The Abdus Salam
International Centre
for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

2453-10

School on Modelling Tools and Capacity Building in Climate and Public Health

15 - 26 April 2013

Generalized Linear Models: An Introduction and General Overview

KAZEMBE Lawrence

University of Malawi, Chancellor College
Faculty of Science Department of Mathematical Sciences
18 Chirunga Road, 0000 Zomba
MALAWI

Generalized Linear Models: An Introduction and General Overview

Lawrence Kazembe
University of Namibia
Windhoek, Namibia

A presentation at
School on Modelling Tools and Capacity Building in
Climate and Public Health

ICTP, Trieste, Italy

Outline

-The Regression Problem

-GLM: motivating examples and its components

-GLM schema

-Models for continuous responses

-Models for binary and binomial responses

-Models for categorical data

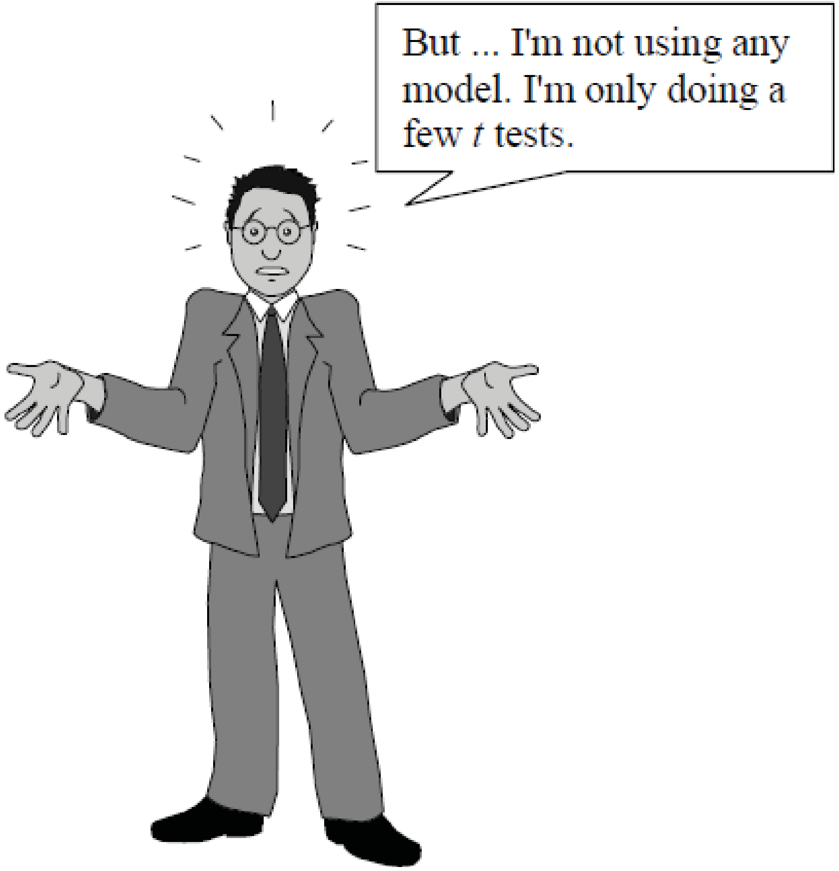
-Models for count data

-Models for duration data

-Inference: estimation and model checking

-Application of GLM.

The role of models-(Olsson, 2002)


But ... I'm not using any model. I'm only doing a few *t* tests.

Regression Problem

The classical linear model assumes a relationship between a set of variables (y_i, x_i) s.t.

$$y_i = \beta x_i + \varepsilon_i, i = 1, \ldots, n$$

where
-β is a regression coefficient
-$\varepsilon \sim N(0, \sigma^2)$
-The observations y_i are normally distributed

$$E(y|x) \sim N(\mu_i, \sigma^2)$$

-The mean μ_i is given as a linear combination with $\mu_i = \beta x_i$.

Regression Problem: Assumptions

-Distributional assumption:
The pairs (y_i, x_i) are assumed conditionally independent.
-Structural assumption:
The expectation μ_i is related to the linear predictor $\eta_i = \beta x_i$ by

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

or

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

where
-h is a known one-to-one response function
-g is a link, i.e., the inverse of h.
-and η is a linear predictor.

Classical Linear Model Fails: Motivating Examples

- Assumption of:
  (i) normally distributed errors fails;
  (ii) When one has multiplicative models, LM fail.

