Statistical Models: general definitions

Random samples

$\mathtt{Y}$	statistical phenomenon of interest (in a given population $\mathtt{P}$)
$\mathbf{y} = (y_1,\ y_2,\ \ldots,\ y_n)^\top$	Observed sample *(numerical) values observed on n statistical units* *randomly drawn from the population* $\mathtt{P}$
$\Downarrow$	
$\boldsymbol{Y} = (Y_1,\ Y_2,\ \ldots,\ Y_n)^\top$	Random sample *set of r.vs. that describe the possible value of* $\mathtt{Y}$ *in each random draw* $\Rightarrow\ \mathbf{y} \in \boldsymbol{\mathcal{Y}} \subseteq \mathbb{R}^n$ *Sample space* $\Rightarrow\ f_0(\mathbf{y})$ *Unknown "true" probability mass/density function of* $\boldsymbol{Y}$

Parametric statistical models

$f_0 \in \mathcal{F} = \left\{f(\cdot;\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k\right\}$	parametric statistical model for $\boldsymbol{Y}_n$ *parametric family containing the "true" probability mass/density function of* $\boldsymbol{Y}$ $\Rightarrow \boldsymbol{\Theta}$ *parameter space* $\Rightarrow\ f_0(\mathbf{y}) = f(\mathbf{y};\boldsymbol{\theta}_0)$ with $\boldsymbol{\theta}_0$ *unknown*

The elements of $\mathcal{F}$ *have the same functional form, and they differ only in the values of* $\boldsymbol{\theta}$

Parametric statistical model specification

The process of defining (*"specifying"*) a parametric statistical model $\mathcal{F}$ suitable for $\boldsymbol{Y}$ can be based on information about:

- the features of the statistical phenomenon `Y` of interest and of the population `P`

  $\Rightarrow$ *composition and properties of the support for each r. v. in the random sample, i. e. composition and properties of the elements in the sample space*

- the sampling scheme

  $\Rightarrow$ *dependence structure among the r. vs. in the random sample* $\boldsymbol{Y}$

Likelihood function of θ

Given:

- *a parametric statistical model for the random sample* $\boldsymbol{Y}$

  $\mathcal{F} = \left\{ f(\cdot; \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k \right\}$

- *an observed sample - realisation of the random sample* $\boldsymbol{Y}$

  $\mathbf{y} = (y_1,\ y_2,\ \ldots,\ y_n)^\top$

$\Rightarrow$ $L(\boldsymbol{\theta})$ *Likelihood function of* θ

$$L(\cdot) = L(\cdot; \mathbf{y}): \quad \begin{array}{ccc} \boldsymbol{\Theta} & \longmapsto & \mathbb{R}^+ \cup 0 \\ \boldsymbol{\theta} & \longmapsto & c(\mathbf{y})\, f(\mathbf{y}; \boldsymbol{\theta}) \end{array}$$

$c(\mathbf{y})$ represents a multiplicative factor that does not depend on $\boldsymbol{\theta}$

What is the likelihood function?

Literally, the likelihood function shows how the probability/density of observing the actually drawn sample changes, as the value of the unknown parameter $\boldsymbol{\theta}$ changes

$\Rightarrow$ it "summarises" all the available information about f_0, the "true" probability distribution of $\boldsymbol{Y}$:

- $f_0 \in \mathcal{F}$ parametric statistical model
  pre-experimental (a priori - before observing the actually drawn sample) information - theoretical assumptions
- $\mathbf{y} = (y_1,\ y_2,\ \ldots,\ y_n)^\top$ observed sample
  empirical evidence

$\Rightarrow$ It can be interpreted as a way to measure the "agreement" between each possible value of $\boldsymbol{\theta}$ and the observed sample, or, in other words, a way to measure the plausibility of each possible value of $\boldsymbol{\theta}$

An example

$Y_i \sim N\left(\mu, \sigma^2\right),\ \ \text{IID}\ \ i = 1, \ldots n, \quad \mu \in \mathbb{R}, \quad \sigma^2 \in \mathbb{R}^+$

$\Rightarrow$ Sample space: $\mathbf{y} \in \boldsymbol{\mathcal{Y}} = \mathbb{R}^n$

$\Rightarrow$ Parameter space: $\boldsymbol{\theta} = \left(\mu, \sigma^2\right)^\top \in \boldsymbol{\Theta} = \mathbb{R} \times \mathbb{R}^+$

$$f\left(\mathbf{y}; \boldsymbol{\theta}\right) \quad = \quad \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\left(y_i - \mu\right)^2}{2\sigma^2}\right] = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{\sum_{i=1}^{n}\left(y_i - \mu\right)^2}{2\sigma^2}\right]$$

Using compact notation:

$$\boldsymbol{Y} \sim MVN_n\left(\underbrace{\boldsymbol{\mu} = \begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix}}_{n \times 1}, \underbrace{\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_n}_{n \times n}\right)$$

Multivariate Gausssian distributions: A quick review

Joint probability density function

$\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$ — n-dimensional random variable

$\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top \in \mathbb{R}^n$ — set of possible values of $\boldsymbol{Y}$ (joint realisations of the n random variables)

$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^\top$ — n-dimensional real-valued vector

$\boldsymbol{\Sigma}$ — $n \times n$ real-valued, symmetric matrix (positive definite - invertible)

$$\boldsymbol{Y} \sim MVN_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \Leftrightarrow \quad f(y_1, \ldots, y_n; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\exp\left[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right]}{(2\pi)^{\frac{n}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}}}$$

$\mathrm{E}[\boldsymbol{Y}] = \boldsymbol{\mu}$

$\mathrm{Var}[\boldsymbol{Y}] = \mathrm{E}\left[\boldsymbol{Y}\boldsymbol{Y}^\top\right] - \mathrm{E}[\boldsymbol{Y}]\,\mathrm{E}[\boldsymbol{Y}]^\top = \boldsymbol{\Sigma}$

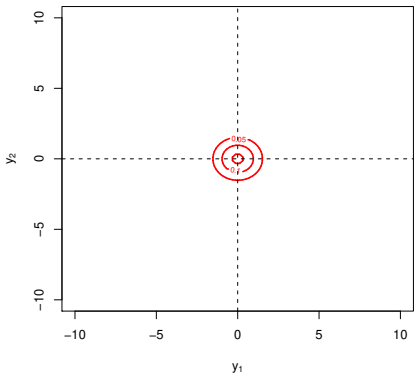
Standardised multivariate Gaussian distribution

$\boldsymbol{\mu} = \mathbf{0}_n = (0, 0, \ldots, 0)^\top$ — n-dimensional null vector

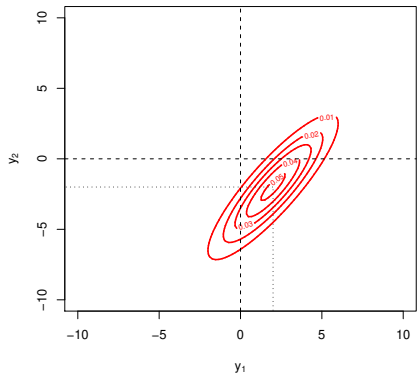
$\boldsymbol{\Sigma} = \mathbf{I}_n$ — $n \times n$ identity matrix

$$\begin{aligned} f(y_1, \ldots, y_n; \mathbf{0}_n, \mathbf{I}_n) &= \frac{\exp\left[-\frac{1}{2}\sum_{i=1}^n y_i^2\right]}{(2\pi)^{\frac{n}{2}}} \\ &= \prod_{i=1}^n \frac{\exp\left[-\frac{y_i^2}{2}\right]}{\sqrt{2\pi}} \end{aligned}$$

Some examples: $n = 2$ - 1

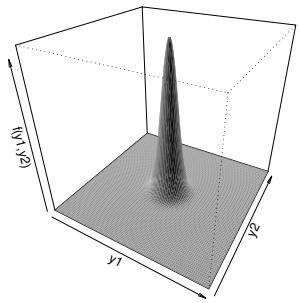


$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$
$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

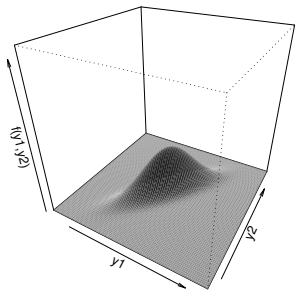


$$\boldsymbol{\mu} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$$
$$\boldsymbol{\Sigma} = \begin{bmatrix} 4.8 & 5.4 \\ 5.4 & 7.95 \end{bmatrix}$$

Some examples: $n = 2$ - 2



$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$
$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\boldsymbol{\mu} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$$
$$\boldsymbol{\Sigma} = \begin{bmatrix} 4.8 & 5.4 \\ 5.4 & 7.95 \end{bmatrix}$$

Some properties

- each marginal distribution of order $q < n$ is a q-dimesional multivariate Gaussian distribution
- each conditional distribution of $Y_{1a}, Y_{2a}, \ldots, Y_{ha}$ given $Y_{1b}, Y_{2b}, \ldots, Y_{lb}$ is an h-dimesional multivariate Gaussian distribution
- $Y_1, Y_2, \ldots, Y_n$ are independent if and only if $\boldsymbol{\Sigma}$ is diagonal (if and only if they are uncorrelated):

$$\begin{aligned} f(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{\exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{\sigma_i^2}\right]}{(2\pi)^{\frac{n}{2}}\left[\prod_{i=1}^{n}\sigma_i^2\right]^{\frac{1}{2}}} \\ &= \prod_{i=1}^{n}\left\{\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left[-\frac{1}{2\sigma_i^2}(y_i-\mu_i)^2\right]\right\} \end{aligned}$$

Linear combinations of multivariate Gaussian random variables

$\boldsymbol{Y}$	n-dimensional Gaussian vector with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$
$\mathbf{A}$	$q \times n$ real-valued matrix
$\mathbf{b}$	n-dimensional real-valued vector
$\mathbf{Z} = \mathbf{A}(\boldsymbol{Y} + \mathbf{b})$	q-dimensional Gaussian vector with parameters $\mathbf{A}(\boldsymbol{\mu} + \mathbf{b})$ and $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top$

Particular example:

$\mathbf{A} = \boldsymbol{\Sigma}^{-\frac{1}{2}}$	such that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}}$ and $\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}} = \mathbf{I}_n$
$\mathbf{b} = -\boldsymbol{\mu}$	
$\boldsymbol{Z} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{Y} - \boldsymbol{\mu})$	n-dimensional standardised Gaussian vector

Gaussian linear models:
an introductory example

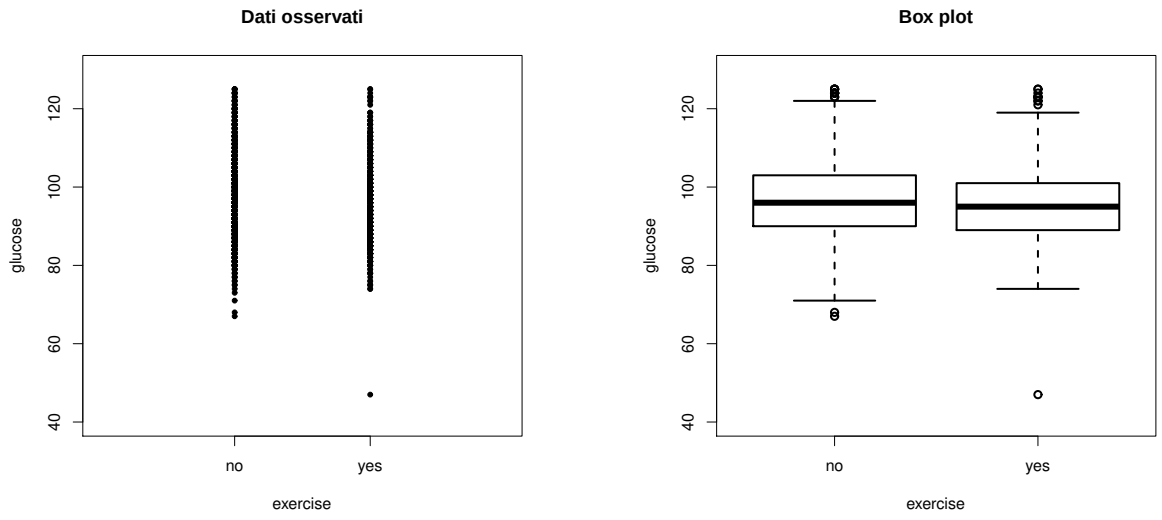
An example from biostatistics

A glucose level in blood between 100 and 125 mg/dL represents a risk factor for developing diabetes.

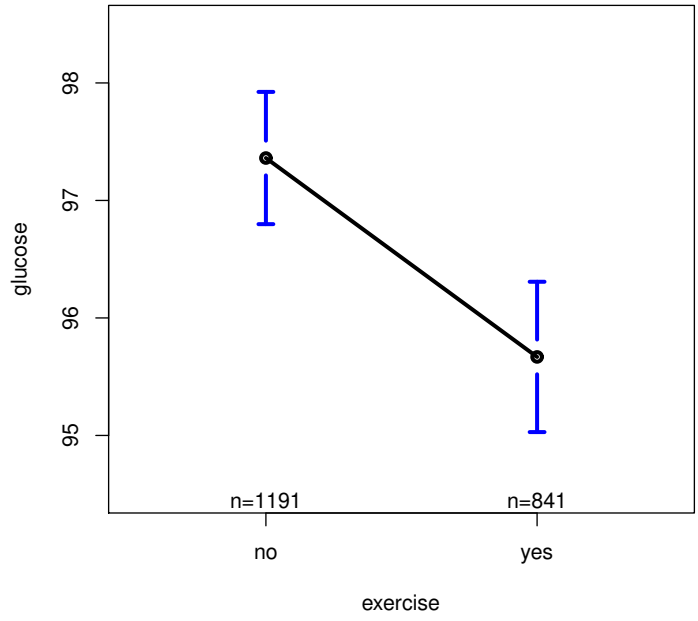
Aim: Does physical activity (a modifiable factor related to life style) contribute to the reduction of the glucose level, thus preventing a severe disease?

available information: data from an observational study
Glucose level and physical activity (yes-no) on a sample of 2032 women not affected by diabetes after menopause

Some graphical displays



Conditional means (C.I. $1 - \alpha = 0.95$)



Simple linear regression models: definition

$f(\texttt{glucose}_i, \texttt{exercise}_i) = f(\texttt{glucose}_i|\texttt{exercise}_i)\, f(\texttt{exercise}_i)$

$i = 1, \ldots, 2032$

A) $\mathrm{E}[\texttt{glucose}_i|\texttt{exercise}_i] = \beta_0 + \beta_1 \mathbf{1}\{\texttt{exercise}_i = \texttt{yes}\}\ \forall i$

$$\mathbf{1}\{\texttt{exercise}_i = \texttt{yes}\} = \begin{cases} 1 & \text{if } \texttt{exercise}_i = \texttt{yes} \\ 0 & \text{otherwise} \end{cases}$$

B) $\mathrm{Var}[\texttt{glucose}_i|\texttt{exercise}_i] = \sigma^2\ \forall i$

C) $\mathrm{Cor}\left[\texttt{glucose}_i|\texttt{exercise}_i, \texttt{glucose}_j|\texttt{exercise}_j\right] = 0\ \forall i \neq j$

D) $\texttt{glucose}_i|\texttt{exercise}_i \sim N\left(\beta_0 + \beta_1 \mathbf{1}\{\texttt{exercise}_i = \texttt{yes}\}, \sigma^2\right)\ \forall i$

Simple linear regression models: some results

```
> summary(modello1)
Call:
lm(formula = glucose ~ exercise, data = hers.nod)

Residuals:
     Min       1Q   Median       3Q      Max
 -48.668   -6.668   -0.668    5.639   29.332
 Coefficients:
                Estimate   Std.  Error   t value   Pr(>|t|)
   (Intercept)    97.3610        0.2815    345.85     0.0000
   exerciseyes    -1.6928        0.4376     -3.87     0.0001


Residual standard error:  9.715 on 2030 degrees of freedom
Multiple R-squared:  0.007318, Adjusted R-squared:  0.006829
F-statistic:  14.97 on 1 and 2030 DF, p-value:  0.000113
```


Multiple linear regression 7

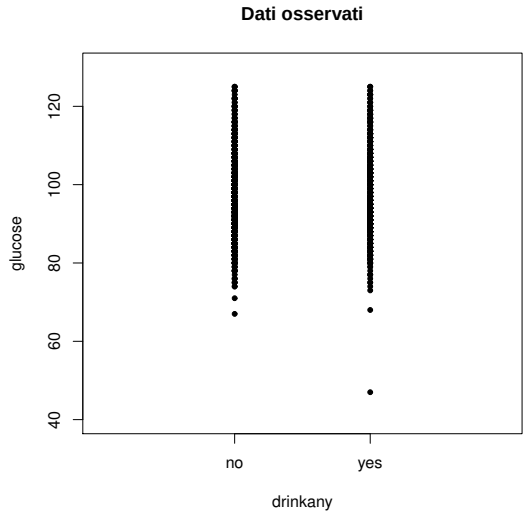
Introducing other regressors

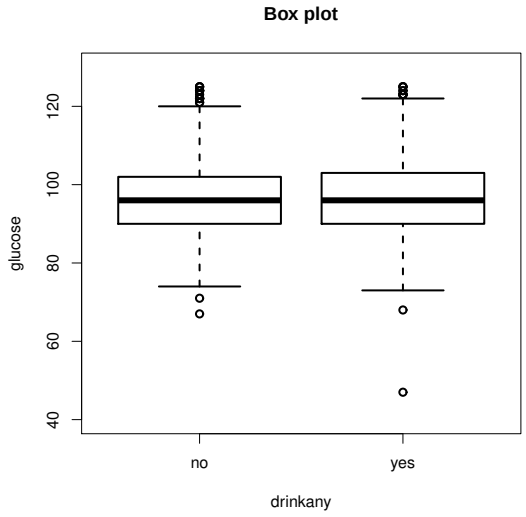
Women that are physically active may completely differ from women that are not, due to a number of other characteristics (socio-economical status, life style, health conditions.
Some of these characteristics could be associated with both the glucose level and physical activity.

Example: women that are physically active could be younger, haltier and have different habits related to alcohol consumption.

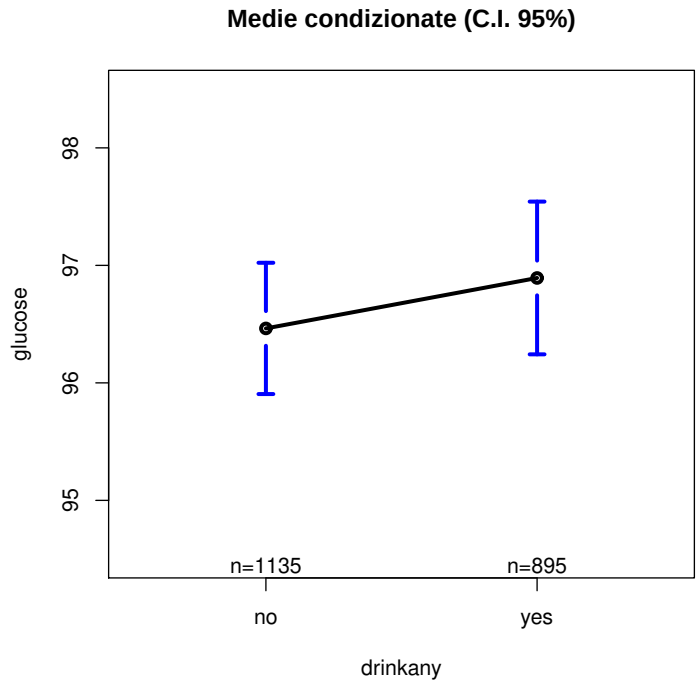
Since data were collected through an observational study, these characteristics could act as confounders, thus preventing a correct evaluation of the effect of physical activity on glucose level.

Alcohol consumption (1)

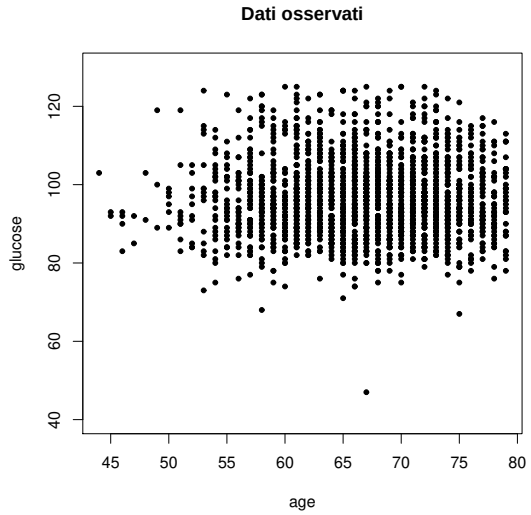


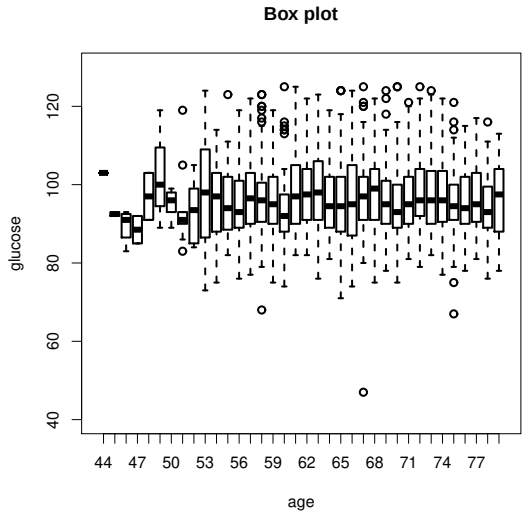


Alcohol consumption (2)

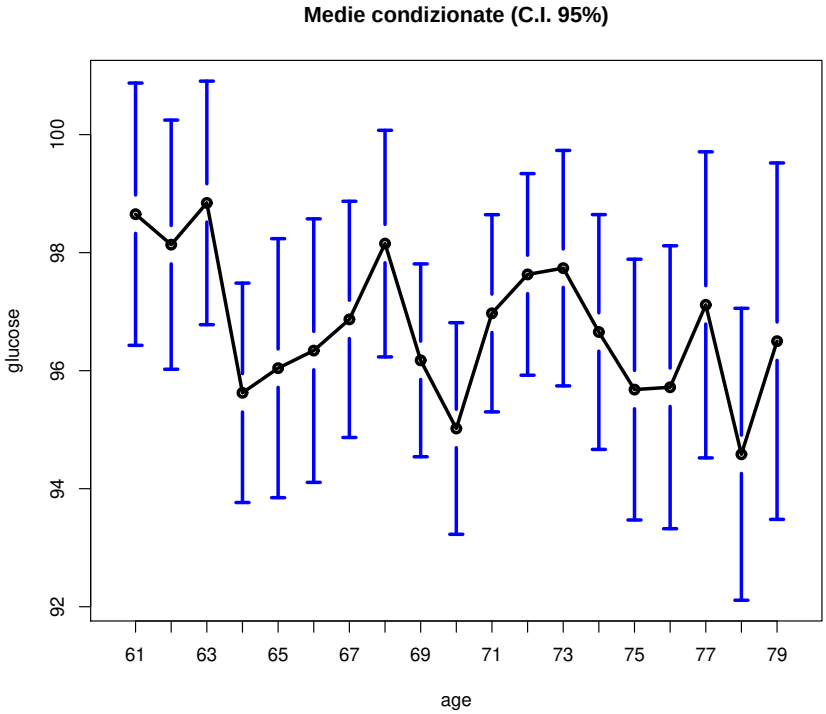


Age (1)

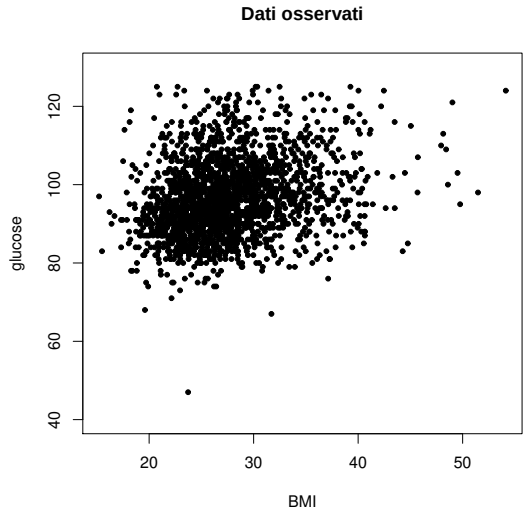


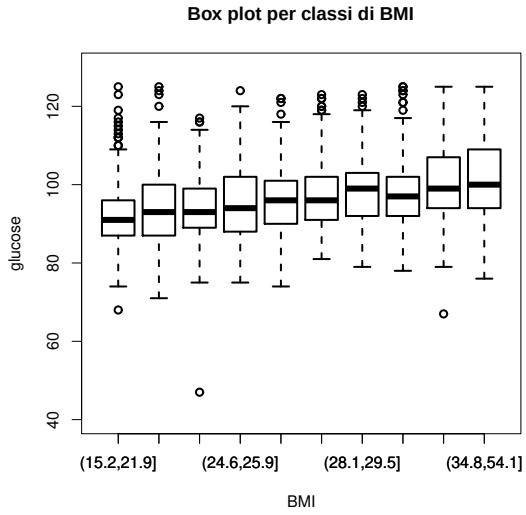


Age - only women 60+ (2)

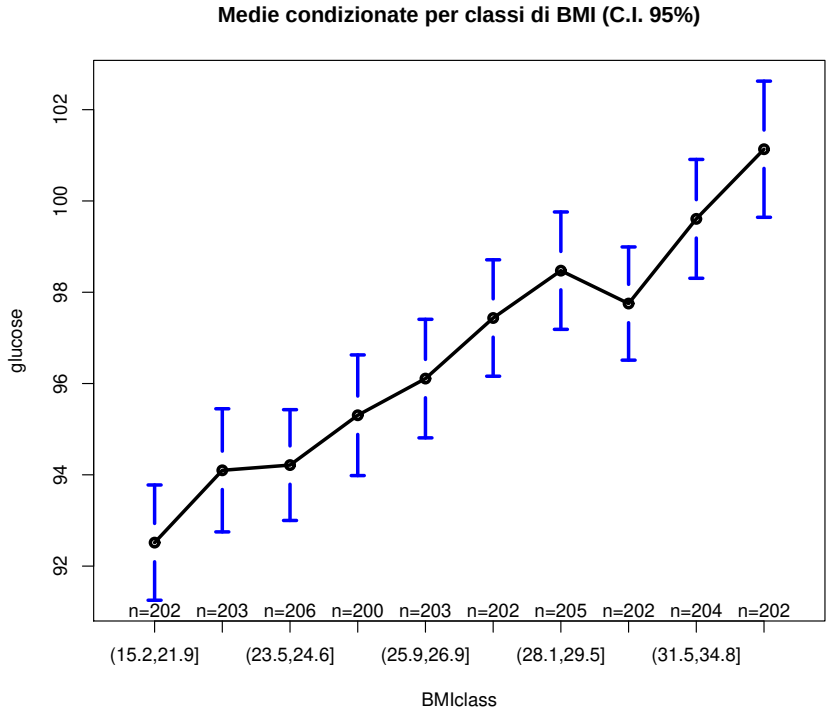


Body mass index (1)





Body mass index (2)



Multiple linear regression models: definition

$f(\texttt{glucose}_i, \texttt{exercise}_i, \texttt{drinkany}_i, \texttt{age}_i, \texttt{BMI}_i)$

$= f(\texttt{glucose}_i | \texttt{exercise}_i, \texttt{drinkany}_i, \texttt{age}_i, \texttt{BMI}_i) \cdot$

$f(\texttt{exercise}_i, \texttt{drinkany}_i, \texttt{age}_i, \texttt{BMI}_i)$

$= f(y_i | x_{1i}, x_{2i}, x_{3i}, x_{4i}) f(x_{1i}, x_{2i}, x_{3i}, x_{4i}) \quad i = 1, \ldots, 2032$

A) $\mathrm{E}[Y_i | x_{1i}, x_{2i}, x_{3i}, x_{4i}] = \mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} =$

$\beta_0 + \beta_1 \mathbf{1}\{\texttt{exercise}_i = \texttt{yes}\} + \beta_2 \mathbf{1}\{\texttt{drinkany}_i = \texttt{yes}\} + \beta_3 \texttt{age}_i + \beta_4 \texttt{BMI}_i \ \forall i$

B) $\mathrm{Var}[Y_i | x_{1i}, x_{2i}, x_{3i}, x_{4i}] = \sigma^2 \ \forall i$

C) $\mathrm{Cor}[Y_i | x_{1i}, x_{2i}, x_{3i}, x_{4i}, Y_h | x_{1h}, x_{2h}, x_{3h}, x_{4h}] = 0 \ \forall i \neq h$

D) $\texttt{glucose}_i | \texttt{exercise}_i, \texttt{drinkany}_i, \texttt{age}_i, \texttt{BMI}_i \sim N(\mu_i, \sigma^2) \ \forall i$

Multiple linear regression models: some results

```
> summary(modello2)
Call:
lm(formula = glucose ~ exercise + drinkany + age + BMI, data = hers.nod)

...
...

 Coefficients:
               Estimate   Std.  Error   t value   Pr(>|t|)
   (Intercept)   78.9624        2.5928     30.45     0.0000
   exerciseyes   -0.9504        0.4287     -2.22     0.0267
   drinkanyyes    0.6803        0.4220      1.61     0.1071
           age    0.0635        0.0314      2.02     0.0431
           BMI    0.4892        0.0416     11.77     0.0000

...
...
```

Gaussian linear models: basic assumptions

Y_i Random variable that describes the value for the dependent variable observed on the $i-$th sample unit $(i = 1, \ldots, n)$

$x_{1i}, x_{2i}, \ldots, x_{pi}$ values of the regressors for the $i-$th sample unit (*covariate pattern*)

A) $\mathrm{E}\left[Y_i | x_{1i}, \ldots, x_{pi}\right] = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} \ \forall i$

B) $\mathrm{Var}\left[Y_i | x_{1i}, \ldots, x_{pi}\right] = \sigma^2 \ \forall i$

C) $\mathrm{Cor}\left[Y_i | x_{1i}, \ldots, x_{pi}, Y_h | x_{1h}, \ldots, x_{ph}\right] = 0 \ \forall i \neq h$

D) $Y_i | x_{1i}, \ldots, x_{pi} \sim N\left(\mu_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}, \sigma^2\right) \ \forall i$

Parameter space and sample space

Model parameters:

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^{(p+1)}$ $(p+1)-$dimensional real-valued vector

$\sigma^2 \in \mathbb{R}^+$ positive scalar value

$\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top, \sigma^2\right)^\top \in \boldsymbol{\Theta} = \mathbb{R}^{(p+1)} \times \mathbb{R}^+$

Conditional sample space:

$\mathbb{R} \times \left\{(x_{1i}, \ldots, x_{pi})^\top, i = 1, \ldots, n\right\}$

Probability density function (1)

$f(y_1, \ldots, y_n | x_{11}, \ldots, x_{p1}, x_{1n}, \ldots, x_{pn})$ joint - for the r.vs. $Y_1, \ldots, Y_n$ conditional - given the regressor values $= \prod_{i=1}^{n} f(y_i | x_{1i}, \ldots, x_{pi})$

conditional independence

$= \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} (y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi})^2 \right] \right\}$ *Gaussian distribution*

$= \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \right\}^n \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi})^2 \right]$

Matrix representation

$\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$ n-dimensional random variable that describes the values for the dependent variable jointly observed on n sample units

$\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$ observed sample values

$\mathbf{x}_i = (x_{0i}, x_{1i} \ldots, x_{pi})^\top$ regressor values for the l'i-th sample unit ($x_{0i} = 1 \ \forall i$ constant regressor associated with the intercept)

$\mathbf{x}_{[j]} = (x_{j1}, x_{j2} \ldots, x_{jn})^\top$ observed values for the j-th regressor ($j = 0, \ldots, p$) $x_{[0]} = (1, 1, \ldots, 1)^\top$

Regressor matrix $n \times (p+1)$ $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{p1} \\ 1 & x_{12} & \ldots & x_{p2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \ldots & x_{pn} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^\top \\ \hline \mathbf{x}_2^\top \\ \hline \vdots \\ \hline \mathbf{x}_n^\top \end{bmatrix} = \left[\mathbf{x}_{[0]} \middle| \, \mathbf{x}_{[1]} \middle| \cdots \mathbf{x}_{[p]} \middle| \right]$

Conditional expected values

$\mathrm{E}[Y_i|x_{1i},\ldots,x_{pi}] = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} = \mathbf{x}_i^\top \boldsymbol{\beta}\ \forall i$

in compact form:

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} \mathbf{x}_1^\top \boldsymbol{\beta} \\ \mathbf{x}_2^\top \boldsymbol{\beta} \\ \vdots \\ \mathbf{x}_n^\top \boldsymbol{\beta} \end{bmatrix}$$

Probability density function (2)

$$\begin{aligned} f\left(\mathbf{y}|\mathbf{X};\boldsymbol{\beta},\sigma^2\right) &= \left\{\frac{1}{\sqrt{2\pi\sigma^2}}\right\}^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi})^2\right] \\ &= \left\{\frac{1}{\sqrt{2\pi\sigma^2}}\right\}^n \exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)^\top\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)\right] \end{aligned}$$

According to Assumptions A) to E), the resulting statistical model leads to a multivariate Gaussian distribution for $\boldsymbol{Y}$, given the regressor values:

$\boldsymbol{Y}|\mathbf{X} \sim MVN_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right)$

An alternative definition (1)

Y_i Random variable that describes the value for the dependent variable observed on the $i-$th sample unit $(i = 1, \ldots, n)$

$x_{1i}, x_{2i}, \ldots, x_{pi}$ values of the regressors for the $i-$th sample unit (*covariate pattern*)

$$Y_i = \underbrace{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}}_{\text{deterministic component}} + \underbrace{\varepsilon_i}_{\text{random error}}$$

A) $\mathrm{E}\left[\varepsilon_i | x_{1i}, \ldots, x_{pi}\right] = 0 \;\; \forall i$

B) $\mathrm{Var}\left[\varepsilon_i | x_{1i}, \ldots, x_{pi}\right] = \sigma^2 \; \forall i$

C) $\mathrm{Cor}\left[\varepsilon_i | x_{1i}, \ldots, x_{pi}, \varepsilon_h | x_{1h}, \ldots, x_{ph}\right] = 0 \; \forall i \neq h$

D) $\varepsilon_i | x_{1i}, \ldots, x_{pi} \sim N\left(0, \sigma^2\right) \;\; \forall i$

An alternative definition (2)

$\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$

$\boldsymbol{\varepsilon} | \mathbf{X} \sim NMV_n\left(\mathbf{0}_n, \sigma^2 \mathbf{I}_n\right)$

$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Thanks to the properties of multivariate Gaussina distributions, setting $\mathbf{A} = \mathbf{I}_n$ and $\mathbf{b} = \mathbf{X}\boldsymbol{\beta}$:

$\boldsymbol{Y} | \mathbf{X} \sim MVN_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n\right)$

Gaussian linear models:
Maximum likelihood estimation

Likelihood function

The unknown parameters in a Gaussian linear regression models are

- $\boldsymbol{\beta}$ regression coefficients (including the intercept)
- σ^2 conditional variance

Given the regressor values in matrix $\mathbf{X}$ and the observed values for the dependent variable on the sample units in vector $\mathbf{y}$:

$$\begin{aligned} L\left(\boldsymbol{\beta}, \sigma^2\right) &= L\left(\boldsymbol{\beta}, \sigma^2; \mathbf{y} \,|\mathbf{X}\right) \\ &= \left\{\frac{1}{\sqrt{2\pi\sigma^2}}\right\}^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi})^2\right] \\ &= \left\{\frac{1}{\sqrt{2\pi\sigma^2}}\right\}^n \exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^\top \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)\right] \end{aligned}$$

Log-likelihood function

$$\begin{aligned} l\left(\boldsymbol{\beta}, \sigma^2\right) &= \ln L\left(\boldsymbol{\beta}, \sigma^2\right) \\ &= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi})^2 \\ &= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^\top \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) \end{aligned}$$

Note that:

$-\frac{n}{2}\ln 2\pi$ additive constant independent from the unknown parameters (it can be ignored)

Score function for β

$U(\boldsymbol{\beta}) = \dfrac{\partial}{\partial \boldsymbol{\beta}} \ln L\left(\boldsymbol{\beta}, \sigma^2\right)$ Gradient of $l\left(\boldsymbol{\beta}, \sigma^2\right)$ with respect to $\boldsymbol{\beta}$

vector with $p+1$ elements

$U_j(\boldsymbol{\beta}) = \dfrac{\partial}{\partial \beta_j} \ln L\left(\boldsymbol{\beta}, \sigma^2\right),\ j = 0, \ldots, p$ First order partial derivative of $l\left(\boldsymbol{\beta}, \sigma^2\right)$ with respect to β_j $(j = 0, \ldots, p)$

generic element of $U(\boldsymbol{\beta})$

$$
\begin{aligned}
U_j(\boldsymbol{\beta}) &= \frac{\partial}{\partial \beta_j}\left\{-\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\textstyle\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi})^2\right\} \\
&= -\frac{1}{2\sigma^2}\textstyle\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi}) \cdot 2 \cdot (-1) \cdot x_{ji} \\
&= \frac{1}{\sigma^2}\textstyle\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi}) x_{ji}
\end{aligned}
$$

Matrix representation for $U(\beta)$

Exploiting the dot product

$$
\begin{aligned}
U_j(\boldsymbol{\beta}) &= \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi}) x_{ji}}{\sigma^2} \\
&= \frac{\mathbf{x}_{[j]}^{\top}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2}
\end{aligned}
$$

$$
U(\boldsymbol{\beta}) = \begin{bmatrix} \frac{\mathbf{x}_{[0]}^{\top}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2} \\ \frac{\mathbf{x}_{[1]}^{\top}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2} \\ \vdots \\ \frac{\mathbf{x}_{[p]}^{\top}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2} \end{bmatrix} = \frac{\mathbf{X}^{\top}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2}
$$

An alternative derivation of $U(\beta)$

Exploiting the differentiation rules for functions with vector arguments

$$
\begin{aligned}
U(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}, \sigma^2) = \frac{\partial}{\partial \boldsymbol{\beta}} \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\} \\
&= -\frac{1}{2\sigma^2} \frac{\partial}{\partial \boldsymbol{\beta}} \left\{ \mathbf{y}^\top \mathbf{y} - \underbrace{\mathbf{y}^\top \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{y}}_{2\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{y}} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} \right\} \\
&\quad \textit{recall that} \quad \frac{\partial}{\partial \boldsymbol{\delta}} \boldsymbol{\delta}^\top \mathbf{A} = \mathbf{A} \quad \textit{e} \quad \frac{\partial}{\partial \boldsymbol{\delta}} \boldsymbol{\delta}^\top \mathbf{A}\boldsymbol{\delta} = 2\mathbf{A}\boldsymbol{\delta} \\
&= -\frac{1}{2\sigma^2} \left\{ \mathbf{0}_{p+1} - 2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} \right\} = \frac{\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2}
\end{aligned}
$$

Observed Fisher information for β

$i(\boldsymbol{\beta}) = -\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \ln L(\boldsymbol{\beta}, \sigma^2)$ Negative of the Hessian matrix $l(\boldsymbol{\beta}, \sigma^2)$ with respect to $\boldsymbol{\beta}$ $(p+1) \times (p+1)$ *matrix*

$i_{jl}(\boldsymbol{\beta}) = -\frac{\partial^2}{\partial \beta_j \partial \beta_l} \ln L(\boldsymbol{\beta}, \sigma^2)$ Second order partial derivative of $l(\boldsymbol{\beta}, \sigma^2)$ with respect to β_j and β_l $(j, l = 0, \ldots, p)$ *generic element of* $i(\boldsymbol{\beta})$

$$
\begin{aligned}
i_{jl}(\boldsymbol{\beta}) &= -\frac{\partial}{\partial \beta_l} U_j(\boldsymbol{\beta}) \\
&= -\frac{\partial}{\partial \beta_l} \frac{1}{\sigma^2} \textstyle\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi}) x_{ji} \\
&= -\frac{1}{\sigma^2} \textstyle\sum_{i=1}^n x_{ji} x_{li} \cdot (-1) = \frac{1}{\sigma^2} \textstyle\sum_{i=1}^n x_{ji} x_{li}
\end{aligned}
$$

Matrix representation for $i(\beta)$

Exploiting the dot product

$$\begin{aligned} i_{jl}(\boldsymbol{\beta}) &= \frac{\sum_{i=1}^{n} x_{ji}x_{li}}{\sigma^2} \\ &= \frac{\mathbf{x}_{[j]}^{\top}\mathbf{x}_{[l]}}{\sigma^2} \end{aligned}$$

$$i(\boldsymbol{\beta}) = \begin{bmatrix} \frac{\mathbf{x}_{[0]}^{\top}\mathbf{x}_{[0]}}{\sigma^2} & \frac{\mathbf{x}_{[0]}^{\top}\mathbf{x}_{[1]}}{\sigma^2} & \dots & \frac{\mathbf{x}_{[0]}^{\top}\mathbf{x}_{[p]}}{\sigma^2} \\ \frac{\mathbf{x}_{[1]}^{\top}\mathbf{x}_{[0]}}{\sigma^2} & \frac{\mathbf{x}_{[1]}^{\top}\mathbf{x}_{[1]}}{\sigma^2} & \dots & \frac{\mathbf{x}_{[1]}^{\top}\mathbf{x}_{[p]}}{\sigma^2} \\ \vdots & & & \\ \frac{\mathbf{x}_{[p]}^{\top}\mathbf{x}_{[0]}}{\sigma^2} & \frac{\mathbf{x}_{[p]}^{\top}\mathbf{x}_{[1]}}{\sigma^2} & \dots & \frac{\mathbf{x}_{[p]}^{\top}\mathbf{x}_{[p]}}{\sigma^2} \end{bmatrix} = \frac{\mathbf{X}^{\top}\mathbf{X}}{\sigma^2}$$

An alternative derivation of $i(\beta)$

Exploiting the differentiation rules for functions with vector arguments

$$\begin{aligned} i(\boldsymbol{\beta}) &= -\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\top}} l(\boldsymbol{\beta},\sigma^2) = -\frac{\partial}{\partial\boldsymbol{\beta}^{\top}} U(\boldsymbol{\beta}) = -\frac{1}{\sigma^2}\frac{\partial}{\partial\boldsymbol{\beta}^{\top}}\left(\mathbf{X}^{\top}\mathbf{y} - \mathbf{X}^{\top}\mathbf{X}\boldsymbol{\beta}\right) \\ &\quad \textit{recall that} \quad \frac{\partial}{\partial\boldsymbol{\delta}^{\top}}\mathbf{A}\boldsymbol{\delta} = \mathbf{A} \\ &= -\frac{1}{\sigma^2}\left[\mathbf{0}_{(p+1)\times(p+1)} - \mathbf{X}^{\top}\mathbf{X}\right] = \frac{\mathbf{X}^{\top}\mathbf{X}}{\sigma^2} \end{aligned}$$

Expected Fisher information for β

$I(\boldsymbol{\beta}) = -\mathrm{E}\left[\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\ln L\left(\boldsymbol{\beta},\sigma^2\right)\right] = \mathrm{E}\left[i(\boldsymbol{\beta})\right]$ $\quad (p+1)\times(p+1)$ *matrix*

$I_{jl}(\boldsymbol{\beta}) = \mathrm{E}\left[i_{jl}(\boldsymbol{\beta})\right]$ *Expected value of the generic element of* $i(\boldsymbol{\beta})$

$$\mathrm{E}\left[i_{jl}(\boldsymbol{\beta})\right] = \mathrm{E}\left[\underbrace{\frac{\sum_{i=1}^n x_{ji}x_{li}}{\sigma^2}}_{\text{independent of } \boldsymbol{Y}}\right] = \frac{\sum_{i=1}^n x_{ji}x_{li}}{\sigma^2}$$

$$\mathrm{E}\left[i(\boldsymbol{\beta})\right] = \frac{\mathbf{X}^\top\mathbf{X}}{\sigma^2}$$

Note that: the expected values are computed considering the conditional distribution of $\boldsymbol{Y}$ *given* $\mathbf{X}$ *(thus holding fixed the values of the regressors)*

Properties of the score function

$U(\boldsymbol{\beta}) = \frac{\mathbf{X}^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2} \Rightarrow$ depends on:

- $\boldsymbol{\beta}, \sigma^2$ the *unknown* model parameters
- $\mathbf{X}$ the regressor values
- $\mathbf{y}$ the observed values of the dependent variable (*realisations of the r. v.* $\boldsymbol{Y}$)

$\Rightarrow$ Conditionally on the regressors values, $U(\boldsymbol{\beta})$ is the realisation of a random vector, that can be expressed as a linear transformation of $\boldsymbol{Y}$:

$$\mathbf{A} = \frac{\mathbf{X}^\top}{\sigma^2}, \quad \mathbf{b} = -\mathbf{X}\boldsymbol{\beta} \quad \Rightarrow \quad U(\boldsymbol{\beta}) = \mathbf{A}(\boldsymbol{Y}+\mathbf{b})$$

$\Rightarrow$ According to the Gaussian linear model assumptions:

$$\boldsymbol{Y}|\mathbf{X} \sim MVN_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right) \quad \Rightarrow \quad U(\boldsymbol{\beta})|\mathbf{X} \sim MVN_{p+1}\left(\underbrace{\frac{\mathbf{X}^\top}{\sigma^2}\left[\mathbf{X}\boldsymbol{\beta}-\mathbf{X}\boldsymbol{\beta}\right]}_{\mathbf{0}_{p+1}}, \underbrace{\frac{\mathbf{X}^\top}{\sigma^2}\sigma^2\mathbf{I}_n\frac{\mathbf{X}}{\sigma^2}}_{\frac{\mathbf{X}^\top\mathbf{X}}{\sigma^2}=I(\boldsymbol{\beta})}\right)$$

Standardising the score function

Let

$I(\boldsymbol{\beta})^{-\frac{1}{2}}$ such that $I(\boldsymbol{\beta}) = I(\boldsymbol{\beta})^{\frac{1}{2}} I(\boldsymbol{\beta})^{\frac{1}{2}}$

$$I(\boldsymbol{\beta})^{\frac{1}{2}} I(\boldsymbol{\beta})^{-\frac{1}{2}} = I(\boldsymbol{\beta})^{-\frac{1}{2}} I(\boldsymbol{\beta})^{\frac{1}{2}} = \mathbf{I}_{p+1}$$

$$I(\boldsymbol{\beta})^{-\frac{1}{2}} U(\boldsymbol{\beta}) \sim MVN_{p+1}(\mathbf{0}_{p+1}, \mathbf{I}_{p+1})$$

Furthermore

$$U(\boldsymbol{\beta})^{\top} I(\boldsymbol{\beta})^{-1} U(\boldsymbol{\beta}) = \frac{(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})^{\top} \mathbf{X} \left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1} \mathbf{X}^{\top} (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2} \sim \chi^2_{p+1}$$

Note that:

$I(\boldsymbol{\beta})^{-\frac{1}{2}} = \sigma \left(\mathbf{X}^{\top}\mathbf{X}\right)^{-\frac{1}{2}}$ *exists if and only if the matrix* $\mathbf{X}^{\top}\mathbf{X}$ *is invertible, that is, if and only if* $\mathbf{X}$ *has full column rank*

Some general results

$L(\boldsymbol{\theta})$	Likelihood function associated with a given parametric statistical model ($\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k$)
$U(\boldsymbol{\theta}) = \dfrac{\partial}{\partial \boldsymbol{\theta}} \ln L(\boldsymbol{\theta})$	Score function

Under general regularity conditions:

- $\mathrm{E}[U(\boldsymbol{\theta})] = \mathbf{0}_k$
- $\mathrm{Var}[U(\boldsymbol{\theta})] = I(\boldsymbol{\theta})$
- $I(\boldsymbol{\theta})^{-\frac{1}{2}} U(\boldsymbol{\theta}) \xrightarrow{d} MVN_k(\mathbf{0}_k, \mathbf{I}_k)$

$\Rightarrow \quad U(\boldsymbol{\theta}) \approx MVN_k(\mathbf{0}_k, I(\boldsymbol{\theta}))$

Maximum likelihood estimate for β (1)

The vector $\hat{\mathbf{b}}$ is the maximum likelihood (ML) estimate for $\boldsymbol{\beta}$ if and only if

$$l\left(\hat{\mathbf{b}}, \sigma^2\right) = \max_{\mathbf{b} \in \mathbb{R}^{p+1}} l\left(\mathbf{b}, \sigma^2\right) \quad \text{or, equivalently} \quad \hat{\mathbf{b}} = \operatorname*{argmax}_{\mathbf{b} \in \mathbb{R}^{p+1}} l\left(\mathbf{b}, \sigma^2\right)$$

- $U\left(\hat{\mathbf{b}}\right) = \dfrac{\partial}{\partial \boldsymbol{\beta}} l\left(\boldsymbol{\beta}, \sigma^2\right)\Big|_{\boldsymbol{\beta}=\hat{\mathbf{b}}} = \mathbf{0}_{p+1}$

  log-likelihood gradient with respect to $\boldsymbol{\beta}$ evaluated at $\hat{\mathbf{b}}$

- $H\left(\hat{\mathbf{b}}\right) = \dfrac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} l\left(\boldsymbol{\beta}, \sigma^2\right)\Big|_{\boldsymbol{\beta}=\hat{\mathbf{b}}}$ negative definite

  log-likelihood Hessian matrix for $\boldsymbol{\beta}$ evaluated at $\hat{\mathbf{b}}$

  $\mathbf{z}^\top H\left(\hat{\mathbf{b}}\right) \mathbf{z} < 0 \quad \forall \mathbf{z} \neq \mathbf{0}_{p+1}$

Maximum likelihood estimate for β (2)

$$U(\mathbf{b}) = \mathbf{0}_{p+1} \iff \frac{\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\mathbf{b})}{\sigma^2} = \mathbf{0}_{p+1}$$

$$\iff \frac{\mathbf{X}^\top \mathbf{y}}{\sigma^2} = \frac{\mathbf{X}^\top \mathbf{X}\mathbf{b}}{\sigma^2}$$

if the matrix $\mathbf{X}$ has full column rank then $\mathbf{X}^\top \mathbf{X}$ is invertible:

$$\mathbf{b} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

Furthermore

$H(\boldsymbol{\beta}) = -i(\boldsymbol{\beta}) = -\dfrac{\mathbf{X}^\top \mathbf{X}}{\sigma^2} \quad \forall \mathbf{b} \in \mathbb{R}^{p+1}$ is negative definite, if the matrix $\mathbf{X}$ has full column rank

$$\Rightarrow \quad \hat{\mathbf{b}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

As far as $\boldsymbol{\beta}$ is concerned, maximum likelihood estimation is equivalent to least square estimation for Gaussian linear models

Properties of the maximum likelihood estimator for β

$\hat{\mathbf{b}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y} \Rightarrow$ depends on:

- $\mathbf{X}$ the regressor values
- $\mathbf{y}$ the observed values of the dependent variable (*realisations of the r. v.* $\boldsymbol{Y}$)

$\Rightarrow$ Conditionally on the regressors values, $\hat{\mathbf{b}}$ is the realisation of a random vector, that can be expressed as a linear transformation of $\boldsymbol{Y}$:

$$\mathbf{A} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top, \quad \mathbf{b} = \mathbf{0}_{p+1} \quad \Rightarrow \quad \hat{\boldsymbol{B}} = \mathbf{A}\boldsymbol{Y} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \boldsymbol{Y}$$

$\Rightarrow$ According to the Gaussian linear model assumptions:

$$\boldsymbol{Y}|\mathbf{X} \sim MVN_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n\right) \quad \Rightarrow \quad \mathrm{E}\left[\hat{\boldsymbol{B}}\middle|\mathbf{X}\right] = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

$$\begin{aligned} \mathrm{Var}\left[\hat{\boldsymbol{B}}\middle|\mathbf{X}\right] &= \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \sigma^2 \mathbf{I}_n \mathbf{X} \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \\ &= \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} = I\left(\boldsymbol{\beta}\right)^{-1} \end{aligned}$$

$$\hat{\boldsymbol{B}}|\mathbf{X} \sim MVN_{p+1}\left(\boldsymbol{\beta}, I\left(\boldsymbol{\beta}\right)^{-1}\right)$$

$\Rightarrow$ *The ML estimator for $\boldsymbol{\beta}$ in a Gaussian linear model is unbiased and efficient (according to the Rao-Cramer lower bound)*

Some general results

$\hat{\boldsymbol{T}}$	Maximum likelihood estimator for $\boldsymbol{\theta}$
$\hat{\mathbf{t}} = \mathrm{argmax}_{\mathbf{t}\in\Theta}\, l(\mathbf{t})$	Maximum likelihood estimatefor $\boldsymbol{\theta}$ (*sample realisation of* $\hat{\boldsymbol{T}}$)

general regularity conditions:

- $I(\boldsymbol{\theta})^{\frac{1}{2}} \left(\hat{\boldsymbol{T}} - \boldsymbol{\theta}\right) \xrightarrow{d} MVN_k\left(\mathbf{0}_k, \mathbf{I}_k\right)$

$\Rightarrow$ $\hat{\boldsymbol{T}} \approx MVN_k\left(\boldsymbol{\theta}, I(\boldsymbol{\theta})^{-1}\right)$

$\Rightarrow$ *Generally speaking, the ML estimator for $\boldsymbol{\theta}$ is asymptotically unbiased and asymptotically efficient, whatever functional form it takes*
(even when an explicit analytical form for computing $\hat{\mathbf{t}}$ does not exist)

Maximum likelihood estimate for σ^2

It is possible to prove that

$$\hat{s}^2 = \underset{s^2 \in \mathbb{R}^+}{\operatorname{argmax}}\, l\left(\hat{\mathbf{b}}, s^2\right) = \frac{\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^\top \hat{\mathbf{b}}\right)^2}{n} = \frac{\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right)^\top \left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right)}{n}$$

$$= \frac{\mathbf{e}^\top \mathbf{e}}{n}$$

$e_i = y_i - \mathbf{x}_i^\top \hat{\mathbf{b}} \quad i = 1, \ldots, n$ raw residuals

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\mathbf{X}^\top \mathbf{y} = \left[\mathbf{I}_n - \underbrace{\mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\mathbf{X}^\top}_{\mathbf{H}}\right]\mathbf{y} = \underbrace{[\mathbf{I}_n - \mathbf{H}]}_{\mathbf{M}}\mathbf{y}$$

Properties of raw residuals

Given the Gaussian linear model assumptions, it is possible to prove that:

- $\mathbf{e}|\mathbf{X} \sim MVN_n\left(\mathbf{0}_n, \sigma^2\mathbf{M}\right)$
- $\left.\frac{\mathbf{e}^\top \mathbf{e}}{\sigma^2}\right| \mathbf{X} \sim \chi^2_{n-p-1}$

$\Rightarrow \quad \mathrm{E}\left[\mathbf{e}^\top \mathbf{e}|\mathbf{X}\right] = \sigma^2(n-p-1)$

- $\frac{\mathbf{e}^\top \mathbf{e}}{\sigma^2}$ is independent of $\hat{\boldsymbol{B}}$

Properties of the maximum likelihood estimator for σ^2

$$\hat{S}^2 = \frac{\left(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{B}}\right)^\top \left(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{B}}\right)}{n}$$

Given the Gaussian linear model assumptions, and exploiting the properties of the raw residuals, it is possible to prove that:

- $\mathrm{E}\left[\hat{S}^2 \middle| \mathbf{X}\right] = \sigma^2 \frac{n-p-1}{n} \neq \sigma^2$

$\Rightarrow \quad \mathrm{E}\left[\hat{S}^2 \middle| \mathbf{X}\right] \underset{n\to\infty}{\longrightarrow} \sigma^2$

$\Rightarrow \quad S^2 = \frac{(\boldsymbol{Y}-\mathbf{X}\hat{\boldsymbol{B}})^\top (\boldsymbol{Y}-\mathbf{X}\hat{\boldsymbol{B}})}{n-p-1}$ *unbiased estimator for* σ^2

- $\hat{S}^2$ (and also S^2) is independent of $\hat{\boldsymbol{B}}$

Standardised residuals

$\mathbf{e}|\mathbf{X} \sim MVN_n\left(\mathbf{0}_n, \sigma^2\mathbf{M}\right)$

- In general:
  - the main diagonal elements of $\mathbf{M}$ differ from one another:

    $\mathbf{M}_{ii} = 1 - \mathbf{x}_i^\top \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{x}_i = 1 - \mathbf{H}_{ii}$
  - $\mathbf{M} = \mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$ is usually not diagonal and is not invertible
- Pearson residuals: $e_i^P = \frac{e_i}{\sqrt{s^2}} \quad i = 1, \ldots, n$
- Standardised residuals: $r_i = \frac{e_i}{\sqrt{s^2(1-\mathbf{H}_{ii})}} \quad i = 1, \ldots, n$

$\Rightarrow$ Given the Gaussian linear model assumptions, it is possible to prove that

$\mathbf{r} = (r_1, r_2, \ldots, r_n)^\top \Big| \mathbf{X} \xrightarrow{d} MVN_n\left(\mathbf{0}_n, \mathbf{I}_n\right)$

Approximately, standardised residuals from a Gaussian linear models are equivalent to an observed sample drawn from an $n-$*dimensional standardised Gaussian random vector*

Linear hypotheses on the parameters of Gaussian linear models: constrained estimation and test statistics

Linear hypotheses on β

$\boldsymbol{Y}|\mathbf{X} \sim NMV_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right)$ Gaussian linear model

$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$ $(p+1) \times 1$ parameter vector

$\mathbf{K}$ $q \times (p+1)$ matrix composed of known constants with full row rank q

$\mathbf{t}$ $q \times 1$ vector composed of known constants

$H_0 : \mathbf{K}\boldsymbol{\beta} = \mathbf{t}$ System of linear hypotheses on $\boldsymbol{\beta}$

Linear hypotheses on β: some examples

$p = 3$

(A) $\mathbf{K} = [0\ 1\ 0\ 0]$, $\mathbf{t} = 0 \Rightarrow H_0 : \beta_1 = 0$

(B) $\mathbf{K} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$, $\mathbf{t} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$\Rightarrow H_0 : \begin{cases} \beta_1 = 0 \\ \beta_3 = 0 \end{cases}$ or $\beta_1 = \beta_3 = 0$

(C) $\mathbf{K} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$, $\mathbf{t} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

$\Rightarrow H_0 : \begin{cases} \beta_1 = 0 \\ \beta_2 = 0 \\ \beta_3 = 0 \end{cases}$ or $\beta_1 = \beta_2 = \beta_3 = 0$ *linear independence*

(D) $\mathbf{K} = [0\ 1\ 0\ -1]$, $\mathbf{t} = 0 \Rightarrow H_0 : \beta_1 = \beta_3$

(E) $\mathbf{K} = [0\ 1\ 0\ 0]$, $\mathbf{t} = 3 \Rightarrow H_0 : \beta_1 = 3$

Nested linear models

Linear hypotheses (A), (B) and (C) lead to Gaussian linear models that can be obtained by removing some regressors from the starting model:

Starting model: $\mathrm{E}\left[Y_i|x_{1i}, x_{2i}, x_{3i}\right] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$

(A) $\Rightarrow \mathrm{E}_{H_0}\left[Y_i|x_{1i}, x_{2i}, x_{3i}\right] = \beta_0 + \beta_2 x_{2i} + \beta_3 x_{3i} = \mathrm{E}\left[Y_i|x_{2i}, x_{3i}\right]$

(B) $\Rightarrow \mathrm{E}_{H_0}\left[Y_i|x_{1i}, x_{2i}, x_{3i}\right] = \beta_0 + \beta_2 x_{2i} = \mathrm{E}\left[Y_i|x_{2i}\right]$

(C) $\Rightarrow \mathrm{E}_{H_0}\left[Y_i|x_{1i}, x_{2i}, x_{3i}\right] = \beta_0 = \mathrm{E}\left[Y_i\right]$

Likelihood ratio test statistics - 1

$\dfrac{L\left(\hat{\mathbf{b}}, \sigma^2\right)}{L\left(\hat{\mathbf{b}}_{H_0}, \sigma^2\right)}$ or, equivalently, $2\left[l\left(\hat{\mathbf{b}}, \sigma^2\right) - l\left(\hat{\mathbf{b}}_{H_0}, \sigma^2\right)\right]$

where:

$\hat{\mathbf{b}} = \underset{\mathbb{R}^{(p+1)}}{\operatorname{argmax}}\, l\left(\mathbf{b}, \sigma^2\right)$

$\hat{\mathbf{b}}_{H_0} = \underset{\{\mathbf{b}:\mathbf{K}\mathbf{b}=\mathbf{t}\}\subset\mathbb{R}^{(p+1)}}{\operatorname{argmax}}\, l\left(\mathbf{b}, \sigma^2\right)$

The Method of Lagrange multipliers

$\hat{\mathbf{b}}_{H_0}$ maximises $l\left(\boldsymbol{\beta}, \sigma^2\right)$ in the parameter subspace $\{\mathbf{b} : \mathbf{K}\mathbf{b} = \mathbf{t}\} \subset \mathbb{R}^{(p+1)}$. The following function can be maximised to obtain it:

$$l^*\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\alpha}\right) = l\left(\boldsymbol{\beta}, \sigma^2\right) - \boldsymbol{\alpha}^\top \left(\mathbf{K}\boldsymbol{\beta} - \mathbf{t}\right)$$

where $\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_q \end{bmatrix}$ $q \times 1$ vector containing (unknown) *Lagrange multipliers*

Constrained maximisation

$l^*\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\alpha}\right)$ must be maximised with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$

$\Rightarrow$ The following system with $p + 1 + q$ equations must be solved:

$$\begin{cases} U\left(\mathbf{b}\right) = \dfrac{\partial}{\partial \boldsymbol{\beta}} l^*\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\alpha}\right)\Big|_{\boldsymbol{\beta}=\mathbf{b}} = \mathbf{0}_{p+1} \\ U\left(\mathbf{a}\right) = \dfrac{\partial}{\partial \boldsymbol{\alpha}} l^*\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\alpha}\right)\Big|_{\boldsymbol{\alpha}=\mathbf{a}} = \mathbf{0}_{q} \end{cases}$$

