

# Statistical Foundations of Generalized Linear Models

$$y \sim f(\mu(\mathbf{X}\boldsymbol{\beta}), \nu)$$

- Exponential Family
- Link Functions
- Variance functions

## Introduction

Linear model make a series of assumptions:

- The mean response is a linear function of the effects  $\mu_i = x'_i\beta$ .

Easy to model effects that result in a constant shift.

However, it might be more reasonable to assume that effects are multiplicative.

Assumption of linearity in the parameters has also lead to heavy reliance on polynomials to model responses.

However, many processes, such as growth, tend to reach an asymptote which are not modeled effectively using polynomials.

- Variability doesn't depend on the mean  $\text{var}(y_i) = \sigma^2$ .
- Normal distribution

## Medical Trial Example:

- A medical trial where interest is in the impact of a drug treatment on the chance of recovery.
- dependent variable:  $(y_{ijk})$   
Recovery (0=no or 1=yes)
- expected response is the chance of recovery

$$E(y_{ijk}) = \Pr(\text{Recovery}).$$

Several patient populations with different chances of recovery without treatment:

- A 10% chance of recovery without treatment
- B 50% chance of recovery without treatment
- C 90% chance of recovery without treatment

Suppose we have an effective treatment that increased the chance for recovery in population A to 90%.

Under a linear effect model and assuming that the “effect” of the treatment is the same for then we would expect:

- A  $10\% \rightarrow 10 + 80 = 90\%$  Recovery
- B  $50\% \rightarrow 50 + 80 = 130\%$  Recovery!
- C  $90\% \rightarrow 90 + 80 = 170\%$  Recovery!.

We will need a “link” function between effects and expected responses.

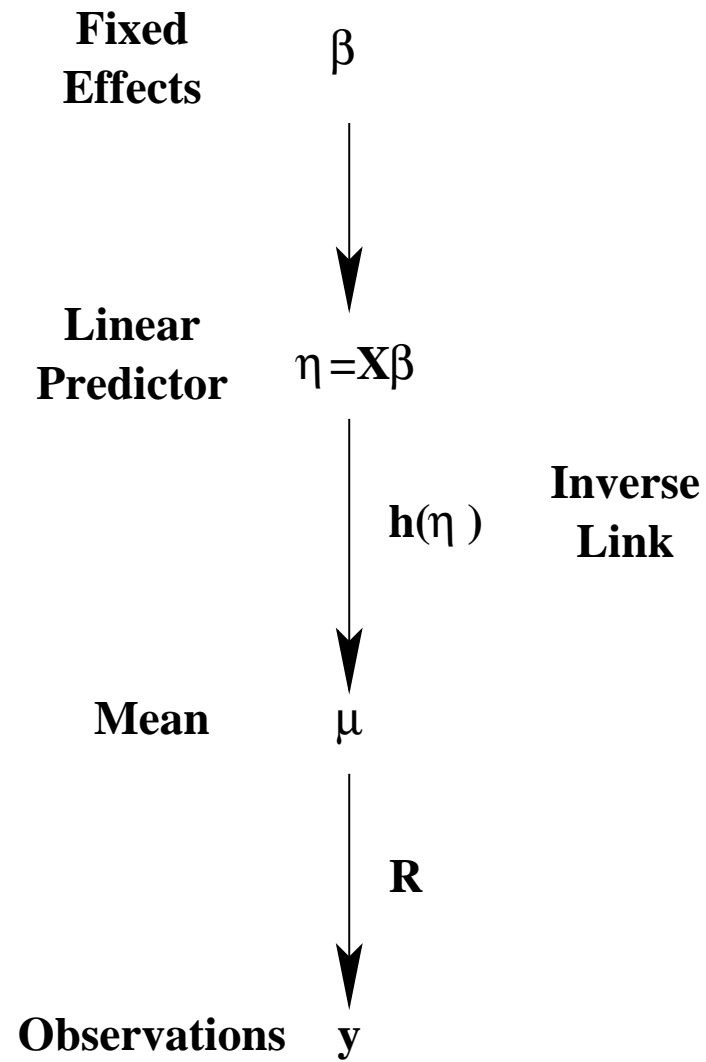
The distribution of recovery for a randomly selected patient in population  $i$  receiving treatment regimen  $j$  will be

$$y_{ijk} \sim \text{Bin}(1, \mu_{ij}).$$

One of the properties of a binomial distribution is that the  $\text{var}(y_{ijk}) = \mu_{ij}(1 - \mu_{ij})$

Which implies that residual variance depends on the mean.