- Motivating examples:
  - Continuous data with nonconstant variance:
    : Medical expenses;
    : in Pharmacology (response to dosage concentration)
  - Binary data: Presence of disease (yes/no);

- Ordinal data: Severity of disease;

- Categorical data (nominal data): type of cause of death;

- Grouped data - Binomial data- (examined for disease, number diseased).

- Response as a proportion;

- Count data: Number of disease cases recorded- $y = 0, 1, 2, \ldots, \infty$;

- Response as a rate;

- (Zero-) Inflated data: -reported number of schistosomiasis eggs in urine sample.

- Skewed data: utilization of Ante-Natal care;

- Duration data: Survival from onset of treatment;

Generalized Linear Models (GLM)

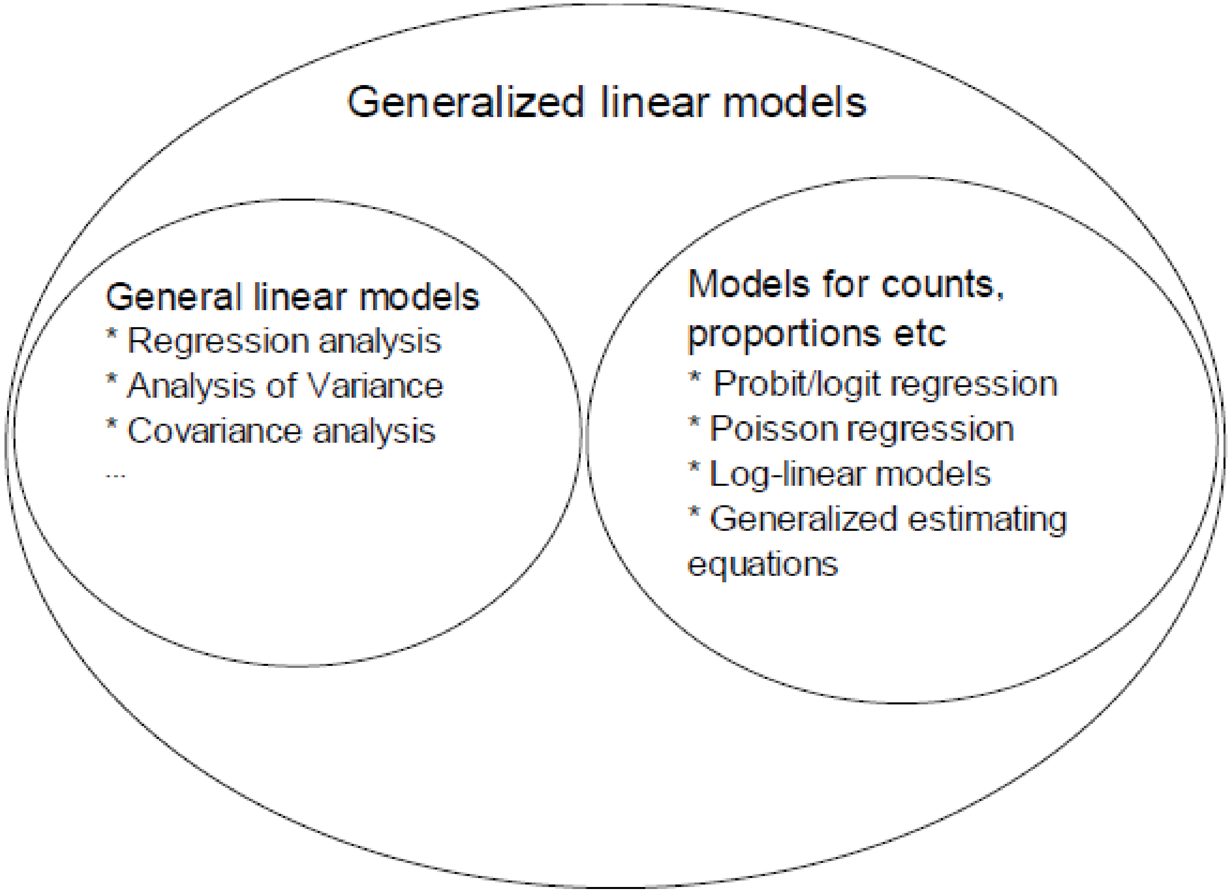
- Generalized linear models (GLMs) is a rich class of statistical methods, which generalizes the classical linear models in two directions, each of which takes care of one of the above mentioned problems:
  - GLMs work with a general class of distributions, which contains a number of discrete and continuous distributions as special cases, in particular the normal, Poisson and gamma distributions.
  - In GLMs some monotone transformation of the mean is a linear function of the x, with the linear and multiplicative models as special cases.