First partial derivatives

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}} l^* \left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\alpha}\right) &= U\left(\boldsymbol{\beta}\right) - \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\alpha}^\top \left(\mathbf{K}\boldsymbol{\beta} - \mathbf{t}\right) \\
&= \frac{\mathbf{X}^\top \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)}{\sigma^2} - \mathbf{K}^\top \boldsymbol{\alpha} \\
\frac{\partial}{\partial \boldsymbol{\alpha}} l^* \left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\alpha}\right) &= -\frac{\partial}{\partial \boldsymbol{\alpha}} \boldsymbol{\alpha}^\top \left(\mathbf{K}\boldsymbol{\beta} - \mathbf{t}\right) \\
&= -\left(\mathbf{K}\boldsymbol{\beta} - \mathbf{t}\right)
\end{aligned}$$

General rules: $\frac{\partial}{\partial \boldsymbol{\delta}} \mathbf{A}\boldsymbol{\delta} = \mathbf{A}^\top$ e $\frac{\partial}{\partial \boldsymbol{\delta}} \boldsymbol{\delta}^\top \mathbf{A} = \mathbf{A}$

Solutions - 1

Consider the first $p+1$ equations:

$$\frac{\mathbf{X}^\top \left(\mathbf{y} - \mathbf{X}\mathbf{b}\right)}{\sigma^2} - \mathbf{K}^\top \mathbf{a} = \mathbf{0}_{p+1}$$

$$\Updownarrow$$

$$\frac{\mathbf{X}^\top \left(\mathbf{y} - \mathbf{X}\mathbf{b}\right)}{\sigma^2} = \mathbf{K}^\top \mathbf{a}$$

$$\Updownarrow$$

$$\mathbf{X}^\top \mathbf{y} - \mathbf{X}^\top \mathbf{X}\mathbf{b} = \sigma^2 \mathbf{K}^\top \mathbf{a}$$

$$\Updownarrow$$

$$\mathbf{X}^\top \mathbf{X}\mathbf{b} = \mathbf{X}^\top \mathbf{y} - \sigma^2 \mathbf{K}^\top \mathbf{a}$$

$\Updownarrow$ if $\mathbf{X}$ has full column rank $(p+1)$

$$\hat{\mathbf{b}}_{H_0} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y} - \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{K}^\top \mathbf{a} = \hat{\mathbf{b}} - \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{K}^\top \mathbf{a}$$

Note that σ^2 and $\mathbf{a}$ are unknown

Solutions - 2

Exploiting the formula for $\hat{\mathbf{b}}_{H_0}$ in the last q equations:

$$\mathbf{K}\left[\hat{\mathbf{b}} - \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{K}^\top \mathbf{a}\right] = \mathbf{t}$$

$$\Updownarrow$$

$$\mathbf{K}\hat{\mathbf{b}} - \sigma^2 \mathbf{K}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{K}^\top \mathbf{a} = \mathbf{t}$$

$$\Updownarrow$$

$$\sigma^2 \mathbf{K}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{K}^\top \mathbf{a} = \mathbf{K}\hat{\mathbf{b}} - \mathbf{t}$$

$\Updownarrow$ if $\mathbf{K}$ has full row rank (q)

$$\hat{\mathbf{a}} = \frac{1}{\sigma^2}\left[\mathbf{K}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{K}^\top\right]^{-1}\left(\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right)$$

Constrained maximum likelihood estimate

Substituting $\hat{\mathbf{a}}$ for $\boldsymbol{\alpha}$ in the formula for $\hat{\mathbf{b}}_{H_0}$:

$$\begin{aligned}\hat{\mathbf{b}}_{H_0} &= \hat{\mathbf{b}} - \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{K}^\top \frac{1}{\sigma^2}\left[\mathbf{K}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{K}^\top\right]^{-1}\left(\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right) \\ &= \hat{\mathbf{b}} - \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{K}^\top \left[\mathbf{K}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{K}^\top\right]^{-1}\left(\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right)\end{aligned}$$

Note that, even the constrained maximum likelihood estimate $\hat{\mathbf{b}}_{H_0}$ can be computed without knowing the true value of σ^2

Furthermore:

$$\begin{aligned}\mathbf{K}\hat{\mathbf{b}}_{H_0} &= \mathbf{K}\hat{\mathbf{b}} - \underbrace{\mathbf{K}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{K}^\top \left[\mathbf{K}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{K}^\top\right]^{-1}}_{\mathbf{I}_q}\left(\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right) \\ &= \mathbf{K}\hat{\mathbf{b}} - \mathbf{K}\hat{\mathbf{b}} + \mathbf{t} = \mathbf{t}\end{aligned}$$

Residuals of the constrained model - 1

$$
\begin{aligned}
\mathbf{e}_{H_0} &= \mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_{H_0} \\
&= \mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_{H_0} - \mathbf{X}\hat{\mathbf{b}} + \mathbf{X}\hat{\mathbf{b}} \\
&= \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} + \mathbf{X}\left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right) = \mathbf{e} + \mathbf{X}\left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right) \\
\mathbf{e}_{H_0}^\top \mathbf{e}_{H_0} &= \left[\mathbf{e} + \mathbf{X}\left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right)\right]^\top \left[\mathbf{e} + \mathbf{X}\left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right)\right] \\
&= \mathbf{e}^\top\mathbf{e} + \mathbf{e}^\top\mathbf{X}\left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right) + \left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right)^\top \mathbf{X}^\top\mathbf{e} + \\
&\quad + \left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right)^\top \mathbf{X}^\top\mathbf{X}\left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right)
\end{aligned}
$$

Residuals of constrained models - 2

As

$$
\begin{aligned}
\mathbf{X}^\top\mathbf{e} &= \mathbf{X}^\top\left[\mathbf{y} - \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}\right] \\
&= \mathbf{X}^\top\mathbf{y} - \underbrace{\mathbf{X}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}}_{\mathbf{I}_{p+1}}\mathbf{X}^\top\mathbf{y} = \mathbf{0}_{p+1}
\end{aligned}
\qquad \mathbf{e}_{H_0}^\top\mathbf{e}_{H_0} = \mathbf{e}^\top\mathbf{e} + \left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right)^\top \mathbf{X}^\top\mathbf{X}\left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right)
$$

if $\mathbf{X}$ has full column rank, $\mathbf{X}^\top\mathbf{X}$ is positive definite and:

$$
\begin{aligned}
\hat{\mathbf{b}} \neq \hat{\mathbf{b}}_{H_0} &\Rightarrow \left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right)^\top \mathbf{X}^\top\mathbf{X}\left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right) > 0 \\
&\Rightarrow \mathbf{e}_{H_0}^\top\mathbf{e}_{H_0} > \mathbf{e}^\top\mathbf{e}
\end{aligned}
$$

Residuals of constrained models - 3

As

$$\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\left[\mathbf{K}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\right]^{-1}\left(\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right)$$

$$\begin{aligned}
\mathbf{e}_{H_0}^\top\mathbf{e}_{H_0} - \mathbf{e}^\top\mathbf{e} &= \left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right)^\top\mathbf{X}^\top\mathbf{X}\left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right) \\
&= \left(\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right)^\top\left[\mathbf{K}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\right]^{-1}\mathbf{K}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\cdot \\
&\quad \cdot\mathbf{X}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\left[\mathbf{K}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\right]^{-1}\left(\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right) \\
&= \left(\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right)^\top\left[\mathbf{K}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\right]^{-1}\left(\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right)
\end{aligned}$$

Likelihood ratio properties 15

Likelihood ratio test statistics - 2

$$\begin{aligned}
\Delta l = 2\ln\left[\frac{L\left(\hat{\mathbf{b}}, \sigma^2\right)}{L\left(\hat{\mathbf{b}}_{H_0}, \sigma^2\right)}\right] &= -n\ln 2\pi\sigma^2 - \frac{\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}\right)^\top\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}\right)}{\sigma^2} + \\
&\quad + n\ln 2\pi\sigma^2 + \frac{\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}_{H_0}\right)^\top\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}_{H_0}\right)}{\sigma^2} \\
&= \frac{\mathbf{e}_{H_0}^\top\mathbf{e}_{H_0} - \mathbf{e}^\top\mathbf{e}}{\sigma^2} \\
&= \frac{\left(\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right)^\top\left[\mathbf{K}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\right]^{-1}\left(\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right)}{\sigma^2}
\end{aligned}$$

$\Rightarrow$ *It is not necessary to know* $\hat{\mathbf{b}}_{H_0}$ *in order to compute the LR test statistic and to derive its distribution*

Likelihood ratio test statistic distribution - σ^2 known

Properties of the maximum likelihood estimator for $\boldsymbol{\beta}$:

$\hat{\boldsymbol{B}} \sim MVN_{p+1}\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\right)$

$\Downarrow$ *Linear transformation*

$\mathbf{K}\hat{\boldsymbol{B}} - \mathbf{t} \sim MVN_q\left(\mathbf{K}\boldsymbol{\beta} - \mathbf{t}, \sigma^2 \mathbf{K}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{K}^\top\right)$

$\Downarrow$ *if H_0 is true*

$\mathbf{K}\hat{\boldsymbol{B}} - \mathbf{t} | H_0 \sim MVN_q\left(\mathbf{0}_q, \sigma^2 \mathbf{K}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{K}^\top\right)$

$\Downarrow$ *Standardisation*

$\boldsymbol{Z} = \frac{1}{\sigma}\left[\mathbf{K}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{K}^\top\right]^{-\frac{1}{2}} \left(\mathbf{K}\hat{\boldsymbol{B}} - \mathbf{t}\right) | H_0 \sim NMV_q\left(\mathbf{0}_q, \mathbf{I}_q\right)$

$\Downarrow$ *Quadratic form (sum of squares)*

$\boldsymbol{Z}^\top \boldsymbol{Z} = \frac{\left(\mathbf{K}\hat{\boldsymbol{B}} - \mathbf{t}\right)^\top \left[\mathbf{K}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{K}^\top\right]^{-1} \left(\mathbf{K}\hat{\boldsymbol{B}} - \mathbf{t}\right)}{\sigma^2} = \Delta l | H_0 \sim \chi^2_q$

Likelihood ratio test statistic distribution - σ^2 unknown

Proprieties of raw residuals:

$\frac{\mathbf{e}^\top \mathbf{e}}{\sigma^2} \sim \chi^2_{n-p-1}$ & independence between $\hat{\boldsymbol{B}}$ and S^2

$\Downarrow$ *If H_0 is true*

$$\frac{\frac{\Delta l}{\frac{\mathbf{e}^\top \mathbf{e}}{\sigma^2}} \frac{n-p-1}{q}}{} = \frac{\frac{\mathbf{e}_{H_0}^\top \mathbf{e}_{H_0} - \mathbf{e}^\top \mathbf{e}}{\sigma^2}}{\frac{\mathbf{e}^\top \mathbf{e}}{\sigma^2}} \frac{n-p-1}{q}$$

$$= \frac{\frac{\left(\mathbf{K}\hat{\boldsymbol{B}} - \mathbf{t}\right)^\top \left[\mathbf{K}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{K}^\top\right]^{-1} \left(\mathbf{K}\hat{\boldsymbol{B}} - \mathbf{t}\right)}{q}}{\frac{\mathbf{e}^\top \mathbf{e}}{n-p-1}} | H_0 \sim \frac{\frac{\chi^2_q}{q}}{\frac{\chi^2_{n-p-1}}{n-p-1}} = F_{(q, n-p-1)}$$

Comparison between complete and reduced models

When linear hypotheses lead to the removal of q regressors, raw residuals $\mathbf{e}_{H_0}$ correspond to the residuals of a reduced model (nested in the complete model):

$$\frac{\Delta l}{\frac{\mathbf{e}^\top \mathbf{e}}{\sigma^2}} \frac{n-p-1}{q} = \frac{\frac{SSE_{M_{H_0}} - SSE_{M_C}}{q}}{\frac{SSE_{M_C}}{n-p-1}}$$

dove:

SSE_{M_C} Residual sum of squares for the complete model (all regressors)

$SSE_{M_{H_0}}$ Residual sum of squares for the reduced model (after excluding q regressors)

Wald test statistics

It is possible to prove that, for hypotheses such $H_0 : \beta_j = 0$:

$$\Delta l = \frac{\hat{B}_j^2}{\sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}_{jj}} = \left[\frac{\hat{B}_j}{\sigma \sqrt{(\mathbf{X}^\top \mathbf{X})^{-1}_{jj}}} \right]^2$$

$(\mathbf{X}^\top \mathbf{X})^{-1}_{jj}$: $j-$th element on the main diagonal of $(\mathbf{X}^\top \mathbf{X})^{-1}$

σ^2 *known* $\Rightarrow \frac{\hat{B}_j}{\sigma \sqrt{(\mathbf{X}^\top \mathbf{X})^{-1}_{jj}}} | H_0 \sim N(0,1)$

σ^2 *unknown* $\Rightarrow \frac{\hat{B}_j}{S \sqrt{(\mathbf{X}^\top \mathbf{X})^{-1}_{jj}}} | H_0 \sim t_{n-p-1}$

Recall the link between t *and* F: $[t_{n-p-1}]^2 = F_{(1,n-p-1)}$

Confidence intervals

$\frac{\hat{B}_j - \beta_j}{\sigma\sqrt{(\mathbf{X}^\top\mathbf{X})^{-1}_{jj}}} \sim N(0,1)$ pivotal quantity for β_j

Gaussian intervals (σ^2 known) at a $1-\alpha$ confidence level:

$$\left[\hat{b}_j - z_{\frac{\alpha}{2}}\sqrt{\sigma^2\,(\mathbf{X}^\top\mathbf{X})^{-1}_{jj}}, \hat{b}_j + z_{\frac{\alpha}{2}}\sqrt{\sigma^2\,(\mathbf{X}^\top\mathbf{X})^{-1}_{jj}}\right]$$

Student-t intervals (σ^2 unknown) at a $1-\alpha$ confidence level:

$$\left[\hat{b}_j - t_{\frac{\alpha}{2},n-p-1}\sqrt{s^2\,(\mathbf{X}^\top\mathbf{X})^{-1}_{jj}}, \hat{b}_j + t_{\frac{\alpha}{2},n-p-1}\sqrt{s^2\,(\mathbf{X}^\top\mathbf{X})^{-1}_{jj}}\right]$$

Gaussian linear models:
use of categorical polytomous regressors

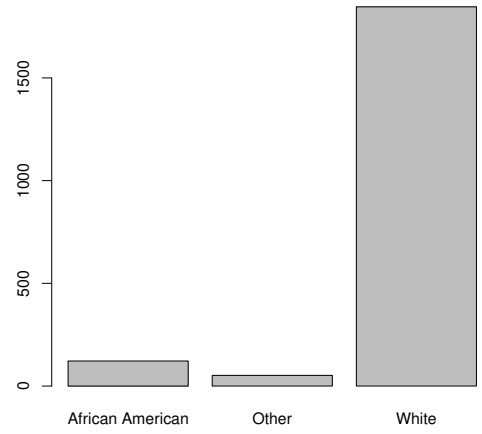
Glucose level in blood and ethnic origin

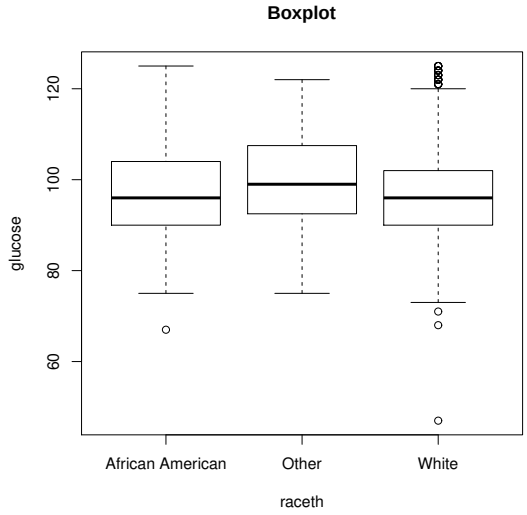
A glucose level in blood between 100 and 125 mg/dL represents a risk factor for developing diabetes.

Aim: Are there systematic differences in the glucose level among people with different ethnic origins?

available information: Glucose level and ethnic origin (White/African American/other) on a sample of 2020 women not affected by diabetes after menopause

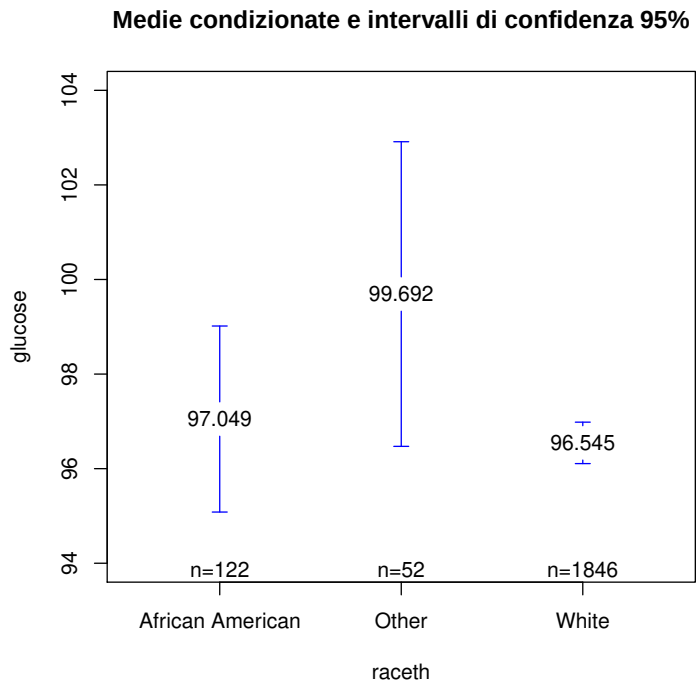
Some graphical displays - 1





Most of the women in the sample are Caucasian

Some graphical displays - 2



Hypothesis of interest

Abscence of significant differences in the average glucose level among different ethnic groups

$$H_0: \quad \mathrm{E}\left[\texttt{glucose}_i | \texttt{raceth}_i = \texttt{White}\right] = \mathrm{E}\left[\texttt{glucose}_i | \texttt{raceth}_i = \texttt{Other}\right] = \\ = \mathrm{E}\left[\texttt{glucose}_i | \texttt{raceth}_i = \texttt{African American}\right]$$

Which inferential tool?

- One-way ANOVA
- Gaussian linear models

One-way ANOVA

```
> summary(aov(glucose ~ raceth,data=hers.nod))

              Df  Sum Sq  Mean Sq  F value  Pr(>F)
raceth         2     521   260.51    2.747  0.0643
Residuals   2017  191259    94.82
```


Numeric coding through indicator/dummy variables

	x_{Ai}	x_{Oi}	x_{Wi}
	African American$_i$	Other$_i$	White$_i$
raceth$_i$ = African American	1	0	0
raceth$_i$ = Other	0	1	0
raceth$_i$ = White	0	0	1

Note that:

- 3 indicator variables allow to code a qualitative regressor with 3 categories
  $\Rightarrow$ it is necessary to consider a multiple linear regression model
- These 3 indicator variables sum up to 1, for any sample unit: $x_{Ai} + x_{Oi} + x_{Wi} = 1$

  $\Rightarrow$ if they are included in a linear model along with an intercept term, the corresponding regressor matrix $\mathbf{X}$ will not have full column rank

  $\Rightarrow$ it is necessary to exclude one of the indicator variables. The corresponding category is termed *baseline/reference category*

Linear regression with a qualitative regressor

$\mathrm{E}\left[\texttt{glucose}_i|\texttt{raceth}_i\right] = \beta_0 + \beta_1\texttt{Other}_i + \beta_2\texttt{White}_i \qquad i = 1, \ldots, 2020$

$\Rightarrow \quad \mathrm{E}\left[\texttt{glucose}_i|\texttt{raceth}_i = \texttt{African American}\right] = \beta_0$

$\Rightarrow \quad \mathrm{E}\left[\texttt{glucose}_i|\texttt{raceth}_i = \texttt{Other}\right] = \beta_0 + \beta_1$

$\Rightarrow \quad \mathrm{E}\left[\texttt{glucose}_i|\texttt{raceth}_i = \texttt{White}\right] = \beta_0 + \beta_2$

Each regression coefficient represents the difference between the conditional expected value given the corresponding category and the conditional expected value given the reference category

Hypothesis of interest - 2

$H_0: \quad \mathrm{E}\left[\texttt{glucose}_i|\texttt{raceth}_i = \texttt{White}\right] = \mathrm{E}\left[\texttt{glucose}_i|\texttt{raceth}_i = \texttt{Other}\right] = \\ = \mathrm{E}\left[\texttt{glucose}_i|\texttt{raceth}_i = \texttt{African American}\right]$

Abscence of significant differences in the average glucose level among different ethnic groups

$$\Longrightarrow H_0: \begin{cases} \beta_0 = \beta_0 + \beta_1 \\ \beta_0 = \beta_0 + \beta_2 \\ (\beta_0 + \beta_1 = \beta_0 + \beta_2) \end{cases} \quad \text{or, equivalently } H_0: \beta_1 = \beta_2 = 0$$

The regression coefficients for the two indicator variables $\texttt{Other}_i$ *and* $\texttt{White}_i$ *are not significantly different from 0*

Gaussian linear regression model: results

```
 > summary(modello1)
Call:
lm(formula = glucose ~ raceth, data = hers.nod)

...
 Coefficients:
                Estimate   Std.  Error   t value   Pr(>|t|)
 (Intercept)     97.0492        0.8816   110.081     <2e-16
 racethOther      2.6431        1.6127     1.639      0.101
 racethWhite     -0.5042        0.9103    -0.554      0.580

Residual standard error:  9.738 on 2017 degrees of freedom
Multiple R-squared:  0.002717, Adjusted R-squared:  0.001728
F-statistic:  2.747 on 2 and 2017 DF, p-value:  0.06434
```

the relevant test statistic is the F test statistic

t test statistics allow to evaluate differences between each category and the reference category

Linear hypothesis - 1

```
> K1
     1   2   3
 1   0   1   0
 2   0   0   1
> t1
[1] 0 0
```

F test - 1

```
> linearHypothesis(modello1,K1,t1,test="F")

Linear hypothesis test

Hypothesis:
racethOther = 0
racethWhite = 0

Model 1:  restricted model
Model 2:  glucose ~ raceth

     Res.Df      RSS   Df   Sum of Sq        F   Pr(>F)
 1     2019   191780
 2     2017   191259    2      521.02   2.7473   0.06434
```


Choice of the reference category

The default choice in R is the first category, in alphabetical order:

```
                    Other  White
  African American      0      0
             Other      1      0
             White      0      1
```

Change of the reference category

```
                   1  2
  African American  1  0
             Other  0  1
             White  0  0
```

The meaning of the regression coefficients changes accordingly:

$$\mathrm{E}\left[\texttt{glucose}_i|\texttt{raceth}_i\right] = \delta_0 + \delta_1\texttt{raceth1}_i + \delta_2\texttt{raceth2}_i + \varepsilon_i \qquad i = 1, \ldots, 2020$$

$$\Rightarrow \quad \mathrm{E}\left[\texttt{glucose}_i|\texttt{raceth}_i = \texttt{African American}\right] = \delta_0 + \delta_1$$

$$\Rightarrow \quad \mathrm{E}\left[\texttt{glucose}_i|\texttt{raceth}_i = \texttt{Other}\right] = \delta_0 + \delta_2$$

$$\Rightarrow \quad \mathrm{E}\left[\texttt{glucose}_i|\texttt{raceth}_i = \texttt{White}\right] = \delta_0$$

Change of the reference category: results

```
 > summary(modello2)
Call:
lm(formula = glucose ~ raceth, data = hers.nod)

...
 Coefficients:
               Estimate   Std.  Error   t value   Pr(>|t|)
 (Intercept)     96.5450        0.2266   425.979     <2e-16
 raceth1          0.5042        0.9103     0.554     0.5797
 raceth2          3.1473        1.3693     2.299     0.0216

Residual standard error:  9.738 on 2017 degrees of freedom
Multiple R-squared:  0.002717, Adjusted R-squared:  0.001728
F-statistic:  2.747 on 2 and 2017 DF, p-value:  0.06434
```

The choice of the reference category is arbitrary:

⇒ *the estimates for the regression coefficients have changed, but the global measures remained the same*

Exclusion of the intercept

If one consider a regression model without intercept, it is possible to include all the 3 indicator variables (without choosing a reference category)

$\mathrm{E}\left[\texttt{glucose}_i|\texttt{raceth}_i\right] = \mu_1\texttt{African American}_i + \mu_2\texttt{Other}_i + \mu_3\texttt{White}_i$

$i = 1, \ldots, 2020$

$\Rightarrow \quad \mathrm{E}\left[\texttt{glucose}_i|\texttt{raceth}_i = \texttt{African American}\right] = \mu_1$

$\Rightarrow \quad \mathrm{E}\left[\texttt{glucose}_i|\texttt{raceth}_i = \texttt{Other}\right] = \mu_2$

$\Rightarrow \quad \mathrm{E}\left[\texttt{glucose}_i|\texttt{raceth}_i = \texttt{White}\right] = \mu_3$

$\Rightarrow \quad H_0 : \mu_1 = \mu_2 = \mu_3 \Longrightarrow H_0 : \begin{cases} \mu_1 = \mu_2 \\ \mu_1 = \mu_3 \end{cases}$

Exclusion of the intercept: results

```
 > summary(modello3)
Call:
lm(formula = glucose ~ raceth - 1, data = hers.nod)

...
 Coefficients:
                          Estimate   Std.  Error   t value   Pr(>|t|)
 racethAfrican American    97.0492        0.8816    110.08     <2e-16
 racethOther               99.6923        1.3504     73.83     <2e-16
 racethWhite               96.5450        0.2266    425.98     <2e-16

Residual standard error:  9.738 on 2017 degrees of freedom
Multiple R-squared:  0.99, Adjusted R-squared:  0.99
F-statistic:  6.634e+04 on 3 and 2017 DF, p-value:  < 2.2e-16
```

WARNING:
⇒ *In this setting the function* `lm` *computes* R^2 *using* $\sum_{i=1}^{n} y_i^2$ *as denominator*
⇒ *the* F *test statistic is referred to the null hypothesis* $H_0 : \mu_1 = \mu_2 = \mu_3 = 0$

Linear hypothesis - 2

```
> K3
     1   2   3
 1   1  -1   0
 2   1   0  -1
> t3
[1] 0 0
```


F test - 2

```
> linearHypothesis(modello3,K3,t3,test="F")
Linear hypothesis test

Hypothesis:
racethAfrican American - racethOther = 0
racethAfrican American - racethWhite = 0

Model 1:  restricted model
Model 2:  glucose ~ raceth

     Res.Df      RSS   Df   Sum of Sq        F   Pr(>F)
 1     2019   191780
 2     2017   191259    2     521.02   2.7473   0.06434
```

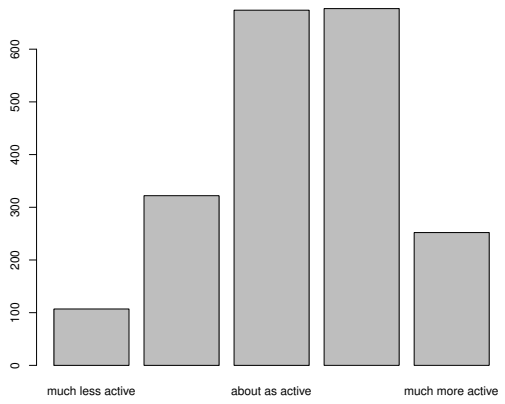
Glucose level in blood and physical activity

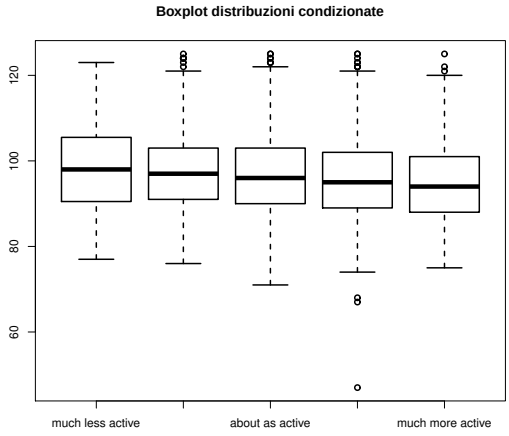
A glucose level in blood between 100 and 125 mg/dL represents a risk factor for developing diabetes.

Aim: Does physical activity (a modifiable factor related to life style) contribute to the reduction of the glucose level, thus preventing a severe disease?

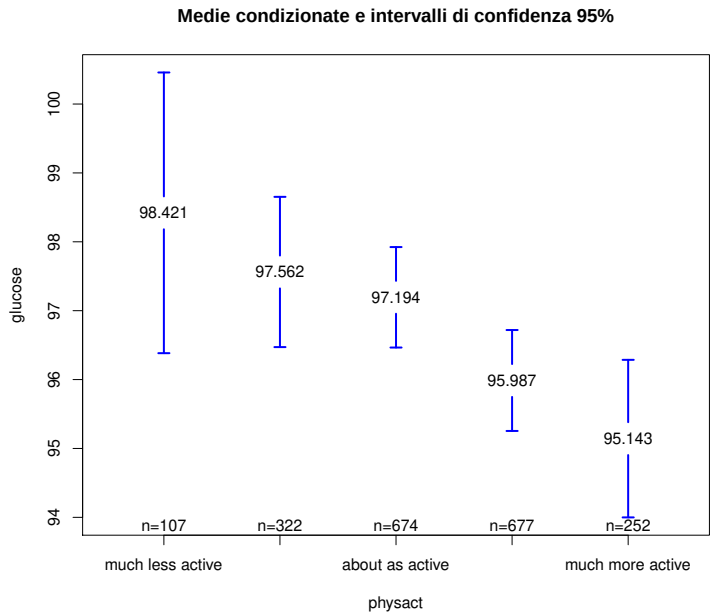
Available information:
Glucose level and physical activity level (`much less active`, `somewhat less active`, `about as active`, `somewhat more active`, `much more active`) on a sample of 2032 women not affected by diabetes after menopause

Some graphical displays - 1





Some graphical displays - 2



Gaussian linear model - 1

```
> physact1<-lm(glucose~physact,data=hers.nod)
> summary(physact1)

...
                              Estimate   Std.  Error   t value   Pr(>|t|)
 (Intercept)                    98.421         0.939   104.784      0.000
 physactsomewhat less active    -0.858         1.084    -0.792      0.429
 physactabout as active         -1.226         1.011    -1.213      0.225
 physactsomewhat more active    -2.434         1.011    -2.408      0.016
 physactmuch more active        -3.278         1.121    -2.924      0.003
...
Residual standard error:  9.716 on 2027 degrees of freedom
Multiple R-squared:  0.008668, Adjusted R-squared:  0.006712
F-statistic:  4.431 on 4 and 2027 DF, p-value:  0.001441
```

Reference category: `much less active`

Linear hypothesis - 1

```
> K1
     1  2  3  4  5
 1   0  1  0  0  0
 2   0  0  1  0  0
 3   0  0  0  1  0
 4   0  0  0  0  1
> t1
[1] 0 0 0 0
```


F test - 1

```
> linearHypothesis(physact1,K1,t1,test="F")
Linear hypothesis test
Hypothesis:
physactsomewhat less active = 0
physactabout as active = 0
physactsomewhat more active = 0
physactmuch more active = 0
Model 1:  restricted model
Model 2:  glucose ~ physact
     Res.Df        RSS  Df  Sum of Sq     F  Pr(>F)
 1     2031  193017.70
 2     2027  191344.61   4    1673.09  4.43  0.0014
```

Incremental/split coding - 1

An alternative coding scheme can be used if there is a "natural" order among the categories

	x_{Bi}	x_{Ci}	x_{Di}	x_{Ei}
much less active	0	0	0	0
somewhat less active	1	0	0	0
about as active	1	1	0	0
somewhat more active	1	1	1	0
much more active	1	1	1	1

It is possible to show that these alternative indicator variables can be obtained by summing subsets of the indicaror variables introduced above

Incremental/split coding - 2

$\mathrm{E}\left[\texttt{glucose}_i|\texttt{physact}_i\right] = \beta_0 + \beta_B x_{Bi} + \beta_C x_{Ci} + \beta_D x_{Di} + \beta_E x_{Ei} \quad i = 1, \ldots, 2032$

$$
\begin{aligned}
\mathrm{E}\left[\texttt{glucose}_i|\texttt{physact}_i = \texttt{much less active}\right] &= \beta_0 \\
\mathrm{E}\left[\texttt{glucose}_i|\texttt{physact}_i = \texttt{somewhat less active}\right] &= \beta_0 + \beta_B \\
\mathrm{E}\left[\texttt{glucose}_i|\texttt{physact}_i = \texttt{about as active}\right] &= \beta_0 + \beta_B + \beta_C \\
\mathrm{E}\left[\texttt{glucose}_i|\texttt{physact}_i = \texttt{somewhat more active}\right] &= \beta_0 + \beta_B + \beta_C + \beta_D \\
\mathrm{E}\left[\texttt{glucose}_i|\texttt{physact}_i = \texttt{much more active}\right] &= \beta_0 + \beta_B + \beta_C + \beta_D + \beta_E
\end{aligned}
$$

Each regression coefficient represents the difference between the conditional expected values associated with two consecutive categories

Guassian linear model - 2

```
>physact2 <-lm(glucose~physact,data=hers.nod)
> summary(physact2)

...
              Estimate   Std.  Error   t value   Pr(>|t|)
 (Intercept)    98.421          0.939   104.784      0.000
 physactd1      -0.858          1.084    -0.792      0.429
 physactd2      -0.368          0.658    -0.559      0.576
 physactd3      -1.208          0.529    -2.284      0.022
 physactd4      -0.844          0.717    -1.177      0.239
...
Residual standard error:  9.716 on 2027 degrees of freedom
Multiple R-squared:  0.008668, Adjusted R-squared:  0.006712
F-statistic:  4.431 on 4 and 2027 DF, p-value:  0.001441
```


Linear hypothesis - 2

```
> K2
     1   2   3   4   5
 1   0   1   0   0   0
 2   0   0   1   0   0
 3   0   0   0   1   0
 4   0   0   0   0   1
> t2
[1] 0 0 0 0
```

F test - 2

```
> linearHypothesis(physact2,K2,t2,test="F")

Linear hypothesis test

Hypothesis:
physactd1 = 0
physactd2 = 0
physactd3 = 0
physactd4 = 0

Model 1:  restricted model
Model 2:  glucose ~ physact

     Res.Df         RSS   Df   Sum of Sq      F   Pr(>F)
 1     2031   193017.70
 2     2027   191344.61    4     1673.09   4.43   0.0014
```


Linear trend hypothesis

The introduction of suitable linear constraints in the regression coefficients associated with the incremental coding scheme allows to test for the existence of a linear trend in the conditional expected values:

$H_0 : \beta_B = \beta_C = \beta_D = \beta_E = \beta (\neq 0)$

$$
\begin{aligned}
&\mathrm{E}\left[\texttt{glucose}_i | \texttt{physact}_i = \texttt{much less active}\right] | H_0 && = \beta_0 + 0 \cdot \beta = \beta_0 \\
&\mathrm{E}\left[\texttt{glucose}_i | \texttt{physact}_i = \texttt{somewhat less active}\right] | H_0 && = \beta_0 + 1 \cdot \beta \\
&\mathrm{E}\left[\texttt{glucose}_i | \texttt{physact}_i = \texttt{about as active}\right] | H_0 && = \beta_0 + 2 \cdot \beta \\
&\mathrm{E}\left[\texttt{glucose}_i | \texttt{physact}_i = \texttt{somewhat as active}\right] | H_0 && = \beta_0 + 3 \cdot \beta \\
&\mathrm{E}\left[\texttt{glucose}_i | \texttt{physact}_i = \texttt{much more active}\right] | H_0 && = \beta_0 + 4 \cdot \beta
\end{aligned}
$$

Testing the linear trend hypothesis is equivalent to testing the adequacy of a numeric coding based on integer scores from 0 to 4

Linear constraints

$$H_0 : \beta_B = \beta_C = \beta_D = \beta_E = \beta(\neq 0) \Longrightarrow H_0 : \begin{cases} \beta_B = \beta_C \\ \beta_C = \beta_D \\ \beta_D = \beta_E \end{cases}$$

```
> K.lin
     1  2  3  4  5
 1   0  1 -1  0  0
 2   0  0  1 -1  0
 3   0  0  0  1 -1
> t.lin
[1] 0 0 0
```


F test for the linear trend hypothesis - 1

```
> linearHypothesis(physact2,K.lin,t.lin,test="F")
Linear hypothesis test
Hypothesis:
physactd1 - physactd2 = 0
physactd2 - physactd3 = 0
physactd3 - physactd4 = 0
Model 1:  restricted model
Model 2:  glucose ~ physact
     Res.Df         RSS  Df  Sum of Sq      F  Pr(>F)
 1     2030   191419.47
 2     2027   191344.61   3       74.86   0.26   0.8511
```

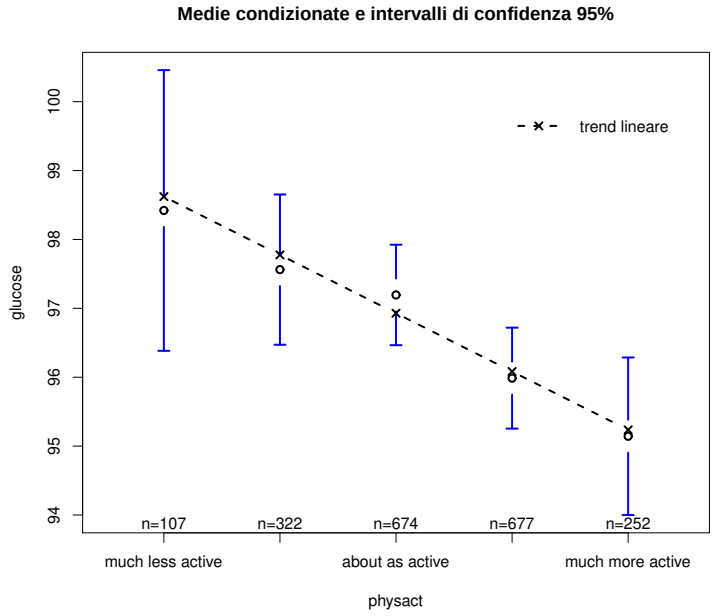
The linear trend hypothesis is not rejected

Linear trend hypotesis - constrained estimation

The parameters of the constrained model can be estimated by coding the ordered categorical regressor using integer scores from 0 to 4 and by fitting a new Gaussian linear model

```
> hers.nod$physact.num<-as.numeric(hers.nod$physact)-1
> physact3<-lm(glucose physact.num,data=hers.nod)
> summary(physact3)

...
              Estimate   Std.  Error   t value   Pr(>|t|)
 (Intercept)     98.622         0.523   188.592      0.000
 physact.num     -0.847         0.206    -4.117      0.000
...

Residual standard error:  9.711 on 2030 degrees of freedom
Multiple R-squared:  0.00828, Adjusted R-squared:  0.007792
F-statistic:  16.95 on 1 and 2030 DF, p-value:  3.993e-05
```


Model comparison



F test for the linear trend hypothesis - 2

Model comparison

```
> anova(physact3,physact2)

Analysis of Variance Table

Model 1:  glucose ~ physact.num
Model 2:  glucose ~ physact

     Res.Df         RSS   Df   Sum of Sq      F   Pr(>F)
 1     2030   191419.47
 2     2027   191344.61    3       74.86   0.26    0.8511
```

Evaluation and comparison criteria for Gaussian linear models

Saturated models

$M : \boldsymbol{Y}|\mathbf{X} \sim MVN_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right)$

⇒ The saturated model for M is a model with a number of parameters for the expected values that is equal to the number of unique covariate patterns in the matrix $\mathbf{X}$ (equal to the unique values for $\mathbf{x}_i^\top\boldsymbol{\beta}$)

⇒ if the number of unique covariate patterns is equal to n (*each sample unit is characterised by a specific combination of regressor values*), the saturated model can be defined as follows:

$M_{sat} : \boldsymbol{Y}|\mathbf{X} \sim MVN_n\left(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n\right)$

con $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^\top \in \mathbb{R}^n$

Maximum likelihood estimation of $\mu_1, \mu_2, \ldots, \mu_n$

- Log-likelihood function for the saturated model:

  $l\left(\mu_1, \ldots, \mu_n, \sigma^2\right) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mu_i)^2$

- Maximum likelihood estimate of μ_i:

  $$\left.\begin{array}{l} \dfrac{\partial}{\partial\mu_i} l\left(\mu_1, \ldots, \mu_n, \sigma^2\right) = \dfrac{y_i - \mu_i}{\sigma^2} \\ \dfrac{\partial^2}{\partial\mu_i^2} l\left(\mu_1, \ldots, \mu_n, \sigma^2\right) = -\dfrac{1}{\sigma^2} \end{array}\right\} \quad \Rightarrow \quad \hat{m}_i = y_i \quad i = 1, \ldots, n$$

  ⇒ $l\left(\hat{m}_1, \ldots, \hat{m}_n, \sigma^2\right) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2$

  Maximum possible value for the log-likelihood associated to Gaussian models for $\boldsymbol{Y}|\mathbf{X}$, given the observed sample $\mathbf{y}$

  ⇒ *any Gaussian linear model for $\boldsymbol{Y}|\mathbf{X}$ shows a maximum value for the log-likelihood that is smaller than that value, given the observed sample $\mathbf{y}$*

Comparisons with the saturated model

Any Gaussian linear model for $\boldsymbol{Y}|\mathbf{X}$ can be seen as a model that introduces some constraints on the parameters of the saturated model. These constraints can be expressed through a linear system:

$$\left.\begin{array}{l} M_{sat} : \boldsymbol{Y}|\mathbf{X} \sim MVN_n\left(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n\right) \\ H_0 : \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \quad \boldsymbol{\beta} \in \mathbb{R}^{p+1} \end{array}\right\} \quad \Rightarrow \quad M : \boldsymbol{Y}|\mathbf{X} \sim MVN_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right)$$

At least theoretically, the adequacy of these constraints can be evaluated through a likelihood ratio test

(Residual) deviance of a Gaussian linear model

$$\begin{aligned} D &= 2\ln\left[\frac{L\left(\hat{m}_1, \ldots, \hat{m}_n, \sigma^2\right)}{L\left(\hat{\mathbf{b}}, \sigma^2\right)}\right] = 2\left[l\left(\hat{m}_1, \ldots, \hat{m}_n, \sigma^2\right) - l\left(\hat{\mathbf{b}}, \sigma^2\right)\right] \\ &= 2\left[-\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 + \frac{n}{2}\ln 2\pi + \frac{n}{2}\ln\sigma^2 + \frac{\mathbf{e}^\top\mathbf{e}}{2\sigma^2}\right] = \frac{\mathbf{e}^\top\mathbf{e}}{\sigma^2} \end{aligned}$$

Note that:

- $\left.\frac{\mathbf{e}^\top\mathbf{e}}{\sigma^2}\right| H_0 : \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \sim \chi^2_{n-p-1}$ (*recall the properties of raw residuals*)
- D depends on σ^2 and cannot be computed if the number of covariate patterns is equal to n (*substituting σ^2 by an estimate derived from M or M_{sat} leads to the values $n-p-1$ or ∞, respectively*)

  $\Rightarrow$ *in general, it is not possible to use D for testing goodness of fit in presence of nuisance parameters*
- some authors/software use the expression "residual deviance" to denote $\mathbf{e}^\top\mathbf{e}$, and the expression "scaled deviance" to denote D

Coefficient of multiple linear determination 6

R^2 coefficient

$$R^2 = 1 - \frac{\mathbf{e}^\top \mathbf{e}}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

- It is possible to prove that $0 \leq \mathbf{e}^\top \mathbf{e} \leq \sum_{i=1}^n (y_i - \bar{y})^2$

  $\Rightarrow$ $R^2 \in [0, 1]$

  $\Rightarrow$ $R^2 = 1$ if and only if $\mathbf{e}^\top \mathbf{e} = 0$

  (*the Gaussian linear model M is "equivalent" to the corresponding saturated model*)

  $\Rightarrow$ $R^2 = 0$ if and only if $\mathbf{e}^\top \mathbf{e} = \sum_{i=1}^n (y_i - \bar{y})^2$

  (*the Gaussian linear model M is "equivalent" to the Gaussian linear model that assumes linear independence of $\boldsymbol{Y}$ from all regressors*)

Comparisons among Gaussian linear models 7

Choice among Gaussian linear models

Aim:
Which is the most adequate Gaussian linear model for a given random sample $\boldsymbol{Y}$?

Simplest situation: two candidate models

$M_A : \boldsymbol{Y}|\mathbf{X}_A \sim MVN_n\left(\mathbf{X}_A\boldsymbol{\beta}_A, \sigma^2\mathbf{I}_n\right), \quad \boldsymbol{\beta}_A \in \mathbb{R}^{p_A+1}$

$M_B : \boldsymbol{Y}|\mathbf{X}_B \sim MVN_n\left(\mathbf{X}_B\boldsymbol{\beta}_B, \sigma^2\mathbf{I}_n\right), \quad \boldsymbol{\beta}_B \in \mathbb{R}^{p_B+1}$

Note that:
the two models differ because they are characterised by different sets of regressors: $\mathbf{X}_A \neq \mathbf{X}_B$ (without loss of generality: $p_A > p_B$)

Situation 1: nested models - 1

Matrix $\mathbf{X}_B$ can be obtained by removing one or more that one columns from matrix $\mathbf{X}_A$

$\Rightarrow$ M_B can be obtained by introducing suitable linear constraints on the parameteres of M_A

$$\left.\begin{array}{l} M_A : \boldsymbol{Y}|\mathbf{X}_A \sim MVN_n\left(\mathbf{X}_A\boldsymbol{\beta}_A, \sigma^2\mathbf{I}_n\right) \\ H_0 : \mathbf{K}_B\boldsymbol{\beta}_A = \mathbf{t}_B \end{array}\right\} \quad \Rightarrow \quad M_B : \boldsymbol{Y}|\mathbf{X}_B \sim MVN_n\left(\mathbf{X}\boldsymbol{\beta}_B, \sigma^2\mathbf{I}_n\right)$$

$q = p_A - p_B$ number of regressors excluded from M_A to obtain M_B

$\mathbf{K}_B$ $(q) \times (p_A + 1)$ matrix
each row of this matrix contains a 1 in a specific position (corresponding to one of the q regressors excluded from M_A), and 0 elsewhere

$\mathbf{t}_B = \mathbf{0}_q$

Situation 1: nested models - 2

A likelihood ratio test can be exploited to choose among M_A and M_B. In particular, such test can be expressedas a function of the two corresponding (scaled) deviances:

$$\Delta l = 2\ln\left[\frac{L\left(\hat{\mathbf{b}}_A, \sigma^2\right)}{L\left(\hat{\mathbf{b}}_{A|H_0}, \sigma^2\right)}\right] = 2\ln\left[\frac{L\left(\hat{\mathbf{b}}_A, \sigma^2\right)}{L\left(\hat{\mathbf{b}}_B, \sigma^2\right)}\right] = \frac{\mathbf{e}_B^\top\mathbf{e}_B - \mathbf{e}_A^\top\mathbf{e}_A}{\sigma^2}$$

$$= D(M_B) - D(M_A) = \Delta D$$

If H_0 is true - if M_B is as "adequate" as M_A:

$\Rightarrow$ $\Delta D|\, M_B \sim \chi^2_q$

$\Rightarrow$ $$\left.\frac{\Delta D}{D_{M_A}}\,\frac{n - p_A - 1}{q}\right| M_B \sim F_{(q, n-p_A-1)}$$

Situation 2: non-nested models

Matrix $\mathbf{X}_B$ cannot be obtained by removing one or more than one columns from matrix $\mathbf{X}_A$

$\Rightarrow$ Model M_B can be obtained by simultaneously excluding some (or all) regressors in model M_A and adding some regressors to model M_A

$\Rightarrow$ *The two models are characterised by two sets of regressors that are only partially overlapping, or non-overlapping*

$\Rightarrow$ The differences between the two deviances does not have a known random distribution, and thus a likelihood ratio test cannot be used to choose between the two models

Adjusted R^2

$$R^2_{adj} = 1 - \left(\frac{\mathbf{e}^\top \mathbf{e}}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \cdot \frac{n-1}{n-p-1} \right)$$

It is possible to prove that $\mathbf{e}^\top \mathbf{e}$ never increase after adding a regressor to a Gaussian linear model (*even if the regressor is irrelevant - see slides on linear hypotheses*)

$\Rightarrow$ *If M_A andM_B have different numbers of regressors, the use of R^2 could favour the model with the largest number of regressors*

Differently from R^2, R^2_{adj} is not affected by the effect of the number of regressors on $\mathbf{e}^\top \mathbf{e}$

$\Rightarrow$ *A reduction in $\mathbf{e}^\top \mathbf{e}$ due to the introduction of an irrelevant regressor can be balanced out by the corresponding increase in $\frac{n-1}{n-p-1}$*

$\Rightarrow$ The best model is the one achieving the **maximum value for** R^2_{adj} (among all the considered models)

Comparing R^2_{adj} for models with the same number of parameters

Consider two models M_A and M_B such that $p_A = p_B = p$. Then:

$$
\begin{aligned}
R^2_{adj}(M_A) > R^2_{adj}(M_B) \quad &\Leftrightarrow \quad 1 - \left(\frac{\mathbf{e}_A^\top \mathbf{e}_A}{\sum_{i=1}^n (y_i - \bar{y})^2} \cdot \frac{n-1}{n-p-1} \right) > 1 - \left(\frac{\mathbf{e}_B^\top \mathbf{e}_B}{\sum_{i=1}^n (y_i - \bar{y})^2} \cdot \frac{n-1}{n-p-1} \right) \\
&\Leftrightarrow \quad \frac{\mathbf{e}_A^\top \mathbf{e}_A}{\sum_{i=1}^n (y_i - \bar{y})^2} \cdot \frac{n-1}{n-p-1} < \frac{\mathbf{e}_B^\top \mathbf{e}_B}{\sum_{i=1}^n (y_i - \bar{y})^2} \cdot \frac{n-1}{n-p-1} \\
&\Leftrightarrow \quad \mathbf{e}_A^\top \mathbf{e}_A < \mathbf{e}_B^\top \mathbf{e}_B \\
&\Leftrightarrow \quad R^2(M_A) > R^2(M_B)
\end{aligned}
$$

Leave-One-Out Cross-Validation

Basic idea:

Between two regression models, choose the one with the smallest prediction error

General definition:

$\hat{m}_i^{[-i]}$ estimate of $\mathrm{E}[Y_i|\mathbf{x}_i]$ obtained after excluding the $i-$th unit from the observed sample (*independent of the $i-$th unit*)

$$LOOCV = \frac{\sum_{i=1}^n \left(y_i - \hat{m}_i^{[-i]} \right)^2}{n}$$ unbiased estimate of the prediction error

- *In order to compute LOOCV for a given regression model, the estimation procedure should be repeated n times after omitting each sample unit. Then, each fitted model is used to compute a prediction for the corresponding omitted sample unit*
- *the quantities $y_i - \hat{m}_i^{[-i]}$ are also referred to as "deleted residuals"*
- *Some authors/softwares use the acronym $PRESS$ (PRedictive Error Sum of Square) to denote $LOOCV$*

$LOOCV$ for Gaussian linear regression models

$\hat{\mathbf{b}}^{[-i]}$ ML estimate of $\boldsymbol{\beta}$ obtained after excluding the $i-$th unit from the observed sample (*independent of the $i-$th unit*)

$\hat{m}_i^{[-i]} = \mathbf{x}_i^\top \hat{\mathbf{b}}^{[-i]}$ estimate of $\mathrm{E}[Y_i|\mathbf{x}_i]$ obtained after excluding the $i-$th unit from the observed sample (*independent of the $i-$th unit*)

It is possible to prove that

$$y_i - \mathbf{x}_i^\top \hat{\mathbf{b}}^{[-i]} = \frac{y_i - \mathbf{x}_i^\top \hat{\mathbf{b}}}{1 - \mathbf{H}_{ii}} = \frac{e_i}{1 - \mathbf{H}_{ii}}$$

$\Rightarrow$ *LOOCV for Gaussian linear regression models can be computed without repeating the fitting process n times*

Differently from $\mathbf{e}^\top\mathbf{e}$, $LOOCV$ may increase if an irrelevant regressor is added to the model

$\Rightarrow$ The best model is the one achieving the **minimum value for** $LOOCV$ (among all the considered models)

Akaike information criterion - 1

General definition:

$\boldsymbol{Y}$ random sample with unknown probability/density function $f_0(\mathbf{y})$

$\mathcal{F}_A = \left\{ f_A(\cdot;\boldsymbol{\theta}_A), \boldsymbol{\theta}_A \in \boldsymbol{\Theta_A} \subseteq \mathbb{R}^{k_A} \right\}$ parametric statistical model A

$\mathcal{F}_B = \left\{ f_B(\cdot;\boldsymbol{\theta}_B), \boldsymbol{\theta}_B \in \boldsymbol{\Theta_B} \subseteq \mathbb{R}^{k_B} \right\}$ parametric statistical model B

Basic idea:

Between two statistical models, choose the one that contains the element that is the most "similar" to $f_0(\cdot)$

$\Rightarrow$ Kullback-Leibler divergence:

$$\mathcal{K}(f_A, f_0) = \mathrm{E}\left[\ln \frac{f_0(\boldsymbol{Y})}{f_A(\boldsymbol{Y};\boldsymbol{\theta}_A)}\right]$$

Amount of information that is lost when $f_0(\cdot)$ is approximated with $f_A(\cdot;\boldsymbol{\theta}_A) \in \mathcal{F}_A$

Akaike information criterion - 2

It is possible to prove that, under suitable regularity conditions,

$$\min_{\mathcal{F}_A} \mathcal{K}(f_A, f_0) < \min_{\mathcal{F}_B} \mathcal{K}(f_B, f_0) \quad \Leftrightarrow \quad \underbrace{-2\ln L_A\left(\hat{\boldsymbol{\theta}}_A\right) + 2k_A}_{AIC(M_A)} < \underbrace{-2\ln L_B\left(\hat{\boldsymbol{\theta}}_B\right) + 2k_B}_{AIC(M_B)}$$

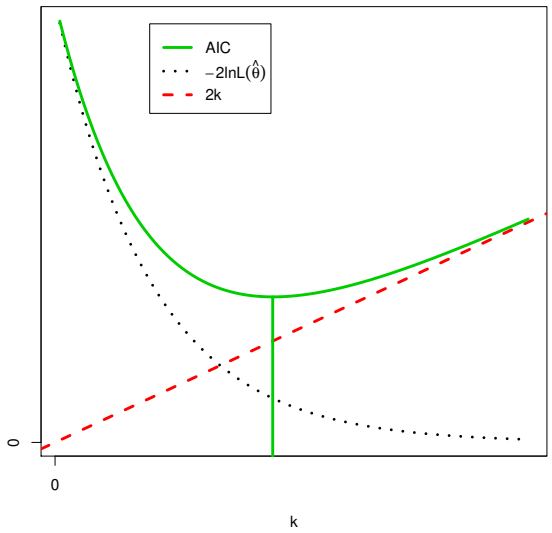
For a generic parametric statistical model:

$$AIC = -2\ln L\left(\hat{\boldsymbol{\theta}}\right) + 2k$$

- $-2\ln L\left(\hat{\boldsymbol{\theta}}\right)$ measures the goodness of fit of a statistical model to the data
  in general, this quantity decreases as the number of parameters increases
- $2k$ measures the complexity of a statistical model
  it increases as the number of the parameters increases

$\Rightarrow$ the best model is the one achieving the **minimum value for** AIC (among all the considered models)
best trade-off between goodness of fit and complexity

AIC - graphical display



The best model is the one achieving the **minimum value for** AIC (among all the considered models)
best trade-off between goodness of fit and complexity

AIC for Gaussian linear models

$$-2\ln L\left(\hat{\mathbf{b}},\hat{s}^2\right) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\frac{\mathbf{e}^\top\mathbf{e}}{n} - \frac{1}{2\frac{\mathbf{e}^\top\mathbf{e}}{n}}\mathbf{e}^\top\mathbf{e} = n\ln 2\pi + n + n\ln\frac{\mathbf{e}^\top\mathbf{e}}{n}$$

$$\Rightarrow \quad AIC = n\ln 2\pi + n + n\ln\frac{\mathbf{e}^\top\mathbf{e}}{n} + 2(p+2)$$

- Maximum likelihood estimates of $\boldsymbol{\beta}$ and σ^2 are considered
- When all the candidate models are Gaussian linear models, the following simplified formula can be used:

$$AIC = n\ln\frac{\mathbf{e}^\top\mathbf{e}}{n} + 2(p+2)$$

(Schwartz) Bayesian criterion - 1

General definition:

$\mathbf{y}$ observed sample

$\mathcal{F}_A$ parametric statistical model A

$\mathcal{F}_B$ parametric statistical model B

In the Bayesian framework, each element in $\mathcal{F}_A$ and in $\mathcal{F}_B$ has an a priori probability of being the true distribution that generated $\mathbf{y}$

$g_A(\boldsymbol{\theta}_A)$ *a priori* probability/density function for distributions belonging to $\mathcal{F}_A$

$g_B(\boldsymbol{\theta}_B)$ *a priori* probability/density function for distributions belonging to $\mathcal{F}_B$

Basic idea:

Between two statistical models, choose the one characterised by the the highest probability of having generated the observed sample

$\Rightarrow \quad \Pr(\mathbf{y}|\,\mathcal{F}_A) = \int g_A(\boldsymbol{\theta}_A) f_A(\mathbf{y};\boldsymbol{\theta}_A) d\boldsymbol{\theta}_A$

(Schwartz) Bayesian criterion - 2

It is possible to prove that, under suitable regularity conditions,

$$\Pr(\mathbf{y}|\mathcal{F}_A) > \Pr(\mathbf{y}|\mathcal{F}_B) \quad \Leftrightarrow \quad \underbrace{-2\ln L_A\left(\hat{\boldsymbol{\theta}}_A\right) + \ln(n)k_A}_{BIC(M_A)} < \underbrace{-2\ln L_B\left(\hat{\boldsymbol{\theta}}_B\right) + \ln(n)k_B}_{BIC(M_B)}$$

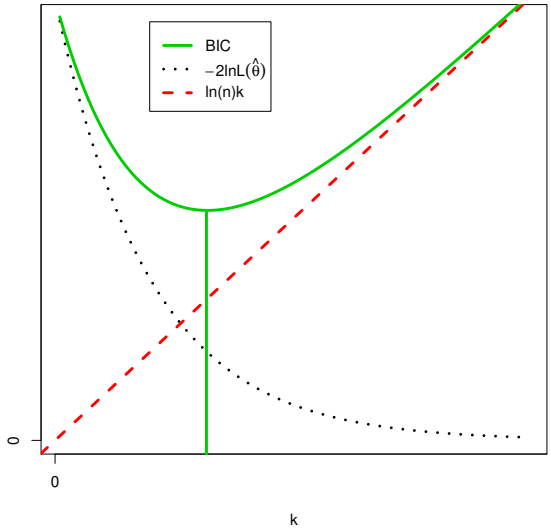
For a generic parametric statistical model:

$$BIC = -2\ln L\left(\hat{\boldsymbol{\theta}}\right) + \ln(n)k$$

- $-2\ln L\left(\hat{\boldsymbol{\theta}}\right)$ measures the goodness of fit of a statistical model to the data
  in general, this quantity decreases as the number of parameters increases
- $\ln(n)k$ measures the complexity of a statistical model
  it increases as the number of the parameters increases

$\Rightarrow$ the best model is the one achieving the **minimum value for** BIC (among all the considered models)
best trade-off between goodness of fit and complexity

BIC - graphical display



The best model is the one achieving the **minimum value for** BIC (among all the considered models)
best trade-off between goodness of fit and complexity

BIC for Gaussian linear models

$$-2\ln L\left(\hat{\mathbf{b}}, \hat{s}^2\right) = n\ln 2\pi + n + n\ln\frac{\mathbf{e}^\top\mathbf{e}}{n}$$

$$\Rightarrow \quad BIC = n\ln 2\pi + n + n\ln\frac{\mathbf{e}^\top\mathbf{e}}{n} + \ln(n)(p+2)$$

- Maximum likelihood estimates of $\boldsymbol{\beta}$ and σ^2 are considered
- When all the candidate models are Gaussian linear models, the following simplified formula can be used:

$$BIC = n\ln\frac{\mathbf{e}^\top\mathbf{e}}{n} + \ln(n)(p+2)$$

AIC or BIC?

