Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Applied Linear Statistical Methods

(short lecturenotes)

Prof. Rozenn Dahyot

School of Computer Science and Statistics
Trinity College Dublin
Ireland
www.scss.tcd.ie/Rozenn.Dahyot

Hilary Term 2016

1. Introduction

Problems in Statistics (or Data Science) often start with a dataset of n observations $\{y^{(i)}, x^{(i)}\}_{i=1,\cdots,N}$ and Mathematics provide several abstract objects (e.g. variable, vector, functions) in which one can plug-in this data. In the models seen in this course, for each observations $(y^{(i)}, x^{(i)})$ we associate (y_i, x_i) with y_i a variable (called response variable) and x_i a vector made up of several variables (called explanatory variables). This one-to-one mapping between data point $(y^{(i)}, x^{(i)})$ and variables (y_i, x_i) makes it pointless to differentiate them in the notations - we only use notation (y_i, x_i) confusing the two meanings. It is worth however pointing out that the likelihood function $p(y_1, y_2, \cdots, y_N | x_1, x_2, \cdots, x_N, \beta)$ is a positive function defined on a N dimensional space such that

$$\int\int \cdots \int p(y_1, y_2, \cdots, y_N | x_1, x_2, \cdots, x_N, \beta)\ dy_1\ dy_2 \cdots dy_N = 1$$

and such writing only make sense when considering variables (e.g. y_i) and not data (e.g. $y^{(i)}$). This chapter gives a quick overview of Linear Regression (section 1.1) and how its premises can be reformulated (section 1.2). We give a quick introduction of this course in paragraph 1.3 pointing out how the premises of Linear Regression will be extended.

1.1 Linear Regression - Reminder

In Linear Regression, we have a set of observations $\{(y_i, x_i)\}_{i=1,\cdots,N}$ such that the following linear relationship holds $\forall i$:

$$\begin{aligned} y_i &= \beta_0 + \beta_1\ x_{1i} + \cdots + \beta_k\ x_{ki} + \epsilon_i \\ &= \beta^T x_i + \epsilon_i \end{aligned} \tag{1.1}$$

where

- $y_i \in \mathbb{R}$ is the outcome or the observed value for the response variable,
- $x_i = (1, x_{1i}, \cdots, x_{ki})^T \in \mathbb{R}^{k+1}$ is a vector collating the values of the explanatory variables associated with the outcome y_i,
- ϵ_i is the noise, residual or error associated with the outcome y_i. It is assumed that the error ϵ has a Normal distribution with mean 0 and variance σ^2:

$$p_\epsilon(\epsilon) = \frac{\exp\left(\frac{-\epsilon^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma}$$

The best parameters β are then estimated as the ones that maximise the joint probability density function of all the residuals:

$$\hat{\beta} = \arg\max_{\beta} p(\epsilon_1, \cdots, \epsilon_N) \tag{1.2}$$

Because the residuals are independent and all follow the same distribution p_ϵ, the joint density function of the residuals corresponds to:

$$p(\epsilon_1,\cdots,\epsilon_N)=\prod_{i=1}^{N}p_\epsilon(\epsilon_i)=\prod_{i=1}^{N}\left(\frac{\exp\left(\frac{-\epsilon_i^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma}\right)=\prod_{i=1}^{N}\left(\frac{\exp\left(\frac{-(y_i-\beta^T x_i)^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma}\right)$$

Using the log transformation of the joint density function of the residuals, the maximum likelihood estimate of the parameters β (cf. eq. 1.2) is in fact computed by minimising the sum of square errors:

$$\hat{\beta}=\arg\min_{\beta}\left\{\sum_{i=1}^{N}\epsilon_i^2=\sum_{i=1}^{N}(y_i-\beta^T x_i)^2\right\}$$

and a forecast for output response $\hat{y}=\hat{\beta}^T x$ can be computed for any chosen input x - confidence intervals can also be computed for modelling uncertainty associated with $\hat{y}$.

1.2 Linear Regression - Reformulation

Consider the equation $y=\beta^T x+\epsilon$ (without the index i) for a moment, if x and β are *given*, then the only uncertainty related to the response y is having the same properties as the uncertainty defined for ϵ. In other words, we have:

$$p_{y|\beta,x}(y|\beta,x)=p_\epsilon(y-\beta^T x)$$

where $p_{y|\beta,x}$ is the probability density function of y *given* x and β (this can be understood as a change of variable $\epsilon=y-\beta^T x$). So saying that the error has a Normal distribution with mean 0 and variance σ^2 is the same as assuming that the conditional probability density function of the response y given the parameters β and the explanatory variables x is:

$$p_{y|\beta,x}(y|\beta,x)=\frac{\exp\left(\frac{-(y-\beta^T x)^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma}$$

which is a Normal distribution with mean $\beta^T x$ and variance σ^2. So we can in fact define the Linear Regression problem without introducing explicitly a variable ϵ as follow:

2.1 Definition (Premises for Linear Regression) Consider a set of observations $\{(y_i,x_i)\}_{i=1,\cdots,N}$ collected independently, such that $\forall i$

- the response y_i is normally distributed
- with mean $\mathbb{E}[y_i]=\beta^T x_i$ (and variance $\mathbb{E}[(y_i-\beta^T x_i)^2]=\sigma^2$).

so $y_i\sim p_{y|x,\beta}(y_i|x_i,\beta)$ is fully defined and the parameters can be estimated using the likelihood function:

$$\hat{\beta}=\arg\max_{\beta}\left\{\mathscr{L}(\beta)=p(y_1,\cdots,y_N|x_1,\cdots,x_N,\beta)=\prod_{i=1}^{N}p_{y|x,\beta}(y_i|x_i,\beta)\right\}$$

providing a parametric model $p_{y|x,\beta}(y|x,\hat{\beta})$.

1.3 Generalised Linear Models

In a nutshell, this course will generalise these premises used for Linear Regression as follow:

- The probability density function $p_{y|x,\beta}$ is a member of the exponential family of distributions. The Normal distribution is a member of that family, but other distributions are also available to deal with various situations such as when the outcome y is not an element of $\mathbb{R}$ (as assumed by the Normal distribution) but is for instance in binary form (e.g. the outcome y indicates a failure 0 or success 1).

- The expectation of y given x and β that is defined by:

$$\mathbb{E}[y] = \int_{\mathbb{Y}} y \; p_{y|x\beta}(y|x\beta) \; dy, \quad \text{with } \mathbb{Y} \text{ domain of definition of the outcome } y$$

  is now related to the explanatory variables with a link function g such that

$$g\left(\mathbb{E}[y]\right) = \beta^T x$$

  For instance in the case of Linear Regression, the link function g that we used is the identity function g defined for $z \in \mathbb{R} \rightarrow g(z) = z \in \mathbb{R}$. g is a link function that will be chosen such that it is bijective and its inverse g^{-1} exists. In general, this function maps the space of the expectation $\mathbb{E}[y]$ to the space $\mathbb{R}$ where $\beta^T x$ takes its value.

