-1}(\beta))$$

where $I(\beta)$ is the expected Fisher information matrix

Remark: This says $\hat{\beta}_i \dot{\sim} N(\beta_i, \sigma_{ii})$ where σ_{ii} is i^{th} diagonal element of $I^{-1}(\beta)$. More generally $\hat{\beta}$ is Multivariate normal (approx) with mean vector β and covariance matrix $I^{-1}(\beta)$.

Here

$$I(\beta) = E\left\{ \underbrace{\frac{\partial L}{\partial \beta}\left(\frac{\partial L}{\partial \beta}\right)'}_{\substack{p \times p \text{ symmetric matrix with } i^{th} \text{ row } \& \\ j^{th} \text{ col element}}} \right\}$$

$$= -E\left\{ \frac{\partial^2 L}{\partial \beta \, \partial \beta'} \right\}$$

$\frac{\partial L}{\partial \beta_i}\frac{\partial L}{\partial \beta_j}$ (top)

$-\frac{\partial^2 L}{\partial \beta_i \partial \beta_j}$ (bottom)

expectation is taken elementwise.

The first expression for $I(\beta)$ is easy to evaluate using standard moment results. If

$$Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_p \end{pmatrix}$$

is a random vector with mean vector

$$\mu_Z = \begin{pmatrix} \mu_{Z_1} \\ \vdots \\ \mu_{Z_p} \end{pmatrix}$$

and covariance matrix

$$E(Z-\mu_Z)(Z-\mu_Z)' = \Sigma_Z = [\sigma_{ij}]$$ (i^{th} row & j^{th} element.)

where $\sigma_{ij} = cov(Z_i, Z_j) \quad i \neq j$

$= var(Z_i) \quad i = j$

and A is an $r \times p$ matrix of constants, then the rv

$$V = \begin{bmatrix} V_1 \\ \vdots \\ V_r \end{bmatrix} = AZ$$

so that

$$v_i = \sum_j a_{ij} z_j$$

has

$$E(V) = E(AZ) = A E(Z) = A \mu_Z$$

and

$$cov(V) = cov(AZ) = A\, cov(Z)\, A' = A \Sigma_Z A'$$

To get $I(\beta)$ note

$$\frac{\partial L}{\partial \beta} = X' W \Delta (y-\mu) \Rightarrow \left(\frac{\partial L}{\partial \beta}\right)' = (y-\mu)' \Delta' W' X$$
$$= (y-\mu)' \Delta W X$$

since W & Δ are symmetric, so

$$E\left\{\frac{\partial L}{\partial \beta}\left(\frac{\partial L}{\partial \beta}\right)'\right\} = E\{ X' W \Delta (y-\mu)(y-\mu)' \Delta W X \}$$

$$= X' W \Delta\, E(y-\mu)(y-\mu)' \Delta W X$$

$$= X' W \Delta\, cov(y)\, \Delta W X$$

Now, the $y_1, y_2, \cdots y_n$ are independent, so

$$cov(y_i, y_j) = 0 \quad \text{for } i \neq j$$

$$var(y_i) = cov(y_i, y_i) = b''(\theta_i) \frac{\phi}{\omega_j}$$

$$= V(\mu_i)\, a_j(\phi)$$

$$= \frac{1}{\omega_{ii}} \cdot \frac{1}{\{g'(\mu_i)\}^2}$$

↑ see p39

= i^{th} diagonal element of W^{-1} * (i^{th} diagonal element of Δ^{-1})^2

That is

$$cov(Y) = \Delta^{-1} \omega^{-1} \Delta^{-1}$$

and so

$$I(\beta) = E\left\{ \frac{\partial L}{\partial \beta} \left(\frac{\partial L}{\partial \beta}\right)' \right\} = X' \omega \Delta \, cov(Y) \, \Delta \omega X$$

$$= X' \omega \Delta \Delta^{-1} \omega^{-1} \Delta^{-1} \Delta \omega X$$

$$= X' \omega X$$

Consequently

$$\boxed{\hat{\beta} \dot{\sim} N(\beta, (X' \omega X)^{-1})}$$

As before, this allows for weights ω_j and scale ϕ even if not usually present!

example: Standard linear regression

$$y_j \sim \text{indep } N(\mu_j, \sigma^2) \qquad \mu_j = x_j'\beta$$

Here: $\phi = \sigma^2$, $\omega_j = 1$, $g(\mu_j) = \mu_j$ (identity link) $\quad a_j(\phi) = \sigma^2$

$$V(\mu_j) = 1$$

$$g'(\mu_j) = 1$$

W has diagonal element $w_{jj} = \dfrac{1}{a_j(\phi)} \dfrac{\omega}{V(\mu_i)} \dfrac{1}{\{g'(\mu_i)\}^2} = \dfrac{1}{\sigma^2}$

Δ " " " $\quad d_{jj} = g'(\mu_i) = 1$

Thus $\Delta = I_n$ and $W = \dfrac{1}{\sigma^2} I_n$ (I_n = $n \times n$ identity matrix)

which implies

$$\mu_i = x_i'\beta \iff \mu = X\beta$$

$$\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix} \beta$$

$$\frac{\partial L}{\partial \beta} = X' W \Delta (y - \mu)$$

$$= \frac{1}{\sigma^2} X'(y - \mu)$$

$$= \frac{1}{\sigma^2} X'(y - X\beta) = 0 \iff X'y = X'X\beta \quad \text{Normal eqns}$$

$$\iff \hat{\beta} = (X'X)^{-1} X'y$$

and

$$I(\beta) = X' W X = \frac{1}{\sigma^2} X'X$$

so

$$\hat{\beta} \dot{\sim} N(\beta, \sigma^2 (X'X)^{-1})$$

$\dot{\sim}$: exact

standard result!

ex: Poisson regression, log link

Y_j ~ indep Poisson (μ_j) $\quad \log \mu_j = x_j'\beta$

write

$$\log(\mu) = \begin{bmatrix} \log \mu_1 \\ \vdots \\ \log \mu_n \end{bmatrix} = \begin{bmatrix} x_1'\beta \\ \vdots \\ x_n'\beta \end{bmatrix} = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix} \beta = X\beta$$

Here

$g(\mu_j) = \log(\mu_j) \Rightarrow g'(\mu_j) = \frac{1}{\mu_j}$

$V(\mu_j) = \mu_j$

$a_j(\phi) = \frac{\phi}{\omega_j}$ will consider $\omega_j = 1$ and leave ϕ here so $a_j(\phi) = \phi$

this allows for "overdispersion" whereas exact Poisson case has $\phi = 1$!

Diagonal elements W: $w_{jj} = \frac{1}{a_j(\phi)} \cdot \frac{1}{V(\mu_i)} \cdot \frac{1}{\{g'(\mu_i)\}^2} = \frac{\mu_j}{\phi}$

$$\Delta : v_{jj} = g'(\mu_j) = \frac{1}{\mu_j}$$

Thus $W\Delta = \frac{1}{\phi} I_n$ and

Likelihood equation: $\frac{\partial L}{\partial \beta} = X'W\Delta(y-\mu) = \frac{1}{\phi} X'(y-\mu)$

$= \frac{1}{\phi} X'(y - \exp(X\beta)) = 0$

here $\exp(X\beta)$ means elementwise exponentiation of vector $X\beta$

$\Leftrightarrow X'(y - \exp(X\beta)) = 0$

Also: $I(\beta) = X'WX = \frac{1}{\phi} X' D(\mu) X$ diagonal matrix of means

Remark

Likelihood equations $X'(y-\mu) = 0$ has same form as for normal linear regression. However $\mu = \exp(X\beta)$ is a non-linear function of β so MLE has to be obtained iteratively for Poisson case

ex: Binomial response model

$$y_j \sim \text{ind } \text{Bin}(n_j, \mu_i) \qquad i = 1, 2, \cdots, n$$

$$g(\mu_i) = x_i' \beta$$

with

$$\mu = (\mu_1, \mu_2, \cdots, \mu_n)'$$

write

$$g(\mu) = \begin{bmatrix} g(\mu_1) \\ \vdots \\ g(\mu_n) \end{bmatrix} = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix} \beta = X\beta$$

Here $\omega_j = n_j$

$$V(\mu_j) = \mu_j(1-\mu_j) \qquad a_j(\phi) = \frac{\phi}{\omega_j} = \frac{\phi}{n_j}$$

Note $\phi = 1$ for Binomial, but will leave in expression here.

Diagonal elements W: $\omega_{jj} = \dfrac{1}{a_j(\phi)} \cdot \dfrac{1}{V(\mu_i)} \dfrac{1}{\{g'(\mu_i)\}^2} = \dfrac{n_j}{\phi} \cdot \dfrac{1}{\mu_j(1-\mu_j)} \cdot \dfrac{1}{[g'(\mu_i)]^2}$

Δ $\quad v_{jj} = g'(\mu_j)$

For logistic regression

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) \Rightarrow g'(\mu_i) = \frac{1}{\mu_i} + \frac{1}{1-\mu_i}$$

$$= \log \mu_i - \log(1-\mu_i) = \frac{1}{\mu_i(1-\mu_i)}$$

For probit regression

$$g(\mu_i) = \Phi^{-1}(\mu_i) \Rightarrow g'(\mu_i) = \frac{1}{\Phi'(\Phi^{-1}(\mu_i))}$$

$$= \frac{1}{f(\Phi^{-1}(\mu_i))}$$

where

$$f(t) = \frac{1}{\sqrt{2\pi}} \exp(-.5t^2) \qquad -\infty < t < +\infty$$

is $N(0,1)$ density.

Thus

$$w_{jj} = \begin{cases} \frac{n_j}{\phi} \frac{1}{\mu_j(1-\mu_j)} \mu_j^2(1-\mu_j)^2 = \frac{n_j}{\phi} \mu_j(1-\mu_j) & \text{Logistic} \\ \frac{n_j}{\phi} \frac{1}{\mu_j(1-\mu_j)} \left[f(\Phi^{-1}(\mu_i)) \right]^2 & \text{Probit} \end{cases}$$

$$v_{jj} = \begin{cases} \frac{1}{\mu_i(1-\mu_i)} & \text{Logistic} \\ \frac{1}{f(\Phi^{-1}(\mu_i))} & \text{Probit} \end{cases}$$

and

$$w_{jj} v_{jj} = \begin{cases} \frac{n_j}{\phi} = \frac{1}{a_j(\phi)} & \text{Logistic} \\ \frac{n_j}{\phi} \frac{1}{\mu_j(1-\mu_j)} f(\Phi^{-1}(\mu_j)) & \text{Probit} \end{cases}$$

$\uparrow$

diag elts WA

Recall that the likelihood equations and information matrices have the forms:

$$X' W \Delta (y - \mu) = 0$$

and

$$I(\beta) = X' W X$$

Note that for logistic regression $W\Delta$ is a diagonal matrix with elements $n_i/\phi = 1/a_i(\phi)$, which is independent of μ, and W has a simple form of $V(\mu_i)/a_i(\phi)$. This is, in fact exactly what we saw for normal theory regression and Poisson regression with a log link, because logistic regression and the other two models are based on <u>canonical link functions</u>.

To see the simplification that results with canonical links, recall that (from p39)

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} \frac{1}{a_i(\phi)} (y_i - \mu_i) \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$

$$= \sum_{i=1}^{n} \frac{1}{a_i(\phi)} (y_i - \mu_i) \frac{1}{V(\mu_i)} \frac{1}{g'(\mu_i)} x_{ij}$$

But with a <u>canonical link</u>, we have

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

So we must have

$$\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{V(\mu_i)}\frac{1}{g'(\mu_i)}$$

$$= \frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \theta_i} = 1$$

i.e.

$$\boxed{\frac{1}{V(\mu_i)} = g'(\mu_i)}$$

With this simplification

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} \frac{1}{a_j(\phi)}(y_i - \mu_i)\frac{\partial \eta_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{1}{a_j(\phi)}(y_i - \mu_i)x_{ij}$$

Since

$$\frac{\partial L}{\partial \beta} = X' W \Delta (y - \mu)$$

it must be that $W\Delta$ has diagonal elements

$$\omega_{jj} v_{jj} = \frac{1}{a_j(\phi)}$$

and W has diagonal element

$$\omega_{jj} = \frac{1}{a_j(\phi)}\frac{1}{V(\mu_i)}\frac{1}{\{g'(\mu_i)\}^2} = \frac{V(\mu_j)}{a_j(\phi)}$$

Other MLEs

At an arbitrary covariate vector x_*, the mean and linear predictor satisfy

$$\eta_* = x_*' \beta \quad \text{and} \quad g(\mu_*) = \eta_* \Rightarrow \mu_* = g^{-1}(\eta_*)$$

Their MLEs are

$$\hat{\eta}_* = x_*' \hat{\beta} \quad \text{and} \quad \hat{\mu}_* = g^{-1}(\hat{\eta}_*)$$

Since

$$\hat{\beta} \dot{\sim} N_p(\beta, I^{-1}(\beta))$$

the large sample distribution of $\hat{\eta}_*$ and $\hat{\mu}_*$ are also normal.