- GLM theory is quite recent-the basic ideas were introduced by Nelder and Wedderburn (1972).

- GLMs constitute a general statistical theory, which has well established techniques for estimating standard errors, constructing confidence intervals, testing, model selection and other statistical features.

- There is standard software for fitting GLMs that can easily be used for a tariff analysis, such as the SAS, GLIM, R or GenStat software packages.

GLM schema



Generalized linear models

General linear models
* Regression analysis
* Analysis of Variance
* Covariance analysis
...

Models for counts, proportions etc
* Probit/logit regression
* Poisson regression
* Log-linear models
* Generalized estimating equations

Three Components of a GLM

- The response or ”error” distribution:
  The $Y_i(i = 1, ..., n)$ are independent random variables with means, μ_i. They share the same distribution from the exponential dispersion family, with a constant scale parameter:
  - $f(y_i|\theta_i, \phi, \omega_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi}\omega_i + c(y_i, \phi, \omega_i)\right\}$, where
  - θ_i is the natural parameter
  - ϕ is a scale or dispersion parameter

- $b(\cdot)$ and $c(\cdot)$ are specific functions corresponding to the type of exponential family, and

- ω_i is a weight with $\omega_i = 1$ for ungrouped data $(i = 1, \ldots, n)$ and $\omega_i = n_i$ for grouped data $(i = 1, \ldots, g)$ or if an average or sum is considered then $\omega_i = 1/n_i$.

- Remark:
  a) If ϕ is fixed then we have a one-parameter exponential family.
  b) If ϕ is unknown, then we have an exponential dispersion model.

- The linear predictor or systematic component: $\eta_i = \sum_1^p \beta_i x_i = \boldsymbol{X\beta}$, relating the response to the covariates.

- The link function: $\eta = g(\mu_i) = g(\boldsymbol{X\beta})$.

- The choice of appropriate link function g depends on the type of response.
  Example of natural links:

  - $\eta = \mu$ for the normal,
  - $\eta = \log \mu$ for the Poisson,
  - $\eta = \log[\frac{\mu}{1-\mu}]$ for the binomial

- Concerning the design vector:
  Nothing new comes here compared to the classical model.

- Categorical covariates, ordered or unordered, have to be coded by a dummy vector.

- Metrical covariates can be incorporated directly or after appropriate transformation like $\log x, x^2, \ldots$ etc.

Standard models: Components of exponential family

	Distribution	$\theta(\mu)$	$b(\theta)$	ϕ
Normal	$N(\mu, \sigma^2)$	μ	$\theta^2/2$	σ^2
Bernoulli	$Binom(1, \pi)$	$\log(\pi/(1-\pi))$	$\log(1+\exp(\theta))$	1
Binomial	$Binom(n, \pi)$	$\log(\pi/(n-\pi))$	$\log(1+\exp(\theta))$	n
Poisson	$P(\lambda)$	$\log \lambda$	$\exp(\theta)$	1
Gamma	$G(\mu, \nu)$	$-1/\mu$	$-\log(-\theta)$	ν^{-1}
Inverse Gaussian	$IG(\mu, \sigma^2)$	$1/\mu^2$	$-(-2\theta)^{1/2}$	σ^2

Standard models: Other links

Other link functions are:

- square root $\sqrt{\mu}$

- exponent $(\mu + c_1)^{c_2}$ (c_1 and c_2) are known.

