GLMMs

GLMs and extensions

L. Egidi - DEAMS, Units (legidi@units.it)

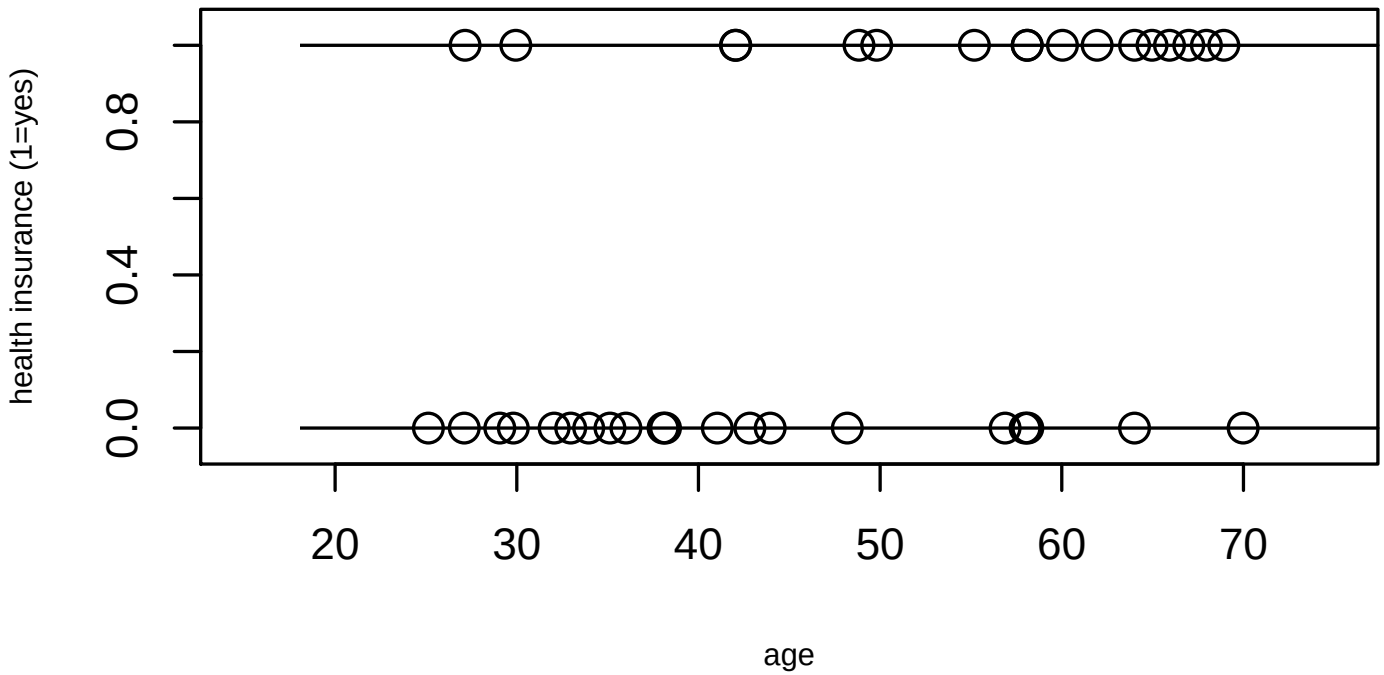
PhD Course in Statistics - XXXVI cycle

Indice

GLM: introduction and basic ideas

- GLMs allow to extend classical normal linear models in many directions:
  - response variables can be assumed non-normal (*including discrete distributions or distributions with support* $[0, \infty)$);
  - The mean and the variance of the response are assumed to vary according to values of observed covariates
  - The impact of covariates on the mean of the response is specified according to a (possibly) *non-linear* function of a linear combination of the covariates
- Main advantages are:
  - Unification of seemingly different models: it makes easy to use, understand and teach the techniques. Many of the standard ways of thinking LM carry over to GLMs;
  - Normal LMs, probit and logit models, log-linear models for contingency tables, Poisson regression, some survival analysis models are GLMs;
  - A single general theory and a single general computational algorithm can be developed for inference.

A first example: Health Insurance coverage



For a sample of 37 individuals we observe the age of any sample unit and whether he/she owns a private health insurance. It seems that older units are more likely to own a health insurance. For these data response variable Y can be assumed Bernoulli

1. $Y_i \sim \text{Bernoulli}(h(x_i))$.
2. and a possibly non linear model can be specified for $h(\cdot) \to [0, 1]$.

A second example: A dose-response analysis

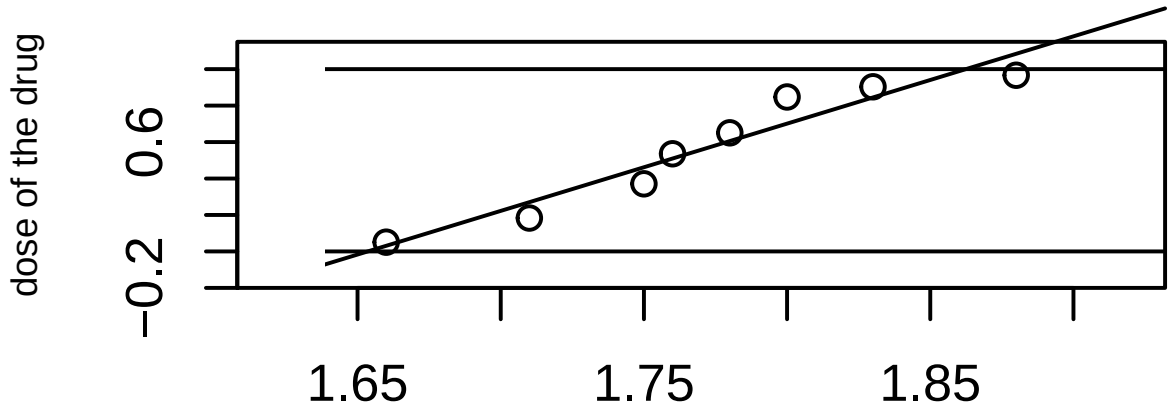
- Consider the data in the table below

dose	1.66	1.74	1.75	1.76	1.78	1.80	1.86	1.88
n. positive	3	9	23	30	46	54	59	58
n. of patients	59	60	62	56	63	59	62	60
proportion	0.051	0.150	0.371	0.536	0.730	0.915	0.951	0.967

- The data refer to 481 individuals who received a drug. For each dose of the drug it has been observed if the individual had a positive response or not.
- Since only 8 different doses have been considered we can obtain the proportion of positive responses for each dose.

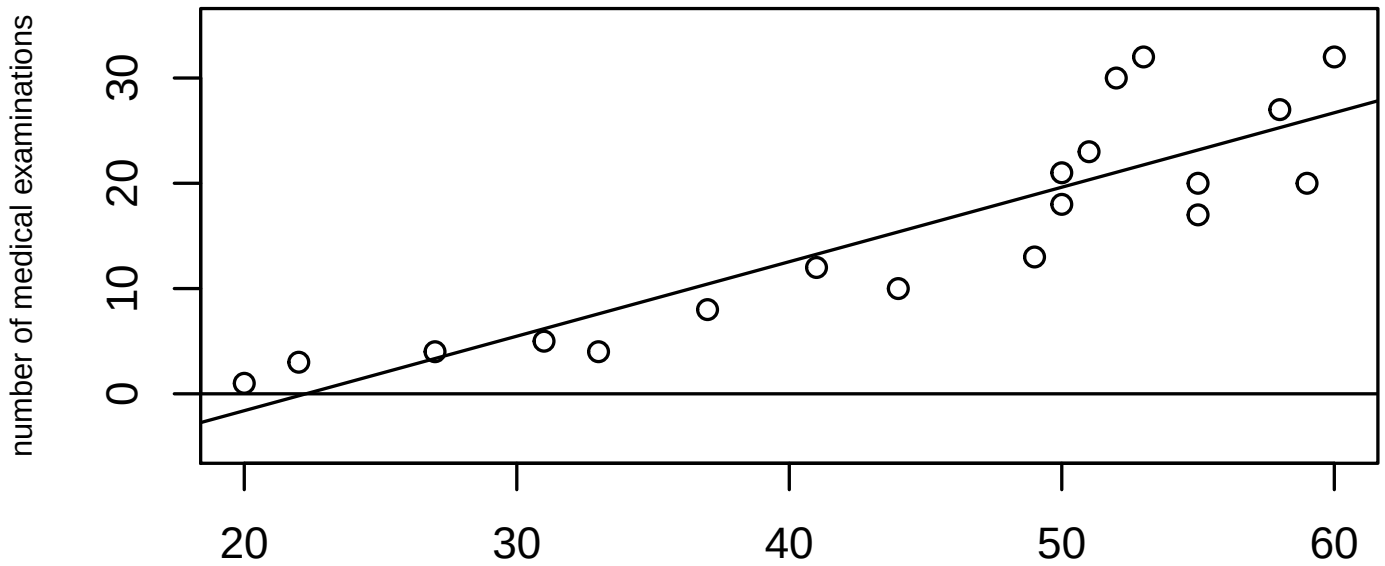
A second example: binomial response

```
clin <- read.table("dati/clintrial.txt", header=T)
prop=clin$num.positive/clin$number
plot(clin$dose, prop, ylim=c(-0.15,1.1),xlim=c(1.62,1.92),
     ylab="dose of the drug", xlab="",  cex.lab=.7)
par(mar=c(3.5,5.5,1,1)); abline(0,0);abline(1,0); abline(lm(prop~clin$dose)
```



- The plot shows that the proportion of positive responses out of m_i on trial increases with the dose of the drug.
- A linear relationship is patently inappropriate. The data are proportions and their values should lie in the [0,1] range
- $Y_i \sim \text{Binomial}(m_i, h(x_i))$. Specify a non linear model for $h(\cdot) \to [0, 1]$.

Count data: an example with medical examinations



- Y_i (number of examinations) can be assumed Poisson: $Y_i \sim \text{Poisson}(\mu_i)$.
- We can assume that $\mu_i = h(x)$, i.e., it is a function of the covariate x.
- A linear specification is clearly inappropriate (also because it will predict negative values). We should choose among functions that $h(\cdot) \to [0, \infty)$.