In particular

$$\hat{\eta}_* = x_*' \hat{\beta} \dot{\sim} N(\underbrace{x_*' \beta}_{\eta}, \underbrace{x_*' I^{-1}(\beta) x_*}_{\text{var}(\hat{\eta}_*)})$$

and letting $h(t) = g^{-1}(t)$

$$\hat{\mu}_* = h(\hat{\eta}_*) \dot{\sim} N(\underbrace{h(\eta_*)}_{\mu_*}, \underbrace{[h'(\eta_*)]^2}_{= \left[\frac{1}{g'(h(\eta_*))}\right]^2} \text{var}(\hat{\eta}_*))$$

$$= \left[\frac{1}{g'(\mu_*)}\right]^2$$

The latter result follows from the Delta Method

This suggests that μ_+ is estimated in 2 steps:

- estimate the linear predictor $\eta_+ = x'\beta$
- transform the linear predictor using the inverse link

$$\mu_+ = g^{-1}(\eta_+)$$

Of course, this could be done in 1 step since $\mu_+ = g^{-1}(x'\beta)$ but in practice CIs for μ are based on inverting the CI for η_+ i.e. the two step process is used in CI estimation - more on this later!

The estimated linear predictors & means for the observed covariates x_i are

$$\hat{\eta}_i = x_i'\hat{\beta} \quad \text{and} \quad \hat{\mu}_i = g^{-1}(\hat{\eta}_i) = g^{-1}(x_i'\hat{\beta}).$$

($\hat{\mu}_i$: Fitted values)

Estimating Information Matrix

Inferences on β require an estimate of

$$I(\beta) = X'\omega X = \begin{cases} E\left(\frac{\partial L}{\partial \beta}\frac{\partial L}{\partial \beta'}\right) \\ E\left(-\frac{\partial^2 L}{\partial \beta \partial \beta'}\right) \end{cases}$$

Several choices are available. The estimated expected Fisher information matrix plugs-in $\hat{\beta}$ into $I(\beta)$:

$$I_E(\hat{\beta}) = X' \hat{\omega} X$$

where $\hat{\omega}$ has diagonal elements

$$\hat{\omega}_{ii} = \frac{1}{a_j(\phi)} \frac{1}{V(\hat{\mu}_i)} \frac{1}{g'(\hat{\mu}_i)^2}$$

Alternative estimate is based on minus the Hessian matrix, which for a GLM can be shown (details omitted!)

$$-H = -\frac{\partial^2 L}{\partial \beta \partial \beta'}$$

$$= X' \omega_0 X$$

where ω_0 is an $n \times n$ diagonal matrix with diagonal elements

$$\omega_{0,ii} = \omega_{ii} + (y_i - \mu_i) \frac{V(\mu_i) g''(\mu_i) + V'(\mu_i) g'(\mu_i)}{[V(\mu_i)]^2 [g'(\mu_i)]^3 a_j(\phi)}$$

Note here that ω_0 depends on <u>data</u> and the means are not estimated (yet)

It is easy to see that

$$E(\omega_{0,ii}) = \omega_{ii}$$

since $E(y_i - \mu_i) = 0$ and thus, as it must be

$$E(-H) = E\left(-\frac{\partial^2 L}{\partial \beta \partial \beta'}\right)$$

$$= X' \omega X = I(\beta)$$

Also note that for <u>canonical links</u>

$$g'(\mu_i) = \frac{1}{V(\mu_i)}$$

so

$$\frac{\partial}{\partial \mu_i} V(\mu_i)\, g'(\mu_i) = V(\mu_i)\, g''(\mu_i) + V'(\mu_i)\, g'(\mu_i)$$

↑ chain rule

$= 0$ for canonical links [which have $V(\mu_i)\, g'(\mu_i) = 1$]

This implies

$$\omega_{0,ii} = \omega_{ii}$$

for <u>canonical links</u> or

$$-H = X' \omega_0 X = X' \omega X = I(\beta)$$

The <u>observed information matrix</u> is

$$I_0(\hat{\beta}) = -H(\hat{\beta}) = X' \hat{\omega}_0 X$$

i.e. minus the Hessian evaluated at $\hat{\beta}$

This result for canonical links easy to get using representation for $\partial l / \partial \beta_j$ given on p 50

In general,

$$I_0(\hat{\beta}) \neq I_E(\hat{\beta})$$

except for canonical links

Noting that $E(-H) = I(\beta)$, the observed information matrix uses $-H$, which is unbiased for $I(\beta)$, but then plugs in $\hat{\beta}$. In particular

$$I_o(\hat{\beta}) = -H(\hat{\beta}) = X' \hat{\omega}_o X$$

where $\hat{\omega}_o$ is a diagonal matrix with entries

$$\hat{\omega}_{o,ii} = \hat{\omega}_{ii} + (y_i - \hat{\mu}_i) \frac{V(\hat{\mu}_i) g''(\hat{\mu}_i) + V'(\hat{\mu}_i) g'(\hat{\mu}_i)}{[V(\hat{\mu}_i)]^2 [g'(\hat{\mu}_i)]^3 a_j(\phi)}$$

There are a number of other choices for estimating $I(\beta)$, but I will stick to these 2 standard choices for now

It is generally accepted that $I_o(\hat{\beta})$ is to be preferred to $I_E(\hat{\beta})$ for estimating the uncertainty in $\hat{\beta}$ (as $I_o(\hat{\beta})$ is related to the observed versus expected curvature of the log-likelihood at $\hat{\beta}$).

I will typically use $I(\hat{\beta})$ to identify $I_o(\hat{\beta})$ or $I_E(\hat{\beta})$.

Inference on β and related quantities

I will first describe some simple inferences based on the distributional approximation

$$\hat{\beta} \dot{\sim} N(\beta, I^{-1}(\hat{\beta}))$$

where $I(\hat{\beta})$ is a version of the estimated information matrix

Let

$$\hat{s_D}(\hat{\beta}_i) = \text{est. std deviation of } \hat{\beta}_i$$
$$= \sqrt{\hat{\sigma}_{ii}}$$

where

$$\hat{\sigma}_{ii} = i^{th} \text{ diagonal element of } I^{-1}(\hat{\beta})$$

A simple test

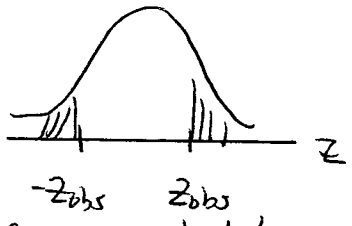
$$H_0: \beta_i = \beta_{i0} \text{ (fixed value)}$$
$$H_A: \neq$$

can be based on

$$z_{obs} = \frac{\hat{\beta}_i - \beta_{i0}}{\hat{s_D}(\hat{\beta}_i)} \dot{\sim} N(0,1) \text{ under } H_0$$

i.e.

$$\text{p-value} \doteq Pr(|Z| \geq |z_{obs}|)$$



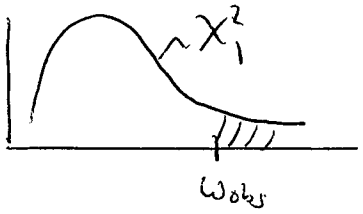
where $Z \sim N(0,1)$. P-values for 1-sided tests are also available.

For a two-sided test it is also common to use the so-called

Wald test statistic

$$w_{obs} \quad z^2_{obs} = \frac{(\hat\beta_i - \beta_{i0})^2}{\hat{SD}^2(\hat\beta_i)}$$

$$= \frac{(\hat\beta_i - \beta_{i0})^2}{\hat\sigma_{ii}} \quad \dot\sim \chi^2_1$$



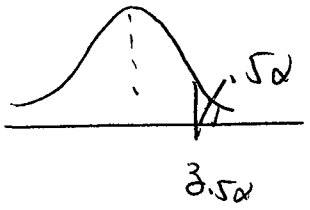
with

$$\text{pvalue} = P(\chi^2_1 \geq w_{obs})$$

This procedure is equivalent to the previous z_{obs}-method.

Similarly, a two-sided CI for β_i is

$$\hat\beta_i \pm z_{.5\alpha}\, \hat{SD}(\hat\beta_i)$$



which has (approximate) coverage of $100(1-\alpha)\%$.

It is also common to compute CI for population means μ_i (or analogously for μ_t at an arbitrary covariate x_t - procedure is same!). To do so, we first calculate a CI for the linear predictor η_i using

$$\hat\eta_i \dot\sim N(\eta_i, \underbrace{\hat{Var}(\hat\eta_i)}_{\equiv \hat{SD}^2(\hat\eta_i)})$$

$$\| \\ x_i' I^{-1}(\hat\beta) x_i$$

i.e. $(1-\alpha)100\%$ CI for η_i is given by

$$\underbrace{\hat{\eta}_i - z_{.5\alpha}\, \hat{SD}(\hat{\eta}_i)}_{\eta_L} \leq \eta_i \leq \underbrace{\hat{\eta}_i + z_{.5\alpha}\, \hat{SD}(\hat{\eta}_i)}_{\eta_U}$$

Then, using $\mu_i = g^{-1}(\eta_i)$ where $g(\cdot)$ and $g^{-1}(\cdot)$ are <u>increasing functions</u>

$$\eta_L \leq \eta_i \leq \eta_U$$

$$\Leftrightarrow \quad \underbrace{g^{-1}(\eta_L)}_{\mu_L} \leq \underbrace{g^{-1}(\eta_i)}_{\mu_i} \leq \underbrace{g^{-1}(\eta_U)}_{\mu_U}$$

i.e. CI for μ_i obtained by "inverting" CI for η_i using inverse link.

<u>Remark</u>: These CI for μ_i are asymmetric about $\hat{\mu}_i$ but generally viewed to work better than CI

$$\hat{\mu}_i \pm z_{.5\alpha}\, \hat{SD}(\hat{\mu}_i)$$

based on normal approx to $\hat{\mu}_i$ given at bottom of p51. If the link is the identity, then the 2 CI are identical.

Most packages will compute CI for η_i and μ_i (using 1ST method!) and corresponding $\hat{SD}$. In SAS, you can set estimates, $\hat{SD}$ and CI at arbitrary x_* by adding a new record to the data set with a missing response and covariate vector x_*.

General Wald Test on β

given the model

$$g(\mu_i) = x_i'\beta$$
$$= \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

suppose we are interested in testing

$$H_0 : L'\beta = 0$$

where L' is a $s \times p$ matrix of constants ($s \leq p$) with linearly independent rows. (L not same as "log-likelihood")

For example, suppose the model includes an intercept β_1 (i.e. $x_{i1} = 1$) so

$$g(\mu_i) = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

We may be interested in whether

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_p = 0$$

holds - i.e. does 1 or more of the predictors affect the mean response. Then

$$L'_{(p-1)\times p} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & & 0 \\ \vdots & \vdots & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & 0 & & \vdots \\ 0 & 0 & 0 & 0 & & 1 \end{bmatrix}$$

i^{th} row has 1 in $(i+1)^{ST}$ spot and zero elsewhere

Similarly, the hypotheses $H_0: \beta_i = 0$ and $H_0: \beta_i = \beta_j$ $(i \neq j)$ for given i and j can be written in this form with

$\Leftarrow H_0: \beta_i - \beta_j = 0$

$$L' = [0\ 0 \cdots 0\ 1\ 0 \cdots 0]$$

(1 in i^{th} spot)

and

$$L' = [0 \cdots 0\ \underset{\uparrow}{1}\ 0 \cdots 0\ \underset{\uparrow}{-1}\ 0 \cdots 0]$$

(1 in i^{th} spot, -1 in j^{th} spot)

respectively.

In general

$$\hat{\beta} \sim N(\beta, I^{-1}(\hat{\beta}))$$

implies

$$L'\hat{\beta} \sim N(L'\beta, L' I^{-1}(\hat{\beta}) L)$$

and

$$L'(\hat{\beta} - \beta) = L'\hat{\beta} - L'\beta \dot{\sim} N(0, L' I^{-1}(\hat{\beta}) L)$$

The <u>quadratic form</u>

$$\{L'(\hat{\beta} - \beta)\}' \{L' I^{-1}(\hat{\beta}) L\}^{-1} \{L'(\hat{\beta} - \beta)\} \dot{\sim} \chi^2_s$$

To test H_0, we note that $L'(\hat{\beta} - \beta) = L'\hat{\beta}$ under H_0

and so

$$W_{obs} = \{L'\hat{\beta}\}' \{L' I^{-1}(\hat{\beta}) L\}^{-1} \{L'\hat{\beta}\}$$

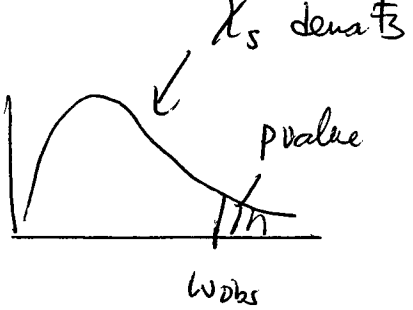
$$= \hat{\beta}' L (L' I^{-1}(\hat{\beta}) L)^{-1} L'\hat{\beta} \dot{\sim} \chi^2_s$$

when H_0 is true. This is the generalized Wald test of $H_0: L'\beta = 0$

In the special case where $H_0: \beta_j = 0$, W_{obs} reduces to the Wald statistic given earlier.