2. Distributions for response variable y

2.1 Exponential family of distributions

1.1 Definition (probability distributions) The probability density function (p.d.f.) p_y of the random variable y has the following properties:

- $\forall y \in \mathbb{Y}$, $p_y(y) \geq 0$ (the function p_y is positive for all possible outcomes, $\mathbb{Y}$ is the notation used for the space of all possible outcomes)
- the function p_y integrates to 1 on the space of all possible outcomes such that:
  - when y is a continuous random variable:
$$\int_{\mathbb{Y}} p_y(y)\ dy = 1$$
  - when y is a discrete random variable:
$$\sum_{y \in \mathbb{Y}} p_y(y) = 1$$

where $\mathbb{Y}$ is the space of all possible outcomes.

Only distributions from the exponential family of distributions, will be considered for the response variable y.

1.2 Definition (Exponential family of distributions) The distribution belongs to the exponential family if it can be written as:

$$p_{y|\theta}(y|\theta) = \exp\left[a(y)b(\theta) + c(\theta) + d(y)\right] \tag{2.1}$$

where a, b, c, d are known functions. If $a(y) = y$ then the distribution is said to be in **canonical form**.

1.3 Definition (Expectation) Consider a random variable y with p.d.f. p_y on the space of outcomes $\mathbb{Y}$, the expectation of y is computed by:

- when y is a continuous random variable:
$$\mathbb{E}[y] = \int_{\mathbb{Y}} y\ p_y(y)\ dy$$
- when y is a discrete random variable:
$$\mathbb{E}[y] = \sum_{y \in \mathbb{Y}} y\ p_y(y)$$

2.2 Some members of the exponential family

Several probability density functions (pdf) are introduced in this section.

Exercises. You should be able to show that these pdf are positive functions, integrating to 1 on the space of outcomes, compute the expectations, and be able to identify the function a, b, c, d to show that they are members of the exponential family of distributions.

2.1 Definition (Gaussian (or Normal) distribution) The normal distribution associated with a random variable $y \in \mathbb{R}$ is defined as:

$$p_{y|\theta}(y|\theta,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y-\theta)^2}{2\sigma^2}\right)$$

$\theta \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$ are resp. the mean and standard deviation. The expectation is $\mathbb{E}[y] = \theta$.

2.2 Definition (Poisson distribution) The Poisson distribution expresses the probability of a given number of events y occurring in a fixed interval of time and/or space, (and/or fixed total population size)

$$p_{y|\theta}(y|\theta) = \frac{\theta^y \exp(-\theta)}{y!}, \quad y \in \mathbb{N},\ \theta \in \mathbb{R}^{+*} \tag{2.2}$$

The expectation of y is $\mathbb{E}[y] = \theta$ represents the average of the number of events.

As an example, the response y associated with the Poisson distribution can be modelling the number of people in line in front of you at the grocery store.

2.3 Definition (Binary variable) A binary variable y has only two possible outcomes ($\mathbb{Y} = \{0, 1\}$):

$$y = \begin{cases} 1 & \text{if the outcome is a } \textit{success} \\ 0 & \text{if the outcome is a } \textit{failure} \end{cases} \tag{2.3}$$

The notions *success* and *failure* are user defined.

2.4 Definition (Bernoulli distribution) Having y a binary variable, lets $p_{y|\theta}(y = 1|\theta) = \theta$ then $p_{y|\theta}(y = 0|\theta) = 1 - \theta$ and more generally:

$$p_{y|\theta}(y|\theta) = \theta^y\ (1-\theta)^{1-y}, \quad y \in \{0, 1\},\ \theta \in [0; 1] \tag{2.4}$$

is the Bernoulli distribution $\text{Bernouilli}(\theta)$ and $\mathbb{E}[y] = \theta$.

A standard example for using the Bernouilli distribution is when the response y is the outcome of a single toss of a coin where head is encoded $y = 1$ and tail is $y = 0$.

2.5 Definition (Binomial distribution) Consider the response y that is the number of successes in n trials (so the space of outcomes is $\mathbb{Y} = \{0, 1, \cdots, n\}$), and the proportion θ is a real number between 0 and 1. The Binomial distribution is defined as:

$$p_{y|\theta}(y|\theta) = \frac{n!}{(n-y)!y!}\ \theta^y (1-\theta)^{n-y} \tag{2.5}$$

The expectation is $\mathbb{E}[y] = n\theta$. The Bernouilli distribution corresponds to the Binomial distribution with the number of trial n is 1.

The Binomial distribution is used when considering the number y of heads when tossing a coin n times. Another example is when y is the number of students passing an exam amongst n students taking that exam.

2.6 Definition (Exponential distribution) The exponential distribution is defined as:

$$p_{y|\theta}(y|\theta) = \theta \ \exp\left(-\theta \ y\right), \quad \text{with } y \in \mathbb{R}^+, \ \theta \in \mathbb{R}^{+*}$$

The expectation is $\mathbb{E}[y] = \frac{1}{\theta}$.

2.7 Definition (Weibull distribution) The Weibull distribution is defined as

$$p_{y|\lambda\theta}(y|\lambda,\theta) = \lambda \ \theta \ y^{\lambda-1} \exp\left[-\theta \ y^{\lambda}\right], \quad \text{with } y \in \mathbb{R}^+, \ \theta \in \mathbb{R}^{+*}, \ \lambda \in \mathbb{R}^{+*}$$

The expectation is $\mathbb{E}[y] = \left(\frac{1}{\theta}\right)^{1/\lambda} \Gamma\left(1+\frac{1}{\lambda}\right)$ where $\Gamma(u) = \int_0^{+\infty} s^{u-1} \ \exp(-s) \ ds$. Note that the exponential distribution is a special case of the Weibull distribution with $\lambda = 1$.

The Weibull distribution (or exponential distribution) is often used to model a time to failure y, for instance the time taken by a new hard drive to eventually fail.

3. Generalized linear Models (GLM)

3.1 Formal structure for the class of generalized Linear Models

Generalized linear Models are a large class of statistical models defined such that:

(1) We have collected independently a set of responses y_i as well as the values for some explanatory variables stored in the vector x_i. In other words, observations that are collected are $\{(y_i, x_i)\}_{i=1,\cdots,N}$.

(2) The response y_i has a distribution $p_{y|\theta}(y_i|\theta_i)$ that is a member of the exponential family, indexed by the parameter θ_i that is related to the expectation of the response $\mathbb{E}[y_i]$.

(3) A model is constructed by linking the expectation of the response $\mathbb{E}[y_i] = \mu_i$ with the linear predictor $x_i^T\beta = \beta_0 + \beta_1\ x_{1i} + \cdots + \beta_k\ x_{ki}$ such that $g(\mu_i) = x_i^T\beta$ with g the link function. The expected response is:

$$\mathbb{E}[y_i] = g^{-1}\left(x_i^T\beta\right)$$

(4) The link function g is a monotonic differentiable function (hence the inverse function g^{-1} exists).