Although derived within two completely different framework, AIC and BIC have a very similar functional form. The main "practical" difference is the way in which model complexity is weighted

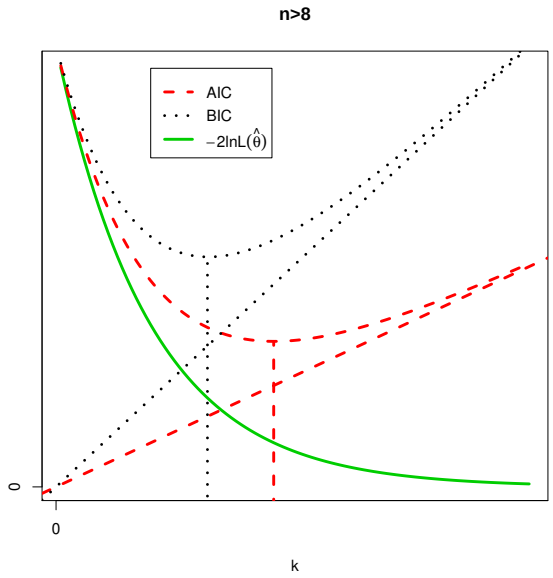
$\Rightarrow$ in general, BIC puts more weight on model complexity
$n > 8 \;\Rightarrow\; ln(n) > 2$

$\Rightarrow$ for a given observed sample, BIC tends to favour less complex models (than those selected according to AIC)

$\Rightarrow$ under suitable conditions, both criteria are consistent (when the sample size is large, they select are able to select "best" model - according to the corresponding conceptual framework)

$\Rightarrow$ there is not any test to evaluate the significance of the difference among AIC (or BIC) values

AIC e BIC - graphical comparison

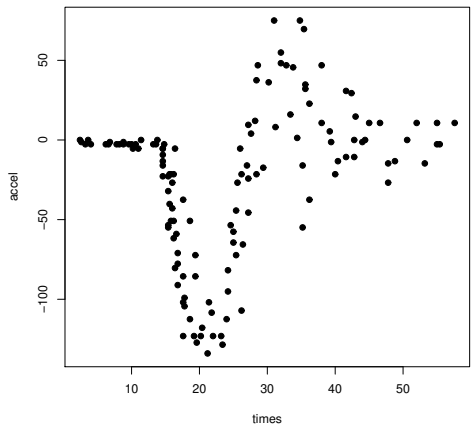


Gaussian regression models: Introducing nonlinearity through polynomials and regression splines

Crash test data - 1

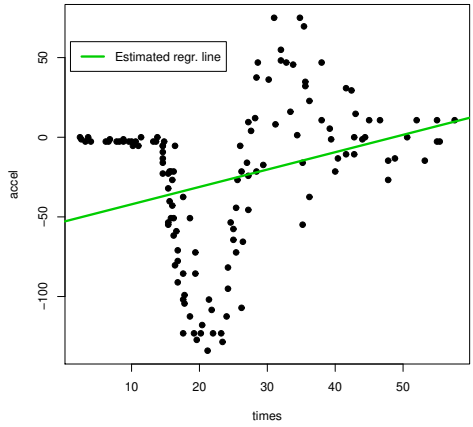
In order to evaluate the efficacy of helmets, a research team performed an experiment. In particular, after applying an accelerometer to the head of a crash test dummy, they simulated a motorcycle crash. A total of $n = 133$ readings were recorded (measured in grams), at different time points after the impact (measured in milliseconds)

Crash test data - 2



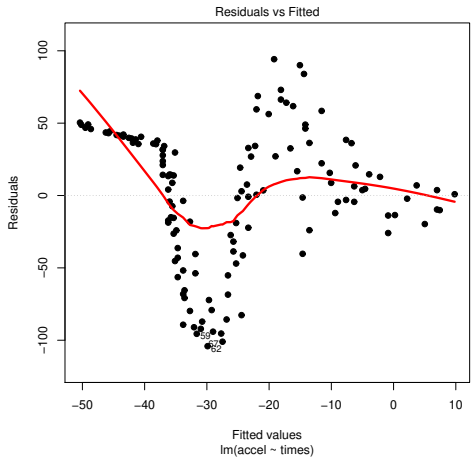
The plot shows a clear nonlinear dependence pattern

Crash test data - results from a linear model




```
Coefficients:
             Estimate   Std.  Error   t value   Pr(>|t|)
  (Intercept)   -53.008          8.712    -6.084      0.000
        times     1.091          0.307     3.552      0.001

Multiple R-squared:  0.08785, Adjusted R-squared:  0.08089
F-statistic:  12.62 on 1 and 131 DF, p-value:  0.0005318
```


Crash test data - residual analysis



The plot suggests a clear violation of the linearity assumption: there is an evident pattern in the average value of the residuals (the red line is the moving average of the residuals).

Gaussian nonlinear regression models

In presence of a single regressor, a Gaussian nonlinear model can be defined by replacing the linearity assumption

$\mathrm{E}[Y_i|x_i] = \beta_0 + \beta_1 x_i$

with the following assumption

$\mathrm{E}[Y_i|x_i] = h(x_i; \beta_0, \ldots, \beta_p)$

where $h(\cdot; \beta_0, \ldots, \beta_p)$ is assumed to have a **known** functional form with $p+1$ **unknown** parameters ($p \geq 1$)

$\Rightarrow$ *depending on the choice of the functional form of $h(\cdot; \beta_0, \ldots, \beta_p)$, different departures from linearity can be accomodated.*

Polynomial regression 7

Introducing nonlinearity through polynomials

Y_i Random variable that describes the value for the dependent variable observed on the $i-$th sample unit ($i = 1, \ldots, n$)

x_i value of the regressor for the $i-$th sample unit

A) $\mathrm{E}[Y_i|x_{1i}, \ldots, x_{pi}] = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p = \sum_{j=0}^{p} \beta_j x_i^j \ \forall i$

B) $\mathrm{Var}[Y_i|x_{1i}, \ldots, x_{pi}] = \sigma^2 \ \forall i$

C) $\mathrm{Cor}[Y_i|x_{1i}, \ldots, x_{pi}, Y_h|x_{1h}, \ldots, x_{ph}] = 0 \ \forall i \neq h$

D) $Y_i|x_{1i}, \ldots, x_{pi} \sim N\left(\sum_{j=0}^{p} \beta_j x_i^j, \sigma^2\right) \ \forall i$

$\Rightarrow$ $h(x; \beta_0, \ldots, \beta_p) = \sum_{j=1}^{p} \beta_j x^j$ *is a nonlinear function in x, but it is still linear in the unknown parameters $\beta_0, \ldots, \beta_p$*

Matrix notation for Gaussian polynomial regression models

$\mathbf{x}_i = \left(1 = x_i^0, x_i^1, \ldots, x_i^p\right)^\top$ powers of the regressor value for the l'i-th sample unit

$\mathbf{x}_{[j]} = \left(x_1^j, x_2^j \ldots, x_n^j\right)^\top$ powers of order j of all the n regressor values ($j = 0, \ldots, p$)
$x_{[0]} = (1, 1, \ldots, 1)^\top$

Regressor matrix $n \times (p+1)$ $\mathbf{X} = \begin{bmatrix} x_1^0 & x_1^1 & \ldots & x_1^p \\ x_2^0 & x_2^1 & \ldots & x_2^p \\ \vdots & \vdots & & \vdots \\ x_n^0 & x_n^1 & \ldots & x_n^p \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^\top \\ \hline \mathbf{x}_2^\top \\ \hline \vdots \\ \hline \mathbf{x}_n^\top \end{bmatrix} = \left[\mathbf{x}_{[0]} \middle| \mathbf{x}_{[1]} \middle| \cdots \mathbf{x}_{[p]}\middle|\right]$

Conditional expected values in compact form

$\mathrm{E}\left[Y_i | x_i\right] = \sum_{j=0}^p \beta_j x^j = \mathbf{x}_i^\top \boldsymbol{\beta}\ \forall i$

$\mathrm{E}\left[\boldsymbol{Y} | \mathbf{X}\right] = \mathbf{X}\boldsymbol{\beta}$

Linear basis expansions

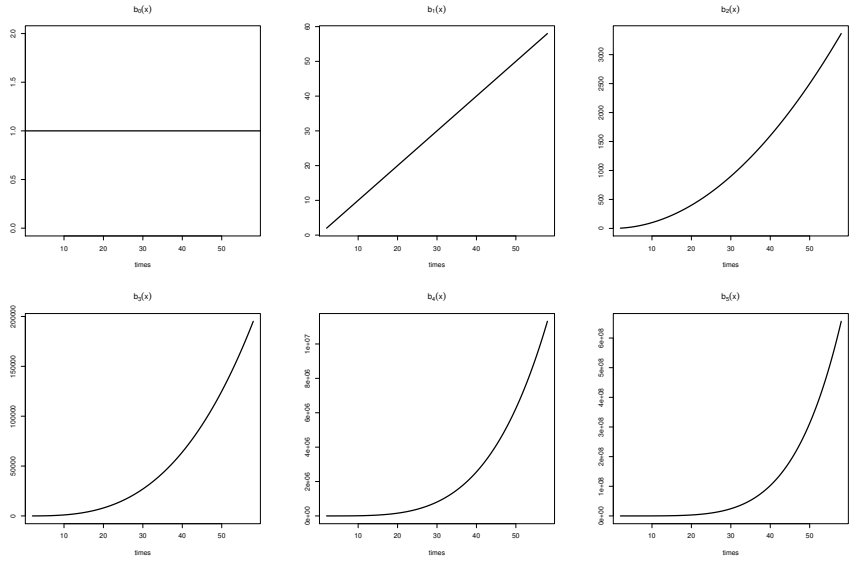
The nonlinear fuctions $h\left(x; \beta_0, \ldots, \beta_p\right)$ used in polynomial regression models can be represented using a linear basis expansion:

$$h\left(x; \beta_0, \ldots, \beta_p\right) = \sum_{j=0}^p \beta_j b_j(x)$$

The functions $b_j(x)$ ($j = 0, \ldots, p$) are called *basis*. They are nonlinear transformations of x with a known functional form and without unknown parameters

$\Rightarrow$ In polynomial regression models, $b_j(x) = x^j$ ($j = 0, \ldots, p$)

Crash test data - polynomial basis - $p = 5$



ML estimation for Gaussian polynomial regression models

$\hat{\mathbf{b}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$

Given their particular structure, the columns of $\mathbf{X}$ tend to be highly correlated (nearly linearly dependent)

- *numerical instability due to the fact that $\mathbf{X}^\top \mathbf{X}$ is nearly singular, as well as possible inflation in the standard error estimates (especially when p is large compared with n)*
- *These issues can be overcome by resorting to orthogonal polynomials: the matrix $\mathbf{X}$ is transformed into a matrix $\tilde{\mathbf{X}}$ such that*
  - $\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} = \mathbf{I}_{p+1}$
  - *for each $\boldsymbol{\beta} \in \mathbb{R}^{(p+1)}$ there exists a unique $\boldsymbol{\theta} \in \mathbb{R}^{(p+1)}$ ensuring that $\mathbf{X}\boldsymbol{\beta} = \tilde{\mathbf{X}}\boldsymbol{\theta}$*

  (the actual recursive formula to be applied at each column of $\mathbf{X}$ to obtain $\tilde{\mathbf{X}}$ is omitted)

Crash test data - polynomial of order 5 - 1

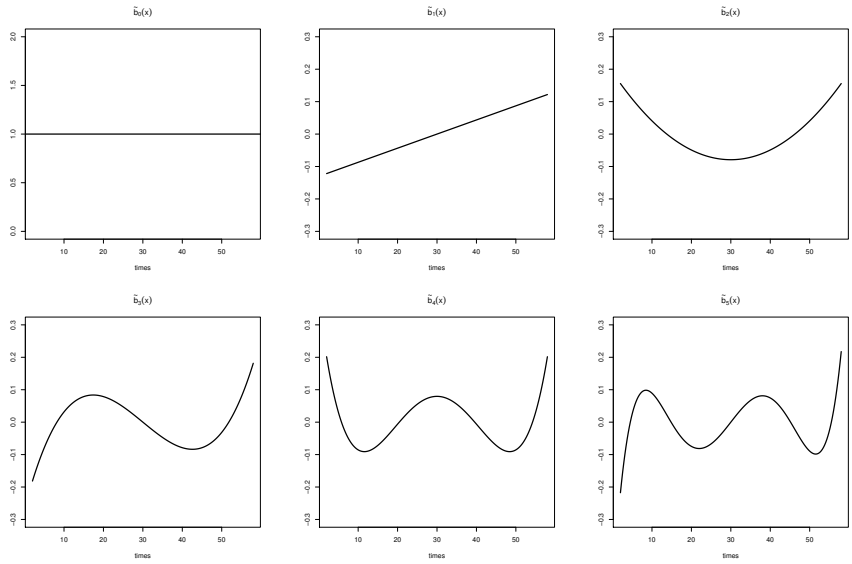
Sample correlation matrix among powers of the regressor

	times1	times2	times3	times4	times5
times1	1.0000	0.9688	0.9112	0.8499	0.7928
times2	0.9688	1.0000	0.9833	0.9479	0.9066
times3	0.9112	0.9833	1.0000	0.9895	0.9662
times4	0.8499	0.9479	0.9895	1.0000	0.9931
times5	0.7928	0.9066	0.9662	0.9931	1.0000

R^2 measuring the linear dependence of each power on the other ones

	times1	times2	times3	times4	times5
R_j^2	0.9995298	0.9999860	0.9999972	0.9999971	0.9999776

Crash test data - ortogonal polynomial basis - $p = 5$



Crash test data - polynomial of order 5 - 2

Parameter estimates

- Original polynomial

```
Coefficients:
              Estimate   Std.  Error   t value   Pr(>|t|)
 (Intercept)  -105.8767       34.9883   -3.0261     0.0030
      times1    49.7816       10.3625    4.8040     0.0000
      times2    -6.3588        1.0149   -6.2655     0.0000
      times3     0.2969        0.0425    6.9819     0.0000
      times4    -0.0057        0.0008   -7.2385     0.0000
      times5     0.0000        0.0000    7.2504     0.0000
```

- Orthogonal polynomial

```
Coefficients:
              Estimate   Std.  Error   t value   Pr(>|t|)
 (Intercept)   -25.5459        2.9396   -8.6902     0.0000
   times1ort   164.5566       33.9014    4.8540     0.0000
   times2ort   131.2271       33.9014    3.8708     0.0002
   times3ort  -239.7898       33.9014   -7.0732     0.0000
   times4ort    -6.7378       33.9014   -0.1987     0.8428
   times5ort   245.7987       33.9014    7.2504     0.0000
```


Crash test data - polynomial of order 5 - 2

Summary statistics

- Original polynomial

```
Residual standard error:  33.9 on 127 degrees of freedom
Multiple R-squared:  0.5264, Adjusted R-squared:  0.5078
F-statistic:  28.24 on 5 and 127 DF, p-value:  < 2.2e-16
```

- Orthogonal polynomial

```
Residual standard error:  33.9 on 127 degrees of freedom
Multiple R-squared:  0.5264, Adjusted R-squared:  0.5078
F-statistic:  28.24 on 5 and 127 DF, p-value:  < 2.2e-16
```

Properties of regression models with orthogonal polynomials

When considering orthogonal polynomials, it is possible to prove that:

- The estimate for the intercept $\tilde{\beta}_0$ coincides with $\bar{y}$
- The estimate for the regression coefficient associated with the j-th orthogonal bases $\tilde{b}_j(x)$ (j-th column of $\tilde{\mathrm{X}}$ coincides with the estimate of the slope of the simple Gaussian linear regression model

  M_j: $Y_i|x_{1i},\ldots,x_{pi} \sim N\left(\tilde{\beta}_0 + \tilde{\beta}_j\tilde{b}_j(x_i), \sigma^2\right)\ \forall i$

  $\Rightarrow$ *the inclusion of an additional term in the orthogonal polynomial does not alter the estimates for the terms already incldued in the model*
- The R^2 for the polynomial model of order p can be decomposed in the sum of the R^2s of the p simple Gaussian linear regression models M_j ($j = 1,\ldots,p$), each involving only one of the orthogonal basis

  $\Rightarrow$ *the contribution of each polynomial term in explaining the variability of the dependent variable can be evaluated independently*

Crash test data - orthogonal polynomial of order 5 - 1

- linear term

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-25.5459	4.0170	-6.36	0.0000
times1ort	164.5566	46.3264	3.55	0.0005

- quadratic term

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-25.5459	4.0868	-6.25	0.0000
times2ort	131.2271	47.1316	2.78	0.0062

- cubic term

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-25.5459	3.7935	-6.73	0.0000
times3ort	-239.7898	43.7484	-5.48	0.0000

- quartic term

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-25.5459	4.2057	-6.07	0.0000
times4ort	-6.7378	48.5026	-0.14	0.8897

- quintic term

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-25.5459	3.7713	-6.77	0.0000
times5ort	245.7987	43.4931	5.65	0.0000

Crash test data - orthogonal polynomial of order 5 - 2

Decomposition of R^2

	R_j^2
linear term	0.08785
quadratic term	0.05587
cubic term	0.18660
quartic term	0.00014
quintic term	0.19600
total	0.5264

Hypothesis testing for Gaussian polynomial regression models

Comparisons between nested polynomials (*choice of the degree of the polynomial*)

$$\left.\begin{array}{l} M_A : \mathrm{E}\left[Y_i|x_i\right] = \sum_{j=0}^{p} \beta_j x^j \\ \\ H_0 : M_B : \mathrm{E}\left[Y_i|x_i\right] = \sum_{j=0}^{p-q} \beta_j x^j \quad (q \leq p) \end{array}\right\} \quad \Rightarrow \quad H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \ldots = \beta_p = 0$$

Likelihood ratio test:

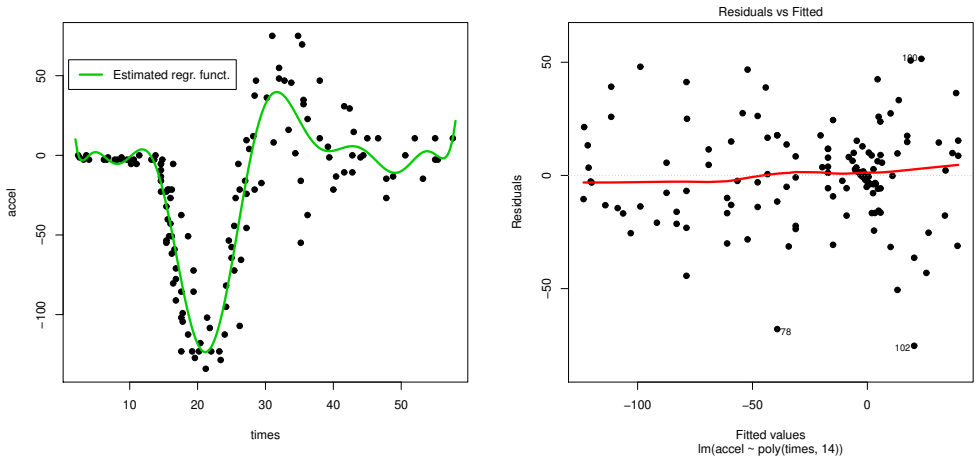
$$\Delta l = \frac{\mathbf{e}_B^\top \mathbf{e}_B - \mathbf{e}_A^\top \mathbf{e}_A}{\sigma^2} = D(M_B) - D(M_A) = \Delta D$$

If H_0 is true - if the polynomial of order $p-q$ is as "adequate" as the polynomial of order p:

$$\Rightarrow \quad \Delta D | \, M_B \sim \chi^2_q$$

$$\Rightarrow \quad \left. \frac{\Delta D}{D_{M_A}} \frac{n - p_A - 1}{q} \right| M_B \sim F_{(q, n-p_A-1)}$$

Crash test data - polynomial of order 14



The plots suggest that a polynomial of order 14 *could be adequate to describe the effect of time on accelaration (no clear pattern in the average value of the residuals)*

Crash test data - comparison between polynomials

```
> K
     1   2   3   4   5   6   7   8   9   10   11   12   13   14   15
 1   0   0   0   0   0   0   0   0   0    0    0    0    0    1    0
 2   0   0   0   0   0   0   0   0   0    0    0    0    0    0    1

> t

[1] 0 0

> linearHypothesis(poly14,K,t,test="F")

Linear hypothesis test

Hypothesis:
poly(times, 14)13 = 0
poly(times, 14)14 = 0

Model 1:  restricted model
                                    Res.Df        RSS   Df   Sum of Sq      F   Pr(>F)
Model 2:  accel ~poly(times, 14)  1      120   61693.46
                                  2      118   61442.12    2      251.33   0.24   0.7860
```

The polynomial of order 14 is not significantly better than the polynomial of order 12

Polynomial regression - some cautionary remarks

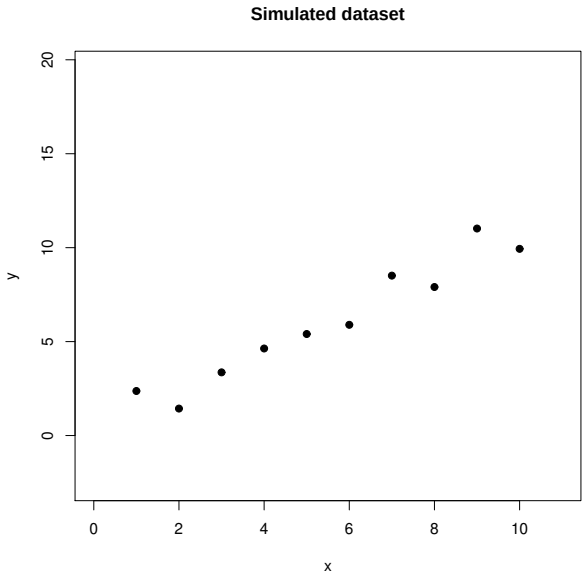
ML estimates of the coefficients of a polynomial regression function are affected by each observation in the sample.

This can lead to undesirable side-effects. In particular, a small change in one observed value for the dependent variable can lead to a dramatic change in the fitted function (even for values of the regressors that are far from that observation).

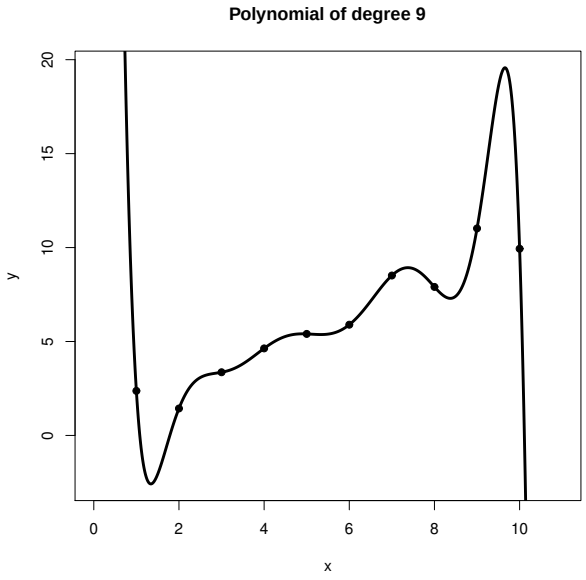
This erratic/unstable behaviour is exacerbated

- near the boundaries of the regressor range
- when the degree of the polynomial is large (compared with the sample size)

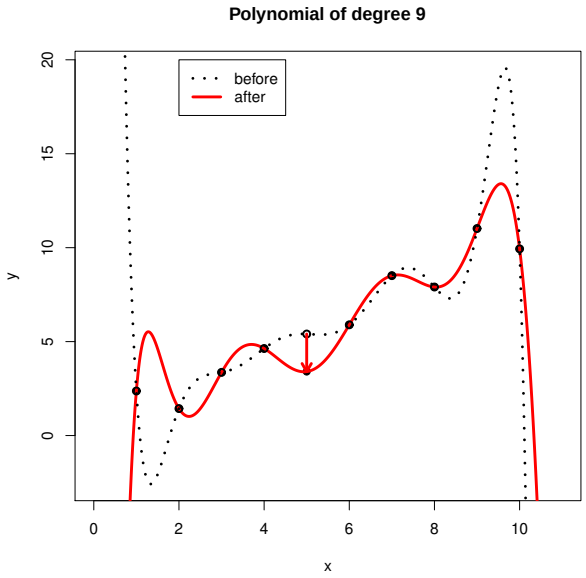
Some cautionary remarks - example - 1



Some cautionary remarks - example - 2



Some cautionary remarks - example - 3



Note the effect on the fitted regression function due to a small change in a single observation

Piecewise linear functions

Suppose that the range of x is partitioned into $K+1$ intervals using a known sequence of K values $l_1 < l_2 < \ldots, < l_K$ (called "knots")

A function $h(x)$ is said to be piecewise linear with fixed knots $l_1 < l_2 < \ldots, < l_K$ if

$$h(x) = \begin{cases} \beta_{01} + \beta_{11}x & x < l_1 \\ \vdots & \vdots \\ \beta_{0k} + \beta_{1k}x & l_{k-1} \leq x < l_k \quad (k = 2, \ldots, K) \\ \vdots & \vdots \\ \beta_{0K+1} + \beta_{1K+1}x & x \geq l_K \end{cases}$$

The total number of free parameters of a piecewise linear function is given by

$2 \cdot (K+1) = 2K + 2$

2 parameters for each interval (1 linear function for each interval)

Continuous piecewise linear functions

A function $h(x)$ is said to be a continuous piecewise linear function with fixed knots $l_1 < l_2 < \ldots, < l_K$, if it is a piecewise continuous linear fuction that satisfies the following additional *continuity* contraints at each knot:

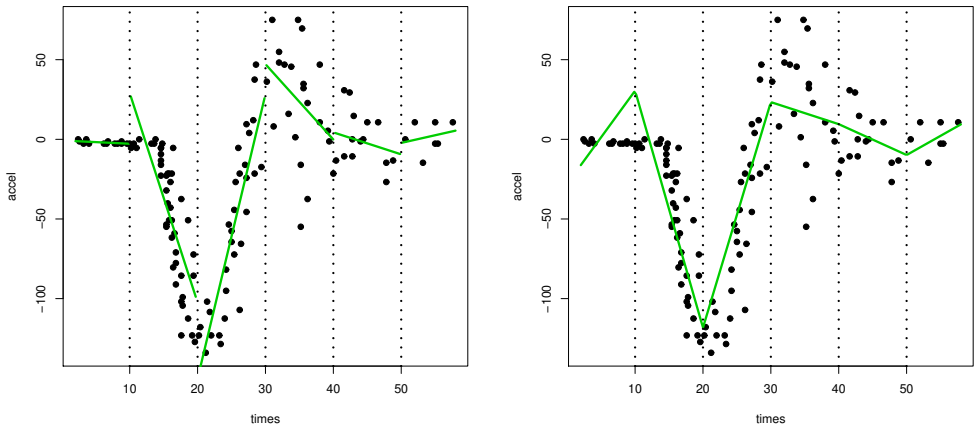
$$\begin{cases} \beta_{01} + \beta_{11}l_1 = \beta_{02} + \beta_{12}l_1 \\ \vdots \\ \beta_{0k} + \beta_{1k}l_k = \beta_{0k+1} + \beta_{1k+1}l_k \quad (k = 2, \ldots, K-1) \\ \vdots \\ \beta_{0K} + \beta_{1K}l_K = \beta_{0K+1} + \beta_{1K+1}l_K \end{cases}$$

The total number of free parameters of a continuous piecewise linear function is given by

$2 \cdot (K+1) - K = 2K + 2 - K = K + 2$

2 parameters for each interval $-$ K constraints

Crash test data - piecewise & cont. piecewise linear functions



The dashed vertical lines denote the location of the knots

Linear basis expansion for cont. piecewise linear functions

It is possible to prove that:

- any continuous piecewise linear function with fixed knots $l_1 < l_2 < \ldots, < l_K$ can be represented using a linear basis expansion:

$$h(x) = \sum_{j=1}^{K+2} \theta_j b_j(x)$$

- this linear basis expansion is not unique, in the sense that there exist several possible choices for the basis functions $b_j(\cdot)$

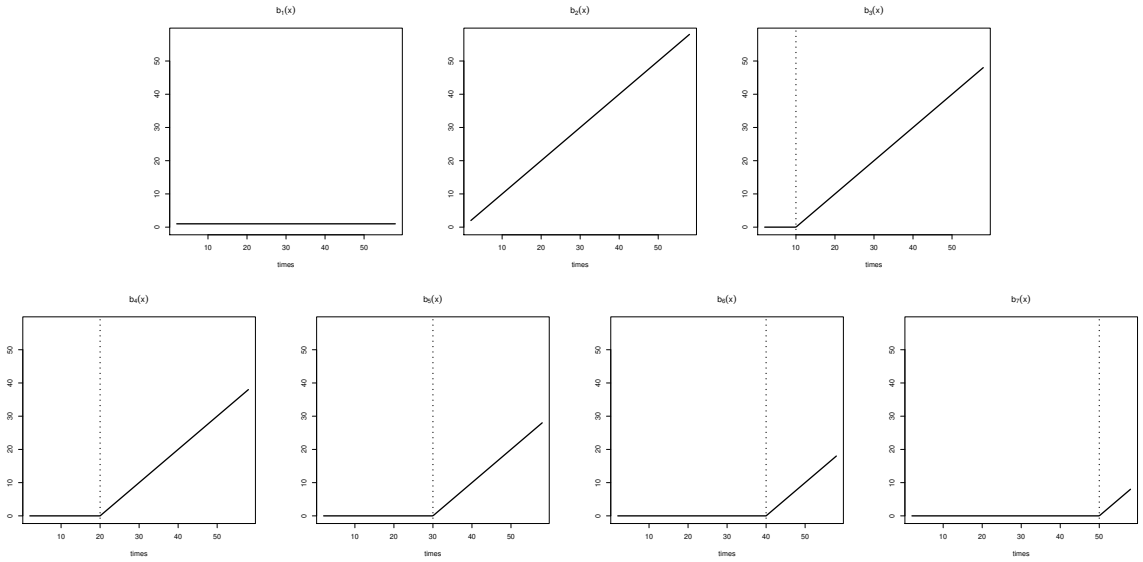
Truncated linear basis

$$b_j(x) = \begin{cases} x^0 = 1 & j = 1 \\ x^1 = x & j = 2 \\ (x - l_{j-2})_+ & j = 3 \ldots, K+2 \end{cases}$$

where $(\cdot)_+$ denotes the positive portion of its argument:

$$(r)_+ = \begin{cases} r & r \geq 0 \\ 0 & r < 0 \end{cases}$$

Crash test data - an example of truncated linear basis



Spline functions

A function $h(x)$ is said to be a spline function of degree m with fixed knots $l_1 < l_2 < \ldots, < l_K$ if

$$h(x) = \begin{cases} \sum_{j=0}^{m} \beta_{j1} x^j & x < l_1 \\ \vdots & \vdots \\ \sum_{j=0}^{m} \beta_{jk} x^j & l_{k-1} \leq x < l_k \quad (k = 2, \ldots, K) \\ \vdots & \vdots \\ \sum_{j=0}^{m} \beta_{jK+1} x^j & x \geq l_K \end{cases}$$

and its partial derivatives with respect to x are continuous up to the order $m - 1$

$\Rightarrow$ *continuous piecewise linear functions are splines of degree 1* $\rightarrow \frac{\partial^0}{\partial x^0} h(x) = h(x)$

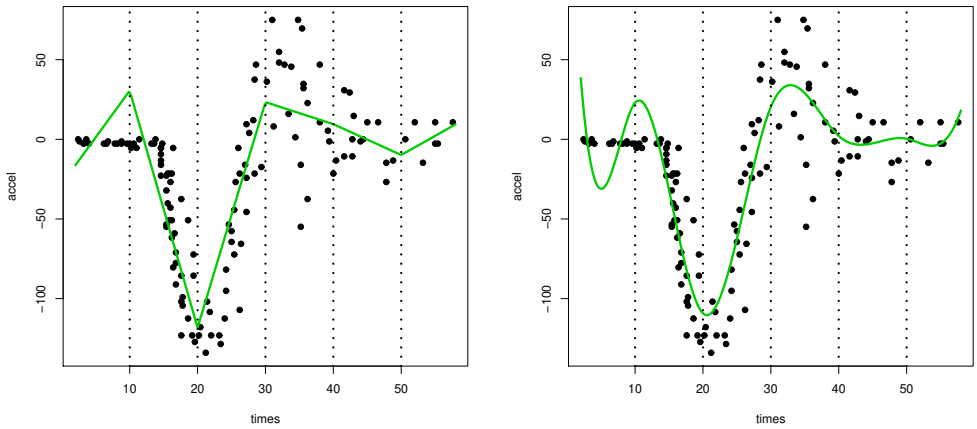
- the total number of free parameters of a spline function is given by $(K + 1)(m + 1) - Km = Km + K + m + 1 - Km = K + m + 1$
  - $\Rightarrow$ $m + 1$ parameters for each interval (*1 polynomial of degree m for each interval*)
  - $\Rightarrow$ m constraints for each knot

Cubic splines

A function $h(x)$ is a cubic spline with fixed knots $l_1 < l_2 < \ldots, < l_K$ if it is a spline function of degree 3

- the total number of free parameters of a cubic spline with K fixed knots is given by $K + 4$
  - $\Rightarrow$ 4 parameters for each interval (*1 cubic function for each interval*)
  - $\Rightarrow$ 3 constraints for each knot
- it is claimed that cubic splines are the lowest order splines for which the discontinuity in the partial derivatives at the knots cannot be noticed by the human eye
- cubic splines have interesting mathematical properties (see upcoming classes)

Crash test data - linear vs cubic regression splines



The dashed vertical lines denote the location of the knots

Linear basis expansion for spline functions

It is possible to prove that:

- any spline function of degree m with fixed knots $l_1 < l_2 < \ldots, < l_K$ can be represented using a linear basis expansion:

$$h(x) = \sum_{j=1}^{K+m+1} \theta_j b_j(x)$$

- this linear basis expansion is not unique, in the sense that there exist several possible choices for the basis functions $b_j(\cdot)$

Truncated power basis for spline functions

$$b_j(x) = \begin{cases} x^{j-1} & j = 1, \ldots, m+1 \\ (x - l_{j-m-1})_+^m & j = m+2, \ldots, K+m+1 \end{cases}$$

where

$$(r)_+^m = \begin{cases} r^m & r \geq 0 \\ 0 & r < 0 \end{cases}$$

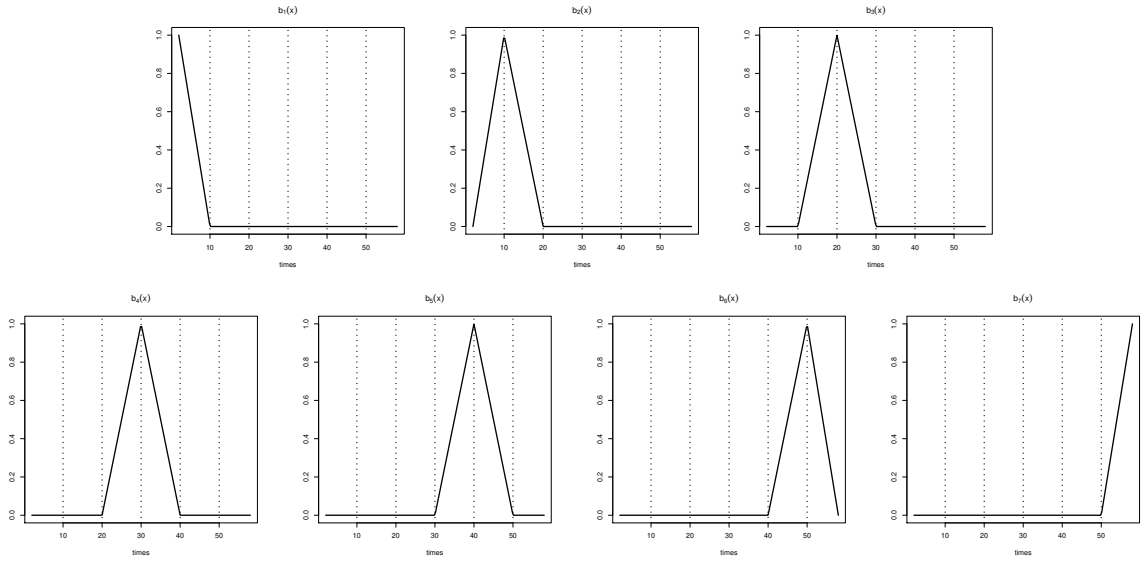
B-spline basis functions

Despite their simple and intuitive structure, truncated power basis are rarely used in practice (they are actually not implemented in R). This is due to the fact that the columns of the corresponding matrix $\mathbf{X}$ tend to be highly correlated (nearly linearly dependent), thus leading to nearly singularity of $\mathbf{X}^\top\mathbf{X}$ and numerical instability in the estimation process

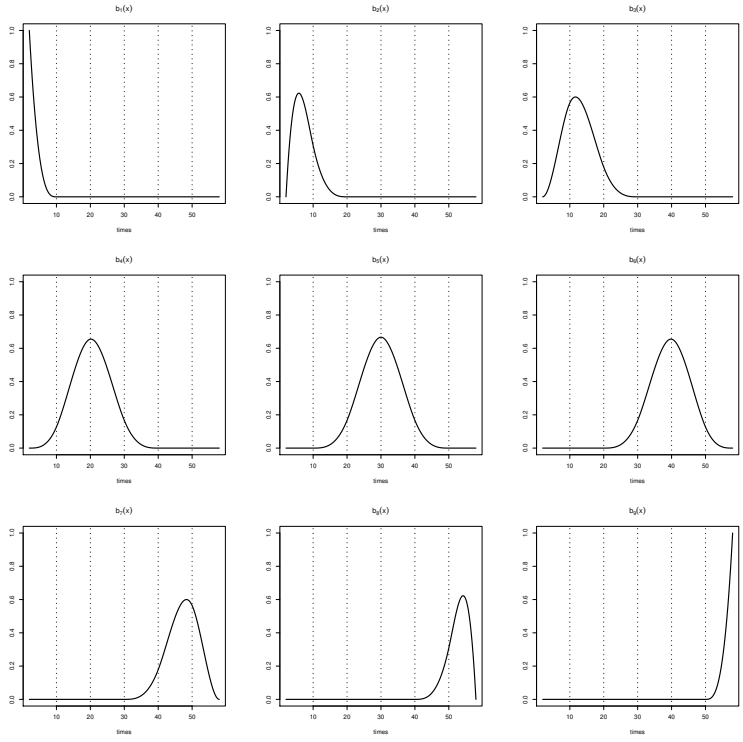
An alternative linear basis expansion representation that does not suffer these problems can be obtained by resorting to the so-called B-spline basis functions. These basis functions are defined using a recursive formula (omitted)

- each B-spline function takes non-zero values only between a pair of knots (the actual definition of this interval depends on m)

Crash test data - an example of B-spline basis for linear splines



Crash test data - an example of B-spline basis for cubic splines



Inference for regression spline models

- Maximum likelihood estimates of the parameters $\theta_1, \ldots, \theta_{K+m+1}$ for a given linear basis expansion can be easily obtained using standard tools for Gaussian regression models with regression functions that are linear in the unknown parameters.
- Comparisons among regression spline models require some caution:
  - nested models can be obtained by adding (or removing) knots, if truncated power basis are used to represent the spline function

    $\Rightarrow$ *likelihood ratio test*
  - a change in the location of one (or more than one) knot leads to a non-nested model

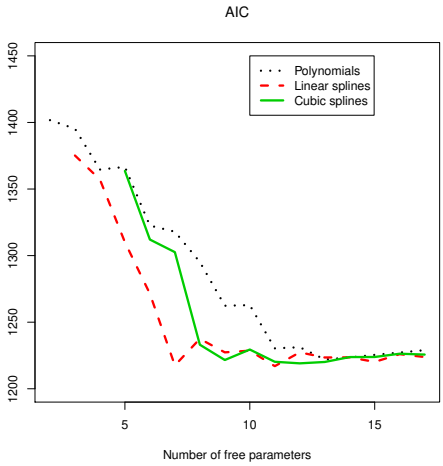
    $\Rightarrow$ *model selection criteria*

Choice of the knots - location

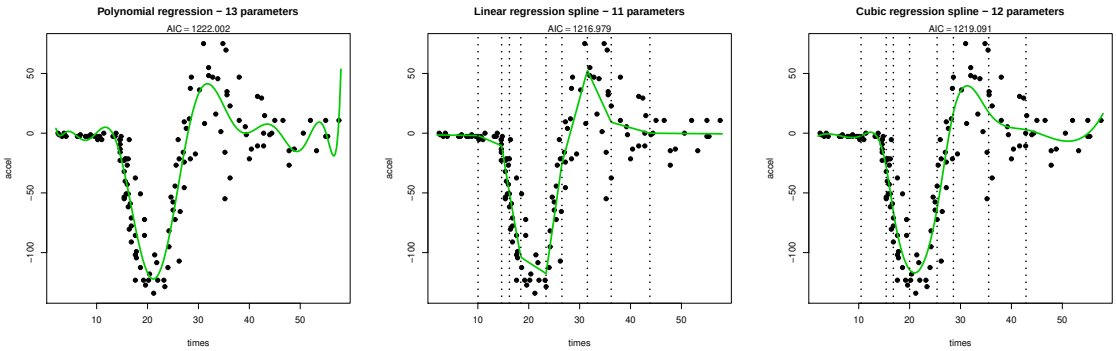
- subjective choice
- equidistant knots
- knots located at the quantiles of the regressor

  $\Rightarrow$ this choice guarantees an approximate constant number of sample units within each interval

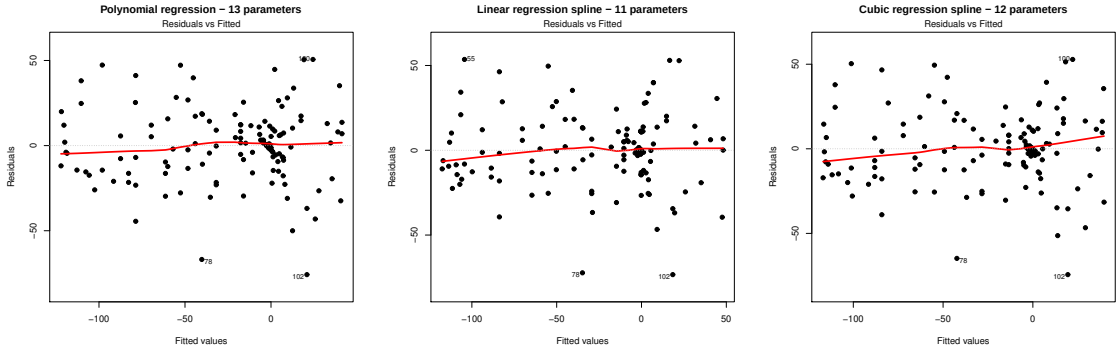
Crash test data - a comparison among alternative models



Crash test data - best models



Crash test data - residuals



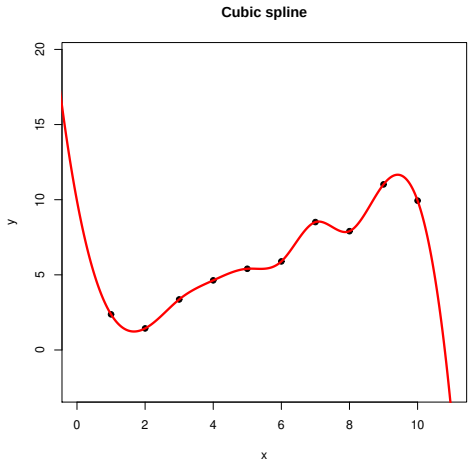
Polynomials & spline functions - some remarks

Differently from polynomials, spline functions have a local nature

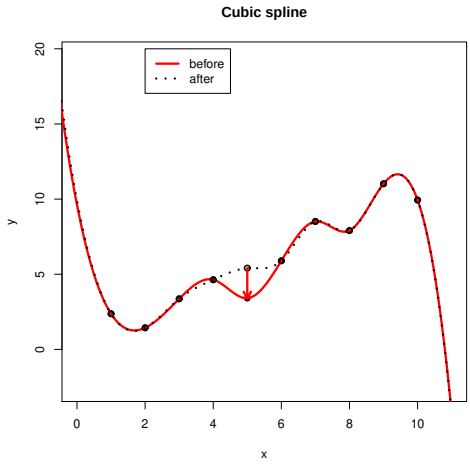
$\Rightarrow$ they are structured as local polynomials defined on non-overlapping intervals

$\Rightarrow$ a change in one observed value for the dependent variable affects only some of the polynomials that compose the spline functions (the ones close to the interval to which the observation belongs), but leaves the other polynomials unchanged

Polynomials & spline functions - some remarks - example - 1



Polynomials & spline functions - some remarks - example - 2



Concluding remarks

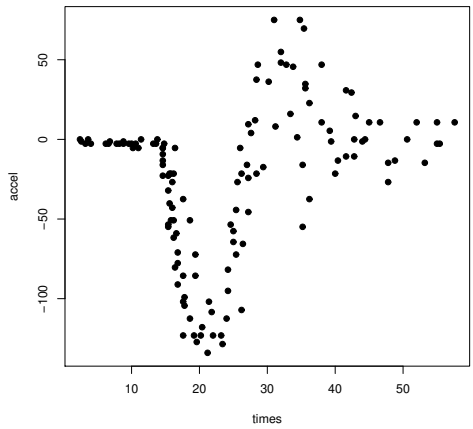
- Besides polynomials and splines, there exist many other examples of functions that admit a linear basis expansion representation
- These techniques can be extended to deal with more than one regressor

  $\Rightarrow$ *additive models*

  $$\mathrm{E}\left[Y_i | x_{1i}, \ldots x_{pi}\right] = h_1\left(x_{1i}; \boldsymbol{\beta}_1\right) + \ldots + h_p\left(x_{1pi}; \boldsymbol{\beta}_p\right)$$

Introducing regularization: smoothing splines and P-splines

Crash test data - continued



Difficulties in using regression splines

Regression (cubic) splines are an attractive solution to enhance the flexibility of Gaussian regression model

HOWEVER, they suffer from a major shortcoming: the choice of the number and of the locations of the knots

⇒ it is basically not possible to completely avoid any amount of subjectivity and to make this choice in a systematic/objective manner

⇒ considering equidistant knots or placing knots at quantiles are suboptimal strategies, leading to complications when performing model comparisons (models with different numbers of knots are not nested)

A (seemingly unrelated) alternative approach

Consider the regression model

A) $\mathrm{E}[Y_i|x_i] = h(x_i)\ \forall i$

B) $\mathrm{Var}[Y_i|x_i] = \sigma^2\ \forall i$

C) $\mathrm{Cor}[Y_i|x_i, Y_h|x_h] = 0\ \forall i \neq h$

⇒ *No explicit assumptions on the functional form of $h(\cdot)$ and of the conditional distribution of $Y_i|x_i$ are introduced (nonparametric model)*

⇒ *The only requirement are:*

- *the existence and continuity of the second partial derivative* $h''(x) = \dfrac{\partial^2}{\partial x^2} h(x)$
- $(0 \leq) \displaystyle\int [h''(t)]^2\, dt < +\infty$

  (the integral is computed on the entire range of x)

Roughness of a function

the quantity

$$\int [h''(t)]^2\, dt$$

can be interpreted as a measure of the roughness/wiggliness (*departure from linearity*) of $h(\cdot)$

⇒ $\displaystyle\int [h''(t)]^2\, dt$ is a measure of the total variability of $h'(\cdot)$, the first partial derivative of $h(\cdot)$

⇒ if $h(\cdot)$ is linear, $h''(x) = 0$ and $\displaystyle\int [h''(t)]^2\, dt = 0$

⇒ $h''(\cdot)$ is not affected if a constant or a linear term is added to $h(\cdot)$

⇒ if $h(\cdot)$ is wiggly, $h'(\cdot)$ will be variable

⇒ the larger the absolute value of $h''(\cdot)$, the larger $\int [h''(t)]^2\, dt$

Penalized least squares estimation

An estimate for $h(\cdot)$ can be obtained by minimizing

$$pls_\lambda\left(h(\cdot)\right) = \sum_i \left(y_i - h(x_i)\right)^2 + \lambda \int \left[h''(t)\right]^2 dt$$

$\Rightarrow$ $\sum_i \left(y_i - h(x_i)\right)^2$ determines the goodness of fit to the data (*the smaller, the better*)

$\Rightarrow$ $\int \left[h''(t)\right]^2 dt$ acts like a penalty for roughness (*the smaller, the better*)

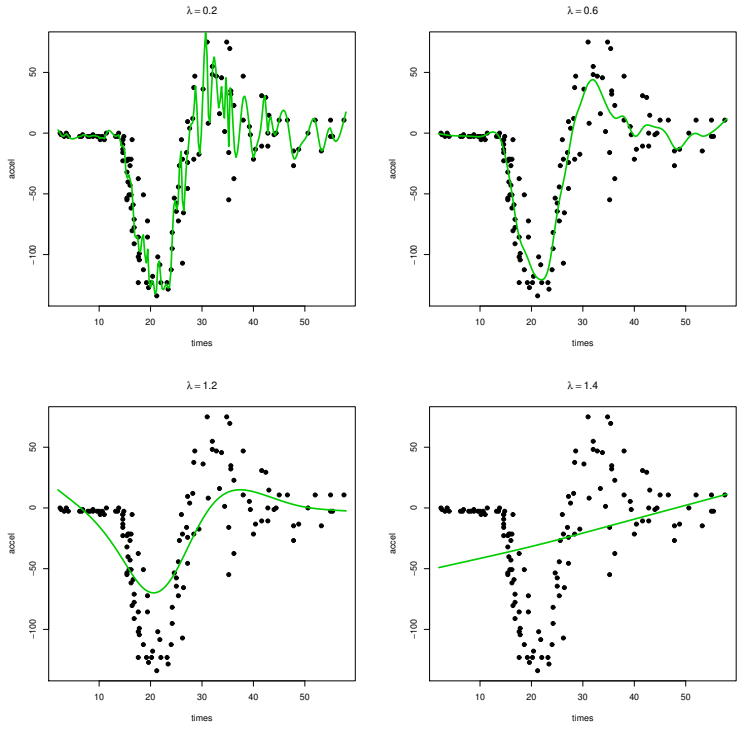
$\Rightarrow$ $\lambda \geq 0$ is the regularization/smoothness parameter controlling the trade-off between goodness of fit and roughness

Role of the smoothing parameter

The smoothing parameter λ controls the trade-off between goodness of fit and roughness

$\lambda = 0$	$\Rightarrow$	no penalization for roughness is imposed (*the resulting fitted model will be equivalent to a saturated model*)
$\lambda \to +\infty$	$\Rightarrow$	any function with $h''(x) \neq 0$ is excluded (*only linear functions are considered*)

Crash test data - penalized LS estimation



Penalized LS estimation and spline functions

It is possible to prove that, for a given value of λ:

- $pls_\lambda\left(h(\cdot)\right)$ admits a unique minimizer
- the minimizer of $pls_\lambda\left(h(\cdot)\right)$ is a natural cubic spline with knots located at the unique values of x_i $(i = 1, \ldots, n)$
  - $\Rightarrow$ for this reason, the penalized LS approach described in the previous slides is also known as **smoothing spline** approach
  - $\Rightarrow$ smoothing splines differ from regression splines due to the presence of the penalty term $\int \left[h''(t)\right]^2 dt$, that implicitly introduce constraints on the parameters of the spline function. The strength of these constraints is controlled by the smoothing parameter λ

Natural cubic splines

A function $h(x)$ is a natural cubic spline with fixed knots $l_1 < l_2 < \ldots, < l_K$ if it is a cubic spline function with the additional constraints that

$h(x) = \beta_{01} + \beta_{11} x$ if $x < l_1$

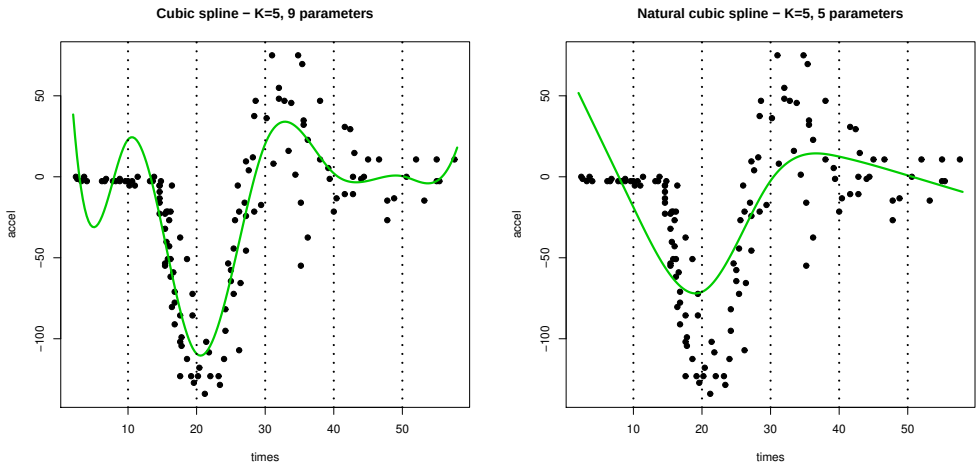
and

$h(x) = \beta_{0K} + \beta_{1K} x$ if $x > l_K$

$\Rightarrow$ *the first and last polynomials are forced to have degree 1 (with second partial derivatives equal to 0)*

$\Rightarrow$ *the total number of free parameters of a natural cubic spline with K fixed knots is given by K (4 additional restrictions are imposed)*

Crash test data - natural cubic splines vs. cubic splines



Linear basis expansion for natural cubic spline functions

It is possible to prove that:

- any natural cubic spline function with fixed knots $l_1 < l_2 < \ldots, < l_K$ can be represented using a linear basis expansion:

$$h(x) = \sum_{j=1}^{K} \theta_j b_j(x)$$

- this linear basis expansion is not unique, in the sense that there exist several possible choices for the basis functions $b_j(\cdot)$, and can be obtained starting from the basis functions of the corresponding cubic spline function with the same fixed knots

Crash test data - an example of basis for natural cubic splines

Starting from the truncated power basis for a cubic splines:

$$b_j(x) = \begin{cases} x^0 = 1 & j = 1 \\ x^1 = x & j = 2 \\ d_{j-2}(x) - d_{K-1}(x) & j = 3, \ldots, K \end{cases}$$

where

$$d_{j-2}(x) = \frac{(x - l_{j-2})_+^3 - (x - l_K)_+^3}{l_K - l_{j-2}}$$

and

$$(r)_+^3 = \begin{cases} r^3 & r \geq 0 \\ 0 & r < 0 \end{cases}$$

Note that natural cubic splines admit constant and linear functions as special cases. Using these basis:

$\Rightarrow$ $h(x)$ is constant if $\theta_j = 0$ for $j = 2, \ldots, K$

$\Rightarrow$ $h(x)$ is linear if $\theta_j = 0$ for $j = 3, \ldots, K$

Penalized LS estimation - matrix notation - 1

When the number of unique values for x_i ($i = 1, \ldots, n$) is equal to n

$$h(x_i) = \sum_{j=1}^{n} \theta_j b_j(x_i) \quad i = 1, \ldots, n$$

$\boldsymbol{\theta}$ n-dimensional vector with unknown parameters

$\mathbf{N}$ $n \times n$ matrix containing the values of the n basis evaluated on each sample unit

$$\mathbf{N} = \begin{bmatrix} b_1(x_1) & \ldots & b_n(x_1) \\ \vdots & \ddots & \vdots \\ b_1(x_n) & \ldots & b_n(x_n) \end{bmatrix}$$

$$\Rightarrow \quad \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_n) \end{bmatrix} = \mathbf{N}\boldsymbol{\theta}$$

Penalized LS estimation - matrix notation - 2

When the number of unique values for x_i ($i = 1, \ldots, n$) is equal to n

If $h(x_i)$ is a natural cubic spline with fixed knots at the unique values for x_i ($i = 1, \ldots, n$), it is possible to prove that

$$\int \left[h''(t)\right]^2 dt = \boldsymbol{\theta}^\top \mathbf{P} \boldsymbol{\theta}$$

where $\mathbf{P}$ is an $n \times n$ symmetric matrix whose entries depend only on the differences between consecutive values of x_i (the actual formulas are omitted)

$\Rightarrow$ the penalized LS criterion $pls_\lambda(h(\cdot))$ can be re-expressed as follows:

$$pls_\lambda(h(\cdot)) = pls_\lambda(\boldsymbol{\theta}) = (\mathbf{y} - \mathbf{N}\boldsymbol{\theta})^\top (\mathbf{y} - \mathbf{N}\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}^\top \mathbf{P} \boldsymbol{\theta}$$

$\Rightarrow$ finding the function $h(\cdot)$ minimizing $pls_\lambda(h(\cdot))$ is equivalent to finding the vector $\boldsymbol{\theta}$ minimizing $pls_\lambda(\boldsymbol{\theta})$

Penalized LS estimates

It is possible to prove that

$$\hat{\mathbf{t}}_\lambda = \underset{\mathbf{t}\in\mathbb{R}^n}{\operatorname{argmin}}\, pls_\lambda(\mathbf{t}) = \left(\mathbf{N}^\top\mathbf{N} + \lambda\mathbf{P}\right)^{-1}\mathbf{N}^\top\mathbf{y}$$

$\Rightarrow$ the estimated conditional expected values $\hat{h}_\lambda(x_i)$ are given by

$$\begin{bmatrix} \hat{h}_\lambda(x_1) \\ \vdots \\ \hat{h}_\lambda(x_n) \end{bmatrix} = \hat{\mathbf{h}}_\lambda = \mathbf{N}\hat{\mathbf{t}}_\lambda = \underbrace{\mathbf{N}\left(\mathbf{N}^\top\mathbf{N} + \lambda\mathbf{P}\right)^{-1}\mathbf{N}^\top}_{\mathbf{S}_\lambda}\mathbf{y}$$

$\Rightarrow$ $\hat{\mathbf{h}}_\lambda$ is an example of linear smoother, obtained using the smoothing matrix $\mathbf{S}_\lambda$

The subscript λ has been added to emphasize the fact that the values of these estimates depend on the specific value of the smoothing parameter

Choice of the smoothing parameter

In the smoothing spline approach the problem of selecting the number and the location of the knots is bypassed

HOWEVER, it is apparent that the smoothing parameter λ plays a crucial role in governing the goodness of fit and the complexity of the estimated regression function

$\Rightarrow$ Leave-One-Out Cross-Validation

$$LOOCV(\lambda) = \frac{1}{n}\sum_i \left(y_i - \hat{h}_\lambda^{[-i]}(x_i)\right)^2$$

$\hat{h}_\lambda^{[-i]}(x_i)$ estimate of $\mathrm{E}\left[Y_i|x_{1i},\ldots,x_{pi}\right]$ obtained after excluding the i-th unit from the observed sample (*independent from* i)

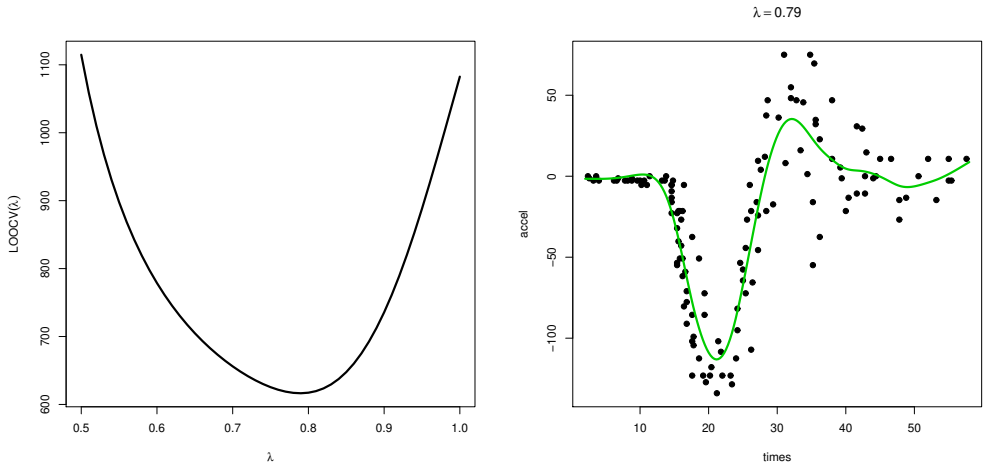
It is possible to prove that

$$y_i - \hat{h}_\lambda^{[-i]}(x_i) = \frac{y_i - \hat{h}_\lambda(x_i)}{1 - \mathbf{S}_{\lambda,ii}}$$

$\mathbf{S}_{\lambda,ii}$ i-th element of the main diagonal of $\mathbf{S}_\lambda$

LOOCV for smoothing splines can be computed without repeating the fitting process n times

Crash test data - optimal value for λ - $LOOCV$



Generalized Cross-Validation

As an alternative to $LOOCV$, some authors suggest the minimization of the following criterion

$$GCV(\lambda) = \frac{1}{n}\sum_i \left(\frac{y_i - \hat{h}_\lambda(x_i)}{1 - \frac{\mathrm{Tr}\,(\mathbf{S}_\lambda)}{n}} \right)^2 = \frac{n}{(n - \mathrm{Tr}\,(\mathbf{S}_\lambda))^2} \sum_i \left(y_i - \hat{h}_\lambda(x_i) \right)^2$$

where $\mathrm{Tr}\,(\cdot)$ is the trace operator (*sum of the diagonal elements*)

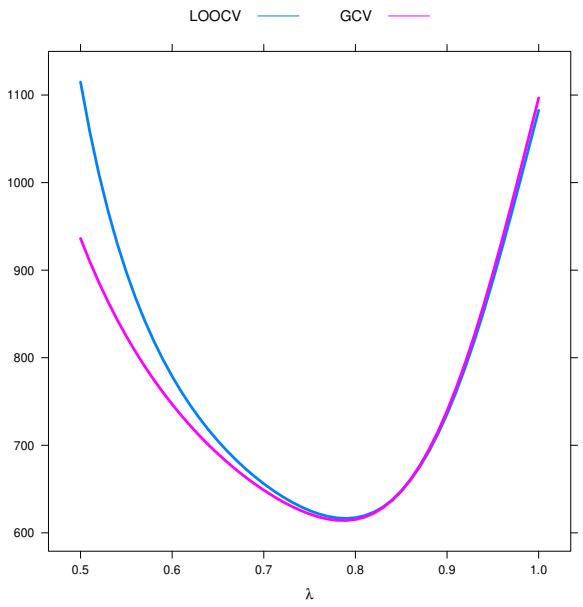
It is possible to prove that:

- $\lambda \to +\infty \Rightarrow \mathrm{Tr}\,(\mathbf{S}_\lambda) \to 2$
- $\lambda \to 0 \Rightarrow \mathrm{Tr}\,(\mathbf{S}_\lambda) \to n$

$\mathrm{Tr}\,(\mathbf{S}_\lambda) = \mathrm{edf}_\lambda$ is used as a measure of the effective degrees of freedom of $\hat{h}_\lambda(\cdot)$

Although $\hat{h}_\lambda(\cdot)$ depends on n parameters, the presence of the penalty term in the penalized LS criterion imposes some restrictions on the estimated parameters (the effective dimension of $\hat{\mathbf{t}}_\lambda$ is lower than n)

Crash test data - optimal value for λ - GCV vs. $LOOCV$



In this example, there is a substantial agreement between the two criteria

Penalized estimation - some remarks

Although the theoretical results linking penalized estimation to splines requires