The pvalue for testing H_0 is

$$\text{pvalue} \doteq \Pr(\chi^2_s \geq W_{obs})$$



Measuring goodness-of-fit

- Deviance & Pearson Statistics

Thinking of the log-likelihood as a function of the mean vector

$$\mu = (\mu_1, \cdots \mu_n)'$$

and keeping in mind that $\mu = \mu(\beta)$ is a function of β we can write

$$L(\mu; \phi, Y) = \sum_{j=1}^{n} \left\{ \frac{y_j \theta_j(\mu) - b_j(\mu)}{a_j(\phi)} - c_j(y_j, \phi) \right\}$$

If we fit the model

$$g(\mu_i) = x_i'\beta = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p$$

by ML, the maximized value of the log-likelihood is

$$L(\hat{\mu}; \phi, y) = \sum_{j=1}^{n} \left\{ \frac{y_j \theta_j(\hat{\mu}) - b_j(\hat{\mu})}{a_j(\phi)} - c_j(y_j, \phi) \right\}$$

where $\hat{\mu}$ is the MLE of μ. Note that $\theta_j(\hat{\mu})$ only depends on $\hat{\mu}_j$ and similarly for $b_j(\hat{\mu})$

Now suppose we fit the alternative model

$$g(\mu_i) = x_i'\beta + z_i'\gamma \quad \left(= [x_i' \; z_i'] \begin{pmatrix} \beta \\ \gamma \end{pmatrix} = \tilde{x}_i' \beta_{new} \right)$$

$$= \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \gamma_1 z_{i1} + \cdots + \gamma_r z_{ir}$$

where for now $r = n - p$. We have (n) observations and (n) regression parameters. If there are no linear dependencies among the predictors then this is the so-called <u>saturated model</u>, which places no constraints on $g(\mu_i)$ and consequently no constraints on μ_i

One can show that the MLE of μ for the saturated model is $\tilde{\mu} = y$. From this the MLEs of β and τ could be computed, if of interest, using $g(\tilde{\mu}_i) = x_i'\beta + z_i'\theta$, which is a full-rank linear system. To see that $\tilde{\mu} = y$ note that the likelihood equation under the saturated model is

$$\tilde{X}' W \Delta (y - \mu) = 0_{n \times 1}$$

where $\tilde{X}$ is the $n \times n$ "extended" design matrix with rows $[x_i' \; z_i']$

By assumption, $\tilde{X}$ is invertible, as are W and Δ so

$$\tilde{X}' W \Delta (y - \mu) = 0 \iff y - \mu = 0$$

so $\tilde{\mu} = y$.

Let $L(\tilde{\mu}; \phi, y) = L(y; \phi, y)$ be the maximized log-likelihood for the saturated model, and note that we must have

$$L(\tilde{\mu}; \phi, y) \geq L(\hat{\mu}; \phi, y)$$

because the alternative model is more general [Can you reason why this is true?]

Then, the likelihood ratio statistic for testing $H_0: \tau = 0$ is just

$$LRT = -2 \log \text{likelihood ratio}$$

$$= -2 \{ L(\hat{\mu}; \phi, y) - L(y; \phi, y) \}$$

$$= 2 \{ L(y; \phi, y) - L(\hat{\mu}; \phi, y) \}$$

Because the alternative model is saturated, this is also viewed as a measure of how well the null model $g(\mu_i) = x_i'\beta$ is fitted by the data.

For a GLM, and writing $a_j(\phi) = \phi / w_j$

$$LRT = \frac{1}{\phi} \underbrace{\sum_{j=1}^{n} 2w_j \{ (\theta_j(y) - \theta_j(\hat{\mu})) y_j - (b_j(y) - b_j(\hat{\mu})) \}}_{D(y; \hat{\mu})}$$

where $D(y; \hat{\mu})$ is called the <u>deviance</u> for the model $g(\mu_i) = x_i'\beta$ and

$$LRT = \frac{1}{\phi} D(y; \hat{\mu}) \equiv D^*(y; \hat{\mu})$$

is called the <u>scaled deviance</u>.

For interpretation purposes, it is more convenient to write the deviance in terms of the mean parameters. One can show that (only including weights for Binomial case)

Distribution	Deviance
Binomial $y_j \sim \text{Bin}(n_j, \mu_j)$	$2 \sum_j y_j \log\left(\frac{y_j}{n_j \hat{\mu}_j}\right) + (n_j - y_j) \log\left(\frac{n_j - y_j}{n_j - n_j \hat{\mu}_j}\right)$
Poisson	$2 \sum_j \left\{ y_j \log\left(\frac{y_j}{\hat{\mu}_j}\right) - (y_j - \hat{\mu}_j) \right\} \cdot w_j$
Normal	$\sum_{j=1}^{n} (y_j - \hat{\mu}_j)^2 \cdot w_j$ ← usually 1

For the Poisson, the second term $(y_j - \hat{\mu}_j)$ is sometimes excluded because $\sum_j (y_j - \hat{\mu}_j) = 0$ if model includes an intercept.

The deviance for the normal model is the Residual Sum of Squares.

In general, the larger the deviance the poorer the fit of the model (relative to the saturated model). If the model fits perfectly then $D(y; \hat{\mu}) = 0$. Large values of $D(y; \hat{\mu})$ suggest a general lack-of-fit of the model.

Remarks

(1.) The standard Poisson & Binomial models have $\phi = 1$. For these models the deviance & scaled deviances are identical.

(1.5) For the Binomial, we can also write Deviance as (with $y_j^* = y_j / n_j$)

$$2 \sum_j n_j \left\{ y_j^* \log\left(\frac{y_j^*}{\hat{\mu}_j}\right) + (1 - y_j^*) \log\left(\frac{1 - y_j^*}{1 - \hat{\mu}_j}\right) \right\}$$

(2.) In certain settings the scaled deviance

$$D^*(y;\hat{\mu}) \dot{\sim} \chi^2_{n-p}$$

$n-p$: # parameters difference between null & saturated models.

For the normal model this result is <u>exact</u>, but of not much practical use since $\phi = \sigma^2$ is typically unknown

For the <u>Binomial</u> model $Y_j \sim$ indep $Bin(n_j, \mu_j)$ $j = 1, 2, \cdots, n$ this assumes the <u>n_j are large</u> and the number of Binomial observations <u>n is fixed</u>. With purely Binary data such as the UNM trauma data ($n > 3000$, $n_j = 1$) this is clearly not the case and the distributional approx is highly suspect

For the <u>Poisson</u> model $Y_j \sim$ indep Poisson (μ_j) $j = 1, 2, \cdots n$ the accuracy of the χ^2 approximation requires the μ_js to be large and n "fixed"

Thus, under suitable conditions for the Binomial & Poisson models, this result provides a means to test adequacy of the model using the pvalue

$$\text{pvalue} = \Pr(\chi^2_{n-p} > D^*_{obs})$$

D^*_{obs} = observed values of scaled & raw deviances

$=$

D_{obs} (i.e. $\phi = 1$)

(3.) Even if the χ^2_{n-p} approximation "breaks down", one can show

$$E\{ D^*(y; \hat{\mu}) \} \approx n - p \qquad \text{(degrees of freedom)}$$

Since large values of $D^*(y; \hat{\mu})$ suggest lack-of-fit, many people recommend qualitatively comparing $D^*(y; \hat{\mu})$ to $n-p$ to provide a rough idea of lack-of-fit. In particular, if

$$\frac{D^*(y; \hat{\mu})}{n-p} < 1 \quad \Rightarrow \text{ no evidence of lack-of-fit}$$

$$>> 1 \quad \Rightarrow \text{ some suggestion of lack of-fit}$$

Unfortunately, there is no accepted "cutoff" for how much greater than 1 the scaled deviance must be to indicate lack-of-fit. Some resolution of lack-of-fit can be based on <u>residual analysis</u>.

As an aside, there are convincing arguments for purely Binary data (each $n_i = 1$) that this convention is silly and that for this case $D^*(y; \hat{\mu})$ provides <u>no information</u> about lack-of-fit [I will explain in words]

④ It is important to recognize that the scaled deviance provides information on whether the model fits the data, while tests on regression coefficients assess the significance of effects <u>assuming the model fits</u>

⑤ An alternative goodness-of-fit statistic is the generalized Pearson statistic

$$X^2 = \sum_{j=1}^{n} w_j \frac{(y_j - \hat{u}_j)^2}{V(\hat{u}_j)}$$

[where it is understood $w_j = n_j$ for Binomial and $y_j = y_j^*$ the fraction of successes] and the scaled Pearson statistic

$$X^2_* = \frac{X^2}{\phi}$$

These are used analogously to the deviance & scaled deviance

<u>Remark:</u> For Binomial data X^2 reduces to the "usual" Pearson statistic, often written in terms of observed and expected (estimated) counts

$$X^2 = \sum \frac{(O-E)^2}{E}$$

where the sum extends over successes and failures

Comparing Nested Models

Suppose you are interested in comparing the fits of two nested models

(1) $\quad g(\mu_i) = x_i'\beta$

(2) $\quad g(\mu_i) = x_i'\beta + z_i'\tau$

Unlike earlier, the alternative model is not necessarily the saturated model. Assuming ϕ is fixed, the likelihood ratio test for comparing (1) to (2), or equivalently for testing $H_0: \tau = 0$ is the difference in scaled deviances between the 2 models

$$LRT = D^*(y; \hat{\mu}_1) - D^*(y; \hat{\mu}_2)$$

($\hat{\mu}_1$, $\hat{\mu}_2$: MLEs under models 1 and 2)

$$\dot{\sim} \chi^2_k$$

where k = # parameters in τ. The χ^2_k approximation tends to hold here even when the χ^2 approximations to the individual deviances fails. A pvalue for the test is

$$\text{pvalue} = \Pr(\chi^2_k \geq LRT_{OBS})$$

In essence, the LRT is just the drop in Deviance obtained by adding the z_i effects to the model, then scaled by ϕ. In practice, sequentially adding effects to an initial model leads to an Analysis of Deviance table, analogous to a table of sequential Sums of Squares in normal theory regression or ANOVA.

As an alternative to the LRT we can use a Wald statistic to test $H_0: \tau = 0$, which would be based on the MLEs from the alternative model (2) (as outlined earlier)

As another alternative we might consider using the difference between the two scaled Pearson statistics for the two models. As with the LRT, this statistic has a "large sample" χ^2_k distribution (but has the possibility of being negative)

As another alternative, we might consider the Score test of $H_0: \tau = 0$, which only requires the MLEs under the null model (1)

In general, the χ^2_k approximation to the LRT is often most accurate among these tests, so the LRT is often preferred

In certain asymptotic frameworks, the LRT, Wald and Score tests are "asymptotically equivalent", so they often behave similarly.

I haven't defined the Score test, but may do so at a later time

Residuals

A residual analysis can often indicate deficiencies with GLMs. As in normal theory regression, residuals are defined on a per-observation basis. By analogy to standard regression, we can write the Deviance & Pearson statistics in the forms

$$D(y; \hat{\mu}) = \sum_{j=1}^{n} d_j^2 \qquad d_j = j^{th} \text{ deviance residual}$$

$$X^2 = \sum_{j=1}^{n} r_j^2 \qquad r_j = j^{th} \text{ Pearson residual}$$

of sums-of-squares of residuals.

Here

$$r_j = \sqrt{\frac{\omega_j}{V(\hat{\mu}_j)}}\,(y_j - \hat{\mu}_j)$$

while

$$d_j = \text{sign}(y_j - \hat{\mu}_j)\sqrt{d_j^2}$$

where d_j^2 is the contribution of the i^{th} observation to the deviance. It is not immediately obvious but the individual contributions to the Deviance are positive, or non-negative!

Thus, for the Poisson model, with $w_j = 1$ and $V(\mu_i) = \mu_i$

$$r_j = \frac{(y_j - \hat{\mu}_j)}{\sqrt{\hat{\mu}_j}}$$

while

$$d_j = \text{sign}(y_j - \hat{\mu}_j) \sqrt{2\{y_j \log\left(\frac{y_j}{\hat{\mu}_j}\right) - (y_j - \hat{\mu}_j)\}}$$

For the normal model with $w_j = 1$ and $V(\mu_i) = 1$

$$r_j = d_j = (y_j - \hat{\mu}_j)$$

In general, the raw residuals are $y_j - \hat{\mu}_j$

These residuals have the property that each residual is zero when $y_j - \hat{\mu}_j$, is negative when $y_j < \hat{\mu}_j$ and positive when $y_j > \hat{\mu}_j$. Moreover, the magnitude of the residuals increases as $|y_j - \hat{\mu}_j|$ increases

The <u>scaled</u> Pearson and Deviance residuals are

$$\frac{r_j}{\sqrt{\phi}} \quad \text{and} \quad \frac{d_j}{\sqrt{\phi}}$$

These have the properties of having mean 0 and variance 1, at least approximately. To see this note

$$\frac{r_j}{\sqrt{\phi}} = \frac{(y_j - \hat{\mu}_j)}{\sqrt{V(\hat{\mu}_i)\frac{\phi}{\omega_j}}}$$

i.e. the scaled Pearson residual centers and scales y_j by its estimated mean and standard deviation. Hence the scaled Pearson residuals are <u>standardized</u>. The same argument applies to $d_j/\sqrt{\phi}$

In reality, the variances of the scaled residuals are slightly smaller than 1 (because my argument ignored the uncertainty in estimating μ_i). To see possibly why, note

$$D^*(y;\hat{\mu}) = \frac{1}{\phi} D(y,\hat{\mu}) = \frac{1}{\phi}\sum_j d_j^2$$

has

$$E\{D^*(y;\hat{\mu})\} \approx n-p \Rightarrow \text{var}\left(\frac{d_j}{\sqrt{\phi}}\right) \approx \frac{n-p}{n} = 1 - p/n$$

$\sum_j \text{var}\left(\frac{d_j}{\sqrt{\phi}}\right)$ if $E(d_j)=0$

on average. This is less than 1, but not much so if p is small relative to n.