(5) The joint density function of the responses given the parameters $\theta_i \propto \mathbb{E}[y_i]$ corresponds to:

$$\begin{aligned}\mathscr{L}(\theta_1,\cdots\theta_N) &= p(y_1,\cdots,y_N|\theta_1,\cdots\theta_N)\\ &= \prod_{i=1}^N p_{y|\theta}(y_i|\theta_i)\end{aligned} \tag{3.1}$$

When the likelihood $\mathscr{L}(\theta_1,\cdots\theta_N)$ is unconstraint (the model is said to be **saturated**) then the maximum likelihood solution corresponds to $(\forall i = 1,\cdots,N)$

$$\begin{aligned}\hat{\theta}_i &= \arg\max_{\theta_i} \mathscr{L}(\theta_1,\cdots\theta_N)\\ &= \arg\max_{\theta_i} p_{y|\theta}(y_i|\theta_i)\end{aligned}$$

The maximum likelihood estimate $\hat{\theta}_i$ of the **saturated model** is found by solving $\frac{\partial p_{y|\theta}(y_i|\theta_i)}{\partial\theta_i} = 0$.

(6) When $\mathbb{E}[y_i]$ (e.g. θ_i) is related to $x_i^T\beta$ using a link function g, then the likelihood function can be rewritten as a function of β (and the explanatory variables):

$$\begin{aligned}\mathscr{L}(\beta) &= p(y_1,\cdots,y_N|x_1,\cdots,x_N,\beta)\\ &= \prod_{i=1}^N p_{y|\theta}(y_i|x_i,\beta)\end{aligned} \tag{3.2}$$

The maximum likelihood estimate $\hat{\beta}$ corresponds to:

$$\hat{\beta} = \arg\max_{\beta} \mathscr{L}(\beta)$$

and this estimate is found by solving $\frac{\partial\mathscr{L}(\beta)}{\partial\beta} = 0$. With this estimate, we can propose the following model for linking $\mathbb{E}[y]$ with any input explanatory variables x:

$$\hat{\theta}(x) = g^{(-1)}(x^T\hat{\beta})$$

3.2 Statistical analysis with GMLs

Different keywords are used for statistical techniques that are GLMs (cf. fig. 3.1):

- Linear regression: the natural link function g is the identity ($\theta \in \mathbb{R}$).
- Poisson regression: the natural link function g is the log ($\theta \in \mathbb{R}^{+*}$).
- Binomial regression: the natural link function g is the logit function ($\theta \in [0,1]$):

$$g(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$$

- Survival analysis: the natural link function will be the log function.

Note how these proposed link functions relate to the function $b(\theta)$ defined for distributions in canonical form in the exponential family of distributions. Other link functions can be used.

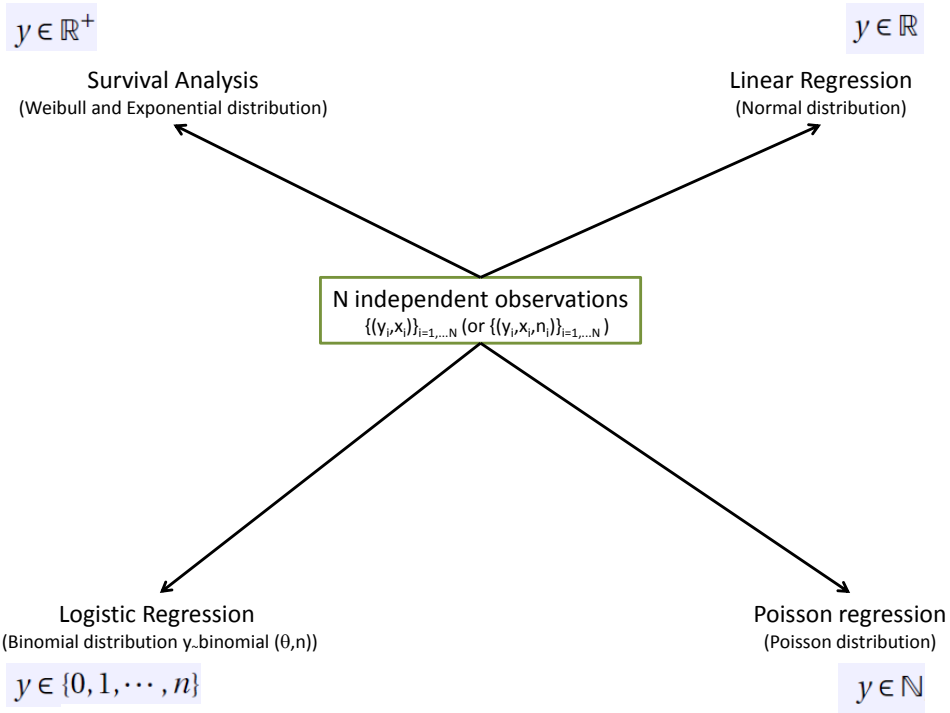

Figure 3.1: Generalised Linear models encapulate several techniques for statistical analysis of data.

Note that in some experiments the observations collected are y_i the response, x_i the explanatory variables and in addition a value n_i is provided e.g. when y_i is the number of successes in n_i trials. This is an indication that the Binomial distribution is appropriate to model the response.

4. Binomial & Poisson Distributions

The Poisson distribution is suitable to model outcomes that represent numbers of events or occurrences. When the recorded data is in the form $\{(y_i, x_i, n_i)\}_{i=1,\cdots,N}$ where the outcome y_i is the number of successes amongst n_i trials, the binomial distribution seems the most suitable distribution to adapt to different exposures n_i because this information appears explicitly in the mathematical definition of the Binomial distribution. The Poisson distribution can also be adapted to deal with various exposures (paragraph 4.1). The relation between Binomial and Poisson distributions is investigated in paragraph 4.2 when $n \to +\infty$.

4.1 Offset and Exposure

Let considers the Poisson distribution for modelling a count y (out of n trial) such that:

- $y \sim \text{Poisson}(\lambda)$ so the distribution for y given parameter λ is:

$$p_{y|\lambda}(y|\lambda) = \frac{\lambda^y \ \exp(-\lambda)}{y!}$$

  In this case, $\mathbb{E}[y] = \lambda$

- Remember that when considering the Binomial distribution $y \sim \text{Bin}(n, \theta)$ then $\mathbb{E}[y] = n\theta$ where θ is a proportion between 0 and 1.

We have the following correspondence with the proportion θ:

$$\mathbb{E}[y] = \lambda = n\theta$$

or $\theta = \frac{\lambda}{n}$ and the link function g is applied to the proportion as follow:

$$g(\theta) = g\left(\frac{\lambda}{n}\right) = x^T\beta$$

The canonical link function for the Poisson distribution is the log function so:

$$\log\left(\frac{\lambda}{n}\right) = x^T\beta \quad \text{or} \quad \log(\lambda) = \log(n) + x^T\beta$$

1.1 Definition (Poisson regression: offset and exposure) When using the Poisson distribution, the population n is called the exposure (e.g. for the Beetle case study, n beetles are exposed to a toxic substance) while $\log(n)$ is called the offset. The log is used as a link function such that

$$\log\left(\mathbb{E}[y]\right) = \log(\lambda) = \log(n) + x^T\beta$$

or

$$\theta = \frac{\lambda}{n} = \exp(x^T\beta)$$

Offsets are used to correct for the group size or differing time periods of observation when using the Poisson distribution.

