

# Lecture 1 - Part 1: Generalized (Non)-Linear Models

GIANCARLO CAMARDA

Institut national d'études démographiques



UGOFILIPPO BASELLINI

Max Planck Institute for Demographic Research



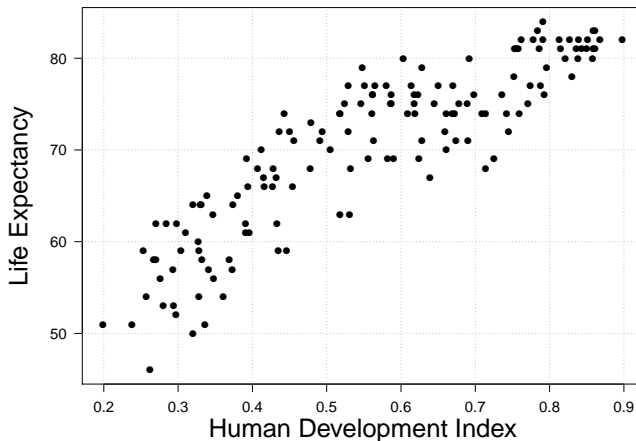
IDEM 117

*Advances in Mortality Forecasting*

*International Advanced Studies in Demography*

17 – 21 January, 2022

# A simple example: $e_0$ and HDI



Life expectancy (both sexes) vs. Human Development Index in 2012.  
Source: World Bank and World Health Organization.

# A (linear) model for the example

- It seems reasonable to assume that the more “developed” a country is, the higher life expectancy would be
- For each country  $i$  we have:

$y_i$  : Life expectancy

$x_i$  : Human Development Index

- A possible model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{with} \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- Or in matrix notation:

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{where} \quad \mathbf{X} = [1 : \mathbf{x}]$$

# Estimating the model

- To estimate  $\beta$  we maximise the log-likelihood which is equivalent to minimizing the residual sum of squares:

$$RSS(\beta) = \sum_i [y_i - X \beta]^2 = (\mathbf{y} - X\beta)' (\mathbf{y} - X\beta)$$

- Taking the derivatives of  $RSS(\beta)$  with respect to  $\beta$  and setting equal to zero, we obtain

$$X'X \beta = X'y \quad \Rightarrow \quad \hat{\beta} = (X'X)^{-1} X'y$$

# Graphical illustration

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$\beta' = [ \beta_0 \quad \beta_1 ]$$

$$\mu = X\beta$$

$$\hat{\beta} = (X'X)^{-1}X'y$$

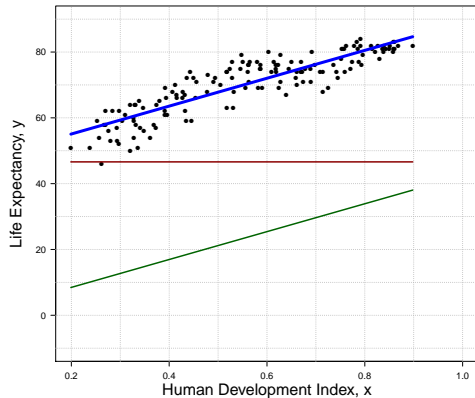
# Graphical illustration

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$\beta' = \begin{bmatrix} \beta_0 & \beta_1 \end{bmatrix}$$

