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

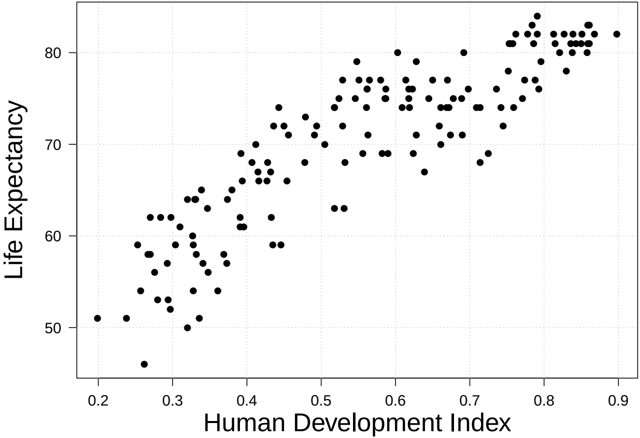
Giancarlo Camarda
Institut national d'études démographiques

Ugofilippo Basellini
Max Planck Institute for Demographic Research

IDEM 117
Advances in Mortality Forecasting

International Advanced Studies in Demography
28 June - 02 July, 2021

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:

$$\begin{aligned} y_i &: \text{Life expectancy} \\ x_i &: \text{Human Development Index} \end{aligned}$$

- A possible model:

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

- Or in matrix notation:

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

Estimating the model

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

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

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

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

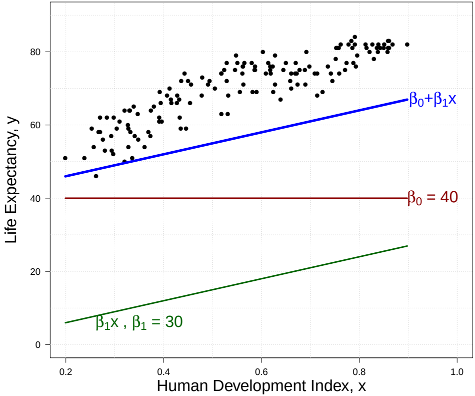
Graphical illustration

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

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

$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$$

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



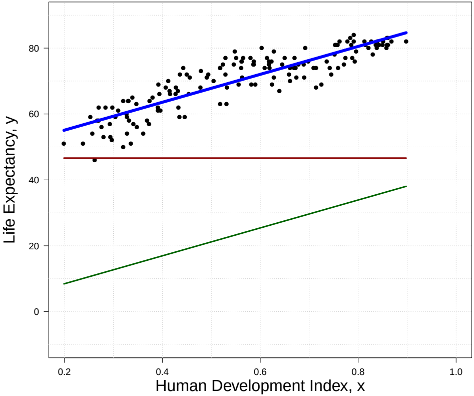
Graphical illustration

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

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

$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$$

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



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

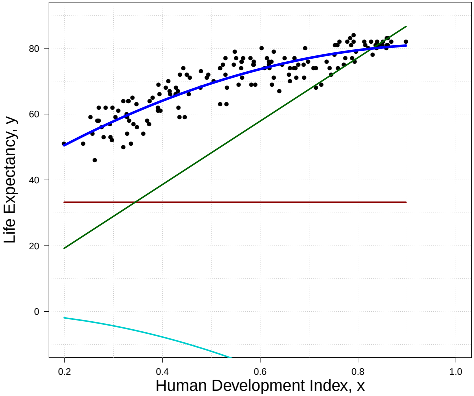
Graphical illustration

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

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

$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$$

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



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

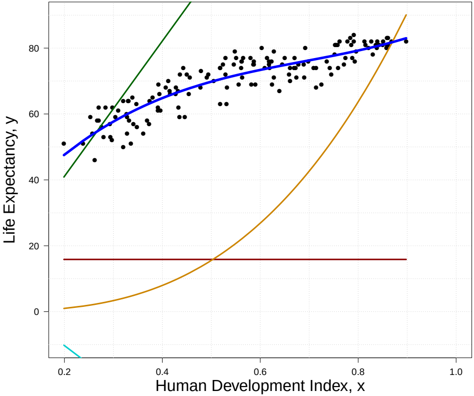
Graphical illustration

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

$$\boldsymbol{\beta}' = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & \beta_3 \end{bmatrix}$$

$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$$

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



$$\hat{\boldsymbol{\beta}}' = \begin{bmatrix} 15.86 & 205.37 & -257.15 & 124.39 \end{bmatrix}$$