4.2 Relation between Poisson and Binomial distributions when $n \to +\infty$

Show that

$$\lim_{n\to\infty} \underbrace{\frac{n!}{(n-y)!y!}\,\theta^y(1-\theta)^{n-y}}_{\text{Binomial}(n,\theta)} = \underbrace{\frac{\lambda^y\,\exp(-\lambda)}{y!}}_{\text{Poisson}(\lambda)}$$

with $\lambda = n\theta$.

Solution. Changing $\theta = \frac{\lambda}{n}$ in the binomial distribution:

$$\frac{n!}{(n-y)!y!}\left(\frac{\lambda}{n}\right)^y\left(1-\frac{\lambda}{n}\right)^{n-y} = \underbrace{\frac{n!}{(n-y)!\;n^y}}_{A_n}\;\frac{\lambda^y}{y!}\;\underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{B_n}\left(1-\frac{\lambda}{n}\right)^{-y}$$

We see that $\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-y} = 1$, hence

$$\lim_{n\to\infty}\frac{n!}{(n-y)!y!}\left(\frac{\lambda}{n}\right)^y\left(1-\frac{\lambda}{n}\right)^{n-y} = \frac{\lambda^y}{y!}\lim_{n\to\infty} A_n\,B_n$$

- $\lim_{n\to\infty} A_n$?

  $$A_n = \frac{n!}{(n-y)!n^y} = \frac{n\;(n-1)\;\cdots(n-y+1)}{n\times n\times\cdots\times n} = 1\;\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{y-1}{n}\right)$$

  so

  $$\lim_{n\to\infty} A_n = 1$$

- $\lim_{n\to\infty}\; B_n$? we know $(1+s)^n = \sum_{k=0}^n \binom{n}{k} s^k$ hence

  $$\left(1-\frac{\lambda}{n}\right)^n = \sum_{k=0}^{n} \underbrace{\frac{n!}{(n-k)!\;n^k}}_{A_n}\frac{(-\lambda)^k}{k!}$$

  The Taylor expansion of $\exp(-\lambda)$ is:

  $$\exp(-\lambda) = \sum_{k=0}^{\infty}\frac{(-\lambda)^k}{k!}$$

  So

  $$\lim_{n\to\infty}\; B_n = \exp(-\lambda)$$

Hence with $\lambda = n\theta$

$$\lim_{n\to\infty} \underbrace{\frac{n!}{(n-y)!y!}\,\theta^y(1-\theta)^{n-y}}_{\text{Binomial}(n,\theta)} = \underbrace{\frac{\lambda^y\,\exp(-\lambda)}{y!}}_{\text{Poisson}(\lambda)}$$

□

5. Akaike Information Criterion

5.1 Likelihood and log-likelihood

Considering the independent observations $\{(y_i, x_i)\}_{i=1,\cdots,N}$ (or $\{(y_i, x_i, n_i)\}_{i=1,\cdots,N}$), the likelihood is defined as:

$$\mathscr{L}(\theta_1, \cdots, \theta_N) = \prod_{i=1}^{N} p_{y|\theta}(y_i|\theta_i)$$

with $p_{y|\theta}$ a probability distribution from the exponential family i.e.:

$$\mathscr{L}(\theta_1, \cdots, \theta_N) = \prod_{i=1}^{N} \exp\big(a(y_i)\ b(\theta_i) + c(\theta_i) + d(y_i)\big)$$

The log transformation of the likelihood is often computed instead:

$$\log \mathscr{L}(\theta_1, \cdots, \theta_N) = \sum_{i=1}^{N} \big(a(y_i)\ b(\theta_i) + c(\theta_i) + d(y_i)\big)$$

and the maximum likelihood estimates for the saturated model are then computed with:

$$(\hat{\theta}_1, \cdots, \hat{\theta}_N) = \operatorname{argmax}\ \log \mathscr{L}(\theta_1, \cdots, \theta_N)$$

When a generalised linear model is used, a link function g is used to constraint the parameters such that $\theta_i \propto\ g^{-1}(x_i^T \beta),\ \forall i = 1, \cdots, N$. In this case the likelihood is written $\mathscr{L}(\beta)$ and the log likelihood is $\log \mathscr{L}(\beta)$. The parameter β has often a lower dimension than θs (i.e. $\dim(\beta) \leq N$) and the maximum likelihood estimate is computed such that:

$$\hat{\beta} = \operatorname{argmax}\ \log \mathscr{L}(\beta)$$

5.2 Comparing Working models with the AIC

For analysing data, it is not rare that more than one distribution from the exponential family can be used, that several link functions can be selected, and also, when several explanatory variables are recorded, several relation $x^T \beta$ can be defined. For instance, some explanatory variables may not be helpful in explaining the responses and should not be integrated in the model, limiting the number of parameters $\beta_0, \beta_1, \cdots$ to estimate.

To compare the different models proposed for the same dataset, several criteria exist to facilitate the selection of the best model. We focus here on the Akaike Information Criterion (AIC) that is given as an output of the function `glm` in R.

2.1 Definition (Akaike Information Criterion) The Akaike Information Criterion is a measure of goodness of fit defined as:

$$AIC = -2\ \log \mathscr{L}(\hat{\beta}) + 2\ p \tag{5.1}$$

where

- $p = \dim \beta$ is the number of parameters to be estimated in the model,
- $\hat{\beta}$ are the estimated parameters that maximize the likelihood (or log likelihood),
- $\log \mathscr{L}(\hat{\beta})$ is the maximum value of the log likelihood.

When fitting several models to the data, the one having the smallest AIC is selected. Note that the best model is a trade off in between one that maximizes the likelihood with also having the minimum number of parameters.

6. Deviance

6.1 Deviance

The deviance is a log-likelihood ratio statistics that compares the saturated model with the proposed GLM model. We have already introduced the following terms:

- the maximum likelihood estimates for the saturated model are computed by:

$$(\hat{\theta}_1, \cdots, \hat{\theta}_N) = \arg\max \{\log \mathscr{L}(\theta_1, \cdots, \theta_N)\}$$

  and the value $\log \mathscr{L}(\hat{\theta}_1, \cdots, \hat{\theta}_N)$ is therefore the maximum value of the log likelihood function of the saturated model.

- When considering a generalised linear model, a link function g is used to constraint the parameters such that $\theta_i \propto g^{-1}(x_i^T \beta),\ \forall i = 1, \cdots, N$. In this case the likelihood is written $\mathscr{L}(\beta)$ and the log likelihood is $\log \mathscr{L}(\beta)$. The parameter β has often a lower dimension than θs (i.e. $\dim(\beta) \leq N$) and the maximum likelihood estimate is computed such that:

$$\hat{\beta} = \arg\max \ \{\log \mathscr{L}(\beta)\}$$

  The maximum log likelihood value for the GLM model is then $\log \mathscr{L}(\hat{\beta})$.

