Introduction to Generalized Linear Models

Dr Rafael de Andrade Moral

https://rafamoral.github.io

Outline

- The normal model: A recap
- Models for binary data (odds, odds ratio, logit link, deviance)
- Models for binomial data (probit, cloglog –end of day 1)
- Models for multinomial data (generalized logit model)
- Models for count data (Poisson, offset, hnp –end of day 2)
- Extensions: overdispersion models (quasi-Poisson, Negbin, quasi-binomial, betabinomial)
- Extensions: zero-inflated models (ZIP, ZINB, hurdle –end of day 3)

The Normal Model: A Recap

The Normal Model: A Recap

- Y_i is a response variable associated with experimental or observational unit i
- We *assume* it comes from a certain probability distribution with pmf/pdf f and vector of parameters $\boldsymbol{\theta}$
- In general, one of the parameters in $\boldsymbol{\theta}$ is the mean of the distribution
- We also have predictors x_i we are interested in studying
- We may link it to a parameter of interest, typically the mean of the distribution

The Normal Model: A Recap

- For the normal model, we typically write for each observation y_i:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

  where $\varepsilon_i \sim \mathsf{N}(0, \sigma^2)$
- Each β coefficient represents the expected mean change in y for a 1-unit increase in its associated predictor
- We can show that, from the equation above, the expected value of Y_i is

$$\mathsf{E}[Y_i] = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

  and the variance is

$$\mathsf{Var}(Y_i) = \sigma^2$$

- Therefore, we are assuming the variance is constant

The Normal Model: A Recap

- The error notation has its advantages...
- However, let's switch things up a bit

Statistical Modelling

What is Statistical Modelling?

Statistical Modelling

What is a Statistical Model?

Statistical Modelling

It's all about Probability!

Statistical Modelling

Building blocks

1. Response variable (Y)
2. Probability distribution
3. Parameters of interest

Statistical Modelling

Building blocks

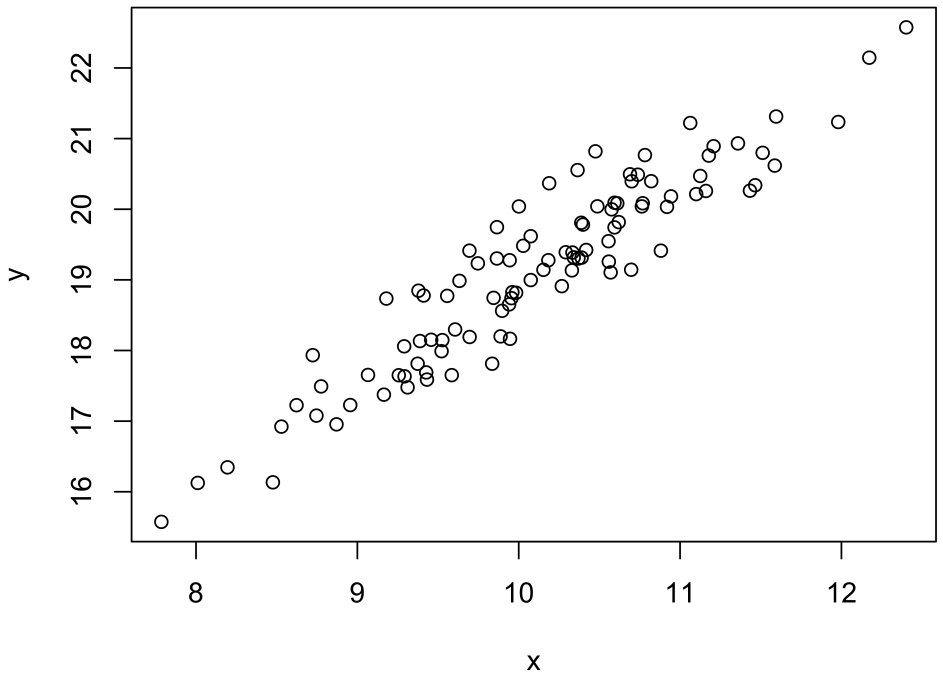
1. Response variable (Y)
2. Probability distribution
3. Parameters of interest $\leftarrow$ covariates / predictors

