# Generalized Linear Models (GLM)

(Extending the linear model)

R. Bellio & N. Torelli

Spring 2018

University of Udine & University of Trieste

**Introduction**

**Inference**

**Model Evaluation**

# Introduction

## 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
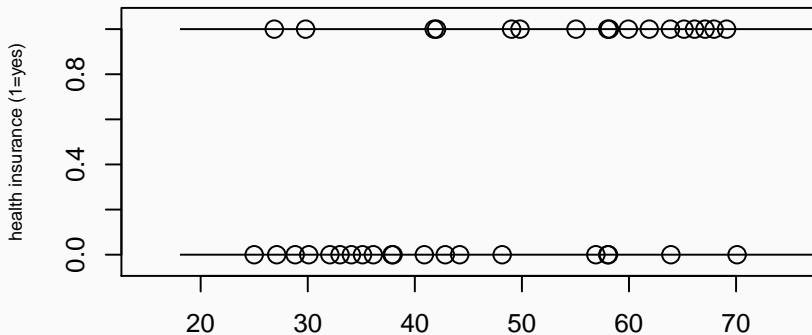


For a sample of 37 individuals we observe the age of any sample unit and whether he/she owns a private health insurance.

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 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))
```
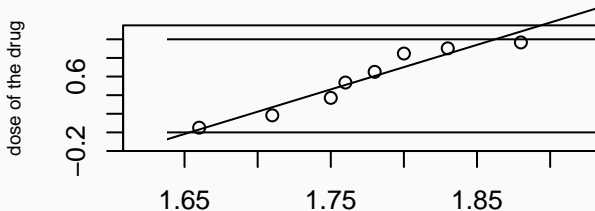


- The plot shows that the proportion of positive responses out of $m_i$ on trial, increases with the dose of the drug.
- A linear relationship is patently inappropriate. The data are proportions and their values should lie in the $[0,1]$ range
- $Y_i \sim$ Binomial$(m_i, h(x_i))$. Specify a non linear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

# Count data: an example with medical examinations
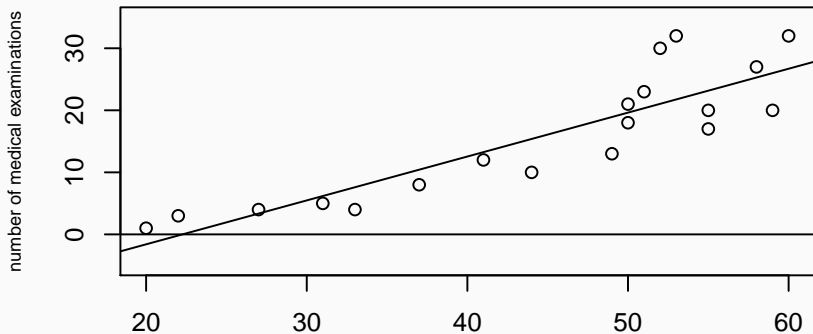


- $Y_i$ (number of examinations) can be assumed Poisson: $Y_i \sim \text{Poisson}(\mu_i)$.
- We can assume that $\mu_i = h(x)$, i.e., it is a function of the covariate $x$.
- A linear specification is clearly inappropriate (also because it will predict negative values). We should choose among functions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

## From LM to GLM

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

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

where

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

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

   ;

- $\boldsymbol{\beta}$ e $\sigma^2$ are unknown parameters.

## Introducing GLMs

GLMs generalize LMs by:

- Considering a class of model of the form

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

  and $\mathbf{x}_i^T$ 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 $\mu$ is determined by $\eta$ that is related to it by a suitable function

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

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

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

$\theta$ e $\phi$ are unknown scalar parameters,
$b(\cdot)$ and $c(\cdot) > 0$ are known functions and domain of $Y$ do not depend on $\theta$ or $\phi$.

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

- $\theta$ is called the *natural or canonical parameter* of the exponential family.
- $\phi$ 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

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

per $Y = 0, 1, \ldots,$.

- 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

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

per $Y = 0, 1, \ldots, n,$.

- It belongs to (1) where $\theta = \log \frac{\pi}{1-\pi}$ natural parameter, $\phi = 1$,

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

e $c(Y, \phi) = \log \binom{n}{Y}$.

- $Y \sim EF(n \log(1 + e^\theta), 1)$.

## Mean and variance for Exponential family

- Functions $b(\cdot)$ 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 $\theta$ is: It follows that:

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

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

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). This becomes clear if we recall that $\mu$ is related to $\theta$, i.e., $\mu = b'(\theta)$.

## Main example

**Poisson**

We have for a Poisson with mean $\lambda$

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

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

**Binomial**

We have for a Binomial with parameters $(n, \pi)$

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

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

## The link function

- The second important step in specifying a GLM is the definition of the function relating $\mu_i$ and the linear predictor $\eta_i$.

- It is assumed that the link between $\mu_i$, the mean of $Y_i$, and $\mathbf{x}_i^T$, the covariate vector, is

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

- $g(\cdot)$ is a known monotone and differentiable function. The function $g(\cdot)$ is the *link function* between $\mu_i$ and $\eta_i$.

- the inverse function $g(\cdot)^{-1} = r(\cdot)$ is also called the response function

- Covariates enter into the model by the linear predictor $\mathbf{x}_i^T \boldsymbol{\beta}$, but the $\mu_i$ and $\eta_i$ are generally non linearly related.

- Appropriate choice 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 $\theta$ 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 software for estimation of GLMs (including R).

## GLM: Complete specification

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

  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** : $\eta_i = \mathbf{x}_i^T \beta$, $\mathbf{x}_i$ is a vector of constants and $\beta$ 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
     $E(Y_i) = g^{-1}(\mathbf{x}_i^T \beta)$.

## Some GLMs

**Normal Linear Regression model**

The standard normal LM is a GLM. In this case

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

$\theta_i = \eta_i = \mu_i$ e $g(\cdot)$ and the canonical link 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 $g(\mu_i) = \mathbf{x}_i^T \beta$. A good choice could be the log
function since $\mu_i = e^{\mathbf{x}_i^T \beta}$ is positive.

The choice: $\mu_i = e^{\mathbf{x}_i^T \beta} = e^{\eta_i}$
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 vary 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 $\mu_i$ 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 $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$ e $\Psi(\cdot)$ 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=.
abline(0,0); abline(lm(numvisite~etavis))
modpo=(glm(numvisite~etavis,poisson))
coepo=modpo$coefficients
curve(exp(coepo[1]+coepo[2]*x),col=4,add=T)
```
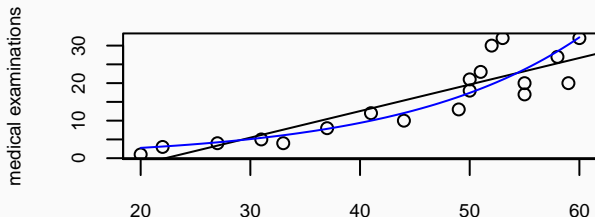