1.1 Definition (Deviance) The deviance, also called the log-likelihood (ratio) statistic, is defined by:

$$D = 2\ \{\log \mathscr{L}(\hat{\theta}_1, \cdots, \hat{\theta}_N) - \log \mathscr{L}(\hat{\beta}))\} = 2\ \log\left(\frac{\mathscr{L}(\hat{\theta}_1, \cdots, \hat{\theta}_N)}{\mathscr{L}(\hat{\beta})}\right) \tag{6.1}$$

where

- $\log \mathscr{L}(\hat{\theta}_1, \cdots, \hat{\theta}_N)$ is the maximum value of the log-likelihood function for the saturated model, and
- $\log \mathscr{L}(\hat{\beta})$ is the value of the log-likelihood function when fitting the model $g(\mathbb{E}[y]) = \mathbf{x}^T \hat{\boldsymbol{\beta}}$.

The deviance is given as an output in R when fitting GLMs.

6.2 Approximation of the log likelihood function near its maximum

Using Taylor expansion, the log likelihood can be approximated near the maximum likelihood estimate

- for the saturated model with notation $(\theta_1, \cdots, \theta_N)^T = \theta$, when θ is close to $\hat{\theta}$:

$$\log \mathscr{L}(\theta) \simeq \log \mathscr{L}(\hat{\theta}) + (\theta - \hat{\theta})^T\ \nabla_{\hat{\theta}} + \frac{1}{2}(\theta - \hat{\theta})^T \mathrm{H}_{\hat{\theta}}\ (\theta - \hat{\theta})$$

  with $\nabla_{\hat{\theta}}$ the gradient of the log likelihood function at $\hat{\theta}$ and $\mathrm{H}_{\hat{\theta}}$ the hessian matrix of the log likelihood function at $\hat{\theta}$ for the saturated model.

- similarly for the GLM model, when β is close to $\hat{\beta}$:

$$\log \mathscr{L}(\beta) \simeq \log \mathscr{L}(\hat{\beta}) + (\beta - \hat{\beta})^T \nabla_{\hat{\beta}} + \frac{1}{2}(\beta - \hat{\beta})^T \mathrm{H}_{\hat{\beta}} (\beta - \hat{\beta})$$

In both case, the gradients $\nabla_{\hat{\beta}}$ and $\nabla_{\hat{\theta}}$ are zero-vectors (since $\hat{\beta}$ and $\hat{\theta}$ are maxima of the log likelihoods !). So the deviance can be approximated by:

$$\begin{aligned} D &\simeq 2 \left\{\log \mathscr{L}(\hat{\theta}) - \log \mathscr{L}(\hat{\beta}))\right\} \\ &\simeq 2 \left\{\log \mathscr{L}(\theta) - \tfrac{1}{2}(\theta - \hat{\theta})^T \mathrm{H}_{\hat{\theta}} (\theta - \hat{\theta}) - \log \mathscr{L}(\beta) + \tfrac{1}{2}(\beta - \hat{\beta})^T \mathrm{H}_{\hat{\beta}} (\beta - \hat{\beta})\right\} \\ &\simeq \underbrace{2 \log\left(\frac{\mathscr{L}(\theta)}{\mathscr{L}(\beta)}\right)}_{v} - (\theta - \hat{\theta})^T \mathrm{H}_{\hat{\theta}} (\theta - \hat{\theta}) + (\beta - \hat{\beta})^T \mathrm{H}_{\hat{\beta}} (\beta - \hat{\beta}) \end{aligned}$$

The term v is positive and it will be near zero if the GLM model fits the data almost as well the saturated model does. Note that Hessian matrices computed at the maxima, $\mathrm{H}_{\hat{\beta}}$ and $\mathrm{H}_{\hat{\theta}}$, are negative.

6.3 Sampling distribution for the deviance

The likelihood for β can be approximated by a Normal distribution near the estimate $\hat{\beta}$ such that $\mathscr{L}(\beta) \propto p_{\hat{\beta}}(\beta)$ with:

$$\begin{gathered} p_{\hat{\beta}}(\beta) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left[-\tfrac{1}{2}(\beta - \hat{\beta})^T \Sigma^{-1} (\beta - \hat{\beta})\right] \\ \text{or} \\ \log p_{\hat{\beta}}(\beta) = -\log\left(\sqrt{2\pi|\Sigma|}\right) - \tfrac{1}{2}(\beta - \hat{\beta})^T \Sigma^{-1} (\beta - \hat{\beta}) \end{gathered} \tag{6.2}$$

Comparing with

$$\log \mathscr{L}(\beta) \simeq \log \mathscr{L}(\hat{\beta}) + \frac{1}{2}(\beta - \hat{\beta})^T \mathrm{H}_{\hat{\beta}} (\beta - \hat{\beta})$$

we can identify the covariance matrix with the Hessian matrix of the log likelihood computed at $\hat{\beta}$:

$$\Sigma^{-1} = -\mathrm{H}_{\hat{\beta}}$$

Remember that the covariance matrix is a real positive definite symmetric matrix that can be rewritten by eigen decomposition as:

$$\Sigma = \mathrm{U}^T \Lambda \mathrm{U}$$

with the orthogonal matrix U and the diagonal matrix $\Lambda = \Upsilon^2$ collecting the positive eigenvalues of Σ. In this case, we have

$$\begin{aligned} (\theta - \hat{\theta})^T \Sigma^{-1} (\theta - \hat{\theta}) &= (\mathrm{U}\ (\theta - \hat{\theta}))^T \Lambda^{-1} (\mathrm{U}\ (\theta - \hat{\theta})) \\ &= (\Upsilon^{-1}\ \mathrm{U}\ (\theta - \hat{\theta}))^T\ \mathrm{I}\ (\Upsilon^{-1}\ \mathrm{U}\ (\theta - \hat{\theta})) \end{aligned}$$

with I is the identity matrix. So by changing variable $\vec{z} = (\Upsilon^{-1}\ \mathrm{U}\ (\theta - \hat{\theta}))$, we have $\vec{z} \sim \mathscr{N}(0, \mathrm{I})$.

3.1 Definition (The central χ^2 distribution) The central χ^2 distribution with k degree of freedom is defined as the distribution of the sum of squares of k independent random variables z_1, ..., z_k such that $z_i \sim \mathcal{N}(0,1), \forall i = 1, \cdots, k$:

$$x = \sum_{i=1}^{k} z_i^2 \quad \text{then } x \sim \chi^2(k)$$

with the distribution $\chi^2(k)$ defined as:

$$p_x(x) = \begin{cases} \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} \exp\left(\frac{-x}{2}\right) & \text{when } x \geq 0 \\ 0 \quad \text{otherwise} \end{cases}$$

with the function Γ defined as

$$\Gamma(k) = \int_0^{+\infty} t^{k-1} \exp(-t)\, dt$$

$\Gamma(1/2) = \sqrt{\pi}$, $\Gamma(1) = 1$ and $\Gamma(y+1) = y\,\Gamma(y)\ \forall y$. The expectation of x is $\mathbb{E}[x] = k$.

For the deviance,

- the term $-(\theta - \hat{\theta})^T \mathrm{H}_{\hat{\theta}}\ (\theta - \hat{\theta})$ is a weighted sum of squares of N variables with distribution $\mathcal{N}(0,1)$, then this term follows a $\chi^2(\dim(\theta))$ distribution with $\dim(\theta) = N$ for the saturated model
- similarly, the term $-(\beta - \hat{\beta})^T \mathrm{H}_{\hat{\beta}}\ (\beta - \hat{\beta})$ will follow a $\chi^2(\dim(\beta))$ distribution with $\dim(\beta) = p$.