From LM to GLM

- Recall that Normal LMs, in matrix notation, are defined by

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

1. $Y_i \sim N(\mu_i, \sigma^2)$, independent, where
   $\mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$ and
   $\boldsymbol{x}_i^T$ is $i-$th row of X, $i = 1, 2, \ldots, n$;
2. The density of Y_i may be written as $f_Y(y_i) = f_\epsilon(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})$ where covariates $\boldsymbol{x_i}$ appear through the **linear predictor**:

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j = \boldsymbol{x}_i^T \boldsymbol{\beta};$$

- $\boldsymbol{\beta}$ and σ^2 are unknown parameters.

Introducing GLMs

GLMs generalize LMs by:

- Considering a class of model of the form

$$f_Y(y) = f_\epsilon(y; \boldsymbol{x}_i^T)$$

  and $\boldsymbol{x}_i^T$ still enters into the model through the **linear predictor**.
- errors can enter the model in more general form (not simply additively).
- Existence of the mean $E(Y) = \mu$ is assumed and μ is determined by η that is related to it by a suitable function

$$g(E(Y_i)) = g(\mu_i) = \eta_i = \boldsymbol{x}_i^T \beta$$

  $g(\cdot)$ is called the **link function**.
- in principle f could be any suitable density (or probability) function, but a family of distributions plays a key role:

Y_i are assumed to be (independent) measurements from a distribution with density (probability) function from the **exponential dispersion family**.

The exponential (dispersion) family

- A random variable Y belongs to exponential (dispersion) family if its density (probability) function can be written as

$$f(Y;\theta,\phi) = \exp\left\{\frac{Y\theta - b(\theta)}{\phi} + c(Y,\phi)\right\} \ , \tag{1}$$

  θ and ϕ are unknown scalar parameters,
  $b(\cdot)$ and $c(\cdot) > 0$ are known functions and the domain of Y does not depend on θ or ϕ.

  We will denote this by $Y \sim EF(b(\theta), \phi)$.

- θ is called the *natural or canonical parameter* of the exponential family.
- ϕ is called the *dispersion parameter*. It can be known in some cases. When it is unknown, the family is more properly called the exponential dispersion family.
- Many of the most common continuous and discrete distributions belong to this family (i.e. Normal, Gamma, Poisson, Binomial, etc)

Example: Poisson

- As we already noted it is the basic choice when modelling count data
- if $Y \sim \text{Poisson}(\lambda)$, its probability function is

$$\begin{aligned} f(Y;\lambda) &= \frac{e^{-\lambda}\lambda^{Y}}{Y!} \\ &= \exp\{Y\log\lambda - \lambda - \log Y!\}\ , \end{aligned}$$

  for $Y = 0, 1, \ldots,$.
- This shows that it is a member of (1) where $\theta = \log\lambda$ is the natural parameter, $\phi = 1$, $b(\theta) = \lambda = e^{\theta}$ and $c(Y,\phi) = -\log Y!$.
- We can write $Y \sim EF(e^{\theta}, 1)$.

Binomial

- Standard distribution when modelling binary responses
- If $Y \sim \text{Bin}(n, \pi)$, its probability function is

$$\begin{aligned} f(Y;\pi) &= \binom{n}{Y}\pi^Y(1-\pi)^{n-Y} \\ &= \exp\{\log\binom{n}{Y} + Y\log\pi + (n-Y)\log(1-\pi)\} \\ &= \exp\left\{Y\log\frac{\pi}{1-\pi} + n\log(1-\pi) + \log\binom{n}{Y}\right\}, \end{aligned}$$

  for $Y = 0, 1, \ldots, n$.
- It belongs to (1) where $\theta = \log\frac{\pi}{1-\pi}$ natural parameter, $\phi = 1$,

$$b(\theta) = -n\log(1-\pi)\big|_{\pi=\frac{e^\theta}{1+e^\theta}} = n\log(1+e^\theta)$$

  and $c(Y, \phi) = \log\binom{n}{Y}$.
- $Y \sim EF(n\log(1+e^\theta), 1)$.

Mean and variance for Exponential family

- The function $b(\cdot)$ is called the *cumulant function* and it is important in evaluating and interpreting first moments of the distribution.
- by using identities related to derivatives of log-likelihood function:

$$E\left(\ell_*(\theta)\right) = E\left(\frac{d}{d\theta}\ell(\theta; Y)\right) = 0$$

and

$$i(\theta) = var\left(\ell_*(\theta)\right) = E\left(-\ell_{**}(\theta)\right) = E\left(-\frac{d^2}{d\theta^2}\ell(\theta; Y)\right),$$

under usual regularity assumptions. If Y is a r.v. member of the exponential family, log-likelihood for θ is: it follows that:

$$E\left(\frac{Y - b'(\theta)}{\phi}\right) = 0 \qquad \text{and} \qquad \boxed{E(Y) = \mu = b'(\theta)}$$

$$var\left(\frac{Y - b'(\theta)}{\phi}\right) = \frac{b''(\theta)}{\phi} \qquad \Rightarrow \qquad \boxed{var(Y) = \phi b''(\theta)}$$

Denote $V(\mu) = b''(\theta)$, we can write $\boxed{var(Y) = \phi V(\mu)}$

- The function $V(\mu)$ is the so called *variance function* since it indicates how the variance depends on the mean of Y (GLM can be heteroscedastic).

Main example

Poisson

We have for a Poisson with mean λ

$$b(\theta) = e^{\theta} \quad \text{and} \quad \phi = 1 \qquad \text{and} \qquad E(Y) = b'(\theta) = e^{\theta} = \lambda \ .$$

$$var(Y) = b''(\theta) = e^{\theta} = \lambda \qquad \text{then} \qquad V(\mu) = \mu$$

Binomial

We have for a Binomial with parameters (n, π)

$$b(\theta) = n\log(1+e^{\theta}), \quad \phi = 1 \quad \text{then} \quad E(Y) = \mu = b'(\theta) = n\frac{e^{\theta}}{1+e^{\theta}} = n\pi \ .$$

$$var(Y) = b''(\theta) = n\frac{e^{\theta}}{(1+e^{\theta})^2} = n\pi(1-\pi) \quad \text{and} \quad V(\mu) = \mu(1-\mu)/n \ .$$

The link function

- The second important step in specifying a GLM is the definition of the function relating μ_i and the linear predictor η_i.
- It is assumed that the link between μ_i, the mean of Y_i, and $\boldsymbol{x}_i^T$, the covariate vector, is

$$g(\mu_i) = \eta_i \qquad \text{and} \qquad \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} \ .$$

- $g(\cdot)$ is a known monotone and differentiable function. The function $g(\cdot)$ is the *link function* between μ_i and η_i.
- the inverse function $g(\cdot)^{-1} = r(\cdot)$ is also called the response function
- Covariates enter into the model by the linear predictor $\boldsymbol{x}_i^T \boldsymbol{\beta}$, but the μ_i and η_i are generally non linearly related.
- Appropriate choices of the link function are such that $\mu_i = g^{-1}(\eta_i)$ takes on values on the appropriate range.

The canonical link

- A typical choice is to write directly the natural parameter θ as a linear function of the covariates Formally,

$$\eta = g(\mu) = g(b'(\theta)) = \theta \ ,$$

  $g(\cdot)$ is then the inverse function of $b'(\cdot)$. This choice of the link function is called *canonical link*.
- Some interesting properties derives from choosing a canonical link. Moreover the canonical link is the default link used in many softwares for estimation of GLMs (including R).

GLM: Complete specification

- A parametric model for the response Y_i and a given vector of covariates $\boldsymbol{x}_i$, $i = 1, 2, \ldots, n$.

  A GLM includes the following components:

  1. **Error structure (or response distribution)**: $Y_i \sim EF(b(\theta_i), \phi)$, independent, where $E(Y_i) = \mu_i = b'(\theta_i)$, $i = 1, 2, \ldots, n$;
  2. **linear predictor**: $\eta_i = \boldsymbol{x}_i^T \beta$, $\boldsymbol{x}_i$ is a vector of constants and β a vector of unknown parameters;
  3. **Link function**: It is defined a function $g(\cdot)$ such that $g(\mu_i) = \eta_i \Leftrightarrow \mu_i = g^{-1}(\eta_i)$, $i = 1, 2, \ldots, n$ and then $E(Y_i) = g^{-1}(\boldsymbol{x}_i^T \beta)$.

Some GLMs

Normal Linear Regression model

The standard normal LM is a GLM. In this case

$$Y_i \sim N(\boldsymbol{x}_i^T \beta, \sigma^2) \ , \quad i = 1, 2, \ldots, n$$

$\theta_i = \eta_i = \mu_i$: the canonical link function $g(\cdot)$ is the identity function.
We can equivalently write $Y_i \sim N(\mu_i, \sigma^2)$ or $Y_i = \mu_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$.

Poisson Regression

Let $Y_i \sim \text{Poisson}(\mu_i) \ , \quad i = 1, 2, \ldots, n$, independent. Let's look for a link function $g(\cdot)$ such that $g(\mu_i) = \boldsymbol{x}_i^T \beta$. A good choice could be the log function since $\mu_i = e^{\boldsymbol{x}_i^T \beta}$ is positive.
The choice: $\mu_i = e^{\boldsymbol{x}_i^T \beta} = e^{\eta_i}$
that is $\eta_i = \log \mu_i$ defines the *Poisson regression* and the log link function is also the canonical one since $\theta_i = \eta_i$.

