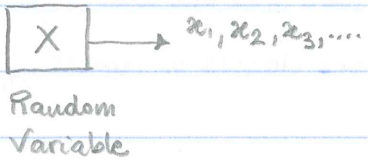
LECTURE 4

So far, we have considered the following problem:



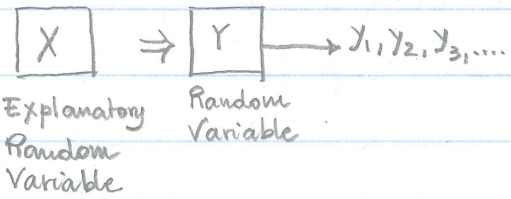
$$? \hat{\theta} : X \sim f_X(x|\hat{\theta})$$

(estimation problem)

We have approached this problem by using ML estimation, i.e., $\hat{\theta} = \arg\max_{\theta} \ell(\theta)$ with: $\ell(\theta) \triangleq \frac{1}{n} \sum_{i=1}^{n} \log\left(f_X(x_i|\theta)\right)$

Let us now consider a more complicated version of this problem:

A)



$$? \hat{\theta} : Y_i \sim f_{Y_i}(y, \hat{\eta}_i)$$
$$\hat{\eta}_i = g(x_i|\hat{\theta})$$

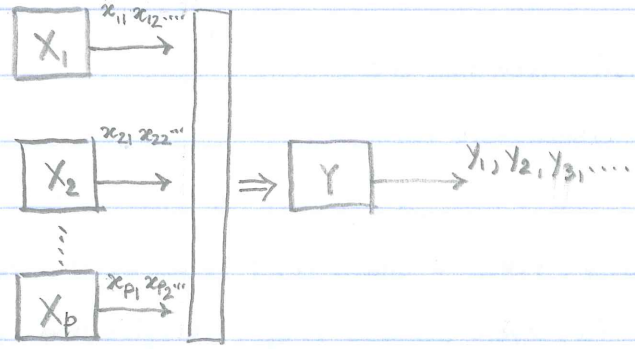
(regression problem)

$x_i \triangleq$ sample of X

In this problem, we assume that the vector of parameters η_i is a function of the value of RV X $\Rightarrow$ We must estimate the parameter vector $\hat{\theta}$ that uniquely defines the function between x_i and $\eta_i \forall i$.

The problem can be also posed as a multivariable problem:

B)



$$? \hat{\theta} : Y_i \sim f_{Y_i}(y|\hat{\eta}_i)$$
$$\hat{\eta}_i = g(x_{i1}, x_{i2}, \ldots, x_{ip}|\hat{\theta})$$

$x_{ij} \triangleq$ sample of X_j $j = 1, 2, \ldots, p$

(multiple regression problem)

Both in problem A) and B), functions $f_{Y_i}(\cdot)$ and $g(\cdot)$ can be linear or nonlinear, and determine the existence of one solution $\hat{\vartheta}$, more than one solution, or no solution.

$\Rightarrow$ To solve these problems, we must provide the classes of models $f_{Y_i}(\cdot)$ and $g(\cdot)$

EX: Let us assume: $Y_i \sim N(\mu_i, \sigma^2)$ with μ_i - unknown, σ^2 - known $\Rightarrow$ We set: $\eta_i \triangleq \mu_i$:

$$\Rightarrow f_{Y_i}(y \mid \eta_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\eta_i)^2}{2\sigma^2}}$$

Let us assume: $\eta_i \triangleq \beta_0 + \beta_1 x_i$, where x_i is a sample of the explanatory RV X, i.e.: $\eta_i = g(x_i \mid \vartheta)$, where $\vartheta = [\beta_0 \ \beta_1]^T$.

