GENERALIZED LINEAR MODELS

Γιώτα Τουλούμη

Καθηγήτρια Βιοστατιστικής και Επιδημιολογίας

Εργ. Υγιεινής, Επιδημιολογίας και Ιατρικής Στατιστικής

Ιατρική Σχολή Πανεπιστημίου Αθήνας

gtouloum@med.uoa.gr

Objectives

- The overall objective of this course is to impact an understanding of Generalized Linear Models (GLM) and their use in practice that allows their application in a wide range of medical settings.

Areas to be covered

- Definition of GLM and the use of maximum likelihood (ML) based inference in the context of GLM
- The main class of GLM and their relevance in medical and epidemiological questions
- The interpretation of parameter's from GLM's
- The use of Stata to model data with GLM
- Comparison and assessment of fit of GLM's

References

- Agresti A (1996). An introduction to categorical data analysis. Wiley.
- McGullagh P, Nelder JA (1989). Generalized linear models, second edition. Chapman & Hall.
- Dobson AJ (1990). An introduction to generalized Linear Models. Chapman & Hall.
- Hosmer DW, Lemeshow S (1989). Applied logistic regression. John Wiley & Sons.
- Clayton D, Hills M (1993). Statistical models in epidemiology. Oxford.

Introduction

Classical statistical inference is the process by which sample data are used to infer the properties of populations

- **The modeling process involves the following steps:**
  - The model
  - Parameters
  - Parameter estimators
  - Hypotheses

The model

- A statistical model is a representation of the population under study. The model also usually reflects the way in which data have been sampled from the population.

- *As an illustration, let the population be some set of individuals, a proportion π of whom correspond to success (i.e. treatment failure, alive etc). Given a random sample of n individuals the number of successes k has a bin(n, π) distribution.*

Parameters

- Parameters in the model correspond to features of the population. In practice the true values of the parameters are unknown and so the sample is to be used to draw inferences about them.

- *The proportion π is the parameter of interest in the previous example (and it is the only parameter in this case).*

Parameter estimators

- There are sample statististics that are used to provide estimates of the unknown parameters. Recall that **estimator** refers to the general form of the statistic (a random variable), while **estimate** refers to the actual numerical realization from a given sample.

- *An obvious estimator of π is* $\hat{\pi} = k / n$

Hypotheses

- There are statements about the unknown parameters, e.g. a parameter takes a particular values π = 1/2 or two parameters are equal

- *There is only one parameter, π, so we are really restricted to hypotheses such as: π = 1/2.*

Model

- Suppose we have a number of measurements or counts, together with some associated structural or contextual information, such as the order in which data were collected, which measuring instruments were used, and other differences in conditions under which individual measurements were made.
- To interpret such data, we search for a *pattern*, i.e. that one measuring instrument has produced consistently higher readings than another.

Model (cont)

- Such systematic effects are likely to be blurred by other variation of a more haphazard nature. The latter variation is described in statistical terms, no attempt being made to model or predict the actual haphazard contribution to each observation.
- **Statistical models contain both elements, which we will call systematic effects and random effects.**

Model (cont)

- The value of the model is that, it often suggests a simple summary of the data in terms of the major systematic effects together with a summary of the nature and magnitude of the unexplained or random variation.
- Thus, the problem of looking intelligently at data demands the formulation of patterns that are thought capable of describing successfully not only the systematic variation in the data under study, but also describing patterns in similar data that might be collected by another investigator at another time and in another place.

Example: regression models

$$Y_i = \beta_0 + \sum \beta_j x_{ij} + e_i$$

- What sort of variable Y is (continuous, discrete, qualitative, …)?
- What is the distribution of Y, what is the range of possible values?
- Does the sample of observed Y's have to fit the distribution exactly? What else is important about the Y's distribution?
- What sort of variables are the x's? (continuous, discrete, qualitative)
- How do we assess how well this model fits a set of data?
- How do we assess how well individual cases conform to the fitted model?
- How do we assess the effect of individual cases on the fitted model?