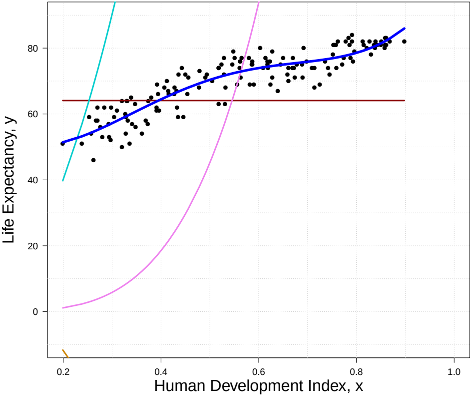
Graphical illustration

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

$$\boldsymbol{\beta}' = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 \end{bmatrix}$$

$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$$

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



$$\hat{\boldsymbol{\beta}}' = \begin{bmatrix} 64.08 & -210.69 & 1003.47 & -1475.64 & 723.46 \end{bmatrix}$$

Can we do better?

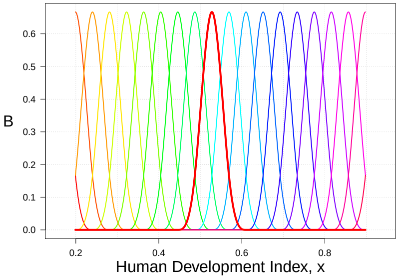
- Simple basis is good for simple example
- Basis function (powers of x) are global
- Moving one end moves the other end too
- Unexpected wiggles
- The higher the degree the more is sensitive

- 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

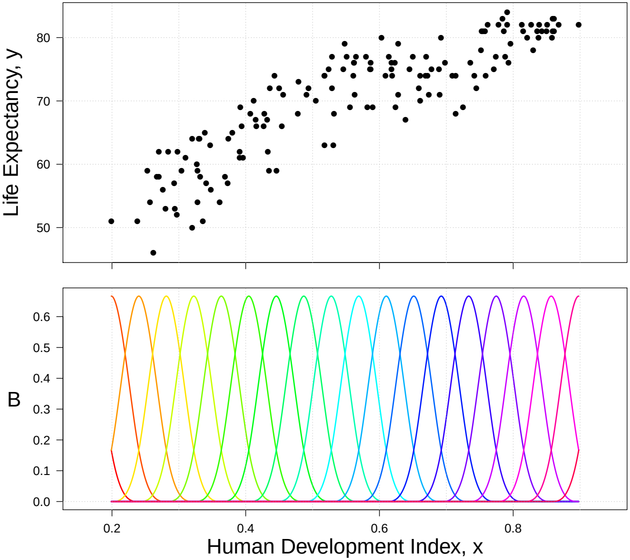
- Create a suitable basis $\Rightarrow$ (equidistant) B-splines:

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

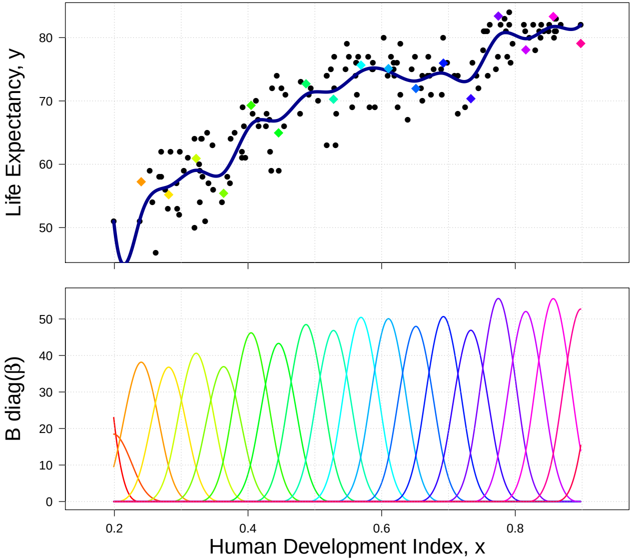


- $E(\boldsymbol{y}) = \boldsymbol{\mu} = \boldsymbol{B}\,\boldsymbol{\beta} \quad \Rightarrow \quad \hat{\boldsymbol{\beta}} = (\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{y}$

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: $S = |\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\beta}|^2$
  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:

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

- Second order differences:

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

Penalizing the coefficients: P-splines

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

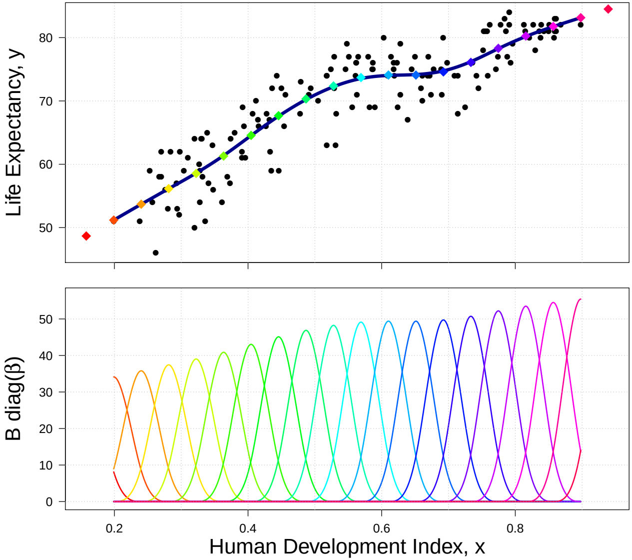
$$S^* = S + \lambda R = |\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\beta}|^2 + \lambda|\boldsymbol{D}\boldsymbol{\beta}|^2$$

- Given a λ, this is again a linear system of equation with explicit solution:

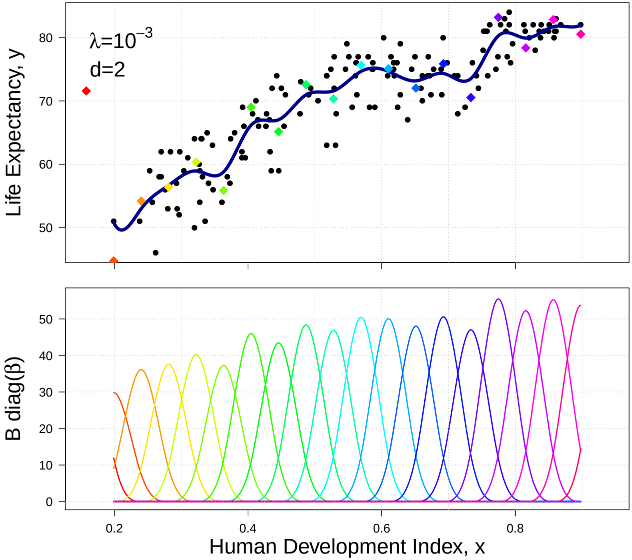
$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{B}'\boldsymbol{y} \\ &= (\boldsymbol{B}'\boldsymbol{B} + \boldsymbol{P})^{-1}\boldsymbol{B}'\boldsymbol{y} \end{aligned}$$

- 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
- $\lambda \to \infty \Rightarrow \boldsymbol{\mu}$ a polynomial of degree d

Fitting with P-splines



Fitting with P-splines

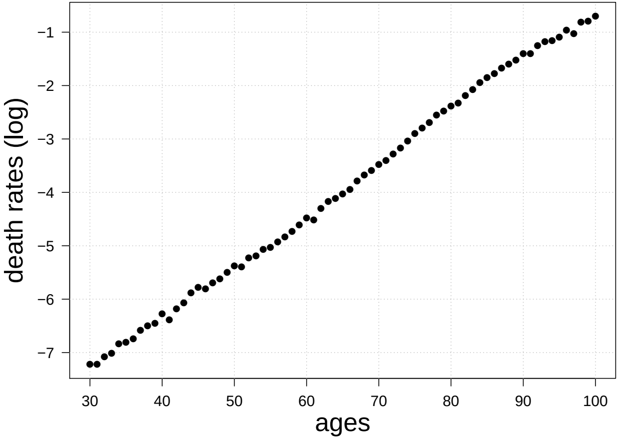


More about P-splines

- We won't see that:
  - number of B-splines is immaterial
  - the degree of B-splines is irrelevant
  - there are objective criteria for selecting λ
- 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



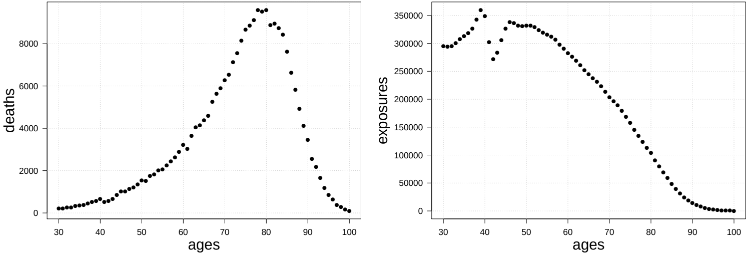
Death rates (in log scale) over age. England & Wales, females, 1960, ages 30-100.
Source: Human Mortality Database.