3.2 Definition (Sampling Distribution of the deviance) The sampling distribution of the deviance is approximated by a χ^2 distribution of degree $\dim(\theta) - \dim(\beta)$ and non-centrality parameter ν:

$$D \sim \chi^2(\dim(\theta) - \dim(\beta), \nu) \tag{6.3}$$

with $\dim(\theta) = N$ the nb of parameters in the saturated model, $\dim(\beta) = p$ is the number of parameters in the GLM model of interest, and ν is a positive constant near 0 when the model of interest fits almost as well as the saturated model:

$$D \sim \chi^2(N - p, 0)$$

Example. Once you have fitted a model (e.g. for $N = 8$ response variables and with $p = 2$ parameters (β_0, β_1) in the Beetles case), read in the statistics tables to find D_{95} (cf. fig. 6.1) and compare the deviance computed with your model with D_{95}:

- if $D < D_{95}$ accept the model,
- if $D > D_{95}$ reject the model,
- you may decide to look for a better model also when D is close to D_{95}.

Exercises

(1) Find the deviance for the following distributions used with their canonical link functions:

- Binomial
- Poisson
- Normal

(2) Assuming that $z_1 \sim \mathcal{N}(0,1)$, show that $x = z_1^2$ has a $\chi^2(1)$ distribution.

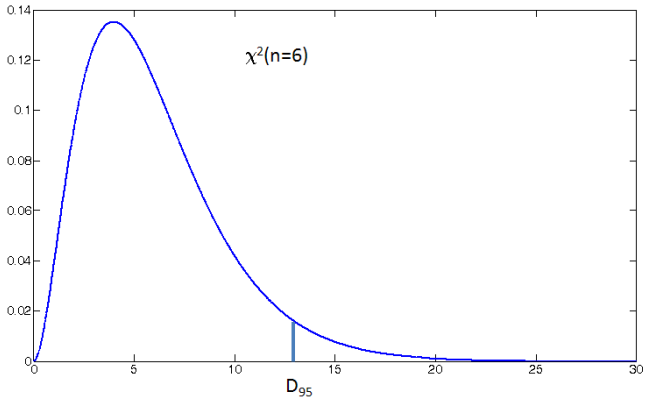

Figure 6.1: $\chi^2(6)$ distribution. Integration of $\chi^2(6)$ between 0 and D_{95} is equal to 0.95 (the value of $D.95 \simeq 12.59159$ and this can be found in the statistical tables or with R). In other words, the interval $[0; D_{95}]$ is the 95% confidence interval.

(3) Assuming that $z_1 \sim \mathcal{N}(0,1)$ and $z_2 \sim \mathcal{N}(0,1)$ such that z_1 and z_2 are independent, show that $x = z_1^2 + z_2^2$ has a $\chi^2(2)$ distribution.

(4) Given $x \sim \chi^2(k)$, show that $\mathbb{E}[x] = k$.

7. Explanatory variables in GLMs

So far, we have found that various GLMs for analysing data can be defined by selecting various distributions, and various link functions. In this chapter, we focus on the design of the linear relation $\beta^T x$.

7.1 Nature of variables

Responses and explanatory variables can be:

- nominal e.g. (red,green,blue); (dead,alive), (male, female)
- ordinal in which there is a natural order or ranking e.g. age categories
- continuous e.g. time, weight, temperature etc.

Categorical variables refer to nominal and ordinal data. When considering a linear form with a continuous variable x, additionnal polynomial explanatory variables can be considered in the model e.g.:

- using x as explanatory variable: $\beta_0 + \beta_1\ x\ (\dim(\beta) = 2)$, $\beta_1\ x\ (\dim(\beta) = 1)$
- using x^2 as explanatory variable: $\beta_0 + \beta_2\ x^2\ (\dim(\beta) = 2)$, $\beta_2\ x^2\ (\dim(\beta) = 3)$
- using x and x^2 as explanatory variables: $\beta_0 + \beta_1 x + \beta_2\ x^2\ (\dim(\beta) = 3)$, $\beta_1 x + \beta_2\ x^2\ (\dim(\beta) = 2)$

7.2 Generalized Mixed Linear Models

2.1 Definition (mixed models) Consider the scenario of having m clusters of data where the response is modeled as a function of a single regressor x. The generalized linear mixed model is:

$$g(\mathbb{E}[y]) = \beta_0 + \beta_1\ x + \sum_{j=1}^{m} \delta_j\ (\alpha_j + \gamma_j\ x)$$

with

- x continuous explanatory variable,
- δ_j is the indicator variable for the j^{th} cluster,
- α_j is the intercept and γ_j is the slope for the j^{th} cluster.

Example. In the case study Brit doctors, the model is

$$g(\mathbb{E}[y]) = \beta_0 + \beta_1\ x + \beta_2\ x^2 + \delta\ (\alpha_1 + \gamma_1\ x)$$

with $\delta = 1$ for smokers and $\delta = 0$ for non-smokers (categorical data), and x is the age. This allows us to model a linear relation for non-smokers ($\delta = 0$):

$$g(\mathbb{E}[y]) = \beta_0 + \beta_1 \; x + \beta_2 \; x^2$$

and for smokers ($\delta = 1$):

$$g(\mathbb{E}[y]) = (\beta_0 + \alpha_1) + (\beta_1 + \gamma_1) \; x + \beta_2 \; x^2$$

8. Survival Analysis

Survival analysis is concerned by the statistical modelling of the time to 'failure' from a well defined origin or starting point. For instance:

- time of hard-drive to fail from the time it has been built or bought (computer science),
- time of a patient to die from the time the disease has been diagnosed (medicine).

8.1 Distributions

Specificity of survival times:

- the times are non-negative and have skewed distributions with long tails,
- some subjects may survive beyond the study and their failure time is not observed. In this case, the data are said to be censored.

1.1 Definition (Exponential distribution) The exponential distribution is defined as

$$p_{y|\theta}(y|\theta) = \theta \ \exp\left(-\theta \ y\right), \quad \text{with } y \in \mathbb{R}^+, \ \theta \in \mathbb{R}^{+*}$$

Exercise: Show that $\mathbb{E}[y] = \frac{1}{\theta}$ for the exponential distribution.

1.2 Definition (Weibull distribution) The Weibull distribution is defined as

$$p_{y|\lambda\theta}(y|\lambda,\theta) = \lambda \ \theta \ y^{\lambda-1} \exp\left[-\theta \ y^{\lambda}\right], \quad \text{width } y \in \mathbb{R}^+, \ \theta \in \mathbb{R}^{+*}, \ \lambda \in \mathbb{R}^{+*}$$

Note that the exponential distribution is a special case of the Weibull distribution with $\lambda = 1$. The Weibull distribution can be rewritten:

$$p_{y|\theta\lambda}(y|\lambda,\theta) = \exp\left[-\theta \ y^{\lambda} + \log(\lambda) + \log\theta + (\lambda-1)\log(y)\right]$$

Exercise: Show that $\mathbb{E}[y] = \left(\frac{1}{\theta}\right)^{1/\lambda} \Gamma(1+\frac{1}{\lambda})$ for the Weibull distribution with $\Gamma(u) = \int_0^{+\infty} s^{u-1} \ \exp(-s) \ ds$.