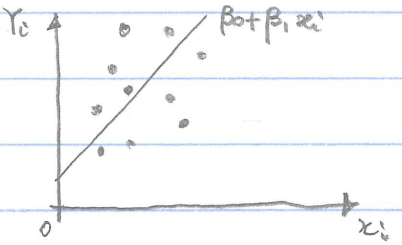
Note: The formulation of f_{Y_i} and g, and the definition of η_i and ϑ imply that we have:

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\mu_i} + \varepsilon_i$$

Y_i: RV of the i-th evaluation; ε_i: RV $\sim N(0, \sigma^2)$

$\Rightarrow$ Linear regression is a special case of "regression problem" that is obtained when:

* $f_{Y_i}(\cdot)$ is Gaussian with fixed variance
* $g(\cdot)$ is linear in the explanatory variables



As a result, Linear regression requires that: (i) Y varies unbounded as the sample x of X changes, (ii) every sample y_i of Y is extracted with the same variance (i.e., $\text{var}(\varepsilon_i) = \sigma^2 \ \forall i$), and (iii) the responses Y_i are all normal

⇒ We want to relax requirements (i)-(iii). In particular, let us assume:

- Y_i is bounded (e.g., discrete variable between 0 and n_i)
- variance of Y_i changes with $i=1,2,3,\ldots$
- Y_i is not Gaussian

} ⇒ For instance: $Y_i \sim B(n_i, p_i)$

Where: $n_i \triangleq$ number of trials used to estimate Y_i

$p_i \triangleq \underset{Y_i}{E}\left(\frac{Y_i}{n_i}\right)$ - expected value of count Y_i over n_i trials

In this case, we need to identify "η_i", "$g(\cdot)$", and "ϑ". Note this:

$0 \le p_i \le 1 \ \forall i \Rightarrow \frac{p_i}{1-p_i} \in [0, +\infty) \Rightarrow \underbrace{\log\left(\frac{p_i}{1-p_i}\right)}_{\triangleq \text{Logit}(p_i)} \in (-\infty, +\infty)$, i.e., Logit could vary unbounded as a function of an explanatory variable x_i ⇒ Hence, consider the following case:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i \Rightarrow p_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \Rightarrow \text{We can assume:}$$

$$\left.\begin{array}{l} \eta_i \triangleq p_i \ ; \ \vartheta \triangleq [\beta_0 \ \beta_1]^T \\ g(x_i|\vartheta) \triangleq \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \end{array}\right\} \Rightarrow \boxed{\begin{array}{l} Y_i \sim f_{Y_i}(y|\eta_i) \\ \eta_i = g(x_i|\vartheta) \end{array}}$$

regression problem

where: $f_{Y_i}(y|\eta_i) \underset{\text{definition of } B(n_i,p_i)}{\triangleq} P(Y_i = y) = \binom{n_i}{y} \eta_i^y (1-\eta_i)^{n_i-y}$

Interestingly, the problem can also be formulated in the compact form:

$$Y_i \sim f_{Y_i}(y|\eta_i) \qquad (*)$$
$$\text{logit}(\eta_i) = [1 \ x_i]\,\vartheta$$

Similarly one can formulate the correspondent multiple regression problem:

$$Y_i \sim f_{Y_i}(y|\eta_i)$$
$$\text{logit}(\eta_i) = [1\ x_{i1} \cdots x_{ip}]\,\theta \qquad (**)$$

Problems (*) and (**) are called "logistic regression" problems.

Note: In both problems, $f_{Y_i}(\cdot)$ is non-Gaussian and $g(\cdot)$ is nonlinear. However $g(x|\eta)$ can be put in a form that is LINEAR in the parameters. Moreover, both problems can be solved by using the ML method:

$Y_i \sim B(n_i, \eta_i)$ $i = 1, 2, 3, \ldots, n \Rightarrow$ log-likelihood function:

$$\ell = \ell(\eta_1, \eta_2, \ldots, \eta_n) = \sum_{i=1}^{n} \log\left(\eta_i^{y_i}(1-\eta_i)^{n_i - y_i}\right)$$

$$= \sum_{i=1}^{n}\left[y_i \log \eta_i + (n_i - y_i)\log(1-\eta_i)\right]$$

By replacing $\eta_i = \dfrac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$, we have:

$$\ell = \ell(\theta) = \sum_{i=1}^{n}\left[y_i(\beta_0 + \beta_1 x_i) - y_i \log\left(1 + e^{\beta_0 + \beta_1 x_i}\right) + \right.$$
$$\left. - (n_i - y_i)\log\left(1 + e^{\beta_0 + \beta_1 x_i}\right)\right]$$

