Tutorial 5

STAT3015/4030/7030 Generalised Linear Modelling

The Australian National University

Week 5, 2017

Overview

Basics of GLM

- Flexible generalization of ordinary linear regression
  - From normal errors to **exponential family errors**
  - From constant variance structure to **non-constant variance structure**
- Allow for exponential family as error distribution
- Generalize the idea of data transformation
- Allow the magnitude of variance to be a function of expectation
- Iterative Re-weighted Least Square Method (well developed)

GLM theory is formulated by John Nelder and Robert Wedderburn in 1970s.

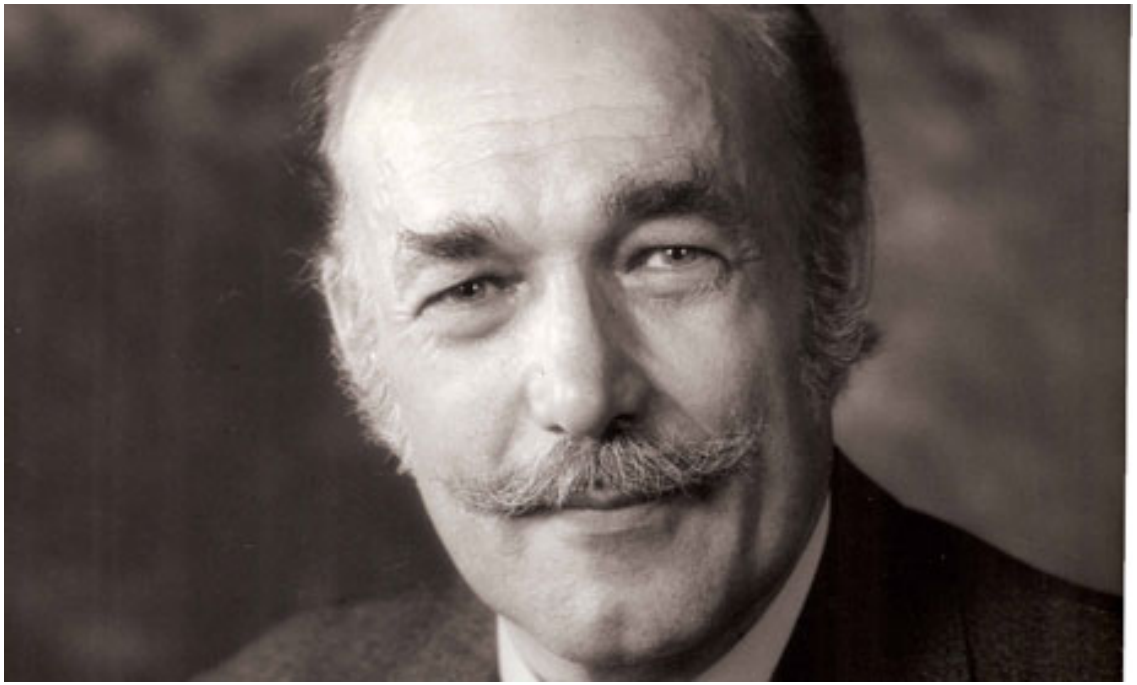

Figure 1: John Ashworth Nelder (8/10/1924 - 7/08/2010)

Weighted regression

Assume Y_i is the proportion of successful events among n_i trails. With n_i trails we have $n_i Y_i$ following a binomial distribution with $p = \pi(x_i)$ probability of success. If we use simple linear regression to model :

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

$$E(Y_i) = \pi(x_i) = \beta_0 + \beta_1 x_i$$

The variance of Y_i is clearly not constant.

$$Var(Y_i) = \frac{1}{n_i}(\beta_0 + \beta_1 x_i)(1 - \beta_0 - \beta_1 x_i)$$

(hint: use the variance of binomially distributed $n_i Y_i$)

Weighted regression

Data transformation is not sufficient to fix the problem of Heteroscedasticity. We can use the weighted linear regression. Let

$$\omega_i^2 = \frac{n_i}{(\beta_0 + \beta_1 x_i)(1 - \beta_0 - \beta_1 x_i)} = \frac{1}{Var(Y_i)}$$

then

$$\begin{aligned} E(\omega_i Y_i) &= \beta_0 \omega_i + \beta_1 x_i \omega_i, \\ Var(\omega_i Y_i) &= 1, \quad \forall\ i. \end{aligned}$$

Thus, the model reaches homoscedasticity.

Weighted regression

ω_i is not always equal to the reciprocal of the variance of the Y_i's.

Estimation of β's can be done by minimising the weighted sum of squares function:

$$d_\omega(\beta_0, \beta_1) = \sum_{i=1} \omega_i^2 (Y_i - \beta_0 - \beta_1 x_i)^2$$

(This relates to Question 1)

Exponential family of probability distributions

Any probability distribution with a density of the following form:

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

for some specified functions $b(\mu), c(\mu)$ and $d(y,\phi)$.

To find $b(\mu), c(\mu)$ and $d(y,\phi)$ of a given pmf or pdf we need to take natural logarithm and then apply exponential function.

The function $b(\mu)$ is often called the **canonical link** function.

Link functions

Apart from the canonical link functions, we have other commonly used link functions:

1. Logit: $g(p) = \log \frac{p}{1-p}$
2. Probit: $g(p) = \Phi^{-1}(p)$
3. Complementary log-log: $g(p) = \log(-\log(1-p))$

In GLM we will model $g(\boldsymbol{\mu}) = \boldsymbol{X}^T \boldsymbol{\beta}$, we need to do back transformation to get $\hat{\boldsymbol{\mu}}$.

(This relates to Question 2)