What if response variable is not continuous

- Let's say that Y_i's are binomial. The response variable will be Y_i = 0 or 1 (**ungrouped data**). For example, case control study where Y_i=0 for control and Y_i=1 for case.
- We can also have **grouped data**. For example a group of n_i cases all have the same values of x_{ij}'s. In that case Y_i=# cases; Y_i=0,1, ..., n_i. Proportion =Y_i/n_i.

Can we use ordinary regression with binomial Y_i?

- Ungrouped data: **NO**
- Grouped data: **In some cases, YES**
- Potential problems:
  - Model is actual $Y_i / n_i = \beta_0 + \sum \beta_j x_{ij} + e_i$
  - Y_i/n_i bounded 0…1
  - Variance of Y_i/n_i not constant (but may be approximately proportional to $1/n_i$)
  - We can use **binomial distribution**

Generalized Linear Models

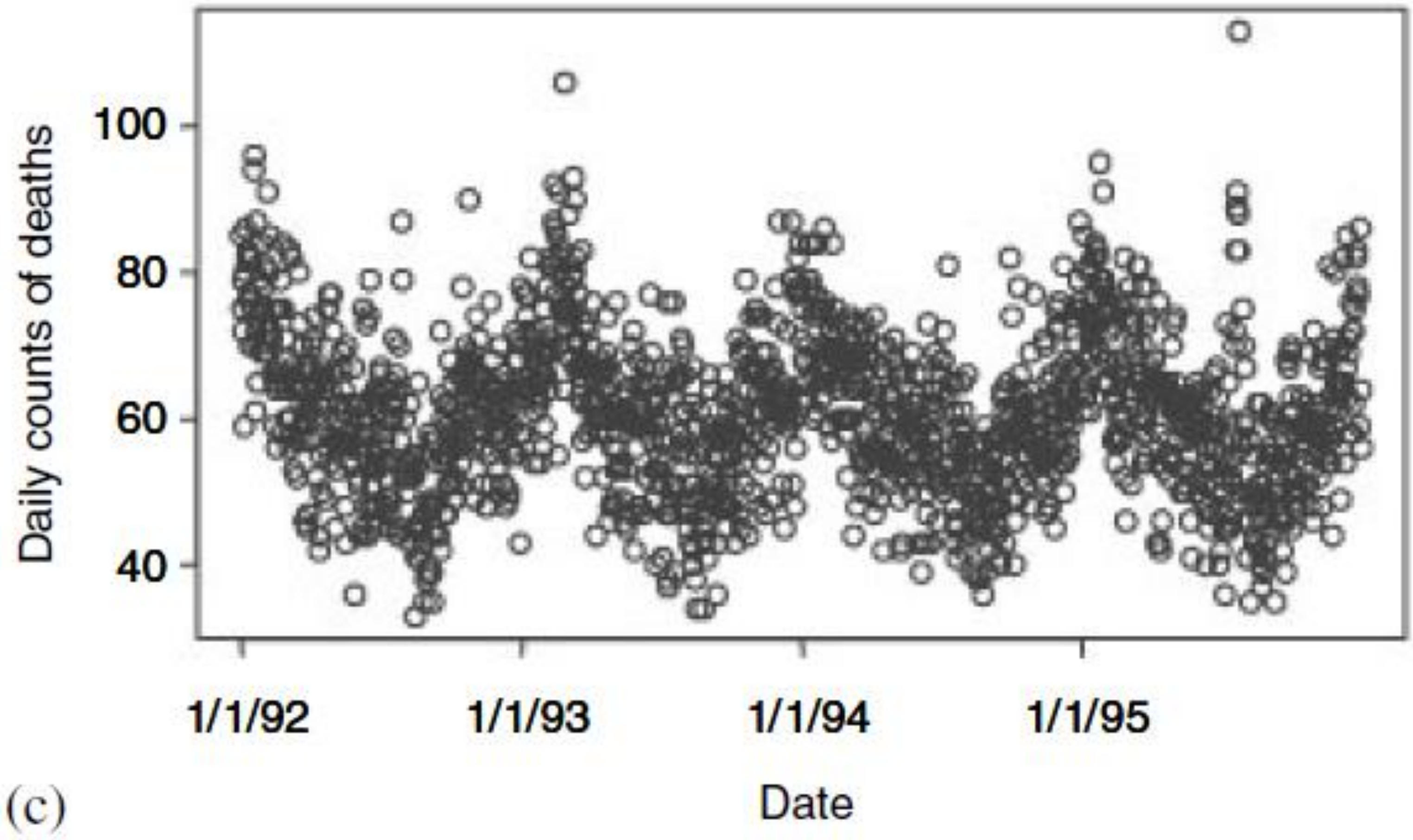
- **GLM** include several known models as a special case: ordinary linear models, ANOVA, logit and probit models for binary data, log-linear models for counts, multinomial response models and some models for survival data.