Binomial Regression (binary response)

- In the simplest case we observe a binary response variable Y and we want to study its dependence on a set of covariates x.
- An appropriate model is still a GLM, where an appropriate distributional assumption is binomial. The goal is to study how probability of success varies with x.
- When data are not grouped and y_i is coded by 0 and 1 the behaviour of y_i is completely determined by $\pi_i = \mu_i$. Note that Y is a Bernoulli implying that $E(Y) = \mu = \pi$ and $var(Y) = \pi(1 - \pi)$.
- If data are grouped (i.e. more observation for any value of x) then the number of successes Z_i for a given x is $Z_i \sim \text{Bin}(m_i, \pi_i)$
- In this case Z_i depend also on m_i but we are still interested in modelling success probability and to this aim it is more natural to use as a response the relative frequency of success. These are scaled binomials and: $Y_i = Z_i/m_i$
- It is still true that $E(Y_i) = \mu_i = \pi_i$ and $var(Y) = \frac{\pi(1-\pi)}{m_i}$. (when estimating the model the known weights m_i should be taken into account)

Link functions for binomial regression

- The parameter μ_i must vary within $[0, 1]$.The link function can be chosen among those functions with this property. A natural choice is $\mu_i = \Psi(\eta_i)$ where $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$ and $\Psi(\cdot)$ is a distribution function.
- The resulting link function is $g(\mu) = \Psi^{-1}(\mu)$ Usual choices for $\Psi(\cdot)$:

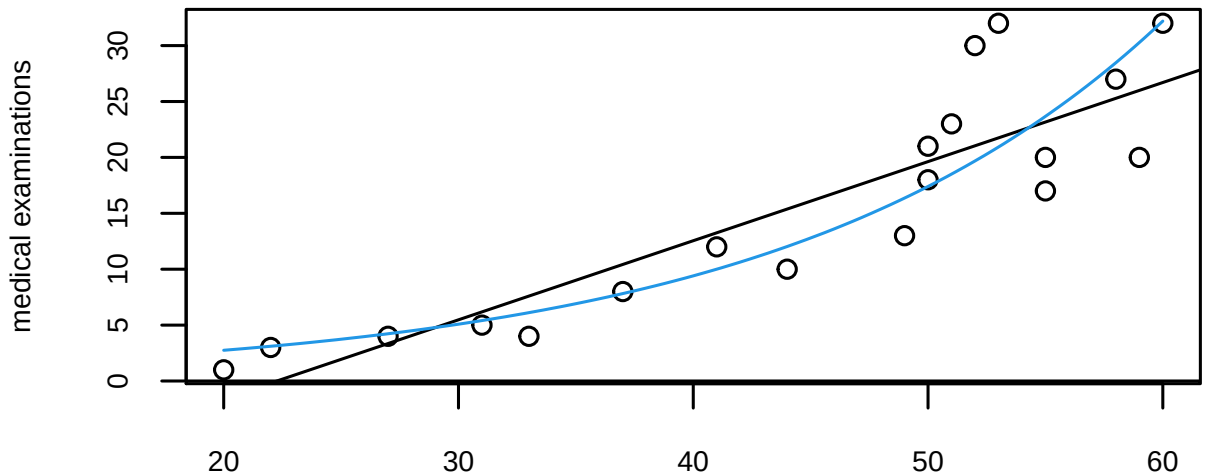
1. $\Psi(\eta) = \Phi(\eta)$, standard normal distribution function. This is the first model proposed for binary response and it is known as **probit regression**.
2. $\Psi(\eta) = \frac{e^\eta}{1+e^\eta}$, and

$$g(\mu) = \Psi^{-1}(\mu) = \log \frac{\mu}{1-\mu} \ .$$

   This gives rise to the well known **logit model** or **logistic regression**.
3. $\Psi(\eta) = 1 - \exp(-e^\eta)$, and $g(\mu) = \Psi^{-1}(\mu) = \log\{-\log(1-\mu)\}$. This link function is the "complementary log-log" (it is related to the distribution function of a type-1 extreme value distribution.)

Poisson regression: Medical examinations

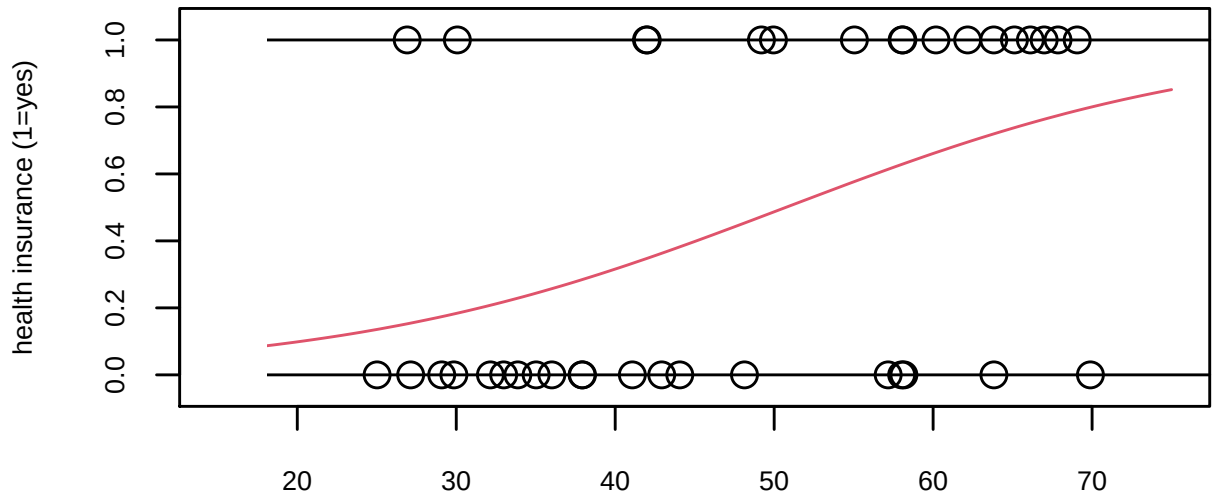
```
etavis=c(20,22,27,31,33,37,41,44,49,51,50,50,53,52,55,58,60,55,59)
numvisite=c(1,3,4,5,4,8,12,10,13,23,21,18,32,30,20,27,32,17,20)
plot(etavis,numvisite,ylab="medical examinations",xlab="",cex.lab=.7,
     cex.axis=.7)
abline(0,0); abline(lm(numvisite~etavis)); modpo=(glm(numvisite~etavis,pois
coepo=modpo$coefficients; curve(exp(coepo[1]+coepo[2]*x),col=4,add=T)
```



- A linear regression would be inappropriate (also because it will predict negative values)
- The blue curve seems to provide a better approximation

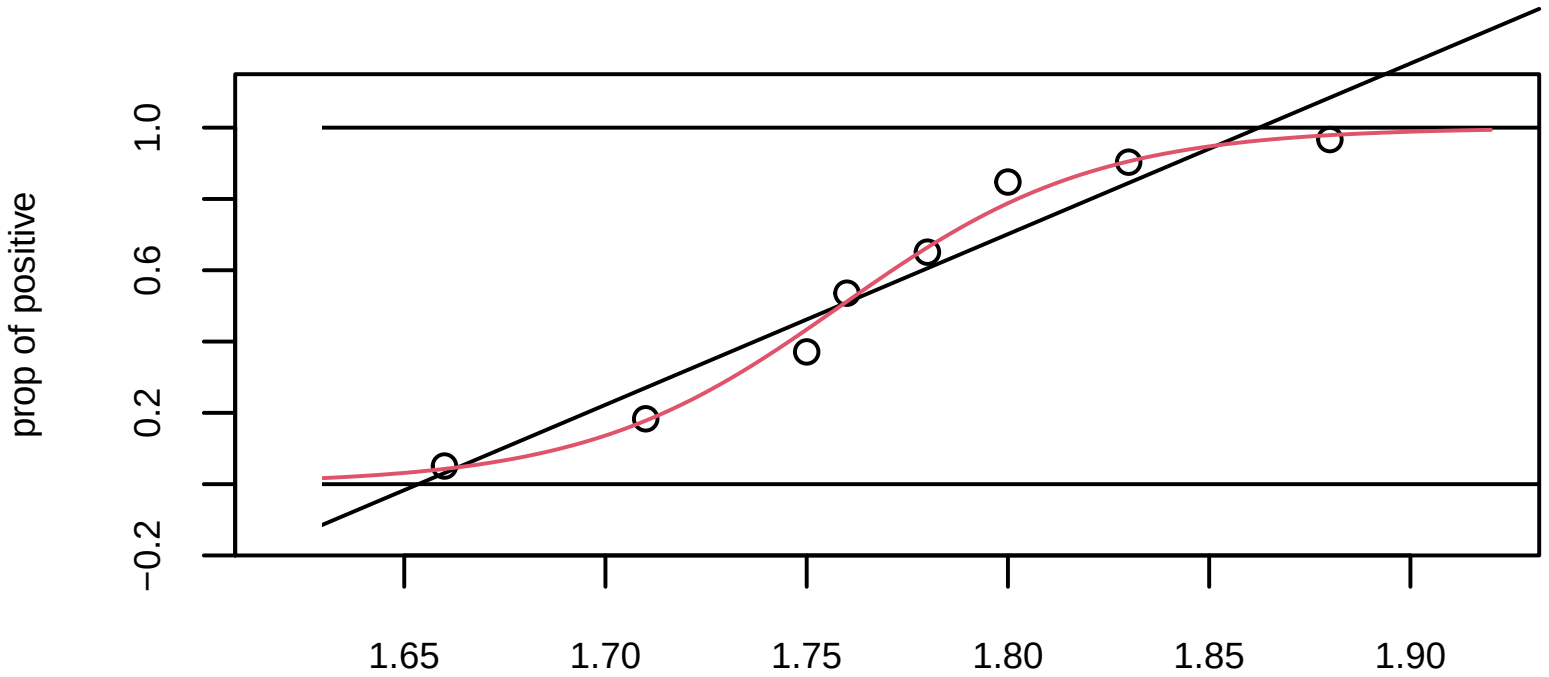
Logistic regression: Health Insurance coverage

```
sanitar <- read.table("dati/sanitar.txt", header=T); attach(sanitar)
plot(jitter(eta),sani,cex=1.5 ,xlim=c(15,75),ylim=c(-0.05,1.05),
     ylab="health insurance (1=yes)",xlab="", cex.lab=.7, cex.axis=.7)
par(mar=c(4,6,1,1)); abline(0,0); abline(1,0)
modall=glm(sani~eta,family=binomial(logit)); coef=modall$coefficients
curve(exp(coef[1]+coef[2]*x)/(1+exp(coef[1]+coef[2]*x)), add=T,col=2)
```



- It seems that older units are more likely to own a health insurance.
- The red curve displays the probability that a unit of a given age has insurance policy

Logistic regression: A dose response model



- The curve represented by logistic regression is by far more appropriate to represent the relationship between the two variables.
- We can select functions that behave similarly to represent this relationship.