- complimentary log-log $\log[-\log[\frac{\mu}{n}]]$

- probit $\Phi^{-1}(\frac{\mu}{n})$

Standard models: Expectations and Variance

Distribution	$E(y) = b'(\theta)$	$var(y) = b''(\theta)$	$var(y) = b''(\theta)\phi/\omega$
Normal	$\mu = \theta$	1	σ^2/ω
Bernoulli	$\pi = \frac{\exp(\theta)}{1+\exp(\theta)}$	$\pi(1-\pi)$	$\pi(1-\pi)/\omega$
Poisson	$\lambda = \exp(\theta)$	λ	λ/ω
Gamma	$\mu = -\theta$	μ^2	$\mu^2\nu^{-1}/\omega$
Inverse Gaussian	$\mu = -(-2\theta)^{-1/2}$	μ^3	$\mu^3\sigma^2/\omega$

Models for Continuous Responses

- Normal distribution:
  - Assuming $\mu = \eta = \beta x$ leads to the classical linear normal model.
  - Sometimes a non-linear relationship $\mu = h(\eta)$, e.g.

    $$h(\eta) = \eta^2, \quad h(\eta) = \log \eta, \quad h(\eta) = \exp \eta$$

    will be more appropriate and can easily be handled within GLM framework.
- Gamma distribution:

- $f(y|\mu,\nu) = \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1} \exp\left(-\frac{\nu}{\mu}y\right), y \geq 0$
- The natural response function is the reciprocal $\mu = \eta^{-1}$ s.t. $\eta = \beta x$.
- The other important response functions are the identity

$$h(\eta) = \eta = \mu,$$

and the exponential response function

$$h(\eta) = \exp(\eta) = \mu,$$

or, equivalently, the log-link

$$g(\mu) = \log(\mu) = \mu.$$

• Inverse Gaussian distribution:

- Right skewed-distribution.

- This distribution can be applied for nonsymmetric regression analysis and for lifetimes.

- Examples include: length of hospital stay, Medical costs, Clotting time of blood, Amount of time plasmodium infection remains in the blood, Claim amount in medical insurance.

- Many continuous, right-skewed distributions are applicable:
  log-gamma, Weibull, Burr, Pareto, generalized Pareto, Makeham and Gompertz, all as candidates for length of hospital stay distributions (Hogg and Klugman, 1984).

Models for Binary and binomial response

- Suppose

$$y = \begin{cases} 1 & \text{if disease present} \\ 0 & \text{otherwise} \end{cases}$$

- $y \sim Bern(\pi)$, where π is the probability or proportion with the event of interest. s.t.

$$P(Y = y) = \pi^y (1 - \pi)^y, \text{ for every } y = 0, 1.$$

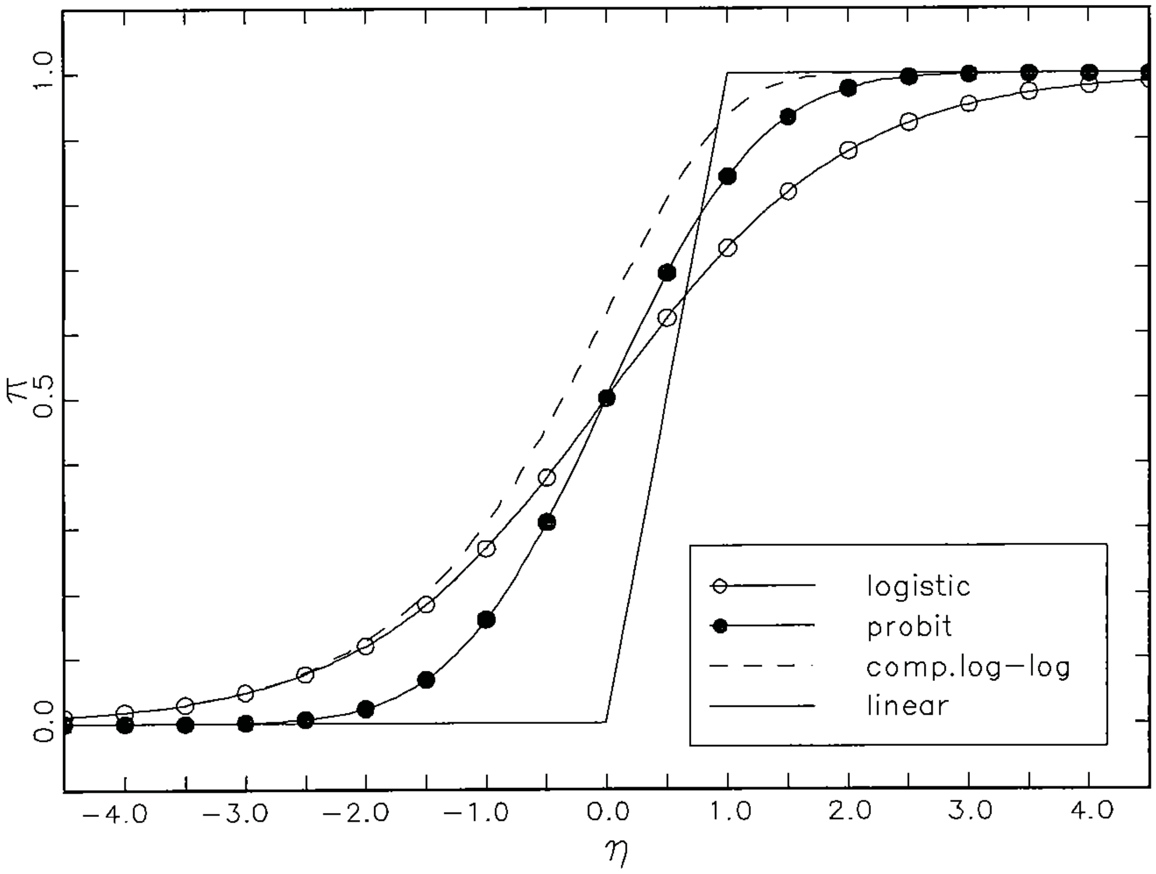
- If binary data are grouped, then $y \sim Bin(n, \pi)$, s.t.