- least squares as a measure of goodness of fit
- natural cubic splines with knots at unique values of x_i $(i = 1, \ldots, n)$
- a penalization term based on second partial derivatives

The use of penalized estimation can be extended beyond smoothing splines

$\Rightarrow$ *different goodness of fit measures can be used (e. g.: loglikelihood functions)*

$\Rightarrow$ *cubic spline with* $1 << K << n$ *knots can be considered*

$\Rightarrow$ *other penalization schemes can be introduced*

Gaussian regression models based on P-splines

Basic idea:

In the context of Gaussian models with cubic regression splines, an alternative strategy to overcome problems related to the selection of the number and of the location of the K knots could be obtained by:

1. choosing a relative large number of equally spaced knots (for example, $K = 20$ or $K = 40$)
2. defining a penalized/regularized log-likelihood function measuring the roughness of the resulting cubic spline

Penalty terms for P-splines (1)

When B-spline basis functions are used, common choices for the penalty term are represented by:

- (squared) first-order differences

$$J_1(\boldsymbol{\theta}) = \sum_{j=2}^{K+4} (\theta_j - \theta_{j-1})^2$$

$\Rightarrow \quad J_1(\boldsymbol{\theta}) = 0 \Leftrightarrow \sum_j \theta_j b_j(x)$ *is constant in* x

- (squared) second-order differences

$$J_2(\boldsymbol{\theta}) = \sum_{j=3}^{K+4} [(\theta_j - \theta_{j-1}) - (\theta_{j-1} - \theta_{j-2})]^2 = \sum_{j=3}^{K+4} (\theta_j - 2\theta_{j-1} + \theta_{j-2})^2$$

$\Rightarrow \quad J_2(\boldsymbol{\theta}) = 0 \Leftrightarrow \sum_j \theta_j b_j(x)$ *is linear in* x

The joint use of B-spline basis functions and penalty terms based on differences leads to the so-called P-spline approach

Penalty terms for P-splines (2)

Both $J_1(\boldsymbol{\theta})$ and $J_2(\boldsymbol{\theta})$ admit a matrix representation:

$J_1(\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{D}_1^\top \mathbf{D}_1 \boldsymbol{\theta} = \boldsymbol{\theta}^\top \mathbf{P}_1 \boldsymbol{\theta}$

$J_2(\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{D}_1^\top \mathbf{D}_1^\top \mathbf{D}_1 \mathbf{D}_1 \boldsymbol{\theta} = \boldsymbol{\theta}^\top \mathbf{P}_2 \boldsymbol{\theta}$

where

$$\mathbf{D}_1 = \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 \\ 0 & 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & -1 & 1 \end{bmatrix} \quad (K+3) \times (K+4) \text{ matrix}$$

Penalized log-likelihood function

$$pl_\lambda\left(\boldsymbol{\theta}, \sigma^2\right) \propto -\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right)^\top \left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right) - \frac{\lambda}{2} \boldsymbol{\theta}^\top \mathbf{P} \boldsymbol{\theta}$$

where

- $\boldsymbol{\theta}$ $(K+4)$-dimensional vector with unknown parameters
- $\mathbf{X}$ $n \times (K+4)$ matrix containing the values of the $K+4$ B-spline basis evaluated on each sample unit

$$\mathbf{X} = \begin{bmatrix} b_1(x_1) & \ldots & b_{K+4}(x_1) \\ \vdots & \ddots & \vdots \\ b_1(x_n) & \ldots & b_{K+4}(x_n) \end{bmatrix}$$

The subscript distinguishing penalty terms defined by first-order differences and second-order differences has been dropped for the sake of simplicity

Relevant quantities related to $\boldsymbol{\theta}$ derived from $pl_\lambda\left(\boldsymbol{\theta}, \sigma^2\right)$

- penalized score function

$$U_\lambda(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} pl_\lambda\left(\boldsymbol{\theta}, \sigma^2\right) = \frac{\mathbf{X}^\top\left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right)}{\sigma^2} - \lambda\mathbf{P}\boldsymbol{\theta} = U(\boldsymbol{\theta}) - \lambda\mathbf{P}\boldsymbol{\theta}$$

- penalized (observed/expected) Fisher information

$$i_\lambda(\boldsymbol{\theta}) = I_\lambda(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top} pl_\lambda\left(\boldsymbol{\theta}, \sigma^2\right) = \frac{\mathbf{X}^\top\mathbf{X}}{\sigma^2} + \lambda\mathbf{P} = I(\boldsymbol{\theta}) + \lambda\mathbf{P}$$

Penalized ML estimation

It is possible to prove that maximizing the penalized log-likelihood $pl_\lambda\left(\boldsymbol{\theta}, \sigma^2\right)$ with respect to θ is equivalent to minimizing the following penaliszed least squares criterion:

$$\left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right)^\top\left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right) + \tilde{\lambda}\boldsymbol{\theta}^\top\mathbf{P}\boldsymbol{\theta}$$

where $\tilde{\lambda} = \lambda\sigma^2$

$$\Rightarrow \quad \hat{\mathbf{t}}_{\tilde{\lambda}} = \left(\mathbf{X}^\top\mathbf{X} + \tilde{\lambda}\mathbf{P}\right)^{-1}\mathbf{X}^\top\mathbf{y}$$

$$\Rightarrow \quad \hat{\mathbf{h}}_{\tilde{\lambda}} = \mathbf{X}\hat{\mathbf{t}}_{\tilde{\lambda}} = \underbrace{\mathbf{X}\left(\mathbf{X}^\top\mathbf{X} + \tilde{\lambda}\mathbf{P}\right)^{-1}\mathbf{X}^\top}_{\mathbf{S}_{\tilde{\lambda}}}\mathbf{y}$$

It is possible to prove that:

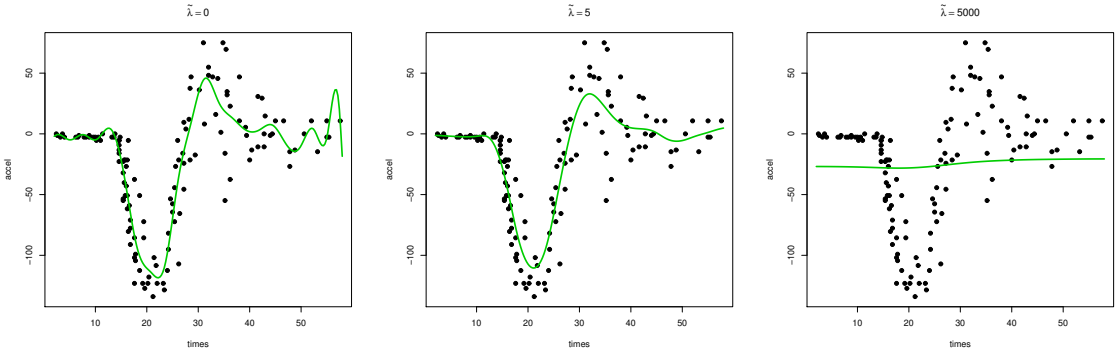
$$\tilde{\lambda} = 0 \Rightarrow \mathrm{Tr}\left(\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\right) = K + 4$$

$$\tilde{\lambda} \to +\infty \quad \Rightarrow \quad \mathrm{Tr}\left(\mathbf{X}\left(\mathbf{X}^\top\mathbf{X} + \tilde{\lambda}\mathbf{P}_1\right)^{-1}\mathbf{X}^\top\right) \to 1$$

$$\Rightarrow \quad \mathrm{Tr}\left(\mathbf{X}\left(\mathbf{X}^\top\mathbf{X} + \tilde{\lambda}\mathbf{P}_2\right)^{-1}\mathbf{X}^\top\right) \to 2$$

Crash test data - P-splines - penalized ML estimation (1)

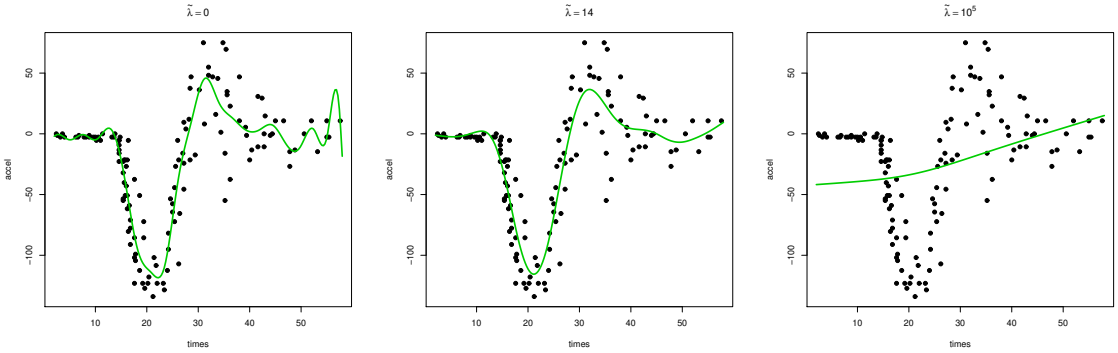
$K = 20$, penalty: $\mathbf{P}_1$ (*squared first-order differences*)



$\hat{h}_{\tilde{\lambda}}(\cdot)$ approaches a constant as $\tilde{\lambda} \to +\infty$

Crash test data - P-splines - penalized ML estimation (2)

$K = 20$, penalty: $\mathbf{P}_2$ (*squared second-order differences*)



$\hat{h}_{\tilde{\lambda}}(\cdot)$ approaches a linear function as $\tilde{\lambda} \to +\infty$

Estimation of σ^2

An estimate of σ^2 can be obtained using the following expression:

$$s^2 = \frac{\sum_i \left(y_i - \hat{h}_{\tilde{\lambda}}(x_i)\right)^2}{n - \mathrm{Tr}\left(\mathbf{S}_{\tilde{\lambda}}\right)}$$

$\Rightarrow$ some authors suggest replacing the effective degrees of freedom $\mathrm{Tr}\left(\mathbf{S}_{\tilde{\lambda}}\right)$ with the equivalent number of parameters: $2\mathrm{Tr}\left(\mathbf{S}_{\tilde{\lambda}}\right) - \mathrm{Tr}\left(\mathbf{S}_{\tilde{\lambda}}\mathbf{S}_{\tilde{\lambda}}\right)$

Choice of the smoothing parameter

The selection of the optimal value for $\tilde{\lambda}$ can be based on a model selection criterion

- $LOOCV$

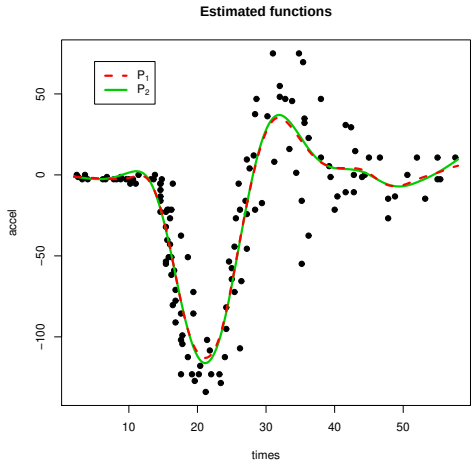
  $\Rightarrow$ it is not necessary to refit the model n times

- GCV
- AIC or BIC

  $\Rightarrow$ the maximum likelihood estimate for σ^2 is needed: $\hat{s}^2 = \frac{1}{n}\sum_i \left(y_i - \hat{h}_{\tilde{\lambda}}(x_i)\right)^2$

  $\Rightarrow$ number of parameter: $\mathrm{Tr}\left(\mathbf{S}_{\tilde{\lambda}}\right) + 1$

Crash test data - P-splines - optimal $\tilde{\lambda}$



Penalty	K	$\tilde{\lambda}$	$\mathrm{Tr}\left(\mathbf{S}_{\tilde{\lambda}}\right)$	GCV	AIC
first-order diff.	20	3.369	12.275	569.540	1222.092
second-order diff.	20	12.182	11.414	562.227	1220.542

Inference for P-splines

Model assumptions:

$$\boldsymbol{Y}|\mathbf{X} \sim MVN_n\left(\mathbf{h}, \sigma^2\mathbf{I}_n\right)$$

$\mathbf{h} = (h(x_1), \ldots, h(x_n))^\top$ n-dimensional vector containing the (unknown) true conditional expected values $\mathrm{E}\left[Y_i|x_i\right]$

Usually, in the context of nonlinear regression, the interest is in the function $\hat{h}(\cdot)$ as a whole rather than in single parameters θ_j

$\Rightarrow$ $\hat{\boldsymbol{h}}_{\tilde{\lambda}}|\mathbf{X} = \mathbf{S}_{\tilde{\lambda}}\boldsymbol{Y}|\mathbf{X} \sim MVN_n\left(\mathbf{S}_{\tilde{\lambda}}\mathbf{h}, \sigma^2\mathbf{S}_{\tilde{\lambda}}\mathbf{S}_{\tilde{\lambda}}\right)$

In general:

$$\mathrm{E}\left[\hat{\boldsymbol{h}}_{\tilde{\lambda}}|\mathbf{X}\right] = \mathbf{S}_{\tilde{\lambda}}\mathbf{h} \neq \mathbf{h}$$

$\Rightarrow$ *P-splines are biased estimators for the unknown function* $h(\cdot)$*:* $\mathbf{h} - \mathbf{S}_{\tilde{\lambda}}\mathbf{h} \neq \mathbf{0}_n$

$\Rightarrow$ *this bias is due both to the constraints implicitly imposed by the penalty term,and to the fact that splines are used as an approximation to* $h(\cdot)$*, but it is usually small/negligible*

$\Rightarrow$ *this distributional result can be exploited to draw approximate inferential conclusions about* $h(\cdot)$

Hypothesis testing for P-splines

The approximate distributional results for $\hat{\mathbf{h}}_\lambda$ can be exploited to test some hypothesis on $h(\cdot)$:

- independence assumption

  $\Rightarrow$ $H_0 : h(\cdot)$ is a constant function

- linearity assumption

  $\Rightarrow$ $H_0 : h(\cdot)$ is a linear function

Due to the fact that constant and linear functions are nested in cubic splines, it is possible to perform an approximate likelihood ratio test/Wald test statistic (usually resulting in approximate F test), after expressing these hypothesis in terms of linear restrictions on the cubic spline coefficients (omitted)

WARNING: the resulting p-values rely on several approximations and do not take into account the uncertainty related to the choice of the smoothing parameter. In particular, they tend to underestimate the actual p-values (anticonservative)

It is also possible to derive approximate pointwise confidence bands for $h(\cdot)$

Crash test data - P-splines - R output (1)

gam function (package: mgcv) - $K = 20$, second-order differences penalty

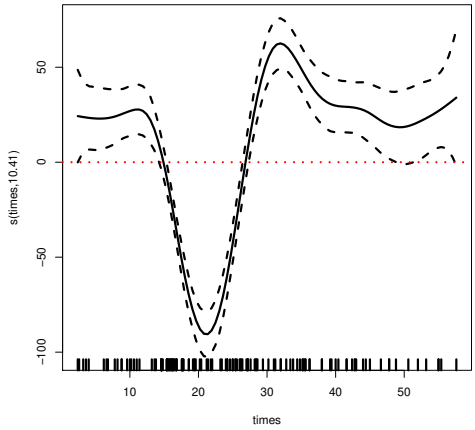
```
Parametric coefficients:
             Estimate  Std.  Error  t value  Pr(>|t|)
 (Intercept)   -25.546        1.966  -12.995     0.000

Approximate significance of smooth terms:
              edf  Ref.df       F  p-value
 s(times)  10.414  12.438  37.198    0.000

R-sq.(adj) = 0.78 Deviance explained = 79.7%
GCV = 562.23 Scale est.  = 513.98 n = 133
```

The estimated function $\hat{h}_{\tilde{\lambda}}$ is decomposed in two parts:

- *the estimated intercept $\Rightarrow$ 1 degrees of freedom*
- *a (nonconstant) function centred around 0 $\Rightarrow$ $\mathrm{Tr}\left(\mathbf{S}_{\tilde{\lambda}}\right) - 1$ effective degrees of freedom*

  $\Rightarrow$ *according to the approximate p-value, one may conclude that the estimated function differ significantly from a constant one*

Crash test data - P-splines - R output (2)



Approximate pointwise confidence bands can be associated to the (centred) estimated function

⇒ *nota that in the constant line at 0 is not contained in the confidence band (consistently with the conclusion drawn in the previous slide)*

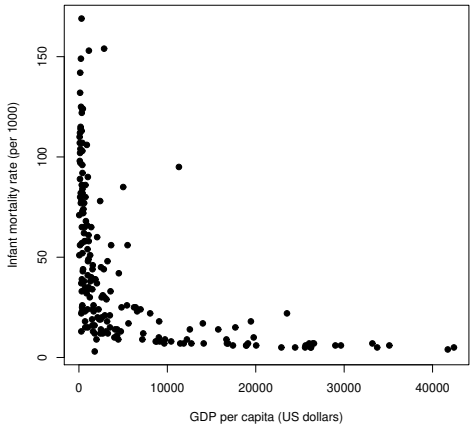
Gaussian regression models and variable transformations: an introduction

Infant mortality vs GDP - 1

A researcher is interested in evaluating the effects of economic conditions on mortality, starting from information about 193 countries. In particular, for each country, the observed quantities are:

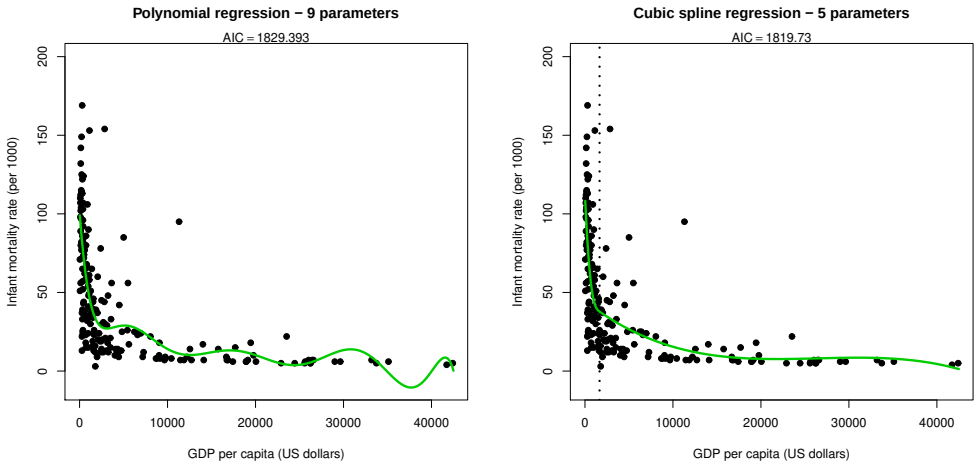
- infant mortality rate (per 1000 live births)
- GDP per capita (US Dollars)

Infant mortality vs GDP - 2



The plot shows a clear nonlinear dependence pattern

Infant mortality vs GDP - polynomial & cubic spline regression



Transformable nonlinearity 5

Multiplicative models: an example

Y_i Random variable that describes the value for the dependent variable observed on the $i-$th sample unit $(i = 1, \ldots, n)$

x_i value of the regressor for the $i-$th sample unit

$$Y_i = \alpha x_i^{\beta_1} \exp(\varepsilon_i), \quad \varepsilon_i | x_i \sim N(0, \sigma^2) \quad \text{IID}$$

Note that this model differs from the previous ones because:

- *it does not have an additive error term*
- *it involves a nonlinear transformation of* ε_i

$\Rightarrow$ *Non-Gaussian regression model for* $Y_i|x_i$ *(*$Y_i|x_i$ *does not have a Gaussian distribution)*

Transformable nonlinearity

Although

$Y_i = h(x_i, \varepsilon_i; \boldsymbol{\theta}) = \alpha x_i^{\beta_1} \exp(\varepsilon_i), \quad \varepsilon_i | x_i \sim N(0, \sigma^2) \quad \text{IID}$

is not linear in (some of) the unknown parameteres, there could exist a function $g(\cdot)$ such that

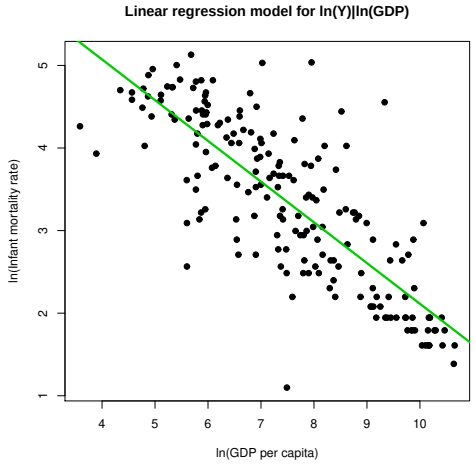
$g(Y_i) = \beta_0 + \beta_1 b(x_i) + \varepsilon_i, \quad \varepsilon_i | x_i \sim N(0, \sigma^2) \quad \text{IID}$

In particular, if $Y_i > 0$ $(i = 1, \ldots, n)$ and $\alpha > 0$, then

$$\begin{aligned} \ln Y_i &= \ln\left[\alpha x_i^{\beta_1} \exp(\varepsilon_i)\right] \\ &= \underbrace{\ln \alpha}_{\beta_0} + \beta_1 \underbrace{\ln x_i}_{b(x_i)} + \varepsilon_i \end{aligned}$$

$\Rightarrow$ Gaussian regression model for $\ln Y_i | x_i$ that is **linear in the parameters**

ln(Infant mortality) vs ln(GDP) - linear regression




```
Coefficients:
                  Estimate   Std. Error   t value   Pr(>|t|)
      (Intercept)     7.045        0.199    35.379      0.000
 log(GDPperCapita)   -0.493        0.026   -19.070      0.000

Residual standard error:  0.5938 on 191 degrees of freedom
Multiple R-squared:  0.6556, Adjusted R-squared:  0.6538
F-statistic:  363.7 on 1 and 191 DF, p-value:  < 2.2e-16
```

Lognormal distributions - 1

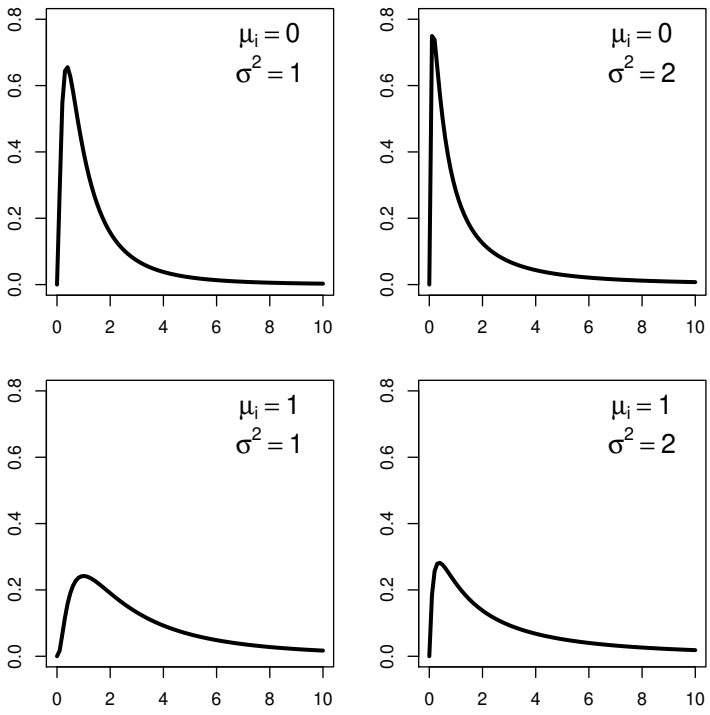
$Y_i > 0$ non-negative independent random variables ($i = 1, \ldots, n$)

$\ln Y_i \sim N(\mu_i, \sigma^2) \Longleftrightarrow Y_i \sim \ln N(\mu_i, \sigma^2)$

$f(y_i; \mu_i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{y_i} \exp\left[-\frac{(\ln y_i - \mu_i)^2}{2\sigma^2}\right]$

$y_i \in (0, +\infty)$
$\mu_i \in (-\infty, +\infty)$
$\sigma^2 \in (0, +\infty)$

Lognormal distributions - 2



Use of Lognormal distributions

Statistical phenomena that:

- take only positive values
- show skewed distributions

Some examples:

- *intensities/densities*
- *durations/waiting times*
- *earnings/expenditures*

Moments of Lognormal distributions

$$\begin{aligned}
\mathrm{E}[Y_i] &= \exp\left(\mu_i + \tfrac{\sigma^2}{2}\right) \\
&> \exp(\mu_i) = \exp(\mathrm{E}[\ln Y_i]) \\
&\quad \text{Jensen inequality} \Rightarrow \ln \mathrm{E}[Y_i] \geq \mathrm{E}[\ln Y_i] \\
\mathrm{Var}[Y_i] &= \exp\left[2\left(\mu_i + \sigma^2\right)\right]\left[1 - \exp\left(-\sigma^2\right)\right] \\
&= \left\{\mathrm{E}[Y_i]\right\}^2\left[\exp\left(\sigma^2\right) - 1\right] \propto \left\{\mathrm{E}[Y_i]\right\}^2
\end{aligned}$$

Although σ^2 does not depend on i, the n random variables are not homoscedastic

However:

$$\mathrm{CV}[Y_i] = \frac{\sqrt{\mathrm{Var}[Y_i]}}{\mathrm{E}[Y_i]} = \frac{\mathrm{E}[Y_i]\sqrt{\exp\left(\sigma^2\right)-1}}{\mathrm{E}[Y_i]} = \sqrt{\exp\left(\sigma^2\right) - 1}$$

the n random variables have the same coefficient of variation (CV)

Gaussian linear models for logarithmic transformations - 1

$\ln Y_i | x_{1i} \dots x_{pi} \sim N(\beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}, \sigma^2)$

$i = 1, \dots, n$ independent

$\Updownarrow$

$Y_i | x_{1i} \dots x_{pi} \sim \ln N(\beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}, \sigma^2)$

$i = 1, \dots, n$ independent

$$\begin{aligned} \mathrm{E}\left[Y_i | x_{1i} \dots x_{pi}\right] &= \exp\left(\beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + \tfrac{\sigma^2}{2}\right) \\ &> \exp\left(\beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}\right) = \exp\left(\mathrm{E}\left[\ln Y_i | x_{1i} \dots x_{pi}\right]\right) \\ \mathrm{Var}\left[Y_i | x_{1i} \dots x_{pi}\right] &\propto \left\{\mathrm{E}\left[Y_i | x_{1i} \dots x_{pi}\right]\right\}^2 \end{aligned}$$

Gaussian linear models for logarithmic transformations - 2

(Gaussian linear) model with additive error term for the logarithmic transformation:

$\ln Y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$

$\Updownarrow$

$Y_i = \exp\left(\beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + \epsilon_i\right), \quad \epsilon_i \sim N(0, \sigma^2)$

$\Updownarrow$

$Y_i = \exp\left(\beta_0\right) \cdot \exp\left(\beta_1 x_{1i}\right) \cdot \dots \cdot \exp\left(\beta_p x_{pi}\right) \cdot \exp\left(\epsilon_i\right), \ \exp\left(\epsilon_i\right) \sim \ln N(0, \sigma^2)$

(Lognormal nonlinear) model with multiplicative error term for the original variable)

Loglikelihood

Consider the original dependent variable:

$$
\begin{aligned}
l_{\ln N}(\boldsymbol{\beta}, \sigma^2|Y) &= \sum_{i=1}^{n} \left\{ \ln \left[\frac{1}{\sqrt{2\pi\sigma^2}} \right] + \ln \left[\frac{1}{y_i} \right] - \frac{\left(\ln y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi} \right)^2}{2\sigma^2} \right\} \\
&= -\sum_{i=1}^{n} \ln y_i + \\
&\quad -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 + \frac{\sum_{i=1}^{n} \left(\ln y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi} \right)^2}{2\sigma^2} \\
&= -\sum_{i=1}^{n} \ln y_i + l_N(\boldsymbol{\beta}, \sigma^2 | \ln Y)
\end{aligned}
$$

Apart from an additive constant that does not involve the model parameters, loglikelihoods for (conditional) independent Lognormal distributions are equivalent to loglikelihoods for (conditional) independent Gaussian distributions (after applying the logarithm to the original variables)
$\Longrightarrow$ *same inferential results and properties as for Gaussian linear models*

Estimation for conditional expected values

Estimation for the logarithmic transformation:

$$\widehat{\ln y_i} = \mathrm{E}\left[\ln Y_i | \widehat{x_{1i} \ldots x_{pi}} \right] = \hat{b}_0 + \hat{b}_1 x_{1i} + \ldots + \hat{b}_p x_{pi}$$

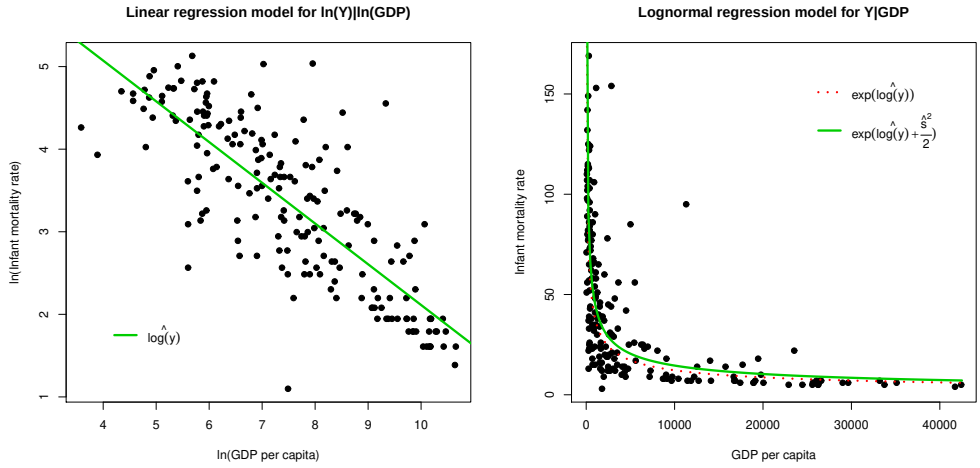
Estimation for the original variable:

$$\widehat{y_i} = \mathrm{E}\left[Y_i | \widehat{x_{1i} \ldots x_{pi}} \right] = \exp\left[\hat{b}_0 + \hat{b}_1 x_{1i} + \ldots + \hat{b}_p x_{pi} + \frac{s^2}{2} \right] > \exp\left[\widehat{\ln y_i} \right]$$

The inverse (exponential) function applied to the estimated conditional expected value of the logarithmic transfomation leads to biased estimates for the conditional expected value of the original variable

It is possible to prove that $\exp\left[\widehat{\ln y_i} \right]$ $(i = 1, \ldots, n)$ *are unbiased estimates for the conditional medians of the original variable*

Infant mortality vs GDP - lognormal regression



Model comparison criteria for lognormal regression models

AIC and BIC can be also exploited to perform model selection among models with different assumptions about the conditional distribution of the dependent variable

$$\begin{aligned} AIC_{\ln N}(M|Y) &= -2l_{\ln N}(\hat{\mathbf{b}}, \hat{s}|Y) + 2(p+2) \\ &= -2\left[-\sum_{i=1}^n \ln y_i + l_N(\hat{\mathbf{b}}, \hat{s}|\ln Y)\right] + 2(p+2) \\ &= -2l_N(\hat{\mathbf{b}}, \hat{s}|\ln Y) + 2(p+2) + 2\sum_{i=1}^n \ln y_i \\ &= AIC_N(M|\ln Y) + 2\sum_{i=1}^n \ln y_i \end{aligned}$$

$$BIC_{\ln N}(M|Y) = BIC_N(M|\ln Y) + 2\sum_{i=1}^n \ln y_i$$

Comparisons among AIC (or BIC) values are admissible if and only if these model comparison criteria are computed with reference to the same random sample

⇒ *it does not make sense to compare tha AIC/BIC of a Gaussian regression model fitted on $\ln Y$ with the AIC/BIC of a Gaussian regression model fitted on Y*

Infant mortality vs GDP - Final comparison among models

Cond. distribution	GDP effect	n. of param.	log-likelihood	AIC	BIC
Gaussian	polynomial	10	-903.696	1829.393	1865.282
Gaussian	cubic spline	6	-903.865	1819.730	1839.306
Lognormal	nonlinear*	3	-816.171	1638.342	1648.130

* the lognormal regression model is fitted using $\ln(GDP)$

Heteroschedasticity in Gaussian linear models: variance-stabilising transformations

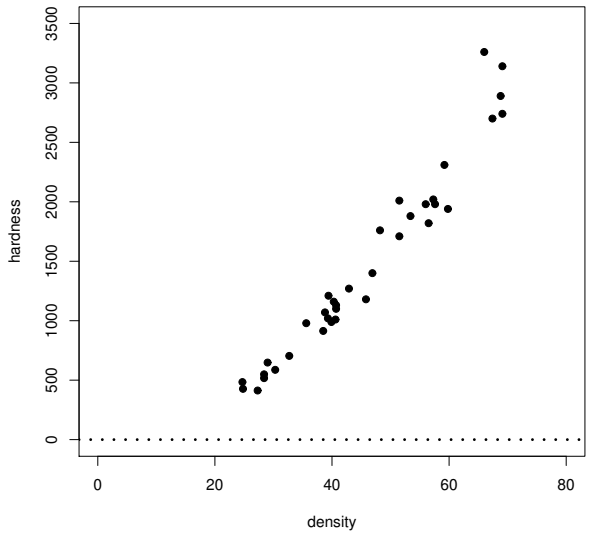
A motivating example

`hardness`: hardness of an hardwood timber (Y)

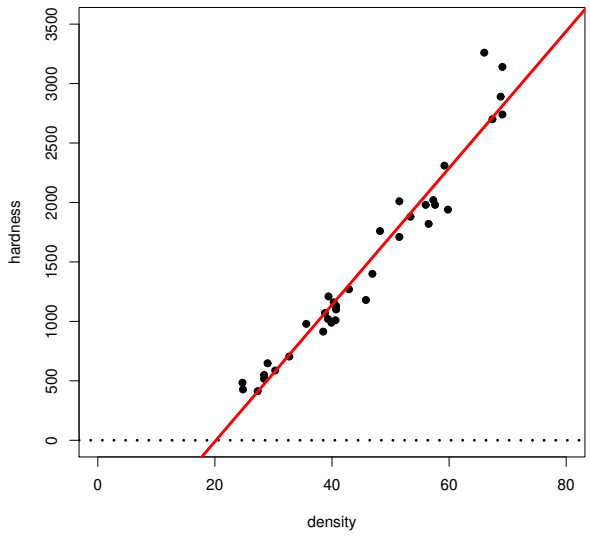
`density`: density of an hardwood timber (X)

$n = 36$

Timber data

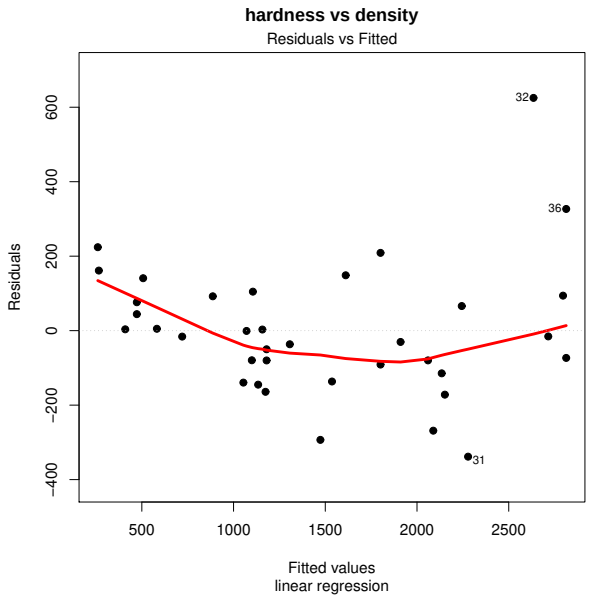


Timber data - Gaussian linear regression - 1



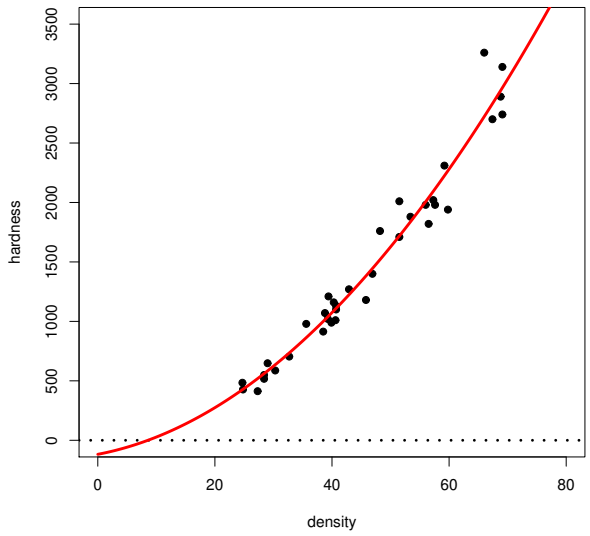
WARNING: hardness cannot take negative values

Timber data - Gaussian linear regression - 2



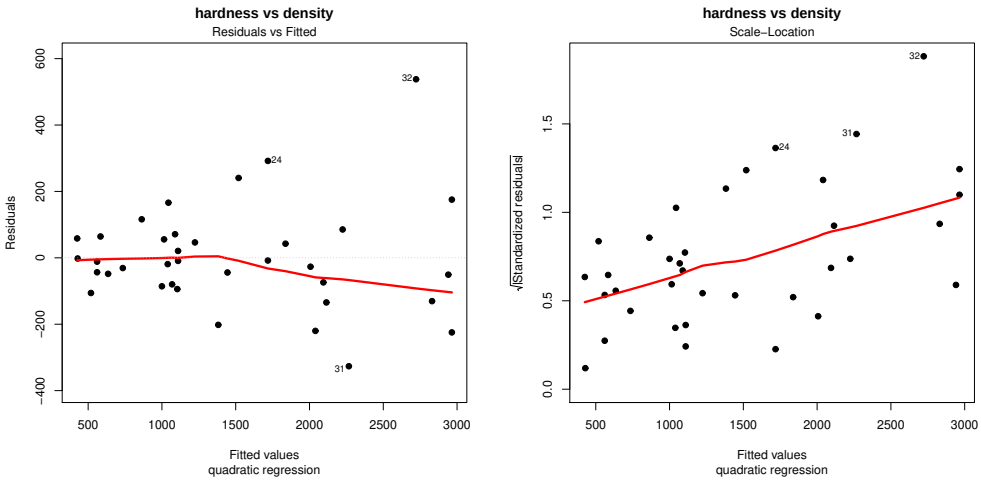
The residuals seems to show a pattern in their average value

Timber data - Gaussian quadratic regression - 1



WARNING: hardness cannot take negative values

Timber data - Gaussian quadratic regression - 2



The inclusion of a quadratic effect seems reasonable, but the model is still inadequate

⇒ *there is a clear pattern in the magnitude of the standardised residuals: it tends to increase as the fitted value increases. This is a symptom of heteroschedasticity of the conditional distributions*

Variance-stabilising transformations

Consider a Gaussian regression model with heteroschedastic conditional distributions

$\mathrm{Var}\left[Y_i|x_{1i},\ldots,x_{pi}\right]=\sigma_i^2\ \forall i$

When the dependent variables Y_i take only positive values, it is possible to prove that:

- $\sigma_i^2\propto \mathrm{E}\left[Y_i|x_{1i},\ldots,x_{pi}\right]\Longrightarrow \mathrm{Var}\left[\sqrt{Y_i}|x_{1i},\ldots,x_{pi}\right]\cong\sigma^2$ CONSTANT
- $\sigma_i^2\propto \left(\mathrm{E}\left[Y_i|x_{1i},\ldots,x_{pi}\right]\right)^2\Longrightarrow \mathrm{Var}\left[\ln\left(Y_i\right)|x_{1i},\ldots,x_{pi}\right]\cong\sigma^2$
- $\sigma_i^2\propto \left(\mathrm{E}\left[Y_i|x_{1i},\ldots,x_{pi}\right]\right)^3\Longrightarrow \mathrm{Var}\left[Y_i^{-0.5}|x_{1i},\ldots,x_{pi}\right]\cong\sigma^2$
- $\sigma_i^2\propto \left(\mathrm{E}\left[Y_i|x_{1i},\ldots,x_{pi}\right]\right)^4\Longrightarrow \mathrm{Var}\left[Y_i^{-1}|x_{1i},\ldots,x_{pi}\right]\cong\sigma^2$
- ...

Box-Cox transformation

$$Y_i^*=\frac{Y_i^\lambda-1}{\lambda},\quad \lambda\in\mathbb{R}$$

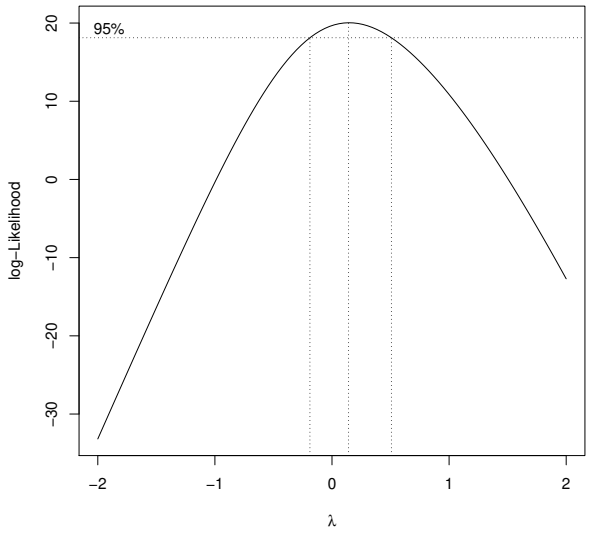
$$\lim_{\lambda\to 0}\frac{Y_i^\lambda-1}{\lambda}=\ln\left(Y_i\right)$$

Choice of $\lambda\Longrightarrow$ maximum likelihood estimation:

$$l\left(\boldsymbol{\beta},\sigma^2,\lambda\right)=-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\frac{y_i^\lambda-1}{\lambda}-\beta_0-\beta_1x_{1i}-\ldots-\beta_px_{pi}\right)^2$$

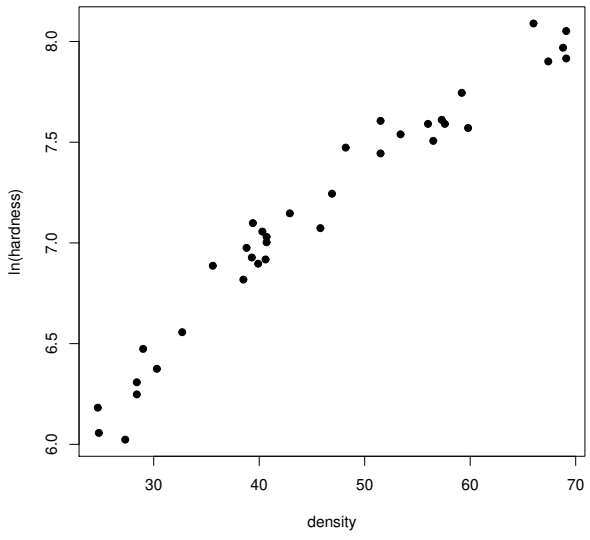
Technical difficulties: there is not an analytical formula for computing $\hat{\lambda}$
$\Longrightarrow\hat{\lambda}$ must be found numerically: for a grid of candidate values, the corresponding log-likelihoods are computed and compared (there will be specific estimates for $\boldsymbol{\beta}$ e σ^2 associated with each point in the grid)

Timber data - choice of λ

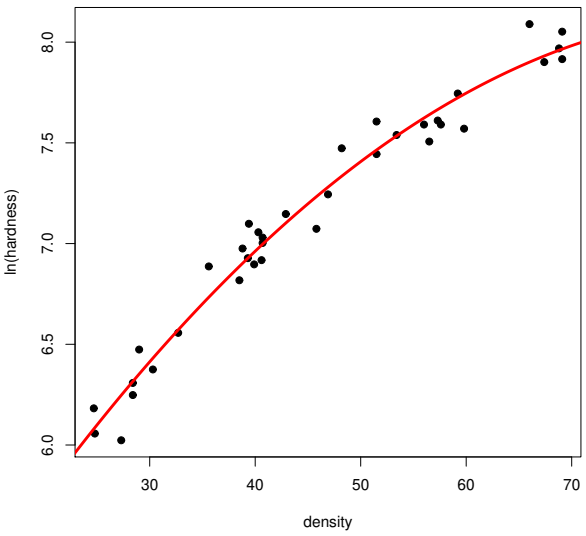


The plot suggests a value for λ close to 0
$\Longrightarrow$ *logarithmic transformation*

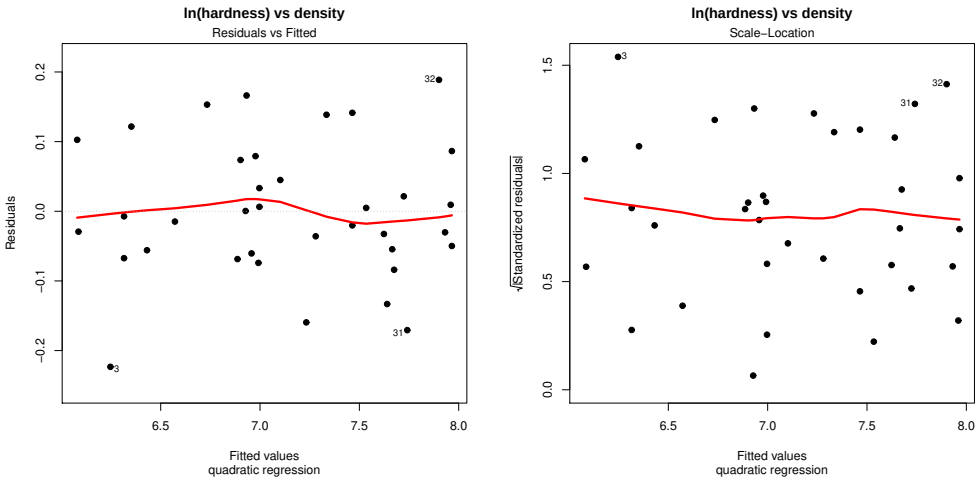
Substitution $\ln(Y)$ for Y



Timber data - Lognormal quadratic regression - 1

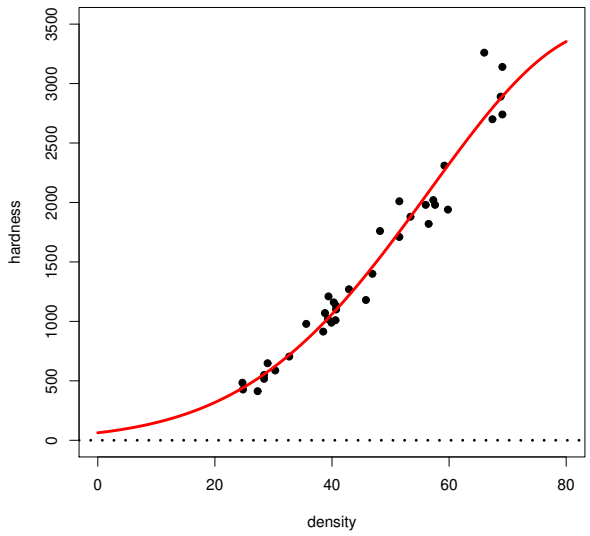


Timber data - Lognormal quadratic regression - 2



No evident pattern in the average of the residuals and in the magnitude of the standardised residuals

Timber data - Lognormal quadratic regression - 3



Variance-stabilising transformations - cautionary remarks

- the Box-Cox transformation is only an example of variance-stabilising transformation
  - ⇒ alternative transformations have been proposed to dela with dependent variables that can also take negative values
- the nonlinear nature of the transformation poses nontrivial issues if one is interested in obtaining information about the original dependent variable
  - ⇒ generally, the form of the conditional distribution for the original variable is not known (the only exception being the logarithmic transformation)

Generalised linear models
Introduction and general definition

An example in insurance

An insurance company is interested in evaluating the main risk factors that affect the number car accident claims made by its policyholders.
In particular, the following factors are investigated:

- age of the policyholder (4 classes)
- district of residence of the policyholder (4 districts)
- type of car (4 classes)

Baxter, L.A., Coutts, S.M. and Ross, G.A.F. (1980). Application of linear models in motor insurance, Proceedings of the 21st International Congress of Actuaries, Zurich, 11–29

Dataset structure

	n	c	age	dist	car
1	197	38	<25	rural	<1
2	284	63	<25	rural	1-1.5
3	133	19	<25	rural	1.5-2
4	24	4	<25	rural	>2
5	85	22	<25	small towns	<1
6	149	25	<25	small towns	1-1.5
7	66	14	<25	small towns	1.5-2
8	9	4	<25	small towns	>2
9	35	5	<25	large towns	<1
10	53	10	<25	large towns	1-1.5
11	24	8	<25	large towns	1.5-2
12	7	3	<25	large towns	>2
13	20	2	<25	major cities	<1
14	31	7	<25	major cities	1-1.5
15	18	5	<25	major cities	1.5-2
16	3	0	<25	major cities	>2
⋮	⋮	⋮	⋮	⋮	⋮

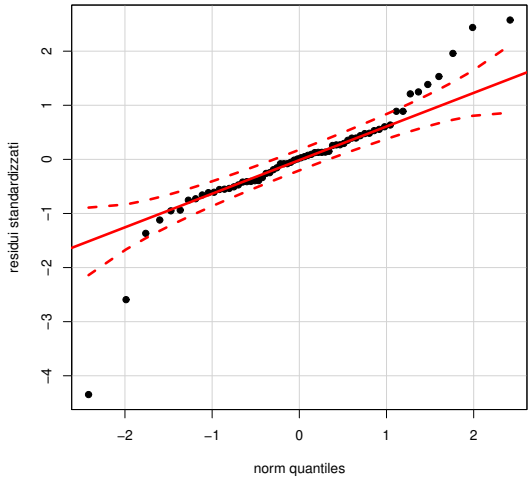
Definition of the dependent variable

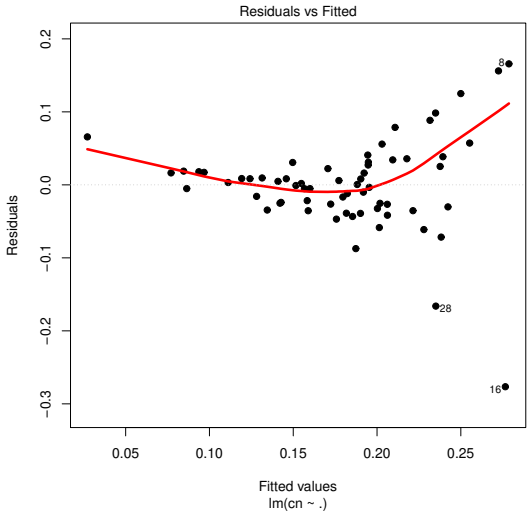
$\mathtt{n} \in \mathbb{R}^+$ number of policyholders characterised by a given covariate pattern - by a given combination of values for the three factors (*exposure level*)

$\mathtt{c} \in \mathbb{N}$ number of car accident claims made by policyholders characterised by a given covariate pattern (*number of events - count variable*)

$Y = \frac{\mathtt{c}}{\mathtt{n}}$ (average) number of claims per policyholder (*rate - number of events per unit of exposure*)

Gaussian linear model: graphical residual analysis





An example in Economics

In a study on women labour force participation, the following information was collected on a random sample of swiss women:

- non-labour income (logarithmic scale)
- age class (decades)
- education level (years)
- number of young children (under 7 years of age)
- number of old children (over 7 years of age)
- nationality (two classes: foreign/non-foreign)

Gerfin, M. (1996). Parametric and Semi-Parametric Estimation of the Binary Response Model of Labour Market Participation. Journal of Applied Econometrics, 11, 321–339

Dataset structure

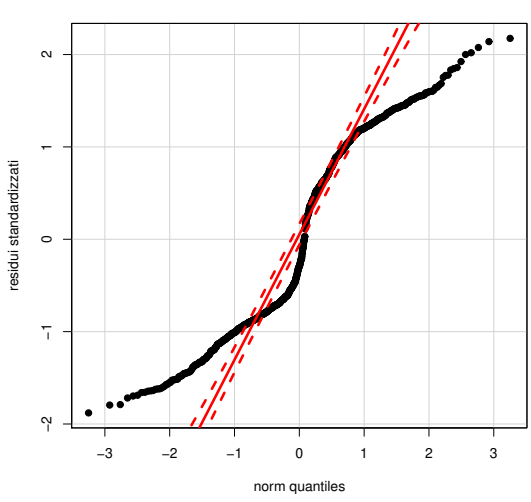
	participation	income	age	education	youngkids	oldkids	foreign
1	no	10.79	3.00	8.00	1.00	1.00	no
2	yes	10.52	4.50	8.00	0.00	1.00	no
3	no	10.97	4.60	9.00	0.00	0.00	no
4	no	11.10	3.10	11.00	2.00	0.00	no
5	no	11.11	4.40	12.00	0.00	2.00	no
6	yes	11.03	4.20	12.00	0.00	1.00	no
7	no	11.45	5.10	8.00	0.00	0.00	no
8	yes	10.49	3.20	8.00	0.00	2.00	no
9	no	10.62	3.90	12.00	0.00	0.00	no
10	no	10.49	4.30	11.00	0.00	2.00	no
11	no	10.66	4.50	11.00	0.00	2.00	no
12	no	10.47	6.00	12.00	0.00	0.00	no
13	no	11.23	3.30	11.00	2.00	0.00	no
14	no	11.91	5.60	14.00	0.00	0.00	no
15	no	11.50	5.60	11.00	0.00	0.00	no
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

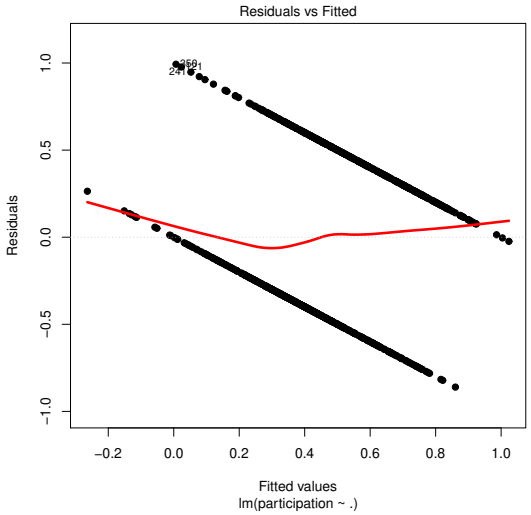
Definition of the dependent variable

$\texttt{participation} \in \{\texttt{yes}, \texttt{no}\}$ Does the individual participate in the labor force?
- Does the individual have a job or is the individual looking for a job?
(*dichotomous variable*)

$$Y = \mathbf{1}\{\texttt{participation} = \texttt{yes}\} = \begin{cases} 0 & \text{if } \texttt{participation} = \texttt{no} \\ 1 & \text{if } \texttt{participation} = \texttt{yes} \end{cases}$$

Gaussian linear model: graphical residual analysis





Limitations of Gaussian linear models

These two examples show some of the "intrinsic" limitations of Gaussian linear models. These limitations are due to:

- features of the dependent variable (support of the random variable used to describe its values)

$\Rightarrow$ *distributional form of* Y

$\Rightarrow$ *coherence of the parameter space*

Furthermore, some other limitations arisen from the examples:

- linearity
- homoschedasticity

$\Rightarrow$ *Sometimes (but not always) these limitations can be overcome by suitably transforming the dependent variable and/or the regressors*

Generalised linear models (GLM)

Class of regression models (including Gaussian linear models as special cases) that allows to overcome some of the previously introduced limitations, that often arise in many practical situations

This class allows a unified approach to describe many specific regression models that where independently developed

This unified approach provide a general common inferential framework (estimation/hypothesis testing) that is "almost" equal to the one associated with Gaussian linear models

Exponential families of distributions of order 1

Y Random variable with probability mass/density function $f(y;\theta)$

$\theta \in \Theta \subseteq \mathbb{R}$ (unknown) parameter

$Y \sim \mathrm{EF}\,(b(\theta))$

The distribution of Y belongs to an exponential family (of order 1) with natural parameter $b(\theta)$ if and only if:

$f(y;\theta) = \exp\{a(y)b(\theta) + c(\theta) + d(y)\}$

- $a(\cdot)$ e $d(\cdot)$ known functions depending only on y
- $b(\cdot)$ e $c(\cdot)$ known functions depending only on θ

The exponential family is expressed in **canonical form** if $a(\cdot)$ is the identity function:

$f(y;\theta) = \exp\{yb(\theta) + c(\theta) + d(y)\}$

Poisson random variables

- $Y \sim \mathrm{Poi}(\theta) \quad y \in \mathbb{N}, \quad \theta \in \mathbb{R}^+$

  $f(y,\theta) = \dfrac{\theta^y \exp(-\theta)}{y!} = \exp\{y\ln\theta - \theta - \ln y!\}$

  - $a(y) = y$, $d(y) = -\ln y!$
  - $b(\theta) = \ln\theta$, $c(\theta) = -\theta$

Bernoulli random variables

- $Y \sim \text{Ber}(\theta) \quad y \in \{0,1\}, \quad \theta \in [0,1]$

$$f(y,\theta) = \theta^y (1-\theta)^{1-y} = \exp\{y \ln\theta + \ln(1-\theta) - y\ln(1-\theta)\}$$

$$= \exp\left\{y \ln\frac{\theta}{1-\theta} + \ln(1-\theta)\right\}$$

  - $a(y) = y$, $d(y) = 0$
  - $b(\theta) = \ln\dfrac{\theta}{1-\theta}$, $c(\theta) = \ln(1-\theta)$

Nuisance parameters and weights

$f(y;\theta,\phi,w)$

$\theta \in \Theta \subseteq \mathbb{R}$ (unknown) parameter
$\phi \in \Phi \subseteq \mathbb{R}^+$ (unknown) parameter
$w \in \mathbb{R}^+$ known quantity

$Y \sim \text{EF}(b(\theta), \phi, w)$

The distribution of Y belongs to an exponential family (of order 1) expressed in canonical form with natural parameter $b(\theta)$, nuisance parameter ϕ and weight w if and only if:

$$f(y;\theta,\phi,w) = \exp\left\{\frac{w}{\phi}[yb(\theta) + c(\theta)] + d(y,\phi,w)\right\}$$

- $b(\cdot)$ e $c(\cdot)$ known functions that do not depend on y and w
- $d(\cdot)$ known function that does not depend on θ

Gaussian random variables

$Y \sim \mathrm{N}(\theta, \phi) \quad y \in \mathbb{R}, \quad \theta \in \mathbb{R},\ \phi \in \mathbb{R}^+$

$$f(y;\theta,\phi) = \frac{1}{\sqrt{1\pi\phi}} \exp\left\{-\frac{(y-\theta)^2}{2\phi}\right\} = \exp\left\{\frac{1}{\phi}\left[y\theta - \frac{\theta^2}{2}\right] + \left[-\frac{y^2}{2\phi} - \frac{\ln 2\pi\phi}{2}\right]\right\}$$

- $w = 1$
- $b(\theta) = \theta$, $c(\theta) = -\dfrac{\theta^2}{2}$
- $d(y,\phi,w) = -\dfrac{y^2}{2\phi} - \dfrac{\ln 2\pi\phi}{2}$

Number of events per unit of exposure

$Y^* \sim \mathrm{Poi}(w\theta) \quad y^* \in \mathbb{N}, \quad \theta \in \mathbb{R}^+,\ w \in \mathbb{R}^+$

$Y = \dfrac{Y^*}{w}$ Number of events per unit of exposure

$\Rightarrow Y^* = wY$

$$f(y;\theta,w) = \frac{(w\theta)^{wy}\exp(-w\theta)}{(wy)!} = \exp\left\{w\left[y\ln\theta - \theta\right] + wy\ln w - \ln(wy)!\right\}$$

- $\phi = 1$
- $b(\theta) = \ln\theta$, $c(\theta) = -\theta$
- $d(y,\phi,w) = wy\ln w - \ln(wy)!$

The relative frequency of successes in Bernoulli trials

$Y^* \sim \mathrm{Bin}(w, \theta) \quad y^* \in \{0, 1, \ldots, w\}, \quad \theta \in [0, 1],\ w \in \mathbb{N}^+$

$Y = \dfrac{Y^*}{w}$ The relative frequency of successes in w Bernoulli trials

$\Rightarrow Y^* = wY$

$$f(y; \theta, w) = \binom{w}{wy} \theta^{wy}(1-\theta)^{w(1-y)} = \exp\left\{ w \left[y \ln \frac{\theta}{1-\theta} + \ln(1-\theta) \right] + \ln \binom{w}{wy} \right\}$$

- $\phi = 1$
- $b(\theta) = \ln \dfrac{\theta}{1-\theta}$, $c(\theta) = \ln(1-\theta)$
- $d(y, \phi, w) = \ln \binom{w}{wy}$

Some relevant properties of exponential families

$Y \sim \mathrm{EF}(b(\theta), \phi, w)$

- $\mathrm{E}[Y] = -\dfrac{c'(\theta)}{b'(\theta)}$

  $b'(\theta) = \dfrac{\partial}{\partial \theta} b(\theta)$, $c'(\theta) = \dfrac{\partial}{\partial \theta} c(\theta)$

- $\mathrm{Var}[Y] = \dfrac{\phi}{w} \dfrac{\mathrm{E}'[Y]}{b'(\theta)} = \dfrac{\phi}{w} \dfrac{c'(\theta) b''(\theta) - c''(\theta) b'(\theta)}{[b'(\theta)]^3}$

  $\mathrm{E}'[Y] = \dfrac{\partial}{\partial \theta} \mathrm{E}[Y]$, $b''(\theta) = \dfrac{\partial^2}{\partial \theta^2} b(\theta)$, $c''(\theta) = \dfrac{\partial^2}{\partial \theta^2} c(\theta)$

Example 1: Poisson random variables

- $b'(\theta) = \frac{\partial}{\partial\theta} \ln\theta = \frac{1}{\theta}$, $c'(\theta) = \frac{\partial}{\partial\theta}(-\theta) = -1$