Indice

Estimation of the parameters

- ML can be used since distributional assumptions on parameters are available (for the normal LM it coincides with LS).
- A property of the exponential families is that they satisfy enough regularity conditions to ensure that the MLE is given uniquely by the solution of the likelihood equations.
- Let us recall some important features of GLM:
  - $g(\mu_i) = \eta_i = \boldsymbol{x}_i^T \beta \Leftrightarrow \mu_i = g^{-1}(\boldsymbol{x}_i^T \beta)$;
  - $\mu_i = b'(\theta_i) \Leftrightarrow \theta_i = (b')^{-1}(\mu_i) = (b')^{-1}(g^{-1}(\eta_i))$;
  - $var(Y_i) = \phi V(\mu_i)$, with $V(\mu_i) = b''(\theta_i)$.
- Assuming independence of $(y_1, \ldots, y_n)$, the log-likelihood $\ell(\beta, \phi)$ is simply given by

$$\ell(\beta, \phi) = \sum_{i=1}^{n} \log f(y_i; \theta_i, \phi) = \sum_{i=1}^{n} \ell_i(\beta, \phi) = \sum_{i=1}^{n} \left\{ \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right\}$$

where θ_i is a function of β through

$$g(\mu_i) = g(b'(\theta_i)) = \eta_i = \boldsymbol{x}_i^T \beta \ .$$

Likelihood equations

- To obtain the MLE of $\boldsymbol{\beta}$ it is necessary to solve the *likelihood equations*:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} = 0 \quad \text{for} \quad j = 1, 2, \ldots, p.$$

- Let us compute the terms

$$\begin{aligned} \frac{\partial \ell_i}{\partial \beta_j} &= \frac{\partial \ell_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} \\ &= \frac{\partial \ell_i}{\partial \theta_i} \left(\frac{\partial \mu_i}{\partial \theta_i} \right)^{-1} \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^{-1} \frac{\partial \eta_i}{\partial \beta_j}, \end{aligned}$$

Likelihood equations (cont.)

- where

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \theta_i} &= \frac{y_i - b'(\theta_i)}{\phi} = \frac{y_i - \mu_i}{\phi} \ , \\
\frac{\partial \mu_i}{\partial \theta_i} &= b''(\theta_i) = \frac{var(Y_i)}{\phi} \ , \\
\frac{\partial \eta_i}{\partial \mu_i} &= g'(\mu_i) \ , \\
\frac{\partial \eta_i}{\partial \beta_j} &= x_{ij} \ .
\end{aligned}$$

- Thus, we have

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \beta_j} &= \frac{y_i - \mu_i}{\phi} \frac{\phi}{var(Y_i)} \frac{1}{g'(\mu_i)} x_{ij} \\
&= \frac{(y_i - \mu_i) x_{ij}}{\phi V(\mu_i) g'(\mu_i)} \ .
\end{aligned}$$

Likelihood equations (cont.)

- The likelihood equations for β are then

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V(\mu_i)g'(\mu_i)} x_{ij} = 0,$$

  $j = 1, 2, \ldots, p$, where $\mu_i = g^{-1}(\boldsymbol{x}_i^T \beta)$.

- Note that the MLE of β for a fixed value of ϕ, does not depend on ϕ and coincides with the unconstrained MLE.

Canonical link

- The use of the *canonical link* ($\eta_i = g(\mu_i) = g(b'(\theta_i)) = \theta_i$) produces some simplifications in the inference based on the log-likelihood $\ell(\beta, \phi)$.
- With the canonical link, we have $g'(\mu_i) = 1/V(\mu_i)$ and the first derivative reduces to

$$\sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\phi V(\mu_i) g'(\mu_i)} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\phi} \ .$$

- This result implies that the likelihood equations simplify and take the form

$$\sum_{i=1}^{n} y_i x_{ij} = \sum_{i=1}^{n} \mu_i x_{ij} \ .$$

  Using matrix notation, $X^T \boldsymbol{y} = X^T \mu$.
- These equations agree with the general structure of the likelihood equations in exponential families: the observed value of the minimal sufficient statistic is equated to its expectation.
- As regards the existence and uniqueness of the MLE of β, if the link is the canonical one, the theory of exponential families applies.
- In general the likelihood equations for β are nonlinear and must be solved with iterative methods. To this end, the expected Fisher information for β is useful.

Fisher information

- Since β and ϕ are orthogonal, we can proceed as if ϕ were known and we can focus only on β.
- Let us consider the second derivatives of ℓ_i:

$$\begin{aligned} -E\left(\frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k}\right) &= E\left(\frac{\partial \ell_i}{\partial \beta_j}\frac{\partial \ell_i}{\partial \beta_k}\right) \\ &= E\left(\left(\frac{(Y_i-\mu_i)x_{ij}}{\phi V(\mu_i) g'(\mu_i)}\right)\left(\frac{(Y_i-\mu_i)x_{ik}}{\phi V(\mu_i) g'(\mu_i)}\right)\right) \\ &= \frac{x_{ij}x_{ik}}{\phi^2 (V(\mu_i))^2 (g'(\mu_i))^2} E\left((Y_i-\mu_i)^2\right) \\ &= \frac{x_{ij}x_{ik}}{\phi V(\mu_i)(g'(\mu_i))^2}\ , \end{aligned}$$

which gives the (j,k)-element of the Fisher information matrix for β. Using matrix notation,

$$i(\beta) = \frac{X^T W X}{\phi}\ ,$$

Fisher information (cont.)

with $W = \text{diag}(w_1, \ldots, w_n)$ and

$$w_i = \frac{1}{V(\mu_i)(g'(\mu_i))^2} \ ,$$

and X is the matrix of the explanatory variables.

- With the **canonical link**, the observed and the expected informations coincide and have (j, k)-element

  $$\frac{x_{ij} x_{ik} V(\mu_i)}{\phi} \ .$$

  In matrix form,

  $$i(\beta) = j(\beta) = \frac{X^T V X}{\phi} \ ,$$

  with $V = \text{diag}(V(\mu_i))$.
- Asymptotic normality of the MLE gives

  $$\hat{\beta} \ \dot{\sim} \ N_p(\beta, \phi(X^T W X)^{-1}) \ ,$$

  for large n.

Fisher information (cont.)

- Therefore, a consistent estimate of the covariance matrix of β is $i(\hat{\beta}) = \phi(X^T \hat{W} X)^{-1}$, where $\hat{W}$ is the matrix W evaluated at $\hat{\beta}$.
- If ϕ is unknown, it should be replaced by a consistent estimator, such as the MLE or the estimator based on the method of moments.
- For normal distribution with identity link we have $g(\mu) = \mu$, so that $g'(\mu) = 1$. Moreover, $V(\mu) = 1$, $\phi = \sigma^2$ and $\mu_i = \boldsymbol{x}_i^T \beta$. The likelihood equations are $\sum_{i=1}^{n} \frac{(y_i - \boldsymbol{x}_i^T \beta) x_{ij}}{\sigma^2} = 0$ that leads to usual LSE

Some models

Normal Linear Model

We have $g(\mu) = \mu$, so that $g'(\mu) = 1$. Moreover, $V(\mu) = 1$, $\phi = \sigma^2$ and $\mu_i = \boldsymbol{x}_i^T \beta$. The likelihood equations are

$$\sum_{i=1}^{n} \frac{(y_i - \boldsymbol{x}_i^T \beta) x_{ij}}{\sigma^2} = 0 \quad j = 1, 2, \ldots, p. \ .$$

Simplifying σ^2 and using matrix notation, the above equations reduce to the usual LS equations: $X^T(\boldsymbol{y} - X\beta) = 0$ or, equivalently,

$$X^T X\beta = X^T \boldsymbol{y} \quad \text{that leads to} \quad \hat{\beta} = (X^T X)^{-1} X^T \boldsymbol{y}$$

Some models (cont.)

Poisson regression

We have $g(\mu) = \log \mu$, so that $g'(\mu) = 1/\mu$. Moreover, $V(\mu) = \mu$, $\phi = 1$ and $\mu_i = \boldsymbol{x}_i^T \beta$. The likelihood equations are

$$\sum_{i=1}^{n} (y_i - e^{\boldsymbol{x}_i^T \beta}) x_{ij} = 0 \ ,$$

which are generally nonlinear in β. In view of this, an explicit solution does not exist in general.

An iterative algorithm

- Likelihood equations for GLMs do not usually have explicit solutions. They should be solved by iterative methods.
- For the GLM there exists the possibility to use a simple algorithm for the solution of the likelihood equations: the MLEs of the parameter β in the linear predictor can be obtained by iterative weighted least squares.
- Starting with appropriate initial value $\hat{\beta}^{(0)}$ and obtaining a sequence $\hat{\beta}^{(1)}, \hat{\beta}^{(2)}, \ldots,$ using a rule to update $\hat{\beta}^{(t+1)}$ with $\hat{\beta}^{(t)}$, until that the value of

$$||\hat{\beta}^{(t+1)} - \hat{\beta}^{(t)}||$$

is sufficiently small ($< \epsilon$).

Newton-Raphson and Fisher scoring

- Let

$$\ell_* = \left(\frac{\partial \ell}{\partial \beta_1}, \ldots, \frac{\partial \ell}{\partial \beta_p} \right)^T$$

be the *score* vector. We want to solve the equation

$$\ell_* = \ell_*(\boldsymbol{\beta}) = 0 \ .$$

- The Newton-Raphson method is based on the updating rule at the $(t+1)-$th iteration

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + (j(\hat{\boldsymbol{\beta}}^{(t)}))^{-1} \ell_*^{(t)} \ , \tag{2}$$

with $\ell_*^{(t)} = \ell_*(\hat{\boldsymbol{\beta}}^{(t)})$.