- **GLM** share a number of properties, such as linearity, and there is a common method for computing parameter estimates.

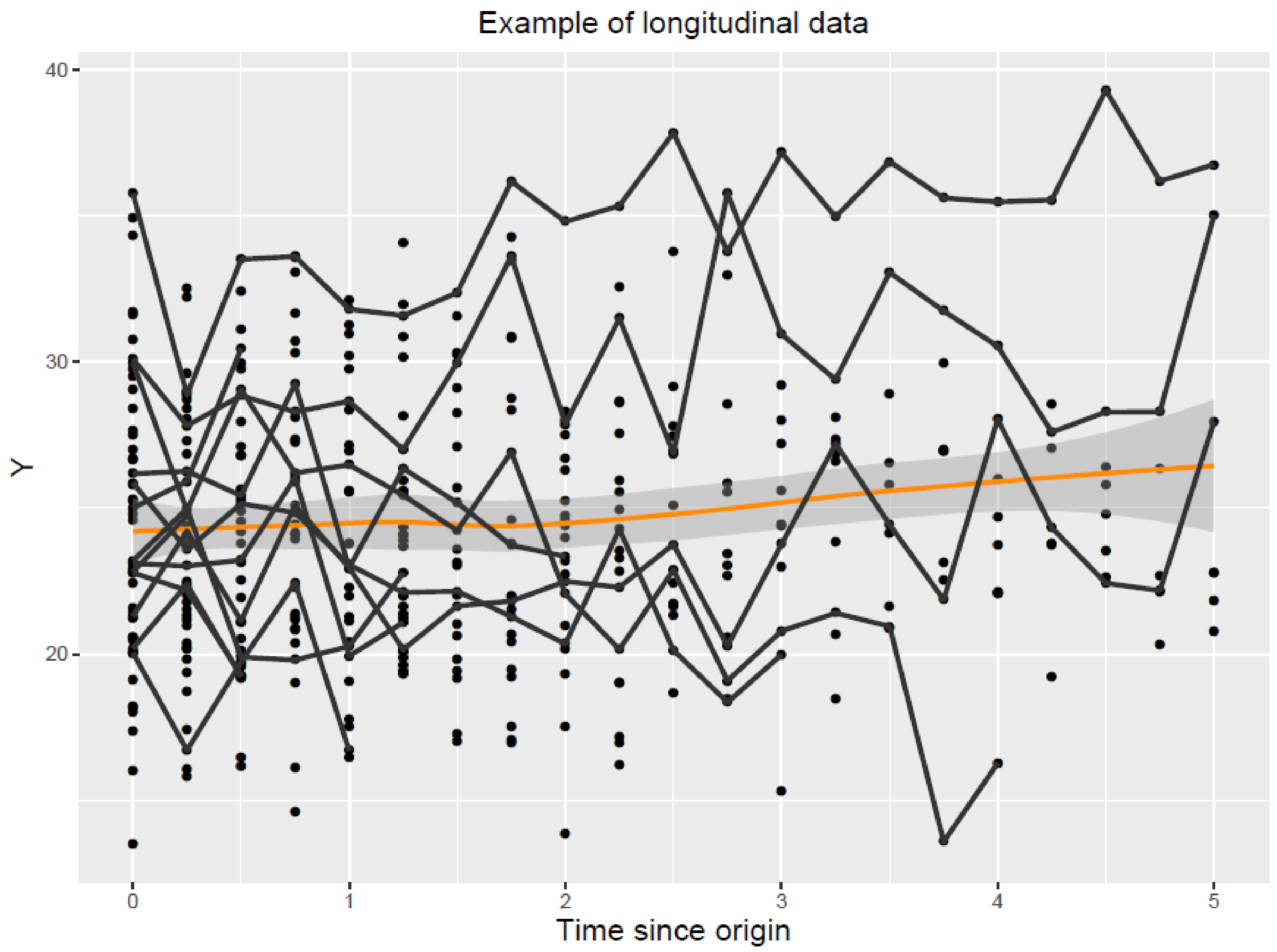
Basic GLM's assumptions

- **Independent observations.**
  - More generally, the observations are independent in blocks of fixed known sizes. *Consequence:* Data exhibiting the autocorrelations of time series are *excluded.*
- **Single Error Term**
  - There is a single error term in the model. This constraint excludes, for example models for the analysis of experiments having more than one error term (split-plot design i.e. between and within-plot variance).

Raw mortality in Madrid: Autocorrelation



Longitudinal data



Relaxing assumptions

- In practice both assumptions can be relaxed for certain kinds of GLM's.
    - For instance autoregression models can be easily fitted using programs designed for ordinary linear models.
    - Through a grouping factor corresponding to a nuisance classification that may induce correlations within groups; a within-groups analysis after elimination of the effects of that nuisance factor can proceed as if the observations were independent.

Estimation

- Having selected a model, it is required to estimate the parameters and to assess the precision of the estimates.
- In the case of GLM estimation proceeds by defining a measure of goodness of fit between the observed data and the fitted values generated by the model.
- The parameter estimates are the values that minimize the goodness of fit criterion.
- In ordinary regression parameters are estimated using the **Least Squares method**. That is by minimizing the sum of the squares of the residuals: $\mathbf{e}^T\mathbf{e} = \sum_i e_i^2 = \sum_i Y_i - (\mathbf{X}_\mathbf{i}\boldsymbol{\beta})$
- The general method used in GLM is that of **maximum likelihood**.