The <u>standardized</u> Pearson & Deviance residuals

$$r_{Pj} = \frac{r_j}{\sqrt{\phi(1-h_j)}} \quad \text{and} \quad r_{Di} = \frac{d_j}{\sqrt{\phi(1-h_j)}}$$

adjust the scaled residuals to have mean 0 and variance 1 (approximately). Here h_j is the j^{th} case leverage, defined as the diagonal elements of

$$W^{.5} X (X' W X)^{-1} X' W^{.5}$$

where $W^{.5}$ is the "square root" of W, that is $W^{.5}$ is a diagonal matrix with diagonal element $w_{ii}^{.5}$, the square roots of the elements of W. Although W may contain ϕ, the leverages do not depend on ϕ. In practice W is estimated using MLEs.

At this point I will not justify the standardization, but will mention that in the case of normal theory regression this yields the usual standardized residuals used in diagnostic analyses.

For logistic & Poisson models, the standardized residuals are "ready to use" - the scale parameter is $\phi = 1$. For normal & gamma models we first have to estimate ϕ.

The techniques you learned for diagnostic analysis of normal linear regression using residuals directly apply to GLMs.

Not surprisingly, other diagnostics such as Cook's Distance can be easily extended to GLMs

In particular, if $\hat{\beta}$ is the MLE of β under the model

$$g(\mu_i) = x_i'\beta$$

and $\hat{\beta}_{(-j)}$ is the MLE based on the data but holding out the j^{th} observation, then Cook's distance for case j is

$$C_j = \frac{1}{p}(\hat{\beta} - \hat{\beta}_{(-j)})' [\widehat{Var}(\hat{\beta})]^{-1} (\hat{\beta} - \hat{\beta}_{(-j)})$$

$$= \frac{1}{p}(\hat{\beta} - \hat{\beta}_{(-j)})\, x'\hat{\omega} x\, (\hat{\beta} - \hat{\beta}_{(-j)}) \qquad j = 1, 2, \ldots, n$$

Some packages do not scale C_j by p (and some do not compute it at all!)

often $\hat{\beta}_{(-j)}$ is computed using a "one-step" approximation (I'll describe in words).

Estimating ϕ

In the normal and gamma models ϕ is unknown and must be estimated from the data. This also applies to Binomial and Poisson models when allowing for overdispersion, a point I will return to later. In practice, simple method-of-moments estimators are used. Noting that

$$E\left(X_i^2\right) = \frac{1}{\phi} E(x^2) \doteq n-p$$

$$E\{D^i(y,\hat{u})\} = \frac{1}{\phi} E\{D(y,\hat{u})\} \doteq n-p$$

when the model holds, if we set the scaled deviance equal to its (approximate) expectation and solve for ϕ we set

$$\frac{D(y;\hat{u})}{\phi} = n-p \Rightarrow \hat{\phi} = \frac{D(y;\hat{u})}{n-p}$$

Alternatively, we might set

$$\hat{\phi} = \frac{X^2}{n-p}$$

given an estimate of ϕ we then replace ϕ by $\hat{\phi}$ in <u>all relevant formulas</u> (variances, CI, residuals etc.). This substitution does not alter large sample properties.

It is important to recognize that the scaled Pearson & Deviance

$$X_+^2 = \frac{X^2}{\hat{\phi}}$$

$$D^+(y;\hat{\mu}) = \frac{D(y;\hat{\mu})}{\hat{\phi}}$$

<u>equal n-p</u> when $\hat{\phi}$ is estimated using the Pearson & Deviance.

Thus, when ϕ needs to be estimated, neither the scaled Pearson or Deviance statistic provides information on lack-of-fit.

Note that for normal linear regression (or even normal GLM with non-identity link!)

$$X^2 = D(y;\hat{\mu}) = \sum_i (y_i - \hat{\mu}_i)^2 = \text{residual SS}$$

so

$$\hat{\phi} = \hat{\sigma}^2 = \frac{1}{n-p}\sum_i (y_i - \hat{\mu}_i)^2$$

is usual Residual MS.

In the case of the LRT comparing 2 nested models

$$LRT = D^+(y;\hat{\mu}_1) - D^+(y;\hat{\mu}_2)$$

$$= \frac{1}{\phi}\{D(y;\hat{\mu}_1) - D(y;\hat{\mu}_2)\}$$

$H_0: g(\mu_i) = x_i'\beta$

$H_A: g(\mu_i) = x_i'\beta + z_i'\tilde{\nu}$

the natural estimate of ϕ is based on the alternative model

$$\hat{\phi} = \frac{D(y; \hat{\mu}_2)}{n - (p+k)}$$

(p: # param in reduced model; k: # parameters added)

In the normal theory regression case this leads to the LRT statistic as the difference in residual SS for the 2 models divided by the residual Mean Square. This has approximately a χ^2_k distribution but often the approximation can be improved by instead using

$$F = \frac{D(y; \hat{\mu}_1) - D(y; \hat{\mu}_2)}{k\,\hat{\phi}} \sim F_{k, n-p-k}$$

where the drop in deviance divided by k is analogous to the Mean Square for the added regression effect.

MLE of ϕ

As an alternative to Method of Moments, one might consider ML estimation. Writing the likelihood as a function of both β and ϕ (as on p36)

$$L(y; \beta, \phi) = \sum_{j=1}^{n} \left\{ \frac{y_j \theta_j(\beta) - b(\theta_j(\beta))}{a_j(\phi)} - c_j(y_j; \phi) \right\}$$

the joint MLE of β and ϕ satisfy:

$$\begin{bmatrix} \frac{\partial L}{\partial \beta} \\ \\ \frac{\partial L}{\partial \phi} \end{bmatrix} = \begin{bmatrix} 0 \\ \\ 0 \end{bmatrix}$$

we saw before that

$$\frac{\partial L}{\partial \beta} = x' \omega \Delta (y - \mu)$$

where Δ does not depend on ϕ and ω has diagonal elements

$$\omega_{jj} = \frac{1}{a_j(\phi)} \frac{1}{V(\mu_i)} \frac{1}{\{g'(\mu_i)\}^2}$$

Noting $a_j(\phi) = \phi / \omega_j$ we can write

$$\omega_{jj} = \frac{1}{\phi} \omega_j \frac{1}{V(\mu_i)} \frac{1}{\{g'(\mu_i)\}^2} = \frac{1}{\phi} \omega_{jj}^*$$

and letting ω^* be the $n \times n$ diagonal matrix with elements ω_{jj}^*, we also have

$$\frac{\partial L}{\partial \beta} = \frac{1}{\phi} x' \omega^* \Delta (y - \mu)$$

Now, using $a_j(\phi) = \phi / \omega_j$

$$\frac{\partial L}{\partial \phi} = -\frac{1}{\phi^2} \sum_j \{ y_j \theta_j(\beta) - b(\theta_j(\beta)) \} - \sum_j c_j'(y_j, \phi)$$

where $c_j'(y_j, \phi) = \partial c_j(y_j, \phi) / \partial \phi$

The solution to

$$\frac{\partial L}{\partial \beta} = 0 = \frac{1}{\phi} X' W^* \Delta (y - \mu)$$

does not depend on ϕ because W^*, Δ and μ are independent of ϕ. However, the solution to $\partial L / \partial \phi = 0$ will depend on β. This implies that computationally the MLE of β does not depend on the MLE of ϕ - i.e. $\hat{\beta}$ is same as before when we assumed ϕ was fixed, but the MLE of ϕ, say $\tilde{\phi}$ does depend on $\hat{\beta}$

A more important property is that the information matrix, as a function of β and ϕ

$$I(\beta, \phi) = -E \begin{bmatrix} \frac{\partial^2 L}{\partial \beta \partial \beta'} & \vdots & \frac{\partial^2 L}{\partial \beta \partial \phi} \\ \cdots & \vdots & \cdots \\ \left(\frac{\partial^2 L}{\partial \beta \partial \phi}\right)' & \vdots & \frac{\partial^2 L}{\partial \phi^2} \end{bmatrix} \begin{matrix} \} \text{ p rows} \\ \} \text{ 1 row} \end{matrix}$$

$$\underbrace{\qquad}_{\text{p cols}} \quad \underbrace{\qquad}_{\text{1 col}}$$

has the property of being block-diagonal. That is,

$$E\left\{ -\frac{\partial^2 L}{\partial \beta \partial \phi} \right\} = 0_{p \times 1}$$

To see this, note that

$$\frac{\partial^2 L}{\partial \beta \partial \phi} = \frac{\partial}{\partial \phi}\left\{\frac{\partial L}{\partial \beta}\right\}$$

$$= \frac{\partial}{\partial \phi}\left\{\underbrace{\frac{1}{\phi} x'\omega^* \Delta (y-\mu)}_{\text{a vector}}\right\}$$

$$= -\frac{1}{\phi^2} x'\omega^* \Delta (y-\mu)$$

so

$$E\left\{-\frac{\partial^2 L}{\partial \beta \partial \phi}\right\} = \frac{1}{\phi^2} x'\omega^* \Delta \underbrace{E(y-\mu)}_{0} = 0_{p\times 1}$$

Letting $I_2(\beta,\phi) = E\left\{-\frac{\partial^2 L}{\partial \phi^2}\right\}$ be unspecified for this argument, we have

$$I(\beta,\phi) = \left[\begin{array}{c|c} \overbrace{x'\omega x}^{\text{same as before!}} & 0_{p\times 1} \\ \hline 0_{1\times p} & I_2(\beta,\phi) \end{array}\right]$$

and

$$I^{-1}(\beta,\phi) = \left[\begin{array}{c|c} (x'\omega x)^{-1} & 0_{p\times 1} \\ \hline 0_{1\times p} & \frac{1}{I_2(\beta,\phi)} \end{array}\right]$$

ML theory suggests

$$\begin{bmatrix} \hat{\beta} \\ \tilde{\phi} \end{bmatrix} \dot{\sim} N\left(\begin{bmatrix} \beta \\ \phi \end{bmatrix}, \left[\begin{array}{c:c} (x'wx)^{-1} & 0_{p\times 1} \\ \hdashline 0_{1\times p} & \frac{1}{I_2(\beta,\phi)} \end{array}\right] \right)$$

Because the inverse information is block-diagonal, $\hat{\beta}$ and $\tilde{\phi}$ are independent (approximately) and further the distribution of $\hat{\beta}$ agrees with that obtained assuming ϕ was known and fixed.

As an aside, I will note that $\tilde{\phi}$ depends on the value of $\hat{\beta}$, but the asymptotic distribution <u>does not</u>, that is $I_2(\beta,\phi)$ does not depend on β. I didn't need this fact here for what I wanted to say.

In practice, any consistent estimator of ϕ may be used in the inferential procedures for β, so the simple estimators suggested earlier are often used instead of MLEs - This is only an issue for gamma regression since the MLE of $\phi = \sigma^2$ in the normal model is

$$\tilde{\phi} = \frac{1}{n}\sum_j (y_j - \hat{\mu}_j)^2 = \frac{1}{n} D(y;\hat{\mu})$$

versus the earlier estimators which divide $D(y;\hat{\mu})$ by $n-p$

Implementation in SAS

Proc GENMOD fits GLMs in SAS. Other Procs fit specific GLMs. For example, Proc Logistic fits binomial response models whereas Proc Reg and Proc GLM (for general linear model not generalized linear model!) fits standard normal theory linear regression and ANOVA models.

My Logistic Regression in SAS II handout extensively discusses the Logistic procedure, which allows certain features not programmed in GENMOD - more options on diagnostics, and automated model selection (i.e. backwards, stepwise etc). See handout for more discussion & online HELP for more details.

GENMOD

Typical syntax

```
Proc GENMOD options;
  CLASS class variables options;
  MODEL response = effects options; ← all cases except Binomial
(or)
  MODEL events/trials = effects options; ← only for Binomial
```

specifies factors

can specify ordering on class vars & create own baseline

↑ for Binomial data with 2 columns in data set: 1 for # successes (event) another sample sizes

CONTRAST ..... ; allows testing on contrasts in β's

weight ..... ; can specify column of weights
- automatically handled for Binomial!

OPTIONS for Model statement

		default link
• DIST = Bin or B	Binomial	logit
gam or G	Gamma	inverse
POI or P	Poisson	log
NOR or N	Normal	identity

specify distribution

• LINK = Log or ID or Logit or CLOGLOG or Probit (others available)
- ID: identity
- CLOGLOG: complementary log-log

• ITPRINT - prints iteration history of ML routine (check convergence)

• LRCI - compute and print likelihood ratio CI for regression coefficient (did not discuss this)

• OBSTATS - table of diagnostic summaries, such as residuals estimated means and linear predictors ($\hat{\mu}_i$ and $\hat{\eta}_i$)

• NOSCALE
Scale = value
} can be used separately or together. Common specification is to use one or other.

NOSCALE ⇒ set $\phi = 1$ i.e. Binomial / Poisson with no overdispersion

SCALE = ML
= P
= D
} estimate ϕ using ML, Pearson or Deviance

When you specify a scale, SAS reports a Dispersion Parameter estimate. For the Normal (and Poisson & Binomial with overdispersion) the reported estimate is $\sqrt{\phi}$. Since $\phi = \sigma^2$ for the Normal, this is the standard deviation. For our representation of the gamma, where σ^2 was the coefficient of variation, the dispersion parameter reported is $1/\phi$ which corresponds to $1/\sigma^2$

SAS will report unscaled & scaled versions of the Deviance and Pearson statistic, and a parameter estimate table with estimated regression coefficients, standard errors, Wald CI (those described in note) and Wald test p-values on individual regression effects. GENMOD does not provide p-values for either the deviance or Pearson statistic, but does report the value of the statistic divided by its corresponding df.

