Chapter 1

Introduction-Generalized Linear Models

1.1 The Basic Components

- Generalized linear models provide a unifying methodology for many common statistical analyses useful in biostatistics including:
  - regression
  - analysis of variance
  - analysis of covariance
  - log linear models
  - logistic regression
  - analysis of rates
  - longitudinal data analysis
The first component of a generalized linear model is the probability or random component which states that we have realized values $y_1, y_2, \ldots, y_n$ of random variables $Y_1, Y_2, \ldots, Y_n$ assumed independent with the probability density function of $Y_i$ given by

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp \left\{ \frac{[y_i \theta_i - b(\theta_i)]}{a(\phi)} + c(y_i, \phi) \right\}$$

- It is easy to show that under weak conditions on $f_{Y_i}$:

$$\mu_i = E(Y_i) = b^{(1)}(\theta_i) \quad \text{where} \quad b^{(1)}(\theta_i) = \left. \frac{d b(\theta)}{d \theta} \right|_{\theta = \theta_i}$$

$$V_i = \text{var}(Y_i) = b^{(2)}(\theta_i)a(\phi) \quad \text{where} \quad b^{(2)}(\theta_i) = \left. \frac{d^2 b(\theta)}{d \theta^2} \right|_{\theta = \theta_i}$$

- Thus the mean depends only on $\theta_i$, the canonical parameter. The variance depends on a function of the canonical parameter (called the variance function) and the dispersion or scale parameter $\phi$.

- These distributional assumptions constitute the probability or random component of a generalized linear model.
\begin{itemize}
\item \textbf{example:} For the normal distribution we have
\[
(2\pi\sigma^2)^{-\frac{1}{2}} \exp \left\{ -\frac{(y_i - \mu_i)^2}{2\sigma^2} \right\} = \exp \left\{ -\frac{y_i^2}{2\sigma^2} + \frac{y_i\mu_i}{\sigma^2} \frac{\mu_i^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2) \right\}
\]
\[
= \exp \left\{ \frac{y_i\mu_i - \mu_i^2}{2\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2) \right\}
\]

Thus:
\[
\theta_i = \mu_i
\]
\[
b(\theta_i) = \frac{\mu_i^2}{2}
\]
\[
= \frac{\theta_i^2}{2}
\]
\[
c(y_i, \phi) = -\frac{y_i^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2)
\]
\[
a(\phi) = \sigma^2
\]
\end{itemize}

\begin{itemize}
\item As an example of the mean variance relationship we have for the normal distribution:
\[
b(\theta_i) = \frac{\theta_i^2}{2} \implies b^{(1)}(\theta_i) = \theta_i \quad \text{so that} \quad E(Y_i) = \theta_i = \mu_i
\]
\[
b^{(2)}(\theta_i) = 1 \quad \text{so that} \quad \text{var}(Y_i) = \sigma^2
\]
\end{itemize}
• The second component of a generalized linear model is the systematic component in which a linear predictor is specified as

\[ \eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \]

○ The \( \beta_j \) are unknown parameters and the \( x_{ij} \) are values of covariates.

○ In the case of the normal distribution, we obtain analysis of variance, analysis of covariance and multiple regression.

○ For the binomial we obtain logistic regression while for the Poisson we obtain log linear models for contingency tables and the analysis of rates.
1.1. THE BASIC COMPONENTS

- The third component of a generalized linear model consists of a link between the random and systematic components.
  - The link is a function relating $\eta$ and $\mu$ and is given by
    \[ \eta_i = g(\mu_i) \]
  - Since $\mu_i = b^{(1)}(\theta_i)$ the link function also relates $\eta_i$ to $\theta_i$. The link function is required to be monotonic and differentiable.
  - While there are many possible link functions the most important are the canonical links defined by
    \[ \eta_i = \theta_i \]
  - In this case the link function is just the function $(b^{(1)})^{-1}$ since
    \[ \eta_i = g(b^{(1)}(\theta_i)) = \theta_i \quad \text{implies} \quad g = (b^{(1)})^{-1} \]
  - The importance of the canonical links is that there are simple sufficient statistics for the $\beta_j$ in this case.
  - Note that in some expositions the link is defined as $g(\eta_i) = \mu_i$.
  - **Example:** For the normal distribution we have $\theta_i = \eta_i$ which implies that
    \[ \theta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \]
    Since $\mu_i = \theta_i$ we have
    \[ \mu_i = E(Y_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \]
    which is the usual general linear model for multiple regression, analysis of variance and analysis of covariance.
Summary: In a generalized linear model we have $y_1, y_2, \ldots, y_n$ which are observed values of independent random variables $Y_1, Y_2, \ldots, Y_n$

- The distribution of $Y_i$ is

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i; \phi) \right\}$$

- The systematic model is specified by a linear predictor of the form

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

- The link between $\eta_i$ and $\mu_i = E(Y_i)$ is defined by

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

The link is called a canonical link if $\theta_i = \eta_i$. In this case $g = (b^{(1)})^{-1}$. 

---

CHAPTER 1. INTRODUCTION-GENERALIZED LINEAR MODELS