($\theta \triangleq [\beta_0\ \beta_1]^T$)

$$\Updownarrow$$

$$\ell(\theta) = \sum_{i=1}^{n}\left[y_i(\beta_0 + \beta_1 x_i) - n_i \log\left(1 + e^{\beta_0 + \beta_1 x_i}\right)\right]$$

$\Rightarrow$ The log-likelihood function is concave $\Rightarrow$ The MLE of the parameter vector θ is obtained by imposing $\ell'(\theta) = 0$ and the S.E. of the ML estimation $\hat{\theta}$ can be obtained from $1/\ell''(\theta)$

* What are the advantages of Logistic regression vs. Linear regression?

Logistic	Linear
$Y_i \sim B(n_i, \eta_i)$	$Y_i \sim N(\eta_i, \sigma^2)$
$\text{logit}(\eta_i) = \beta_0 + \beta_1 x_i$	$\eta_i = \beta_0 + \beta_1 x_i$

- The pdf is NOT Gaussian ($\Rightarrow$ good to model discrete/quantized events) and the variance needs NOT to be constant across trials

- The function $g(\cdot)$ needs NOT to be linear $\Rightarrow$ It can be part of the more general class of nonlinear functions (e.g., exponential, logarithmic, etc.)

- Despite the nonlinearities, the logistic regression problem can be solved by using the ML method and the solution is unique (i.e., the log-likelihood function is concave)

Note: The advantages reported above are NOT exclusive to the pair $(B(\cdot,\cdot), \text{logit}(\cdot))$ $\Rightarrow$ We can consider numerous alternatives and still preserve the advantages $\Rightarrow$ We can define new types of regression problems. For instance:

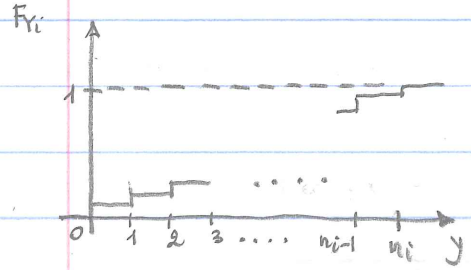
(***) $Y_i \sim B(n_i, \eta_i)$ where: $\Phi(z) \triangleq P(Z \le z)$ and $Z \sim N(0,1)$

$\Phi^{-1}(\eta_i) = \beta_0 + \beta_1 x_i$ ($\Phi^{-1} \triangleq \text{probit}(\cdot)$) i.e., $\Phi(z) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2}\, du$

Note: This choice is not casual but it rather reflects a property of the logistic regression:

$p_i \equiv \underset{Y_i}{E}\left(\frac{Y_i}{n_i}\right) \Rightarrow F_{Y_i}(y) = \sum_{k} \binom{n_i}{k} p_i^k (1-p_i)^{n_i-k} \Rightarrow F_{Y_i}$ is a combination of expit functions

(F_{Y_i}: CDF; k: number of integers $< y$)



where $\text{expit}(p) \triangleq \frac{e^p}{1+e^p} \Rightarrow F_{Y_i}(\cdot)$ has a sigmoidal shape $\Rightarrow$ Hence, model (***) simply replaces a sigmoidal cdf with another sigmoidal cdf $\Rightarrow$ This may help to better fit experimental data

Ex.: (Latent Variable) Let us assume that the value of a binary RV Y depends on the value of another (continuous) RV W (e.g., Y may model the "perception" event while W captures the intensity that is required to the perception process in order to generate an event):

$$Y = \begin{cases} 1 & W > c \\ 0 & W \leq c \end{cases} \quad \text{where } c \text{ - given and } W \sim N(\mu_W, 1)$$

By defining $Z \triangleq \mu_W - W$ we have: $P(Y=1) = P(W>c) = P(Z < \mu_W - c)$ and $Z \sim N(0,1)$. Hence, we can describe the process with the model:

$$\begin{cases} Y \sim B(1, \eta) \\ \Phi^{-1}(\eta) = \beta_0 + \beta_1 x \end{cases}$$