Revision of likelihood based inference

- If $Y_1, \ldots, Y_n$ are independent random variables each with probability density function (pdf) $f_i(y_i;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a (possibly vector-valued) parameter, then, by virtue of independence, the joint pdf of the vector $\mathbf{Y}$ is:

$$\mathrm{L}(\boldsymbol{\theta};\mathbf{y}) = \prod_{i=1}^{n} f_i\left(y_i;\boldsymbol{\theta}\right)$$

- This function, is called the *likelihood.*
- Note that the pdf function $f_i(y_i;\boldsymbol{\theta})$ is considered as a function of y for fixed θ, whereas the likelihood is considered as a function of θ for the particular data set observed.

The log-likelihood function

- Usually we work with the logarithm of the likelihood function and under the assumption of independence of the observations we have

$$l(\boldsymbol{\theta};\mathbf{y}) = \log \mathrm{L}(\boldsymbol{\theta};\mathbf{y}) = \sum_{i=1}^{n} \log f_i\left(y_i;\boldsymbol{\theta}\right)$$

Maximum likelihood estimation

- A way to estimate ***θ*** , is by finding a value $\hat{\theta}$ such that

  $$\mathrm{L}(\hat{\boldsymbol{\theta}};\mathbf{y}) \geq \mathrm{L}(\boldsymbol{\theta};\mathbf{y}), \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

  where Θ is the space of ***θ***. This value is called the ***maximum-likelihood estimate (MLE) of θ***.
- We can more easily calculate ***θ*** by maximizing the log-likelihood. The MLE of ***θ*** also maximizes this function. That is,

  $$l\left(\hat{\boldsymbol{\theta}};\mathbf{y}\right) \geq l(\boldsymbol{\theta};\mathbf{y}), \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$$
- Working with the log-likelihood is preferred because it is easier to maximize sums of functions versus products (think how much easier it is differentiating a sum versus a product of functions).