$$\binom{n}{y} \pi^y (1 - \pi)^{n-y}$$

- Four possible links:

Response function	Link	Mean	Variance
Linear	$\pi = \eta = \beta x$	0.5	1/12
Logit	$\log[\frac{\pi}{1-\pi}] = \eta$ s.t. $\pi = \frac{\exp(\theta)}{1+\exp(\theta)}$	0	$\pi^2/3$
Probit	$\pi = \Phi(\eta) = \Phi(\beta x)$	0	1
Compl.log-log	$h(\eta) = 1 - \exp(-\exp(\eta))$	-0.5772	$\pi^2/6$

Response functions for binary responses



Parameter Interpretation

- For the logistic model we have a linear model for the ”log odds” s.t.

  $$\exp(\beta)$$

  is called ODDS RATIO.

- For other link transformations this is not straight forward.

- General rule:
  -Interpret covariate effects as in the linear models.
  -Back-transform the linear effect on η into a nonlinear effect on π, i.e. use response function $\pi = h(\eta)$.

Models for Categorical data

- Suppose Anemia presence is classified as:

$$y = \begin{cases} 1 & \text{not present} \\ 2 & \text{mild anemia} \\ 3 & \text{severe anemia} \end{cases}$$

- Or Cause of death

$$y = \begin{cases} 1 & \text{HIV} \\ 2 & \text{TB} \\ 3 & \text{DIABETES} \end{cases}$$

-In the first example the response y is ordered;
-In the second example, the response y is nominal.

Models for Nominal Responses

- First nominal (unordered) response:
  This is an extension of the dichotomous response variable $y_i \in \{0, 1\}$, where each category versus some reference category is fitted as category-specific logistic or probit model.

- Multicategorical logit model is given as

$$P(Y = r) = \frac{\exp(\beta_{0r} + \beta_r x_i)}{1 + \sum_{s=1}^{q} \exp(\beta_{0s} + \beta_s x_i)}$$

  or equivalently written as

$$\log \frac{P(Y = r)}{P(Y = k)} = \beta_{0r} + \beta_r x_i$$

which is the log odds for category r with respect to the reference category k.

- From the above one immediately gets the response function $h = (h_1, \ldots, h_q)$ with

$$h_r(\eta_1, \ldots, \eta_q) = \frac{\exp(\eta_r)}{1 + \sum_{s=1}^{q} \exp(\eta_s)}, r = 1, \ldots, q.$$

Models for Ordinal Responses

- Ordinal responses stem from a different mechanism, and can at times be seen as a mere categorization of continuous variable, which can be observable (manifest) or unobservable (latent) variables.

- It is stipulated that Y is a categorized version of the latent variable $U = \eta + \varepsilon$ obtained through the threshold mechanism

$$Y = r \Leftrightarrow \theta_{r-1} < U < \theta_r, r = 1, \ldots, k,$$

with thresholds $-\infty = \theta_0 < \theta_1 < \ldots < \theta_k = \infty$.

- Assuming the error variable ε has the distribution function F, then Y obeys a cumulative model

  $$P(Y \leq r) = F(\theta_r - \eta)$$

  where η is the predictor.