- The observed information can be replaced by the expected Fisher information $i(\boldsymbol{\beta})$. This algorithm takes the name of Fisher *scoring* method. This maintains the convergence of the algorithm and simplifies the expressions (if the canonical link function is used, the two expressions coincide).

Developing the algorithm

Expression (2) is equivalent to

$$i(\hat{\beta}^{(t)})\hat{\beta}^{(t+1)} = i(\hat{\beta}^{(t)})\hat{\beta}^{(t)} + \ell_*^{(t)} \ .$$

Remember that the (j,k)-th element of $i(\beta)$ is

$$\sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \ ,$$

which gives $i(\beta) = \frac{X^T W X}{\phi}$, with $w_{ii} = \frac{1}{V(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$.

In view of this, the right hand term can be written as

$$(i^{(t)})\hat{\beta}^{(t)} + \ell_*^{(t)}$$

$$= \sum_{k=1}^{p}\sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \hat{\beta}_k^{(t)} + \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)$$

$$= X^T W^{(t)} \boldsymbol{s}^{(t)} \ ,$$

Weighted Least Squares

- where $\boldsymbol{s}^{(t)}$ is a vector with elements

$$s_i^{(t)} = \sum_{k=1}^{p} x_{ik}\hat{\beta}_k^{(t)} + (y_i - \mu_i)\left(\frac{\partial \mu_i}{\partial \eta_i}\right) \ ,$$

  and all the involved quantities are evaluated at $\hat{\boldsymbol{\beta}}$.
- Therefore, it is possible to arrive to the expression

$$X^T W^{(t)} X \hat{\boldsymbol{\beta}}^{(t+1)} = X^T W^{(t)} \boldsymbol{s}^{(t)} \ . \tag{3}$$

- Clearly, the parameter ϕ simplifies.
- The above expression has the form of the normal equations for a LM obtained with weighted least squares, except that the equation above has to be solved iteratively because in general $\boldsymbol{s}$ and W depend on $\boldsymbol{\beta}$.

Iterative Weigthed Least Squares (IWLS)

- Indeed, the Newton-Raphson iteration is

$$\hat{\beta}^{(t+1)} = (X^T W^{(t)} X)^{-1} X^T W^{(t)} \boldsymbol{s}^{(t)} \ . \tag{4}$$

- Each iteration of the algorithm is equivalent to a weighted least squares estimate, in which the adjusted dependent variable and the weights depend on the fitted values, for which only current estimates are available.
- The algorithm has two main steps:
  1. Given $\hat{\beta}^{(t)}$, compute $\boldsymbol{s}^{(t)}$ and $W^{(t)}$;
  2. Obtain $\hat{\beta}^{(t+1)}$ through (4).

To start the algorithm a simple and convenient choice of the starting values is $\boldsymbol{s}^{(0)} = g(Y_i)$ and $W^{(0)}$ equals to the identity matrix.

Estimating the dispersion parameter ϕ

- For the LM, the estimation of β is independent from the value of the variance σ^2. A similar situation holds for the dispersion parameter ϕ in GLMs.
- Obviously, the MLE of ϕ, with β replaced by $\hat{\beta}$, could be used.
- Also estimators based on the method of moments are often used for ϕ.
- Since $var(Y_i) = \phi V(\mu_i)$ or, equivalently, since $\frac{E((Y_i-\mu_i)^2)}{V(\mu_i)} = \phi$ if β is known, an unbiased estimator of ϕ is

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{V(\mu_i)} \ .$$

If the expected values μ_i are replaced with their estimates based on $\hat{\beta}$, then the following adjusted consistent estimator is obtained

$$\widehat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i-\widehat{\mu}_i)^2}{V(\widehat{\mu}_i)},$$

where

$$\widehat{\mu}_i = g^{-1}(\boldsymbol{x}_i^T\widehat{\beta}) \ .$$

Exploiting asymptotic normality of $\hat{\beta}$

- For n large, the asymptotic distribution of the MLE is

$$\hat{\beta} \dot{\sim} N_p(\beta, [i(\hat{\beta})]^{-1}) \quad \text{where} \quad i(\hat{\beta}) = \frac{X^T \hat{W} X}{\phi}$$

  with $\hat{W}$ computed at $\hat{\beta}$. The estimated asymptotic variances are the diagonal elements of the matrix $(X^T \hat{W} X)^{-1}\phi$.
- Using the asymptotic distribution of $\hat{\beta}$, a confidence interval for β_j with approximate level $1 - \alpha$ is

$$\hat{\beta}_j \pm z_{1-\alpha/2} \sqrt{\phi [(X^T \hat{W} X)^{-1}]_{j,j}} \ .$$

- and the statistic $\dfrac{\hat{\beta}_j}{\sqrt{\phi [(X^T \hat{W} X)^{-1}]_{j,j}}}$ can be used to test significance of a single β_j

Indice

Comparing nested models

- Let us start by considering two nested GLMs. Let denote the models by M_C and M_R, such that $M_R \subset M_C$. Specifically, the current model M_C contains p parameters and the reduced model M_R contains p_0 parameters, where $p > p_0$.
- Consider the following partition of $\boldsymbol{\beta} = (\boldsymbol{\beta}_{MR}, \boldsymbol{\beta}_{MC})$, where $\boldsymbol{\beta}_{MR} = (\beta_1, \ldots, \beta_{p_0})$ and $\boldsymbol{\beta}_{MC} = (\beta_{p_0+1}, \ldots, \beta_p)$. Suppose we want to test the following hypothesis

$$H_0 : \boldsymbol{\beta}_{MC} = 0 \quad \text{against} \quad H_1 : \boldsymbol{\beta}_{MC} \neq 0 \ .$$

- The criterion we will adopt to compare M_C and M_R is the likelihood ratio

$$W = 2\{\ell(\hat{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}_{MR})\} \ .$$

The deviance in LMs

- In normal LMs, with σ^2 known, the likelihood ratio is a function of the deviance (sum of square of residuals) $D = SSE = \sum_i (y_i - \hat{\mu}_i)^2$ of the two models. When comparing two nested models ($M_R \subset M_C$), the likelihood ratio criterion will lead to rejection of H_0 for large values of the following statistic

$$W = 2\{\ell(\hat{\beta}) - \ell(\hat{\beta}_{MR})\} = \frac{D_{MR} - D}{\sigma^2} \ ,$$

  where $D_{MR} = SSE_{H_0}$ and $D = SSE$ are sums of square of residuals in the reduced and current models respectively.
- When H_0 holds this statistic has a $\chi^2_{p-p_0}$ distribution.

LR test

- Like Normal LMs, we look for an interpretation of (log-)likelihood ratio in GLMs so that the relationship between the two classes of models is clear. It will help if we can define an analogous quantity as deviance in LMs.
- Log-likelihood for a GLM is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}) \ ,$$

where

$$\ell_i(\boldsymbol{\beta}) = \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \ .$$

- With nested GLM, the statistic

$$W = 2\{\ell(\hat{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}_{MR})\}$$

is asymptotically distributed as a $\chi^2_{p-p_0}$ when H_0 holds.

The saturated model

- Analogy with Normal LM can be kept by introducing likelihood associated to a model where there are as many parameters as observations. This model will be denoted as **saturated** or **full**.
- At the other extreme there is a model as simple as possible, *i.e.*, a model where a single parameter represents a common μ for all the y_i

A "'good"' model usually stands between these two extremes since a saturated model is uniformative being unable to summarize data: it just repeats them in full, and a null model is usually too simple to be useful. We should seek a balance between conflicting goals of parsimony and goodness of fit.

The saturated model (cont.)

- Saturated model is defined as:
  - $\mapsto$ a GLM having the same distribution and link function of the current model;
  - $\mapsto$ but a number of parameter equal to n (or to the number of different groups sharing the same $\boldsymbol{x}$ vector).
- We can evaluate likelihood function for the saturated model and the current model at the value of the MLE obtained in both cases ($\tilde{\theta}$ and $\hat{\theta}$ respectively). If the current model fits the data, $\ell(\tilde{\theta})$ should be very similar to $\ell(\hat{\theta})$. In case of a poor fit then $\ell(\hat{\theta})$ should be much smaller than $\ell(\tilde{\theta})$.

The deviance in GLMs

- Formally, the quantity

$$D(y;\hat{\theta}) \quad = \quad 2\phi\{\ell(\tilde{\theta}) - \ell(\hat{\theta})\} = \phi \sum_{i=1}^{n} D_i$$

with $D_i = 2\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\}$, is called *deviance function* of the model and

$$\frac{D(y;\hat{\theta})}{\phi} = \sum_{i=1}^{n} D_i \tag{5}$$

is the *scaled deviance*: note that it is always non negative.

The deviance in GLMs (cont.)

This quantity is small for good models and is large when the current model gives a poor fit. Behaviour of deviance is equivalent to that of SSE in LMs.

- $\ell(\tilde{\theta})$ is the log-likelihood obtained by letting $\mu_i = b'(\theta_i) = y_i (\Leftrightarrow (\partial \ell_i / \partial \theta_i) = 0)$, so the saturated model has $p = n$ parameters.
- The saturated model is useless but $\ell(\tilde{\theta})$ provides a benchmark to compare log-likelihood of the current model.

Example: Normal regression model

Since Normal LMs are GLMs with identity link functions we can show that calculating the above defined deviance we give in this case the same result obtained by standard theory for goodness of fit evaluation in Normal LMs.

- $Y_i \sim N(\mu_i, \sigma^2)$, $b'(\theta) = \frac{\theta^2}{2}$, $\theta = \mu = b'(\theta)$ and $\phi = \sigma^2$.
- $\ell(\theta) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2$
- For the saturated model $\tilde{\mu}_i = y_i$, and

$$\ell(\tilde{\theta}) = -\frac{n}{2}\log\sigma^2 \ .$$

- For the current model $\hat{\mu}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$, and

$$\ell(\hat{\theta}) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$$

- Scaled deviance is

$$D(y;\hat{\theta}) = \sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\sigma^2}$$

the same expression for SSE of the current model, divided by σ^2.