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:

$$\boldsymbol{y} = (y_i) \quad : \quad \text{death counts over age } x_i$$

$$\boldsymbol{e} = (e_i) \quad : \quad \text{exposure population over age } x_i$$



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 (μ_i)
- In a regression setting, μ_i depends on a vector of explanatory variables
- Here, we have no external covariates $\Rightarrow$ μ_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):

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$$

commonly called *linear predictor*

A simple example: our friend Gompertz

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

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

with

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

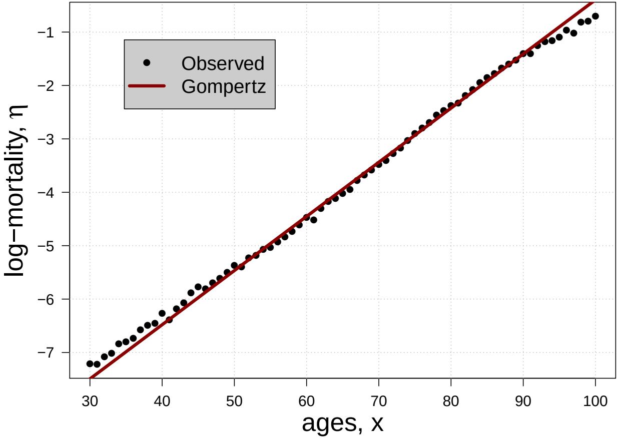
Estimating a Poisson GLM

- Unlike in Linear model, no an explict solution for $\boldsymbol{\beta}$
- 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 $\boldsymbol{\beta}$ by regressing the working response on the covariates in $\boldsymbol{X}$ using the weights w_i:

     $$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{z}$$

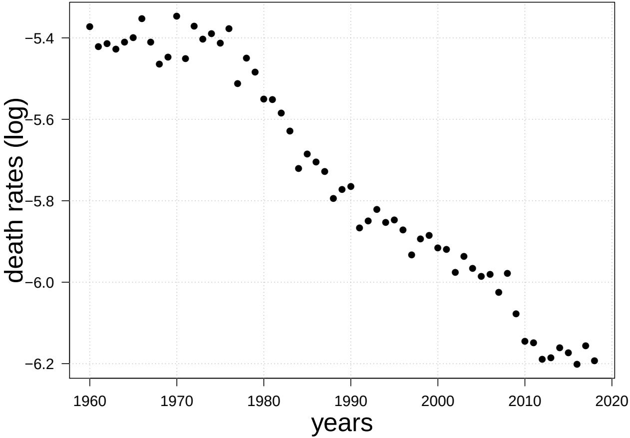
     where $\boldsymbol{W}$ is the diagonal matrix with entries w_i
  6. Update the linear predictor: $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$
  7. Repeat 2 to 6 until convergence, e.g. small difference in subsequent $\boldsymbol{\eta}$

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



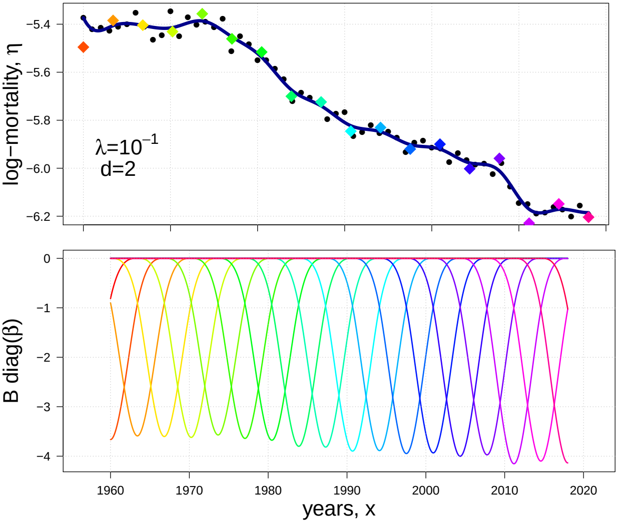
Death rates (in log scale) over years.
England & Wales, females, age 50, years 1960-2018.
Source: Human Mortality Database.

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

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

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

  - Variance-covariance matrix:

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

  from which we could compute standard errors:

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