Generalized Linear Models

- Classical linear regression
  - $\Rightarrow$ complicated formulation of simple model, structural and random component of the model
- Generalized linear models
  - $\Rightarrow$ general description and examples
- Parameter estimation (extra matetrial)[1]
  - $\Rightarrow$ maximum likelihood method, computational issues
- Statistical inference (extra matetrial)[1]
  - $\Rightarrow$ goodness of fit, analysis of deviance

[1]Not obligatory

Wikipedia

- In statistics, the **generalized linear model** (**GLM**) is a useful generalization of ordinary least squares regression. It relates the random distribution of the measured variable of the experiment (the *distribution function*) to the systematic (non-random) portion of the experiment (the *linear predictor*) through a function called the *link function*.
- The subject of generalized linear models was formulated by John Nelder and Robert Wedderburn as a way of unifying various other statistical models under one framework, **allowing for one general method of efficiently performing maximum likelihood estimation for these models**.

Classical Linear Regression

Why easy formulation if complicated formulation exists?

Response variable Y has a normal distribution
Expected value EY of Y depends on explanatory variables

$$\begin{cases} (i) \quad Y_i \sim N(\mu_i, \sigma^2) & \textit{random component} \\ (ii) \quad \eta_i = x_i^T \beta & \textit{systematic component, linear predictor} \\ (iii) \quad \eta_i = g(\mu_i) = \mu_i & \textit{link function, linking (i) and (ii)} \end{cases}$$

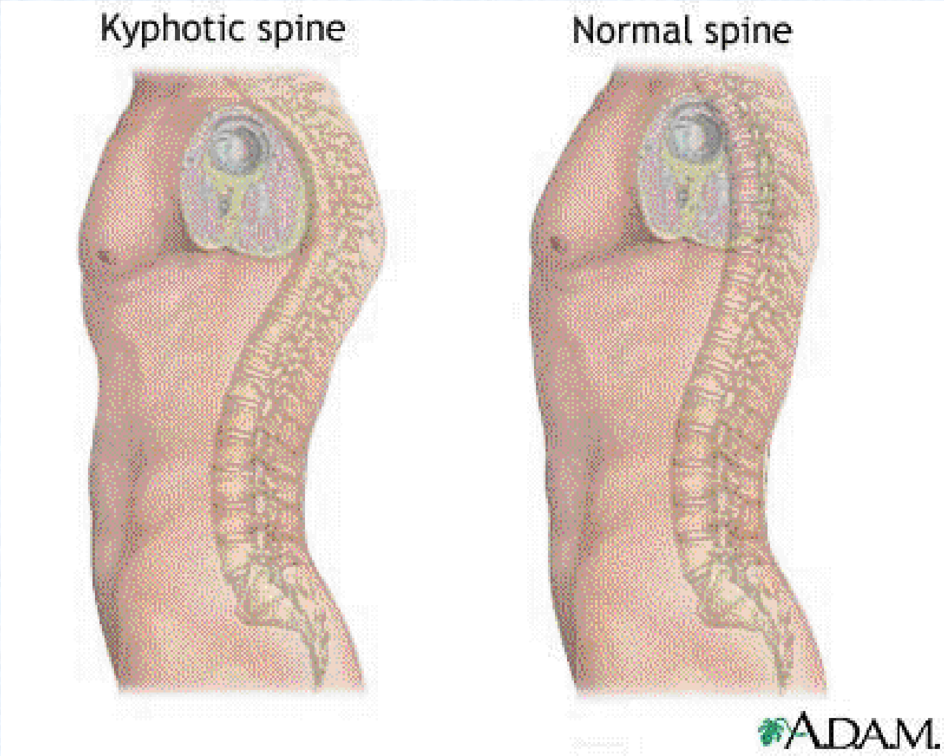
Taking more general distribution in (*i*) and a more general link function *g* in (*iii*) sticking to this linear form in (*ii*)

⇒ Generalized Linear Model

Generalized Linear Models

Kyphosis, medical context

Kyphosis is a deformation that can occur with children that underwent corrective spinal surgery.