□

Similarly, one can replace the probability function $B(n_i; \eta_i)$ with another probability function and still preserve the same advantages. For instance, consider:

$y_i \sim \mathcal{G}(\eta_i)$

$\log \eta_i = [1 \; x_i]\,\theta$

where: $\theta \triangleq [\beta_0 \; \beta_1]^T$

$\eta_i \triangleq \lambda_i$

$f_{Y_i}(y|\eta) \triangleq e^{-\eta}\frac{\eta^y}{y!}$

In this case, by applying the ML method, we have:

$$\ell(\eta_1, \eta_2, \dots, \eta_n) \triangleq \sum_{i=1}^{n} \log\left(e^{-\eta_i}\eta_i^{y_i}\right) = -\sum_{i=1}^{n}\eta_i + \sum_{i=1}^{n} y_i \log \eta_i$$

$$\Rightarrow \quad \ell(\theta) = -\sum_{i=1}^{n} e^{\beta_0+\beta_1 x_i} + \sum_{i=1}^{n} y_i(\beta_0+\beta_1 x_i)$$

($\eta_i = e^{\beta_0+\beta_1 x_i}$)

$$\doteq -e^{\beta_0}\sum_{i=1}^{n} e^{\beta_1 x_i} + n\beta_0\bar{y} + \beta_1\sum_{i=1}^{n} x_i y_i$$

($\bar{y} \triangleq$ sample mean)

Hence, we have: $\frac{\partial \ell}{\partial \beta_0} = 0 \Leftrightarrow \left(-\sum_{i=1}^{n} e^{\beta_1 x_i}\right)e^{\beta_0} + n\bar{y} = 0 \Leftrightarrow \hat{\beta}_0 = \log\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} e^{\beta_1 x_i}}$

$$\frac{\partial \ell}{\partial \beta_1} = 0 \Leftrightarrow \left(-\sum_{i=1}^{n} x_i e^{\beta_1 x_i}\right)e^{\beta_0} + \sum_{i=1}^{n} x_i y_i = 0$$

$$\Leftrightarrow \frac{\sum_{i=1}^{n} x_i e^{\beta_1 x_i}}{\sum_{i=1}^{n} e^{\beta_1 x_i}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i}$$

- A solution can be found numerically

□

Note this: In every example considered thus far, the possibility of using the ML method stems from the fact that (i) the function $f_{Y_i}(\cdot)$ is a combination of exponential functions and (ii) the function $g(\cdot)$ is manipulated to get a new relationship that is linear in the parameters. Finally, note that in all these examples, the variable η_i is the expected value of the data $\Rightarrow$ We can generalize:

(8)

(a) $\begin{cases} Y_i \sim f_{Y_i}(y|\eta_i) = h(y)\, e^{(A(\eta_i)T(y) - B(\eta_i))} \\ g^{-1}(\eta_i) = \underbrace{[1 \; x_i]\theta}_{\text{linear predictor}} \end{cases}$

Model (a) is called "Generalized Linear Model" (GLM) and function $g(\cdot)$ is called the "link" function, as it links the random and systematic components of the model. In fact:

- $Y_i \sim \mathcal{P}(\eta_i) \Leftrightarrow f_{Y_i}(y|\eta_i) = \frac{1}{y!}\eta_i^y e^{-\eta_i} = h(y)\, e^{(y\log\eta_i - \eta_i)}$ ($h(y)$: $\frac{1}{y!}$; $T(y)$: y; $A(\eta_i)$: $\log\eta_i$; $B(\eta_i)$: η_i)

  and

  $g^{-1}(\cdot) = \log(\cdot) \Leftrightarrow g(\cdot) = \exp(\cdot)$

- $Y_i \sim B(n_i, \eta_i) \Leftrightarrow f_{Y_i}(y|\eta_i) = \binom{n_i}{y}\eta_i^y(1-\eta_i)^{n_i - y} =$

  and

  $g^{-1}(\cdot) = \text{logit}(\cdot)$

  $= \binom{n_i}{y} e^{(y\log\eta_i + (n_i - y)\log(1-\eta_i))}$

  $= \binom{n_i}{y} e^{\left(y\underbrace{(\log\eta_i - \log(1-\eta_i))}_{A(\eta_i)} + n_i\underbrace{\log(1-\eta_i)}_{B(\eta_i)}\right)}$ ($h(y)$: $\binom{n_i}{y}$; $T(y)$: y)