  $\mathrm{E}[Y] = -\frac{-1}{\frac{1}{\theta}} = \theta$

- $b''(\theta) = \frac{\partial}{\partial\theta}\frac{1}{\theta} = -\frac{1}{\theta^2}$, $c''(\theta) = \frac{\partial}{\partial\theta}(-1) = 0$, $\phi = 1$, $w = 1$

  $\mathrm{Var}[Y] = \frac{1}{1}\frac{-1 \cdot -\frac{1}{\theta^2} - \frac{1}{\theta} \cdot 0}{\left[\frac{1}{\theta}\right]^3} = \frac{\theta^3}{\theta^2} = \theta$

Example 2: number of events per unit of exposure

- $\mathrm{E}[Y] = -\frac{-1}{\frac{1}{\theta}} = \theta$
- $\mathrm{Var}[Y] = \frac{1}{w}\frac{-1 \cdot -\frac{1}{\theta^2} - \frac{1}{\theta} \cdot 0}{\left[\frac{1}{\theta}\right]^3} = \frac{\theta^3}{w\theta^2} = \frac{\theta}{w}$

Example 3: Gaussian random variables

- $b'(\theta) = \dfrac{\partial}{\partial\theta}\theta = 1$, $c'(\theta) = \dfrac{\partial}{\partial\theta}\left(-\dfrac{\theta^2}{2}\right) = -\theta$

  $\mathrm{E}[Y] - \dfrac{-\theta}{1} = \theta$

- $b''(\theta) = \dfrac{\partial}{\partial\theta}1 = 0$, $c''(\theta) = \dfrac{\partial}{\partial\theta}(-\theta) = -1$, $w = 1$

  $\mathrm{Var}[Y] = \dfrac{\phi}{1}\dfrac{(-\theta)\cdot 0 - 1\cdot(-1)}{[1]^3} = \phi$

Generalised linear models 25

Definition of a generalised linear model (GLM)

- Y_i r.v. that describes the possible value of the dependent variable on the i-th sample unit ($i = 1, \ldots, n$)
- $\mathbf{x}_i = (x_{0i}, x_{1i}, \ldots, x_{pi})^\top$ $p+1$-dimensional vector containing the values of the regressors for the i-th sample unit ($x_{0i} = 1$ $\forall i$ constant regressor associated with the model intercept)

A generalised linear model (GLM) is a statistical model for the random sample $\boldsymbol{Y}$

characterised by:

$\Rightarrow$ a **probabilistic component**

conditional probability mass/density funcion for Y_i, given the regressors

$\Rightarrow$ a **systematic/deterministic component**

functional relationship between the regressors and the (conditional) expevted value of Y_i

Probabilistic component of a GLM

- $Y_i|\mathbf{x}_i \sim \mathrm{EF}(b(\theta_i), \phi, w_i)$
  - the conditional distributions belong to the same exponential family expressed in canonical form with natural parameter $b(\theta)$
  - the functions $b(\cdot)$, $c(\cdot)$ and $d(\cdots)$ are the same for all Y_i (they do not depend on i)
  - each conditional distribution can take a different value for the unknown parameter θ_i
  - each conditional distribution can have a different (known) weight w_i
  - all the conditional distributions share the same value for the nuisance parameter ϕ, that can be either known or unknown
- (conditional) independence among the n r. v. $Y_1|\mathbf{x}_1, \ldots, Y_n|\mathbf{x}_n$

Systematic/deterministic component of a GLM

- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^{p+1}$ vector of unknown parameters
- $\eta_i = \beta_0 x_{0i} + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} = \mathbf{x}_i^\top \boldsymbol{\beta}$ linear predictor
- $g\left(\mathrm{E}\left[Y_i|\mathbf{x}_i\right]\right) = \eta_i$ link function
  - with known functional form
  - differentiable
  - invertible, with $g^{-1}(\cdot) = h(\cdot)$

    $\mathrm{E}\left[Y_i|\mathbf{x}_i\right] = h\left(\eta_i\right)$

Gaussian linear models

$\mathbf{Y}|\mathbf{X} \sim NMV_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right)$

- Probabilistic component
  - $Y_i|\mathbf{x}_i \sim N(\theta_i, \phi) \qquad i = 1, \ldots, n$
  - (conditional) independence among the n r. v. $Y_1|\mathbf{x}_1, \ldots, Y_n|\mathbf{x}_n$
- Systematic component
  - $\eta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} \qquad i = 1, \ldots, n$
  - $\theta_i = \mathrm{E}\left[Y_i|\mathbf{x}_i\right] = \eta_i \quad \Rightarrow \quad g(\cdot)$ identity function

Poisson regression models

- Probabilistic component
  - $Y_i|\mathbf{x}_i \sim \mathrm{Poi}(\theta_i) \qquad i = 1, \ldots, n$
  - (conditional) independence among the n r. v. $Y_1|\mathbf{x}_1, \ldots, Y_n|\mathbf{x}_n$
- Systematic component
  - $\ln\left(\mathrm{E}\left[Y_i|\mathbf{x}_i\right]\right) = \eta_i \quad \Rightarrow \quad \mathrm{E}\left[Y_i|\mathbf{x}_i\right] = \exp\left(\eta_i\right)$

A class of regression models for rates

- Probabilistic component
  - $Y_i^*|\mathbf{x}_i \sim \text{Poi}(w_i\theta_i) \qquad i = 1, \ldots, n$
  - $Y_i = \dfrac{Y_i^*}{w_i}$
  - (conditional) independence among the n r. v. $Y_1|\mathbf{x}_1, \ldots, Y_n|\mathbf{x}_n$
- Systematic component
  - $\ln\left(\text{E}\left[Y_i|\mathbf{x}_i\right]\right) = \eta_i \quad \Rightarrow \quad \text{E}\left[Y_i|\mathbf{x}_i\right] = \exp\left(\eta_i\right)$

    $\Rightarrow \quad \ln\left(\text{E}\left[Y_i^*|\mathbf{x}_i\right]\right) = \ln\left(w_i\text{E}\left[Y_i|\mathbf{x}_i\right]\right) = \underbrace{\eta_i + \ln\left(w_i\right)}_{\eta_i^*}$

    $\Rightarrow \quad \text{E}\left[Y_i^*|\mathbf{x}_i\right] = \exp\left(\eta_i^*\right) = w_i \exp\left(\eta_i\right)$

    η_i^* linear predictor in Poisson regression models for Y_i^*, with an **offset** (compensating term - *regressor with regression coefficient set to 1*) equal to $\ln\left(w_i\right)$

Logistic regression models

- Probabilistic component
  - $Y_i^*|\mathbf{x}_i \sim \text{Bin}(w_i, \theta_i) \qquad i = 1, \ldots, n$
  - $Y_i = \dfrac{Y_i^*}{w_i}$
  - (conditional) independence among the n r. v. $Y_1|\mathbf{x}_1, \ldots, Y_n|\mathbf{x}_n$
- Systematic component
  - $\text{logit}\left(\text{E}\left[Y_i|\mathbf{x}_i\right]\right) = \ln \dfrac{\text{E}\left[Y_i|\mathbf{x}_i\right]}{1 - \text{E}\left[Y_i|\mathbf{x}_i\right]} = \eta_i \quad \Rightarrow \quad \text{E}\left[Y_i|\mathbf{x}_i\right] = \dfrac{\exp\left(\eta_i\right)}{1 + \exp\left(\eta_i\right)}$

    $\Rightarrow \quad \text{E}\left[Y_i^*|\mathbf{x}_i\right] = w_i \dfrac{\exp\left(\eta_i\right)}{1 + \exp\left(\eta_i\right)}$

Choice of the link function

Each probabilistic component is characterised by a specific set of possible value for $\mathrm{E}[Y_i|\mathbf{x}_i]$

$\mathrm{E}[Y_i|\mathbf{x}_i] \in \Omega \subseteq \mathbb{R}$

the range of values for $h(\cdot) = g^{-1}(\cdot)$ should coincide with Ω

$h(\cdot) : \mathbb{R} \longmapsto \Omega$

Canonical link functions

Each exponential family has a specific link function (known as the **canonical** link function), that is obtained by equating the natural parameter to the linear predictor:

$b(\theta_i) = \eta_i$

- Gaussian r. vs.
  $\Rightarrow \theta_i = \eta_i$ identity function
- Poisson r. vs. / number of events per unit of exposure
  $\Rightarrow \ln \theta_i = \eta_i$ logarithm
- Bernoulli r. vs./ relative frequency
  $\Rightarrow \ln \dfrac{\theta_i}{1-\theta_i} = \eta_i$ logistic function

Some comments about the use of canonical link functions

$\Rightarrow$ the use of canonical link functions lead to some analytical simplifications and to some theoretical properties

$\Rightarrow$ not necessarily canonical link functions are the best choice

- they could not be adequate to describe the effect of the regressors on $\mathrm{E}[Y_i|\mathbf{x}_i]$
- when $b(\theta_i) \in \Psi \subset \mathbb{R}$, there could be some compatibility problems with the range of η_i

GLM and Gaussian linear models

GLMs allow to overcome some limitations of Gaussian linear models

- use of probability mass/density functions that differ from the Gaussian one
  - with supports that differ from $\mathbb{R}$
  - with (conditional) variance that is not constant (linked to the conditional expected values)
- non-linearity of the conditional expected values (wrt both the regressors and the parameters $\beta_0, \beta_1, \ldots, \beta_p$
  - use of link functions that differ from the identity one

One of the key features of generalised linear models is the possibility to match any probabilistic componet with any systematic component (at least in principle)

Log-likelihood for a GLM

$$g\left(-\frac{c'(\theta_i)}{b'(\theta_i)}\right) = \eta_i \iff -\frac{c'(\theta_i)}{b'(\theta_i)} = h(\eta_i)$$

$\Rightarrow$ There is an implicit functional relationship betweeen θ_i and the values of the regressors, with a known functional form and unknown parameters $\beta_0, \ldots, \beta_p$

$l(\beta_0, \ldots, \beta_p, \phi) = \sum_{i=1}^{n} l_i(\beta_0, \ldots, \beta_p, \phi)$

where

$$\begin{aligned} l_i(\beta_0, \ldots, \beta_p, \phi) &= \ln f(y_i; \theta_i, \phi, w_i) \\ &= \frac{w_i}{\phi}\left[y_i b(\theta_i) + c(\theta_i)\right] + d(y_i, \phi, w_i) \end{aligned}$$

Score function of a GLM (1)

$$\begin{aligned} U_j(\boldsymbol{\beta}) &= \frac{\partial l(\beta_0, \ldots, \beta_p, \phi)}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \frac{\partial l_i(\beta_0, \ldots, \beta_p, \phi)}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \frac{\partial l_i(\beta_0, \ldots, \beta_p, \phi)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mathrm{E}[Y_i|\mathbf{x}_i]} \frac{\partial \mathrm{E}[Y_i|\mathbf{x}_i]}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \end{aligned}$$

Score function of a GLM (2)

$$\begin{aligned}
\frac{\partial l_i(\beta_0, \ldots, \beta_p, \phi)}{\partial \theta_i} &= \frac{\partial}{\partial \theta_i} \left\{ \frac{w_i}{\phi} \left[y_i b(\theta_i) + c(\theta_i) \right] + d(y_i, \phi, w_i) \right\} \\
&= \frac{w_i}{\phi} \left[y_i b'(\theta_i) + c'(\theta_i) \right] \\
&= \frac{w_i}{\phi} b'(\theta_i) \left[y_i + \frac{c'(\theta_i)}{b'(\theta_i)} \right] \\
&= \frac{w_i}{\phi} b'(\theta_i) \left\{ y_i - \mathrm{E}\left[Y_i | \mathbf{x}_i \right] \right\}
\end{aligned}$$

Score function of a GLM (3)

Recall that

$$\mathrm{Var}\left[Y_i | \mathbf{x}_i \right] = \frac{1}{b'(\theta)} \frac{\phi}{w_i} \frac{\partial \mathrm{E}\left[Y_i | \mathbf{x}_i \right]}{\partial \theta_i}$$

Thus:

$$\frac{\partial \theta_i}{\partial \mathrm{E}\left[Y_i | \mathbf{x}_i \right]} = \frac{1}{\frac{\partial \mathrm{E}\left[Y_i | \mathbf{x}_i \right]}{\partial \theta_i}} = \frac{1}{\frac{w_i}{\phi} b'(\theta_i) \mathrm{Var}\left[Y_i | \mathbf{x}_i \right]}$$

furthermore:

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ji}$$

Score function of a GLM (4)

$$
\begin{aligned}
U_j(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{\partial l_i(\beta_0, \ldots, \beta_p, \phi)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mathrm{E}\left[Y_i|\mathbf{x}_i\right]} \frac{\partial \mathrm{E}\left[Y_i|\mathbf{x}_i\right]}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \\
&= \sum_{i=1}^{n} \frac{\frac{w_i}{\phi} b'(\theta_i) \left\{y_i - \mathrm{E}\left[Y_i|\mathbf{x}_i\right]\right\}}{\frac{w_i}{\phi} b'(\theta_i) \mathrm{Var}\left[Y_i|\mathbf{x}_i\right]} \frac{\partial \mathrm{E}\left[Y_i|\mathbf{x}_i\right]}{\partial \eta_i} x_{ji} \\
&= \sum_{i=1}^{n} \frac{y_i - \mathrm{E}\left[Y_i|\mathbf{x}_i\right]}{\mathrm{Var}\left[Y_i|\mathbf{x}_i\right]} \frac{\partial \mathrm{E}\left[Y_i|\mathbf{x}_i\right]}{\partial \eta_i} x_{ji}
\end{aligned}
$$

Score function of a Gaussian linear model

$\mathbf{Y}|\mathbf{X} \sim NMV_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n\right)$

Recall that, for $i = 1, \ldots, n$:

- $\mathrm{E}\left[Y_i|\mathbf{x}_i\right] = \eta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} \quad \Rightarrow \quad \frac{\partial \mathrm{E}\left[Y_i|\mathbf{x}_i\right]}{\partial \eta_i} = 1$
- $\mathrm{Var}\left[Y_i|\mathbf{x}_i\right] = \sigma^2$

$$
\begin{aligned}
U_j(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{(y_i - \eta_i)}{\sigma^2} \cdot 1 \cdot x_{ji} \\
&= \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi}\right) x_{ji}
\end{aligned}
$$

Observed Fisher information of a GLM (1)

$$
\begin{aligned}
i_{jl}(\boldsymbol{\beta}) &= -\frac{\partial^2}{\partial\beta_j\beta_l} l(\beta_0,\ldots,\beta_p,\phi) = -\frac{\partial}{\partial\beta_l} U_j(\boldsymbol{\beta}) \\
&= -\frac{\partial}{\partial\beta_l} \sum_{i=1}^{n} \frac{x_{ji}}{\mathrm{Var}\left[Y_i|\mathbf{x}_i\right]} \frac{\partial \mathrm{E}\left[Y_i|\mathbf{x}_i\right]}{\partial\eta_i} \left\{y_i - \mathrm{E}\left[Y_i|\mathbf{x}_i\right]\right\} \\
&= -\sum_{i=1}^{n} \frac{x_{ji}}{\mathrm{Var}\left[Y_i|\mathbf{x}_i\right]} \frac{\partial \mathrm{E}\left[Y_i|\mathbf{x}_i\right]}{\partial\eta_i} \frac{\partial}{\partial\beta_l} \left\{y_i - \mathrm{E}\left[Y_i|\mathbf{x}_i\right]\right\} + \\
&\quad -\sum_{i=1}^{n} \left\{y_i - \mathrm{E}\left[Y_i|\mathbf{x}_i\right]\right\} \frac{\partial}{\partial\beta_l} \left(\frac{x_{ji}}{\mathrm{Var}\left[Y_i|\mathbf{x}_i\right]} \frac{\partial \mathrm{E}\left[Y_i|\mathbf{x}_i\right]}{\partial\eta_i} \right)
\end{aligned}
$$

Observed Fisher information of a GLM (2)

Since

$$
\frac{\partial}{\partial\beta_l} \left\{y_i - \mathrm{E}\left[Y_i|\mathbf{x}_i\right]\right\} = -\frac{\partial \mathrm{E}\left[Y_i|\mathbf{x}_i\right]}{\partial\eta_i} \frac{\partial\eta_i}{\partial\beta_l} = -\frac{\partial \mathrm{E}\left[Y_i|\mathbf{x}_i\right]}{\partial\eta_i} x_{li}
$$

$$
\begin{aligned}
i_{jl}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{x_{ji}x_{li}}{\mathrm{Var}\left[Y_i|\mathbf{x}_i\right]} \left(\frac{\partial \mathrm{E}\left[Y_i|\mathbf{x}_i\right]}{\partial\eta_i} \right)^2 + \\
&\quad -\sum_{i=1}^{n} \left\{y_i - \mathrm{E}\left[Y_i|\mathbf{x}_i\right]\right\} \frac{\partial}{\partial\beta_l} \left(\frac{x_{ji}}{\mathrm{Var}\left[Y_i|\mathbf{x}_i\right]} \frac{\partial \mathrm{E}\left[Y_i|\mathbf{x}_i\right]}{\partial\eta_i} \right)
\end{aligned}
$$

Observed Fisher information of a Gaussian linear model

Similarly to what seen previously regarding the score function:

$$\begin{aligned} i_{jl}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{x_{ji}x_{li}}{\sigma^2}(1)^2 + \\ &\quad - \sum_{i=1}^{n} \{y_i - \mathrm{E}[Y_i|\mathbf{x}_i]\} \underbrace{\frac{\partial}{\partial \beta_l}\left(\frac{x_{ji}}{\sigma^2}\cdot 1\right)}_{=0 \quad \forall\ j} \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n} x_{ji}x_{li} \end{aligned}$$

Expected Fisher information of a GLM

Note that the expected valuesa are computed conditionally on the regressor values (in the conditional sample space)

$$\begin{aligned} I_{jl}(\boldsymbol{\beta}) &= \mathrm{E}[i_{jl}(\boldsymbol{\beta})] = \sum_{i=1}^{n} \mathrm{E}\left[\frac{x_{ji}x_{li}}{\mathrm{Var}[Y_i|\mathbf{x}_i]}\left(\frac{\partial \mathrm{E}[Y_i|\mathbf{x}_i]}{\partial \eta_i}\right)^2\right] + \\ &\quad - \sum_{i=1}^{n} \mathrm{E}\left[\{Y_i - \mathrm{E}[Y_i|\mathbf{x}_i]\}\frac{\partial}{\partial \beta_l}\left(\frac{x_{ji}}{\mathrm{Var}[Y_i|\mathbf{x}_i]}\frac{\partial \mathrm{E}[Y_i|\mathbf{x}_i]}{\partial \eta_i}\right)\right] \\ &= \sum_{i=1}^{n} \frac{x_{ji}x_{li}}{\mathrm{Var}[Y_i|\mathbf{x}_i]}\left(\frac{\partial \mathrm{E}[Y_i|\mathbf{x}_i]}{\partial \eta_i}\right)^2 + \\ &\quad - \sum_{i=1}^{n} \frac{\partial}{\partial \beta_l}\left(\frac{x_{ji}}{\mathrm{Var}[Y_i|\mathbf{x}_i]}\frac{\partial \mathrm{E}[Y_i|\mathbf{x}_i]}{\partial \eta_i}\right)\underbrace{\mathrm{E}[Y_i - \mathrm{E}[Y_i|\mathbf{x}_i]]}_{0} \\ &= \sum_{i=1}^{n} \frac{x_{ji}x_{li}}{\mathrm{Var}[Y_i|\mathbf{x}_i]}\left(\frac{\partial \mathrm{E}[Y_i|\mathbf{x}_i]}{\partial \eta_i}\right)^2 \end{aligned}$$

Canonical link function and log-likelihood

$$
\begin{aligned}
l(\beta_0, \ldots, \beta_p, \phi) &= \sum_{i=1}^{n} \frac{w_i}{\phi} \left[y_i \eta_i + c(\theta_i) \right] + d(y_i, \phi, w_i) \\
&= \sum_{i=1}^{n} \frac{w_i}{\phi} y_i \eta_i + \sum_{i=1}^{n} \frac{w_i}{\phi} c(\theta_i) + \sum_{i=1}^{n} d(y_i, \phi, w_i) \\
&= \frac{1}{\phi} \left[\beta_0 \sum_{i=1}^{n} w_i y_i x_{0i} + \ldots + \beta_p \sum_{i=1}^{n} w_i y_i x_{pi} \right] + \\
&\quad + \sum_{i=1}^{n} \frac{w_i}{\phi} c(\theta_i) + \sum_{i=1}^{n} d(y_i, \phi, w_i)
\end{aligned}
$$

$\sum_{i=1}^{n} w_i y_i x_{0i}, \ldots, \sum_{i=1}^{n} w_i y_i x_{pi}$ are (minimal) sufficient statistics for $\boldsymbol{\beta}$

Canonical link function and score function

$$
\begin{aligned}
\frac{\partial \mathrm{E}\left[Y_i|\mathbf{x}_i\right]}{\partial \eta_i} &= \frac{\partial \mathrm{E}\left[Y_i|\mathbf{x}_i\right]}{\partial \theta_i} \frac{\partial \theta_i}{\partial \eta_i} \\
&= \frac{w_i}{\phi} b'(\theta_i) \mathrm{Var}\left[Y_i|\mathbf{x}_i\right] \frac{1}{\frac{\partial \eta_i}{\partial \theta_i}} \\
&= \frac{w_i}{\phi} b'(\theta_i) \mathrm{Var}\left[Y_i|\mathbf{x}_i\right] \frac{1}{\frac{\partial b(\theta_i)}{\partial \theta_i}} \\
&= \frac{w_i}{\phi} \mathrm{Var}\left[Y_i|\mathbf{x}_i\right] \\
U_j(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{y_i - \mathrm{E}\left[Y_i|\mathbf{x}_i\right]}{\mathrm{Var}\left[Y_i|\mathbf{x}_i\right]} \frac{w_i}{\phi} \mathrm{Var}\left[Y_i|\mathbf{x}_i\right] x_{ji} \\
&= \frac{1}{\phi} \sum_{i=1}^{n} w_i \left\{ y_i - \mathrm{E}\left[Y_i|\mathbf{x}_i\right] \right\} x_{ji}
\end{aligned}
$$

Canonical link function and Fisher information

$$
\begin{aligned}
i_{jl}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{x_{ji}x_{li}}{\mathrm{Var}\left[Y_i|\mathbf{x}_i\right]} \left(\frac{w_i}{\phi}\mathrm{Var}\left[Y_i|\mathbf{x}_i\right]\right)^2 + \\
&\quad - \sum_{i=1}^{n} \left\{y_i - \mathrm{E}\left[Y_i|\mathbf{x}_i\right]\right\} \underbrace{\frac{\partial}{\partial\beta_l}\left(\frac{x_{ji}}{\mathrm{Var}\left[Y_i|\mathbf{x}_i\right]}\frac{w_i}{\phi}\mathrm{Var}\left[Y_i|\mathbf{x}_i\right]\right)}_{0} \\
&= \sum_{i=1}^{n} \left(\frac{w_i}{\phi}\right)^2 \mathrm{Var}\left[Y_i|\mathbf{x}_i\right] x_{ji}x_{li} \\
&= I_{jl}(\boldsymbol{\beta})
\end{aligned}
$$

Canonical link function and Gaussian linear models

- $\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} y_i x_{1i}, \ldots, \sum_{i=1}^{n} y_i x_{pi}$ are (minimal) sufficient statistics $\boldsymbol{\beta}$
- $U_j(\boldsymbol{\beta}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi}\right)x_{ji}$
- $i_{jl}(\boldsymbol{\beta}) = I_{jl}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(\frac{1}{\sigma^2}\right)^2 \sigma^2 x_{ji}x_{li} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_{ji}x_{li}$

Inference for a GLM

- Properties of the score function

  $\Rightarrow U(\boldsymbol{\beta}) \xrightarrow{d} NMV_{p+1}\left(\mathbf{0}_{p+1}, I(\boldsymbol{\beta})\right)$

- Maximum likelihood estimation for $\boldsymbol{\beta}$
  $\Rightarrow$ *Numerical optimization techniques*
  (Newton-Raphson and Fisher scoring algorithms)

- Properties of the maximum likelihood estimators

  $\Rightarrow \hat{\boldsymbol{B}} \xrightarrow{d} NMV_{p+1}\left(\boldsymbol{\beta}, I(\boldsymbol{\beta})^{-1}\right)$

- Goodness of fit /adequacy of the model
  $\Rightarrow$ *graphical analysis of residuals*
  $\Rightarrow$ *hypothesis testing based on the residual deviance (if ϕ is known)*

- Linear hypothesis on $\boldsymbol{\beta}$: $H_0 : \mathbf{K}\boldsymbol{\beta} = \mathbf{t}$

  $\Rightarrow$ *likelihood ratio test statistic:* $2\ln\frac{L(\hat{\mathbf{b}},\hat{\phi})}{L(\hat{\mathbf{b}}_{H_0},\hat{\phi}_{H_0})}\Big| H_0 \xrightarrow{d} \chi^2_q$

  $\Rightarrow$ *Wald test statistic:* $\left[\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right]^\top \left[\mathbf{K}\widehat{I(\boldsymbol{\beta})}^{-1}\mathbf{K}^\top\right]^{-1}\left[\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right]\Big| H_0 \xrightarrow{d} \chi^2_q$

Poisson regression models: Definition and maximum likelihood estimation

Counts as dependent variables - Examples

Some examples:

- number of car accident claims made by policyholders to an insurance company
- number of points scored by a basketball player during a regular season
- number of imperfections on a glass plate
- number of patients with a given disease hospitalised in a given town
- ...

Model definition - 1

- Y_i r.v. describing the observed value for the dependent variable on the i-th sample unit ($i = 1, \ldots, n$)
- $\mathbf{x}_i = (x_{0i}, x_{1i} \ldots, x_{pi})^\top$ $p+1$-dimensional vector containing the regressor values observed on the i-th sample unit ($x_{0i} = 1$ $\forall i$ constant regressor associated with the model intercept)
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ $p+1$-dimensional vector containing the model parameters (including the intercept)
- $\eta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} = \mathbf{x}_i^\top \boldsymbol{\beta}$ linear predictor associated with the i-th statistical unit
- $Y_i|\mathbf{x}_i \sim \mathrm{Poi}\left(\exp(\eta_i)\right)$ or, equivalently, $f\left(y_i|\mathbf{x}_i\right) = \dfrac{\exp(\eta_i)^{y_i} \exp(-\exp(\eta_i))}{y_i!}$
- $Y_i|\mathbf{x}_i$ independent ($i = 1, \ldots, n$)

Model definition - 2

- Probabilistic component

  $$f(y_i|\mathbf{x}_i) = \frac{\mu_i^{y_i}\exp(-\mu_i)}{y_i!} = \{y_i \ln \mu_i - \mu_i - \ln y_i!\}$$

  - $Y_i|\mathbf{x}_i \sim \text{EF}(\ln \mu_i, \phi = 1, w_i = 1)$
  - $\text{E}[Y_i|\mathbf{x}_i] = \mu_i$
  - $\text{Var}[Y_i|\mathbf{x}_i] = \mu_i$
- Systematic component

  $\mu_i = \exp(\eta_i) \quad \Leftrightarrow \quad \ln \mu_i = \eta_i$ Canonical link function

Count variables and exposure levels - Examples

Some examples:

- number of car accident claims made by policyholders to an insurance company

  $\Rightarrow$ *number of policyholders with a specific covariate pattern*
- number of points scored by a basketball player during a regular season

  $\Rightarrow$ *number of minutes (time) played during the regular season*
- number of imperfections on a glass plate

  $\Rightarrow$ *dimension (surface) of the glass plate*
- number of patients with a given disease hospitalised in a given town

  $\Rightarrow$ *number of inhabitants of the town* $\times$ *time length*
- ...

Introduction of offsets (compensating terms)

- w_i exposure level of the i-th sample unit
- $\eta_i^* = \eta_i + \ln w_i$ linear predictor associated with the i-th statistical unit, with offset equal to $\ln w_i$
- $Y_i|\mathbf{x}_i \sim \text{Poi}\left(\exp(\eta_i^*)\right)$ or, equivalently, $f\left(y_i|\mathbf{x}_i\right) = \dfrac{[w_i \exp(\eta_i)]^{y_i} \exp(-w_i \exp(\eta_i))}{y_i!}$
- $Y_i|\mathbf{x}_i$ independent $(i = 1, \ldots, n)$

Note that:
The inclusion of an offset in the linear predictor allows to deal with Poisson regression models and regression models for rate derived from Poisson distributions within a "unified" framework - the attention will be focused on the former in the following

Log-likelihood

$$
\begin{aligned}
l\left(\boldsymbol{\beta}\right) &= \sum_{i=1}^{n} \left[y_i \eta_i - \exp(\eta_i) - \ln(y_i!)\right] \\
&= \left[\beta_0 \sum_{i=1}^{n} y_i + \ldots + \beta_p \sum_{i=1}^{n} y_i x_{pi}\right] + \\
&\quad - \sum_{i=1}^{n} \exp(\eta_i) - \sum_{i=1}^{n} \ln y_i!
\end{aligned}
$$

$\sum_{i=1}^{n} y_i, \ldots, \sum_{i=1}^{n} y_i x_{pi}$ are (minimal) sufficient statistics for $\boldsymbol{\beta}$

Note that:
When exposure levels w_i are considered, it is possible to prove that the log-likelihood is only changed by the additive constant $\sum_{i=1}^{n} y_i \ln w_i$, that does not involve $\boldsymbol{\beta}$

Score function

Considering the general formula for the generic element of the score function for a GLM with canonical link function:

$$\begin{aligned} U_j(\boldsymbol{\beta}) &= \frac{1}{\phi}\sum_{i=1}^{n} w_i \left\{ y_i - \mathrm{E}\left[Y_i | \mathbf{x}_i \right] \right\} x_{ji} \\ &= \frac{1}{1}\sum_{i=1}^{n} 1 \cdot \left[y_i - \exp\left(\eta_i \right) \right] x_{ji} \\ &= \sum_{i=1}^{n} \left[y_i - \exp\left(\eta_i \right) \right] x_{ji} \end{aligned}$$

Note that:
When exposure levels w_i are considered, it is possible to prove that the score function remains unchanged, provided that the offset terms $\ln w_i$ are included in the linear predictors

Fisher information

Considering the general formula for the generic element of the (observed and expected) Fisher information matrix for a GLM with canonical link function:

$$\begin{aligned} i_{jl}(\boldsymbol{\beta}) &= I_{jl}(\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} \left(\frac{w_i}{\phi} \right)^2 \mathrm{Var}\left[Y_i | \mathbf{x}_i \right] x_{ji} x_{li} \\ &= \sum_{i=1}^{n} \left(\frac{1}{1} \right)^2 \exp\left(\eta_i \right) x_{ji} x_{li} \\ &= \sum_{i=1}^{n} \exp\left(\eta_i \right) x_{ji} x_{li} \end{aligned}$$

Nota bene:
When exposure levels w_i are considered, it is possible to prove that the Fisher infromation matrices remain unchanged, provided that the offset terms $\ln w_i$ are included in the linear predictors

Hessian matrix

$\Rightarrow$ (j, l)-th element of the Hessian matrix of the log-likelihood function

$$H_{jl}(\boldsymbol{\beta}) = \frac{\partial^2}{\partial \beta_j \partial \beta_l} l(\boldsymbol{\beta}) = -i_{jl}(\boldsymbol{\beta})$$

$$= -\sum_{i=1}^{n} \exp(\eta_i) x_{ji} x_{li}$$

Matrix representation (1)

- $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$ n-dimensional vector that contains the observed values of the dependent variable on the n sample units
- $\mathbf{X}$ $n \times (p+1)$ matrix that contains the observed values of the regressors on the n sample units
- $\boldsymbol{\mu} = (\exp(\eta_1), \exp(\eta_2), \ldots, \exp(\eta_n))^\top$ n-dimensional vector that contains the conditional expected values of the dependent variable for the n sample unit
- $$\mathbf{W} = \begin{bmatrix} \exp(\eta_1) & 0 & 0 & \ldots & 0 \\ 0 & \exp(\eta_2) & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \exp(\eta_n) \end{bmatrix}$$

  $n \times n$ diagonal matrix that contains the conditional variances of the dependent variable for the n sample unit on the main diagonal

Matrix representation (2)

- Score function

  $U(\boldsymbol{\beta}) = \mathbf{X}^\top [\mathbf{y} - \boldsymbol{\mu}]$

- Hessian matrix of the log-likelihood function

  $H(\boldsymbol{\beta}) = -\mathbf{X}^\top \mathbf{W} \mathbf{X}$

- (observed and expected) Fisher information matrix

  $i(\boldsymbol{\beta}) = I(\boldsymbol{\beta}) = \mathbf{X}^\top \mathbf{W} \mathbf{X}$

Maximum likelihood estimation (1)

$\hat{\mathbf{b}}$ is the maximum likelihood estimate of $\boldsymbol{\beta}$ if and only if

$l\left(\hat{\mathbf{b}}\right) = \max_{\mathbf{b}} l(\mathbf{b})$

or, equivalently, if and only if

- $U\left(\hat{\mathbf{b}}\right) = \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\hat{\mathbf{b}}} = \mathbf{0}_{p+1}$
  log-likelihood gradient evaluated at $\hat{\mathbf{b}}$
- $H\left(\hat{\mathbf{b}}\right) = \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} l(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\hat{\mathbf{b}}}$ negative definite
  log-likelihood hessian matrix evaluated at $\hat{\mathbf{b}}$

Maximum likelihood estimation (2)

The maximum likelihood estimate $\hat{\mathbf{b}}$ can be obtained by solving the following system of equations wrt $\mathbf{b}$

$$U(\mathbf{b}) = \mathbf{X}^\top [\mathbf{y} - \mathbf{m}] = \mathbf{0}_{p+1}$$

where $\mathbf{m} = \begin{bmatrix} \exp(b_0 + b_1 x_{11} + b_2 x_{21} + \ldots + b_p x_{p1}) \\ \exp(b_0 + b_1 x_{12} + b_2 x_{22} + \ldots + b_p x_{p2}) \\ \vdots \\ \exp(b_0 + b_1 x_{1n} + b_2 x_{2n} + \ldots + b_p x_{pn}) \end{bmatrix} = \begin{bmatrix} \exp(\mathbf{x}_1^\top \mathbf{b}) \\ \exp(\mathbf{x}_2^\top \mathbf{b}) \\ \vdots \\ \exp(\mathbf{x}_n^\top \mathbf{b}) \end{bmatrix}$

System of non-linear equations in $\mathbf{b}$:

$\Rightarrow$ In general, this system does not have an explicit solution (*it is not possible to obtain an analytical formula to compute* $\hat{\mathbf{b}}$)

The Newton-Raphson algorithm 15

General case

- $\boldsymbol{\theta}$ k-dimensional parameter vector
- Log-likelihood function

  $l(\boldsymbol{\theta})$ *scalar*
- Log-likelihood gradient

  $U(\boldsymbol{\theta}) = \dfrac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta})$ $k \times 1$ *vector*
- Log-likelihood Hessian matrix

  $H(\boldsymbol{\theta}) = \dfrac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} l(\boldsymbol{\theta})$ $k \times k$ *matrix*

Newton-Raphson algorithm

The Newton-Raphson algorithm finds a sequence of approximated solutions $\hat{\mathbf{t}}$ of a (non-linear) system

$U(\mathbf{t}) = \mathbf{0}_k$

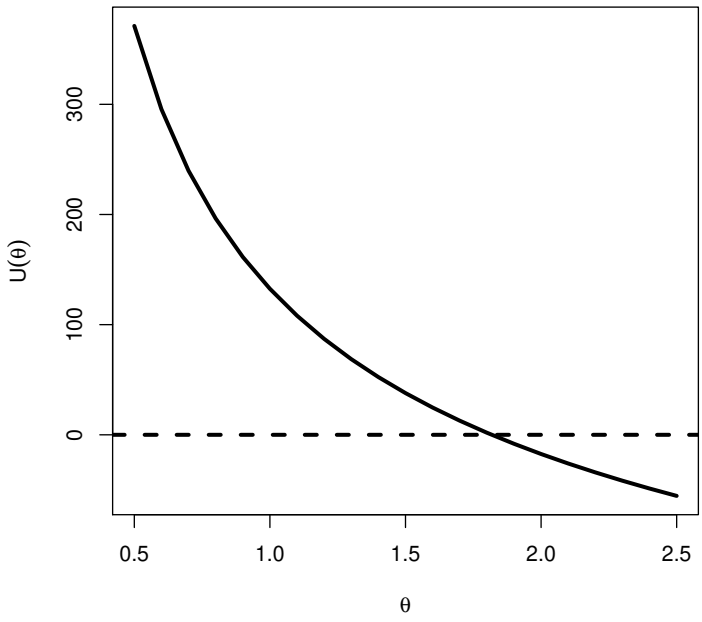
Given an initial solution $\mathbf{t}^{(1)}$ for $\hat{\mathbf{t}}$, this system is locally approximated with a linear system

$U(\mathbf{t}) \cong U\left(\mathbf{t}^{(1)}\right) + H\left(\mathbf{t}^{(1)}\right)\left(\mathbf{t} - \mathbf{t}^{(1)}\right) = \mathbf{0}_k$

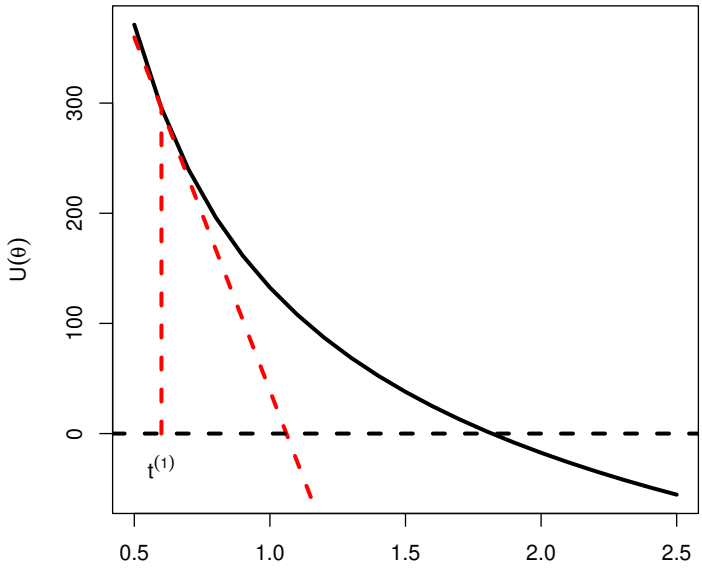
First-order Taylor series expansion

- $U\left(\mathbf{t}^{(1)}\right) = \frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\mathbf{t}^{(1)}}$
  log-likelihood gradient evaluated at $\mathbf{t}^{(1)}$
- $H\left(\mathbf{t}^{(1)}\right) = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} l(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\mathbf{t}^{(1)}}$
  log-likelihood Hessian matrix evaluated at $\mathbf{t}^{(1)}$

Example: $k = 1$



Example: $k = 1$



Example 1

A new approximation $\mathbf{t}^{(2)}$ to $\hat{\mathbf{t}}$ is obtained by solving the linear system

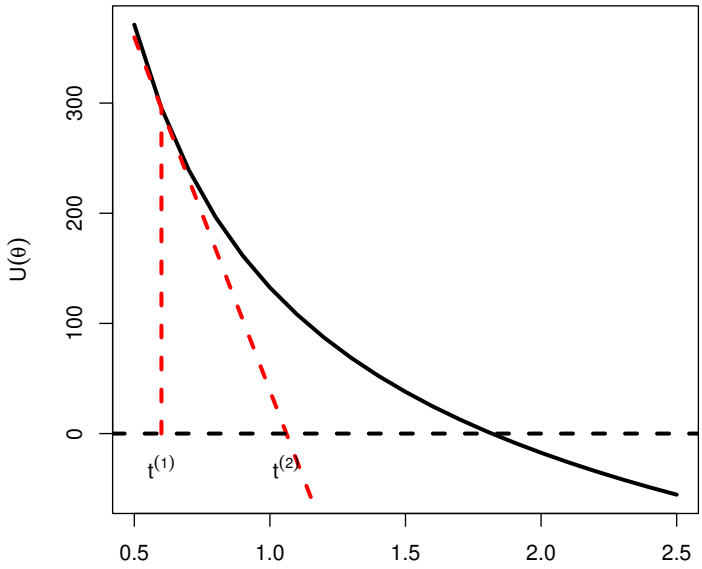
$$U\left(\mathbf{t}^{(1)}\right) + H\left(\mathbf{t}^{(1)}\right)\left(\mathbf{t} - \mathbf{t}^{(1)}\right) = \mathbf{0}_k$$

In particular:

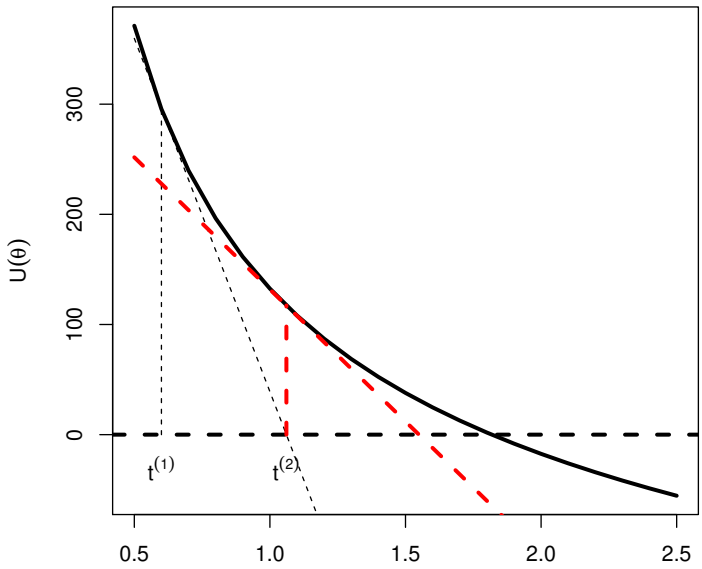
$$\begin{aligned} & U\left(\mathbf{t}^{(1)}\right) + H\left(\mathbf{t}^{(1)}\right)\left(\mathbf{t} - \mathbf{t}^{(1)}\right) = \mathbf{0}_k \\ & \Updownarrow \\ & H\left(\mathbf{t}^{(1)}\right)\mathbf{t} = H\left(\mathbf{t}^{(1)}\right)\mathbf{t}^{(1)} - U\left(\mathbf{t}^{(1)}\right) \\ & \Updownarrow \\ & \mathbf{t}^{(2)} = \mathbf{t}^{(1)} - H\left(\mathbf{t}^{(1)}\right)^{-1} U\left(\mathbf{t}^{(1)}\right) \end{aligned}$$

Warning: $H\left(\mathbf{t}^{(1)}\right)$ *must be invertible*

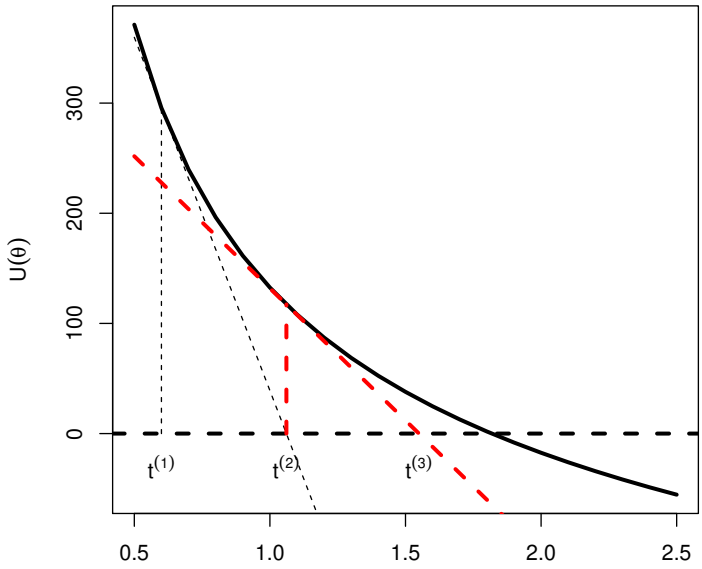
Example: $k = 1$



Example: $k = 1$



Example: $k = 1$



r-th step

The $(r+1)$-th approximation $\mathbf{t}^{(r+1)}$ to $\hat{\mathbf{t}}$ is obtained using the recursive formula

$$\mathbf{t}^{(r+1)} = \mathbf{t}^{(r)} - H\left(\mathbf{t}^{(r)}\right)^{-1} U\left(\mathbf{t}^{(r)}\right)$$

If some regularity conditions are met, it is possible to prove that

$$\lim_{r\to\infty} \mathbf{t}^{(r)} = \hat{\mathbf{t}}$$

The sequence of approximations converges to the solution of the non-linear system $U(\mathbf{t}) = \mathbf{0}_k$

Stopping criteria

The recursive formula

$\mathbf{t}^{(r+1)} = \mathbf{t}^{(r)} - H\left(\mathbf{t}^{(r)}\right)^{-1} U\left(\mathbf{t}^{(r)}\right)$ is repeatedly applied until a stopping criterion is met

- $\|\mathbf{t}^{(r+1)} - \mathbf{t}^{(r)}\|_2 < \epsilon$ with $\epsilon > 0$

  euclidean norm between two consecutive approximations
- $l\left(\mathbf{t}^{(r+1)}\right) - l\left(\mathbf{t}^{(r)}\right) < \epsilon$ with $\epsilon > 0$

  difference in the log-likelihood evaluated at two consecutive approximations

The final approximation to $\hat{\mathbf{t}}$ is given by the last element of the sequence

Fisher scoring algorithm

The log-likelihood Hessain matrix evaluated at $\mathbf{t}^{(r)}$ is replaced with its expected value in the recursive formula

$$E\left[H\left(\mathbf{t}^{(r)}\right)\right] = -I\left(\mathbf{t}^{(r)}\right)$$

where $I\left(\mathbf{t}^{(r)}\right)$ is the expected Fisher information evaluated at $\mathbf{t}^{(r)}$

The $(r+1)$-th approximation $\mathbf{t}^{(r+1)}$ to $\hat{\mathbf{t}}$ is obtained using the recursive formula

$$\mathbf{t}^{(r+1)} = \mathbf{t}^{(r)} + I\left(\mathbf{t}^{(r)}\right)^{-1} U\left(\mathbf{t}^{(r)}\right)$$

A comparison between the two algorithm

Generally speaking:

- The Newton-Raphson has a faster convergence rate
- The Fisher scoring is more stable
- it is possible to define mixed strategies: the Fisher scoring algorithm is used for a given number of steps, then the final steps are performed according to the Newton-Raphson algorithm

Poisson regression models (1)

Given the initial approximation $\mathbf{b}^{(1)}$ to $\hat{\mathbf{b}}$:

$$U\left(\mathbf{m}^{(1)}\right) = \mathbf{X}^\top \left[\mathbf{y} - \mathbf{m}^{(1)}\right] \text{ e } H\left(\mathbf{m}^{(1)}\right) = -\mathbf{X}^\top \mathbf{W}^{(1)} \mathbf{X}$$

where

$$\mathbf{m}^{(1)} = \begin{bmatrix} \exp(\mathbf{x}_1^\top \mathbf{b}^{(1)}) \\ \exp(\mathbf{x}_2^\top \mathbf{b}^{(1)}) \\ \vdots \\ \exp(\mathbf{x}_n^\top \mathbf{b}^{(1)}) \end{bmatrix} \quad \mathbf{W}^{(1)} = \begin{bmatrix} \exp(\mathbf{x}_1^\top \mathbf{b}^{(1)}) & 0 & 0 & \dots & 0 \\ 0 & \exp(\mathbf{x}_2^\top \mathbf{b}^{(1)}) & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \exp(\mathbf{x}_n^\top \mathbf{b}^{(1)}) \end{bmatrix}$$

Poisson regression models (2)

The system of non-linear equations in $\mathbf{b}$

$$\mathbf{X}^\top [\mathbf{y} - \mathbf{m}] = \mathbf{0}_{p+1}$$

is locally approximated with the system of linear equations in $\mathbf{b}$

$$\mathbf{X}^\top [\mathbf{y} - \mathbf{m}] \cong \mathbf{X}^\top \left[\mathbf{y} - \mathbf{m}^{(1)}\right] - \mathbf{X}^\top \mathbf{W}^{(1)} \mathbf{X} \left[\mathbf{b} - \mathbf{b}^{(1)}\right] = \mathbf{0}_{p+1}$$

Due to the use of the canonical link function, the Newton-Raphson coincides with the Fisher scoring algorithm

A new approximation $\mathbf{b}^{(2)}$ to $\hat{\mathbf{b}}$ is obtained solving the latter system wrt $\mathbf{b}$

Step 1

$$\mathbf{X}^\top \left[\mathbf{y} - \mathbf{m}^{(1)}\right] - \mathbf{X}^\top \mathbf{W}^{(1)} \mathbf{X} \left[\mathbf{b} - \mathbf{b}^{(1)}\right] = \mathbf{0}_{p+1}$$

$$\Updownarrow$$

$$\mathbf{X}^\top \left[\mathbf{y} - \mathbf{m}^{(1)}\right] - \mathbf{X}^\top \mathbf{W}^{(1)} \mathbf{X} \mathbf{b} + \mathbf{X}^\top \mathbf{W}^{(1)} \mathbf{X} \mathbf{b}^{(1)} = \mathbf{0}_{p+1}$$

$$\Updownarrow$$

$$\mathbf{X}^\top \mathbf{W}^{(1)} \mathbf{X} \mathbf{b} = \mathbf{X}^\top \mathbf{W}^{(1)} \mathbf{X} \mathbf{b}^{(1)} + \mathbf{X}^\top \left[\mathbf{y} - \mathbf{m}^{(1)}\right]$$

$$\Updownarrow$$

$$\begin{array}{rcl} \mathbf{b}^{(2)} & = & \mathbf{b}^{(1)} \underbrace{+ \left(\mathbf{X}^\top \mathbf{W}^{(1)} \mathbf{X}\right)^{-1}}_{-H\left(\mathbf{t}^{(1)}\right)^{-1}} \underbrace{\mathbf{X}^\top \left[\mathbf{y} - \mathbf{m}^{(1)}\right]}_{U\left(\mathbf{t}^{(1)}\right)} \\ \mathbf{t}^{(2)} & = & \mathbf{t}^{(1)} \end{array}$$

Note that:
$\mathbf{X}^\top \mathbf{W}^{(1)} \mathbf{X}$ *is invertible if and only if:*

- $\mathbf{X}$ *has full column rank*
- $\mathbf{W}^{(1)}$ *has strictly positive elements on its main diagonal*

Step r

The $(r+1)$-th approximation $\mathbf{b}^{(r+1)}$ to $\hat{\mathbf{b}}$ is obtained using the recursive formula

$$\begin{array}{rcl} \mathbf{b}^{(r+1)} & = & \mathbf{b}^{(r)} \underbrace{+\left(\mathbf{X}^\top \mathbf{W}^{(r)} \mathbf{X}\right)^{-1}}_{-H\left(\mathbf{t}^{(r)}\right)^{-1}} \underbrace{\mathbf{X}^\top \left[\mathbf{y} - \mathbf{m}^{(r)}\right]}_{U\left(\mathbf{t}^{(r)}\right)} \\ \mathbf{t}^{(r+1)} & = & \mathbf{t}^{(r)} \end{array}$$

Stopping criteria

- $\|\mathbf{b}^{(r+1)} - \mathbf{b}^{(r)}\|_2 < \epsilon$ with $\epsilon > 0$

  euclidean norm between two consecutive approximations

- $l\left(\mathbf{b}^{(r+1)}\right) - l\left(\mathbf{b}^{(r)}\right) < \epsilon$ with $\epsilon > 0$

  difference in the log-likelihood evaluated at two consecutive approximations

Iterative reweighted least squares (1)

The system

$$\mathbf{X}^\top \left[\mathbf{y} - \mathbf{m}^{(r)}\right] - \mathbf{X}^\top \mathbf{W}^{(r)} \mathbf{X} \left[\mathbf{b} - \mathbf{b}^{(r)}\right] = \mathbf{0}_{p+1}$$

can be re-expressed as

$$\mathbf{X}^\top \mathbf{W}^{(r)} \mathbf{X}\mathbf{b} = \mathbf{X}^\top \mathbf{W}^{(r)} \mathbf{X}\mathbf{b}^{(r)} + \mathbf{X}^\top \left[\mathbf{y} - \mathbf{m}^{(r)}\right]$$

$$\Updownarrow$$

$$\begin{array}{rcl} \mathbf{X}^\top \mathbf{W}^{(r)} \mathbf{X}\mathbf{b} & = & \mathbf{X}^\top \mathbf{W}^{(r)} \left\{\mathbf{X}\mathbf{b}^{(r)} + \left[\mathbf{W}^{(r)}\right]^{-1} \left[\mathbf{y} - \mathbf{m}^{(r)}\right]\right\} \\ & = & \mathbf{X}^\top \mathbf{W}^{(r)} \mathbf{z}^{(r)} \end{array}$$

where $\mathbf{z}^{(r)} = \mathbf{X}\mathbf{b}^{(r)} + \left[\mathbf{W}^{(r)}\right]^{-1} \left[\mathbf{y} - \mathbf{m}^{(r)}\right]$

Iterative reweighted least squares (2)

The $(r+1)$-th $\mathbf{b}^{(r+1)}$ to $\hat{\mathbf{b}}$ can be re-expressed as

$$\begin{aligned}\mathbf{b}^{(r+1)} &= \mathbf{b}^{(r)} + \left(\mathbf{X}^\top \mathbf{W}^{(r)} \mathbf{X}\right)^{-1} \left[\mathbf{y} - \mathbf{m}^{(r)}\right] \\ &= \left(\mathbf{X}^\top \mathbf{W}^{(r)} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W}^{(r)} \mathbf{z}^{(r)}\end{aligned}$$

weighted least squares estimate for the regression coefficients of a linear regression model that links the dependent variable $Z^{(r)}$ to the regressors $X_1, \ldots, X_p$, with individual weights equal to $m_i^{(r)}$

Adjusted dependent variable - pseudo-dependent variable

the i-th element of $\mathbf{z}^{(r)}$ is equal to

$$z_i^{(r)} = \mathbf{x}_i^\top \mathbf{b}^{(r)} + \frac{1}{m_i^{(r)}} \left[y_i - m_i^{(r)}\right]$$

is the value of the adjusted dependent variable - pseudo-dependent variable at the r-th step of the Newton-Raphson algorithm

$z_i^{(r)}$ *can be interpreted as an approximation to* $\ln(y_i)$ *(value of the link function applied to the observed value of the dependent variable), obtained using a first order Taylor series expansion at* $m_i^{(r)}$

$$\ln(y_i) \cong \ln\left(m_i^{(r)}\right) + \left.\frac{\partial \ln(y_i)}{\partial y_i}\right|_{y_i = m_i^{(r)}} \left[y_i - m_i^{(r)}\right]$$

Initialisation

The Newton-Raphson can be initialised by setting

- $\mathbf{W}^{(0)} = \mathbf{I}_n$ identity matrix
- $z_i^{(0)} = \ln(y_i + 0.5) \quad (0.5$ *is added in order to avoid* $\ln(0))$
- $\mathbf{b}^{(1)} = \left(\mathbf{X}^\top \mathbf{W}^{(0)} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W}^{(0)} \mathbf{z}^{(0)} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{z}^{(0)}$

Maximum likelihood estimator (1)

In absence of an analytical formula for $\hat{\mathbf{B}}$, the properties of the maximum likelihood estimator must be investigated starting from the (asymptotic) properties of the score function $U(\boldsymbol{\beta})$

Under regularity conditions (*that hold for Poisson regression models*)

$$U(\boldsymbol{\beta}) = \mathbf{X}^\top [\mathbf{y} - \boldsymbol{\mu}] \xrightarrow{d} MVN_{p+1}\left(\mathbf{0}_{p+1}, \mathbf{X}^\top \mathbf{W} \mathbf{X}\right)$$

Exploiting the following Taylor series expansion

$$\mathbf{0}_{p+1} = U\left(\hat{\mathbf{B}}\right) \cong U(\boldsymbol{\beta}) - I(\boldsymbol{\beta}) \left[\hat{\mathbf{B}} - \boldsymbol{\beta}\right]$$

it is possible to prove that

$$U(\boldsymbol{\beta}) \cong I(\boldsymbol{\beta}) \left[\hat{\mathbf{B}} - \boldsymbol{\beta}\right]$$

$U(\boldsymbol{\beta})$ *is approximately equivlaent to a linear transformation of* $\hat{\mathbf{B}}$

this implies that

$$\mathbf{X}^\top \mathbf{W} \mathbf{X} \left[\hat{\mathbf{B}} - \boldsymbol{\beta}\right] \xrightarrow{d} MVN_{p+1}\left(\mathbf{0}_{p+1}, \mathbf{X}^\top \mathbf{W} \mathbf{X}\right)$$

Maximum likelihood estimator (2)

Thus

- $\hat{\mathbf{B}} \xrightarrow{d} MVN_{p+1}\left(\boldsymbol{\beta}, I\left(\boldsymbol{\beta}\right)^{-1}\right)$

  The maximum likelihood estimator is asymptotically unbiased, efficient and has a multivariate Gaussian distribution

- $\left(\mathbf{X}^\top \mathbf{W}\mathbf{X}\right)^{1/2}\left[\hat{\mathbf{B}} - \boldsymbol{\beta}\right] \xrightarrow{d} MVN_{p+1}\left(\mathbf{0}_{p+1}, I_{p+1}\right)$

  Vector of asymptotical pivotal for $\boldsymbol{\beta}$ with a multivariate Gaussian distribution

- $\left[\hat{\mathbf{B}} - \boldsymbol{\beta}\right]^\top \left(\mathbf{X}^\top \mathbf{W}\mathbf{X}\right)\left[\hat{\mathbf{B}} - \boldsymbol{\beta}\right] \xrightarrow{d} \chi^2_{(p+1)}$

Estimation of the asymptotic variance

$\mathbf{X}^\top \mathbf{W}\mathbf{X}$ is uknown (it depends on $\boldsymbol{\beta}$)

$\Rightarrow$ it can be estimated using:

$$\mathbf{X}^\top \hat{\mathbf{W}}\mathbf{X}$$

where

$$\hat{\mathbf{W}} = \begin{bmatrix} \exp(\mathbf{x}_1^\top \hat{\mathbf{b}}) & 0 & 0 & \dots & 0 \\ 0 & \exp(\mathbf{x}_2^\top \hat{\mathbf{b}}) & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \exp(\mathbf{x}_n^\top \hat{\mathbf{b}}) \end{bmatrix}$$

Poisson regression models: Deviance, residuals and model selection criteria

Saturated model

$$M: \begin{cases} Y_i|\mathbf{x}_i \sim \text{Poi}\left(\exp(\eta_i)\right) & \\ & \text{independent} \quad i = 1, \ldots, n \\ \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} & \end{cases}$$

$\Rightarrow$ The saturated model for M is a model with a number of parameters for the expected values that is equal to the number of unique covariate patterns in the matrix $\mathbf{X}$ (equal to the number of unique values for η_i)

$\Rightarrow$ if the number of unique covariate patterns is equal to n (*each sample unit is characterised by a specific combination of regressor values*), the saturated model can be defined as follows:

$M_{sat}: Y_i|\mathbf{x}_i \sim \text{Poi}\left(\mu_i\right) \quad \text{indipendenti} \quad i = 1, \ldots, n$

Maximum likelihood estimation of $\mu_1, \mu_2, \ldots, \mu_n$

- Log-likelihood function of the saturated model:

  $l\left(\mu_1, \ldots, \mu_n\right) = \sum_{i=1}^n (y_i \ln \mu_i - \mu_i) - \sum_{i=1}^n \ln y_i!$

- Maximum likelihood estimate of μ_i:

$$\left.\begin{array}{l} \dfrac{\partial}{\partial \mu_i} l\left(\mu_1, \ldots, \mu_n\right) = y_i \dfrac{1}{\mu_i} - 1 = \dfrac{y_i - \mu_i}{\mu_i} \\ \dfrac{\partial^2}{\partial \mu_i^2} l\left(\mu_1, \ldots, \mu_n\right) = -\dfrac{y_i}{\mu_i^2} \end{array}\right\} \quad \Rightarrow \quad \hat{m}_i = y_i \quad i = 1, \ldots, n$$

$\Rightarrow \quad l\left(y_1, \ldots, y_n\right) = \sum_{i=1}^n (\underbrace{y_i \ln y_i}_{0 \ln 0 \equiv 0} - y_i) - \sum_{i=1}^n \ln y_i!$

Maximum possible value for the log-likelihood associated with Poisson models for $\boldsymbol{Y}|\mathbf{X}$, given the observed sample $\mathbf{y}$

$\Rightarrow$ *Any Poisson regression model for $\boldsymbol{Y}|\mathbf{X}$ shows a maximum value for the log-likelihood that is smaller than that value, given the observed sample $\mathbf{y}$*

(Residual) deviance for a Poisson regression model

$$
\begin{aligned}
D &= 2\ln\left[\frac{L\left(y_1,\ldots,y_n\right)}{L\left(\hat{\mathbf{b}}\right)}\right] = 2\left[l\left(y_1,\ldots,y_n\right) - l\left(\hat{\mathbf{b}}\right)\right] \\
&= 2\left\{\sum_{i=1}^{n}\left[y_i\ln y_i - y_i\right] - \sum_{i=1}^{n}\ln y_i! + \right. \\
&\quad \left. - \sum_{i=1}^{n}\left[y_i\ln\exp(\mathbf{x}_i^\top\hat{\mathbf{b}}) - \exp(\mathbf{x}_i^\top\hat{\mathbf{b}})\right] + \sum_{i=1}^{n}\ln y_i!\right\} \\
&= 2\sum_{i=1}^{n}\left[y_i\ln\frac{y_i}{\hat{m}_i} - (y_i - \hat{m}_i)\right]
\end{aligned}
$$

where $\hat{m}_i = \exp(\mathbf{x}_i^\top\hat{\mathbf{b}})$

$\Rightarrow$ observed counts y_i and (estimated) expected counts $\hat{m}_i$ are compared using a metric that exploits not only the difference but also the ratio

An approximation to D

Second order Taylor series expansion of $y_i\ln\frac{y_i}{\hat{m}_i}$ at $\hat{m}_i$:

$$
\begin{aligned}
y_i\ln\frac{y_i}{\hat{m}_i} &\cong \hat{m}_i\ln\frac{\hat{m}_i}{\hat{m}_i} + \\
&\quad + \frac{\partial}{\partial y_i}y_i\ln\frac{y_i}{\hat{m}_i}\bigg|_{y_i=\hat{m}_i}(y_i - \hat{m}_i) + \\
&\quad + \frac{1}{2}\,\frac{\partial^2}{\partial y_i^2}y_i\ln\frac{y_i}{\hat{m}_i}\bigg|_{y_i=\hat{m}_i}(y_i - \hat{m}_i)^2
\end{aligned}
$$

First and second order derivatives

$$\begin{aligned}
\frac{\partial}{\partial y_i} y_i \ln \frac{y_i}{\hat{m}_i} &= \frac{\partial}{\partial y_i} y_i \ln y_i - \frac{\partial}{\partial y_i} y_i \ln \hat{m}_i \\
&= y_i \cdot \frac{1}{y_i} + 1 \cdot \ln y_i - \ln \hat{m}_i = 1 + \ln y_i - \ln \hat{m}_i \\
\left.\frac{\partial}{\partial y_i} y_i \ln \frac{y_i}{\hat{m}_i}\right|_{y_i = \hat{m}_i} &= 1 + \ln \hat{m}_i - \ln \hat{m}_i = 1 \\
\frac{\partial^2}{\partial y_i^2} y_i \ln \frac{y_i}{\hat{m}_i} &= \frac{\partial}{\partial y_i} \left[1 + \ln y_i - \ln \hat{m}_i\right] = \frac{1}{y_i} \\
\left.\frac{\partial^2}{\partial y_i^2} y_i \ln \frac{y_i}{\hat{m}_i}\right|_{y_i = \hat{m}_i} &= \frac{1}{\hat{m}_i}
\end{aligned}$$

Pearson X^2 statistics

$$y_i \ln \frac{y_i}{\hat{m}_i} \cong \underbrace{\hat{m}_i \ln \frac{\hat{m}_i}{\hat{m}_i}}_{0} + (y_i - \hat{m}_i) + \frac{1}{2} \frac{(y_i - \hat{m}_i)^2}{\hat{m}_i}$$

Plugging this approximation in D:

$$\begin{aligned}
D &= 2 \textstyle\sum_{i=1}^n \left[y_i \ln \frac{y_i}{\hat{m}_i} - (y_i - \hat{m}_i) \right] \\
&\cong 2 \textstyle\sum_{i=1}^n \left[(y_i - \hat{m}_i) + \frac{1}{2} \frac{(y_i - \hat{m}_i)^2}{\hat{m}_i} - (y_i - \hat{m}_i) \right] \\
&\cong \textstyle\sum_{i=1}^n \frac{(y_i - \hat{m}_i)^2}{\hat{m}_i} = X^2
\end{aligned}$$

Goodness of fit test (1)

If the systematic component of a Poisson regression model is adequate (*correctly specified*), or, equivalently, if the null hypothesis

$$\begin{cases} Y_i|\mathbf{x}_i \sim \text{Poi}\,(\mu_i) \quad \text{independent} \quad i = 1, \ldots, n \\ H_0 : \ln \mu_i = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} \quad (+ \ln w_i) \end{cases}$$

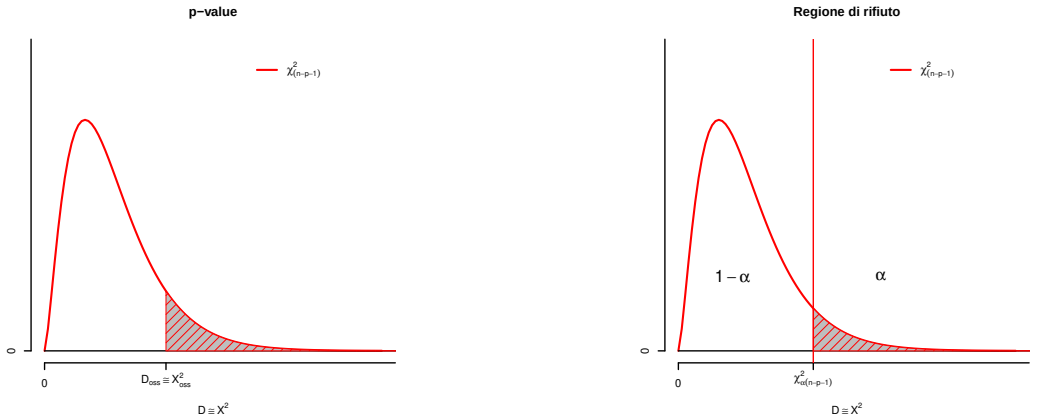
is true, then

- $X^2 \xrightarrow{d} D$ (*th two statistics are asymptotically equivalent*)
- D and X^2 are asymptotically independent from $\hat{\mathbf{B}}$
- $D|\, H_0 \xrightarrow{d} \chi^2_{n-p-1}$ & $X^2|\, H_0 \xrightarrow{d} \chi^2_{n-p-1}$

Note that:
differently from asymptotic results previously described, those related to D and X^2 hold for n fixed, and $w_i \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right) \longrightarrow \infty \;\; \forall i$

$\Rightarrow$ *This goodness of fit test should be exploited only when* $\hat{m}_i = w_i \exp\left(\mathbf{x}_i^\top \hat{\mathbf{b}}\right)$ *(or* w_i*) are "reasonably" large*

Goodness of fit test (2)



For a given significance level α

- $p-$value$< \alpha \Leftrightarrow D_{oss} \cong X^2_{oss} > \chi^2_{\alpha(n-p-1)} \Rightarrow H_0$ rejected
  $\Rightarrow$ *the model is not adequate*
- $p-$value$> \alpha \Leftrightarrow D_{oss} \cong X^2_{oss} < \chi^2_{\alpha(n-p-1)} \Rightarrow H_0$ not rejected
  $\Rightarrow$ *the model is adequate*

Residuals for Poisson regression models

Differently for the Gaussian case, there is not a straightforward definition of residual for Poisson regression models (and, in general, for GLMs)

$\Rightarrow$ *it is not possible to exploit an additive structure (systematic component + error term)*

Different definitions of residual are present in the statistical literature.
In particular, the most used residuals are:

- deviance residuals
- Pearson residuals

Other GLM residuals:

- *Ascombe residuals*
- *pseudo-residuals*

Deviance residuals

Squared root (with sign) of the generic term of the deviance:

$$e_i^D = \text{sign}\,(y_i - \hat{m}_i)\sqrt{2\left[y_i \ln \frac{y_i}{\hat{m}_i} - (y_i - \hat{m}_i)\right]} \quad i = 1, \ldots, n$$

where

$$\text{sign}\,(y_i - \hat{m}_i) = \begin{cases} -1 & \text{if} \quad y_i < \hat{m}_i \\ +1 & \text{if} \quad y_i \geq \hat{m}_i \end{cases}$$

$\Rightarrow \quad \sum_{i=1}^n \left(e_i^D\right)^2 = D$

Pearson residuals

Squared root (with sign) of the generic term of the Pearson X^2 statistic:

$e_i^P = \dfrac{y_i - \hat{m}_i}{\sqrt{\hat{m}_i}} \quad i = 1, \ldots, n$

$\Rightarrow \quad \sum_{i=1}^{n} \left(e_i^P\right)^2 = X^2$

Properties of the residuals

If a Poisson regression model is adequate (correctly specified):

- $\mathrm{E}\left[e_i^D \middle| \mathbf{x}_i\right] \cong \mathrm{E}\left[e_i^P \middle| \mathbf{x}_i\right] \cong 0$
- $\mathrm{Var}\left[e_i^D \middle| \mathbf{x}_i\right] \cong \mathrm{E}\left[e_i^P \middle| \mathbf{x}_i\right] \cong 1 - H_{ii}$

  H_{ii} is the l'i-th element of the main diagonal of the matrix

  $\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}\left(\mathbf{X}^\top\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{W}^{\frac{1}{2}}$
- Asymptotically, Pearson residuals and deviance residuals are equivalent and they are distributed as independent Gaussian random variables

  $\Rightarrow$ *as for D and X^2, the asymptotic properties of deviance and Pearson residuals hold for fixed n and $w_i \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right) \longrightarrow \infty \;\; \forall i$*

Graphical analysis of residuals

If a Poisson regression model is not adequate, according to a goodness of fit test, the graphical analysis of residuals can help in finding the misspecification source

In particular, misspecification errors in the systematic component can be detected by examining the plots

- e_i^P vs. $\mathbf{x}_i^\top \hat{\mathbf{b}}$
- e_i^P vs. $\hat{m}_i = w_i \exp\left(\mathbf{x}_i^\top \hat{\mathbf{b}}\right)$

Comparisons among Poisson regression models 15

Choice among Poisson regression models

Aim:
Which is the most adequate Poisson regression model for a given random sample $\boldsymbol{Y}$?