Poisson

- $Y_i \sim Poisson(\mu_i), b(\theta_i) = e^{\mu_i} = b'(\theta_i), \phi = 1, \log \mu_i = \boldsymbol{x}_i^T \beta$
- $\ell(\theta) = \sum_{i=1}^n y_i \log \mu_i - \sum_{i=1}^n \mu_i$
- For the saturated model $\tilde{\mu}_i = y_i$, and

$$\ell(\tilde{\theta}) = \sum_{i=1}^n y_i \log y_i - \sum_{i=1}^n y_i \ .$$

- For the current model $\log \hat{\mu}_i = \boldsymbol{x}_i^T \hat{\beta}$, and

$$\ell(\hat{\theta}) = \sum_{i=1}^n y_i \log \hat{\mu}_i - \sum_{i=1}^n \hat{\mu}_i \ .$$

- So deviance is $D(y; \hat{\theta}) = 2 \left(\sum_{i=1}^n y_i \log \frac{y_i}{\hat{\mu}_i} - \sum_{i=1}^n y_i + \sum_{i=1}^n \hat{\mu}_i \right)$

Binomial

- $Y_i \sim Bin(1, \pi_i)$, con $\pi_i = Pr(Y_i = 1) = E(Y_i) = \mu_i$
- $\ell(\theta) = \sum_{i=1}^{n} (y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i))$
- For the saturated model $\tilde{\mu}_i = y_i$ and

$$\ell(\tilde{\theta}) = \sum_{i=1}^{n} (y_i \log y_i + (1 - y_i) \log(1 - y_i)) \ .$$

- For the current model $logit(\hat{\mu}_i) = \boldsymbol{x}_i^T \hat{\beta}$ and

$$\ell(\hat{\theta}) = \sum_{i=1}^{n} (y_i \log \hat{\pi}_i + (1 - y_i) \log(1 - \hat{\pi}_i)) \ .$$

- The deviance is

$$D(y; \hat{\theta}) \quad = \quad 2 \sum_{i=1}^{n} \left(y_i \log \frac{y_i}{\hat{\pi}_i} + (1 - y_i) \log \frac{1 - y_i}{1 - \hat{\pi}_i} \right) \ .$$

Comparing nested models

- Considering two nested models M_C and M_R, likelihood ratio test is

$$\begin{aligned} W &= 2\left\{\ell(\hat{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}_{MR})\right\} \\ &= \frac{D(Y, \hat{\theta}_{MR}) - D(Y, \hat{\theta})}{\phi}, \end{aligned}$$

  as $n \to \infty$ it is distributed $\chi^2_{p-p_0}$ when H_0 holds.
- So to test if reduced model can be accepted we can compare

$$W = \frac{D(Y, \hat{\theta}_{MR}) - D(Y, \hat{\theta})}{\phi}$$

  with the quantiles of the distribution $\chi^2_{p-p_0}$. We reject H_0 for large values of the statistic (or for a small *p-value*).

Residual Deviance

- It is important to note that since also deviance is defined as a function of the difference arising from a log-likelihood ratio of two nested model one is tempted to use the same criteria for evaluating if deviance of the current model is significantly small. One can look if value of deviance is not large enough when compared to a χ^2_{n-p}.
- In this last case standard asymptotic theory could not work when the number of parameter in the saturated model is not fixed as n goes to infinity.

Nonetheless the criterion could work when the number of parameters is fixed: this is, for instance, the case of a binomial model for grouped data or a Poisson model with factors as the only covariates (as it happens in log linear model from contingency tables).

- In some cases (the most notable being binomial and Poisson) the dispersion parameter is fixed to 1.
- When dispersion parameter ϕ is not known another consistent estimate of it must be considered

$$\widehat{\phi} = \frac{D(Y, \hat{\theta})}{(n-p)}$$

and under mild conditions the result stated above still works.

Model selection

- Model selection strategies can exploit the tools defined above to explore which combination of explanatory variables leads to a satisfactory model.
- So one can consider a stepwise backward search by starting with a model that includes all the covariates and then consider a set of reduced sub models obtained by removing certain variables (backward selection). In order to choose among models, one can consider the sub-model obtained by deleting variables with a large p-value.
- A forward search starts from the null model (usually the one including only the intercept) and (groups of) variables are included if the p-values associated are small.
- A combination of the two strategies can also be considered.
- To compare models also the well known criteria AIC and BIC can be used. For instance, in this case $AIC = -2\ell(\hat{\theta}) + 2p$ where p is the number of parameters of the model (when dispersion parameter is known) and one chooses the model where AIC is smaller.

Residuals in GLM

- Let us recall the basic ideas in using residual analysis in LMs:
  - residuals are easily defined as the difference between the observed datum and the estimated systematic part of the model: this step is less natural in GLM.
  - residuals tell us if there are symptoms of systematic differences between observed and fitted values (i.e. plot of residuals against fitted values, or against covariates)
  - residuals help us recognizing discrepancies between few data and the rest (ouliers detection, evaluation of leverage: hat matrix, case deletion measures -Cook's distance-, jacknife residuals, etc.)

Residuals in GLM (cont.)

- Some of these ideas can be generalized in GLMs.
- A straight extension of the concept of standardized residual is given by

$$r_{Pi} = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\phi} V(\hat{\mu}_i)}} \,, \tag{6}$$

called *Pearson residuals*. The definition (6) resembles that for residuals in LMs based on the estimation of the error term ϵ_i.

Deviance residuals

- Recall that in GLMs ϵ_i does not exist in general, so we can measure the contribution of each observation to deviance. This is analogous to LMs where SSE is defined as

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})^2 \ ,$$

  while in GLMs a similar quantity is the deviance. Recall that deviance is defined as

$$D(y, \hat{\theta}) = \sum_{i=1}^{n} D_i \ .$$

  Large individual contributions to total deviance D_i reflect data that are not properly reproduced by the model. Let us define

$$r_{Di} = \text{sgn}(y_i - \hat{\mu}_i)\sqrt{D_i} \ ,$$

  that is called *deviance residual* of the model.

For large n it is possible to show that $r_{Pi} \approx r_{Di}$.

Residual analysis

- Actually if the model is valid, residuals of any type, possibly scaled by $\hat{\phi}$, will have a distribution that can be (loosely) approximated by a $N(0,1)$. This suggest to use standard graphical tools, like
  - $\mapsto$ normal probability plot of the residuals;
  - $\mapsto$ plot of residuals against the fitted values $\hat{Y}_i$;
  - $\mapsto$ plot of residuals against explanatory variables

  to check assumptions.
- It is also possible to generalize the Hat matrix H to check influence and leverage of residuals. Recall that H in LMs is such that $\hat{\boldsymbol{y}} = \mathbf{H}\boldsymbol{y}$ and $H = X(X^TX)^{-1}X^T$.
- Generalized hat matrix is similarly obtained as $H = W^{\frac{1}{2}}X(X^TWX)^{-1}X^TW^{\frac{1}{2}}$ where W is substituted by $\hat{W}$.
- A generalization of the Cook's distances is also possible.

Indice

Offset in Poisson regression/modelling rates

There are many cases where the observed counts should be interpreted as relative to some baseline.

- The parameter λ of a Poisson regression model can be interpreted with reference to a specific unit of time or space. And the number of cases y_i in a process is then Poisson$(e_i\lambda)$. λ is the rate of the process and e_i is the exposure.
- Suppose we want to model the number of those with a specific disease within a geographical area: this clearly depends on the rate and on the number of units living in that area.
- One could then model the rate λ_i/e_i. In this case

$$\frac{\lambda_i}{e_i} = exp(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}) \rightarrow \log\left(\frac{\lambda_i}{e_i}\right) = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} \rightarrow \log(\lambda_i) = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} + \log(e_i)$$

- Then y_i is again modelled as in Poisson regression by specifying its mean as λ_i=exp$(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})$ but also $\log(e_i)$ is introduced into the model.
- $\log(e_i)$ is included in the model as a *deterministic* predictor whose coefficient is fixed to 1 and it is called the **offset** (in `R` is introduced in `glm` with the option `offset=log(....)`).

Overdispersed count data

- Poisson regression models in a GLM context imply that the dispersion parameter is fixed to 1.
- Variance function is then functionally related to the mean function (it is actually the same).
- For a Poisson model the standardized residuals are

$$z_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

  where $\hat{\mu}_i = \exp(x_i^T \hat{\beta})$
- If the Poisson model holds the z_i are approximately independent and will have mean equal to 0 and variance equal to 1. Approximately $\sum_{i=1}^{n} z_i^2$ is a χ^2_{n-k} distribution if the model holds. This can be used for detecting overdispertion.
- Considering a model for counts with overdispersion will be more realistic in many practical cases.

Logistic regression

Binomial regression with logit link is by far the most popular.
Some reasons are:

1. it is the canonical link
2. it can be interpreted as log-odds of probability of success
3. statistical analysis is simplified
4. appropriate for data collected in a retrospective study or when oversampling one of the classes.

The last property deserves some more words since it could be relevant when this model is used for classification (prediction).

It states that if we oversample one of the two classes (as typically done in retrospective studies) the estimates of the β_j $(j = 1, 2, \ldots, p-1)$ parameters are unchanged with the exception of the intercept β_0.

This can bias, but in a predictable direction, the estimated probability of success.

Overdispersion in binomial regression

- Also in the case of Binomial regression the mean $\mu = np$ completely defines the variance function $v(\mu) = np(1-p)$.
- Also in this context data can appear to have more variance than expected under binomial variation.
- Again, one can use standardized residuals to reveal this.
- The simplest and more common mechanism that gives rise to overdispersion is clustering in the population. In the case of a binomial response assume data are clustered and that cluster size k is fixed. Since we have m individuals in the sample, there are m/k clusters. Now if we assume that in each cluster the number of successes Z_i follows a $Bi(k, \pi_i)$ which varies across clusters, the response $Y = Z_1 + Z_2 + \cdots + Z_{m/k}$.