Further options are described on the handouts and in the online help.

Overdispersion

In many settings responses are collected on a cluster of units, for example a family, a litter of mice or seeds from a batch. A natural feature of such data is that responses within a cluster are often correlated. A cluster may also correspond to multiple measures on a single individual at a specific time or single measurements at multiple times, or both.

When responses within a cluster are totaled, the variability in the total often differs from that obtained with independent responses. This leads to overdispersion or underdispersion relative to a model based on independent responses within a cluster. We will study a variety of ways to deal with dependence. At this time we will consider a standard means of dealing with <u>overdispersed</u> (or underdispersed) clustered data. The method does not specifically require data to be clustered, but rather deals with a specific model deviation for the mean and variance structure in a GLM.

To illustrate the idea of overdispersion, let us consider the Challenger shuttle data. Let

$$Y_i = \#\text{ O-rings of 6 that failed on } i^{th} \text{ flight}$$

Also consider the individual O-rings via

$$Y_{ij} = \begin{cases} 1 & \text{if } j^{th} \text{ O-ring on } i^{th} \text{ flight failed} \\ 0 & \text{else} \end{cases}$$

Then, with $n_i = 6$

$$Y_i = \sum_{j=1}^{n_i} Y_{ij}$$

As a first pass, it might be reasonable to assume each O-ring has the same failure probability on a given flight

$$\mu_i = pr(Y_{ij} = 1) \qquad j = 1, 2, \cdots, n_i$$

which implies that $Y_{i1}, Y_{i2}, \cdots, Y_{in_i}$ are identically distributed Bernoulli (μ_i) random variables with

$$E(Y_{ij}) = \mu_i \quad \text{and} \quad var(Y_{ij}) = \mu_i(1-\mu_i)$$

Note

$$E(Y_i) = E\left\{\sum_{j=1}^{n_i} Y_{ij}\right\} = \sum_{j=1}^{n_i} E(Y_{ij}) = n_i \mu_i$$

and in general

$$\text{Var}(Y_i) = \text{var}\left(\sum_{j=1}^{n_i} Y_{ij}\right)$$

$$= \sum_{j=1}^{n_i} \text{var}(Y_{ij}) + \sum_{j \neq k} \text{cov}(Y_{ij}, Y_{ik})$$

Now $\text{var}(Y_{ij}) = \mu_i(1-\mu_i)$ and

$$\text{cov}(Y_{ij}, Y_{ik}) = \text{correlation}(Y_{ij}, Y_{ik}) \sqrt{\text{var}(Y_{ij})\,\text{var}(Y_{ik})}$$

$$= \rho_{jk} \qquad \mu_i(1-\mu_i)$$

ρ — assume does not depend on flight

Thus

$$\text{var}(Y_i) = n_i \mu_i(1-\mu_i) + \mu_i(1-\mu_i) \sum_{j \neq k} \rho_{jk}$$

There are $2\binom{n_i}{2}$ terms in the sum, or $n_i(n_i-1)$ terms. If the correlation is constant, say $\rho_{jk} = \rho$ across pairs of observations then

$$\text{var}(Y_i) = n_i \mu_i(1-\mu_i) + n_i(n_i-1)\rho\,\mu_i(1-\mu_i)$$

$$= n_i \mu_i(1-\mu_i)\{1 + (n_i-1)\rho\}$$

If the Y_{ij}'s are mutually independent then $Y_i \sim \text{Binomial}(n_i, \mu_i)$ with $\text{var}(Y_i) = n_i \mu_i(1-\mu_i)$, which also follows from above because $\rho = 0$ in this case.

If we define

$$\phi_i = 1 + (n_i - 1)\rho$$

then

$$\text{var}(Y_i) = n_i \mu_i (1-\mu_i) \phi_i$$

In the EF framework, if $Y_i \sim \text{Bin}(n_i, p_i)$ the density is defined in terms of $Y_i^* = Y_i / n_i$ for which

$$E(Y_i^*) = \mu_i \quad \text{and} \quad \text{var}(Y_i^*) = \frac{\phi_i V(\mu_i)}{\omega_i}$$

where $V(\mu_i) = \mu_i(1-\mu_i)$; $\omega_i = n_i$ and $\phi_i = 1$.

For our model with correlation as above and with

$$Y_i^* = \#\text{ O-rings fail on flight } i \,/\, n_i$$

we have

$$E(Y_i^*) = \mu_i$$

$$\text{var}(Y_i^*) = \frac{1}{n_i^2}\text{var}(Y_i) = \frac{\phi_i \mu_i(1-\mu_i)}{n_i} = \frac{\phi_i V(\mu_i)}{\omega_i}$$

Regardless of the distribution of Y_i^*, it has the same <u>mean</u> and <u>variance function</u> as a Binomial RV.

If $\rho > 0$ then $\phi_i > 1 \Rightarrow Y_i$ is "over-dispersed" (relative to Binomial)

$<$ $<$ $\Rightarrow$ "under-dispersed"

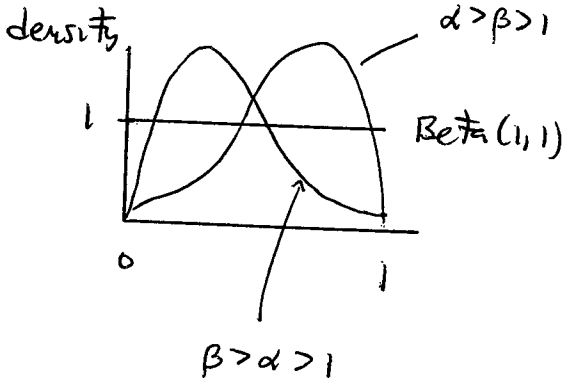
Except for the scale parameter ϕ_i being potentially different from 1, the first 2 moments of Y_i^* agree with the moments of the Binomial (i.e. structural dependence of moments on μ_i)

<u>Overdispersion</u> will often also result from <u>heterogeneity</u> (often modelled using random effects). Suppose for example

Y = # seeds that germinate from a packet of n seeds

Assume that $Y \mid p \sim$ Binomial (n, p) but that the germination probability varies by packet according to the Beta (α, β) distribution

$$p \sim \text{Beta}(\alpha, \beta)$$



With this distribution

$E(p)$ = average success probability

$$= \frac{\alpha}{\alpha+\beta} \equiv \mu$$

and

$$\text{var}(p) = \frac{\alpha}{\alpha+\beta}\frac{\beta}{\alpha+\beta} \cdot \underbrace{\frac{1}{\alpha+\beta+1}}_{\text{call this } \rho}$$

$$= \mu(1-\mu)\rho$$

Using well-known rules

$E(Y)$ = expected # successes from a randomly selected packet

$$= E_p\, E(Y|p)$$

$$= E_p\{np\} = nE(p) = n\mu$$

$$\text{var}(Y) = E_p\, \text{var}(Y|p) + \text{var}_p\, E(Y|p)$$

$$= E_p\{np(1-p)\} + \text{var}_p(np)$$

$$= nE\{p(1-p)\} + n^2\,\text{var}(p)$$

$$= n^2\,\text{var}(p) + n\{E(p) - E(p^2)\}$$

$$= n^2\,\text{var}(p) + n\{\underbrace{E(p)}_{\mu} - \underbrace{[E(p)]^2}_{\mu^2} - \underbrace{\{[E(p^2) - [E(p)]^2\}}_{\text{var}(p)}\}$$

$$= (n^2-n)\,\text{var}(p) + n(\mu-\mu^2)$$

$$= n(n-1)\,\text{var}(p) + n\mu(1-\mu)$$

$$= n(n-1)\mu(1-\mu)\rho + n\mu(1-\mu)$$

$$= n\mu(1-\mu)\{1 + (n-1)\rho\}$$

This is same form as we had before - As it should since here we could have written $Y = \sum_{j=1}^{n} Y_j$ where $Y_j|p$ are indep Bernoulli(p) random variables, so marginally $Y_j \sim$ Bernoulli(μ) and

ρ = correlation(Y_j, Y_h)

Overdispersion can also occur with Poisson-like count data.

Suppose $Y|T \sim$ Poisson (T) where T has mean μ and variance $\chi\mu$. Then

$$E(Y) = E_T E(Y|T) = E_T(T) = \mu$$

$$\begin{aligned} \text{var}(Y) &= \text{var}_T E(Y|T) + E_T \text{var}(Y|T) \\ &= \text{var}_T(T) + E_T(T) \\ &= \chi\mu + \mu = \mu(1+\chi) = \mu\phi \end{aligned}$$

where $\phi = 1 + \chi$.

We can think of Y as a Poisson rv that is observed for a random length of time T, where T has mean & variance $(\mu$ & $\chi\mu)$ You can set this structure on T if it has a gamma distribution

If Y were Poisson (μ) it would have an EF distribution with mean μ and variance function $\phi V(\mu)/w$ where $\phi = w = 1$ and $V(\mu) = \mu$. As defined above the random variable Y has mean μ and variance $\phi V(\mu)/w$ with $V(\mu) = \mu$; $w = 1$ and $\phi = 1 + \chi > 1$. Thus rv has same mean & variance structure but is overdispersed relative to the Poisson distribution.

In GLMs we are primarily interested in EF distributions. The notion of over and underdispersion relative to Binomial and Poisson make sense, as discussed above. The normal and gamma have scale parameters, so overdispersion is not considered in those settings, although it is reasonable to consider the gamma as overdispersed relative to the exponential.

Quasi-Likelihood

If we have independent rv's $Y_1, Y_2, \ldots, Y_n$ with EF distributions having means μ_j and variances $\phi V(\mu_j)/w_j$ where

$$g(\mu_j) = x_j'\beta$$

then the score function for β is

$$\frac{\partial L}{\partial \beta} = \frac{1}{\phi} \sum_{j=1}^{n} \frac{w_j (y_j - \mu_j)}{V(\mu_j)\, g'(\mu_j)}$$

$$= X' W \Delta (y - \mu)$$

where μ_j, $V(\mu_j)$, W and Δ are functions of β. Furthermore, the solution $\hat{\beta}$ to the likelihood equation

$$X' W \Delta (y - \mu) = 0 \qquad (1)$$

has a large sample normal distribution

$$\hat{\beta} \dot{\sim} N(\beta, (X' W X)^{-1}) \qquad (2)$$

We can think of (1) as an estimating equation for β. That is, (1) implicitly defines a method for obtaining an estimate of β. This estimating equation is motivated by ML, that is we obtain (1) from assuming that the Y_j's have EF distributions.

Estimating equations are at the core of most statistical methods. For example, the sample mean $\bar{x}$ solves the estimating equation $\sum_i (x_i - \mu) = 0$, our method of moments estimator of ϕ given by $\hat{\phi} = D(y;\hat{\mu})/(n-p)$ solves

$$D(y;\hat{\mu}) - (n-p)\phi = 0$$

and so on...

An interesting theoretical question with enormous practical application is "For what distributions of $Y_1, Y_2, \ldots, Y_n$ does the estimator $\hat{\beta}$ obtained by solving (1) have the large sample distribution (2)?"

The theory of estimating equations, called quasi-likelihood in the context of GLMs establishes that (2) applies regardless of the distribution of the Y_j's provided

- $Y_1, Y_2, \ldots, Y_n$ are independent
- $E(Y_i) = \mu_i$ where $g(\mu_i) = x_i'\beta$

and

- $\text{var}(Y_i) = \frac{\phi}{\omega_i} v(\mu_i)$ for some ϕ and fixed ω_i

Put another way, (2) holds provided the Y_i's have the same mean and variance function as a GLM for an EF distribution with link $g(\cdot)$. Further, the scale ϕ may be estimated with either the corresponding Pearson or (quasi-) Deviance statistics divided by the df and standard GLM results for models with scale parameters may be applied to make inferences about β.

The implications here are that we can use standard GLM software to fit GLMs to overdispersed Binomial and Poisson data and only need to specify that a scale parameter should be estimated in order for the resulting inferences to be valid.

Remarks

1. The discussion of quasi-likelihood assumed that the scale ϕ was constant across observations. This constraint is made in most statistical packages, for example SAS and R.

2. Our development of overdispersion for Binomial type data had a variance of the form $\phi_j V(\mu_j)/w_j$ with $w_j = n_j$ $V(\mu_j) = \mu_j(1-\mu_j)$ and $\phi_j = 1 + (n_j - 1)\rho$. Here ϕ_j is constant only if $n_1 = n_2 = \cdots = n_n$ which holds in the shuttle example. There are ways to allow ϕ to be modelled as a function of sample sizes and parameters but this approach is not often taken, and not available in most packages. In the "unequal" n_i case where we assume a constant scale ϕ we are thinking of ϕ as capturing the average overdispersion or correlation

3. In the shuttle / seed examples, an overdispersed Binomial model is sensible when the <u>total</u> # of O-ring failures and germinating seeds is used as the basis for the analysis. Although individual responses have the same success probabilities as the batch totals and satisfy the regression model, the individual responses are correlated. Thus, using batch totals as basis for analysis eliminates issues of dependence among individual observations.

Clearly, a different approach would be needed if individual responses within a batch had different covariate patterns. We will deal with this issue at some point, time permitting.