Notes

- For regular problems, $\hat{\theta}$ can be obtained by equating the first derivative of the likelihood function (or equivalently of the log-likelihood) to zero. Provided the second derivative at this point is negative, the resulting value is the MLE.
- The likelihood is a function of the parameters and we are interested in its behavior (or shape) with respect to them. We are therefore concerned with the likelihood up to a constant multiplier (log-likelihood up to an additive constant) so when working with these, multiplicative (additive) terms not involving the parameters can be dropped.

Notes (cont)

- MLE's have a number of important properties that make them desirable.
  - MLE's are asymptotically unbiased i.e. the expectation of $\hat{\theta}$, E($\hat{\theta}$), becomes equal to θ *as* $n \to \infty$
  - A MLE has a sampling distribution that is asymptotically normal with variance the inverse of minus the information, $\{\mathrm{d}^2 l / \mathrm{d}\theta^2\}^{-1}$
  - MLE's are invariant under transformation i.e. if $\hat{\theta}$ is the MLE of θ then any function of $\hat{\theta}$ will be the MLE of the same function of θ.

Example: The Binomial distribution

Consider for example the Binomial distribution, of counts $y_i, i = 1,\dots,n$ and each $y_i = 0,1,2,\dots,n_i$

$$f_i(y_i;\pi_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i}$$

The likelihood function is, from above, $\mathrm{L}(\theta;\mathbf{y}) = \prod_{i=1}^{n} f_i(y_i;\theta)$

That is $\mathrm{L}(\theta;\mathbf{y}) = \prod_{i=1}^{n} f_i(y_i;\pi_i) = \prod_{i=1}^{n} \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i}$

Maximizing this likelihood involves taking derivatives with respect to $\theta = (\pi_1,\dots,\pi_n)$

The log likelihood

Consider the log likelihood

$$
\begin{aligned}
l(\theta;\mathbf{y}) &= \sum_{i=1}^{n} \log f_i\left(y_i;\pi_i\right) \\
&= \sum_{i=1}^{n} \log\left\{\binom{n_i}{y_i}\pi_i^{y_i}\left(1-\pi_i\right)^{n_i-y_i}\right\} \\
&= \sum_{i=1}^{n} \left\{y_i \log\pi_i + (n_i - y_i)\log\left(1-\pi_i\right) + \log\binom{n_i}{y_i}\right\} \\
&= \sum_{i=1}^{n} y_i \log\pi_i + \sum_{i=1}^{n} (n_i - y_i)\log\left(1-\pi_i\right) + \sum_{i=1}^{n} \log\binom{n_i}{y_i} \\
&= \sum_{i=1}^{n} y_i \log\pi_i - \sum_{i=1}^{n} y_i \log\left(1-\pi_i\right) + \sum_{i=1}^{n} n_i \log\left(1-\pi_i\right) + C
\end{aligned}
$$

Maximizing the log likelihood

To maximize the above expression, we must take *n* derivatives with respect to π_i, set them all equal to zero and solve a system of *n* equations with *n* unknowns. Let's consider the much simpler case, where $\pi_1 = \pi_2 = \cdots = \pi_n = \pi$