- If $F(x) = 1/(1 + \exp(-x))$ is a logistic distribution we have a cumulative logit model:

  $$P(Y \leq r|x) = \frac{\exp(\theta_r + \gamma x)}{1 + \exp(\theta_r + \gamma x)}$$

- If $F(x) = \Phi(x)$ is a cumulative normal distribution, one has a cumulative probit:

  $$P(Y \leq r|x) = \Phi(\theta_r - \eta),$$

- Other specifications of ordinal response:
  - extreme-minimal-value distribution:
    $F(x) = 1 - \exp(-\exp(x))$.
  - extreme-maximal-value distribution:
    $F(x) = \exp(-\exp(-x))$.
  - log-log links:
    $\log[-\log P(Y \leq r|x)] = -(\theta_r + x'\gamma)$.
  - sequential regression models.
  - stereotype regression models.
  - adjacent categories regression models

Count data (1)

- Count data- Recorded over time or space or both:
  -e.g. No. of deaths, hospitalized cases, accidents
  -Such data are recorded as: $0, 1, 2, \ldots, \infty$.

- Generally the Poisson distribution or some modification should be the first choice.

- Log-linear Poisson model:

$$\log(\mu) = \eta = \beta x, \quad \mu = \exp(\eta)$$

- If all covariates are categorical, this leads to modelling of frequencies in contingency tables.

Count data (2)

Married	Male	Female
Yes	y_{11}	y_{12}
No	y_{21}	y_{22}

Translates to

i	Married	Gender	Observation
1	Yes	M	y_1
2	Yes	F	y_2
3	No	M	y_3
4	No	F	y_4

- Therefore we model

$$\log(y_i) = \beta_0 + \beta_1 * \text{Gender} + \beta_2 * \text{Married}.$$

- If y is exactly Poisson-distributed, its variance equals its expectation:

$$var(y|x) = \mu$$

- Overdispersion: The case where the variance is larger than expected. This is called extra-Poisson variation.

- For count data we denote by

$$var(y|x) = \sigma(\mu) = \phi\mu$$

- More complex models that account for extra-variation in the data are available:

- Quasi-Poisson model
- Negative binomial model
- zero-inflated models
- hurdle models
- finite mixture models

Duration data analysis

- Analysis of survival time, lifetime or failure data has received considerable attention.

- Challenges of duration data:

  - censoring,

  - time-varying covariates,

  - multiple failures

- Censoring means that the survival time is not known for all individuals when the study is finished.

-For right censored observations we only know that the survival time is at least the time at which censoring occurred.
-Left censoring, i.e. observations for which we do not know e.g. the duration of disease when the study started, is also possible.

- Denote the density function for the survival time with $f(t)$,
  Let the corresponding distribution function be $F(t) = \int_{-\infty}^{t} f(s)ds$.
  The survival function is defined as

$$S(t) = 1 - F(t),$$

and the hazard function is defined as

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d \log(S(t))}{dt}.$$

The cumulating hazard function is $H(t) = \int_{-\infty}^{t} h(s)ds$.

- Modelling of survival data includes choosing a suitable distribution for the survival times or, which is equivalent, choosing a hazard function.

  (1). In nonparametric modelling, the survival function is not specified, but is estimated nonparametrically through the observed survival distribution. This is the basis for the so called Kaplan–Meier estimates of the survival function.

(2). In parametric models, the distribution of survival times is assumed to have some specified parametric form. The exponential distribution, Weibull distribution or extreme value distribution are often used to model survival times.
(3). A semiparametric approach is to leave the distribution unspecified but to assume that the hazard function changes in steps which occur at the observed events.

- Most commonly used survival models: Cox Regression model (or Proportional hazard models- 1972).

Likelihood Inference

Estimation:

- Ordinary least squares/Iterative Weighted Least Squares
- Maximum likelihood/Restricted maximum likelihood
- Bayesian methods

Maximum Likelihood Estimation

-For the Maximum likelihood in GLM, the estimates are those parameter values that maximize the log likelihood,

$$\ell = \log[L(\theta, \phi; y)] = \frac{y\theta - b(\theta)}{a(\theta)} + c(y, \phi)$$

-The parameters of regression coefficient β are function of θ.