- Note that overdispersion cannot arise in case of a Bernoulli model. Then this issue should be taken into account only with grouped data.

Quasi-likelihood

- For LMs the method of LS allows to obtain estimates of the regression parameters without the specification of a probabilistic model.
- The method of LS requires only the specification of the relation between the expected value of the response variable and the linear predictor, and the specification of the variance of the error term, which is not related to the expected value:

$$E(Y_i) = \mu_i = \eta_i \quad var(Y_i) = \sigma^2$$

- Also for the GLMs it is possible to specify only these two relations (assuming that the variance function $V(\mu_i)$ is known).
- Indeed, the likelihood equation for β

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i) g'(\mu_i)} x_{ij} = 0 \ , \quad j = 1, \ldots, p \ ,$$

is an unbiased estimating equation provided that $E(Y_i) = \mu_i = g^{-1}(\eta_i)$.

Quasi-likelihood (cont.)

- In other words, this means that the parametric assumption $Y_i \sim EF(\cdot, \phi)$ could not even be satisfied. Only the assumption about expectations is essential: $\mu_i = E(Y_i) = g^{-1}(\eta_i)$
- The only distributional feature that must be known in order to calculate the estimating equation is the variance function $V(\mu)$.

- Under suitable regularity conditions, the likelihood equations for a GLM give estimates for the coefficients β which maintain several properties, also if the parametric assumptions of Y_i are substituted with weaker **second order assumptions**:

1. $g(\mu_i) = g(E(Y_i)) = \eta_i, \qquad i = 1, \ldots, n$
2. $var(Y_i) = \phi V(\mu_i), \qquad i = 1, \ldots, n$
3. $cov(Y_i, Y_j) = 0$, if $i \neq j$.

Quasi-likelihood (cont.)

- The semi-parametric statistical model specified by assumptions 1–3 is called **quasi-likelihood model**.
- If $V(\mu) = 1$ and $g(\mu) = \mu$, the assumptions 1–3 match the usual second order assumptions of the classical LM.
- On the other hand, if $V(\mu) = \mu^2$ we obtain a multiplicative model, $Y_i = \mu_i \epsilon_i$, with $E(\epsilon_i) = 1$ and $var(\epsilon_i) = \phi$.
- Gauss-Markov (BLUE) optimality of LS extends to quasi-likelihood estimates and it has minimum asymptotic variance among estimating equations that are linear (in Y) and unbiased

- Indeed, the likelihood equation for β

$$q(y;\beta) = \sum_{i=1}^{n} q(y_i;\beta) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)g'(\mu_i)} x_{ij} = 0 \ , \quad j = 1, \ldots, p \ ,$$

  behaves like a score vector. Specifically:

$$E(q(Y;\beta)) = 0, \quad \text{and} \quad var(q(Y;\beta)) = -E(\partial q(Y;\beta)/\partial\beta) \ .$$

Quasi-likelihood (cont.)

- Quasi likelihood estimators shares many properties of a proper likelihood: the quasi-MLE β is asymptotically normal, the quasi-likelihood ratio statistic has a null chi-squared distribution.

Quasi-likelihood and overdispersion

- The assumptions 1–3 offer an increase in flexibility with respect to the usual parametric specifications based, respectively, on the Poisson, binomial or exponential distributions.
- In practice, there are situations in which the dispersion parameter does not agree with the assumed exponential family.
- For example, for the binomial or Poisson distributions we have $\phi = 1$, but data could show agreement with $\phi > 1$.
- In this case we have *overdispersion*, i.e. the variance of Y is greater than its theoretical value, and it is more plausible to assume $var(Y_i) = \phi V(\mu_i)$, with $\phi > 1$. For example, for proportions, it can be assumed that $var(Y) = \phi n\pi(1-\pi) > n\pi(1-\pi)$, with $\phi > 1$, where $n\pi(1-\pi)$ is the variance of a binomial distribution.
- In general, the quasi-likelihood approach allows to deal with *overdispersion problems*: it is possible to specify $var(Y_i)$ so that there is more variability with respect to the exponential family.
- The case of *underdispersion*, i.e. $\phi < 1$, is less important in applications, but can be dealt with by the QL model as well.

Using quasi-likelihood in `glm`

- When estimating a GLM by using quasi-likelihood one can use the same variance function derived from a Binomial or from a Poisson model and using the canonical link for those models. In R this leads to a specification of the `family` that is called `quasibinomial` or `quasipoisson`.
- Estimates of the β are the same since the estimating equations do not change
- But standard errors of estimates will change since a value different from 1 is estimated for ϕ. In `quasipoisson` one should take into account that variance is modelled as $Var(y_i) = \phi\mu_i$
- The parameter ϕ can be also estimated as

$$\hat{\phi} = \frac{1}{n-p}\sum \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

- In those cases also the Deviance of the model has to be corrected because it is computed assuming $\phi = 1$. The deviance reported has to be divided by $\hat{\phi}$
- Also the standardized residuals are different. E.g., for the Poisson:

$$z_i^{QL} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi}\hat{\mu}_i}} \quad vs \quad z_i^{GLM} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

Indice

Negative binomial regression

- It is an alternative model that can be considered when data exhibits overdispersion. Its probability function is

$$Pr(Z = z) = \binom{z + k - 1}{z} p^k (1 - p)^z \qquad z = 0, 1, \ldots$$

where $E(Z) = k(1 - p)/p$ and $Var(Z) = k(1 - p)/p^2$.

- Interpretation: probability to observe z *failures* until the pre-specified number of *successes* k is observed.
- Compared with Poisson
  - since it has an extra parameter it proves to be more flexible
  - mean is larger than variance and then it accommodates overdispersion
  - Poisson is a limiting case of negative binomial (if $p \to 1$ and $k \to 0$ then $kp \to \lambda$)
- Recall that negative binomial emerges as a mixture of Poisson when each unit Y is Poisson with mean λ and λ are drown from a *Gamma* distribution.

Negative Binomial regression

- When building a model for Negative Binomial a different parametrization is more appropriate, by defining $Y = Z - k$ and $p = \frac{1}{1+\alpha}$
- $$Pr(Y = y) = \binom{y + k - 1}{k - 1} \frac{\alpha^y}{(1 + \alpha)^{y+k}} \qquad y = 0, 1, \ldots$$
- Then
  - $E(Y) = \mu = k\alpha$
  - $Var(Y) = k\alpha + k\alpha^2 = \mu + \mu^2/k$
- and the following link can be used $\log \frac{\alpha}{1+\alpha} = \log \frac{\mu}{k+\mu}$
- Note that BiN is not a member of the EF. Then it is not a proper GLM. Classical IWLS cannot be used as it is.
- In `R` a specific function has to be used: `glm.nb(...)` included in the package MASS

Zero inflated Poisson

- Zero inflation means that we have far more zeros than what would be expected for a Poisson or BiN distribution
- Ignoring zero inflation can have two consequences:
  - the estimated parameters and standard errors may be biased
  - the excessive number of zeros can cause overdispersion
- A possible model hypotizes that the observed counts derive from a mixture of two populations:
  - for a part of the population (with probability p) Y can only be 0
  - for the remaining part (with probability $1-p$) Y is distributed as a Poisson or a BiN.
- Distribution of counts is then, in case of Poisson

$$P(y_i = 0) = p_i + (1-p_i)e^{-\mu_i}$$

$$P(y_i = y_i | y_i > 0) = (1-p_i)\frac{\mu_i^{y_i} e^{\mu_i}}{y_i!}$$

- Covariates can be introduced, like in GLM, for modelling p_i and μ_i

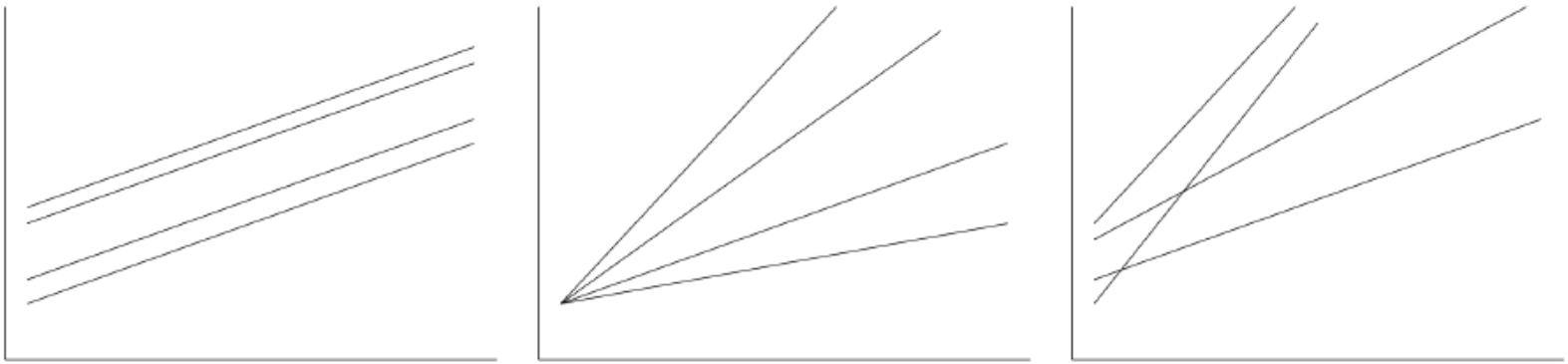
Indice

Multilevel structures

- **Hierarchical/Multilevel** models are extensions of regression in which data are structured in groups and coefficients can vary by group.
- Example of multilevel structures:
  - Simple grouped data—persons within cities—where some information is available on persons and some information is at the city level.
  - Repeated measurements.
  - Time-series cross sections.
  - Non-nested structures.

Varying-intercept and varying-slope models

- With grouped data, a regression that includes indicators for groups is called a *varying-intercept model* because it can be interpreted as a model with a different intercept within each group



(a) Varying intercept (b) Varying slope (c) Varying intercept and slope

Varying-intercept and varying-slope models (cont.)