- $Y_i \sim N(\eta_i, \sigma^2) \Leftrightarrow f_{Y_i}(y|\eta_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\eta_i)^2/2\sigma^2}$

  $= \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{y(y-2\eta_i)}{2\sigma^2} + \frac{\eta_i^2}{2\sigma^2}\right)}$ ($h(y)$: $\frac{1}{\sqrt{2\pi}\sigma}$; $T(y)$: y; $A(\eta_i)$: $\frac{y-2\eta_i}{2\sigma^2}$; $B(\eta_i)$: $\frac{\eta_i^2}{2\sigma^2}$)

In this way, regression methods are extended to exponential families (e.g., binomial, Poisson, normal, inverse Gaussian distributions, etc.). Moreover, the choice of the link function is NOT dictated by the chosen exponential family

The introduction of the GLMs leads to the following question: for a given pair (GLM, link function), how to choose the size of ϑ?

In this, one can consider the likelihood function $L(\vartheta)$ and exploit the fact that the higher $L(\vartheta)$, the better the fit $\Rightarrow$ One can choose a set of parameters ϑ_1 over another set ϑ_2 based on the ratio:

$$\frac{L(\vartheta_1)}{L(\vartheta_2)} \iff \text{or, equivalently, } r = \log(L(\vartheta_1)) - \log(L(\vartheta_2))$$

In particular, if $\vartheta_1 = \beta_0$ and $\vartheta_2 = [\beta_0 \ \beta_1]^T$ we have:

$D_n \triangleq -2 \log L(\beta_0)$ - null deviance

$D_r \triangleq -2 \log L(\vartheta_2)$ - residual deviance

$d \triangleq -2\log(r) = D_n - D_r \Rightarrow$ The decision about whether or not $\beta_1 = 0$ depends on the value of the log-likelihood ratio

(d: log-likelihood ratio)

One last generalization from the linear regression case:

- In linear regression, we assume that the variance remains the same $\forall i$ $\Rightarrow$ The sample estimation is: $S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ with: $\hat{y}_i = \hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$\Rightarrow$ The sum: $SSE \triangleq \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ (sum of squares for error) accounts for the prediction error

Analogously, one can define:

$SST \triangleq \sum_{i=1}^{n}(y_i - \bar{y})^2$ (total sum of squares) $\Rightarrow$ It accounts for the entire variability in the data

with $\bar{y} \triangleq$ sample mean

The ratio: $R^2 \triangleq \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$ is the proportion of variability in Y that is attributable to the regression line

$\Rightarrow R^2$ is the proportion of variability in Y that is explained by X

The definition of R^2, however, is not valid in the GLM case $\Rightarrow$ An alternative is the Nagelkerke R^2 measure:

$$R_N^2 \triangleq 1 - \left(\frac{L(\beta_0)}{L(\vartheta)} \right)^{2/n}$$

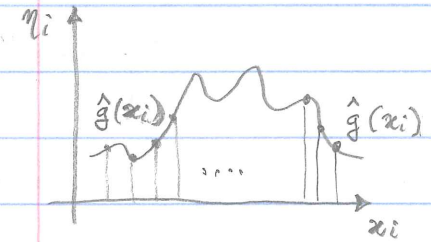
with the max value given for: $R_{N\,max}^2 = 1 - \left(L(\beta_0) \right)^{2/n}$

* Nonparametric Regression

Consider the original model we started from: $Y_i \sim f_{Y_i}(y|\eta_i)$

$$\eta_i = g(x_{1i}, x_{2i}, \dots, x_{pi})$$

In this case, let us assume that the class of functions for $g(\cdot)$ is NOT defined, i.e, there is no set of parameters β to estimate and a general form for $g(\cdot)$ must be determined $\Rightarrow$ It is a "nonparametric regression" problem