- A linear regression would be inappropriate (also because it will predict negative values)
- The blue curve seems to provide a better approximation

# Logistic regression: Health Insurance coverage

```r
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)
```
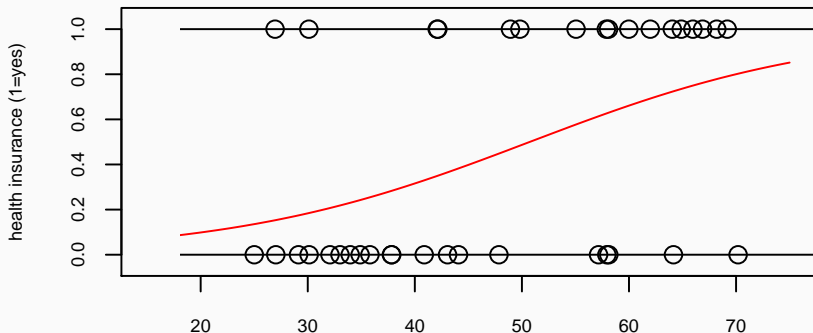


- It seems that older units are more likely to own a health insurance.
- The red curve displays the probability that a unit of a given age has insurance policy

# Inference

## Parameters estimation

- 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:
  - $g(\mu_i) = \eta_i = \mathbf{x}_i^T \beta \Leftrightarrow \mu_i = g^{-1}(\mathbf{x}_i^T \beta)$;
  - $\mu_i = b'(\theta_i) \Leftrightarrow \theta_i = (b')^{-1}(\mu_i) = (b')^{-1}(g^{-1}(\eta_i))$;
  - $var(Y_i) = \phi V(\mu_i)$, 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 $\theta_i$ is a function of $\beta$ through

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

## Likelihood equations

- To obtain the MLE of $\beta$ 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

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

- where the terms can be written as

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

## Likelihood equations

- Thus, we have

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

- The likelihood equations for $\beta$ 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 $\mu_i = g^{-1}(\mathbf{x}_i^T \beta)$.

- Note that the MLE of $\beta$ for a fixed value of $\phi$, does not depend on $\phi$ 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, $X^T y = X^T \mu$.

- 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 $\beta$, if the link it the canonical one, the theory of exponential families applies.

- In general the likelihood equations for $\beta$ are nonlinear and must be solved with iterative methods. To this end, the expected Fisher information for $\beta$ is useful.

## Fisher information

- Since $\beta$ and $\phi$ are orthogonal, we can proceed as if $\phi$ were known and we can focus only on $\beta$.
- Let us consider the second derivatives of $\ell_i$:

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

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

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

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.

# Fisher information

- 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 $V = \mathrm{diag}(V(\mu_i))$.

- Asymptotic normality of the MLE gives

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

  for large $n$.