We will now lay down the statistical foundations.



# Linear Predictor

$$\eta = X\beta$$

- The linear predictor is the systematic component.
- It differs from a model equation by dropping the residual ( $e$ ) term.
- Unlike a linear model the mean response does not have to be  $\eta = X\beta$ .
- It does need to be a function of the linear predictor. The function is called an INVERSE LINK FUNCTION.

## (Inverse) Link Function

- A inverse link function is the function which maps a linear predictor to a mean

$$\mu_i = h(\eta_i).$$

- The function which maps from a mean to a linear predictor is called a link function

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

- Link functions will only exist when there is a 1-1 mapping between linear predictors and the mean.

- An inverse link function allows for the effect of a one unit change to depend on where it is expressed.
- Medical Trial: A reasonable link function to consider would be logit link:

$$\mu_i = h(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}.$$

The effect of a one unit change is given below:

$\eta_i$	$h(\eta_i)$	$h(\eta_i + 1)$	$h(\eta_i + 1) - h(\eta_i)$
-3	.05	.12	.07
0	.50	.73	.23
3	.95	.98	.03

- Order is preserved for the common univariate link functions

$$a > b \Rightarrow h(a) > h(b).$$

- The magnitude of a difference can differ on the linear predictor and mean scales.
- The direction of the difference will be the same.

## Common Link Functions

Distribution	Link	Inverse Link
Normal	Identity	Identity
Binomial	$\ln[\mu/(1 - \mu)]$	$e^\eta/(1 + e^\eta)$
	Probit	$\Phi(\eta)$
Poisson	$\ln(\mu)$	$e^\eta$
Gamma	$1/\mu$	$1/\eta$
	$\ln(\mu)$	$e^\eta$

## Selection of link functions:

- The ability to interpret treatment effects.
- Simple functions that have a functional interpretation will be easier to interpret.

For example, the use of a nonlinear growth function to model growth instead of polynomials.

- Select a link function that will not lead to problems in the analysis.

For example, a link function that can lead to negative probabilities is usually a poor choice for modeling chance of recovery.

# Variance Function

- The variance function allows the residual variability to depend on the mean.
- The residual variance can also include an over-dispersion parameter  $\phi$ .
- The residual variance is given by

$$\text{var}(y_i) = \phi^2 v(\mu_i).$$

## Common Variance Functions

Distribution	$v(\mu)$
Normal	1
Binomial	$\mu(1 - \mu)$
Poisson	$\mu$
Gamma	$\mu^2$

## Exponential Family

- Generalized linear models are not restricted to distributions which are members of the exponential family.

The Weibull distribution being a counter example.

In addition, quasi-likelihood can be used when no specific distribution is given.

However, many of common distributions are members of the exponential family and the resulting estimating formulas have a nice form.

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}, \phi) &= \exp \left\{ \frac{\mathbf{y}'\boldsymbol{\theta} - b(\boldsymbol{\theta})}{\phi^2} - c(\mathbf{y}, \phi) \right\} \\ &= \exp \left\{ \sum_{i=1}^N \frac{y_i\theta_i - b(\theta_i)}{\phi^2} - c(\mathbf{y}, \phi) \right\} \end{aligned}$$

where  $\boldsymbol{\theta}$  is a  $N \times 1$  vector of canonical parameters and  $\phi$  is a scalar scale parameter.

Before deriving the maximum likelihood estimating equations. We will obtain some useful results.