Option #1: We can define $g(\cdot)$ as the sequence of values obtained for each eval. of the variables x's. For instance:

$g(x)$ is defined by $(\hat{g}(x_1) \dots \hat{g}(x_n))$ where $\hat{g}(\cdot)$ is an estimation

$(\hat{g}(x_1), \hat{g}(x_2), \ldots, \hat{g}(x_n))$ is a "smoother"

One possible smoother is obtained by assuming that each estimation $\hat{g}(x_i)$ is a linear combination of all the measurements $y_1, y_2, \ldots, y_n$ according to some weights $(h_{i1}\ h_{i2} \ldots h_{in})$, i.e.,

$$\hat{g}(x_i) = [h_{i1}\ h_{i2} \ldots h_{in}] \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Hence, in matricial form:

$$[\hat{g}(x_1)\ \hat{g}(x_2) \ldots \hat{g}(x_n)]^T = Hy \qquad y \triangleq [y_1\ y_2 \ldots y_n]^T$$

$$H = \{h_{ij}\} \text{ - } n \times n \text{ matrix}$$

Note this: In linear regression we have $\eta_i = [1\ x_i] \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \Rightarrow$ Denoted with:

$$X \triangleq \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix} \qquad \eta \triangleq \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{bmatrix}$$

$$\theta \triangleq [\beta_0\ \beta_1]^T$$

We have: $\eta = X\theta$

An estimator $\hat{\theta}$ of θ is computed by using the least-squares method and is given by: $\hat{\theta} = (X^TX)^{-1}X^Ty \Rightarrow$ Hence we have: $\hat{\eta} \triangleq [\hat{g}(x_1) \ldots \hat{g}(x_n)] = \underbrace{X(X^TX)^{-1}X^T}_{H \text{ - "hat" matrix}} y$

Moreover, since $\eta_1, \eta_2, \ldots, \eta_3$ are predictions, one can also compute the variance of the estimation:

$$\text{var}\left([\hat{g}(x_1)\ \hat{g}(x_2) \ldots \hat{g}(x_n)]^T\right) = \text{var}(Hy) = H\,\text{var}(y)\,H^T = \sigma^2 HH^T$$

↑ If the variance is the same $\forall i = 1, 2, \ldots, n$

Option #2: We can approximate $g(x)$ with a polynomical function of adequate order s:

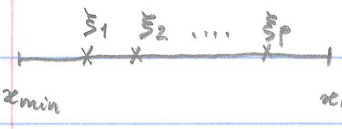
$$g(x) \cong b_0 + b_1 x + b_2 x^2 + \dots + b_s x^s \qquad (b)$$

In this case, we can define $w_1 \triangleq \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ $w_2 \triangleq \begin{bmatrix} x_1^2 \\ x_2^2 \\ \vdots \\ x_n^2 \end{bmatrix}$ $\dots$ $w_s \triangleq \begin{bmatrix} x_1^s \\ x_2^s \\ \vdots \\ x_n^s \end{bmatrix}$ and have:

$$[\hat{g}(x_1) \;\; \hat{g}(x_2) \dots \hat{g}(x_n)] = [1 \; w_1 \; w_2 \dots w_s] \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_s \end{bmatrix}$$

$w_1, w_2, \dots, w_s$ are called "basis functions" $\Rightarrow$ We can use the least-squares method to estimate $[b_0 \; b_1 \dots b_s]^T$ from data $[y_1 \; y_2 \dots, y_n]^T$ as done above.