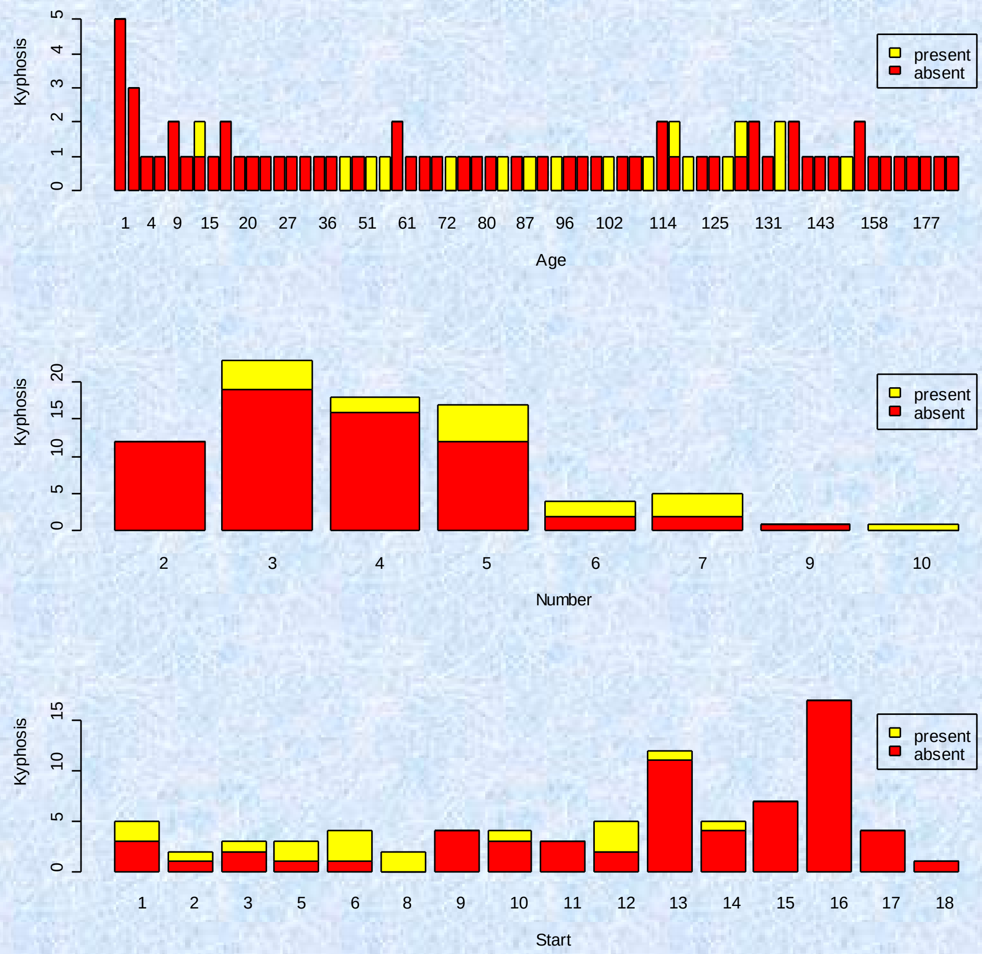
Question: given the age of the child at the time of surgery, what is the probability of occurrence of Kyphosis?

```
> library(rpart)
> data(kyphosis)
> kyphosis
   Kyphosis Age Number Start
1    absent  71      3     5
2    absent 158      3    14
...
```

Kyphosis

Data visualisation

```
> kyphosis
   Kyphosis Age Number Start
1   absent   71    3      5
2   absent  158    3     14
3   present 128    4      5
 :     :     :     :      :
83 absent   36    4     13
> par(mfrow=c(3,1))
> plot(Kyphosis~Age,
+ data=kyphosis)
> plot(Kyphosis~Number,
+ data=kyphosis)
> plot(Kyphosis~Start,
+ data=kyphosis)
```



Kyphosis

towards a model

Notation: Y_i indicator of presence of Kyphosis (1 means present, 0 not)
x_i age of child i at time of surgery

Initial naive model:

$$Y_i \sim \text{Bernoulli}(\mu_i), 1 \le i \le n, \text{ i.e., } Y_i = \begin{cases} 0 & \text{w.p.} \quad 1-\mu_i \\ 1 & \text{w.p.} \quad \mu_i \end{cases}$$

μ_i depends on x_i in some way

$Y_i \perp Y_j, i \neq j$

Kyphosis

Generalized linear model

Notation: Y_i indicator of presence of Kyphosis (1 means present, 0 not)

x_i age of child *i* at time of surgery

μ_i does not depend on x_i linearly, but a nonlinear function g (link function) of μ_i depends on x_i linearly

$$\begin{cases} Y_i \overset{\text{indep}}{\sim} \text{Bernoulli}(\mu_i), 1 \le i \le n, \text{i.e., } Y_i = \begin{cases} 0 & \text{w.p.} \quad 1-\mu_i \\ 1 & \text{w.p.} \quad \mu_i \end{cases} \\ \\ \eta_i = x_i^T \beta \\ \eta_i = g(\mu_i) = \log\big(\mu_i / (1-\mu_i)\big) \end{cases}$$

random component

systematic component, linear predictor

link function (logit)

$$\left(Y_i \overset{\text{indep}}{\sim} \text{Bern}\left(\frac{\exp(x_i^T \beta)}{1+\exp(x_i^T \beta)} \right), 1 \le i \le n \right)$$