-Differentiation of ℓ with respect to the elements of β using the chain rule yields,

$$\frac{\partial \ell}{\partial \beta_j} = \frac{\partial \ell}{\partial \theta}\frac{d\theta}{d\mu}\frac{d\mu}{d\eta}\frac{\partial \eta}{\partial \beta_j}$$

-We have shown earlier that $b'(\theta) = \mu$ and that $b''(\theta) = V$, the variance function.

-Thus $\frac{d\mu}{d\theta} = V$,

and $\eta = \sum \beta_j x_j$ we obtain

$$\frac{\partial \eta}{\partial \beta_j} = x_j$$

-Putting together:

$$\frac{\partial \ell}{\partial \beta_j} = \frac{(y-\mu)}{a(\phi)} \frac{1}{V} \frac{d\mu}{d\eta} x_j.$$

-The likelihood equation for the one parameter β_j is given by setting the above equation equal to zero, i.e,

$$\sum_i \frac{W_i(y_i - \mu_i)}{a(\phi)} \frac{d\eta_i}{d\mu_i} x_{ij} = 0,$$

where $W^{-1} = \left(\frac{\partial \eta}{\partial \mu}\right)^2 V$.

Model checking: Residuals and goodness-of-fit statistics

- Deviance:
  -GLM models can be assessed through the *deviance*
  -*Null* model- one with one parameter only (mainly the mean of all observations).
  -*Saturated (full)* model has *n* parameters- one for each observation.
  -Full model used as a benchmark for assessing the fit of any model to the data.
  -This is done by calculating the deviance as

$$D = 2[\ell(y, \phi; y) - \ell(\widehat{\mu}, \phi; y)]$$

where $\ell(y, \phi; y)$ is for the log-likelihood for the full model and $\ell(\widehat{\mu}, \phi; y)$ is for the current model.

-D is distributed as χ^2 as n increases.
-For *competing* models M_1 and M_2 use $D(M_2) - D(M_1)$ with the difference compared to χ^2 with $df_2 - df_1$ degrees of freedom.

- Residual analysis:
  -An alternative to the deviance is to use deviance residuals e.g. Pearson's χ^2 residuals

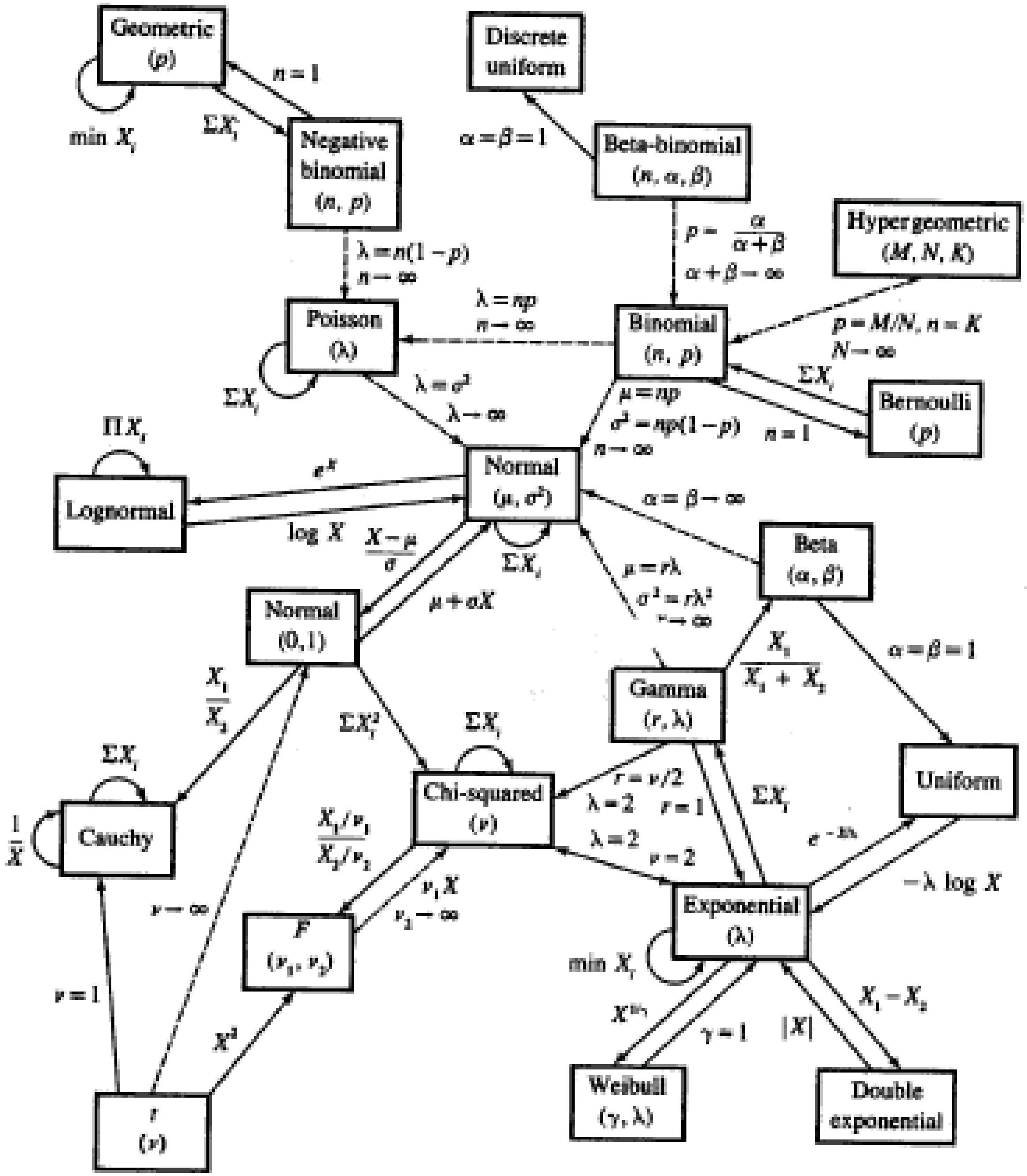
$$\chi^2 = \sum \frac{(y - \widehat{\mu})^2}{V(\widehat{\mu})}$$