Then taking a derivative with respect to π becomes

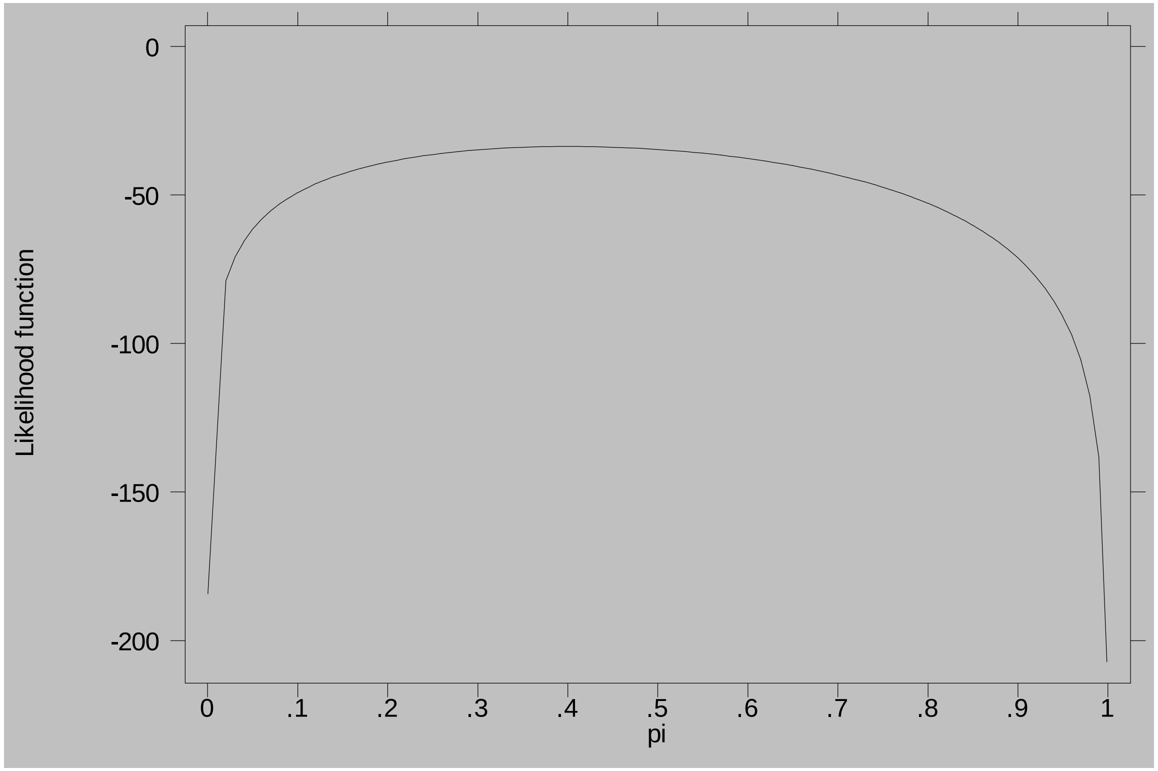
$$\frac{dl(\pi;\mathbf{y})}{d\pi} = \frac{1}{\pi}\sum_{i=1}^{n} y_i + \frac{1}{1-\pi}\sum_{i=1}^{n} y_i - \frac{1}{1-\pi}\sum_{i=1}^{n} n_i = 0$$

which becomes $\dfrac{\sum_{i=1}^{n} y_i - \pi N}{\pi(1-\pi)} = 0 \Leftrightarrow \sum_{i=1}^{n} y_i = \pi N$

since $\pi \in (0,1)$, and finally $\hat{\pi} = \dfrac{1}{N}\sum_{i=1}^{n} y_i = \bar{y}$

Example

Consider the situation $n_1=n_2=n_3=n_4=n_5=10$ and $y_1=2$, $y_2=1$, $y_3=1$, $y_4=3$, $y_5=3$. A plot of the log-likelihood is as follows:



The log likelihood is maximized at

$$\hat{\pi}=\overline{y}=0.2$$

Example: Linear regression

Suppose that two measurements (y_i,x_i) are made on each of n individuals. A simple linear regression model of *y* on *x* can be written: $y_i = \alpha + \beta x_i + e_i$

where e_i are independent with $e_i \sim N(0,\sigma^2)$

Alternatively, we can say that the random variable Y_i has a conditional distribution $Y_i \mid x_i \sim \mathrm{N}(\alpha + \beta x_i, \sigma^2)$

This shows more clearly what is actually happening in a regression model. In general, in any regression model, the conditional expectation of a random variable (say Y) given the values of one or more other variables ($x=\{x_1, x_2, \ldots, x_p\}$) is expressed as some function of these fixed variables (covariates, independent variables) and some parameters

$$\beta = \{\beta_1, \beta_2, \ldots, \beta_p\}^T$$

$$E(\mathbf{Y} \mid \mathbf{X}) = \mathrm{g}(\mathbf{X}, \boldsymbol{\beta})$$

The log likelihood

The log-likelihood can be written as:

$$l(\alpha,\beta,\sigma^2 \mid \mathbf{Y},\mathbf{X}) = -\frac{\mathrm{n}}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2$$

The MLE's of α and β can be obtained from the **score equations:**

$$U(\alpha) = l'(\alpha) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)$$

$$U(\beta) = l'(\beta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i) * x_i$$

$$U(\alpha) = 0 \Rightarrow \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$U(\beta) = 0 \Rightarrow \hat{\beta} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

Note that we do not need to know σ² to solve these. These estimates are equivalent to the least squares estimators derived by minimizing the residual sum of squares. Hence, in this case, maximum likelihood is equivalent to lest squares.