Basic Computing with GLMs

Recall the following notation. For a GLM we have independent responses $Y_1, Y_2, \ldots, Y_n$ from an EF distribution where $E(Y_i) = \mu_i$ satisfies

$$g(\mu_i) = x_i'\beta.$$

for some known link function $g(\cdot)$ and covariates.

Using matrix notation, and emphasizing the dependence on β, the score function satisfies

$$S(\beta) = \frac{\partial L}{\partial \beta} = X' W(\beta) \Delta(\beta) \{ y - \mu(\beta) \}$$

while the expected and observed information matrix are (when evaluated at β)

$$I_E(\beta) = X' W(\beta) X$$

and

$$I_O(\beta) = X' W_O(\beta) X$$

The diagonal matrices $W(\beta)$ and $\Delta(\beta)$ have diagonal elements

$$w_{jj} = \frac{1}{a_j(\phi)} \frac{1}{V(\mu_j)} \frac{1}{g'(\mu_j)^2}$$

$$v_{jj} = g'(\mu_j)$$

where $\mu_j \equiv \mu_j(\beta)$ i.e. μ_j is a function of β. The form for W_O was given on p52 of the notes, but remember that

$$W_O = -\frac{\partial^2 L}{\partial \beta \, \partial \beta'} \quad \text{(minus the Hessian)}$$

To calculate the MLE of β, we solve the score or likelihood equation

$$S(\beta) = 0$$

This typically must be done iteratively. There are 3 standard approaches for GLMs: IRWLS (iteratively reweighted least squares), Newton-Raphson, and Fisher's method of scoring. The latter two methods are general in that they can be applied to arbitrary maximum likelihood settings, while IRWLS was developed primarily for GLMs. I will discuss the basic ideas of Newton-Raphson and Fisher's method of scoring.

Consider solving

$$S(\beta) = 0$$

where β is $p \times 1$, as is $S(\beta)$. We have, in general, p non-linear equations in p unknowns. A simple iterative scheme can be derived based on a one-step Taylor series expansion about (an arbitrary point) β_0

$$S(\beta) \doteq S(\beta_0) + [S'(\beta_0)](\beta - \beta_0)$$

Here $[S'(\beta_0)]$ is a $p \times p$ matrix of partial derivatives with elements

$$\frac{dS_i(\beta)}{\partial \beta_j} \quad \text{in } i^{th} \text{ row and } j^{th} \text{ column.}$$

If in fact β is the solution, i.e. $S(\beta) = 0$ then

$$0 \doteq S(\beta_0) + [S'(\beta_0)](\beta - \beta_0)$$

or

$$\beta \doteq \beta_0 - [S'(\beta_0)]^{-1} S(\beta_0)$$

assuming the matrix inverse exists. Alternatively, we can write this as

$$\beta \doteq \beta_0 + [-S'(\beta_0)]^{-1} S(\beta_0)$$

This suggests an iterative scheme where we nominate an initial guess $\hat{\beta}_0$ to the solution and compute successive approximations to the solution $\hat{\beta}$ by computing

$$\hat{\beta}_{i+1} = \hat{\beta}_i + [-S'(\hat{\beta}_i)]^{-1} S(\hat{\beta}_i) \qquad i = 0, 1, \ldots \tag{1}$$

The iteration continues until convergence, which we can take to mean that

$$\| \hat{\beta}_{i+1} - \hat{\beta}_i \|^2 = \sum_{j=1}^{p} \{ \hat{\beta}_{i+1,j} - \hat{\beta}_{i,j} \}^2 < \varepsilon$$

($\hat{\beta}_{i+1,j}$: step; $\hat{\beta}_{i,j}$: element of β)

where ε is some prespecified small tolerance.

To understand why this approach might actually work, suppose that for some value j we get that

$$\hat{\beta}_{j+1} = \hat{\beta}_j$$

That is, no change in the approximation to the solution from the j^{th} to $(j+1)^{ST}$ step. Noting that

$$\hat{\beta}_{j+1} = \hat{\beta}_j + [-S'(\hat{\beta}_j)]^{-1} S(\hat{\beta}_j)$$

where $[S'(\hat{\beta}_j)]^{-1}$ was assumed to exist, we must have $S(\hat{\beta}_j) = 0$

Thus, $\hat{\beta}_j$ is a solution to $S(\beta) = 0$.

The iterative scheme (1) used to solve $S(\beta)=0$ is called the Newton-Raphson method. It applies to arbitrary non-linear equations. In the context of maximizing a log-likelihood $L(\beta)$ where $S(\beta) = \partial L(\beta)/\partial \beta$ we have

$$\hat{\beta}_{i+1} = \hat{\beta}_i + \left[-\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta'} \right]^{-1}_{\beta=\hat{\beta}_i} S(\hat{\beta}_i)$$

$$= \hat{\beta}_i + I_o^{-1}(\hat{\beta}_i)\, S(\hat{\beta}_i) \qquad \text{general statistical model}$$

GLM
$$= \hat{\beta}_i + (x' \omega_o(\hat{\beta}_i) x)^{-1} x' \omega(\hat{\beta}_i) \Delta(\hat{\beta}_i) \{y - \mu(\hat{\beta}_i)\}$$

A standard variation on Newton-Raphson (NR) is Fisher's method of scoring, which replaces the observed information matrix by the expected information matrix, leading to

$$\hat{\beta}_{i+1} = \hat{\beta}_i + I_E^{-1}(\hat{\beta}_i)\, s(\hat{\beta}_i)$$

$$= \hat{\beta}_i + (x'\omega(\hat{\beta}_i)x)^{-1} x'\omega(\hat{\beta}_i)\Delta(\hat{\beta}_i)\{y - \mu(\hat{\beta}_i)\}$$

It is interesting to note that one might consider instead the iterative scheme

$$\hat{\beta}_{i+1} = \hat{\beta}_i + H(\hat{\beta}_i)\, s(\hat{\beta}_i)$$

for an <u>arbitrary</u> nonsingular matrix $H(\hat{\beta}_i)$. If this method leads to convergence, then the point of convergence is a solution to $s(\beta) = 0$ - to see this note that the "argument" suggesting why NR works did not actually use the fact that $[-s'(\hat{\beta}_i)]$ was the matrix of partial derivatives but only that this matrix is invertible. Thus, one might study whether alternatives to $[-s'(\hat{\beta}_i)]$ might be preferable from a computational perspective. This issue is closely examined in courses on <u>numerical analysis</u>.

<u>General Remarks</u>

1. There is an extensive literature on existence & uniqueness of MLEs and correspondingly, on existence & uniqueness of solutions to $s(\beta) = 0$.

2. Convergence and/or speed of convergence can depend critically on the choice of the initial guess $\hat{\beta}_0$. I will discuss this a bit later for Binomial response models.

   It is often the case that convergence is quicker if $\hat{\beta}_0$ is near the solution, but this is not always the case - and further how do you choose $\hat{\beta}_0$ near $\hat{\beta}$ if the whole point is to find $\hat{\beta}$?

3. For maximizing the log-likelihood, an alternative convergence criterion is to iterate until successive estimates are close in log-likelihood i.e. iterate until

$$|L(\hat{\beta}_{i+1}) - L(\hat{\beta}_i)| < \delta$$

   for some small tolerance δ

4. An alternative convergence criterion might be to iterate until successive estimates are close in a <u>relative</u> sense

$$\frac{\|\hat{\beta}_{i+1} - \hat{\beta}_i\|^2}{\|\hat{\beta}_i\|^2} < \varepsilon$$

   This may not be such a good criterion if many of the regression coefficients are near 0!

5. When the iterative scheme, say NR, corresponds to maximizing a log-likelihood, we may wish to modify the iteration

$$\hat{\beta}_{j+1} = \hat{\beta}_j + \left[-\frac{\partial^2 \ell(\beta)}{\partial \beta \, \partial \beta'} \right]^{-1}_{\beta = \hat{\beta}_i} s(\hat{\beta}_i)$$

via

$$\hat{\beta}_{j+1}(\alpha) = \hat{\beta}_j + \alpha \left[\frac{-\partial^2 L(\beta)}{\partial \beta \, \partial \beta'} \right]^{-1}_{\beta = \hat{\beta}_i} s(\hat{\beta}_i)$$

where α is a <u>scalar</u>. The basic idea is that NR does not guarantee that the log-likelihood at $\hat{\beta}_{j+1}$ exceeds that at $\hat{\beta}_j$; but it does point you in the direction of steepest ascent. A simple modification is to search places in this direction, and choose the scalar α that maximizes $L(\hat{\beta}_{j+1}(\alpha))$. Note that $\alpha = 1$ gives NR.

There are some minor practical issues to work out with this method - for example, what do you do if the optimal $\alpha = 0$? This leads you to stopping the iteration, but you likely have not converged! So a choice of $\alpha \neq 0$ is needed.

I will refer to this as NR with a <u>line search</u> (for the optimal α). The same idea applies to Fisher's scoring.

Fisher Scoring for Binomial Response Models

I will show you a series of Matlab programs I wrote to fit Binomial response models in MATLAB. The programs allow the user to specify a choice for the link function: logistic, Probit or complementary log-log.

As a reminder, for Y_i indep $\text{Bin}(m_i, \mu_i)$ these are defined via

$$g(\mu_i) = \begin{cases} \log\left(\frac{\mu_i}{1-\mu_i}\right) & \text{logit} \\ \Phi^{-1}(\mu_i) & \text{probit} \\ \log(-\log(1-\mu_i)) & \text{complementary log-log (cloglog)} \end{cases}$$

given the linear predictor

$$\eta_i = x_i'\beta$$

the probabilities expressed in terms of η_i are

$$\mu_i = g^{-1}(\eta_i) = \begin{cases} \exp(\eta_i)/\{1+\exp(\eta_i)\} & \text{logit} \\ \Phi(\eta_i) & \text{probit} \\ 1-\exp(-\exp(\eta_i)) & \text{clog log} \end{cases}$$

For Fisher scoring, I need the score function

$$X' W \Delta (y - \mu)$$

and expected Fisher information matrix.

$$I_E(\beta) = X' W X$$

we showed earlier in the semester (p47-8) that W and Δ had diagonal elements (taking $\phi = 1$)

$$w_{jj} = \frac{1}{a_j(\phi)} \cdot \frac{1}{V(\mu_j)} \cdot \frac{1}{g'(\mu_j)^2} = m_j \cdot \frac{1}{\mu_j(1-\mu_j) g'(\mu_j)^2}$$

$$v_{jj} = g'(\mu_j)$$

so $W\Delta$ has diagonal elements

$$w_{jj} v_{jj} = \frac{m_j}{\mu_j(1-\mu_j) g'(\mu_j)}$$

Also

$$g'(\mu_i) = \begin{cases} \dfrac{1}{\mu_i(1-\mu_i)} & \text{logit} \\ \dfrac{1}{f(\Phi^{-1}(\mu_i))} & \text{probit} \\ \dfrac{-1}{(1-\mu_i)\log(1-\mu_i)} & \text{cloglog} \end{cases}$$

Here $f(t)$ is the $N(0,1)$ density

My general routine for Fisher Scoring (a Matlab function glmle2) calls several functions that provide relevant summaries. All routines do calculations for vectors a matrices of input

① pt : Implement the following idea. given a design matrix X and a generic value of β, compute vector of mean responses via

$$\eta = X\beta \quad \text{linear predictor}$$
$$\mu = g^{-1}(\eta)$$

where inverse function operates elementwise on η. Choice of link is passed as a parameter to the function

② gp : Computes the derivative of the link function for each binomial response. That is, compute

$$g'(\mu)$$

for every element of mean vector μ. Choice of link is passed as a parameter to the function

③ starts : Provides simple starting values for the iterative procedure. Idea used is the following: If we have the model

$$g(\mu_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$$

but the regression effects were weak, so that

$$\beta_1 \approx \beta_2 \approx \cdots \approx \beta_{p-1} \approx 0$$

we might consider fitting the model with just an intercept

$$g(\mu_i) = \beta_0$$

For this model, all the probabilities are identical, so one can show the MLE of β_0 is

$$\hat{\beta}_0 = g(\hat{\mu})$$

where

$$\hat{\mu} = \frac{\sum y_i}{\sum m_i} = \text{overall proportion from collection of Binomial samples}$$

As a starting value in the iteration, I use

$$(\hat{\beta}_0, \underbrace{0, 0, \cdots, 0}_{p-1 \text{ entries}})$$

as a starting value. This approach often works well.

As with previous functions, the link is passed as a parameter.