$Y_i|\mathbf{x}_i \sim \text{Poi}(\mu_i)$, independent $(i = 1, \ldots, n)$

Simplest situation: two candidate models

$M_A : \ln \mu_i = \eta_{Ai} = \mathbf{x}_{Ai}^\top \boldsymbol{\beta}_A \quad (+\ln w_i), \quad \boldsymbol{\beta}_A \ \ (p_A + 1) - \text{dimensional vector}$

$M_B : \ln \mu_i = \eta_{Bi} = \mathbf{x}_{Bi}^\top \boldsymbol{\beta}_B \quad (+\ln w_i), \quad \boldsymbol{\beta}_B \ \ (p_B + 1) - \text{dimensional vector}$

Note that:
the two models differ because they are characterised by different sets of regressors: $\mathbf{x}_{Ai} \neq \mathbf{x}_{Bi} \ \forall i$ *(without loss of generality:* $p_A > p_B$*)*

Situation 1: nested models - 1

Vectors $\mathbf{x}_{Bi}$ can be obtained by removing one or more that one regressor from vectors $\mathbf{x}_{Ai}$

$\Rightarrow$ M_B can be obtained by introducing suitable linear constraints on the parameteres of M_A

$$\left.\begin{array}{l} M_A : \ln \mu_i = \eta_{Ai} = \mathbf{x}_{Ai}^\top \boldsymbol{\beta}_A \quad (+\ln w_i) \\ H_0 : \mathbf{K}_B \boldsymbol{\beta}_A = \mathbf{t}_B \end{array}\right\} \quad \Rightarrow \quad M_B : \ln \mu_i = \eta_{Bi} = \mathbf{x}_{Bi}^\top \boldsymbol{\beta}_B \quad (+\ln w_i)$$

$q = p_A - p_B$ number of regressors excluded from M_A to obtain M_B

$\mathbf{K}_B$ $(q) \times (p_A + 1)$ matrix
each row of this matrix contains a 1 in a specific position (corresponding to one of the q regressors excluded from M_A), and 0 elsewhere

$\mathbf{t}_B = \mathbf{0}_q$

Situation 1: nested models - 2

A likelihood ratio test can be exploited to choose among M_A and M_B. In particular, such test can be expressed as a function of the two corresponding deviances:

$$\Delta l = 2 \ln \left[\frac{L\left(\hat{\mathbf{b}}_A\right)}{L\left(\hat{\mathbf{b}}_{A|H_0},\right)} \right] = 2 \ln \left[\frac{L\left(\hat{\mathbf{b}}_A\right)}{L\left(\hat{\mathbf{b}}_B\right)} \right] =$$

$$= D(M_B) - D(M_A) = \Delta D$$

If H_0 is true - if M_B is as "adequate" as M_A:

$\Rightarrow$ $\Delta D | \, M_B \xrightarrow{d} \chi^2_q$

Furthermore

$\Rightarrow$ ΔD is asymptotically equivalent to $\left[\mathbf{K}_B \hat{\mathbf{b}}_A - \mathbf{t}_B\right]^\top \left[\mathbf{K}_B \widehat{I(\boldsymbol{\beta}_A)}^{-1} \mathbf{K}_B^\top\right]^{-1} \left[\mathbf{K}_B \hat{\mathbf{b}}_A - \mathbf{t}_B\right]$

$\Rightarrow$ $\left[\mathbf{K}_B \hat{\mathbf{b}}_A - \mathbf{t}_B\right]^\top \left[\mathbf{K}_B \widehat{I(\boldsymbol{\beta}_A)}^{-1} \mathbf{K}_B^\top\right]^{-1} \left[\mathbf{K}_B \hat{\mathbf{b}}_A - \mathbf{t}_B\right] \Big| \, M_B \xrightarrow{d} \chi^2_q$

Situation 2: non-nested models

Vectors $\mathbf{x}_{Bi}$ cannot be obtained by removing one or more that one regressor from vectors $\mathbf{x}_{Ai}$

$\Rightarrow$ Model M_B can be obtained by simultaneously excluding some (or all) regressors in model M_A and adding some regressors to model M_A

$\Rightarrow$ *The two models are characterised by two sets of regressors that are only partially overlapping, or non-overlapping*

$\Rightarrow$ The differences between the two deviances does not have a known random distribution, and thus a likelihood ratio test cannot be used to choose between the two models

AIC and BIC for Poisson regression models

$\Rightarrow$ Akaike information criterion

$$\begin{aligned} AIC &= -2\ln L\left(\hat{\mathbf{b}}\right) + 2(p+1) \\ &= -2\sum_{i=1}^{n}\left[y_i(\mathbf{x}_i^\top\hat{\mathbf{b}}) - \exp(\mathbf{x}_i^\top\hat{\mathbf{b}})\right] + 2\sum_{i=1}^{n}\ln y_i! + 2(p+1) \end{aligned}$$

$\Rightarrow$ (Schwartz) Bayesian criterion

$$\begin{aligned} BIC &= -2\ln L\left(\hat{\mathbf{b}}\right) + \ln(n)\cdot(p+1) \\ &= -2\sum_{i=1}^{n}\left[y_i(\mathbf{x}_i^\top\hat{\mathbf{b}}) - \exp(\mathbf{x}_i^\top\hat{\mathbf{b}})\right] + 2\sum_{i=1}^{n}\ln y_i! + \ln(n)\cdot(p+1) \end{aligned}$$

The additive constant $2\sum_{i=1}^{n}\ln y_i!$ *can be ignored when all competing models have a Poisson probabilistic component*

Poisson regression models: an example

Introduction

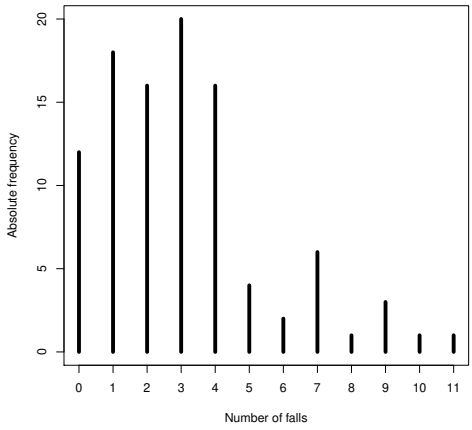
A researcher in geriatrics designed a prospective study to investigate the effects of an aerobic exercise training on the frequency of falls.

One hundred subjects (at least 65 years old and in reasonably good health) were randomly splitted into two group, and only subjects in the second groups were enrolled in the training program.

For each subject, the gender and the values of a balance index and of a strenght index were also registered (The higher the balance index, the more stable the subject, and the higher the strength index, the stronger subject).

Each subject kept a diary recording the number of falls during the six months of the study.

Number of falls - observed values



Dummy variable coding schemes

- train:

	trainYES
NO	0
YES	1

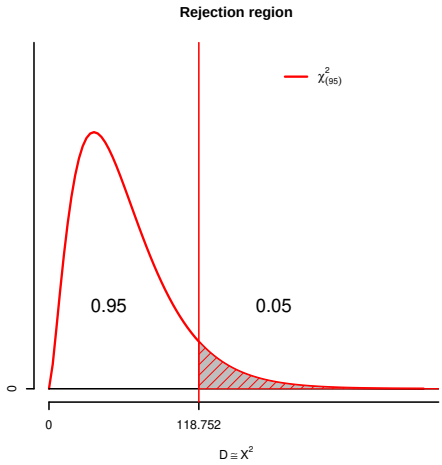
- gender:

	genderMALE
FEMALE	0
MALE	1

falls vs. train, gender, balance & strength - R summary output

```
Coefficients:
              Estimate   Std.  Error   z value   Pr(>|z|)
 (Intercept)     0.489         0.337     1.453      0.146
 trainYES       -1.069         0.133    -8.031      0.000
 genderMALE     -0.047         0.120    -0.388      0.698
 bal             0.009         0.003     3.207      0.001
 str             0.009         0.004     1.986      0.047

    Null deviance:  199.19 on 99 degrees of freedom
Residual deviance:  108.79 on 95 degrees of freedom
AIC: 377.29
```

falls vs. train, gender, balance & strength - goodness of fit

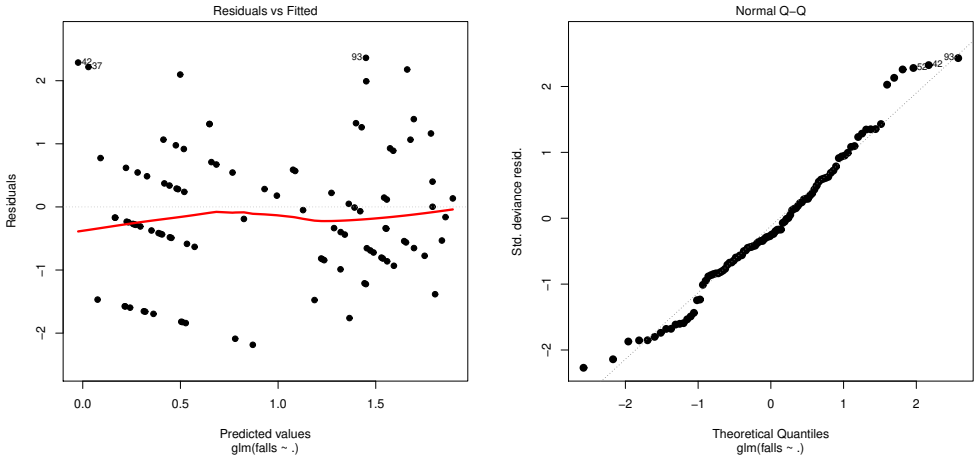


$D = 108.790,\ \Pr\left(\chi^2_{(95)} \geq 108.79\right) = 0.158$

$X^2 = 105.547,\ \Pr\left(\chi^2_{(95)} \geq 105.547\right) = 0.216$

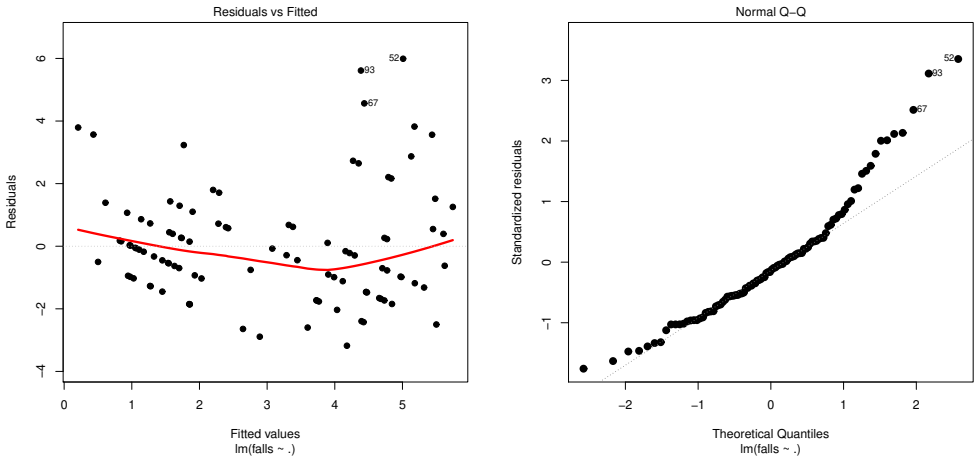
The model can be considered adequate

falls vs. train, gender, balance & strength - residuals



No evident patterns emerge from the plots, consistently with the result of the goodness of fit test

Residuals of a Gaussian linear model



These plots highlight the inadequacy of the Gaussianity and homoschedasticity assumption

Interpretation of the model parameters - 1

$\mathrm{E}[Y_i|\mathbf{x}_i] = \exp(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi})$

Each regressor has a non-linear effect on $\mathrm{E}[Y_i|\mathbf{x}_i]$:

$$\frac{\partial}{\partial x_{ji}} \mathrm{E}[Y_i|\mathbf{x}_i] = \beta_j \exp(\eta_i)$$

$\Rightarrow$ the direction of the change in $\mathrm{E}[Y_i|\mathbf{x}_i]$ due to a unit increase in x_{ji} depends on the sign of β_j

$\Rightarrow$ the magnitude of the change depends also on the values of all the regressors

Interpretation of the model parameters - 2

Multiplicative models for $\mathrm{E}[Y_i|\mathbf{x}_i]$:

- Conditional expected value *before* a unit increase in x_{ji}

  $\mathrm{E}[Y_i|\mathbf{x}_i] = \exp(\beta_0) \cdot \ldots \cdot \exp(\beta_j x_{ji}) \cdot \ldots \cdot \exp(\beta_p x_{pi})$

- Conditional expected value *after* a unit increase in x_{ji}

  $\mathrm{E}[Y_i|\mathbf{x}_i+] = \exp(\beta_0) \cdot \ldots \cdot \exp[\beta_j (x_{ji}+1)] \cdot \ldots \cdot \exp(\beta_p x_{pi})$

$\Rightarrow$ Multiplicative change due to a unit increase in x_{ji}

$$\frac{\mathrm{E}[Y_i|\mathbf{x}_i+]}{\mathrm{E}[Y_i|\mathbf{x}_i]} = \exp(\beta_j) = \begin{cases} < 1 & \text{if } \beta_j < 0 \\ = 1 & \text{if } \beta_j = 0 \\ > 1 & \text{if } \beta_j > 0 \end{cases}$$

$\Rightarrow$ Percentage change due to a unit increase in x_{ji}

$[\exp(\beta_j) - 1]\,\%$

falls vs. train, gender, balance & strength - Estimated effects

	$\hat{b}_j$	$\exp\left(\hat{b}_j\right)$
trainYES	-1.069	0.343
genderMALE	-0.047	0.954
bal	0.009	1.010
str	0.009	1.009

- *the expected number of falls for individuals enrolled in the training program is approximately one third of the expected number of falls for untrained individuals, for given gender, balance index and strength index*
- *the expected number of falls for males is approximately 5 percent lower than the expected number of falls for female, after controlling for all the other regressors in the model*
- *the expected number of falls increases of about 1 percent for each additional point in the balance index, holding fixed the values for the other regressors in the model*
- *the expected number of falls increases of about 1 percent for each additional point in the strength index, holding fixed the values for the other regressors in the model*

Hypothesis testing

$H_0 : \beta_{\texttt{train}} = \beta_{\texttt{gender}} = \beta_{\texttt{balance}} = \beta_{\texttt{strength}} = 0$

- Full model:

```
Coefficients:
               Estimate   Std.  Error   z value   Pr(>|z|)
 (Intercept)      0.489         0.337     1.453      0.146
 trainYES        -1.069         0.133    -8.031      0.000
 genderMALE      -0.047         0.120    -0.388      0.698
 bal              0.009         0.003     3.207      0.001
 str              0.009         0.004     1.986      0.047

    Null deviance:  199.19 on 99 degrees of freedom
Residual deviance:  108.79 on 95 degrees of freedom
AIC: 377.29
```

- Reduced model:

```
Coefficients:
               Estimate   Std.  Error   z value   Pr(>|z|)
 (Intercept)      1.112         0.057    19.386      0.000

    Null deviance:  199.19 on 99 degrees of freedom
Residual deviance:  199.19 on 99 degrees of freedom
AIC: 459.69
```


Likelihood ratio test

```
Model                        Resid.  Df   Resid.  Dev   Df   Deviance   Pr(>Chi)
falls~1                      99        199.19
falls~train+gender+bal+str   95        108.79    4      90.40     0.0000
```

$$2 \ln \frac{L(F)}{L(R)} = -2 \ln \left[L(R) - L(F) \right] = 199.19 - 108.79 = 90.40$$

At least one of the three regressors is significantly associated with the number of falls

Wald test

```
Hypothesis:
trainYES = 0
genderMALE = 0
balance = 0
strength = 0

Model 1:  restricted model
Model 2:  falls ~train + gender + balance + strength

     Res.Df      Df    Chisq   Pr(>Chisq)
 1   99.000
 2   95.000   4.000   80.250        0.000
```

The Wald test leads to the same conclusion

$H_0 : \beta_{\texttt{train}} = 0$

- Full model:

```
Coefficients:
               Estimate   Std.  Error   z value   Pr(>|z|)
 (Intercept)      0.489         0.337     1.453      0.146
 trainYES        -1.069         0.133    -8.031      0.000
 genderMALE      -0.047         0.120    -0.388      0.698
 bal              0.009         0.003     3.207      0.001
 str              0.009         0.004     1.986      0.047

    Null deviance:  199.19 on 99 degrees of freedom
Residual deviance:  108.79 on 95 degrees of freedom
AIC: 377.29
```

- Reduced model:

```
Coefficients:
               Estimate   Std.  Error   z value   Pr(>|z|)
 (Intercept)      0.366         0.322     1.138      0.255
 genderMALE      -0.228         0.117    -1.944      0.052
 bal              0.009         0.003     2.979      0.003
 str              0.006         0.004     1.466      0.143

    Null deviance:  199.19 on 99 degrees of freedom
Residual deviance:  182.31 on 96 degrees of freedom
AIC: 448.81
```

Likelihood ratio test

Model	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
falls~gender+bal+str	96	182.31			
falls~train+gender+bal+str	95	108.79	1	73.52	0.0000

$$2\ln\frac{L(F)}{L(R)} = -2\ln\left[L(R)-L(F)\right] = 182.31 - 108.79 = 73.52$$

$$\cong \left(\frac{\hat{\beta}^2_{\texttt{train}}}{s^2\left[\hat{\beta}_{\texttt{train}}\right]}\right) = \frac{-1.069^2}{0.133^2} = 64.502$$

The aerobic training significantly reduced the average number of falls, after controlling for all the other subject characteristics

Generalised linear models for binary outcomes: Introduction

Binary (dicothoumous) outcomes - Examples

Some examples:

- presence/absence of a given feature
- success/failure of a given trial
- positive/negative answer to a question
- ...

$\Rightarrow$ employed/unemployed woman

$\Rightarrow$ satisfied/unsatisfied costumer

$\Rightarrow$ presence/absence of side-effects after a given treatment

$\Rightarrow$ recovery from a disease (yes/no)

$\Rightarrow$ In favor/against a given law

$\Rightarrow$ ...

Dummy/indicator variables

Any dichotomous outcomes can be numerically coded using a dummy/indicator variable

$$Z = \begin{cases} 0 & \text{absence/failure/negative answer} \\ 1 & \text{presence/success/positive answer} \end{cases}$$

Z_j	r.v. describing the observed value for the dependent variable on the j-th sample unit ($j = 1, \ldots, N$)
$\mathbf{x}_j = (x_{0j}, x_{1j} \ldots, x_{pj})^\top$	$p+1$-dimensional vector containing the regressor values observed on the j-th sample unit ($x_{0j} = 1\ \forall j$ constant regressor associated with the model intercept)

Note the change in notation - the "usual" notation Y_i, $i = 1, \ldots, n$ will appear again later, with a slightly different meaning...

Bernoulli distributions and exponential families

$Z_j \sim \text{Ber}(\pi_j) \quad z_j \in \{0,1\}, \quad \pi_j \in [0,1]$

$$f(z_j, \pi_j) = \pi_j^{z_j}(1-\pi_j)^{1-z_j} = \exp\left\{\frac{1}{1}\left[z_j \ln\frac{\pi_j}{1-\pi_j} + \ln(1-\pi_j)\right] + 0\right\}$$

$Z_j \sim \text{EF}(\text{logit}(\pi_j), \phi = 1, w_j = 1)$

- $b(\pi_j) = \ln\frac{\pi_j}{1-\pi_j} = \text{logit}(\pi_j)$ natural parameter,
- $c(\theta) = \ln(1-\pi_j)$
- $d(z_j, \phi, w_j) = 0$

Bernoulli distributions - some properties

Thanks to the properties of exponential families of distributions:

- $$\text{E}[Z_j] = -\frac{c'(\pi_j)}{b'(\pi_j)} = -\frac{-\frac{1}{1-\pi_j}}{\frac{1}{\pi_j} + \frac{1}{1-\pi_j}} = \frac{\pi_j(1-\pi_j)}{1-\pi_j} = \pi_j$$
- $$\text{Var}[Z_j] = \frac{\phi}{w_j}\frac{\text{E}'[Z_j]}{b'(\pi_j)} = \frac{1}{1}\frac{1}{\frac{1}{\pi_j} + \frac{1}{1-\pi_j}} = \frac{1}{\frac{1}{\pi_j(1-\pi_j)}} = \pi_j(1-\pi_j)$$

GLM for binary outcomes - basic definition

- Probabilistic component

  $Z_j|\mathbf{x}_j \sim \text{Ber}(\pi_j)$ independent $j = 1, \ldots, N$

- Systematic component

  $\text{E}[Z_j|\mathbf{x}_j] = \pi_j = h(\eta_j) = h\left(\mathbf{x}_j^\top \boldsymbol{\beta}\right)$

Since $\pi_j \in [0, 1]$, the link function $h(\cdot)$ should be chosen among all functions satisfying the following requirement:

$h(\cdot) : \mathbb{R} \longmapsto [0, 1]$

Log-likelihood function for Bernoulli GLMs

Irrespective of the choice for the link function, any GLM with a probabilistic component based on the Bernoulli distribution is characterised by the following log-likelihood function

$$l(\beta_0, \ldots, \beta_p) = \sum_{j=1}^{N} \left\{ z_j \ln \frac{h\left(\mathbf{x}_j^\top \boldsymbol{\beta}\right)}{1 - h\left(\mathbf{x}_j^\top \boldsymbol{\beta}\right)} + \ln\left[1 - h\left(\mathbf{x}_j^\top \boldsymbol{\beta}\right)\right] \right\}$$

Underlying data structure

- $\mathbf{z} = (z_1, \ldots, z_N)^\top$ 0/1 values (*1 value for each sample unit*)
- $\mathbf{X}$ $N \times (p + 1)$ matrix (*1 row for each sample unit*)

Bernoulli GLMs & covariate patterns

Suppose that some sample units are characterised by the same covariate pattern (*the matrix* $\mathbf{X}$ *contains some identical rows*)

- n number of unique covariate patterns in the sample (*number of unique rows in* $\mathbf{X}$)
- $\mathbf{x}_i = (x_{0i}, x_{1i} \ldots, x_{pi})^\top$ $i-$th covariate pattern ($i = 1, \ldots, n$)
  - n_i number of sample units showing the $i-$th covariate pattern

    For each of these units

    - $\mathbf{x}_j^\top \boldsymbol{\beta} = \beta_0 x_{0j} + \beta_1 x_{1j} + \ldots + \beta_p x_{pj} = \beta_0 x_{0i} + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} = \mathbf{x}_i^\top \boldsymbol{\beta}$
    - $\pi_j = h\left(\mathbf{x}_j^\top \boldsymbol{\beta}\right) = h\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right) = \pi_i$

    Units characterised by the same covariate pattern show the same value for the linear predictor and for the conditional expected value
  - number of sample units showing the $i-$th covariate pattern **and** a value of z_j equal to 1

    $$y_i = \sum_{j:\eta_j = \eta_i} z_j$$

Log-likelihood for Bernoulli GLMs & covariate patterns

In presence of repeated covariate patterns, the log-likelihood function of any GLM with a Bernoulli probabilistic component can be re-expressed as the sum of n elements, one for each unique covariate pattern, as follows

$$l(\beta_0, \ldots, \beta_p) = \sum_{i=1}^{n} \underbrace{\left\{ y_i \ln \frac{h\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)}{1 - h\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)} + n_i \ln \left[1 - h\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right] \right\}}_{\sum_{j:\eta_j = \eta_i} \left\{ z_j \ln \frac{h(\eta_i)}{1 - h(\eta_i)} + \ln[1 - h(\eta_i)] \right\}}$$

The "relevant" information about $\boldsymbol{\beta}$ *is provided by the number of presences/successes/positive answers associated with each covariate pattern rather than the individual precence/success/positive answer*

Binomial distributions

Apart from an additive constant depending only on the values y_i, $l(\beta_0, \ldots, \beta_p)$ coincides with the log-likelihood function associated with the following n binomial random variables:

$$Y_i|\mathbf{x}_j \sim \text{Bin}\left(n_i, \pi_i = h\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right) \quad \text{independent} \quad i = 1, \ldots, n$$

Recall that if $Z_1, \ldots, Z_{n_i}$ are i.i.d. Bernoulli random variables with parameter π_i, then

$$Y_i = \sum_{j=1}^{n_i} Z_j \sim \text{Bin}(n_i, \pi_i)$$

An "alternative" data structure

$\mathbf{y} = (y_1, \ldots, y_n)^\top$	number of presences/successes/positive answers (*1 value for each unique covariate pattern*)
$\mathbf{n} = (n_1, \ldots, n_n)^\top$	number of "trials" (*1 value for each unique covariate pattern*)
	or, equivalently
$(n_1 - y_1, \ldots, n_n - y_n)^\top$	number of absences/failures/negative answers (*1 value for each unique covariate pattern*)
$\mathbf{X}$	$n \times (p+1)$ matrix (*1 row for each each unique covariate pattern*)

Some comments about the two data structures

- The two data structures coincide if $n_i = 1 \ \forall i$ (if $n = N$)

  the 0/1 structure is a special case of the successes/trials structure

- The distiction between the two data structures is crucial for some asymptotic properties of GLMs with a Bernoulli probabilistic component. In particular two different asymptotic concepts can be exploited

  A) $N = \sum_i n_i \longrightarrow \infty$

  B) $n_i \longrightarrow \infty \ \forall i$ (*fixed cell asymptotics*)

  Note that B) $\Rightarrow$ A), but A) $\not\Rightarrow$ B)

Relative frequencies

$p_i = \dfrac{y_i}{n_i}$ relative frequency of presences/successes/positive answers out of n_i trials

(*1 value for each unique covariate pattern*)

The log-likelihood function of any GLM with a Bernoulli probabilistic component can also be re-expressed in terms of p_i

$$l(\beta_0, \ldots, \beta_p) = \sum_{i=1}^{n} n_i \left\{ \underbrace{\frac{y_i}{n_i}}_{p_i} \ln \frac{h\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)}{1 - h\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)} + \ln\left[1 - h\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right] \right\}$$

Recall that

$Y_i | \mathbf{x}_i \sim \text{Bin}\left(n_i, \pi_i = h\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right)$ independent $i = 1, \ldots, n$

implies

$$P_i = \left. \frac{Y_i}{n_i} \right| \mathbf{x}_i \sim \text{EF}\left(b(\pi_i) = \text{logit}\left[h\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right], \phi = 1, w_i = n_i\right) \text{ independent } i = 1, \ldots, n$$

Relative frequencies - some properties

Thanks to the properties of exponential families of distributions:

- $\mathrm{E}[P_i] = -\dfrac{c'(\pi_i)}{b'(\pi_i)} = -\dfrac{-\dfrac{1}{1-\pi_i}}{\dfrac{1}{\pi_i} + \dfrac{1}{1-\pi_i}} = \dfrac{\pi_i(1-\pi_i)}{1-\pi_i} = \pi_i$
- $\mathrm{Var}[P_i] = \dfrac{\phi}{w_i}\dfrac{\mathrm{E}'[Z_i]}{b'(\pi_i)} = \dfrac{1}{n_i}\dfrac{1}{\dfrac{1}{\pi_i} + \dfrac{1}{1-\pi_i}} = \dfrac{1}{\dfrac{n_i}{\pi_i(1-\pi_i)}} = \dfrac{\pi_i(1-\pi_i)}{n_i}$

Logistic regression models:
Definition and maximum likelihood estimation

Model definition - 1

$\mathbf{x}_i = (x_{0i}, x_{1i} \ldots, x_{pi})^\top$	$p+1$-dimensional vector containing the regressor values characterising the i-th covariate pattern - $x_{0i} = 1 \ \forall i$ $(i = 1, \ldots, n)$
$\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$	linear predictor associated with the $i-$th covariate pattern
n_i	number of sample units showing the $i-$th covariate pattern
Y_i	r. v. describing the observed number of presences/successes/positive answers on the n_i sample units showing the $i-$th covariate pattern

$$\Rightarrow \quad Y_i|\mathbf{x}_i \sim \text{Bin}\left(n_i, \pi_i = \frac{\exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)}{1 + \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)}\right) \quad \text{independent} \quad i = 1, \ldots, n$$

Model definition - 2

- Probabilistic component

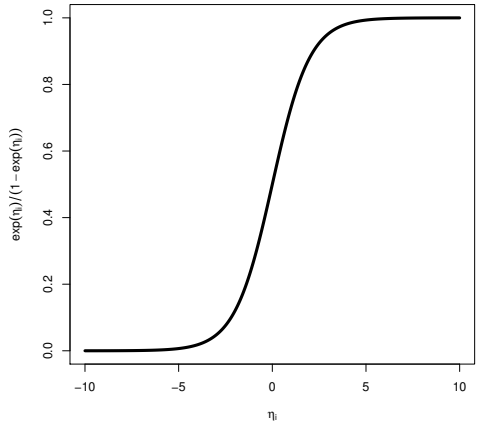
$$Y_i|\mathbf{x}_i \sim \text{Bin}\left(n_i, \pi_i\right) \iff P_i = \left.\frac{Y_i}{n_i}\right| \mathbf{x}_i \sim \text{EF}\left(b(\pi_i) = \text{logit}(\pi_i), \phi = 1, w_i = n_i\right)$$

  - $\text{E}\left[P_i|\mathbf{x}_i\right] = \pi_i \implies \text{E}\left[Y_i|\mathbf{x}_i\right] = n_i\pi_i$
  - $\text{Var}\left[P_i|\mathbf{x}_i\right] = \dfrac{\pi_i(1-\pi_i)}{n_i} \implies \text{Var}\left[Y_i|\mathbf{x}_i\right] = n_i\pi_i(1-\pi_i)$

- Systematic component

$$\pi_i = \frac{\exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)}{1 + \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)} \quad \text{Logistic function}$$

Logistic link function - graphical display



$$\frac{\exp(\cdot)}{1+\exp(\cdot)} : \mathbb{R} \longmapsto (0,1)$$

Canonical link function

$$\pi_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)} \iff \pi_i\left[1+\exp(\eta_i)\right] = \exp(\eta_i)$$

$$\iff \pi_i = \exp(\eta_i) - \pi_i \exp(\eta_i)$$

$$\iff \pi_i = \exp(\eta_i)(1-\pi_i)$$

$$\iff \frac{\pi_i}{1-\pi_i} = \exp(\eta_i)$$

$$\iff \text{logit}(\pi_i) = \eta_i$$

Log-likelihood

$$
\begin{aligned}
l(\boldsymbol{\beta}) &= \sum_{i=1}^{n} n_i \left\{ \frac{y_i}{n_i}\eta_i - \ln\left[1 - \frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right]\right\} + c \\
&= \left[\beta_0 \sum_{i=1}^{n} y_i + \ldots + \beta_p \sum_{i=1}^{n} y_i x_{pi}\right] - \sum_{i=1}^{n} n_i \ln\left[1 - \frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right] + c
\end{aligned}
$$

$\sum_{i=1}^{n} y_i, \ldots, \sum_{i=1}^{n} y_i x_{pi}$ are (minimal) sufficient statistics for $\boldsymbol{\beta}$

c is an additive constant that does not depend on $\boldsymbol{\beta}$

Score function

Considering the general formula for the generic element of the score function for a GLM with canonical link function:

$$
\begin{aligned}
U_j(\boldsymbol{\beta}) &= \frac{1}{\phi}\sum_{i=1}^{n} w_i \{p_i - \mathrm{E}[P_i|\mathbf{x}_i]\} x_{ji} \\
&= \frac{1}{1}\sum_{i=1}^{n} n_i \left[p_i - \frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right] x_{ji} \\
&= \sum_{i=1}^{n} \left[y_i - n_i \frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right] x_{ji} \\
&= \sum_{i=1}^{n} \left\{ y_i - \underbrace{\mathrm{E}[Y_i|\mathbf{x}_i]}_{\mu_i} \right\} x_{ji}
\end{aligned}
$$

Fisher information

Considering the general formula for the generic element of the (observed and expected) Fisher information matrix for a GLM with canonical link function:

$$
\begin{aligned}
i_{jl}(\boldsymbol{\beta}) &= I_{jl}(\boldsymbol{\beta}) \\
&= \sum_{i=1}^{n} \left(\frac{w_i}{\phi}\right)^2 \mathrm{Var}\left[P_i|\mathbf{x}_i\right] x_{ji} x_{li} \\
&= \sum_{i=1}^{n} \left(\frac{n_i}{1}\right)^2 \frac{\pi_i(1-\pi_i)}{n_i} x_{ji} x_{li} \\
&= \sum_{i=1}^{n} n_i \frac{\exp\left(\eta_i\right)}{\left[1+\exp\left(\eta_i\right)\right]^2} x_{ji} x_{li} \\
&= \sum_{i=1}^{n} \mathrm{Var}\left[Y_i|\mathbf{x}_i\right] x_{ji} x_{li}
\end{aligned}
$$

Hessian matrix

$\Rightarrow$ (j,l)-th element of the Hessian matrix of the log-likelihood function

$$
\begin{aligned}
H_{jl}(\boldsymbol{\beta}) &= \frac{\partial^2}{\partial\beta_j \partial\beta_l} l\left(\boldsymbol{\beta}\right) = -i_{jl}(\boldsymbol{\beta}) \\
&= -\sum_{i=1}^{n} n_i \frac{\exp\left(\eta_i\right)}{\left[1+\exp\left(\eta_i\right)\right]^2} x_{ji} x_{li}
\end{aligned}
$$

Matrix representation (1)

- $\mathbf{y} = (y_1, \ldots, y_n)^\top$ n-dimensional vector that contains the observed number of presences/successes/positive answers on the n unique covariate patterns
- $\mathbf{X}$ $n \times (p+1)$ matrix that contains the n unique covariate patterns
- $\boldsymbol{\mu} = \left(n_1 \frac{\exp(\eta_1)}{1+\exp(\eta_1)}, \ldots, n_n \frac{\exp(\eta_n)}{1+\exp(\eta_n)} \right)^\top$ n-dimensional vector that contains the conditional expected values of $Y_1, \ldots, Y_n$ for the n unique covariate patterns
- $\mathbf{W} = \begin{bmatrix} n_1 \frac{\exp(\eta_1)}{[1+\exp(\eta_1)]^2} & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & n_n \frac{\exp(\eta_n)}{[n+\exp(\eta_n)]^2} \end{bmatrix}$

  $n \times n$ diagonal matrix that contains the conditional variances of $Y_1, \ldots, Y_n$ for the n unique covariate patterns on the main diagonal

Matrix representation (2)

- Score function

  $U(\boldsymbol{\beta}) = \mathbf{X}^\top [\mathbf{y} - \boldsymbol{\mu}]$
- Hessian matrix of the log-likelihood function

  $H(\boldsymbol{\beta}) = -\mathbf{X}^\top \mathbf{W} \mathbf{X}$
- (observed and expected) Fisher information matrix

  $i(\boldsymbol{\beta}) = I(\boldsymbol{\beta}) = \mathbf{X}^\top \mathbf{W} \mathbf{X}$

For logistic regression models, the Newton-Raphson algorithm is equivalent to the Fisher scoring algorithm

Properties of the score function

- $\mathrm{E}\left[U\left(\boldsymbol{\beta}\right)\right] = \mathbf{0}_{p+1}$
- $\mathrm{Var}\left[U\left(\boldsymbol{\beta}\right)\right] = I(\boldsymbol{\beta})$
- $I(\boldsymbol{\beta})^{-\frac{1}{2}} U(\boldsymbol{\beta}) \xrightarrow{d} MVN_{p+1}\left(\mathbf{0}_{p+1}, \mathbf{I}_{p+1}\right)$

$\Rightarrow \quad U(\boldsymbol{\beta}) \approx MVN_{p+1}\left(\mathbf{0}_{k}, I(\boldsymbol{\beta})\right)$

The asymptotic properties of the score function hold both if

A) $N = \sum_i n_i \longrightarrow \infty$
and if

B) $n_i \longrightarrow \infty \ \forall i$

Maximum likelihood estimation 13

Maximum likelihood estimation (1)

$\hat{\mathbf{b}}$ is the maximum likelihood estimate of $\boldsymbol{\beta}$ if and only if

$$l\left(\hat{\mathbf{b}}\right) = \max_{\mathbf{b}} l\left(\mathbf{b}\right)$$

or, equivalently, if and only if

- $U\left(\hat{\mathbf{b}}\right) = \frac{\partial}{\partial \boldsymbol{\beta}} l\left(\boldsymbol{\beta}\right)|_{\boldsymbol{\beta}=\hat{\mathbf{b}}} = \mathbf{0}_{p+1}$
  log-likelihood gradient evaluated at $\hat{\mathbf{b}}$
- $H\left(\hat{\mathbf{b}}\right) = \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} l\left(\boldsymbol{\beta}\right)|_{\boldsymbol{\beta}=\hat{\mathbf{b}}}$ negative definite
  log-likelihood hessian matrix evaluated at $\hat{\mathbf{b}}$

Maximum likelihood estimation (2)

The maximum likelihood estimate $\hat{\mathbf{b}}$ can be obtained by solving the following system of equations wrt $\mathbf{b}$

$$U(\mathbf{b}) = \mathbf{X}^\top [\mathbf{y} - \mathbf{m}] = \mathbf{0}_{p+1}$$

where $\mathbf{m} = \begin{bmatrix} n_1 \dfrac{\exp(\mathbf{x}_1^\top \mathbf{b})}{1+\exp(\mathbf{x}_1^\top \mathbf{b})} \\ \vdots \\ n_n \dfrac{\exp(\mathbf{x}_n^\top \mathbf{b})}{1+\exp(\mathbf{x}_n^\top \mathbf{b})} \end{bmatrix}$

System of non-linear equations in $\mathbf{b}$:

$\Rightarrow$ In general, this system does not have an explicit solution (*it is not possible to obtain an analytical formula to compute* $\hat{\mathbf{b}}$)

Newton-Raphson/Fisher scoring algorithm

The $(r+1)$-th approximation $\mathbf{b}^{(r+1)}$ to $\hat{\mathbf{b}}$ is obtained using the recursive formula

$$\begin{aligned} \mathbf{b}^{(r+1)} &= \mathbf{b}^{(r)} + \left(\mathbf{X}^\top \mathbf{W}^{(r)} \mathbf{X}\right)^{-1} \mathbf{X}^\top \left[\mathbf{y} - \mathbf{m}^{(r)}\right] \\ &= \left(\mathbf{X}^\top \mathbf{W}^{(r)} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W}^{(r)} \mathbf{z}^{(r)} \end{aligned}$$

where $\mathbf{m} = \begin{bmatrix} n_1 \pi_1^{(r)} \\ \vdots \\ n_n \pi_n^{(r)} \end{bmatrix}$, $\mathbf{W}^{(r)} = \begin{bmatrix} n_1 \pi_1^{(r)}(1-\pi_1^{(r)}) & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & n_n \pi_n^{(r)}(1-\pi_n^{(r)}) \end{bmatrix}$,

$$\pi_i^{(r)} = \frac{\exp(\mathbf{x}_i^\top \mathbf{b}^{(r)})}{1+\exp(\mathbf{x}_i^\top \mathbf{b}^{(r)})} \quad (i = 1, \dots, n),$$

$$\mathbf{z}^{(r)} = \mathbf{X}\mathbf{b}^{(r)} + \left[\mathbf{W}^{(r)}\right]^{-1} \left[\mathbf{y} - \mathbf{m}^{(r)}\right]$$

Iterative reweighted least squares

the i-th value of the pseudo-dependent variable at the $r-$th step of the Newton-Raphson/Fisher scoring algorithm is equal to

$$\begin{aligned} z_i^{(r)} &= \mathbf{x}_i^\top \mathbf{b}^{(r)} + \frac{1}{n_i \pi_i^{(r)} \left(1 - \pi_i^{(r)}\right)} \left[y_i - n_i \pi_i^{(r)}\right] \\ &= \mathbf{x}_i^\top \mathbf{b}^{(r)} + \frac{1}{\pi_i^{(r)} \left(1 - \pi_i^{(r)}\right)} \left[p_i - \pi_i^{(r)}\right] \end{aligned}$$

$z_i^{(r)}$ can be interpreted as an approximation to $\mathrm{logit}(p_i)$ (value of the link function applied to the observed relative frequency), obtained using a first order Taylor series expansion at $\pi_i^{(r)}$

$$\mathrm{logit}(p_i) \cong \mathrm{logit}\left(\pi_i^{(r)}\right) + \left.\frac{\partial \mathrm{logit}(p_i)}{\partial y_i}\right|_{p_i = \pi_i^{(r)}} \left[p_i - \pi_i^{(r)}\right]$$

Initialisation

The Newton-Raphson/Fisher scoring algorithm can be initialised by setting

- $\mathbf{W}^{(0)} = \mathbf{I}_n$ identity matrix
- $z_i^{(0)} = \ln \frac{y_i + 0.5}{n_i - y_i + 1}$ (0.5 *and* 1 *are added in order to avoid* $\ln(0)$ or $\ln(+\infty)$)
- $\mathbf{b}^{(1)} = \left(\mathbf{X}^\top \mathbf{W}^{(0)} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W}^{(0)} \mathbf{z}^{(0)} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{z}^{(0)}$

Maximum likelihood estimator

In absence of an analytical formula for $\hat{\mathbf{B}}$, the properties of the maximum likelihood estimator must be investigated starting from the (asymptotic) properties of the score function $U(\boldsymbol{\beta})$ (*as for Poisson regression models*)

If a logistic regression model is correctly specified, it is possible to prove that

$$U(\boldsymbol{\beta}) \cong I(\boldsymbol{\beta}) \left[\hat{\mathbf{B}} - \boldsymbol{\beta}\right]$$

$U(\boldsymbol{\beta})$ *is approximately equivalent to a linear transformation of* $\hat{\mathbf{B}}$

this implies that

$$\hat{\mathbf{B}} \xrightarrow{d} MVN_{p+1}\left(\boldsymbol{\beta}, I(\boldsymbol{\beta})^{-1}\right)$$

The asymptotic properties of the maximum likelihood estimator hold both if

A) $N = \sum_i n_i \longrightarrow \infty$

and if

B) $n_i \longrightarrow \infty \; \forall i$

Estimation of the asymptotic variance

$\mathbf{X}^\top \mathbf{W} \mathbf{X}$ is uknown (it depends on $\boldsymbol{\beta}$)

$\Rightarrow$ it can be estimated using:

$$\mathbf{X}^\top \hat{\mathbf{W}} \mathbf{X}$$

where

$$\hat{\mathbf{W}} = \begin{bmatrix} n_1 \dfrac{\exp(\mathbf{x}_1^\top \hat{\mathbf{b}})}{\left[\exp(\mathbf{x}_1^\top \hat{\mathbf{b}})\right]^2} & 0 & 0 & \dots & 0 \\ 0 & n_2 \dfrac{\exp(\mathbf{x}_2^\top \hat{\mathbf{b}})}{\left[\exp(\mathbf{x}_2^\top \hat{\mathbf{b}})\right]^2} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & n_n \dfrac{\exp(\mathbf{x}_n^\top \hat{\mathbf{b}})}{\left[\exp(\mathbf{x}_n^\top \hat{\mathbf{b}})\right]^2} \end{bmatrix}$$

Logistic regression models:
Deviance, residuals, hypothesis testing and model selection criteria

Saturated model

$$M: \begin{cases} Y_i|\mathbf{x}_i \sim \text{Bin}\left(n_i, \frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right) \\ \\ \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} \end{cases} \quad \text{independent} \quad i = 1, \ldots, n$$

$\Rightarrow$ The saturated model for M is a model with a number of parameters for the expected values that is equal to the number of unique covariate patterns in the matrix $\mathbf{X}$ (equal to the number of unique values for η_i):

$$M_{sat}: Y_i|\mathbf{x}_i \sim \text{Bin}\left(n_i, \pi_i\right) \quad \text{independent} \quad i = 1, \ldots, n$$

Maximum likelihood estimation of $\pi_1, \pi_2, \ldots, \pi_n$

- Log-likelihood function of the saturated model:

$$l\left(\pi_1, \ldots, \pi_n\right) = \sum_{i=1}^n \left[y_i \ln \frac{\pi_i}{1-\pi_i} + n_i \ln(1-\pi_i) \right] + c$$

- Maximum likelihood estimate of π_i:

$$\left.\begin{array}{l} \dfrac{\partial}{\partial \pi_i} l\left(\pi_1, \ldots, \pi_n\right) = \dfrac{y_i - n_i\pi_i}{\pi_i(1-\pi_i)} \\ \\ \dfrac{\partial^2}{\partial \pi_i^2} l\left(\pi_1, \ldots, \pi_n\right) = \dfrac{-y_i + 2y_i\pi_i - n_i\pi_i^2}{\pi_i^2(1-\pi_i)^2} \end{array}\right\} \quad \Rightarrow \quad \hat{\pi}_i = \frac{y_i}{n_i} = p_i \quad i = 1, \ldots, n$$

$\Rightarrow \quad \hat{\text{E}}\left[Y_i|\mathbf{x}_i\right] = n_i p_i = y_i \quad i = 1, \ldots, n$

Evaluation of the log-likelihood function for saturated models

$$
\begin{aligned}
l(p_1,\ldots,p_n) &= \sum_{i=1}^{n}\left[y_i \ln \frac{\frac{y_i}{n_i}}{1-\frac{y_i}{n_i}} + n_i \ln\left(1-\frac{y_i}{n_i}\right)\right] \\
&= \sum_{i=1}^{n}\left[y_i \ln \frac{y_i}{n_i-y_i} + n_i \ln \frac{n_i-y_i}{n_i}\right] \\
&= \sum_{i=1}^{n}\left[y_i \ln \frac{\hat{\mathrm{E}}\left[Y_i|\mathbf{x}_i\right]}{n_i-\hat{\mathrm{E}}\left[Y_i|\mathbf{x}_i\right]} + n_i \ln \frac{n_i-\hat{\mathrm{E}}\left[Y_i|\mathbf{x}_i\right]}{n_i}\right]
\end{aligned}
$$

- $y_i = 0 \implies \hat{\pi}_i = 0 \implies 0 \ln \frac{0}{n_i} \equiv 0$
- $y_i = n_i \implies \hat{\pi}_i = 1 \implies n_i \ln \frac{n_i}{0} + n_i \ln \frac{0}{n_i} \equiv 0$

Maximum possible value for the log-likelihood associated with binomial models for $\boldsymbol{Y}|\mathbf{X}$, given the observed sample $\mathbf{y}$

$\Rightarrow$ Any logistic regression model for $\boldsymbol{Y}|\mathbf{X}$ shows a maximum value for the log-likelihood that is smaller than that value, given the observed sample $\mathbf{y}$

(Residual) deviance for a logistic regression model - 1

$$
\begin{aligned}
D &= 2 \ln \left[\frac{L(p_1,\ldots,p_n)}{L\left(\hat{\mathbf{b}}\right)}\right] = 2\left[l(p_1,\ldots,p_n) - l\left(\hat{\mathbf{b}}\right)\right] \\
&= 2\left\{\sum_{i=1}^{n}\left[y_i \ln \frac{y_i}{n_i-y_i} + n_i \ln \frac{n_i-y_i}{n_i}\right] + \right. \\
&\quad \left. - \sum_{i=1}^{n}\left[y_i \ln \frac{\hat{m}_i}{n_i-\hat{m}_i} + n_i \ln \frac{n_i-\hat{m}_i}{n_i}\right]\right\}
\end{aligned}
$$

where $\hat{m}_i = n_i \hat{\pi}_i = n_i \dfrac{\exp(\mathbf{x}_i^\top \hat{\mathbf{b}})}{1+\exp(\mathbf{x}_i^\top \hat{\mathbf{b}})}$

(Residual) deviance for a logistic regression model - 2

$$
\begin{aligned}
D &= 2\left\{\sum_{i=1}^{n}\left[y_i \ln y_i - y_i \ln(n_i - y_i) + n_i \ln(n_i - y_i) - n_i \ln n_i\right] + \right. \\
&\quad \left. - \sum_{i=1}^{n}\left[y_i \ln \hat{m}_i - y_i \ln(n_i - \hat{m}_i) + n_i \ln(n_i - \hat{m}_i) - n_i \ln n_i\right]\right\} \\
&= 2\sum_{i=1}^{n}\left\{y_i\left[\ln y_i - \ln \hat{m}_i\right] - y_i\left[\ln(n_i - y_i) - \ln(n_i - \hat{m}_i)\right] + \right. \\
&\quad \left. + n_i\left[\ln(n_i - y_i) - \ln(n_i - \hat{m}_i)\right] - n_i\left[\ln n_i - \ln n_i\right]\right\} \\
&= 2\sum_{i=1}^{n}\left[y_i \ln \frac{y_i}{\hat{m}_i} + (n_i - y_i) \ln \frac{n_i - y_i}{n_i - \hat{m}_i}\right]
\end{aligned}
$$

If $y_i = 0$ or $y_i = n_i$, then $0 \ln 0 \equiv 0$

Comparing observed and fitted values - 1

For each covariate pattern:

- observed presences/successes/positive answers y_i are compared with (estimated) expected presences/successes/positive answers $\hat{m}_i = n_i \frac{\exp(\mathbf{x}_i^\top \hat{\mathbf{b}})}{1+\exp(\mathbf{x}_i^\top \hat{\mathbf{b}})}$ through the quantity

  $$y_i \ln \frac{y_i}{\hat{m}_i}$$

- observed absences/failures/negative answers $n_i - y_i$ are compared with (estimated) expected absences/failures/negative answers $n_i - \hat{m}_i = n_i \left(1 - \frac{\exp(\mathbf{x}_i^\top \hat{\mathbf{b}})}{1+\exp(\mathbf{x}_i^\top \hat{\mathbf{b}})}\right)$ through the quantity

  $$(n_i - y_i) \ln \frac{n_i - y_i}{n_i - \hat{m}_i}$$

Comparing observed and fitted values - 2

Note that:

- if $y_i = \hat{m}_i$, then $n_i - y_i = n_i - \hat{m}_i$ and

  $$y_i \ln \frac{y_i}{\hat{m}_i} = (n_i - y_i) \ln \frac{n_i - y_i}{n_i - \hat{m}_i} = 0$$

- the larger $|y_i - \hat{m}_i|$, the larger $y_i \ln \frac{y_i}{\hat{m}_i} + (n_i - y_i) \ln \frac{n_i - y_i}{n_i - \hat{m}_i}$

(Residual) deviance for a logistic regression model - 3

Note that

$$y_i \ln \frac{y_i}{\hat{m}_i} = y_i \ln \frac{y_i}{n_i \hat{\pi}_i} = \left(\sum_{j:\eta_j = \eta_i} z_j \right) \ln \frac{\left(\sum_{j:\eta_j = \eta_i} z_j \right)}{\left(\sum_{j:\eta_j = \eta_i} \hat{\pi}_j \right)} \neq \sum_{j:\eta_j = \eta_i} z_j \ln \frac{z_j}{\hat{\pi}_j}$$

Similarly

$$(n_i - y_i) \ln \frac{n_i - y_i}{n_i - \hat{m}_i} \neq \sum_{j:\eta_j = \eta_i} (1 - z_j) \ln \frac{1 - z_j}{1 - \hat{\pi}_j}$$

Thus

$$2 \sum_{i=1}^{n} \left[y_i \ln \frac{y_i}{\hat{m}_i} + (n_i - y_i) \ln \frac{n_i - y_i}{n_i - \hat{m}_i} \right] \neq 2 \sum_{i=1}^{N} \left[z_j \ln \frac{z_j}{\hat{\pi}_j} + (1 - z_j) \ln \frac{1 - z_j}{1 - \hat{\pi}_i} \right]$$

$\Rightarrow$ *while $l(\beta_0, \ldots, \beta_p)$ is unaffected by the data structure (sample units/covariate patterns), D must always be computed using the covariate pattern data structure (unless $n_i = 1 \ \forall i$)*

$\Rightarrow$ *observed and fitted values must be compared at covariate pattern level, and not at sample unit level (unless $n_i = 1 \ \forall i$)*

An approximation to D

Second order Taylor series expansion of $y_i \ln \frac{y_i}{\hat{m}_i}$ at $\hat{m}_i$ (*as for Poisson regression models*):

$$y_i \ln \frac{y_i}{\hat{m}_i} \cong \underbrace{\hat{m}_i \ln \frac{\hat{m}_i}{\hat{m}_i}}_{0} + (y_i - \hat{m}_i) + \frac{1}{2} \frac{(y_i - \hat{m}_i)^2}{\hat{m}_i}$$

Exploiting similar arguments, it is possible to prove that:

$$\begin{aligned} n_i - y_i \ln \frac{n_i - y_i}{n_i - \hat{m}_i} &\cong [(n_i - y_i) - (n_i - \hat{m}_i)] + \frac{1}{2} \frac{[(n_i - y_i) - (n_i - \hat{m}_i)]^2}{n_i - \hat{m}_i} \\ &\cong -(y_i - \hat{m}_i) + \frac{1}{2} \frac{[(n_i - y_i) - (n_i - \hat{m}_i)]^2}{n_i - \hat{m}_i} \end{aligned}$$