8.2 Survivor and hazard functions

The random variable y denotes the survival time and $f(\cdot)$ is its p.d.f. (i.e. either the exponential $p_{y|\theta}$ or Weibull $p_{y|\lambda\theta}$).

- The **probability of failure** before a specific time is:

$$F(y) = \mathbb{P}(0 \leq t \leq y) = \int_0^y f(t) \ dt$$

The **median survival time** is given by the solution of the equation $F(y) = .5$ and it serves as the *average survival time*.

- The **survivor function** is the probability of survival beyond time y:

$$S(y) = \int_y^{+\infty} f(t)\ dt = 1 - F(y)$$

- The **hazard function** is defined as:

$$h(y) = \frac{f(y)}{S(y)} = -\frac{d\log(S(y))}{dy}$$

$h(y)$ can be understood as the probability of failure in between time y and $y+\delta y$ (with $\delta y \to 0$) given that the subject has survived up to time y. The primitive $H(y) = -\log(S(y)$ is called the **cumulative hazard function**.

8.2.1 Survivor and hazard functions for the exponential distribution

- $F_\theta(y) = 1-\exp(-\theta\ y)$. The median survival time is $\log(2)/\theta$ and this is a more appropriate description of the average survival time than $\mathbb{E}(y) = 1/\theta$ because of the skewness of the exponential distribution.
- $S_\theta(y) = \exp(-\theta\ y)$
- $h_\theta(y) = \theta$. The hazard function does not depend on y so the probability of failure in the interval $[y; y+\delta y]$ is not related to how long the subject has already survived. This lack of memory property may be a limitation in some cases when probability of failure increases with time.
- $H_\theta(y) = \theta\ y$

8.2.2 Survivor and hazard functions for the Weibull distribution

- $F_{\theta,\lambda}(y) = 1 - \exp\left(-\theta\ y^\lambda\right)$. The median survival time is then $\theta^{-1/\lambda}\ (\log(2))^{1/\lambda}$.
- $S_{\theta,\lambda}(y) = \exp\left(-\theta\ y^\lambda\right)$
- $h_{\theta,\lambda}(y) = \lambda\ \theta\ y^{\lambda-1}$. When $\lambda \neq 1$, the hazard function does depend on y so the probability of failure in the interval $[y; y+\delta y]$ is related to how long the subject has already survived. This allows for **accelerated failure time**.
- $H_{\theta,\lambda}(y) = \theta\ y^\lambda$

8.3 Link function

The expectation of the survival time, $\mathbb{E}[y] = \left(\frac{1}{\theta}\right)^{1/\lambda}\Gamma(1+\frac{1}{\lambda})$, is an element of $\mathbb{R}^+$. Hence a convenient link (invertible) function g to map $\mathbb{R}^+$ to $\mathbb{R}$ where $x^T\beta$ is, is the log function:

$$g\left(\mathbb{E}[y]\right) = \log\left(\mathbb{E}[y]\right) = x^T\beta$$

or

$$\theta \propto \exp\left(x^T\beta\right)$$

8.4 Estimation of β with the Likelihood

The likelihood function can be used as an objective function to maximise to estimate the *best* parameters β.

8.4.1 With uncensored data

Having collected N responses with their explanatory variables values $\{(y_i, x_i)\}_{i=1,\cdots,N}$, the likelihood is:

$$\mathscr{L}(\beta) = \prod_{i=1}^{N} p(y_i|\theta_i, \lambda) \quad \text{contrained by } \theta_i = \exp(x_i^T \beta) \tag{8.1}$$

8.4.2 With censored data

Unfortunatly, in many applications, the values of some responses can be only partially known. For instance, when the study is ending before failure time has been observed or when a subject is not followed up during study period. In this case, the collected responses are censored. For a censored response, the survival time is at least y_i and the probability associated to this response is then $S(y_i)$ (ie probability to have survived behond time y_i). Lets define δ_i an indicator variable that is 0 when the response y_i is censored and 1 if it is not censored. The likelihood can then be written:

$$\mathscr{L}(\boldsymbol{\beta}) = \prod_{i=1}^{N} p(y_i|\theta_i, \lambda)^{\delta_i} \; S(y_i)^{1-\delta_i} \quad \text{with } \theta_i = \exp(x_i^T \beta) \tag{8.2}$$

The data to analyse is then better expressed as $\left\{\left((y_i, \delta_i), x_i\right)\right\}_{i=1,\cdots,N}$ and, in R using `survreg` for fitting the data the expression is written as $(y, \delta) \sim x$ (the indicator variable δ is encapsulated with the response).

8.5 Proportional hazard models

For the Exponential distribution, with x is an indicator variable, the hazard function:

$$h(y) = \theta = \exp(x^T \beta) = \exp(\beta_0 + \beta_1 \; x) \propto \exp(\beta_1 \; x)$$

For the Weibull distribution, the hazard function is:

$$h(y) \propto \lambda\theta y^{\lambda-1} = \lambda y^{\lambda-1} \; \exp(\beta_0 + \beta_1 \; x) = h_0(y) \; \exp(\beta_1 \; x)$$

The ratio

$$\frac{h^{(x=1)}(y)}{h^{(x=0)}(y)} = \frac{\exp(\beta_1 \; 1)}{\exp(\beta_1 \; 0)} = \exp(\beta_1)$$

is called the hazard ratio for presence or absence of exposure to x.

Example of remission time. $x = 0$ indicates that placebo has been given to the patient and $x = 1$ indicates that a treatment (specific drug) has been given to the patient. The hazard ratio $\frac{h^{(x=1)}(y)}{h^{(x=0)}(y)} \simeq 1/4$ indicates that the drugs is helping patients as they are 4 times more likely to 'Fail' (i.e. remission time ending) when they take the placebo.

9. Multinomial Distribution

9.1 From Binomial Distribution to Multinomial Distribution

The Binomial distribution has been defined as the joint distribution of Bernouilli random variables. Bernouilli variables have only two possible outcomes (e.g. success and failure, or yes and no). The Multinomial Distribution defined below extends the number of categories for the outcomes from 2 to J (e.g. for $J=3$: yes, maybe, no).

1.1 Definition (Multinomial Distribution) Consider J categories. Having collected the outcomes of n experiments, y_1 indicates the number of experiments with outcomes in category 1, y_2 indicates the number of experiments with outcomes in category 2, ..., y_J indicates the number of experiments with outcomes in category J. Then the joint density function of $(y_1,\ y_2,\cdots,\ y_J)$ is:

$$p(y_1,y_2,\cdots,y_J;\theta_1,\cdots,\theta_J,n)=\frac{n!}{y_1!\ y_2!\ \cdots y_J!}\ \theta_1^{y_1}\ \theta_2^{y_2}\ \cdots\theta_J^{y_J}$$

with $\theta_1,\cdots,\ \theta_J$ are the respective probabilities of the categories then $\theta_1+\theta_2+\cdots+\theta_J=1$. and $n=y_1+y_2+\cdots+y_J$. The Multinomial Distribution is noted:

$$\text{Multinomial}(n,\theta_1,\cdots,\theta_J)$$

and when $J=2$, we get the Binomial distribution.