Statistical Modelling

- Y_i is a response variable associated with experimental or observational unit i
- We *assume* it comes from a certain probability distribution with pmf/pdf f and vector of parameters $\boldsymbol{\theta}$
- Very often one of the parameters in $\boldsymbol{\theta}$ is the mean of the distribution (or a function of the mean)
  - e.g.[1] for the normal distribution, $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$, where μ is the mean of the distribution
  - e.g.[2] for the Poisson distribution, $\boldsymbol{\theta} = \mu$, where μ is the mean of the distribution
- We also have predictors x_i we are interested in studying
- We may link these predictors to any parameter of interest, but we typically do it for the **mean**

Statistical Modelling

Example



Statistical Modelling

Example

$$\begin{aligned} Y_i &\sim \mathsf{N}(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 x_i \end{aligned}$$

Statistical Modelling

Example

$$\begin{aligned} Y_i &\sim \mathsf{N}(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 x_i \end{aligned}$$

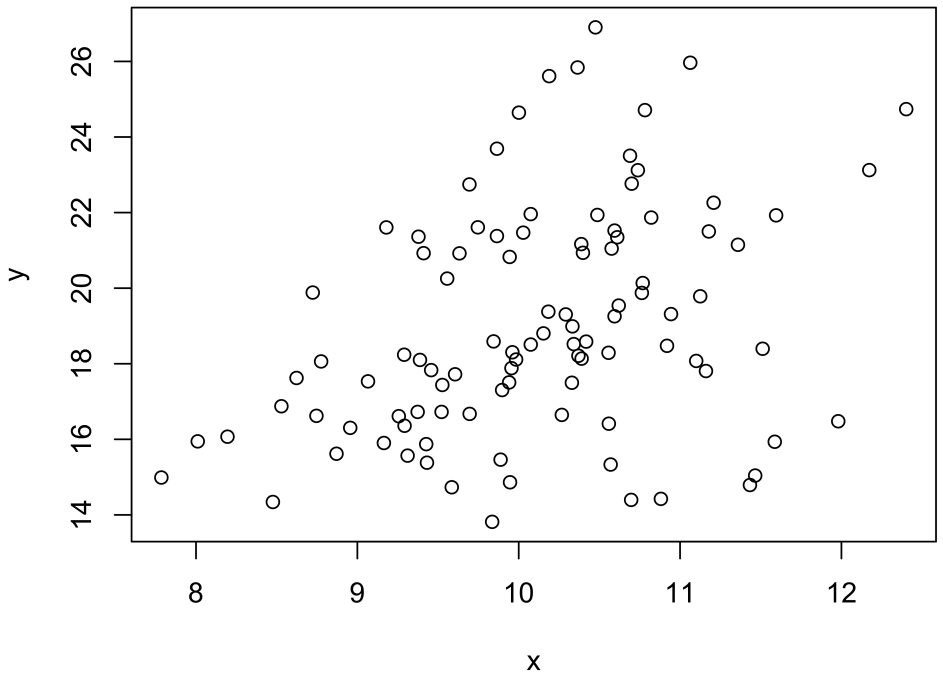
What about

- normality of residuals?
- homogeneity of variances?

This *only* makes sense for the model above!

Statistical Modelling

What if you had



Statistical Modelling

Maybe then you'd assume

$$\begin{aligned} Y_i &\sim \mathsf{N}(\mu_i, \sigma_i^2) \\ \mu_i &= \beta_0 + \beta_1 x_i \\ \log \sigma_i^2 &= \gamma_0 + \gamma_1 x_i \end{aligned}$$

No homogeneity of variances here!

Statistical Modelling

Y can be assumed to have *any* distribution

Statistical Modelling

Y can be assumed to have *any* distribution

Why is the normal distribution used to often then?