Pearson X^2 statistics - 1

Plugging these two approximations in D:

$$\begin{aligned} D &= 2 \textstyle\sum_{i=1}^{n} \left[y_i \ln \frac{y_i}{\hat{m}_i} + (n_i - y_i) \ln \frac{n_i - y_i}{n_i - \hat{m}_i} \right] \\ &\cong 2 \textstyle\sum_{i=1}^{n} \left[(y_i - \hat{m}_i) + \frac{1}{2} \frac{(y_i - \hat{m}_i)^2}{\hat{m}_i} - (y_i - \hat{m}_i) + \frac{1}{2} \frac{[(n_i - y_i) - (n_i - \hat{m}_i)]^2}{n_i - \hat{m}_i} \right] \\ &\cong \textstyle\sum_{i=1}^{n} \left[\frac{(y_i - \hat{m}_i)^2}{\hat{m}_i} + \frac{[(n_i - y_i) - (n_i - \hat{m}_i)]^2}{n_i - \hat{m}_i} \right] = X^2 \end{aligned}$$

Comparing observed and fitted values - 3

	Observed	
	Successes	*Failures*
Cov. pattern $\mathbf{x}_1$	y_1	$n_1 - y_1$
Cov. pattern $\mathbf{x}_2$	y_2	$n_2 - y_2$
$\vdots$	$\vdots$	$\vdots$
Cov. pattern $\mathbf{x}_n$	y_n	$n_n - y_n$

	Expected	
	Successes	*Failures*
Cov. pattern $\mathbf{x}_1$	$\hat{m}_1$	$n_1 - \hat{m}_1$
Cov. pattern $\mathbf{x}_2$	$\hat{m}_2$	$n_2 - \hat{m}_2$
$\vdots$	$\vdots$	$\vdots$
Cov. pattern $\mathbf{x}_n$	$\hat{m}_n$	$n_n - \hat{m}_n$

The residual deviance can be approximated by applying the Pearson X^2 statistic to compare these two contingency tables (observed vs theoretical)

Pearson X^2 statistics - 2

Plugging these two approximations in D:

$$
\begin{aligned}
X^2 &= \sum_{i=1}^{n} \left[\frac{(y_i - \hat{m}_i)^2}{\hat{m}_i} + \frac{[(n_i - y_i) - (n_i - \hat{m}_i)]^2}{n_i - \hat{m}_i} \right] \\
&= \sum_{i=1}^{n} \left[\frac{(y_i - \hat{m}_i)^2}{\hat{m}_i} + \frac{(y_i - \hat{m}_i)^2}{n_i - \hat{m}_i} \right] = \sum_{i=1}^{n} \left[\frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i} + \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i(1 - \hat{\pi}_i)} \right] \\
&= \sum_{i=1}^{n} \frac{(y_i - n_i\hat{\pi}_i)^2 - (y_i - n_i\hat{\pi}_i)^2 \hat{\pi}_i^2 + (y_i - n_i\hat{\pi}_i)^2 \hat{\pi}_i^2}{n_i\hat{\pi}_i(1 - \hat{\pi}_i)} \\
&= \sum_{i=1}^{n} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1 - \hat{\pi}_i)} = \sum_{i=1}^{n} \left[\frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1 - \hat{\pi}_i)}} \right]^2
\end{aligned}
$$

Deviance residuals

Squared root (with sign) of the generic term of the deviance:

$$e_i^D = \text{sign}\,(y_i - \hat{m}_i)\sqrt{2\left[y_i \ln \frac{y_i}{\hat{m}_i} + (n_i - y_i)\ln \frac{n_i - y_i}{n_i - \hat{m}_i}\right]} \quad i = 1, \ldots, n$$

where

$$\text{sign}\,(y_i - \hat{m}_i) = \begin{cases} -1 & \text{if } \; y_i < \hat{m}_i \\ +1 & \text{if } \; y_i \geq \hat{m}_i \end{cases}$$

$\Rightarrow \quad \sum_{i=1}^n \left(e_i^D\right)^2 = D$

Pearson residuals

Squared root (with sign) of the generic term of the Pearson X^2 statistic:

$$\begin{aligned} e_i^P &= \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}} \quad i = 1, \ldots, n \\ &= \frac{y_i - \hat{\text{E}}\,[Y_i|\mathbf{x}_i]}{\sqrt{\hat{\text{Var}}\,[Y_i|\mathbf{x}_i]}} \\ &= \frac{p_i - \hat{\text{E}}\,[P_i|\mathbf{x}_i]}{\sqrt{\hat{\text{Var}}\,[P_i|\mathbf{x}_i]}} \end{aligned}$$

$\Rightarrow \quad \sum_{i=1}^n \left(e_i^P\right)^2 = X^2$

Properties of the residuals

If a logistic regression model is adequate (correctly specified):

- $\mathrm{E}\left[e_i^D \middle| \mathbf{x}_i\right] \cong \mathrm{E}\left[e_i^P \middle| \mathbf{x}_i\right] \cong 0$
- $\mathrm{Var}\left[e_i^D \middle| \mathbf{x}_i\right] \cong \mathrm{Var}\left[e_i^P \middle| \mathbf{x}_i\right] \cong 1 - H_{ii}$

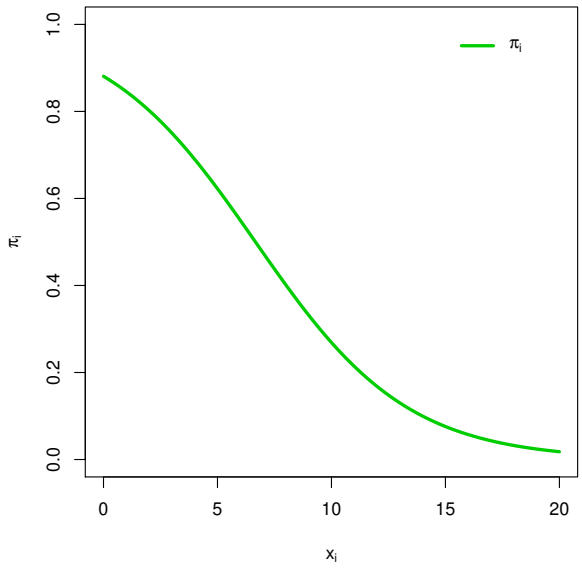
  H_{ii} is the l'i-th element of the main diagonal of the matrix

  $\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}\left(\mathbf{X}^\top\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{W}^{\frac{1}{2}}$
- Asymptotically, Pearson residuals and deviance residuals are equivalent and they are distributed as independent Gaussian random variables

  $\Rightarrow$ **WARNING**: the asymptotic properties of deviance and Pearson residuals hold if n is considered fixed and $n_i \longrightarrow \infty \; \forall i$ (*fixed cell asymptotics*)

Residual analysis when data are sparse 17

A trivial example

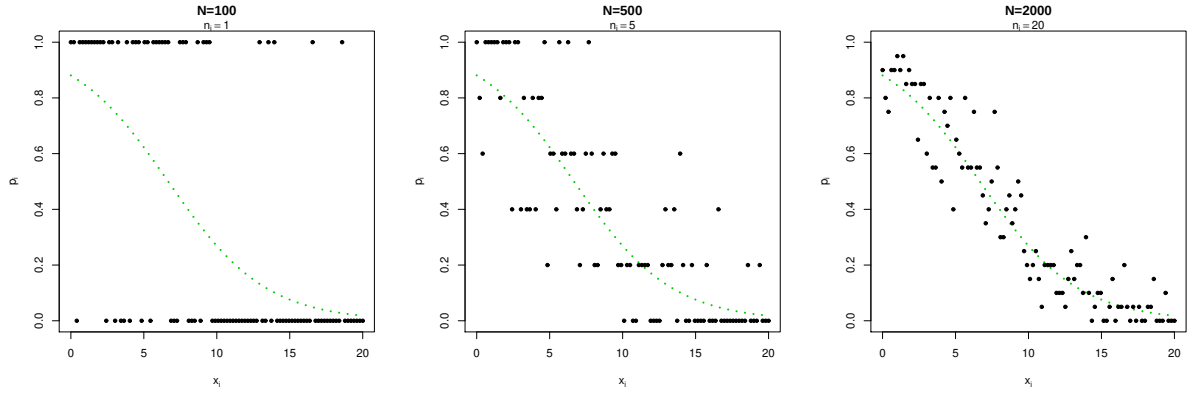


$\pi_i = \dfrac{\exp(2 - 0.3x_i)}{1 + \exp(2 - 0.3x_i)} \quad i = 1, \ldots, n$

Sampling scheme

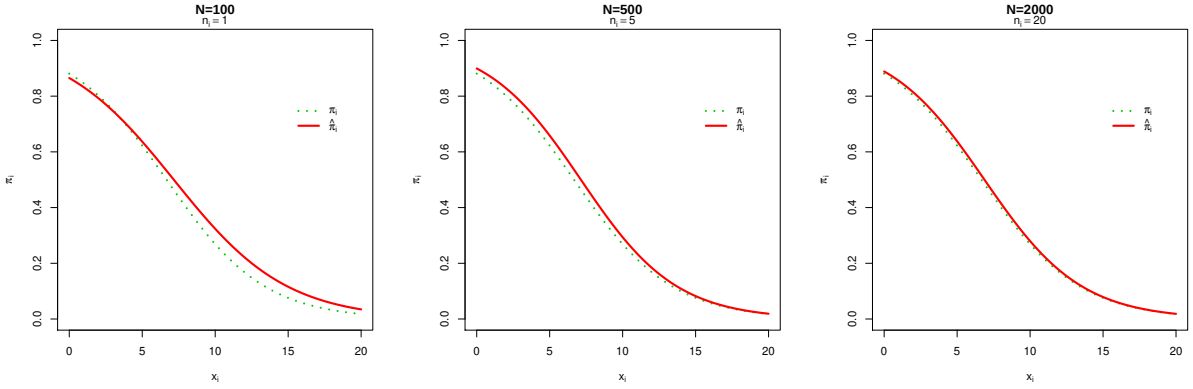
- 100 covariate patterns (*100 different values for the regressor*) $\{x_i, i = 1, \ldots, 100\}$
- 3 possible values for n_i: 1, 5, 20
- Three observed samples are simulated considering three total sample sizes $\sum_{i=1}^{n} n_i = N$ (equal to 100, 500 and 2000) using as π_i $(j = 1, \ldots, n)$ the values obtained from the function described in the previous slide
- For each observed samples,a logistic regression model is fitted (*no mispecification*)

Observed relative frequencies for each covariate pattern



Each relative frequency can take up to $n_i + 1$ *unique values:* $0, \frac{1}{n_i}, \ldots, \frac{n_i - 1}{n_i}, 1$

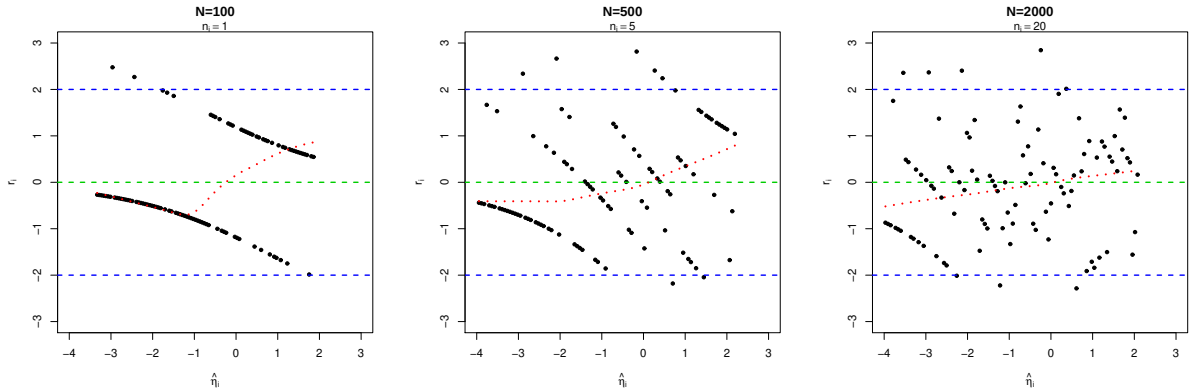
Maximum likelihood estimation



n_i	$\hat{b}_0$	$\hat{b}_1$
1	1.858	-0.259
5	2.191	-0.307
20	2.078	-0.302

All the three models provide satisfactory results (the estimated probabilities are very close to the true ones) - they can be considered adequate

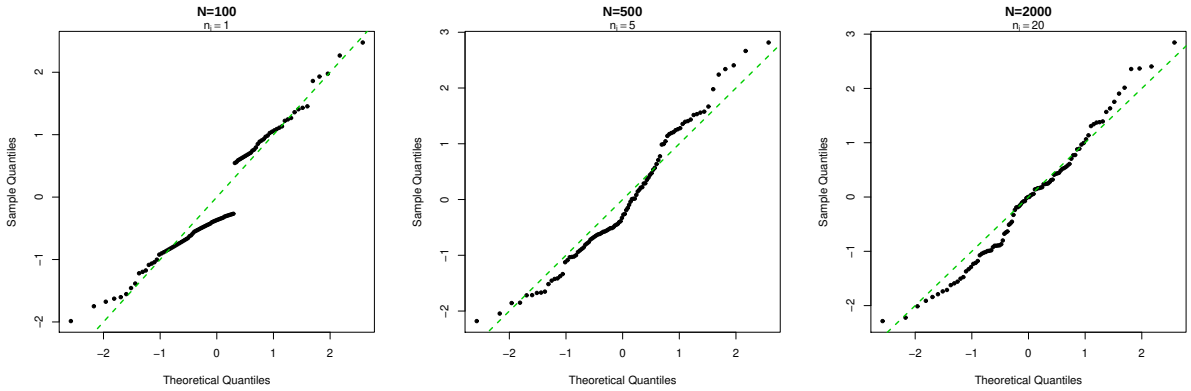
Residual plots: residuals vs (estimated) linear predictors



Despite the adequacy of the three models, these plots seems to show anomalous patterns

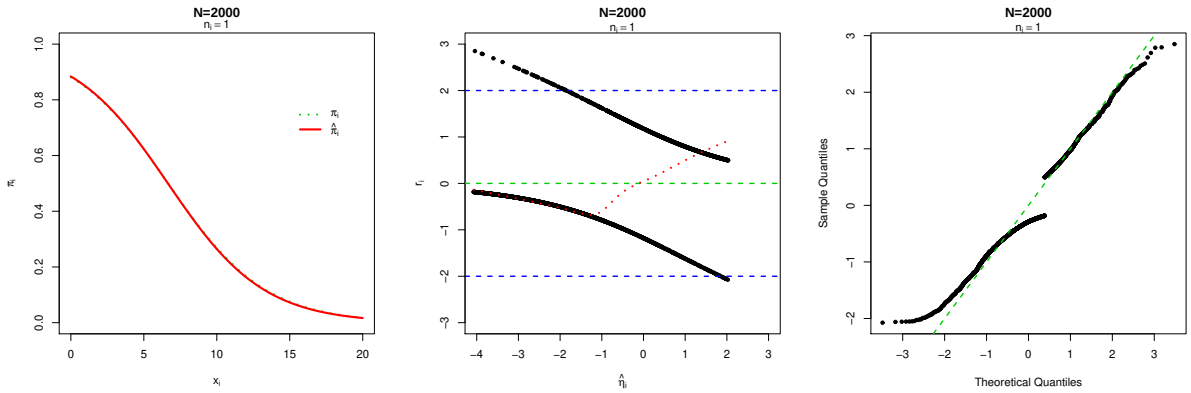
- *Residuals lies on curves, whose number depends on* n_i
- *the anomaly seems to become less evident as* n_i *increases*

Quantile-Quantile plots



The anomalous patterns have an impact also on the quantile quantile plots

An "extreme" situation: $n_i = 1$, $N = 2000$



The anonaly in the residuals does not seem to be related with the overall sample size, but it is connected withthe number of sample units associated with each covariate pattern

Residuals and data sparsity

In the context of regression models for binary outcomes, the expression "data sparsity" denotes situations in which n_i is small for any covariate pattern

$\Rightarrow$ limiting situation: $n_i = 1 \ \forall i$

Pearson residuals: $$\frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}} = \begin{cases} -\sqrt{\dfrac{\hat{\pi}_i}{1-\hat{\pi}_i}} = -\sqrt{\exp\left(\mathbf{x}_i^\top \hat{\mathbf{b}}\right)} & \text{if } y_i = 0 \\ \sqrt{\dfrac{1-\hat{\pi}_i}{\hat{\pi}_i}} = \sqrt{\exp\left(\mathbf{x}_i^\top \hat{\mathbf{b}}\right)^{-1}} & \text{if } y_i = 1 \end{cases}$$

- for each covariate pattern, the Pearson residual can take only one of these two values
- these values lie on two curves, both being continuous functions in $\hat{\eta}_i = \mathbf{x}_i^\top \hat{\mathbf{b}}$
- this behaviour is not related to the adequacy of the fitted model

$\Rightarrow$ similar conclusions can be drawn when $n_i > 1$ ($n_i + 1$ curves can be identified)

$\Rightarrow$ this sparsity problem arises also with deviance residuals

Residuals and the central limit theorem

Recall that, when $Z_1, \ldots, Z_{n_i}$ are i.i.d. Bernoulli variables

$Z_j \sim \text{Ber}(\pi_i) \ \ \text{independent} \ \ j = 1, \ldots, n_i$

then, if $n_i \longrightarrow \infty$,

$$\frac{\sum_{j=1}^{n_i} Z_j - \text{E}\left[\sum_{j=1}^{n_i} Z_j\right]}{\sqrt{\text{Var}\left[\sum_{j=1}^{n_i} Z_j\right]}} = \frac{Y_i - n_i\pi_i}{\sqrt{n_i\pi_i(1-\pi_i)}} \xrightarrow{d} N(0,1)$$

Aggregated Pearson residuals

In presence of data sparsity, Pearson (and deviance) residuals do not provide any valuable information about the adequacy of a logistic regression model

An alternative definition of residual can be used, based on aggregating sample units in homogeneous groups

$\Rightarrow$ the estimated values $\hat{\eta}_i$ (or, equivalently the estimated probabilities $\hat{\pi}_i$) can be used for defining such homogeneus groups: sample units showing similar $\hat{\eta}_i$ are assigned to the same group

$\Rightarrow$ the range of values for $\hat{\eta}_i$ can be split into G sub-intervals, according to $G-1$ thresholds
these threshold should be chosen according to the quantiles of $\hat{\eta}_i$ in order to obtain groups with similar sizes

$\Rightarrow$ when $n_i > 1$, all sample units showing the $i-$th covariate pattern are assigned to the same group

Aggregated Pearson residuals - definition

$$\bar{e}_l^P = \frac{y_l - n_l \bar{\pi}_l}{\sqrt{n_l \bar{\pi}_l (1 - \bar{\pi}_l)}} \quad l = 1, \ldots, G$$

$n_l \cong \frac{N}{G}$ number of sample units assigned to the $l-$th group

y_l observed presences/successes/positive answers for sample units assigned to the $l-$th group

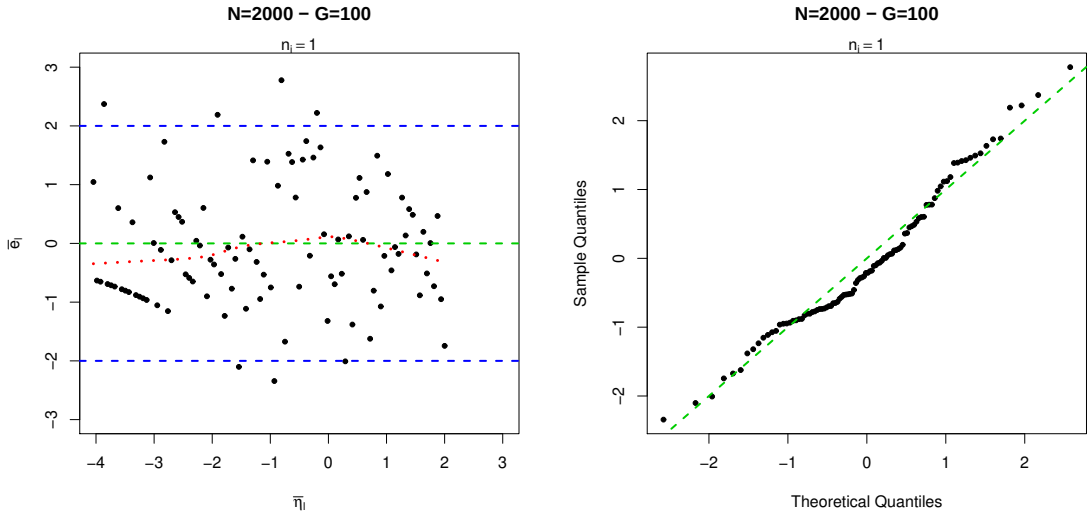
$\bar{\pi}_l$ (weighted) average of the estimated probabilities $\hat{\pi}_i$ for sample units assigned to the $l-$th group (the corresponding n_i are used as weights)

$\bar{\eta}_l$ (weighted) average of the estimated values $\hat{\eta}_i$ for sample units assigned to the $l-$th group (the corresponding n_i are used as weights)

Aggregated Pearson residuals - $n_i = 1,\ N = 2000,\ G = 10$

Intervals for $\hat{\eta}_i$	n_l	y_l	$\bar{\pi}_l$ (average $\hat{\pi}_i$)	Pearson aggregated residual $\bar{e}_l^P$	$\bar{\eta}_l$ (average $\hat{\eta}_i$)
$(-\infty$,-3.46]	200	4	0.024	-0.406	-3.770
(-3.46,-2.85]	200	5	0.044	-1.290	-3.160
(-2.85,-2.24]	200	15	0.077	-0.090	-2.549
(-2.24,-1.63]	200	24	0.131	-0.474	-1.938
(-1.63,-1.02]	200	39	0.216	-0.708	-1.328
(-1.02,-0.412]	200	74	0.333	1.112	-0.717
(-0.412,0.198]	200	105	0.475	1.407	-0.106
(0.198,0.809]	200	117	0.622	-1.073	0.504
(0.809,1.42]	200	154	0.749	0.683	1.115
(1.42,$+\infty$)	200	163	0.844	-1.145	1.725

Aggregated Pearson residuals - $n_i = 1,\ N = 2000,\ G = 100$



The anomalous pattern seems less evident after the aggregation - the plots suggest that the model can be considered adequate

Goodness of fit test (1)

If the systematic component of a logistic regression model is adequate (*correctly specified*), or, equivalently, if the null hypothesis

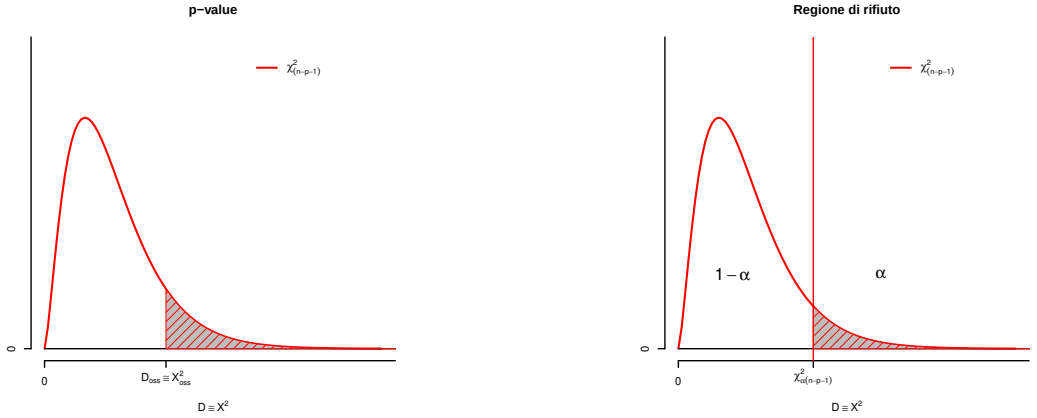
$$\begin{cases} Y_i|\mathbf{x}_i \sim \text{Bin}\,(n_i, \pi_i) \quad \text{independent} \quad i = 1, \ldots, n \\ H_0 : \text{logit}(\pi_i) = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} \end{cases}$$

is true, then

- $X^2 \xrightarrow{d} D$ (*the two statistics are asymptotically equivalent*)
- D and X^2 are asymptotically independent from $\hat{\mathbf{B}}$
- $D|\, H_0 \xrightarrow{d} \chi^2_{n-p-1}$ & $X^2 \big|\, H_0 \xrightarrow{d} \chi^2_{n-p-1}$

$\Rightarrow$ **WARNING**: the asymptotic properties of the residual deviance and of the Pearson X^2 statistic hold if n is considered fixed and $n_i \longrightarrow \infty \; \forall i$ (*fixed cell asymptotics*)

Goodness of fit test (2)



For a given significance level α

- $p-$value$< \alpha \Leftrightarrow D_{oss} \cong X^2_{oss} > \chi^2_{\alpha(n-p-1)} \Rightarrow H_0$ rejected
  $\Rightarrow$ *the model is not adequate*
- $p-$value$> \alpha \Leftrightarrow D_{oss} \cong X^2_{oss} < \chi^2_{\alpha(n-p-1)} \Rightarrow H_0$ not rejected
  $\Rightarrow$ *the model is adequate*

Goodness of fit and data sparsity

In presence of data sparsity, the distribution of D and X^2 is unknown, even if $N \longrightarrow \infty$

In particular, both D and X^2 can take large values even if the null hypothesis H_0 is true (even if the model is adequate)

it is possible to prove that, when $n_i = 1 \ \forall i$, D has a degenerate distribution, given $\hat{\mathbf{b}}$ - it takes a fixed value

$\Rightarrow$ An alternative goodness of fit statistic should be exploited

Hosmer-Lemeshow goodness of fit test

Given a partition of the sample units in G groups (as for the aggregated Pearson residuals)

$$D_{HL} = \sum_{l=1}^{G} \frac{(y_l - n_l \bar{\pi}_l)^2}{n_l \bar{\pi}_l (1 - \bar{\pi}_l)} = \sum_{l=1}^{G} \left(\bar{e}_l^P \right)^2$$

If the systematic component of a logistic regression model is adequate (*correctly specified*), or, equivalently, if the null hypothesis

$$\begin{cases} Y_i | \mathbf{x}_i \sim \text{Bin}\,(n_i, \pi_i) \quad \text{independent} \quad i = 1, \ldots, n \\ H_0 : \text{logit}(\pi_i) = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} \end{cases}$$

is true, then

$D_{HL} | \, H_0 \xrightarrow{d} \chi^2_{G-2}$ as $N \longrightarrow \infty$

$\Rightarrow$ Hosmer and Lemeshow suggest to set $G = 8$ or $G = 10$

Testing linear hypotheses on the parameters of a logistic regression model
36

Linear hypotheses on β

$H_0 : \mathbf{K}\boldsymbol{\beta} = \mathbf{t}$

$\Rightarrow$ likelihood ratio test statistic: $2 \ln \frac{L(\hat{\mathbf{b}})}{L(\hat{\mathbf{b}}_{H_0})} \Big| H_0 \xrightarrow{d} \chi^2_q$

$\Rightarrow$ Wald test statistic: $\left[\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right]^\top \left[\mathbf{K}\widehat{I(\boldsymbol{\beta})}^{-1}\mathbf{K}^\top\right]^{-1} \left[\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right] \Big| H_0 \xrightarrow{d} \chi^2_q$

- *the two statistics are asymptotically equivalent*
- These asymptotic properties hold if

  A) $N = \sum_i n_i \longrightarrow \infty$

  or

  B) $n_i \longrightarrow \infty \; \forall i$

  $\Rightarrow$ *data sparsity does not affect the asympotic behaviour of these two statistics for logistic regression models*

Testing $H_0 : \beta_h = 0$ ($h = 1, \ldots, p$)

The Wald test statistic becomes

$$\frac{b_h^2}{\widehat{I(\boldsymbol{\beta})}^{-1}_{h+1,h+1}} \Big| H_0 \xrightarrow{d} \chi^2_1$$

or, equivalently

$$\frac{b_h}{\sqrt{\widehat{I(\boldsymbol{\beta})}^{-1}_{h+1,h+1}}} \Big| H_0 \xrightarrow{d} N(0,1)$$

$\widehat{I(\boldsymbol{\beta})}^{-1}_{h+1,h+1}$ $h+1-$th element on the main diagonal of $\widehat{I(\boldsymbol{\beta})}^{-1}$
(*recall that the first element on the main diagonal refers to* β_0)

Choice among logistic regression models

Aim:
Which is the most adequate logistic regression model for a given random sample $\boldsymbol{Y}$?

$Y_i|\mathbf{x}_i \sim \text{Bin}\,(n_i, \pi_i)$, independent $(i = 1, \ldots, n)$

Simplest situation: two candidate models

$M_A : \text{logit}(\pi_i) = \eta_{Ai} = \mathbf{x}_{Ai}^\top \boldsymbol{\beta}_A, \quad \boldsymbol{\beta}_A \in \mathbb{R}^{p_A+1}$

$M_B : \text{logit}(\pi_i) = \eta_{Bi} = \mathbf{x}_{Bi}^\top \boldsymbol{\beta}_B \quad \boldsymbol{\beta}_B \in \mathbb{R}^{p_B+1}$

Note that:
the two models differ because they are characterised by different sets of regressors: $\mathbf{x}_{Ai} \neq \mathbf{x}_{Bi}\ \forall i$ (without loss of generality: $p_A > p_B$)

Situation 1: nested models - 1

Vectors $\mathbf{x}_{Bi}$ can be obtained by removing one or more that one regressor from vectors $\mathbf{x}_{Ai}$

$\Rightarrow$ M_B can be obtained by introducing suitable linear constraints on the parameteres of M_A

$$\left.\begin{array}{l} M_A : \text{logit}(\pi_i) = \eta_{Ai} = \mathbf{x}_{Ai}^\top \boldsymbol{\beta}_A \\ H_0 : \mathbf{K}_B \boldsymbol{\beta}_A = \mathbf{t}_B \end{array}\right\} \quad \Rightarrow \quad M_B : \text{logit}(\pi_i) = \eta_{Bi} = \mathbf{x}_{Bi}^\top \boldsymbol{\beta}_B$$

$q = p_A - p_B$ number of regressors excluded from M_A to obtain M_B

$\mathbf{K}_B$ $(q) \times (p_A + 1)$ matrix
each row of this matrix contains a 1 in a specific position (corresponding to one of the q regressors excluded from M_A), and 0 elsewhere

$\mathbf{t}_B = \mathbf{0}_q$

Situation 1: nested models - 2

A likelihood ratio test can be exploited to choose among M_A and M_B. In particular, such test can be expressed as a function of the two corresponding deviances:

$$\Delta l = 2\ln\left[\frac{L\left(\hat{\mathbf{b}}_A\right)}{L\left(\hat{\mathbf{b}}_{A|H_0},\right)}\right] = 2\ln\left[\frac{L\left(\hat{\mathbf{b}}_A\right)}{L\left(\hat{\mathbf{b}}_B\right)}\right] = D(M_B) - D(M_A) = \Delta D$$

If H_0 is true - if M_B is as "adequate" as M_A:

$\Rightarrow \quad \Delta D|\, M_B \xrightarrow{d} \chi^2_q$

The asymptotic properties of ΔD hold if

A) $N = \sum_i n_i \longrightarrow \infty$ (it is not necessary that $n_i \longrightarrow \infty\ \forall i$)

$\Rightarrow$ *data sparsity does not affect the asympotic properties of differences between residual deviances of logistic regression models*

Situation 1: nested models - 3

ΔD is asymptotically equivalent to

$$\left[\mathbf{K}_B\hat{\mathbf{b}}_A - \mathbf{t}_B\right]^\top \left[\mathbf{K}_B\widehat{I(\boldsymbol{\beta}_A)}^{-1}\mathbf{K}_B^\top\right]^{-1}\left[\mathbf{K}_B\hat{\mathbf{b}}_A - \mathbf{t}_B\right]$$

Also a Wald test can be exploited to choose among M_A and M_B. Such test can be expressed as a function of the ML estimator for $\boldsymbol{\beta}_A$
In particular:

$$\left[\mathbf{K}_B\hat{\mathbf{b}}_A - \mathbf{t}_B\right]^\top \left[\mathbf{K}_B\widehat{I(\boldsymbol{\beta}_A)}^{-1}\mathbf{K}_B^\top\right]^{-1}\left[\mathbf{K}_B\hat{\mathbf{b}}_A - \mathbf{t}_B\right]\Big| M_B \xrightarrow{d} \chi^2_q$$

The asymptotic properties of the Wald test statistic hold if

A) $N = \sum_i n_i \longrightarrow \infty$ (it is not necessary that $n_i \longrightarrow \infty\ \forall i$)

$\Rightarrow$ *data sparsity does not affect the asympotic properties of the Wald test statistic for logistic regression models*

Situation 2: non-nested models

Vectors $\mathbf{x}_{Bi}$ cannot be obtained by removing one or more that one regressor from vectors $\mathbf{x}_{Ai}$

$\Rightarrow$ Model M_B can be obtained by simultaneously excluding some (or all) regressors in model M_A and adding some regressors to model M_A

$\Rightarrow$ *The two models are characterised by two sets of regressors that are only partially overlapping, or non-overlapping*

$\Rightarrow$ The differences between the two deviances does not have a known random distribution, and thus a likelihood ratio test cannot be used to choose between the two models

AIC and BIC for logistic regression models

$\Rightarrow$ Akaike information criterion

$$\begin{aligned} AIC &= -2\ln L\left(\hat{\mathbf{b}}\right) + 2(p+1) \\ &= -2\sum_{i=1}^{n}\left\{ y_i \exp(\mathbf{x}_i^\top \hat{\mathbf{b}}) + n_i \ln\left[1 - \frac{\exp\left(\exp(\mathbf{x}_i^\top \hat{\mathbf{b}})\right)}{1+\exp\left(\exp(\mathbf{x}_i^\top \hat{\mathbf{b}})\right)}\right]\right\} - 2\sum_{i=1}^{n}\ln\left(\begin{array}{c} n_i \\ y_i \end{array}\right) + \\ &\quad +2(p+1) \end{aligned}$$

$\Rightarrow$ (Schwartz) Bayesian criterion

$$\begin{aligned} BIC &= -2\ln L\left(\hat{\mathbf{b}}\right) + \ln(N)\cdot(p+1) \\ &= -2\sum_{i=1}^{n}\left\{ y_i \exp(\mathbf{x}_i^\top \hat{\mathbf{b}}) + n_i \ln\left[1 - \frac{\exp\left(\exp(\mathbf{x}_i^\top \hat{\mathbf{b}})\right)}{1+\exp\left(\exp(\mathbf{x}_i^\top \hat{\mathbf{b}})\right)}\right]\right\} - 2\sum_{i=1}^{n}\ln\left(\begin{array}{c} n_i \\ y_i \end{array}\right) + \\ &\quad +\ln(N)\cdot(p+1) \end{aligned}$$

The additive constant $-2\sum_{i=1}^{n}\ln\left(\begin{array}{c} n_i \\ y_i \end{array}\right)$ *can be ignored when all competing models have a binomial probabilistic component*

Logistic regression models: an example

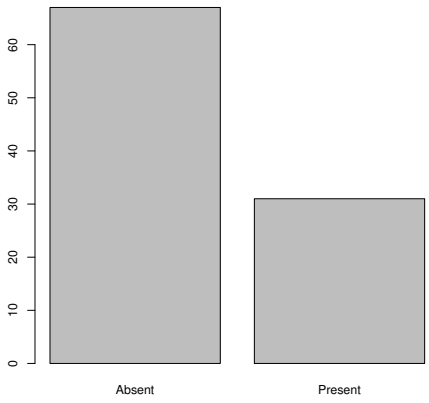
Introduction

In a health study to investigate an epidemic outbreak of a disease that is spread by mosquitoes in a city, 98 individuals were randomly sampled. For each individual, information about the following variables was collected:

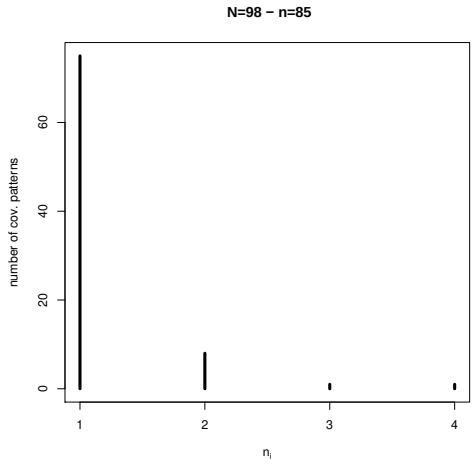
- `disease`: absence/presence of specific symptoms associated with the disease,
- `age`: age of the individual (years),
- `area`: sector of the city in which the individual lives (two categories: sector 1/sector 2),
- `status`: socio-economic status of the household to which the individual belongs (three categories: lower/medium/upper)

Is there a significant association between the presence of the disease syptoms and any of the regressors?

Dependent variable - observed values



Observations for each covariate pattern



Most of the covariate patterns are associated with only one sample unit: this data set is sparse

ML estimation 5

Dummy variable coding schemes

- `disease`:

	z
absence	0
presence	1

$\Rightarrow \quad \pi_j = \Pr(z_j = 1) = \Pr(\texttt{disease}_j = \texttt{present}) \quad j = 1, \ldots, N$

- `area`:

	areaSect2
Sector 1	0
Sector 2	1

- `status`:

	statusMiddle	statusUpper
Lower	0	0
Middle	1	0
Upper	0	1

disease vs. age, area & status - R summary output - 1

Covariate pattern data structure

```
Call:
glm(formula = cbind(yi, ni - yi) ~age + area + status, family = "binomial", data = disease.cov)

Coefficients:
              Estimate   Std.  Error   z value   Pr(>|z|)
 (Intercept)    -2.618         0.613    -4.270      0.000
 age             0.030         0.014     2.203      0.028
 areaSect2       1.575         0.502     3.139      0.002
 statusMiddle    0.714         0.654     1.092      0.275
 statusUpper     0.305         0.604     0.505      0.613

    Null deviance:  115.726 on 84 degrees of freedom
Residual deviance:  94.462 on 80 degrees of freedom
```

Note that:

- *the null deviance corresponds, up to a constant, to minus twice the maximized log-likelihood for a logistic regression model that contains only the intercept (without regressors)*
- *the residual deviance corresponds, up to a constant, to minus twice the maximized log-likelihood of the fitted model*

disease vs. age, area & status - R summary output - 2

Sample unit data structure

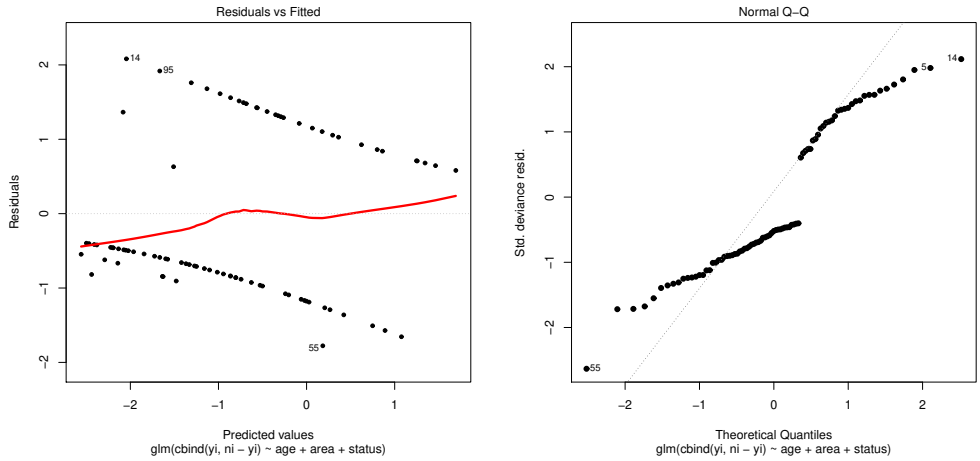
```
Call:
glm(formula = disease ~age + area + status, family = "binomial", data = disease)

Coefficients:
              Estimate   Std.  Error   z value   Pr(>|z|)
 (Intercept)    -2.618         0.613    -4.270      0.000
 age             0.030         0.014     2.203      0.028
 areaSect2       1.575         0.502     3.139      0.002
 statusMiddle    0.714         0.654     1.092      0.275
 statusUpper     0.305         0.604     0.505      0.613

    Null deviance:  122.32 on 97 degrees of freedom
Residual deviance:  101.05 on 93 degrees of freedom
```

Note that:

- *the maximum likelihood estimates (and the other related quantities) are not affected by the particular data structure*
- *the deviances obtained from the sample unti data structure are different frome the previous one (also in the corresponding degrees of freedom)*

$\Rightarrow$ *the correct values can be obtained only considering the covariate pattern data structure*

disease vs. age, area & status - residuals



Due to data sparsity, these residuals do not provide any valuable information about the adequacy of the fitted model

disease vs. age, area & status - goodness of fit test

Hosmer-Lemeshow test ($G = 10$)

Intervals for $\hat{\pi}_i$	n_l	Observed Absence	Observed Presence		Expected Absence	Expected Presence	$\bar{e}_l^P$
[0.0718,0.0897]	10.00	10.00	0.00		9.21	0.79	0.923
(0.0897,0.111]	10.00	9.00	1.00		8.98	1.02	0.020
(0.111,0.163]	10.00	9.00	2.00		9.49	1.51	-0.428
(0.163,0.185]	9.00	8.00	1.00		7.42	1.58	0.509
(0.185,0.258]	9.00	7.00	2.00		7.01	1.99	-0.006
(0.258,0.323]	10.00	7.00	3.00		7.03	2.97	-0.023
(0.323,0.42]	10.00	3.00	6.00		5.71	3.29	-1.877
(0.42,0.513]	10.00	6.00	4.00		5.36	4.64	0.404
(0.513,0.659]	10.00	5.00	5.00		4.31	5.69	0.439
(0.659,0.845]	10.00	3.00	7.00		2.47	7.53	0.390

$\chi^2_{HL} = 5.327$

$\Rightarrow \quad \Pr\left[\chi^2_{(8)} \geq 5.327\right] = 0.722$

$\Rightarrow$ *The fitted logistic regression model can be considered adequate for the data*

Interpretation of the model parameters - 1

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi})}$$

Each regressor has a non-linear effect on π_i:

$$\frac{\partial}{\partial x_{ji}} \pi_i = \beta_j \frac{\exp(\eta_i)}{[1 + \exp(\eta_i)]^2}$$

$\Rightarrow$ the direction of the change in π_i due to a unit increase in x_{ji} depends on the sign of β_j

$\Rightarrow$ the magnitude of the change depends also on the values of all the regressors

Interpretation of the model parameters - 2

Multiplicative models for the odds $\frac{\pi_i}{1-\pi_i}$:

- odds *before* a unit increase in x_{ji}

$$\frac{\pi_i}{1-\pi_i} = \exp(\beta_0) \cdot \ldots \cdot \exp(\beta_j x_{ji}) \cdot \ldots \cdot \exp(\beta_p x_{pi})$$

- odds *after* a unit increase in x_{ji}

$$\frac{\pi_{i+}}{1-\pi_{i+}} = \exp(\beta_0) \cdot \ldots \cdot \exp[\beta_j (x_{ji} + 1)] \cdot \ldots \cdot \exp(\beta_p x_{pi})$$

$\Rightarrow$ Multiplicative change due to a unit increase in x_{ji}

$$\frac{\pi_{i+}}{1-\pi_{i+}} = \exp(\beta_j) \frac{\pi_i}{1-\pi_i}$$

$\Rightarrow$ Odds ratio associated with a unit increase in x_{ji}

$$\frac{\frac{\pi_{i+}}{1-\pi_{i+}}}{\frac{\pi_i}{1-\pi_i}} = \exp(\beta_j) = \begin{cases} < 1 & \text{if } \beta_j < 0 \\ = 1 & \text{if } \beta_j = 0 \\ > 1 & \text{if } \beta_j > 0 \end{cases}$$

disease vs. age, area & status - Estimated odds ratios

	b_k	$\exp(b_k)$
age	0.030	1.0302
areaSect2	1.575	4.8295
statusMiddle	0.714	2.0422
statusUpper	0.305	1.3570

- *the odds of an individual having contracted the disease increase by about 3.0 percent with each additional year of age, for given city sector location and socio-economic status*
- *the odds of an individual from sector 2 having contracted the disease are almost five times as great as for an individual from sector 1, for given age and socio-economic status*
- *the odds of an individual with middle socio-economic status having contracted the disease are almost twice times as great as for an individual with lower socio-economic status, for given age and city sector location*
- *the odds of an individual with upper socio-economic status having contracted the disease are about 35 percent larger than the odds of an individual with lower socio-economic status, for given age and city sector location*

Hypothesis testing 14

$H_0 : \beta_{\texttt{age}} = \beta_{\texttt{Sect2}} = \beta_{\texttt{Middle}} = \beta_{\texttt{Upper}} = 0$

- Full model:

```
Coefficients:
               Estimate   Std.  Error   z value   Pr(>|z|)
 (Intercept)     -2.618         0.613    -4.270      0.000
 age              0.030         0.014     2.203      0.028
 areaSect2        1.575         0.502     3.139      0.002
 statusMiddle     0.714         0.654     1.092      0.275
 statusUpper      0.305         0.604     0.505      0.613

    Null deviance:  115.726 on 84 degrees of freedom
Residual deviance:  94.462 on 80 degrees of freedom
AIC: 107.47
```

- Reduced model:

```
Coefficients:
             Estimate   Std.  Error   z value   Pr(>|z|)
 (Intercept)    -0.771         0.217    -3.548      0.000

    Null deviance:  115.73 on 84 degrees of freedom
Residual deviance:  115.73 on 84 degrees of freedom
AIC: 120.73
```

Likelihood ratio test

```
Model                       Resid.  Df   Resid.  Dev   Df   Deviance   Pr(>Chi)
disease~1                           84         115.73
disease~age+sector+status           80          94.46    4      21.26     0.0003
```

$$2 \ln \frac{L(F)}{L(R)} = -2 \ln \left[L(R) - L(F) \right] = 115.73 - 94.46 = 21.26$$

- *At least one of the three regressors is significantly associated with the presence of the disease (at a significance level $\alpha = 0.01$)*
- *Note that the degrees of freedom for this test statistic are equal to 4, since 4 regression coefficients are set equal to 0, according to H_0*
- *the same result (in terms of test statistic and p-value) can be obtained considering models fitted using the sample unit data structure*

Wald test

```
Hypothesis:
age = 0
areaSect2 = 0
statusMiddle = 0
statusUpper = 0

 Model 1:  restricted model
Model 2:  cbind(yi, ni - yi) ~age + area + status

     Res.Df   Df   Chisq   Pr(>Chisq)
 1       84
 2       80    4   16.65       0.0023
```

The Wald test leads to the same conclusion - even when applied to the model fitted on the sample unit data structure

$H_0 : \beta_{\texttt{Middle}} = \beta_{\texttt{Upper}} = 0$

- Full model:

```
Coefficients:
                Estimate   Std.  Error   z value   Pr(>|z|)
 (Intercept)      -2.618          0.613    -4.270      0.000
 age               0.030          0.014     2.203      0.028
 areaSect2         1.575          0.502     3.139      0.002
 statusMiddle      0.714          0.654     1.092      0.275
 statusUpper       0.305          0.604     0.505      0.613

    Null deviance:  115.726 on 84 degrees of freedom
Residual deviance:  94.462 on 80 degrees of freedom
AIC: 107.47
```

- Reduced model:

```
Coefficients:
               Estimate   Std.  Error   z value   Pr(>|z|)
 (Intercept)     -2.335          0.511    -4.569      0.000
 age              0.029          0.013     2.224      0.026
 areaSect2        1.673          0.487     3.434      0.001

    Null deviance:  115.726 on 84 degrees of freedom
Residual deviance:  95.668 on 82 degrees of freedom
AIC: 104.68
```


Likelihood ratio test

```
Model                       Resid.  Df   Resid.  Dev   Df   Deviance   Pr(>Chi)
disease~age+sector                  82          95.67
disease~age+sector+status           80          94.46    2       1.21     0.5474
```

$$2 \ln \frac{L(F)}{L(R)} = -2 \ln \left[L(R) - L(F) \right] = 95.67 - 94.46 = 1.21$$

- *There are not significant differences in the probability of having the disease among the three categories of socio-economic status, for given age and city sector location*
- *Note that the degrees of freedom for this test statistic are equal to 2, since 2 regression coefficients are set equal to 0, in order to exclude the socio-economic status from the full model*

$H_0 : \beta_{\text{age}} = 0$

- Full model:

```
Coefficients:
                Estimate   Std.  Error   z value   Pr(>|z|)
 (Intercept)      -2.618          0.613    -4.270      0.000
 age               0.030          0.014     2.203      0.028
 areaSect2         1.575          0.502     3.139      0.002
 statusMiddle      0.714          0.654     1.092      0.275
 statusUpper       0.305          0.604     0.505      0.613

    Null deviance:  115.726 on 84 degrees of freedom
Residual deviance:  94.462 on 80 degrees of freedom
AIC: 107.47
```

- Reduced model:

```
Coefficients:
                Estimate   Std.  Error   z value   Pr(>|z|)
 (Intercept)      -1.917          0.481    -3.984      0.000
 areaSect2         1.620          0.486     3.336      0.001
 statusMiddle      0.713          0.636     1.120      0.263
 statusUpper       0.478          0.583     0.820      0.412

    Null deviance:  115.726 on 84 degrees of freedom
Residual deviance:  99.612 on 81 degrees of freedom
AIC: 110.62
```


Likelihood ratio test

```
Model                       Resid.  Df   Resid.  Dev   Df   Deviance   Pr(>Chi)
disease~sector+status               81          99.61
disease~age+sector+status           80          94.46    1       5.15     0.0233
```

$$2 \ln \frac{L(F)}{L(R)} = -2 \ln [L(R) - L(F)] = 99.61 - 94.46 = 5.15$$

$$\cong \left(\frac{b_{\text{age}}^2}{s^2 [b_{\text{age}}]} \right) = \frac{0.03^2}{0.014^2} = 4.854$$

The age of an individual has a significant effect on the probability of having the disease, for given city sector location and socio-economic status, but only if one considers a significance level equal to $\alpha = 0.05$

Change in the coding scheme for the dependent variable - 1

- Original coding scheme:

	z
absence	0
presence	1

- Alternative coding scheme:

	z*
absence	1
presence	0

$\Rightarrow \quad z_j^* = 1 - z_j, \quad j = 1, \ldots, N$

$\Rightarrow \quad \pi_j^* = \Pr\left(z_j^* = 1\right) = \Pr\left(\texttt{disease}_j = \texttt{absent}\right) = 1 - \Pr\left(\texttt{disease}_j = \texttt{present}\right) = 1 - \Pr\left(z_j = 1\right) = 1 - \pi_j$

disease vs. age, area & status - R summary output

- Original coding scheme:

```
Coefficients:
               Estimate   Std.  Error   z value   Pr(>|z|)
 (Intercept)     -2.618          0.613    -4.270      0.000
 age              0.030          0.014     2.203      0.028
 areaSect2        1.575          0.502     3.139      0.002
 statusMiddle     0.714          0.654     1.092      0.275
 statusUpper      0.305          0.604     0.505      0.613

Residual deviance:  94.462 on 80 degrees of freedom
```

- Alternative coding scheme:

```
Coefficients:
               Estimate   Std.  Error   z value   Pr(>|z|)
 (Intercept)      2.618          0.613     4.270      0.000
 age             -0.030          0.014    -2.203      0.028
 areaSect2       -1.575          0.502    -3.139      0.002
 statusMiddle    -0.714          0.654    -1.092      0.275
 statusUpper     -0.305          0.604    -0.505      0.613

Residual deviance:  94.462 on 80 degrees of freedom
```

The two models are equivalent (they have the same residual deviance): the change in the coding scheme affects only the signs of the regression coefficients

Change in the coding scheme for the dependent variable - 2

$$
\begin{aligned}
\pi_j^* = \Pr\left(z_j^* = 1\right) &= \frac{\exp\left(\mathbf{x}_j^\top \boldsymbol{\beta}^*\right)}{1+\exp\left(\mathbf{x}_j^\top \boldsymbol{\beta}^*\right)} = \Pr\left(z_j = 0\right) \\
&= 1 - \Pr\left(z_j = 1\right) = 1 - \frac{\exp\left(\mathbf{x}_j^\top \boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_j^\top \boldsymbol{\beta}\right)} \\
&= \frac{1+\exp\left(\mathbf{x}_j^\top \boldsymbol{\beta}\right) - \exp\left(\mathbf{x}_j^\top \boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_j^\top \boldsymbol{\beta}\right)} = \frac{1}{1+\exp\left(\mathbf{x}_j^\top \boldsymbol{\beta}\right)} \\
&= \frac{\exp\left(-\mathbf{x}_j^\top \boldsymbol{\beta}\right)}{\exp\left(-\mathbf{x}_j^\top \boldsymbol{\beta}\right)} \frac{1}{1+\exp\left(\mathbf{x}_j^\top \boldsymbol{\beta}\right)} = \frac{\exp\left(-\mathbf{x}_j^\top \boldsymbol{\beta}\right)}{1+\exp\left(-\mathbf{x}_j^\top \boldsymbol{\beta}\right)}
\end{aligned}
$$

$\Rightarrow \quad \boldsymbol{\beta}^* = -\boldsymbol{\beta}$

Change in the coding scheme for the dependent variable - 3

$$
\begin{aligned}
l^*\left(\beta_0^*, \ldots, \beta_p^*\right) &= \sum_{j=1}^{N} \left\{ z_j^* \ln \frac{\pi_j^*}{1-\pi_j^*} + \ln\left[1-\pi_j^*\right] \right\} \\
&= \sum_{j=1}^{N} \left\{ (1-z_j) \ln \frac{1-\pi_j}{\pi_j} + \ln\left[\pi_j\right] \right\} \\
&= \sum_{j=1}^{N} \left\{ -z_j \ln \frac{1-\pi_j}{\pi_j} + \ln \frac{1-\pi_j}{\pi_j} + \ln\left[\pi_j\right] \right\} \\
&= \sum_{j=1}^{N} \left\{ z_j \ln \frac{\pi_j}{1-\pi_j} + \ln\left[1-\pi_j\right] \right\} \\
&= l\left(\beta_0, \ldots, \beta_p\right)
\end{aligned}
$$

Generalised linear models for binary outcomes: Choice of the link function

GLM for binary outcomes - basic definition

- Probabilistic component

  $Y_i|\mathbf{x}_i \sim \text{Bin}(n_i, \pi_i) \iff P_i = \left.\frac{Y_i}{n_i}\right| \mathbf{x}_i \sim \text{EF}(b(\pi_i) = \text{logit}(\pi_i), \phi = 1, w_i = n_i)$

  - $\text{E}[P_i|\mathbf{x}_i] = \pi_i \implies \text{E}[Y_i|\mathbf{x}_i] = n_i\pi_i$
  - $\text{Var}[P_i|\mathbf{x}_i] = \frac{\pi_i(1-\pi_i)}{n_i} \implies \text{Var}[Y_i|\mathbf{x}_i] = n_i\pi_i(1-\pi_i)$

- Systematic component

  $\text{E}[P_i|\mathbf{x}_i] = \pi_i = h\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)$, or, equivalently, $g(\pi_i) = \mathbf{x}_i^\top\boldsymbol{\beta}$

  $\Rightarrow$ the choice of the link function affects only the systematic component

  $\Rightarrow$ since $\pi_i \in [0,1]$, the link function $h(\cdot)$ should be chosen among all functions satisfying the following requirement: $h(\cdot) : \mathbb{R} \longmapsto [0,1]$

  $\Rightarrow$ furthermore, $h(\cdot)$ should be differentiable and invertible, so that $\mathbf{x}_i^\top\boldsymbol{\beta} = h^{-1}(\pi_i) = g(\pi_i)$

Common choices for $g(\cdot)$

- logit function: $\text{logit}(\pi_i) = \ln\frac{\pi_i}{1-\pi_i} = \mathbf{x}_i^\top\boldsymbol{\beta}$ (*canonical link function*)
- probit function: $\text{probit}(\pi_i) = \Phi^{-1}(\pi_i) = \mathbf{x}_i^\top\boldsymbol{\beta}$

  $\Phi^{-1}(\cdot)$ denotes the inverse (cumulative) probability function of a standard Gaussian r.v.
- c-log-log function: $\ln[-\ln(1-\pi_i)] = \mathbf{x}_i^\top\boldsymbol{\beta}$

Historical perspective

- 1922: Fisher introduces the c-log-log link function
- 1933: first use of the probit link function
- 1944: the logit link function is proposed
- 1972: Nelder and Wedderburn present their first work on GLM

Probability functions as link functions - 1

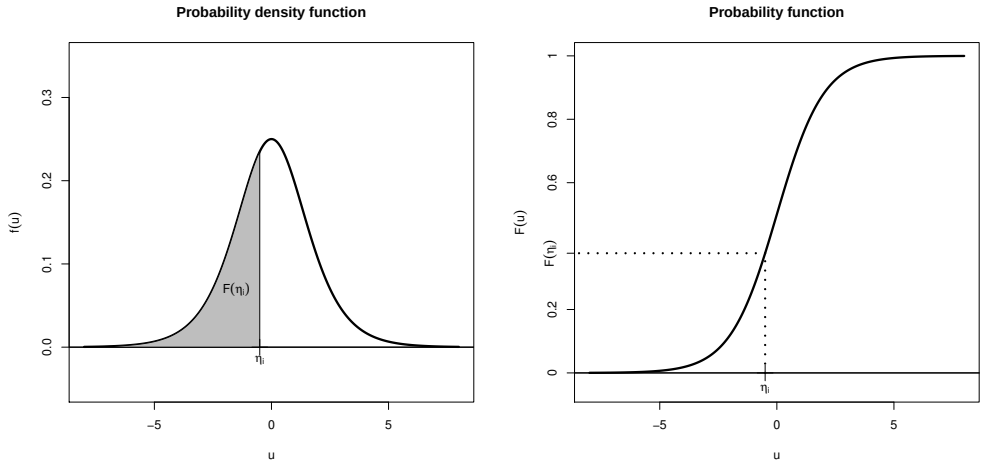
U	absolutely continuous r.v.
$f(u)$	(known) probability density function of U such that $f(u) > 0 \ \forall u \in \mathbb{R}$
$F(u) = \int_{-\infty}^{u} f(t)dt$	(cumulative) probability function of U

$\Rightarrow$ $F(\cdot) : \mathbb{R} \longmapsto [0, 1]$

$\Rightarrow$ $F(\cdot)$ is differentiable: $\dfrac{\partial}{\partial u} F(u) = f(u)$

$\Rightarrow$ $F(\cdot)$ is invertible: $f(u) > 0$ implies that $F(u)$ is monotonically increasing

Probability functions as link functions - 2



Probability functions as link functions: some examples

- *Logistic distribution*

  $f(u) = \dfrac{\exp(u)}{[1+\exp(u)]^2} \quad \Rightarrow \quad F(u) = \dfrac{\exp(u)}{1+\exp(u)}$

  $\Rightarrow \quad \mathrm{E}[U] = 0, \quad \mathrm{Var}[U] = \dfrac{\pi^2}{3}$

- *Standard normal distribution*

  $f(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \quad \Rightarrow \quad F(u) = \Phi(u)$

  Recall that $\Phi(u)$ *does not have a closed form expression*

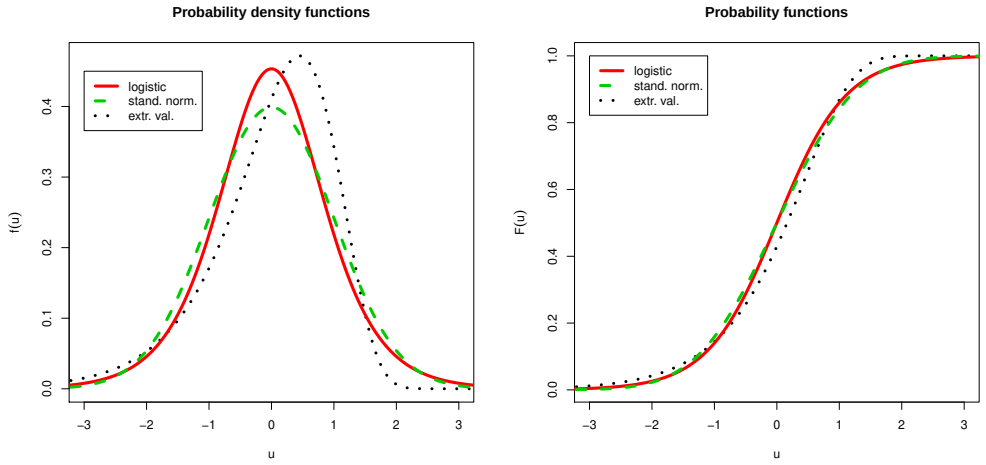
  $\Rightarrow \quad \mathrm{E}[U] = 0, \quad \mathrm{Var}[U] = 1$

- *Extreme (minimum) value distribution*

  $f(u) = \exp\left[u - \exp(u)\right] \quad \Rightarrow \quad F(u) = 1 - \exp\left[-\exp(u)\right]$

  $\Rightarrow \quad \mathrm{E}[U] = -0.5772, \quad \mathrm{Var}[U] = \dfrac{\pi^2}{6}$

Probability functions as link functions - graphical comparisons



Note that the probability density functions and the corresponding probability functions have been rescaled in order to have expected values equal to 0 and variances equal to 1

Link functions for GLMs for binary outcomes

- *logit function*

  $$\pi_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)} \quad \Rightarrow \quad \eta_i = \ln \frac{\pi_i}{1-\pi_i}$$