A variation of the option with basis functions is given by "splines": The interval $x \in [x_{min}, x_{max}]$ is divided into $p+2$ sub-intervals (no overlap) and formula (b) is used in each interval with different values $[b_0 \; b_1 \dots b_s]^T$ and a continuity constrain for the extremes of adjacent intervals, i.e.:



$$(x - \xi_i)_+ \triangleq \begin{cases} x - \xi_i & \text{if } x \geq \xi_i \\ 0 & \text{otherwise} \end{cases}$$

$$\forall x \in [x_{min}, x_{max}] \quad i = 1, 2, \dots, p$$

$$g(x) \cong [b_0 \; b_1 \dots b_s] \begin{bmatrix} 1 \\ x \\ \vdots \\ x^s \end{bmatrix} + \sum_{i=1}^{p} [b_{s+(i-1)s+1} \dots b_{2s+(i-1)s}] \begin{bmatrix} (x - \xi_i)_+ \\ \vdots \\ (x - \xi_i)_+^s \end{bmatrix}$$

The function $g(x)$ is polynomial on each segment $[\xi_i; \xi_{i+1}]$ and, defined the variables: $x_1 \triangleq x$; $x_2 \triangleq x^2$; ....; $x_S \triangleq x^S$; $x_{S+1} \triangleq (x-\xi_1)_+$; ....

$x_{2S} \triangleq (x-\xi_1)_+^S$; ...; $x_{PS+1} \triangleq (x-\xi_P)_+$; ....; $x_{(P+1)S} \triangleq (x-\xi_P)_+^S$,

one can estimate the parameters $b_0, b_1, b_2, \ldots, b_{(P+1)S}$ by using a multiple linear regression method, under the assumption that the variance σ^2 is the same at each sample and the samples are independent.

Note: The solution with basis functions (in particular, splines) can be easily extended to the link function of a GLM, i.e., we can consider:

$Y_i \sim f_{Y_i}(y|\eta_i)$

$g^{-1}(\eta_i) = W\vartheta$ where W is a matrix of basis functions:

$$W \triangleq \begin{bmatrix} 1 & w_{11} & w_{12} & \ldots & w_{1S} \\ 1 & w_{21} & w_{22} & \ldots & w_{2S} \\ \vdots & & & \ddots & \\ 1 & & & & w_{SS} \end{bmatrix} \qquad w_{ij} \triangleq x_i^j$$

Option #3: In the GLM template, η_i is the expected value of the RV $Y_i \Rightarrow$ We can consider: $\left.\begin{aligned} \eta_i &= \underset{Y_i}{E}(y_i) \\ \eta_i &= g(x_i) \end{aligned}\right\} \Rightarrow \eta_i = \underset{Y_i}{E}(Y_i | x_i)$

This suggests that, if the values $x_1, x_2, \ldots, x_n$ are close to a certain value x and $g(\cdot)$ is a smooth function, then it is reasonable to expect that:

$\bar{y} \triangleq \frac{1}{n}\sum_{i=1}^{n} y_i$ is an estimator for $\underset{Y_i}{E}(Y_i | X = x) \Rightarrow$ We can consider a neighborogh of x and average the values y_i that are obtained for x_i in that neighborogh $\Rightarrow$ $g(\cdot)$ can be estimated by using the local averages $\bar{y}$. Also, in a more general form, one can use all y_i to compute $\bar{y}$, but different weights are given to the

$g(x) = \bar{y}$ ↑

values y_i depending on the distance of x_i from x (i.e., the farther x_i, the lower the weight w_i):

$$\bar{y}_w = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \qquad (c)$$

Formula (c) is particularly useful when $w_i = k\left(\frac{x - x_i}{h}\right)$ $i=1,2,\ldots,n$, where $k(u)$ is a smooth function (e.g., a pdf) and is called KERNEL, and h controls the sensitivity to $x - x_i$ and is called BANDWIDTH ($k(u) \sim N(0,1) \Rightarrow h$ plays the role of s.d.). In this case, we have:

$$g(x) = \frac{\sum_{i=1}^{n} k\left(\frac{x-x_i}{h}\right) y_i}{\sum_{i=1}^{n} k\left(\frac{x-x_i}{h}\right)}$$

□

References:

Textbook: ch. 14

ch. 12 (section 12.1, 12.5.1, 12.5.3 up to page 343)

ch. 15 (section 15.1, 15.2.1, 15.2.2, 15.2.3, 15.2.4, 15.2.7, 15.3.1)

Recommended Reading:

McCullagh P. & Nelder J.A. "Generalized Linear Models", 2nd ed., CRC, 1990: read ch. 2 (a copy is on HuskyCT)