$$\mu = X\beta$$

$$\hat{\beta} = (X'X)^{-1}X'y$$



$$\hat{\beta}' = \begin{bmatrix} 46.61 & 42.36 \end{bmatrix}$$

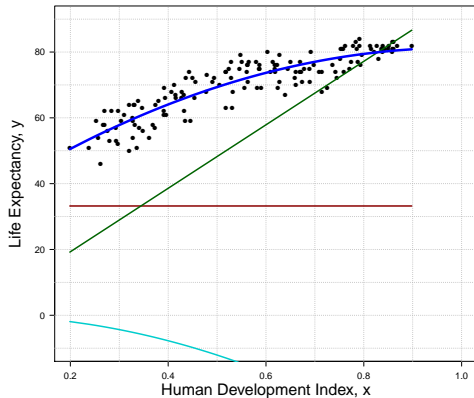
# Graphical illustration

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}$$

$$\beta' = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 \end{bmatrix}$$

$$\mu = X\beta$$

$$\hat{\beta} = (X'X)^{-1}X'y$$



$$\hat{\beta}' = \begin{bmatrix} 33.22 & 96.47 & -48.38 \end{bmatrix}$$

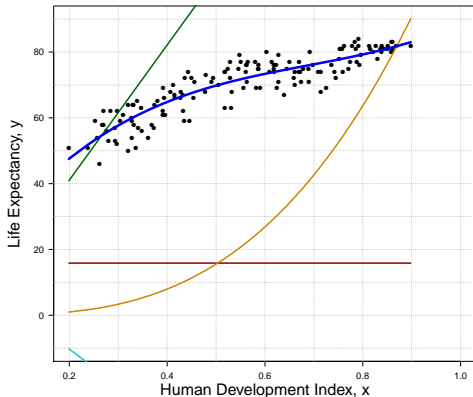
# Graphical illustration

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 \end{bmatrix}$$

$$\beta' = [\beta_0 \quad \beta_1 \quad \beta_2 \quad \beta_3]$$

$$\mu = X\beta$$

$$\hat{\beta} = (X'X)^{-1}X'y$$



$$\hat{\beta}' = [15.86 \quad 205.37 \quad -257.15 \quad 124.39]$$



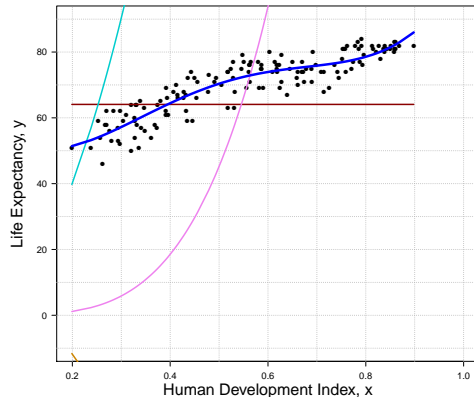
# Graphical illustration

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & x_1^4 \\ 1 & x_2 & x_2^2 & x_2^3 & x_2^4 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 & x_n^4 \end{bmatrix}$$

$$\beta' = [\beta_0 \quad \beta_1 \quad \beta_2 \quad \beta_3 \quad \beta_4]$$

$$\mu = X\beta$$

$$\hat{\beta} = (X'X)^{-1}X'y$$



$$\hat{\beta}' = [64.08 \quad -210.69 \quad 1003.47 \quad -1475.64 \quad 723.46]$$

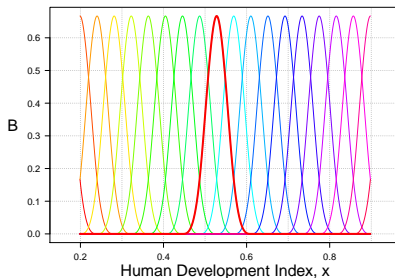
# Can we do better?

- Simple basis is good for simple example
- Basis function (powers of  $x$ ) are global
- Moving one end moves the other end too
- Unexpected wiggles
- The higher the degree the more is sensitive
- We seek for local basis
- Useful for more complex data
- No assumptions on the trend (let the data speak by themselves!)
- Smooth outcomes