- *probit function*

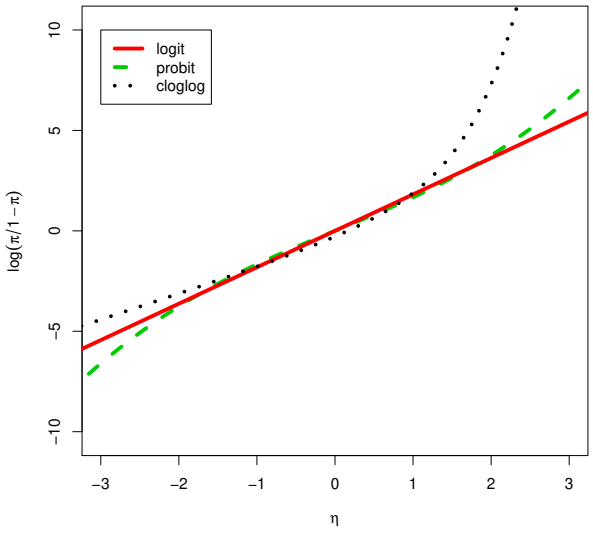
  $$\pi_i = \Phi(\eta_i) \quad \Rightarrow \quad \eta_i = \Phi^{-1}(\pi_i)$$

  Recall that $\Phi(\eta_i)$ *and* $\Phi^{-1}(\pi_i)$ *do not have closed form expressions*

- *c-log-log function*

  $$\pi_i = 1 - \exp\left[-\exp(\eta_i)\right] \quad \Rightarrow \quad \eta_i = \ln\left[-\ln(1-\pi_i)\right]$$

Link functions - graphical comparison on the logit scale



Note that the values of the linear predictors have been rescaled to remove possible differences in the scale of the regression coefficients

Some comments

- when considering values for π_i the range from 0.1 to 0.9 (corresponding to values for $\text{logit}(\pi_i)$ in the range from -2.197 to 2.197), the three functions are almost equivalent, and show an almost linear behaviour
- the probit function approaches estreme values for π_i (0 and 1) at a faster rate than the logit function

  $\Rightarrow$ *the logistic distribution has thicker tails than the standard normal distribution*
- the c-log-log function has different tail behaviours:
  - it approaches the value $\pi_i = 0$ at a slower rate than the other two functions
  - it approaches the value $\pi_i = 1$ at a faster rate than the other two functions

  $\Rightarrow$ *the extreme (minimum) value distribution is skewed to the left*

Log-likelihood

$$l(\beta_0, \ldots, \beta_p) = \sum_{i=1}^{n} \left\{ y_i \ln \frac{h\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)}{1 - h\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)} + n_i \ln \left[1 - h\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right] \right\} + c$$

If $h\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right) \neq \dfrac{\exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)}{1 + \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)}$, then $\ln \dfrac{h\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)}{1 - h\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)} \neq \mathbf{x}_i^\top \boldsymbol{\beta}$

$\Rightarrow$ $\sum_{i=1}^{n} y_i, \ldots, \sum_{i=1}^{n} y_i x_{pi}$ are not sufficient statistics for $\boldsymbol{\beta}$

Score function

Considering the general formula for the generic element of the score function for a GLM:

$$\begin{aligned}
U_j(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{p_i - \mathrm{E}\left[P_i|\mathbf{x}_i\right]}{\mathrm{Var}\left[P_i|\mathbf{x}_i\right]} \frac{\partial \mathrm{E}\left[P_i|\mathbf{x}_i\right]}{\partial \eta_i} x_{ji} \\
&= \sum_{i=1}^{n} \frac{\frac{y_i}{n_i} - \pi_i}{\frac{\pi_i(1-\pi_i)}{n_i}} \frac{\partial \pi_i}{\partial \eta_i} x_{ji} \\
&= \sum_{i=1}^{n} \frac{1}{n_i} \frac{y_i - n_i\pi_i}{\frac{\pi_i(1-\pi_i)}{n_i}} \frac{\partial \pi_i}{\partial \eta_i} x_{ji} \\
&= \sum_{i=1}^{n} \frac{y_i - \mathrm{E}\left[Y_i|\mathbf{x}_i\right]}{\pi_i(1-\pi_i)} \frac{\partial \pi_i}{\partial \eta_i} x_{ji}
\end{aligned}$$

Observed Fisher information

Considering the general formula for the generic element of the observed Fisher information matrix for a GLM:

$$
\begin{aligned}
i_{jl}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{x_{ji}x_{li}}{\mathrm{Var}\left[P_i|\mathbf{x}_i\right]} \left(\frac{\partial \mathrm{E}\left[P_i|\mathbf{x}_i\right]}{\partial \eta_i}\right)^2 + \\
&\quad - \sum_{i=1}^{n} \left\{p_i - \mathrm{E}\left[P_i|\mathbf{x}_i\right]\right\} \frac{\partial}{\partial \beta_l} \left(\frac{x_{ji}}{\mathrm{Var}\left[P_i|\mathbf{x}_i\right]} \frac{\partial \mathrm{E}\left[P_i|\mathbf{x}_i\right]}{\partial \eta_i}\right) \\
&= \sum_{i=1}^{n} \frac{n_i x_{ji}x_{li}}{\pi_i(1-\pi_i)} \left(\frac{\partial \pi_i}{\partial \eta_i}\right)^2 + \\
&\quad - \sum_{i=1}^{n} \left\{y_i - n_i\pi_i\right\} \frac{\partial}{\partial \beta_l} \left(\frac{x_{ji}}{\pi_i(1-\pi_i)} \frac{\partial \pi_i}{\partial \eta_i}\right)
\end{aligned}
$$

Expected Fisher information

Considering the general formula for the generic element of the expected Fisher information matrix for a GLM:

$$
\begin{aligned}
I_{jl}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{x_{ji}x_{li}}{\mathrm{Var}\left[P_i|\mathbf{x}_i\right]} \left(\frac{\partial \mathrm{E}\left[P_i|\mathbf{x}_i\right]}{\partial \eta_i}\right)^2 \\
&= \sum_{i=1}^{n} \frac{n_i x_{ji}x_{li}}{\pi_i(1-\pi_i)} \left(\frac{\partial \pi_i}{\partial \eta_i}\right)^2 \neq i_{jl}(\boldsymbol{\beta})
\end{aligned}
$$

Matrix representation

- Score function

$$U(\boldsymbol{\beta}) = \mathbf{X}^\top \mathbf{Q}\left[\mathbf{y} - \boldsymbol{\mu}\right]$$

- Expected Fisher information matrix

$$I(\boldsymbol{\beta}) = \mathbf{X}^\top \mathbf{Q}\mathbf{W}\mathbf{Q}\mathbf{X} \neq i(\boldsymbol{\beta})$$

where

$$\mathbf{Q} = \begin{bmatrix} \frac{1}{\pi_1(1-\pi_1)}\frac{\partial \pi_1}{\partial \eta_1} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\pi_n(1-\pi_n)}\frac{\partial \pi_n}{\partial \eta_n} \end{bmatrix}$$

If $\pi_i = \frac{\exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)}$, then

$\Rightarrow \quad \frac{\partial \pi_i}{\partial \eta_i} = \pi_i(1-\pi_i)$

$\Rightarrow \quad \mathbf{Q} = \mathbf{I}_n$

Maximum likelihood estimation

Generally, maximum likelihood estimates of $\boldsymbol{\beta}$ can be obtained only using numerical optimisation techniques

$\Rightarrow$ since $I(\boldsymbol{\beta}) \neq i(\boldsymbol{\beta})$, the Newton-Raphson algorithm does not coincide with the Fisher Scoring algorithm, but the differences in the final results are usually negligible

$\Rightarrow$ the Fisher Scoring algorthm is usually preferred, due to its numerical stability ($I(\boldsymbol{\beta})$ is always invertible)

$\Rightarrow$ the recursive formual associated with the Fisher Scoring algorithm can be expressed as the solution of an iterative reweighted least square problem

$$\begin{aligned} \mathbf{b}^{(r+1)} &= \mathbf{b}^{(r)} + \left(\mathbf{X}^\top \mathbf{Q}^{(r)} \mathbf{W}^{(r)} \mathbf{Q}^{(r)} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Q}^{(r)} \left[\mathbf{y} - \mathbf{m}^{(r)}\right] \\ &= \left(\mathbf{X}^\top \mathbf{Q}^{(r)} \mathbf{W}^{(r)} \mathbf{Q}^{(r)} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Q}^{(r)} \mathbf{W}^{(r)} \mathbf{z}^{(r)} \end{aligned}$$

where

$$\mathbf{z}^{(r)} = \mathbf{Q}^{(r)}\mathbf{X}\mathbf{b}^{(r)} + \left[\mathbf{W}^{(r)}\right]^{-1}\left[\mathbf{y} - \mathbf{m}^{(r)}\right]$$

and $\mathbf{Q}^{(r)}$ is obtained by evaluating $\mathbf{Q}$ at $\mathbf{b}^{(r)}$

Maximum likelihood estimator

If a GLM for a binary outcome is correctly specified, it is possible to prove that

$$\hat{\mathbf{B}} \xrightarrow{d} MVN_{p+1}\left(\boldsymbol{\beta}, I\left(\boldsymbol{\beta}\right)^{-1}\right)$$

The asymptotic properties of the maximum likelihood estimator hold both if $N = \sum_i n_i \longrightarrow \infty$

(it is not necessary that $n_i \longrightarrow \infty \; \forall i$)

The asymptotic variance of $\hat{\mathbf{B}}$ can be estimated using

$$\widehat{I\left(\boldsymbol{\beta}\right)}^{-1} = \left[\mathbf{X}^\top \hat{\mathbf{Q}} \hat{\mathbf{W}} \hat{\mathbf{Q}} \mathbf{X}\right]^{-1}$$

where $\hat{\mathbf{Q}}$ is obtained by evaluating $\mathbf{Q}$ at $\hat{\mathbf{b}}$

Deviance, residuals and goodness of fit tests

$$\begin{aligned} D &= 2\sum_{i=1}^{n}\left[y_i \ln \frac{y_i}{\hat{m}_i} + (n_i - y_i)\ln\frac{n_i - y_i}{n_i - \hat{m}_i}\right] \\ &\cong \sum_{i=1}^{n} \frac{\left(y_i - n_i\hat{\pi}_i\right)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)} = X^2 \\ D_{HL} &= \sum_{l=1}^{G} \frac{\left(y_l - n_l\bar{\pi}_l\right)^2}{n_l\bar{\pi}_l(1-\bar{\pi}_l)} = \sum_{l=1}^{G} \bar{e}_l^P \end{aligned}$$

$\Rightarrow$ the choice of the link function affects only the formula for computing $\hat{\pi}_i = h\left(\mathbf{x}_i^\top \hat{\mathbf{b}}\right)$

$\Rightarrow$ the effects and remedies to deal with data sparsity are the same, even when using non-canonical link functions

$\Rightarrow$ *a change in the link function may improve/deteriorate the adequacy of a model*

Hypothesis testing and model comparisons

⇒ Linear hypotheses on $\boldsymbol{\beta}$ can be tested using either the likelihood ratio test statistic or the Wald test statistic, no matter which link function has been chosen

Recall that both test statistics can be used assuming that the model - and hence the link function - is adequate

⇒ Two GLMs for a binary outcome can be nested if and only if they have the same link function

⇒ GLMs for a binary outcome with different link functions can be conpared only through model selectionc citeria such as the AIC or the BIC

An example - continued 22

Introduction

In a health study to investigate an epidemic outbreak of a disease that is spread by mosquitoes in a city, 98 individuals were randomly sampled. For each individual, information about the following variables was collected:

- `disease`: absence/presence of specific symptoms associated with the disease,
- `age`: age of the individual (years),
- `area`: sector of the city in which the individual lives (two categories: sector 1/sector 2),
- `status`: socio-economic status of the household to which the individual belongs (three categories: lower/medium/upper)

Is there a significant association between the presence of the disease syptoms and any of the regressors?

disease vs. age, area & status - maximum likelihood estimates

Logistic regression model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.618	0.613	-4.270	0.000
age	0.030	0.014	2.203	0.028
areaSect2	1.575	0.502	3.139	0.002
statusMiddle	0.714	0.654	1.092	0.275
statusUpper	0.305	0.604	0.505	0.613

Probit regression model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.593	0.339	-4.696	0.000
age	0.018	0.008	2.316	0.021
areaSect2	0.953	0.295	3.225	0.001
statusMiddle	0.442	0.383	1.156	0.248
statusUpper	0.187	0.352	0.532	0.594

Bernoulli GLM with c-log-log link function

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.475	0.497	-4.976	0.000
age	0.021	0.009	2.283	0.022
areaSect2	1.215	0.405	3.002	0.003
statusMiddle	0.685	0.518	1.322	0.186
statusUpper	0.287	0.495	0.581	0.561

Some comments

- the link function affects the scale of the regression coefficients: it does not make sense to compare the estimated regression coefficient for the same regressor obtained using different link functions
- $\dfrac{\partial \pi_i}{\partial x_{ji}} = \dfrac{\partial \pi_i}{\partial \eta_i}\dfrac{\partial \eta_i}{\partial x_{ji}} = \beta_j \dfrac{\partial h(\eta_i)}{\partial \eta_i}$

  When $\dfrac{\partial h(\eta_i)}{\partial \eta_i} > 0$ (*the link function is monotonically increasing*):

  $\Rightarrow$ the direction of the change in π_i due to a unit increase in x_{ji} depends on the sign of β_j

  $\Rightarrow$ the magnitude of the change depends also on the values of all the regressors
- there is not any connection between the regression coefficients obtained using non-canonical link functions and the odds ratio

  $$\frac{\dfrac{\pi_{i+}}{1-\pi_{i+}}}{\dfrac{\pi_i}{1-\pi_i}}$$

Probit link function - goodness of fit test

Hosmer-Lemeshow test ($G = 10$)

Intervals for $\hat{\pi}_i$	n_l	Observed Absence	Observed Presence	Expected Absence	Expected Presence
[0.0598,0.0796]	10.00	10.00	0.00	9.33	0.67
(0.0796,0.103]	10.00	9.00	1.00	9.07	0.93
(0.103,0.162]	11.00	9.00	2.00	9.54	1.46
(0.162,0.186]	9.00	8.00	1.00	7.42	1.58
(0.186,0.261]	9.00	7.00	2.00	6.98	2.02
(0.261,0.33]	10.00	6.00	4.00	6.97	3.03
(0.33,0.421]	9.00	4.00	5.00	5.68	3.32
(0.421,0.513]	10.00	6.00	4.00	5.35	4.65
(0.513,0.659]	10.00	5.00	5.00	4.32	5.68
(0.659,0.852]	10.00	3.00	7.00	2.45	7.55

$\chi^2_{HL} = 3.5298$

$\Rightarrow \quad \Pr\left[\chi^2_{(8)} \geq 3.5298\right] = 0.8969$

$\Rightarrow$ *The fitted probit regression model can be considered adequate for the data*

C-log-log link function - goodness of fit test

Hosmer-Lemeshow test ($G = 10$)

Intervals for $\hat{\pi}_i$	n_l	Observed Absence	Observed Presence	Expected Absence	Expected Presence
[0.0841,0.0993]	10.00	10.00	0.00	9.10	0.90
(0.0993,0.119]	10.00	9.00	1.00	8.89	1.11
(0.119,0.168]	10.00	8.00	2.00	8.59	1.41
(0.168,0.185]	10.00	9.00	1.00	8.20	1.80
(0.185,0.252]	9.00	8.00	1.00	7.03	1.97
(0.252,0.303]	10.00	4.00	6.00	7.23	2.77
(0.303,0.399]	9.00	5.00	4.00	5.93	3.07
(0.399,0.514]	10.00	5.00	5.00	5.49	4.51
(0.514,0.67]	10.00	7.00	3.00	4.40	5.60
(0.67,0.911]	10.00	2.00	8.00	2.25	7.75

$\chi^2_{HL} = 10.86$

$\Rightarrow \quad \Pr\left[\chi^2_{(8)} \geq 10.86\right] = 0.2098$

$\Rightarrow$ *The fitted Bernoulli GLM model with c-log-log link function can be considered adequate for the data*

Comparison among link functions

Link function	D	D_{HL}	AIC
logit	94.462	5.3267	107.47
probit	94.081	3.5298	107.09
c-log-log	94.689	10.8600	107.70

$\Rightarrow$ D should not be used to compare GLMs with different link functions, as they are not nested models (*the distribution of the difference between two residual deviances is not known for non-nested models*)

$\Rightarrow$ D_{HL} should not be used to compare Bernoulli GLMs with different link functions, as the values of D_{HL} have been obtained using different splittings of the data into subgroups (*even if the number of subgroups is the same, the composition of the subgroups depends on the estimated* $\hat{p}_i$)

$\Rightarrow$ the "best" link function can be selected according to the AIC (*note that there is no guarantee that the model with the smallest* AIC *is an adequate model*)

Generalised linear models
with Gamma probabilistic component

Gamma distributions (1)

$Y \sim \mathrm{Ga}(\mu, \nu)$

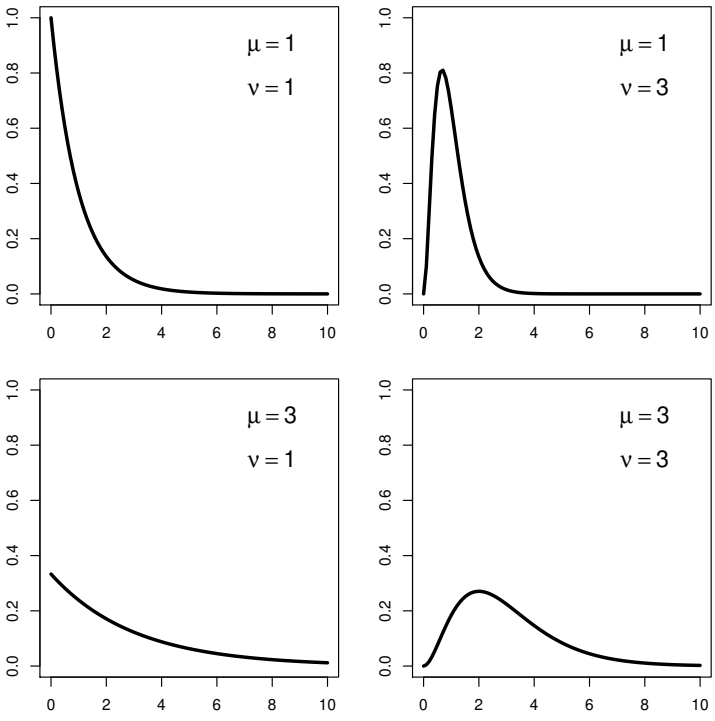
$$f(y; \mu, \nu) = \frac{1}{\left(\frac{\mu}{\nu}\right)^{\nu} \Gamma(\nu)} y^{\nu-1} \exp\left(-\frac{\nu}{\mu} y\right)$$

$y \in \mathbb{R}^+$
$\mu \in \mathbb{R}^+$
$\nu \in \mathbb{R}^+$

Gamma itegral:

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

Gamma distributions (2)



Use of Gamma distributions

Statistical phenomena that:

- take only positive values
- show skewed distributions

Some examples:

- *intensities/densities*
- *durations/waiting times*
- *earnings/expenditures*

Moments of Gamma distributions

$\mathrm{E}[Y] = \mu$

$\mathrm{Var}[Y] = \dfrac{\mu^2}{\nu}$

Any Gamma random variable has variance proportional to the squared expected value

$\mathrm{CV}[Y] = \dfrac{\sqrt{\mathrm{Var}[Y]}}{\mathrm{E}[Y]} = \dfrac{\sqrt{\frac{\mu^2}{\nu}}}{\mu} = \dfrac{1}{\sqrt{\nu}}$

$\Rightarrow$ the coefficient of variation does not depend on the expected value

Some special cases

- *Exponential variables*

  $Y \sim \text{Exp}(\mu) \Longrightarrow Y \sim \text{Ga}(\mu, \nu = 1)$ $(\Gamma(1) = 1)$

- χ^2 *variables*

  $Y \sim \chi^2(g)$ with $g \in \mathbb{N}^+ \Longrightarrow Y \sim \text{Ga}(\mu = g, \nu = g/2)$

Alternative parameterisations

- $\frac{\mu}{\nu} = \alpha$ *scale parameter*

  $$Y \sim \text{Ga}(\alpha, \nu) \Longrightarrow f(y; \alpha, \nu) = \frac{1}{(\alpha)^\nu \, \Gamma(\nu)} y^{\nu-1} \exp\left(-\frac{y}{\alpha}\right)$$

- $\frac{\nu}{\mu} = \theta$ *rate*

  $$Y \sim \text{Ga}(\theta, \nu) \Longrightarrow f(y; \theta, \nu) = \frac{\theta^\nu}{\Gamma(\nu)} y^{\nu-1} \exp\left(-\theta y\right)$$

GLM con componente casuale Gamma

$Y_i \sim \text{Ga}(\mu_i, \nu) \quad i = 1, \ldots, n$ conditionally indenpendent

$$\begin{aligned} f(y_i; \mu_i, \nu) &= \frac{1}{\left(\frac{\mu_i}{\nu}\right)^{\nu} \Gamma(\nu)} y_i^{\nu-1} \exp\left(-\frac{\nu}{\mu_i} y_i\right) \\ &= \exp\left\{ \frac{1}{\frac{1}{\nu}} \left[y_i \left(-\frac{1}{\mu_i}\right) - \ln \mu_i \right] + \left[(\nu - 1) \ln y_i - \ln \Gamma(\nu) + \nu \ln \nu\right] \right\} \end{aligned}$$

- $\theta_i = \mu_i$, $\phi = \frac{1}{\nu}$, $w_i = 1$
- $a(y_i) = y_i$
- $b(\mu_i) = -\frac{1}{\mu_i}$
- $c(\mu_i) = -\ln \mu_i$
- $d(y_i, \nu) = (\nu - 1) \ln y_i - \ln \Gamma(\nu) + \nu \ln \nu$

  $\Rightarrow \quad Y_i \sim \text{Ef}\left(-\frac{1}{\mu_i}, \frac{1}{\nu}, w_i = 1\right) \quad i = 1, \ldots, n$ conditionally indenpendent

Systematic component

- linear predictor

  $\eta_i = \mathbf{x}_i^{\top} \boldsymbol{\beta} = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$
- Link function

  $g(\text{E}[Y_i | \mathbf{x}_i]) = g(\mu_i) = \eta_i$

  $h(\eta_i) = \mu_i$

  $g(\cdot) : \mathbb{R}^+ \longmapsto \mathbb{R}; \quad h(\cdot) : \mathbb{R} \longmapsto \mathbb{R}^+$

  - *Canonical link function*: $b(\mu_i) = -\frac{1}{\mu_i} = \eta_i$

    WARNING: η_i must be strictly negative, in order to ensure that $\mu_i \in \mathbb{R}^+$
  - $\ln \mu_i = \eta_i$
  - $\ldots$

Log-likelihood and score function

$$
\begin{aligned}
l(\beta_0, \ldots, \beta_p, \nu) &= \sum_{i=1}^{n} \left\{ -\frac{\nu}{h(\eta_i)} y_i - \nu \ln h(\eta_i) + (\nu - 1) \ln y_i - \ln \Gamma(\nu) + \nu \ln \nu \right\} \\
U_j(\boldsymbol{\beta}) &= \frac{\partial}{\partial \beta_j} l(\beta_0, \ldots, \beta_p, \nu) \\
&= \sum_{i=1}^{n} \left\{ \frac{y_i - \mathrm{E}[Y_i|\mathbf{x}_i]}{\mathrm{Var}[Y_i|\mathbf{x}_i]} \frac{\partial \mathrm{E}[Y_i|\mathbf{x}_i]}{\partial \eta_i} x_{ji} \right\} \\
&= \sum_{i=1}^{n} \left\{ \nu \frac{y_i - \mu_i}{\mu_i^2} \frac{\partial \mu_i}{\partial \eta_i} x_{ji} \right\}
\end{aligned}
$$

Fisher information matrices

$$
\begin{aligned}
i_{jl}(\boldsymbol{\beta}) &= -\frac{\partial^2}{\partial \beta_j \beta_l} l(\beta_0, \ldots, \beta_p, \nu) \\
&= \sum_{i=1}^{n} \left\{ \frac{x_{ji} x_{li}}{\mathrm{Var}[Y_i|\mathbf{x}_i]} \left(\frac{\partial \mathrm{E}[Y_i|\mathbf{x}_i]}{\partial \eta_i} \right)^2 \right\} + \\
&\quad - \sum_{i=1}^{n} \left\{ (y_i - \mathrm{E}[Y_i|\mathbf{x}_i]) \frac{\partial}{\partial \beta_l} \left(\frac{x_{ji}}{\mathrm{Var}[Y_i|\mathbf{x}_i]} \frac{\partial \mathrm{E}[Y_i|\mathbf{x}_i]}{\partial \eta_i} \right) \right\} \\
&= \sum_{i=1}^{n} \left\{ \frac{\nu x_{ji} x_{li}}{\mu_i^2} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \right\} - \sum_{i=1}^{n} \left\{ (y_i - \mu_i) \frac{\partial}{\partial \beta_l} \left(\frac{\nu x_{ji}}{\mu_i^2} \frac{\partial \mu_i}{\partial \eta_i} \right) \right\} \\
I_{jl}(\boldsymbol{\beta}) &= E[i_{jl}(\boldsymbol{\beta})] = \sum_{i=1}^{n} \left\{ \frac{\nu x_{ji} x_{li}}{\mu_i^2} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \right\}
\end{aligned}
$$

Matrix representation

- Score function

$$U(\boldsymbol{\beta}) = \mathbf{X}^\top \mathbf{Q}\left[\mathbf{y} - \boldsymbol{\mu}\right]$$

- Expected Fisher information matrix

$$I(\boldsymbol{\beta}) = \mathbf{X}^\top \mathbf{Q}\mathbf{W}\mathbf{Q}\mathbf{X}$$

where

$$\mathbf{Q} = \begin{bmatrix} \frac{\nu}{\mu_1^2}\frac{\partial \mu_1}{\partial \eta_1} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{\nu}{\mu_n^2}\frac{\partial \mu_n}{\partial \eta_n} \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} \frac{\mu_1^2}{\nu} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{\mu_n^2}{\nu} \end{bmatrix}$$

If $\mu_i = -\dfrac{1}{\eta_i}$, then $\dfrac{\partial \mu_i}{\partial \eta_i} = \mu_i^2$

$\Rightarrow \quad \mathbf{Q} = \nu \mathbf{I}_n$

$\Rightarrow \quad U(\boldsymbol{\beta}) = \nu \mathbf{X}^\top \left[\mathbf{y} - \boldsymbol{\mu}\right]$

$\Rightarrow \quad I(\boldsymbol{\beta}) = \nu^2 \mathbf{X}^\top \mathbf{W}\mathbf{X}$

Maximum likelihood estimation of $\beta_0, \dots, \beta_p$

- Numerical optimisation techniques (Newton-Raphson/Fisher scoring)
- If the model is correctly specified (*adequate*), then the maximum likelihood estimators are asymptotically Gaussian, unbiased and efficient

$$\hat{\mathbf{B}} \xrightarrow{d} NMV_{p+1}\left(\boldsymbol{\beta}, I(\boldsymbol{\beta})^{-1}\right)$$

- $\hat{\mathrm{E}}[Y_i|\mathbf{x}_i] = \hat{m}_i = h\left(\mathbf{x}_i^\top \hat{\mathbf{b}}\right)$

Saturated model and deviance

Assuming that each unit is characterised by a different covariate pattern:

$$
\begin{aligned}
l(\mu_1, \ldots, \mu_n, \nu) &= \sum_{i=1}^{n} \left\{ -\frac{\nu}{\mu_i} y_i - \nu \ln \mu_i + (\nu - 1) \ln y_i - \ln \Gamma(\nu) + \nu \ln \nu \right\} \\
U_i(\boldsymbol{\mu}) &= \frac{\partial}{\partial \mu_i} l(\mu_1, \ldots, \mu_n, \nu) = \nu \frac{y_i - \mu_i}{\mu_i^2} \\
\hat{m}_i &= y_i \\
D &= 2 \left[l(y_1, \ldots, y_n, \nu) - l(\hat{\mathbf{b}}, \nu) \right] \\
&= 2\nu \sum_{i=1}^{n} \left[\frac{y_i - \hat{m}_i}{\hat{m}_i} - \ln \frac{y_i}{\hat{m}_i} \right] \cong \nu \sum_{i=1}^{n} \left[\frac{y_i - \hat{m}_i}{\hat{m}_i} \right]^2 \\
&\quad \text{where } \hat{m}_i = h\left(\mathbf{x}_i^\top \hat{\mathbf{b}} \right)
\end{aligned}
$$

NOTE THAT: due to the presence of a nuisance parameter, the residual deviance of a GLM with Gamma probabilistic component cannot be used as a goodness of fit test statistic.

Residuals and estimation of ν

- Deviance residuals $e_i^D = \text{sign}(y_i - \hat{m}_i) \sqrt{2\hat{\nu} \left[\frac{y_i - \hat{m}_i}{\hat{m}_i} - \ln \frac{y_i}{\hat{m}_i} \right]}$
- Pearson residuals $e_i^P = \sqrt{\hat{\nu}} \frac{(y_i - \hat{m}_i)}{\hat{m}_i}$
- $\hat{\nu} = \frac{n - p - 1}{\sum_{i=1}^{n} \left[\frac{y_i - \hat{m}_i}{\hat{m}_i} \right]^2}$ or, equivalently $\frac{1}{\hat{\nu}} = \frac{\sum_{i=1}^{n} \left[\frac{y_i - \hat{m}_i}{\hat{m}_i} \right]^2}{n - p - 1}$

If the model is correctly specified adn the true value of ν is large, it is possible to prove that the (standardised) residuals are asymptotically independent, homoscedastic and with a Gaussian distribution.

Testing linear hypotheses on $\beta_0, \ldots, \beta_p$

$H_0 : \mathbf{K}\boldsymbol{\beta} = \mathbf{t}$

- Likelihood ratio test statistic

$$-2 \ln \frac{l(\hat{\mathbf{b}}, \hat{\nu})}{l(\hat{\mathbf{b}}_{H_0}, \hat{\nu})} \Bigg| H_0 \xrightarrow{d} \chi^2_{(q)}$$

- Wald test statistic

$$\left[\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right]^\top \left[\mathbf{K}\left(\widehat{I(\boldsymbol{\beta})}\right)^{-1}\mathbf{K}^\top\right]^{-1} \left[\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right] \Bigg| H_0 \xrightarrow{d} \chi^2_{(q)}$$

If the model is correctly secified and the sample size is sufficiently large, χ^2 distributions (with properly defined degrees of freedom) can be used to approximate the rejection region for H_0/to approximate the $p-$value associated with the observed value of the test statistic.

An example 17

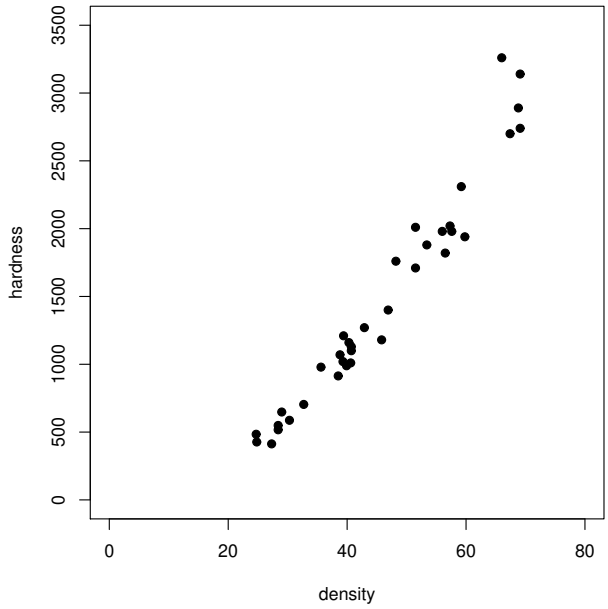
An example

`hardness`: hardness of an hardwood timber (Y)

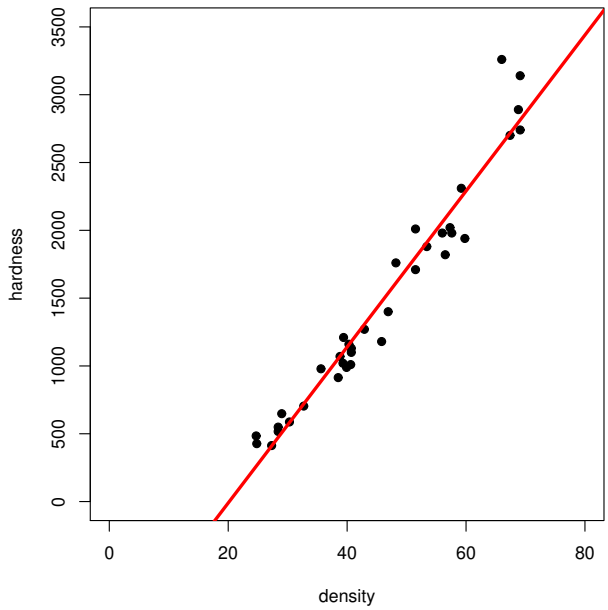
`density`: density of an hardwood timber (X)

$n = 36$

The data set

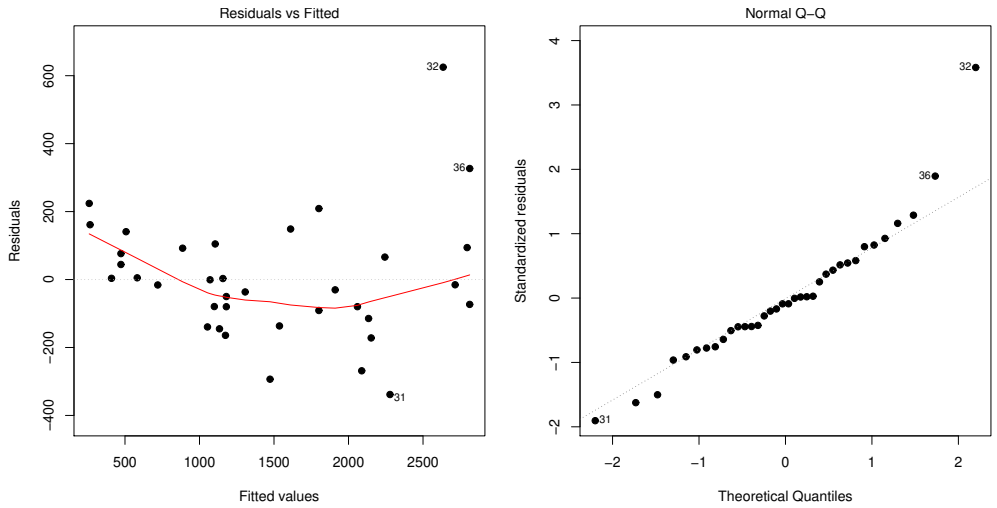


Gaussian linear model - parameter estimates



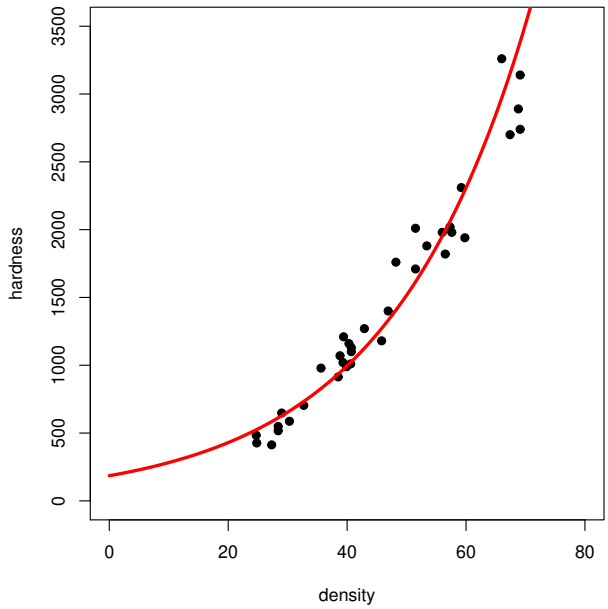
WARNING: hardness cannot take negative values

Gaussian linear model - graphical residual analysis



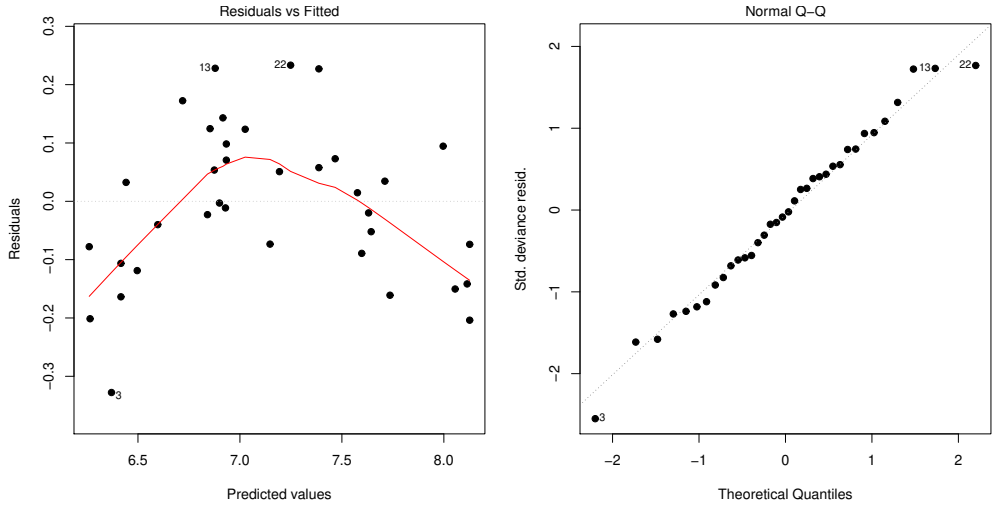
WARNING: possible violation of the omoschedasticity assumption

Gamma 1 - parameter estimates



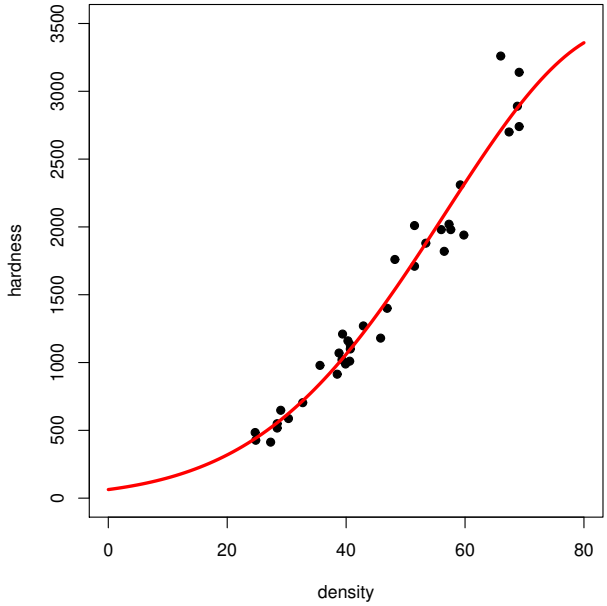
Systematic component: $g(\mu_i) = \ln \mu_i = \beta_0 + \beta_1 x_i$

Gamma GLM 1 - graphical residual analysis - $\hat{\nu} = 55.669$



WARNING: possible nonlinear contribution of the regressor to η_i (residuals have constant variability, but nonconstant moving average)

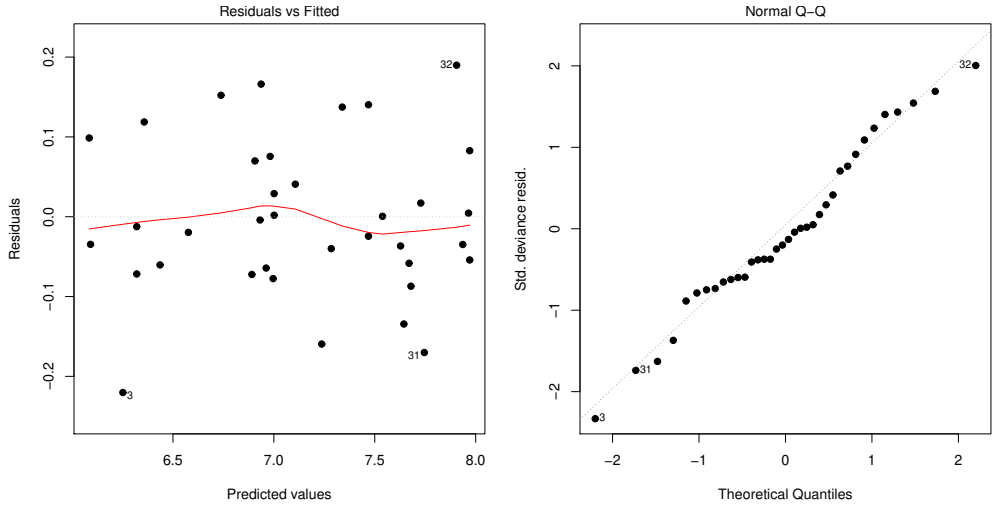
Gamma GLM 2 - parameter estimates



Systematic component: $g(\mu_i) = \ln \mu_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$

Note that the predictor is still linear in the parameters

GLM Gamma 2 - Graphical residual analysis - $\hat{\nu} = 98.242$



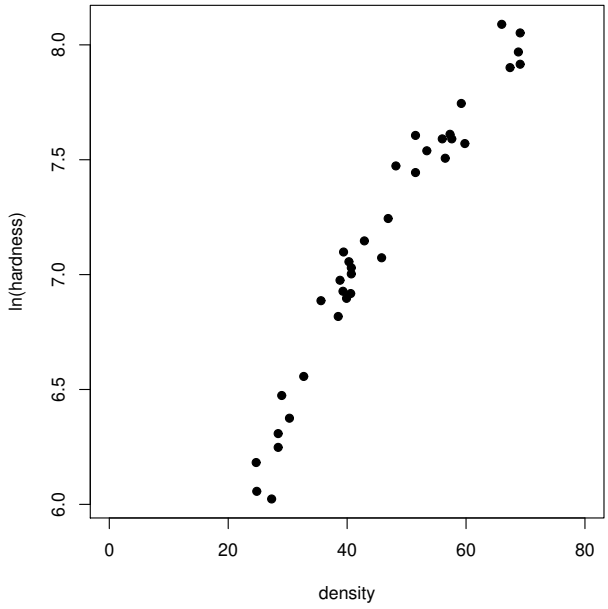
According to this plots, a model with a quadratic predictor, which is still linear in the parameter, combined with a logarithmic link function, seems adequate

Gamma GLM 2 - Summary output

```
Coefficients
              Estimate   Std.  Error   t value   Pr(>|t|)
   (Intercept)   4.1470         0.2089     19.85     0.0000
       density   0.0913         0.0093      9.80     0.0000
  I(density^2)  -0.0005         0.0001     -5.33     0.0000
```

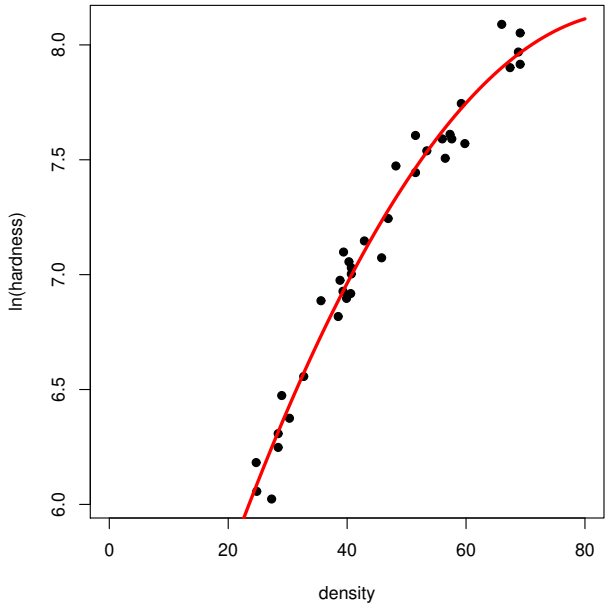
All regression coefficients are significantly larger than zero

An alternative strategy: transformation of Y



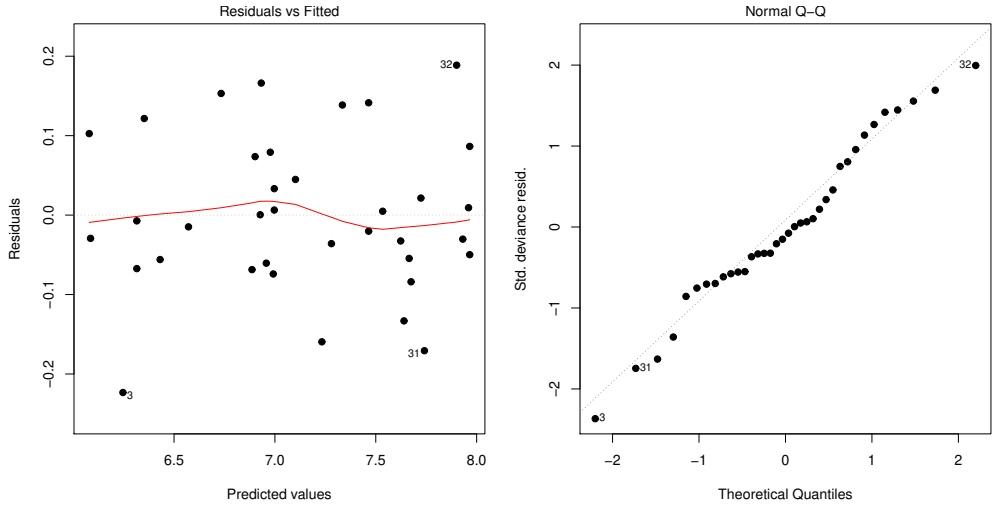
The log-transform of the dependent variable is considered

Lognormal regression model - parameter estimates



$\ln Y \sim N\left(\delta_0 + \delta_1 x_i + \delta_2 x_i^2, \sigma^2\right)$

Lognormal regression model - graphical residual analysis



According to this plots, a lognormal regression model with a quadratic component seems adequate

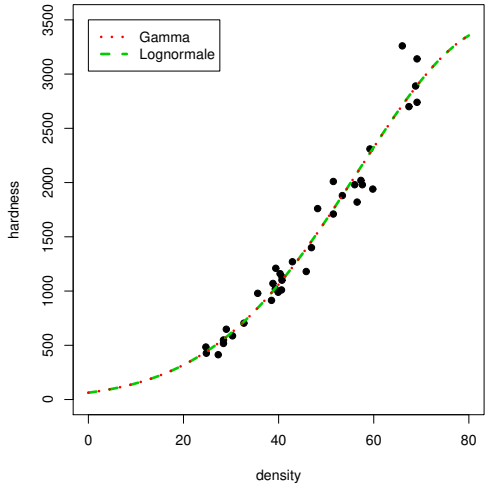
Comparison between Gamma and Lognormal random varaibles (1)

- The lognormal distribution does not belong to an exponential family expressed in canonical form

  $\Longrightarrow$ lognormal regression models do not belong to the GLM family to study the relationship between a dependent variable Y and a set of regressors
- lognormal regression models imply non-constant conditional variance, but constant conditional coefficient of variation

  $\Longrightarrow \mathrm{CV}[Y_i] = \sqrt{\exp(\sigma^2) - 1}$
- When the (conditional) coefficient of variation is lower than 0.8, the lognormal distribution and the Gamma distribution are very close
  In this example
  - $\hat{\nu} = 98.242 \Longrightarrow \widehat{\mathrm{CV}} = \frac{1}{\sqrt{98.242}} = 0.101$
  - $\hat{\sigma}^2 = 0.010 \Longrightarrow \widehat{\mathrm{CV}} = \sqrt{\exp(0.010) - 1} = 0.100$

Comparison between Gamma and Lognormal random varaibles (2)



The two models are almost equivalent. According to the AIC, the Gamma GLM seems slightly better than the lognormal regression model (455.6448 *vs* 455.7336*)*

Data transformation vs GLM (1)

Data transformation (in particular when applied to the dependent variable) is a practical solution to (simultaneously) address three issues:

- obtaining a new dependent variable whose range is comparable to the range of the linear predictor
- stabilising the conditional variance of the dependent variable
- obtaining a new dependent variable with a symmetric (or, even better, Gaussian) conditional distribution

The first issue is substantial and it is related to a proper definition of the effects of the regressors on the conditional expecetd value of the dependent variable

The second and third issue are mostly related to the efficiency of the inferential procedures

Data transformation and GLM (2)

When the interpretation of the results on the original scale of the dependent variable is important, transforming the dependent variable may introduce some difficulties

Nevertheless, data transformation can be useful, especially for explorative data analysis

Generalised linear models allow to:

- address the first issue through the choice of the systematic component
- deal with heteroschedastic and asymmetric conditional distribution through the choice of the probabilistic component

Within the GLM framework, the systematic component can be chosen (almost) regardless of the probabilistic components $\Longrightarrow$ *enhanced flexibility*

Enhancing the flexibility of GLMs via regularization and additive modelling: an introduction

GLMs based on P-splines

The P-spline approach can be extended beyond Gaussian models exploiting the GLM framework, by:

- replacing the linear predictor in the systematic component of a GLM with

$$g\left(\mathrm{E}\left[Y_i|x_i\right]\right) = \sum_{j=1}^{K+4} \theta_j b_j(x)$$

(note that this quantity is still linear in the unknown parameteres)

- exploiting penalized maximum likelihood estimation

  ⇒ general expressions for the penalized score function and the penalized expected Fisher information matrix can be easily obtained

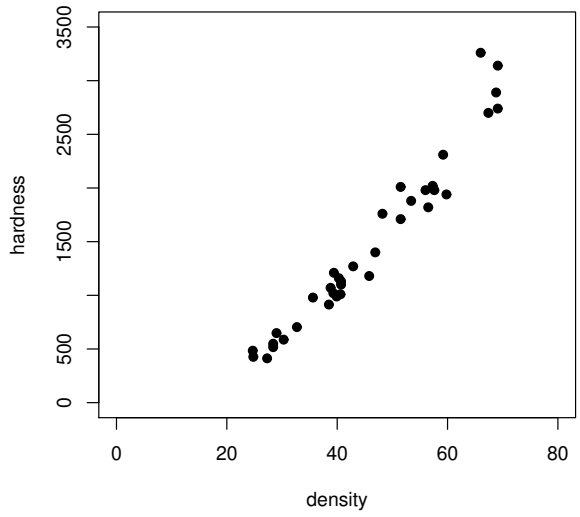
  ⇒ pseudo-Fisher scoring algorithms can be defined using the penalized score function and the penalized expected Fisher information matrix

  ⇒ the GCV criterion can be easily extended, starting from the residual deviance

  ⇒ hypothesis testing can be based on approximate asymptotic results

Formulas are omitted for the sake of simplicity

Timber data - continued



Probabilistic component: gamma distribution

Systematic component: $\ln \mu_i = \sum_{j=1}^{K+4} \theta_j b_j(x)$

Timber data - P-splines - R output (1)

gam function (package: mgcv) - $K = 20$, second-order differences penalty

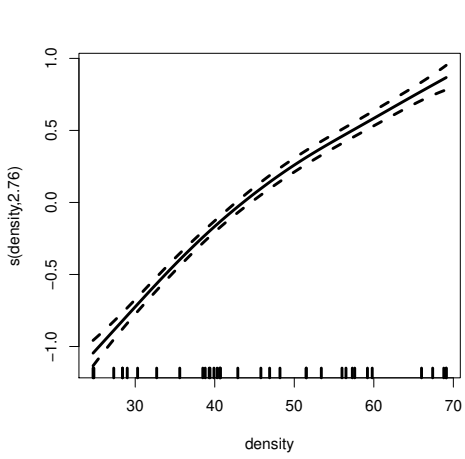
```
Parametric coefficients:
             Estimate   Std.  Error   t value   Pr(>|t|)
 (Intercept)     7.141          0.017   431.680      0.000

Approximate significance of smooth terms:
               edf   Ref.df         F   p-value
 s(density)  2.763    3.373   352.900     0.000
```

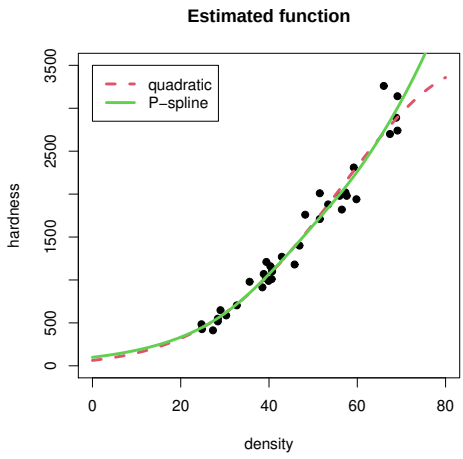
Systematic component	$\ln L$	n. parameters	AIC
quadratic function	-223.8224	3(+1)	455.6448
P-spline	-222.7507	3.7632(+1)	455.0279

⇒ Note that the effective number of parameters for the P-spline model are only slightly larger than 3 (the fitted function is approximately quadratic)

⇒ the model based on P-spline is only slighlty better than the model with a quadratic effect

Timber data - P-splines - R output (2)





Definition of a generalised additive model (GAM)

Y_i r.v. that describes the possible value of the dependent variable on the i-th sample unit ($i = 1, \ldots, n$)

$x_{1i}, \ldots, x_{pi}$ values of the regressors for the i-th sample unit

Generalized additive models (GAM) are statistical model for random samples $\boldsymbol{Y}$

characterised by:

⇒ **probabilistic components** similar to those of GLM

conditional distributions belonging to the same exponential family + conditional independence

⇒ a more flexible **systematic/deterministic component**

linear predictors are replaced with **additive predictors**

Systematic/deterministic component of a GAM

- Additive predictor:

  $\eta_i = \alpha_0 + s_1(x_{1i}) + s_2(x_{2i}) + \ldots + s_p(x_{pi})$

  $s_l(\cdot)$ (nonlinear) unknown smooth function depending only on the l-th regressor ($i = 1, \ldots, p$)

  ⇒ (additional) source of nonlinearity in the model

  ⇒ no assumption on the functional form of $s_l(\cdot)$ are required (nonparametric model)

- link function:

  $g(\cdot)$

  with known functional form, differentiable and invertible, such that $g^{-1}(\cdot) = h(\cdot)$

⇒ $g\left(\mathrm{E}\left[Y_i | x_{1i}, \ldots, x_{pi}\right]\right) = \alpha_0 + s_1(x_{1i}) + s_2(x_{2i}) + \ldots + s_p(x_{pi})$

or, equivalently

$\mathrm{E}\left[Y_i | x_{1i}, \ldots, x_{pi}\right] = h\left(\alpha_0 + s_1(x_{1i}) + s_2(x_{2i}) + \ldots + s_p(x_{pi})\right)$

Identifiability of additive predictors

Any additive predictor is identifiable up to constant shifts in the smooth functions $s_l(\cdot)$

Example: consider

$s_1^*(x_{1i}) = s_1(x_{1i}) + c$ and $s_2^*(x_{2i}) = s_1(x_{1i}) - c$

then

$$\begin{aligned} \eta_i^* &= \alpha_0 + s_1^*(x_{1i}) + s_2^*(x_{2i}) + \ldots + s_p(x_{pi}) \\ &= \alpha_0 + s_1(x_{1i}) + s_2(x_{2i}) + \ldots + s_p(x_{pi}) \\ &= \eta_i \end{aligned}$$

$\Rightarrow$ the following restrictions are usually introduced:

$$\sum_i s_l(x_{li}) = 0, \quad l = 1, \ldots, p$$

Penalized ML estimation for GAM (1)

A popular strategy to fit GAMs is based on an extension of the penalized ML approach described in the previous slides

$\Rightarrow$ each smooth function $s_l(\cdot)$ is approximated using a spline function and represented using a specific set of m_l basis

$$s_l(x_{lj}) = \sum_{j=1}^{m_l} \theta_{lj} b_{lj}(x_{li})$$

example: B-splines with K knots (with K large)

$\Rightarrow$ a specific penalty term with a matrix represenation is chosen for the parameters vectors $\boldsymbol{\theta}_l = (\theta_{l1}, \ldots, \theta_{lm_l})^\top$

$$J_l(\boldsymbol{\theta}_l) = \boldsymbol{\theta}_l^\top \mathbf{P}_l \boldsymbol{\theta}_l$$

example: first- or second-order differnces

$\Rightarrow$ in order to control smoothness, a specific smoothing parameter is allowed for each smooth function

$$\lambda_l \geq 0$$

Penalized ML estimation for GAM (2)

Penalized log-likelihood function

$$pl_{\boldsymbol{\lambda}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p) = \ln L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p) + \sum_{l=1}^{p} \lambda_l J_l(\boldsymbol{\theta}_l)$$

$\Rightarrow$ the actual form of $pl_{\boldsymbol{\lambda}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p)$ and the possible presence of nuisance parameters/weights depend on the specific probabilistic component

$\Rightarrow$ general expressions for the penalized score function and the penalized expected Fisher information matrix can be easily obtained

$\Rightarrow$ pseudo-Fisher scoring algorithms can be defined using the penalized score function and the penalized expected Fisher information matrix

$\Rightarrow$ the GCV criterion can be easily extended, starting from the residual deviance (*note that an optimal value for the smoothing parameter of each $s_l(\cdot)$ can be selected*)

$\Rightarrow$ hypothesis testing can be based on approximate asymptotic results (*although also in this context p-values are usually anticonservative*)

Formulas are omitted for the sake of simplicity

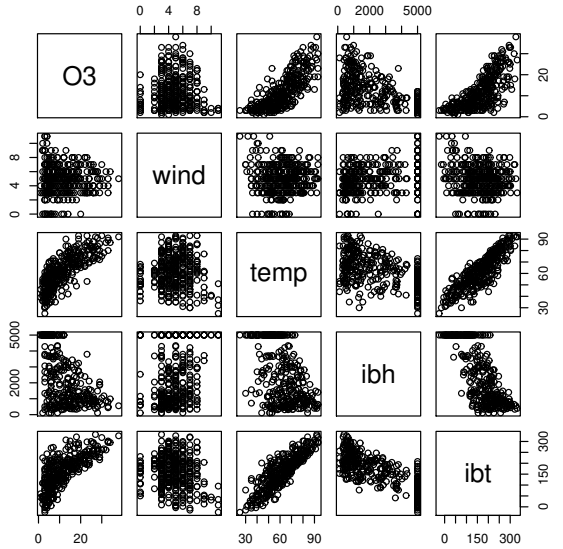
Ozone data (1)

In a study on air pollution in the Los Angels area, a researcher is interested in evaluating the effects of some regressors on the daily atmospheric ozone concentration O3 (ppm - *count variable*). In particular, the following regressors are considered:

- `temp`: temperature (degree F)
- `ibt`: inversion base height (feet)
- `ibh`: inversion base temperature (feet)

Data refer to 330 different days.

Ozone data (2)

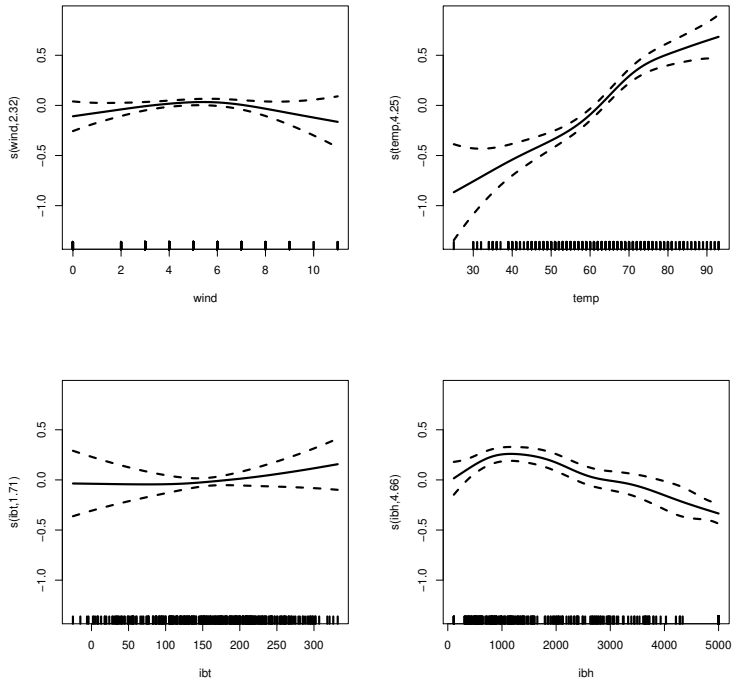


Ozone data - Poisson GAM - R output (1)

gam function (package: mgcv) - default settings

```
Parametric coefficients:
             Estimate  Std.  Error  t value  Pr(>|t|)
 (Intercept)     2.291         0.019  119.903     0.000

Approximate significance of smooth terms:
            edf  Ref.df         F  p-value
 s(wind)  2.324   2.972     5.060    0.145
 s(temp)  4.251   5.238   101.571    0.000
  s(ibt)  1.714   2.201     1.769    0.431
  s(ibh)  4.658   5.644    68.202    0.000
```

Ozone data - Poisson GAM - R output (2)



Ozone data - Comparison between GLM and GAM

Recall that linear functions are a special case of spline functions

⇒ *a linear predictor can be obtained by introducing suitable linear constraints on the parameters of the linear basis expansions associated to a given additive predictor*

```
Model 1:  O3 ~ wind + temp + ibt + ibh
Model 2:  O3 ~ s(wind) + s(temp) + s(ibt) + s(ibh)

     Resid. Df   Resid. Dev        Df   Deviance   Pr(>Chi)
 1     325.0000      519.0159
 2     312.9443      456.3987   12.0557    62.6172     0.0000
```

⇒ *at least one of the regressors included in the GAM model has a significantly nonlinear effect on the dependent variable*

⇒ *the GAM model is also better in terms of* AIC

Predictor	$\ln L$	n. parameters	AIC
linear	-931.8987	5	1873.797
additive	-900.5901	13.947	1829.075

Further extensions

Several extensions of GAM models are possible:

- semiparametric models can be defined, by defining systematic components in which only some of the regressors have a nonlinear smooth effect:

  $\eta_i = \alpha_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + s_3(x_{3i}) \ldots + s_p(x_{pi})$

  This, for example, allows the inclusion of categorical regressors in GAM models

- the additivity assumption can be overcome by allowing interactions, which can also be represented using smooth multivariate functions depending on more than one regressor

  $\eta_i = \alpha_0 + s_{12}(x_{1i}, x_{2i}) + s_3(x_{3i}) \ldots + s_p(x_{pi})$