The log likelihood of the canonical parameters is

$$\ell = \sum_{i=1}^N \frac{y_i \theta_i - b(\theta_i)}{\phi^2} - c(\mathbf{y}, \phi).$$

The score function of the canonical parameters is obtained by taking the first partial derivatives of the log likelihood

$$\frac{\partial \ell}{\partial \theta_i} = \frac{y_i - \frac{\partial b(\theta_i)}{\partial \theta_i}}{\phi^2}.$$

Recalling that the expected value of a score function equal 0 yields

$$E(y_i) = \mu_i = \frac{\partial b(\theta_i)}{\partial \theta_i}.$$

The information is then obtained

$$\frac{\partial^2 \ell}{\partial \theta_i^2} = - \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} \frac{1}{\phi^2}$$
$$i(\theta_i) = - \mathbb{E} \left[ \frac{\partial^2 \ell}{\partial \theta_i^2} \right] = \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} \frac{1}{\phi^2}$$
$$i(\theta_i) = \frac{\partial \mu_i}{\partial \theta_i} \frac{1}{\phi^2}.$$

Recalling that the variance of a score function is equal to  $i(\theta_i)$

$$i(\theta_i) = \text{var}(y_i) \frac{1}{\phi^4}$$

$$\text{var}(y_i) \frac{1}{\phi^4} = \frac{\partial \mu_i}{\partial \theta_i} \frac{1}{\phi^2}$$

$$\text{var}(y_i) = \frac{\partial \mu_i}{\partial \theta_i} \phi^2$$

$$\phi^2 \nu(\mu_i) = \phi^2 \left( \frac{\partial \mu_i}{\partial \theta_i} \right)$$

$$\nu(\mu_i) = \frac{\partial \mu_i}{\partial \theta_i} = \left( \frac{\partial \theta_i}{\partial \mu_i} \right)^{-1} .$$

$$\begin{aligned}\frac{\partial \ell}{\partial \boldsymbol{\beta}'} &= \frac{\partial \ell}{\partial \boldsymbol{\theta}'} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}'} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}'} \\ &= (\mathbf{y} - \boldsymbol{\mu})' \text{Diag} (\phi^2 \nu(\mu_i))^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}'}\end{aligned}$$

$$\begin{aligned}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}'} &= \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}'} \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}'} \\ &= \text{Diag} \left( \frac{\partial \mu_i}{\partial \eta_i} \right) \mathbf{X}\end{aligned}$$

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}'} = (\mathbf{y} - \boldsymbol{\mu})' \text{Diag}(\phi^2 \nu(\mu_i))^{-1} \text{Diag}\left(\frac{\partial \mu_i}{\partial \boldsymbol{\eta}_i}\right) \mathbf{X}$$

$$= (\mathbf{y} - \boldsymbol{\mu})' \mathbf{R}^{-1} \mathbf{H} \mathbf{X}$$

$$\mathbf{R} = \text{var}(\mathbf{y})$$

$$\mathbf{H} = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}'}$$

The information matrix is obtained directly as the variance of the score function without the need for any additional partial derivatives:

$$\begin{aligned}i(\boldsymbol{\beta}) &= \text{var} \left( \frac{\partial \ell}{\partial \boldsymbol{\beta}} \right) \\&= \text{var} \left( \mathbf{X}' \mathbf{H}' \mathbf{R}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right) \\&= \mathbf{X}' \mathbf{H}' \mathbf{R}^{-1} \text{var}(\mathbf{y}) \mathbf{R}^{-1} \mathbf{H} \mathbf{X} \\&= \mathbf{X}' \mathbf{H}' \mathbf{R}^{-1} \mathbf{H} \mathbf{X}\end{aligned}$$

The Fisher scoring estimating equations can now be obtained from the score function and the information matrix:

$$i(\boldsymbol{\theta}^{[i]})(\boldsymbol{\theta}^{[i+1]} - \boldsymbol{\theta}^{[i]}) = \frac{\partial \ell}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}}$$

$$[\mathbf{X}'\mathbf{H}'\mathbf{R}^{-1}\mathbf{H}\mathbf{X}] (\boldsymbol{\beta}^{[i+1]} - \boldsymbol{\beta}^{[i]}) = \mathbf{X}'\mathbf{H}'\mathbf{R}^{-1}(\mathbf{y} - \boldsymbol{\mu}^{[i]})$$

$$[\mathbf{X}'\mathbf{H}'\mathbf{R}^{-1}\mathbf{H}\mathbf{X}] \boldsymbol{\beta}^{[i+1]} = \mathbf{X}'\mathbf{H}'\mathbf{R}^{-1}(\mathbf{y} - \boldsymbol{\mu}^{[i]} + \mathbf{H}\boldsymbol{\eta}^{[i]})$$