# Introducing $B$ -splines

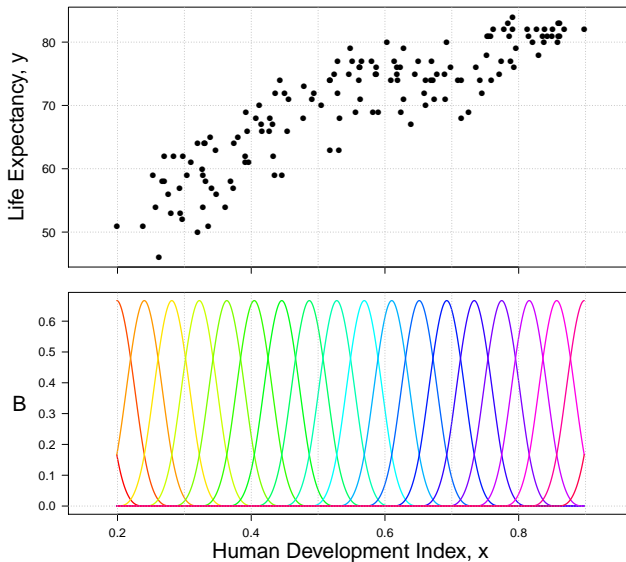
- Create a suitable basis  $\Rightarrow$  (equidistant)  $B$ -splines:

$$B = \begin{bmatrix} B_1(x_1) & B_2(x_1) & \dots & B_r(x_1) & \dots & B_k(x_1) \\ B_1(x_2) & B_2(x_2) & \dots & B_r(x_2) & \dots & B_k(x_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ B_1(x_n) & B_2(x_n) & \dots & B_r(x_n) & \dots & B_k(x_n) \end{bmatrix}$$

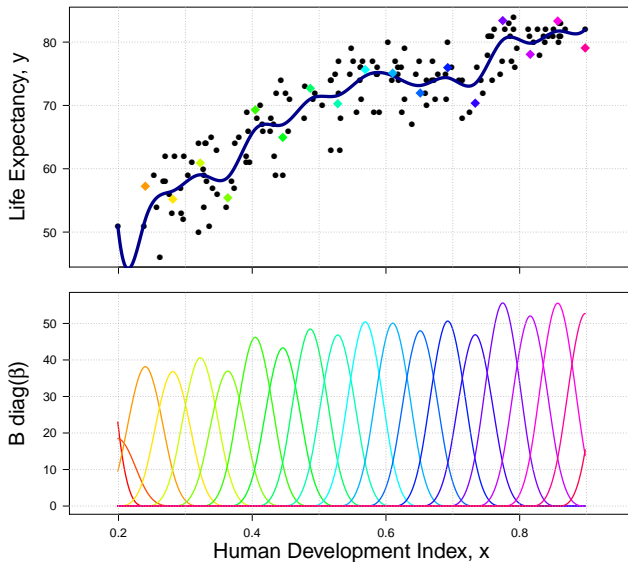


$$\bullet \quad E(y) = \mu = B\beta \quad \Rightarrow \quad \hat{\beta} = (B'B)^{-1}B'y$$

# Fitting with $k = 20$ B-splines



# Fitting with $k = 20$ B-splines



# In the pursuit of smoothness

- Outcomes are not smooth, we could:
  - take less  $B$ -splines
  - place each  $B$ -splines in specific positions
  - set a double goal:
    - 1 good fit to the data, i.e. low least-squares:  $S = |\mathbf{y} - \mathbf{B}\boldsymbol{\beta}|^2$
    - 2 smooth curve, i.e. low roughness:  $R$
- How to measure roughness? By summing up squared differences of  $\boldsymbol{\beta}$
- Simplest case: first differences ( $k = 4$ )

$$R = (\beta_2 - \beta_1)^2 + (\beta_3 - \beta_2)^2 + (\beta_4 - \beta_3)^2$$

- In matrix notation:

$$R = |\mathbf{D}_1\boldsymbol{\beta}|^2 \quad \text{with} \quad \mathbf{D}_1 = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

- Second order differences:

$$R = |\mathbf{D}_2\boldsymbol{\beta}|^2 \quad \text{with} \quad \mathbf{D}_2 = \begin{bmatrix} 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \end{bmatrix}$$