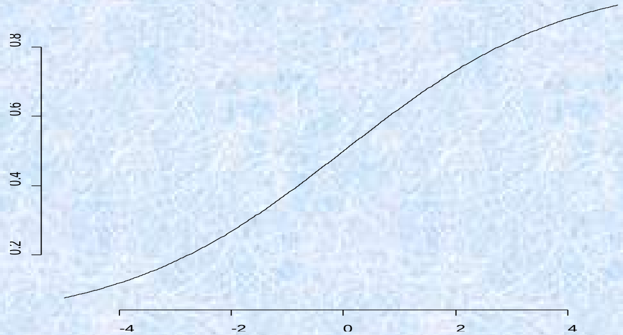
Logistic Regression model

Logistic Regression

$$\begin{cases} Y_i \overset{\text{indep}}{\sim} \text{Bernoulli}(\mu_i), 1 \le i \le n, \text{ i.e., } Y_i = \begin{cases} 0 & \text{w.p.} \quad 1 - \mu_\text{i} \\ 1 & \text{w.p.} \quad \mu_\text{i} \end{cases} \\ \qquad\qquad \textit{random component} \\ \eta_i = x_i^T \beta \quad \textit{systematic component, linear predictor} \\ \eta_i = g(\mu_i) = \log\left(\mu_i / (1 - \mu_i)\right) \quad \textit{link function} \end{cases}$$

$$Y_i \overset{\text{indep}}{\sim} \text{Bern}\left(\frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} \right), 1 \le i \le n$$

Bernoulli $\mu \uparrow$



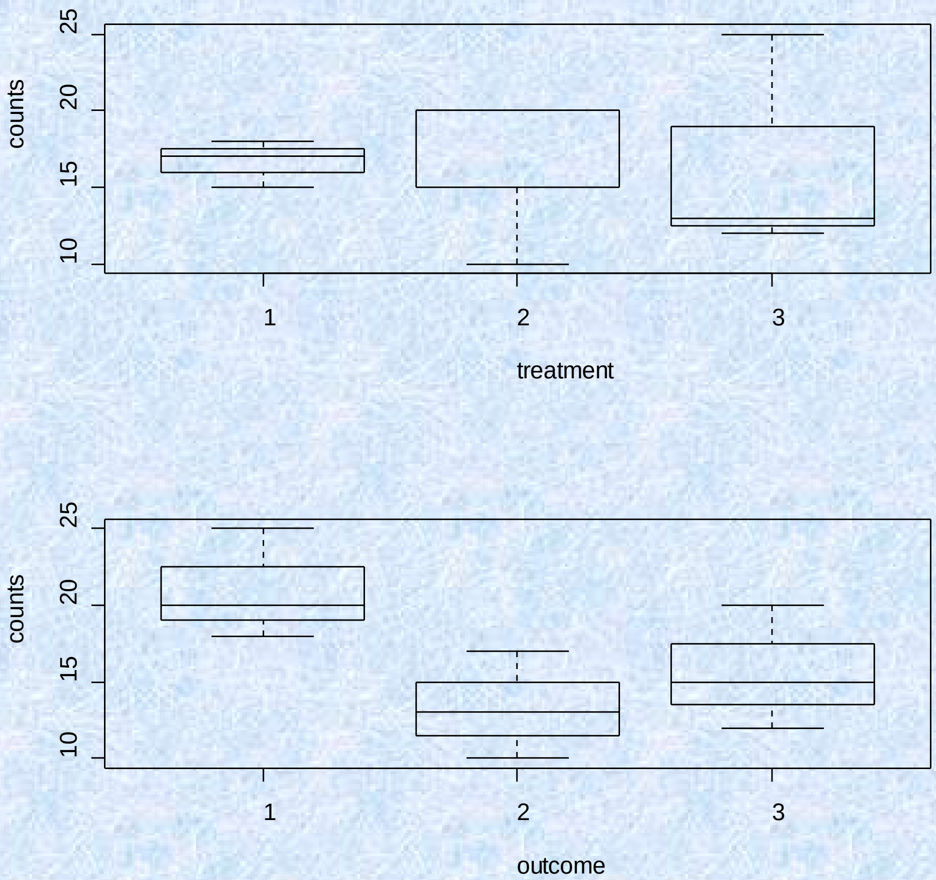
$x^T \beta \rightarrow$

Estimate parameter vector β based on the available data ⇒ estimated model specified