- Or use measures of fit:
  -Akaike Information Criterion (AIC),

-Bayesian Information Criterion (BIC),
-Deviance Information Criterion (DIC).

- Testing for significance of variables:
  -Likelihood ratio test
  -Wald's test
  -Score test.

Relationship between common distributions (Leemis, 1986)



Geometric (p)
min X_i
$n=1$
ΣX_i
Negative binomial (n, p)
Discrete uniform
$\alpha=\beta=1$
Beta-binomial (n, α, β)
$p=\frac{\alpha}{\alpha+\beta}$
$\alpha+\beta\to\infty$
Hypergeometric (M, N, K)
$p=M/N,\ n=K$
$N\to\infty$
$\lambda=n(1-p)$
$n\to\infty$
Poisson (λ)
$\lambda=np$
$n\to\infty$
Binomial (n, p)
ΣX_i
Bernoulli (p)
$n=1$
ΣX_i
$\lambda=\sigma^2$
$\lambda\to\infty$
$\mu=np$
$\sigma^2=np(1-p)$
$n\to\infty$
ΠX_i
Lognormal
e^X
log X
Normal (μ, σ²)
$\alpha=\beta\to\infty$
Beta (α, β)
$\frac{X-\mu}{\sigma}$
ΣX_i
$\mu+\sigma X$
$\mu=r\lambda$
$\sigma^2=r\lambda^2$
$r\to\infty$
Normal (0,1)
$\frac{X_1}{X_2}$
ΣX_i^2
ΣX_i
Gamma (r, λ)
$\frac{X_1}{X_1+X_2}$
$\alpha=\beta=1$
ΣX_i
Cauchy
$\frac{1}{X}$
Chi-squared (ν)
$r=\nu/2$
$\lambda=2$
$r=1$
$\lambda=2$
$\nu=2$
ΣX_i
Uniform
$e^{-\lambda X}$
$-\lambda\log X$
$\frac{X_1/\nu_1}{X_2/\nu_2}$
$\nu_1 X$
$\nu_2\to\infty$
$\nu\to\infty$
F (ν₁, ν₂)
Exponential (λ)
min X_i
$\nu=1$
X^2
$X^{1/\gamma}$
$\gamma=1$
$|X|$
X_1-X_2
t (ν)
Weibull (γ, λ)
Double exponential

References

- A. Agresti: Categorical Data Analysis.
- A. DeMarias: Regression with Social Data-Modelling continuous and limited response variables
- A. Dobson: An introduction to generalized linear models
- L. Fahrmeir and G. Tutz: Multivariate statistical modelling using generalized linear models
- P. McCollough and J.A. Nelder: Generalized linear models.