- Therefore, a consistent estimate of the covariance matrix of $\beta$ is $i(\hat{\beta}) = \phi(X^T \hat{W} X)^{-1}$, where $\hat{W}$ is the matrix $W$ evaluated at $\hat{\beta}$.

- If $\phi$ 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 $\mu_i = x_i^T \beta$. The likelihood equations are $\sum_{i=1}^n \frac{(y_i - x_i^T \beta) x_{ij}}{\sigma^2} = 0$ 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 $\mu_i = \mathbf{x}_i^T \beta$. The likelihood equations are

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

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

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

### Poisson regression

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

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

which are generally nonlinear in $\beta$. In view of this, an explicit solution does not exists 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 $\beta$ in the linear predictor can be obtained by iterative weighted least squares.

- starting with appropriate initial value $\hat{\beta}^{(0)}$ and obtains 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
$$\ell_* = \left( \frac{\partial \ell}{\partial \beta_1}, \ldots, \frac{\partial \ell}{\partial \beta_p} \right)^T$$
the *score* vector. We want to solve the equation
$$\ell_* = \ell_*(\beta) = 0 \ .$$

- The Newton-Raphson method is based on the updating rule at the $(t+1)-$th iteration
$$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + (j(\hat{\beta}^{(t)}))^{-1}\ell_*^{(t)} \ , \tag{2}$$
with $\ell_*^{(t)} = \ell_*(\hat{\beta}^{(t)})$.

- The observed information can be replaced by the expected Fisher information $i(\beta)$. 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

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

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

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

## Weighted Least Squares

- where $s^{(t)}$ 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 $\hat{\beta}$.

- Therefore, it is possible to arrive to the expression

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

- Clearly, the parameter $\phi$ 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 $s$ and $W$ depend on $\beta$.

## 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 $\phi$

- For the LM, the estimation of $\beta$ is independent from the value of the variance $\sigma^2$. A similar situation holds for the dispersion parameter $\phi$ in GLMs.

- Obviously, the MLE of $\phi$, with $\beta$ replaced by $\hat{\beta}$, could be used.

- Also estimators based on the method of moments are often used for $\phi$.

- Since $var(Y_i) = \phi V(\mu_i)$ or, equivalently, since $\frac{E((Y_i - \mu_i)^2)}{V(\mu_i)} = \phi$ if $\beta$ is known, an unbiased estimator of $\phi$ is

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

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

$$\widehat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu_i})^2}{V(\widehat{\mu_i})} \quad \text{where} \quad \widehat{\mu_i} = g^{-1}(\boldsymbol{x}_i^T \widehat{\beta})$$

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 $\beta_j$ 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 $\beta_j$

# Model Evaluation

## 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 $\beta = (\beta_{MR}, \beta_{MC})$, where $\beta_{MR} = (\beta_1, \ldots, \beta_{p_0})$ and $\beta_{MC} = (\beta_{p_0+1}, \ldots, \beta_p)$. Suppose we want to test the following hypothesis

$$H_0 : \beta_{MC} = 0 \quad \text{against} \quad H_1 : \beta_{MC} \neq 0 \ .$$

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

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

## The residual deviance in LMs

- In normal LMs, with $\sigma^2$ 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 class 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 represent a common $\mu$ for all the $y_i$

A "'good'" model usually stands between these two extremes since a saturated model is uninformative 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.

- 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 **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 fit 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

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

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.

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.

## Example: Normal regression model

Since Normal LMs are GLMs with identity link 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 $\hat{\mu}_i = \mathbf{x}_i^T \hat{\beta}$, 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 $\sigma^2$.

## Poisson

- $Y_i \sim Poisson(\mu_i), b(\theta_i) = e^{\mu_i} = b'(\theta_i), \phi = 1, \log \mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$
- $\ell(\theta) = \sum_{i=1}^n y_i \log \mu_i - \sum_{i=1}^n \mu_i$
- 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 $\log \hat{\mu}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$, 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 $logit(\hat{\mu}_i) = \mathbf{x}_i^T \hat{\beta}$ 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

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

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 $\chi^2_{n-p}$.

- 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 $\phi$ is not known another consistent estimate of it must be considered

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

and, under mild condition 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-value 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 discrepances between few data and the rest (ouliers detection, evaluation of leverage: hat matrix, case deletion measures -Cook's distance-, jacknife residuals, etc.)
- Some of these ideas can be generalized in GLMs.
- A straight extension of the concept of standardized residual is given by

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

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

## Deviance residuals

- Recall that in GLMs $\epsilon_i$ 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

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

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 $r_{Pi} \approx r_{Di}$.

Other residuals, such as Anscombe residuals, are also defined for GLMs.

## 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
  $H = X(X^T X)^{-1} X)^T$.

- Generalized hat matrix is similarly obtained as
  $H = W^{\frac{1}{2}} X (X^T W X)^{-1} X^T W^{\frac{1}{2}}$ where $W$ is substitude by $\hat{W}$.

- A generalization of the Cook's distances is also possible.