- Model with one continuous predictor x and indicators for $J = 5$ groups. The model can be written as a regression with 6 predictors or, equivalently, as a regression with two predictors (x and the constant term), with the intercept varying by group (left figure panel):

$$y_i = \alpha_{j(i)} + \beta x_i + \epsilon_i, \quad \text{varying-intercept.}$$

- Another option (central panel) is to let the slope vary with constant intercept:

$$y_i = \alpha + \beta_{j(i)} x_i + \epsilon_i, \quad \text{varying-slope.}$$

Varying-intercept and varying-slope models (cont.)

- Finally, the right panel shows a model in which both the intercept and the slope vary by group:

$$y_i = \alpha_{j(i)} + \beta_{j(i)} x_i + \epsilon_i, \qquad \text{varying-intercept and slope.}$$

  The varying slopes are interactions between the continuous predictor x and the group indicators.
- It can be challenging to estimate all these α_j's and β_j's, especially when inputs are available at the group level.

Clustered data

- With multilevel modeling we need to go beyond the classical setup of a data vector y and a matrix of predictors X. Each level of the model can have its own matrix of predictors.
- Observational study from Gelman and Hill, (2006): effect of city-level policies on enforcing child support payments from unmarried fathers.
- The treatment is at the group (city) level, but the outcome is measured on individual families.
- To estimate the effect of child support enforcement policies, the key "treatment" predictor is a measure of enforcement policies, which is available at the city level.
- Aim: estimate the probability that the mother received informal support, given the city-level enforcement measure and other city- and individual-level predictors.

Clustered data (cont.)

ID	dad age	mom race	informal support	city ID	city name	enforce intensity	benefit level	city indicators 1	2	···	20
1	19	hisp	1	1	Oakland	0.52	1.01	1	0	···	0
2	27	black	0	1	Oakland	0.52	1.01	1	0	···	0
3	26	black	1	1	Oakland	0.52	1.01	1	0	···	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
248	19	white	1	3	Baltimore	0.05	1.10	0	0	···	0
249	26	black	1	3	Baltimore	0.05	1.10	0	0	···	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
1366	21	black	1	20	Norfolk	−0.11	1.08	0	0	···	1
1367	28	hisp	0	20	Norfolk	−0.11	1.08	0	0	···	1

Figure: Table 1: compact table for clustered data

Clustered data (cont.)

ID	dad age	mom race	informal support	city ID
1	19	hisp	1	1
2	27	black	0	1
3	26	black	1	1
⋮	⋮	⋮	⋮	⋮
248	19	white	1	3
249	26	black	1	3
⋮	⋮	⋮	⋮	⋮
1366	21	black	1	20
1367	28	hisp	0	20

city ID	city name	enforce-ment	benefit level
1	Oakland	0.52	1.01
2	Austin	0.00	0.75
3	Baltimore	−0.05	1.10
⋮	⋮	⋮	⋮
20	Norfolk	−0.11	1.08

Figure: Table 2: two data-matrices for clustered data

Clustered data (cont.)

- First table: data for the analysis as it might be stored in a computer package, with information on each of the 1367 mothers surveyed.
- Second table: to make use of the **multilevel structure** of the data, however, we need to construct two data matrices, one for each level of the model (city and mothers).
- Conceptually, the two-matrix, or multilevel, data structure has the advantage of clearly showing which information is available on individuals and which on cities.
- It also gives more flexibility in fitting models, allowing us to move beyond the classical regression framework.

Clustered data (cont.)

We briefly outline several possible ways of analyzing these data, as a motivation and lead-in to multilevel modeling.

- *Individual-level regression*: $\Pr(Y_i = 1) = \text{logit}^{-1}(X_i\beta)$ where X includes the constant term, the treatment (enforcement intensity), and the other predictors (father's age and indicators for mother's race at the individual level; and benefit level at the city level). X is thus constructed from the data matrix of Table 1.

  Problem: it ignores city-level variation beyond that explained by enforcement intensity and benefit level,which are the city-level predictors in the model.

- *Group-level regression on city averages*: perform a city-level analysis, with individual-level predictors included using their group-level averages. The outcome, y_j, would be the average total support among the respondents in city j, the enforcement indicator would be the treatment, and the other variables would also be included as predictors. Such a regression—in this case, with 20 data points—has the advantage that its errors are automatically at the city level.

  Problem: however, by aggregating, it removes the ability of individual predictors to predict individual outcomes.

Clustered data (cont.)

- *Individual-level regression with city indicators, followed by group-level regression of the estimated city effects*: two-steps analysis, first fitting a logistic regression to the individual data y given individual predictors (in this example, father's age and indicators for mother's race) along with indicators for the 20 cities. Then, the next step is to perform a linear regression at the city level, considering the estimated coefficients of the city indicators (in the individual model that was just fit) as the "data" y_j . This city-level regression has 20 data points and uses, as predictors, the city-level data (in this case, enforcement intensity and benefit level).

  Problem: can run into problems when sample sizes are small in particular groups, or when there are interactions between individual- and group-level predictors.

Clustered data (cont.)

Multilevel modeling is a more general approach that can include predictors at both levels at once.

- The multilevel model looks something like the two-step model we have described, except that both steps are fitted at once.
- Two components: a logistic regression with 1369 data points predicting the binary outcome given individual-level predictors and with an intercept that can vary by city, and a linear regression with 20 data points predicting the city intercepts from city-level predictors.

$$\Pr(Y_i = 1) = \text{logit}^{-1}(\alpha_{j(i)} + X_i\beta), \quad i = 1, \ldots, n,$$

where X is the matrix of individual-level predictors and $j(i)$ indexes the city where person i resides.

Clustered data (cont.)

The second part of the model—what makes it "multilevel"—is the regression of the city coefficients:

$$\alpha_j \sim \mathcal{N}(U_j\gamma, \sigma_\alpha^2), \quad j = 1, \ldots, 20,$$

where U is the matrix of city-level predictors, γ is the vector of coefficients for the city-level regression, and σ_α is the standard deviation of the unexplained group-level errors.

- The key is the group-level variation parameter σ_α , which is estimated from the data (along with α, β).
- The model for the α allows us to include all 20 of them in the model *without having to worry about collinearity*.

Repeated measurements

- Another kind of multilevel data structure involves repeated measurements on persons (or other units)—thus, measurements are clustered within persons, and predictors can be available at the measurement or person level.
- Suppose a dataset where some people who bought an insurance policy are every year asked either to renew or to interrupt the policy. We basically have as many repeated measurements for each person as many years that person is observed/asked.
- A naive multilevel logistic regression could then be similar to the previous model, with each α_j defined here in terms of the j-th ensured for which the i-th policy was observed.
- Here also, we can work with a more rectangular-structured data matrix (similarly as Table 1) or with two-data matrices: the choice is done in terms of users' convenience.

Indicator variables and fixed or random effects

- When including an input variable with J categories into a classical regression, standard practice is to choose one of the categories as a baseline and include indicators for the other $J-1$ categories (in the child enforcement example, one could set city 1 (Oakland) as the baseline and include indicators for the other 19. The coefficient for each city then represents its comparison to Oakland.)
- In a multilevel model it is unnecessary to do this arbitrary step of picking one of the levels as a baseline. For example, in the child support study, one would include indicators for all 20 cities in the model. In a classical regression these could not all be included because they would be collinear with the constant term, but in a multilevel model this is not a problem because they are themselves modeled by a group-level distribution.

Indicator variables and fixed or random effects (cont.)

- The varying coefficients (α_j's or β_j's) in a multilevel model are sometimes called **random effects**, a term that refers to the randomness in the probability model for the group-level coefficients.
- The term **fixed effects** is used in contrast to random effects—but not in a consistent way! Fixed effects are usually defined as varying coefficients that are not themselves modeled.
- As an interpretation issue, fixed effects are constant across individuals, and random effects vary.

Costs and benefits of multilevel modeling

Before we go to the effort of learning multilevel modeling, it is helpful to briefly review what can be done with classical regression:

- Prediction for continuous or discrete outcomes,
- Fitting of nonlinear relations using transformations,
- Inclusion of categorical predictors using indicator variables,
- Modeling of interactions between inputs,
- Causal inference (under appropriate conditions).

Costs and benefits of multilevel modeling (cont.)

Motivations for moving to multilevel models:

- Accounting for individual- and group-level variation in estimating group-level regression coefficients.
- Modeling variation among individual-level regression coefficients. In classical regression, one can do this using indicator variables, but multilevel modeling is convenient when we want to model the variation of these coefficients across groups, make predictions for new groups, or account for group-level variation in the uncertainty for individual-level coefficients.
- Estimating regression coefficients for particular groups
- A potential drawback to multilevel modeling is the additional complexity of coefficients varying by group.

Costs and benefits of multilevel modeling (cont.)

- A multilevel model requires additional assumptions beyond those of classical regression—basically, each level of the model corresponds to its own regression with its own set of assumptions such as additivity, linearity, independence, equal variance, and normality.
- The usual alternative to multilevel modeling is classical regression—either ignoring group-level variation, or with varying coefficients that are estimated classically (and not themselves modeled)—or combinations of classical regressions.
- In various limiting cases, the classical and multilevel approaches coincide. When there is very little group-level variation, the multilevel model reduces to classical regression with no group indicators; conversely, when group-level coefficients vary greatly (compared to their standard errors of estimation), multilevel modeling reduces to classical regression with group indicators.

Costs and benefits of multilevel modeling (cont.)

- When the number of groups is small (less than five, say), there is typically not enough information to accurately estimate group-level variation. As a result, multilevel models in this setting typically gain little beyond classical varying-coefficient models.
- Computational softwares: `lme4`, `WinBUGS`, `JAGS`, Stan.

Further reading

Further reading

- *Generalized Linear Models*, P. McCullagh, J. A. Nelder.
- Chapter 11 from *Data Analysis using Regression and Multilevel/Hierarchical models*, A. Gelman and Jennifer Hill.