# Penalizing the coefficients: *P*-splines

- In general  $D_d$  is a  $(k - d) \times k$  matrix with  $d$  order of difference
- We balance the two object-functions:

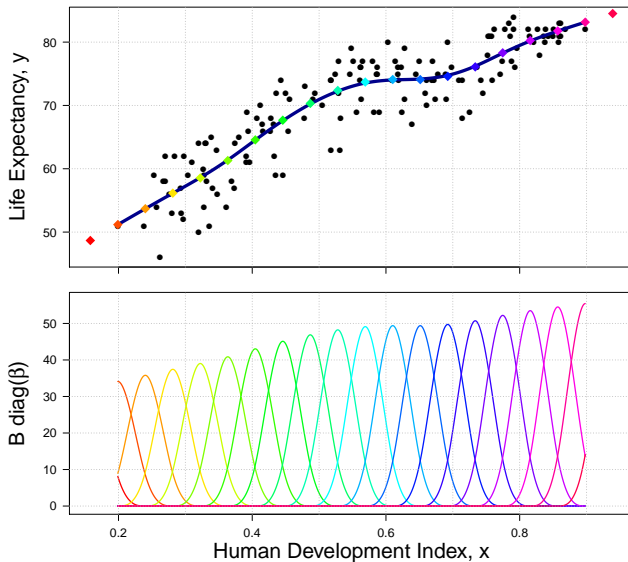
$$S^* = S + \lambda R = |\mathbf{y} - \mathbf{B}\boldsymbol{\beta}|^2 + \lambda |\mathbf{D}\boldsymbol{\beta}|^2$$

- Given a  $\lambda$ , this is again a linear system of equation with explicit solution:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{B}'\mathbf{B} + \lambda \mathbf{D}'\mathbf{D})^{-1} \mathbf{B}'\mathbf{y} \\ &= (\mathbf{B}'\mathbf{B} + \mathbf{P})^{-1} \mathbf{B}'\mathbf{y}\end{aligned}$$

- Higher  $\lambda \Rightarrow R$  more important than  $S \Rightarrow$  smoother  $\boldsymbol{\beta} \Rightarrow$  smoother  $\boldsymbol{\mu}$
- $\lambda = 0 \Rightarrow$  simple least-squares with *B*-splines
- $\lambda \rightarrow \infty \Rightarrow \boldsymbol{\mu}$  a polynomial of degree  $d$

# Fitting with *P*-splines



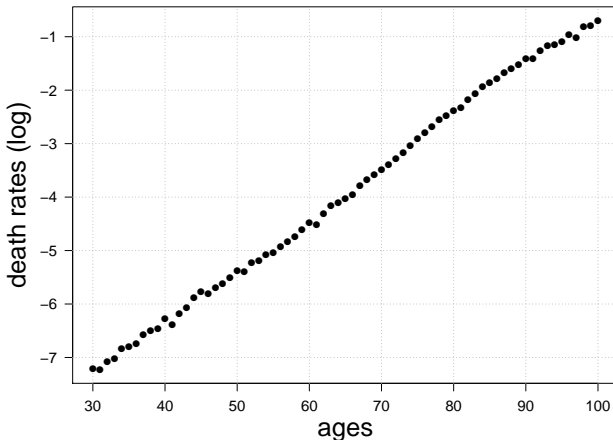


# Fitting with $P$ -splines

# More about *P*-splines

- We won't see that:
  - number of *B*-splines is immaterial
  - the degree of *B*-splines is irrelevant
  - there are objective criteria for selecting  $\lambda$
- In the following (lectures), we will see how
  - to interpolate/extrapolate
  - the order of difference is crucial for forecasting
  - to generalize *P*-splines for non-Normal data
  - to integrate prior knowledge
  - higher-dimensional generalization is straightforward
- If you want to know more:  
Eilers & Marx (2021).  
*Practical Smoothing. The Joy of P-splines*  
Cambridge University Press.