④ loglike : Computes the log likelihood for a Binomial sample $Y_1, Y_2, \cdots Y_n$

$$\ell = \sum_{i=1}^{n} y_i \log \mu_i + (m_i - y_i) \log(1 - \mu_i) \qquad \text{subscript notation}$$

$$= y' \log(\mu) + (m - y)' \log(1 - \mu) \qquad \text{vectors}$$

glmb2 : Computes MLE by Fisher scoring and line search. Link function is passed as a parameter to this function. The routine computes MLEs, estimated variance-covariance matrix and saves iteration history

Some general comments

a. Function calls loglike, pt, gp. Starting values passed to function.

b. GLM iteration written previously as

$$\hat{\beta}_{i+1} = \hat{\beta}_i + \alpha (x' w(\hat{\beta}_i) x)^{-1} x' w(\hat{\beta}_i) \Delta(\hat{\beta}_i) \{ y^* - \mu(\hat{\beta}_i) \}$$

but we have to interpret y^* as the standardized counts with elements y_i / m_i (proportions)! In my routine, I factored the m_i out of $w(\hat{\beta}_i) \Delta(\hat{\beta}_i)$ and replaced $y_i^* - \mu_i(\hat{\beta}_i)$ by $y_i - m_i \mu_i(\hat{\beta}_i)$ so that the score function is defined in terms of the counts rather than proportions. The iteration is implemented in a WHILE loop. Iteration continues until either $\sqrt{\|\hat{\beta}_{i+1} - \hat{\beta}_i\|^2}$ or $|L(\hat{\beta}_{i+1}) - L(\hat{\beta}_i)|$ is small or until a prespecified # of iterations is exceeded. The optimal α is found by a line search over a grid with $-1 \le \alpha \le 2$.

c. Calculations use vector or matrix expressions

d. I don't invert the information matrix in the iteration. It is more stable to define the increment

$$(x' w(\hat{\beta}_i) x)^{-1} x' w(\hat{\beta}_i) \Delta(\hat{\beta}_i) \{ y^* - \mu(\hat{\beta}_i) \}$$

via solving the system

$$(X'\omega(\hat{\beta}_i)X)\,\delta_{i+1} = X'\omega(\hat{\beta}_i)\Delta(\hat{\beta}_i)\{y^{+} - u(\hat{\beta}_i)\}$$

for δ_{i+1}, then setting

$$\hat{\beta}_{i+1} = \hat{\beta}_i + \alpha\,\delta_{i+1}$$

and find the optimal α. In the program, I update the score function (i.e. the right hand side of above equation), and solve for the Fisher increment δ_{i+1} using Matlab's built-in function for solving linear equations.

This discussion is not meant to be complete, but rather to give you an idea of the logic that goes into creating the suite of programs!

To use these programs in an analysis, I typically write a MATLAB script (i.e. program) that enters the data, defines the appropriate data structures needed for ermle2. In the Matlab computing handout, I have included a script to fit Binomial response models to the Beetle mortality data, described below. Let's take a look at the data, the script and the associated output generated by the script.

Example 4.5 Mortality of confused flour beetles

The aim of an experiment originally reported by Strand (1930) and quoted by Bliss (1935) was to assess the response of the confused flour beetle, *Tribolium confusum*, to gaseous carbon disulphide (CS_2). In the experiment, prescribed volumes of liquid carbon disulphide were added to flasks in which a tubular cloth cage containing a batch of about thirty beetles was suspended. Duplicate batches of beetles were used for each concentration of CS_2. At the end of a five-hour period, the proportion killed was recorded and the actual concentration of gaseous CS_2 in the flask, measured in mg/l, was determined by a volumetric analysis. The mortality data are given in Table 4.2.

Table 4.2 The number of beetles killed, y, out of n exposed to different concentrations of gaseous carbon disulphide

Concentration of CS_2	Replicate 1		Replicate 2	
	y	n	y	n
49.06	2	29	4	30
52.99	7	30	6	30
56.91	9	28	9	34
60.84	14	27	14	29
64.76	23	30	29	33
68.69	29	31	24	28
72.61	29	30	32	32
76.54	29	29	31	31

In a number of articles that refer to these data, the responses from the first two concentrations are omitted because of apparent non-linearity. Bliss himself remarks that

> ... in comparison with the remaining observations, the two lowest concentrations gave an exceptionally high kill. Over the remaining concentrations, the plotted values seemed to form a moderately straight line, so that the data were handled as two separate sets, only the results at 56.91 mg of CS_2 per litre being included in both sets.

However, there does not appear to be any biological motivation for this and so here they are retained in the data set.

Combining the data from the two replicates and plotting the empirical logit of the observed proportions against concentration gives the graph shown in Fig. 4.4.

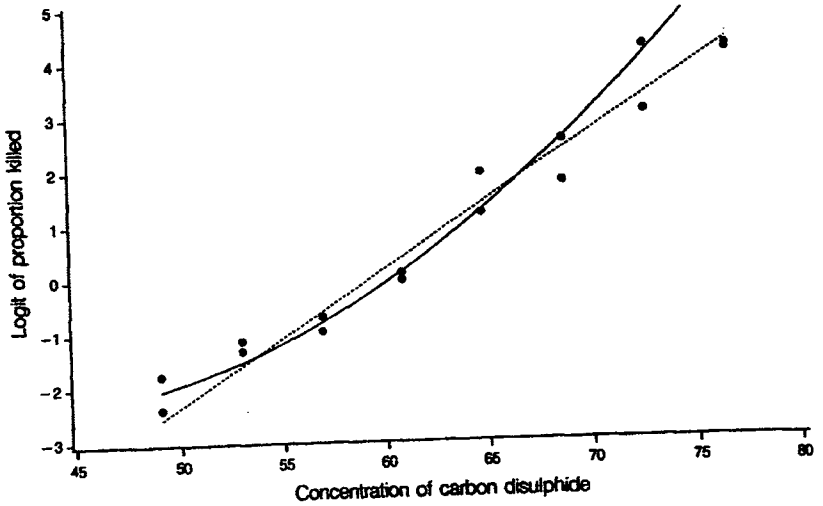


Fig. 4.4 Plot of the logistic transform of the proportion of beetles killed against concentration of CS_2 with the fitted linear (·····) and quadratic (——) logistic regression lines.

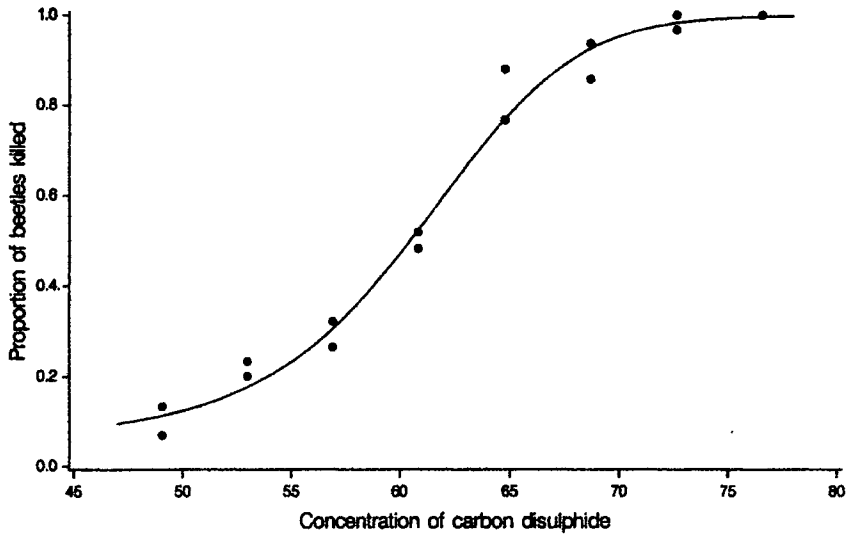


Fig. 4.5 Relationship between the proportion of beetles killed and the concentration of CS_2, with the fitted quadratic logistic regression model.

The description does not include the entire discussion. The important points are that

a) data from the two replicates were combined (at a given concentration)

b) A logistic regression was considered, but a plot of the empirical logits against concentration suggested a quadratic relationship

c) The model

$$\text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 \text{conc} + \beta_2 \text{conc}^2$$

was fitted to the data

d) A plot of the observed and fitted proportions from the quadratic fit shows fairly close agreement - see above

I will also fit the quadratic model, but will also consider probit and complementary log-log links. Let us take a look at the output!

Smoothing or Modelling Rates

I emailed to you a chapter from Selvin's book "Practical Biostatistical Methods" dealing with Poisson regression analysis. The chapter has a nice discussion of analyzing rates, which are a primary focus in a considerable number of epidemiologic analyses.

Selvin discusses a variety of basic summaries and provides a number of nice examples. I will use the following example to highlight important points. See Selvin for a more complete discussion of ideas and methods.

The following table compares Hodgkins disease mortality rates in males to corresponding mortality rates in females for residents of California in 1989.

Hodgkins disease mortality data from California for males and females, 1989

	Males			Females			
Age	Person-Years	Deaths	Rates ($\hat{r}_{i1}$)	Person-Years	Deaths	Rates ($\hat{r}_{i2}$)	$\hat{r}_{i2}/\hat{r}_{i1}$
30–34	1,299,868	55	4.23	1,300,402	37	2.85	0.67
35–39	1,240,595	49	3.95	1,217,896	29	2.38	0.60
40–44	1,045,453	38	3.63	1,045,801	23	2.20	0.61
45–49	795,776	26	3.26	810,260	12	1.48	0.45
50–54	645,991	19	2.94	665,612	7	1.05	0.36
55–59	599,729	17	2.83	633,646	12	1.89	0.67
60–64	568,109	22	3.87	650,686	9	1.38	0.36
65–69	506,475	21	4.15	600,455	19	3.16	0.76
70–74	368,751	18	4.88	474,609	13	2.74	0.56
75–79	252,581	11	4.36	376,781	14	3.72	0.85
80–84	140,053	10	7.14	255,412	5	1.96	0.27
85+	81,850	4	4.87	313,603	3	1.40	0.29
Total	7,545,231	290	3.84	8,345,163	183	2.19	0.58

Note: Rates per 100,000 person-years.

To make comparisons between sexes or between age categories, some form of standardization of counts (deaths) is needed - it is not appropriate to simply compare counts ignoring the differences in group sizes. In this example the counts are standardized in the form of counts per 100,000 person-years, a concept that I will describe in words, but for which Selvin gives careful explanation. The number of person years per age-by-sex categories is approximately the number of individuals in that cell, with a caveat that some individuals "switch" age categories during the study period. In the person-year calculation, you contribute 1 year to a specific age category if you don't switch categories during the study period. If you do switch categories, the two age categories split the 1 year proportionally. If you die, your age at death is used to identify which age category receives credit.

So for example, the table reports 55 deaths among males in ages 30-34. The total person-years for this category is 1,299,868. To calculate the rate per 100,000 PY (person years) use

$$\text{observed count} = \text{rate per } 100{,}000 \text{ py} * \frac{\text{observed \# py}}{100{,}000}$$

↑ observed # of 100,000 unit py

or

$$\text{rate per } 100{,}000 \text{ py} = \text{observed count} * \frac{100{,}000}{\text{observed \# py}}$$

For this category

$$\text{rate} = 55 * \frac{100{,}000}{1{,}299{,}868} = 4.23 \quad (\text{per } 100{,}000 \text{ py})$$

The summaries for other categories follow analogously.

This table might be described as containing the raw rates or unadjusted rates, stratified by sex and age.

Inference on raw rates if fairly straightforward. Suppose

$$Y_{ij} = \#\text{ deaths for age group } i, \text{ sex } j$$

we might assume

$$Y_{ij} \sim \text{indep Poisson}(\mu_{ij})$$

where

$$\mu_{ij} = r_{ij} \, c_{ij}$$

μ_{ij}: expected count

r_{ij}: population rate per 100,000 py

$$c_{ij} = \frac{\#\text{ py for age } i, \text{ sex } j}{100{,}000} \equiv \frac{py_{ij}}{10^5}$$

i.e. apply formula relating observed counts & observed rates to expected counts and expected rates.

With this model, it is usually the case that the c_{ij}'s are treated as fixed. The outcome, how many die, is treated as random but dependent on the # of person years.

With no constraints on the r_{ij} (or μ_{ij}), the MLE of μ_{ij} is the observed count:

$$\hat{\mu}_{ij} = \hat{r}_{ij}\, c_{ij} = y_{ij}$$

so the MLE of r_{ij} satisfies

$$\hat{r}_{ij} = \frac{y_{ij}}{c_{ij}}$$

Well, this is what we had before, so nothing new here! However, using the Poisson model we know

$$E(\hat{r}_{ij}) = \frac{1}{c_{ij}} E(y_{ij}) = \frac{\mu_{ij}}{c_{ij}} = r_{ij} \qquad \text{(unbiased)}$$

and

$$\text{var}(\hat{r}_{ij}) = \frac{1}{c_{ij}^2} \text{var}(y_{ij}) = \frac{\mu_{ij}}{c_{ij}^2} = \frac{r_{ij}}{c_{ij}}$$

We can estimate

$$SD(\hat{r}_{ij}) = \sqrt{\frac{r_{ij}}{c_{ij}}}$$

using

$$SE = \sqrt{\frac{\hat{r}_{ij}}{c_{ij}}}$$

given that $Y_{ij} \sim \text{Poisson}(\mu_{ij})$, exact CI's exist for μ_{ij}. This leads to an exact CI for $\Gamma_{ij} = \mu_{ij} / C_{ij}$. Alternatively since a Poisson distribution is approximated by a normal (assuming μ_{ij} is large), a simple approximate 95% CI for Γ_{ij} is just

$$\hat{\Gamma}_{ij} \pm 1.96\, SE$$

or

$$\hat{\Gamma}_{ij} \pm 1.96 \sqrt{\frac{\hat{\Gamma}_{ij}}{C_{ij}}}$$

For the single category we looked at earlier $\hat{\Gamma}_{ij} = 4.23$ and

$$C_{ij} = \frac{1{,}299{,}868}{100{,}000} = 13.00$$

Our approx 95% CI is

$$4.23 \pm \underbrace{1.96 \sqrt{\frac{4.23}{13}}}_{1.12}$$

$SE(\hat{\Gamma}_{ij}) = 0.57$

or (3.11, 5.35) per 100,000 py.