Concept of GLM's: exponential family of distributions

In GLMs the observations are assumed to arise from the *exponential family* of distributions:

$$f(y;\theta,\varphi)=\exp\{[y\theta-b(\theta)]/a(\varphi)+c(y,\varphi)\}$$

where $a(\varphi), b(\theta)$ and $c(y,\varphi)$ are known functions (McCullagh and Nelder, 1989, p. 28). The parameter **θ** is known as the *canonical parameter*. In general, it can be shown that:

$E(Y)=b'(\theta)$ and $\mathrm{var}(Y)=b''(\theta)a(\varphi)$

The variance is thus a product of two terms, $b''(\theta)$ which depends on the mean (through **θ**) which is called the *variance function V(μ)* , and the other on a(φ), a function of the form $a(\varphi)=\varphi/\omega=\sigma^2/\omega$ where φ is called the dispersion or scale parameter, is **constant over observations** and ω known prior weights that vary from observation to observation.

Example 1 – The Normal distribution

The density of a N(μ,σ²) random variable Ycan be written:

$$f(y) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(y-\mu)^2\}$$

and the logarithm of it is

$$\ln\{f(y)\} = -\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2) =$$

$$= (y\mu - \frac{\mu^2}{2})/\sigma^2 - \frac{1}{2}\{\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\}$$

so $\theta = \mu$, $b(\theta) = \frac{\theta^2}{2}$ $, a(\phi) = \alpha = \sigma^2$ $, and\ c(y,\phi) = \{\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\}$

The mean of the normal distribution $E(y) = b'(\theta) = \theta = \mu$, i.e. it is equal to the canonical parameter θ. The variance $\mathrm{var}(Y) = b''(\theta)a(\phi) = \phi = \sigma^2$ and it is of the form α(φ)=φ/ω with prior weights equal to 1.

Example 2 - The Poisson distribution

The density of a *P(λ)* random variable Y can be written:

$$f(y) = \Pr(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!} \quad y = 0,1,2,\ldots$$

and the logarithm of it is

$$\ln\{f(y)\} = y\ln(\lambda) - \lambda - \ln(y!)$$

so $\theta = \ln(\lambda)$, $b(\theta) = \lambda = e^{\theta}$, $a(\phi) = 1$, *and* $c(y,\phi) = \ln(y!)$

The mean of the Poisson distribution $E(y) = b^{'}(\theta) = e^{\theta} = \lambda$

The variance is $\mathrm{var}(Y) = b^{''}(\theta)a(\phi) = e^{\theta} * 1 = \lambda$

Example 3 – The Binomial distribution

The density of a Bin(n,π) random variable Ycan be written:

$$f(y) = \Pr(Y = y) = \binom{n}{y} \pi^{y} (1-\pi)^{n-y} \quad y = 0,1,2,\dots,n$$

and the logarithm of it is

$$\ln\{f(y)\} = y\ln(\pi) + (n-y)\ln(1-\pi) + \ln\{\binom{n}{y}\} =$$

$$= y\ln(\frac{\pi}{1-\pi}) + n\ln(1-\pi) + \ln\{\binom{n}{y}\}$$

so $\theta = \ln(\frac{\pi}{1-\pi})$, $b(\theta) = -n*\ln(1-\pi) = n*\ln(1+e^{\theta})$, $a(\phi) = 1$, *and* $c(y,\phi) = \ln\{\binom{n}{y}\}$

The mean of the Binomial distribution $E(y) = b'(\theta) = n*\frac{e^{\theta}}{1+e^{\theta}} = n*\pi$

The variance is $\mathrm{var}(Y) = b''(\theta)a(\phi) = n*\frac{e^{\theta}}{(1+e^{\theta})^2} = n*\pi(1-\pi)$