# A simple non-normal example: mortality



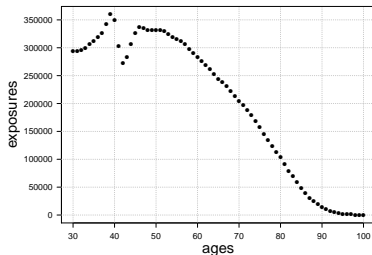
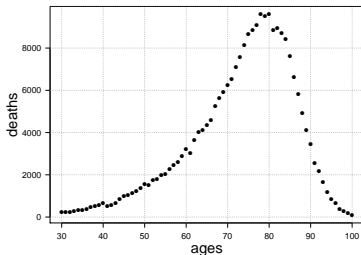
Death rates (in log scale) over age. England & Wales, females, 1960, ages 30-100.  
Source: Human Mortality Database.

# A simple non-normal example: mortality

- Death rates are *by product* of death counts and exposures
- Real random variables are deaths  $\Rightarrow$  model deaths
- Data:

$y = (y_i)$  : death counts over age  $x_i$

$e = (e_i)$  : exposure population over age  $x_i$



Deaths and exposures over age. England & Wales, females, 1960, ages 30-100.

Source: Human Mortality Database.

# Assumptions in mortality setting

- Observed deaths are realization of a Poisson distribution:

$$y_i \sim \mathcal{P}(e_i \mu_i)$$

- The expected values are the product of
  - some given measure of exposure ( $e_i$ ). Commonly called *offset*
  - force of mortality ( $\mu_i$ )
- In a regression setting,  $\mu_i$  depends on a vector of explanatory variables
- Here, we have no external covariates  $\Rightarrow \mu_i$  depends on age and/or time
- In a Poisson regression, we model the logarithm of the force of mortality

$$\ln \mu_i = \eta_i$$

- In a Generalized **Linear** Model (GLM):

$$\eta = X\beta$$

commonly called *linear predictor*

# A simple example: our friend Gompertz

- For demographers:  $\mu(x) = ae^{bx}$
- Equivalent to a Poisson GLM model:

$$d_i \sim \mathcal{P}(e_i \mu_i) \quad \ln \mu = \eta = X\beta$$

with

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \beta' = [ \ln(a) \quad b ]$$

# Estimating a Poisson GLM

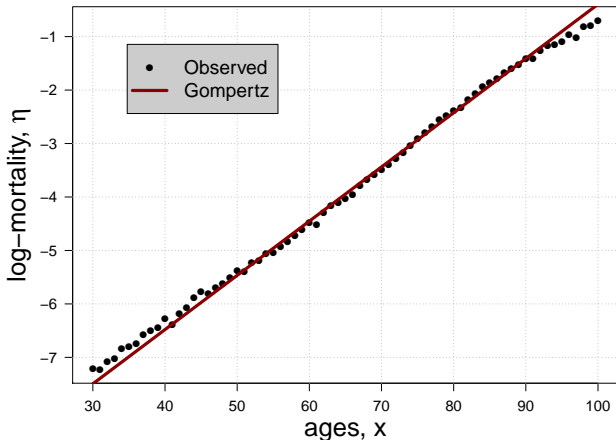
- Unlike in Linear model, no an explicit solution for  $\beta$
- The *iteratively reweighted least-squares* is used (common for all GLMs)
- For Poisson data with offset, in mortality setting:
  - 1 Start with trial estimate of  $\eta_i^{(0)}$ . Commonly  $\eta_i^{(0)} = \ln(\frac{y_i+1}{e_i+1})$
  - 2 Evaluate the current force of mortality:  $\mu_i = \exp(\eta_i)$
  - 3 Compute the so-called working response:  $z_i = \eta_i + \frac{y_i - e_i \mu_i}{e_i \mu_i}$
  - 4 Compute the iterative weights:  $w_i = e_i \mu_i$
  - 5 Update  $\beta$  by regressing the working response on the covariates in  $X$  using the weights  $w_i$ :

$$\tilde{\beta} = (X'WX)^{-1}X'Wz$$

where  $W$  is the diagonal matrix with entries  $w_i$

- 6 Update the linear predictor:  $\eta = X\tilde{\beta}$
- 7 Repeat 2 to 6 until convergence, e.g. small difference in subsequent  $\eta$