Is the multinomial distribution a member of the exponential family of distribution? no, it is not.

Notation conventions. The data collected for analysis are now $\{(y_{i,1},\cdots,y_{i,J},x_i,n_i)\}_{i=1,\cdots,N}$ for N groups (we can use the vectorial notation $\mathbf{y}_i=(y_{i,1},\cdots,y_{i,J})$ leading equivalently to the observations $\{(\mathbf{y}_i,x_i,n_i)\}_{i=1,\cdots,N}$). The likelihood function for the saturated model is then:

$$\mathscr{L}\left(\{\theta_{i,1},\cdots\theta_{i,J}\}_{i=1,\cdots,N}\right)=\prod_{i=1}^{N}\text{Multinomial}(n_i,\theta_{i,1},\cdots,\theta_{i,J})=\prod_{i=1}^{N}\frac{n_i!}{y_{i,1}!\ y_{i,2}!\ \cdots y_{i,J}!}\ \theta_{i,1}^{y_{i,1}}\ \theta_{i,2}^{y_{i,2}}\ \cdots\theta_{i,J}^{y_{i,J}}$$

The degree of freedom for this saturated model is $(J-1)\ N$ (remember that we have the constraints $\theta_{i,1}+\theta_{i,2}+\cdots+\theta_{i,J}=1,\ \forall i$).

9.2 Multinomial Distribution & Poisson random variables

Preliminary exercise: consider two independent random variables that follow a Poisson distribution i.e. $y_1\sim\mathscr{P}_o(\lambda_1)$ and $y_2\sim\mathscr{P}_o(\lambda_2)$. Show that $n=y_1+y_2$ follows a Poisson distribution with parameter $\lambda_1+\lambda_2$ (or $n\sim\mathscr{P}_o(\lambda_1+\lambda_2$).

Multinomial distribution & joint distribution of Poisson variables. Consider $y_1, y_2, \cdots, y_J$ independent random variables with distributions:

$$y_j \sim \mathscr{P}_o(\lambda_j), \ \forall j = 1, \cdots, J$$

Then the joint density function of $(y_1, y_2, \cdots, y_J)$ is:

$$p(y_1, y_2, \cdots, y_J; \lambda_1, \cdots, \lambda_J) = \prod_{j=1}^{J} \frac{\lambda_j^{y_j}}{y_j!} \exp(-\lambda_j) \tag{9.1}$$

Lets define $n = y_1 + y_2 + \cdots + y_J$ then $n \sim \mathscr{P}_o(\lambda_1 + \lambda_2 + \cdots + \lambda_J)$ (see exercise). Using Bayes, and ignoring the parameters of the distributions for now:

$$p(y_1, y_2, \cdots, y_J | n) = \frac{p(n|y_1, \cdots, y_J) \ p(y_1, y_2, \cdots, y_J)}{p(n)} \tag{9.2}$$

with $p(n|y_1, \cdots, y_J) = \delta(n - y_1 - \cdots - y_J)$ (the Kronecker delta) since $n = y_1 + y_2 + \cdots + y_J$. Hence

$$\begin{aligned} p(y_1, y_2, \cdots, y_J | n) &= \frac{\delta(n - y_1 - \cdots - y_J) \ \prod_{j=1}^{J} \frac{\lambda_j^{y_j}}{y_j!} \exp(-\lambda_j)}{\frac{(\sum_{j=1}^{J} \lambda_j)^n}{n!} \exp(-\sum_{j=1}^{J} \lambda_j)} \\ &= \delta(n - y_1 - \cdots - y_J) \ n! \ \prod_{j=1}^{J} \frac{1}{y_j!} \left(\frac{\lambda_j}{\sum_{j=1}^{J} \lambda_j} \right)^{y_j} \end{aligned} \tag{9.3}$$

This is equivalent to the multinomial distribution:

$$p(y_1, y_2, \cdots, y_J | n) = \frac{n!}{y_1! \cdots y_J!} \theta_1^{y_1} \cdots \theta_J^{y_J} \tag{9.4}$$

with the convention $\theta_j = \frac{\lambda_j}{\sum_{j=1}^{J} \lambda_j}$ and the constraint $n = y_1 + \cdots + y_J$ (this is when the kronecker delta is 1, otherwise it is 0). So the Multinomial distribution can be regarded as the joint distribution of Poisson random variables conditionnal upon their sum n. This justifies the use of generalized linear models.

9.3 Nominal logistic regression

3.1 Definition (Nominal logistic Regression) The outcomes of experiments are in J categories and there is no natural order amongst the response categories. One category is arbitrarily chosen as the reference category e.g. θ_1. Then the logits for the other categories are defined by:

$$\text{logit}(\theta_j) = \log\left(\frac{\theta_j}{\theta_1}\right) = x^T \beta_j \quad , \ \forall j = 2, \cdots, J$$

having the constraints $\sum_{j=1}^{J} \theta_j = 1$. When the estimates $\hat{\beta}_j$ are computed, then

$$\begin{cases} \hat{\theta}_j = \hat{\theta}_1 \ \exp\left(x^T \hat{\beta}_j\right) & \forall j = 2, \cdots, J \\ \hat{\theta}_1 = \frac{1}{1 + \sum_{j=2}^{J} \exp\left(x^T \hat{\beta}_j\right)} \end{cases}$$

or

$$\hat{\theta}_j = \frac{\exp\left(x^T \hat{\beta}_j\right)}{1 + \sum_{j=2}^{J} \exp\left(x^T \hat{\beta}_j\right)} \quad \forall j = 2, \cdots, J$$

Having observations $\{(y_{i,1},\cdots,y_{i,J},x_i,n_i)\}_{i=1,\cdots,N}$ for N groups then the estimates of the proportions using the model are:

$$\hat{\theta}_{i,j}=\frac{\exp\left(x_i^T\hat{\beta}_j\right)}{1+\sum_{j=2}^{J}\exp\left(x_i^T\hat{\beta}_j\right)} \quad \forall j=2,\cdots,J$$

Bibliography

[1] Hirotugu Akaike. The interpretation of improper prior distributions as limits of data dependent proper prior distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(1):46–52, 1980.

[2] A. J. Dobson and A. G. Barnett. *An Introduction to Generalized Linear Models*. CRC Press, Third Edition, 2008.

[3] R. H. Myers, D. C. Montgomery, G. G. Vining, and T. J. Robinson. *Generalized Linear Models with Applications in Enginnering and the Sciences*. Wiley, 2nd edition, 2010.

[4] Yudi Pawitan. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford Science Publications, 2001.

[5] M. A. Tanner. *Tools for Statistical Inference- Methods for the exploration of Posterior Distributions and Likelihood functions*. Springer, 3rd Edition, 1996.