A Brief History of GLMs

- Multiple linear regression: a normal model with the identity link (Legendre, Gauss, Galton, 19th century)
- Analysis of variance (ANOVA): a normal model with the identity link (Fisher, 1918)
- Analysis of dilution assays: a binomial model with the complementary log-log link (Fisher, 1922)
- The exponential family class of distributions (Fisher, 1934)
- Probit analysis: a binomial distribution with the probit link (Bliss, 1935)
- Logistic regression: a binomial distribution with the logit link (Berkson, 1944; Dyke and Patterson, 1952)
- Item analysis: a Bernoulli distribution with the logit link (Rasch, 1960)
- Log-linear models: a Poisson distribution with the log link (Birch, 1963)
- Regression for survival data: an exponential distribution with the inverse or log links (Feigl and Zelen, 1965; Zippin and Armitage, 1966; Gasser, 1967)
- Inverse polynomials: a gamma distribution with the inverse link (Nelder, 1966)

A Brief History of GLMs

J. R. Statist. Soc. A,
(1972), **135**, *Part* 3, p. 370

Generalized Linear Models

By J. A. NELDER and R. W. M. WEDDERBURN

Rothamsted Experimental Station, Harpenden, Herts

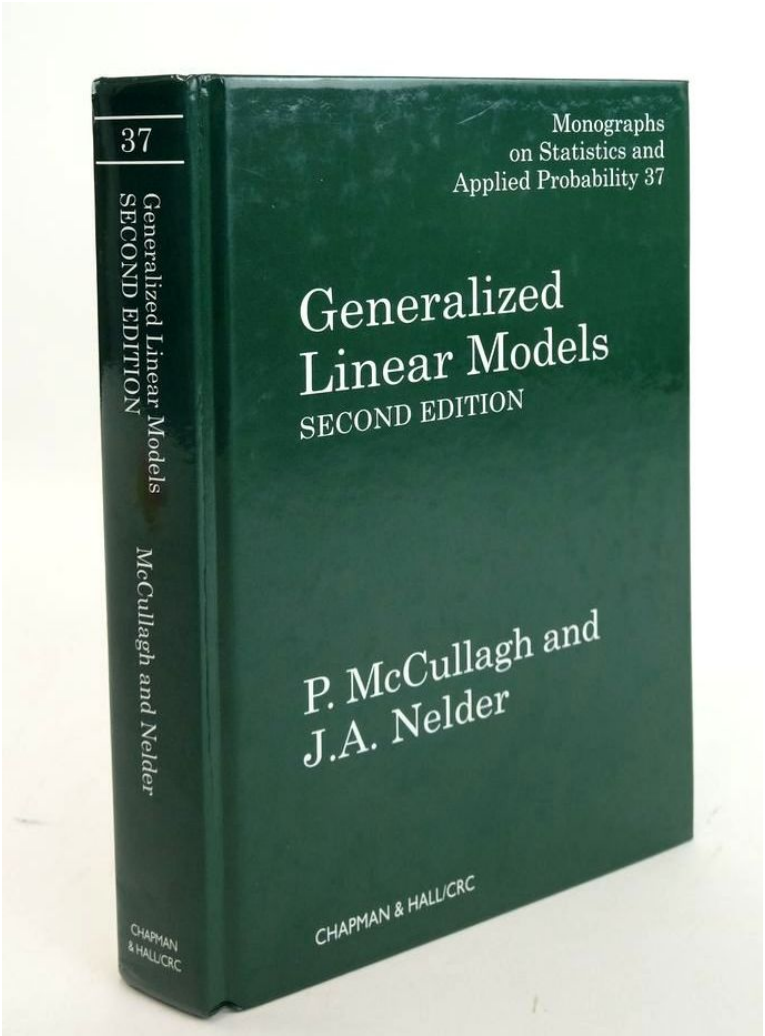
SUMMARY

The technique of iterative weighted linear regression can be used to obtain maximum likelihood estimates of the parameters with observations distributed according to some exponential family and systematic effects that can be made linear by a suitable transformation. A generalization of the analysis of variance is given for these models using log-likelihoods. These generalized linear models are illustrated by examples relating to four distributions; the Normal, Binomial (probit analysis, etc.), Poisson (contingency tables) and gamma (variance components).

The implications of the approach in designing statistics courses are discussed.

Keywords: ANALYSIS OF VARIANCE; CONTINGENCY TABLES; EXPONENTIAL FAMILIES; INVERSE POLYNOMIALS; LINEAR MODELS; MAXIMUM LIKELIHOOD; QUANTAL RESPONSE; REGRESSION; VARIANCE COMPONENTS; WEIGHTED LEAST SQUARES

A Brief History of GLMs



Models for Binary Data

Models for Binary Data