# A Gompertz fit as Poisson GLM

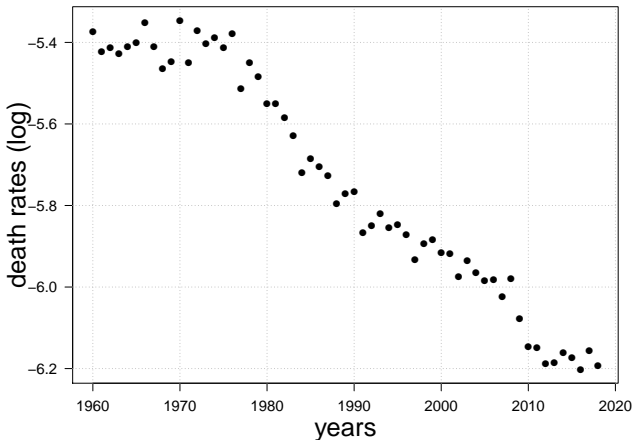


Actual and estimated death rates (in log scale) using a Gompertz over age.  
England & Wales, females, 1960, ages 30-100.

Source: Human Mortality Database.



# A slightly more complex example



Death rates (in log scale) over years.  
England & Wales, females, age 50, years 1960-2018.  
Source: Human Mortality Database.

# A slightly more complex example

- We are still in a Poisson framework
- We cannot assume linearity of the log-mortality  $\Rightarrow$  we assume smoothness
- Likewise in the linear model, we use P-splines
- Generalization is achieved straightforwardly:

Underlying distribution		$y \sim \mathcal{N}(\mu, \sigma^2)$	$y \sim \mathcal{P}(e\mu)$
Linear	Model	$\mu = X\beta$	$\ln \mu = \eta = X\beta$
	Estimation	$\hat{\beta} = (X'X)^{-1}X'y$	$\tilde{\beta} = (X'WX)^{-1}X'Wz$
Non-linear	Model	$\mu = B\beta$	$\ln \mu = \eta = B\beta$
	Estimation	$\hat{\beta} = (B'B + P)^{-1}B'y$	$\tilde{\beta} = (B'WB + P)^{-1}B'Wz$



# $P$ -splines on mortality over years

# A great advantage: *P*-splines = a penalized GLM

- What we know for GLMs can be used for *P*-splines
  - Deviance residuals

$$r_i = \text{sign}(y_i - e_i \hat{\mu}_i) \sqrt{2 \left[ y_i \ln \left( \frac{y_i}{e_i \hat{\mu}_i} \right) - y_i + e_i \hat{\mu}_i \right]}$$

- Model complexity from the hat-matrix, here called effective dimension

$$\begin{aligned} ED &= \text{tr}(\mathbf{H}) \quad \text{with} \\ \mathbf{H} &= \mathbf{W}^{\frac{1}{2}} \mathbf{B} (\mathbf{B}' \mathbf{W} \mathbf{B} + \mathbf{P})^{-1} \mathbf{B}' \mathbf{W}^{\frac{1}{2}} \end{aligned}$$

- Variance-covariance matrix:

$$\mathbf{V} = (\mathbf{B}' \mathbf{W} \mathbf{B} + \mathbf{P})^{-1}$$

from which we could compute standard errors:

$$\begin{aligned} se(\hat{\beta}) &= \sqrt{\text{diag}(\mathbf{V})} \\ se(\hat{\eta}) &= \sqrt{\text{diag}(\mathbf{B} \mathbf{V} \mathbf{B}')} \end{aligned}$$