Generalized Linear Models

next example

```
> d.AD
  treatment outcome counts
1         1       1     18
2         1       2     17
3         1       3     15
4         2       1     20
5         2       2     10
6         2       3     20
7         3       1     25
8         3       2     13
9         3       3     12
> par(mfrow=c(2,1))
> plot(counts~treatment,d.AD)
> plot(counts~outcome,d.AD)
```



Data: randomized control trial.

Check first example for R function glm (type ?glm)

Counts

Generalized linear model

Notation: Y_i measurement on counts, assumed to be Poisson
x_i vector of explanatory variables for experiment i

μ_i does not depend on x_i linearly, but a nonlinear function g (link function) of μ_i depends on x_i linearly

$$\begin{cases} Y_i \overset{\text{indep}}{\sim} \text{Poisson}(\mu_i), 1 \le i \le n & \textit{random component} \\ \eta_i = x_i^T \beta & \textit{systematic component, linear predictor} \\ \eta_i = g(\mu_i) = \log \mu_i & \textit{link function} \end{cases}$$

$$\left(Y_i \overset{\text{indep}}{\sim} \text{Poisson}\left(\exp(x_i^T \beta)\right), 1 \le i \le n \right)$$

Log-linear Regression model

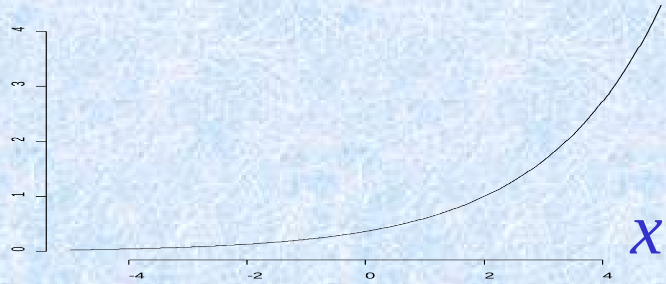
Log-linear Regression

$$\begin{cases} Y_i \overset{\text{indep}}{\sim} \text{Poisson}(\mu_i), 1 \le i \le n \\ \eta_i = x_i^T \beta \\ \eta_i = g(\mu_i) = \log \mu_i \end{cases}$$

Estimate parameter vector β based on the available data $\Rightarrow$ estimated model specified

$$Y_i \overset{\text{indep}}{\sim} \text{Poisson}\left(\exp(x_i^T \beta)\right), 1 \le i \le n$$

Poisson mean $\uparrow$



$x^T \beta \rightarrow$

How to estimate β in the logistic and log-linear regression model?

First: “general” generalized linear model

$\Rightarrow$ ML estimation in GLM’s

Generalized linear model

General structure and examples

$$\begin{cases} (i) \quad Y_i \sim f_i, \text{ a probability density function with } EY_i = \mu_i; \text{ see below} \\ (ii) \quad \eta_i = x_i^T \beta, \text{ linear predictor} \\ (iii) \quad \eta_i = g(\mu_i), \text{ with } g \text{ a general, monotonic link function} \end{cases}$$

Here f_i is the probability density of a one-dimensional exponential family distribution

θ_i natural parameter

$$f_i(y) = \exp\left(\frac{y\theta_i - b(\theta_i)}{\phi/A_i} + c(y, \phi/A_i) \right)$$

ϕ dispersion parameter

Scale parameter, in statistical problems *known*

Looks complicated, but…………..

Generalized linear model

examples of exponential family

$$f_i(y) = \exp\left(\frac{y\theta_i - b(\theta_i)}{\phi/A_i} + c(y, \phi/A_i) \right)$$

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

$$\frac{\mu^y}{y!}\exp(-\mu) = \exp\left(y\log\mu - \log y! - \mu\right)$$

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

Note: in exponential family the canonical link function is the function mapping the mean μ to the natural parameter. In other words: in a GLM with exponential family density *f* and canonical link, the natural parameter is modeled as linear function of the parameter vector β!!

Generalized linear model

Background material[1]

- Moments for exponential family
- Maximum Likelihood Estimation
- Newton-Rhapson
- Testing
- Confidence intervals

[1]Not obligatory