We may also consider unadjusted rates for each age group, ignoring sex. These can be obtained two ways. First, and most easily, we just collapse counts and PY across sexes, and then use earlier methods. To illustrate, consider age 30-34

$$\hat{r}_i = \frac{y_{i\cdot}}{c_{i\cdot}}$$

$\hat{r}_i$: estimated rate age i per 10^5 PY

$y_{i\cdot}$: observed # deaths age i $= y_{i1} + y_{i2}$

$c_{i\cdot}$: # PY age i $= \frac{PY_{i1} + PY_{i2}}{10^5} = c_{i1} + c_{i2}$

For this age category

$$y_{i\cdot} = 55 + 37 = 92$$

$$c_{i\cdot} = \frac{1{,}299{,}868 + 1{,}300{,}402}{10^5} = \frac{2{,}600{,}270}{10^5} = 26$$

so

$$\hat{r}_i = \frac{92}{26} = 3.54$$

as before

$$SE(\hat{r}_i) = \sqrt{\frac{\hat{r}_i}{c_{i\cdot}}} = \sqrt{\frac{3.54}{26}} = .37$$

An alternative way to view $\hat{r}_i$ is as a weighted average of the sex-specific rates for males $\hat{r}_{i1}$ and females $\hat{r}_{i2}$. Convince

yourself (not that hard!) that

$$\hat{r}_i = a_{i1}\hat{r}_{i1} + a_{i2}\hat{r}_{i2}$$

where $$a_{ij} = \frac{c_{ij}}{c_{i1}+c_{i2}} = \frac{PY_{ij}}{PY_{i1}+PY_{i2}} \quad (a_{i1}+a_{i2}=1)$$

Except for the 10^5 multiplier, c_{ij} is essentially a sample size so the weights reflect the relative sizes of males & females within an age group.

Recognizing that the (marginal) raw rate for each age-group is a weighted average of sex-specific rates, where the weights vary from age-to-age, is it reasonable to ask whether a comparison across ages makes sense if the sex distributions vary markedly by age. For example consider a comparison of ages 50-54 and 85+. The table below gives the sex specific rates for each age group, and the overall rate ignoring sex:

age	M	F	overall
50-54	2.94	1.05	1.98
85+	4.87	1.40	1.77

We observe a classical illustration of Simpson's paradox!

The sex specific rates for males & females are higher for 85+ year olds than for 50-54 year olds, but the overall rate for 85+ year olds is lower than for 50-54 year olds.

The explanation why is clear - among 50-54 year olds males & females are found in close to a 50:50 ratio, but an overwhelming percentage of very old people are women. Thus the overall rate for 85+ year olds is heavily influenced by the lower rates observed for women, whereas the overall rate for 45-49 year olds is about half-way between the male & female rate!

Note that Simpson's paradox can not happen if the male and female weights are constant across age categories.

Also note that any concern over comparing ages ignoring sex also applies to a comparison of sexes within an age group because these may be obtained implicitly by averaging over all other factors not listed (with weights that vary across sexes)

One approach to this problem is to use a specific set of weights <u>for all</u> age categories. That is, compute a rate via

$$\tilde{r}_i = a_1 \hat{r}_{i1} + a_2 \hat{r}_{i2} \qquad a_1 + a_2 = 1$$

where as before, $\hat{r}_{i1}$ and $\hat{r}_{i2}$ are sex-specific rates within the age category. A popular choice might be to use the weights for a particular age category, say 55-59 or to use equal weights $a_1 = a_2 = .50$. We would interpret the rate as what we expect the overall age group rate to be if this age group had the same sex distribution as the 55-59's or, in the latter case, if males and females contribute the same number of person years.

This is an example of what is called <u>direct standardization of rates</u>.

Under the Poisson model, inference based on $\tilde{r}_i$ are straightforward because $\hat{r}_{i1}$ & $\hat{r}_{i2}$ are independent. So for example

$$\text{var}(\tilde{r}_i) = a_1^2 \text{var}(\hat{r}_{i1}) + a_2^2 \text{var}(\hat{r}_{i2})$$

$$= a_1^2 \frac{r_{i1}}{c_{i1}} + a_2^2 \frac{r_{i2}}{c_{i2}}$$

This can be easily estimated, for example

$$SE(\tilde{r}_i) = \sqrt{a_1^2 \frac{\hat{r}_{i1}}{c_{i1}} + a_2^2 \frac{\hat{r}_{i2}}{c_{i2}}}$$

leading to an approx CI of the form

$$\hat{r}_i \pm 1.96\, SE(\tilde{r}_i)$$

In general, <u>direct standardization</u> of rates can be significantly more complex than illustrated here, but in all cases the estimated rate has the form of a weighted average of stratum-specific rates

$$\tilde{r}_i = \sum_{j=1}^{k} a_j \hat{r}_{ij} \qquad a_j > 0 \quad a_1 + a_2 + \cdots + a_k = 1$$

The number k of strata can be large. For example, dialysis providers compute standardized mortality rates, where a_j's correspond to weights associated with the USRDS (US renal disease service). The strata correspond to all possible combinations of age (categorized), race, cause of end stage renal disease (ESRD) & sex.

One potential limitation of <u>direct standardization</u> is that the variability can be relatively large if certain stratum-specific rates $\hat{r}_{ij}$ are estimated poorly and these receive large weight in the weighted average. As an alternative, one might consider

an indirect standardization of rates.

A classic example of <u>indirect standardization</u> is the standardized mortality rate (SMR). In an SMR, one identifies a reference population. The rate for a specific group is computed, and divided by what the rate for the reference population would have been had the reference population have the same weight as the specific group. For example, suppose in the Hodgkin's study we have data from each state for 1989 and we know that in the US as a whole the sex-specific mortality rates for males & females were 3.20 and 2.70, respectively. The SMR for California is computed as follows. First, compute CA raw rate (ignoring sex and age) as a weighted average of the sex specific rates, which are 3.84 & 2.19 - see the total row in Table.

$$\hat{r}_{CA} = a_{M,CA}\,\hat{r}_{M,CA} + a_{F,CA}\,\hat{r}_{F,CA}$$

$$= .4748 * 3.84 + .5252 * 2.19$$

$$= 2.9955$$

$$a_{M,CA} = \frac{PY_{M,CA}}{PY_{M,CA} + PY_{F,CA}}$$

$$= \frac{7,545,231}{15,890,394}$$

$$= .4748$$

$$a_{F,CA} = 1 - .4748$$

$$= .5252$$

Next, compute the **US** rate using the CA weights

$$\hat{r}_{US} = .4748 * 3.20 + .5252 * 2.70$$

$$= 2.9647$$

Here, the US is the <u>reference population</u>

$\hat{r}_{US}$ is the US rate standardized relative to CA. The CA SMR is

$$\hat{SMR} = \frac{\hat{r}_{CA}}{\hat{r}_{US}} = \frac{2.9955}{2.9647} = 1.01$$

That is, we expect CA's mortality rate to be 1.01 the US rate had the US had the same sex distribution as CA.

The same computation could then be repeated for each state, leading to 50 SMR's. Each state would standardize relative to the US, but by using <u>their own state specific weights</u> in the calculation of the denominator. Thus, each state is compared to the US, but using state specific weights in the assessment of the US rate. Put another way, the denominator in SMR varies from state to state.

In practice, the SMRs might be compared across states, but this can be problematic - one state can have a lower SMR than another state yet have higher stratum specific rates - Simpson's paradox strikes again!

Inference on SMRs is fairly straightforward; see Selvin for details.

Smoothing Rates

We have described a variety of summaries involving raw rates, weighted averages of raw rates or ratios of weighted averages of raw rates. One common feature of each summary is that there was no "structure" assumed for the rates. For example, in the Hodgkin's disease analysis we have 12 age * 2 sex = 24 age by sex rates r_{ij} and these were estimated without making any assumptions about how they might be dependent on age and sex i.e. without a specific model for the r_{ij}'s. One might expect that better estimates of the r_{ij}'s can be obtained (compared to using the "unrestricted" MLE $\hat{r}_{ij} = y_{ij}/c_{ij}$) through the fitting of an appropriate model for the rates. Then estimates of the r_{ij}'s (MLEs) under the model could be used, if desired, in direct and indirect standardized rates, instead of raw rates $\hat{r}_{ij}$.

One possible model for the rates is the 2-way interaction model

$$\log(r_{ij}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

α_i: year i effect; β_j: sex j effect; $(\alpha\beta)_{ij}$: year-by-sex interaction

This log-linear model makes no restrictions on the r_{ij}'s.

Thus, if we assume $y_{ij} \sim$ indep Poisson (μ_{ij}) where

$$\mu_{ij} = r_{ij} c_{ij}$$

as before, then the MLE's are just the raw rates $\hat{r}_{ij} = y_{ij} / c_{ij}$.

The last column of the Hodgkin's data table provides the ratio of the female to male mortality rates by age: $\hat{r}_{i2} / \hat{r}_{i1}$. These ratios are roughly constant, or equivalently

$$\log \left(\frac{\hat{r}_{i2}}{\hat{r}_{i1}} \right) = \log \hat{r}_{i2} - \log \hat{r}_{i1} \doteq \text{constant}$$

If we plotted $\log \hat{r}_{i2}$ and $\log \hat{r}_{i1}$ against the age category, we should see roughly parallel profiles - suggestive of no-interaction between age and sex on the log rates.

This leads us to suggest

$$\log (r_{ij}) = \mu + \alpha_i + \beta_j$$

as a model for the rates. Coupling this with the "sampling model"

$$y_{ij} \sim \text{indep Poisson} (\mu_{ij})$$

where

$$\mu_{ij} = r_{ij} c_{ij}$$

we have a special type of GLM model - Poisson responses where the means satisfy

$$\begin{aligned}\log(\mu_{ij}) &= \log(c_{ij}\, r_{ij}) \\ &= \log(c_{ij}) + \log(r_{ij}) \\ &= \log(c_{ij}) + \mu + \alpha_i + \beta_j\end{aligned}$$

Here the first term $\log(c_{ij})$ is a constant that varies across cells - in essence, it is a predictor in the model but with a known regression coefficient of 1. This first term is known as an <u>offset</u>. The other parts of the mean specification are parameters to be estimated. Our model is a <u>Poisson GLM with a log-link and an offset</u>. Most packages (SAS & R) allow for offsets in a GLM. Further, standard theory that we outlined for GLMs without offsets extends to models with offsets "seamlessly".

Noting that $c_{ij} = PY_{ij}/10^5$, the model for log-means can be written equivalently using $\log(PY_{ij})$ as the offset.

$$\begin{aligned}\log(\mu_{ij}) &= \log\left(\frac{PY_{ij}}{10^5}\right) + \mu + \alpha_i + \beta_j \\ &= \log(PY_{ij}) + \mu^* + \alpha_i + \beta_j\end{aligned}$$

where $\mu^* = \mu - \log(10^5)$. Because the model includes an intercept, any 2 offsets that differ by a constant correspond to same model

Standard output for this GLM will provide estimates, SE and CI for the means μ_{ij} rather than the rates Γ_{ij}. However using the multiplicative relationship

$$\mu_{ij} = \Gamma_{ij} C_{ij}$$

we can easily get corresponding summaries for Γ_{ij}. In particular

- $\tilde{\Gamma}_{ij} \equiv$ MLE of $\Gamma_{ij} = \dfrac{\tilde{\mu}_{ij}}{C_{ij}}$ ← MLE of μ_{ij}
- $SE(\hat{\Gamma}_{ij}) = \dfrac{1}{C_{ij}} SE(\hat{\mu}_{ij})$
- $\mu_{ij,low} \leq \mu_{ij} \leq \mu_{ij,up}$ with $100(1-\alpha)\%$ confidence $\Rightarrow$

$$\frac{\mu_{ij,low}}{C_{ij}} \leq \Gamma_{ij} \leq \frac{\mu_{ij,up}}{C_{ij}} \text{ with } 100(1-\alpha)\% \text{ confidence}$$

These summaries, with the exception of $SE(\hat{\mu}_{ij})$ are available in SAS

However, SE are available for the <u>linear predictor</u>

$$\hat{\eta}_{ij} = \log(C_{ij}) + \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j$$

Using the delta method, with

$$\hat{\mu}_{ij} = \exp(\hat{\eta}_{ij})$$

we can show

$$\text{var}(\hat{\mu}_{ij}) \doteq [\exp(\eta_{ij})]^2 \text{ var}\{\hat{\eta}_{ij}\}$$

Or

$$\text{Var}(\hat{u}_{ij}) \doteq u_{ij}^2 \, \text{Var}(\hat{n}_{ij})$$

Thus

$$SD(\hat{u}_{ij}) \doteq u_{ij} \, SD(\hat{n}_{ij})$$

So

$$\boxed{SE(\hat{u}_{ij}) \doteq \hat{u}_{ij} \, SE(\hat{n}_{ij})}$$

These basic calculations are illustrated in the SAS document.

In SAS code, I computed $SE(\hat{r}_{ij})$ via

$$SE(\hat{r}_{ij}) = \hat{r}_{ij} \, SE(\hat{n}_{ij})$$

This follows because

$$SE(\hat{r}_{ij}) = \frac{1}{c_{ij}} SE(\hat{u}_{ij}) \qquad \text{previous page}$$

$$= \frac{1}{c_{ij}} \hat{u}_{ij} \, SE(\hat{n}_{ij}) \qquad \text{above}$$

$$= \hat{r}_{ij} \, SE(\hat{n}_{ij})$$