LECTURE 4

So far, we have considered the following problem:

\[ X \rightarrow x_1, x_2, x_3, \ldots \]

Random Variable

\[ ? \hat{\theta} : X \sim f_X(x | \hat{\theta}) \quad \text{(estimation problem)} \]

We have approached this problem by using ML estimation, i.e.,

\[ \hat{\theta} = \arg \max_{\theta} l(\theta) \]

with:

\[ l(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log \left( f_X(x_i | \theta) \right) \]

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

A) \[ X \Rightarrow Y \rightarrow y_1, y_2, y_3, \ldots \]

Random Variable

\[ ? \hat{\theta} : y_i \sim f_{Y|\hat{\theta}}(y_i | \hat{\eta}_i) \]

\[ \hat{\eta}_i = g(x_i | \hat{\theta}) \]

\[ x_i \triangleq \text{sample of } X \]

(regression problem)

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

The problem can be also posed as a multivariable problem:

B) \[ X_1 \rightarrow x_{i1}, x_{i2}, \ldots \]

\[ X_2 \rightarrow x_{i1}, x_{i2}, \ldots \]

\[ \vdots \]

\[ X_p \rightarrow x_{i1}, x_{i2}, \ldots \]

\[ Y \rightarrow y_1, y_2, y_3, \ldots \]

\[ ? \hat{\theta} : y_i \sim f_{Y|\hat{\theta}}(y_i | \hat{\eta}_i) \]

\[ \hat{\eta}_i = g(x_{i1}, x_{i2}, \ldots, x_{ip} | \hat{\theta}) \]

\[ x_{ij} \triangleq \text{sample of } X_j, \quad j=1,2,\ldots,p \]

(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 \( \Theta \), more than one solution, or no solution.

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 \( \mu_i \) unknown, \( \sigma^2 \) known \( \Rightarrow \) We set: \( \eta_i = \mu_i \).

\[ f_{\eta_i}(y | \eta_i) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(y-\eta_i)^2}{2\sigma^2}} \]

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

Note: The formulation of \( f_{\eta_i} \) and \( g \), and the definition of \( \eta_i \) and \( \Theta \) imply that we have:

\[ Y_i = \beta_0 + \beta_1 x_i + \epsilon_i; \quad RV \sim N(0, \sigma^2) \]

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., \( \text{var}(\epsilon_i) = \sigma^2 \)), 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 \( i = 1, 2, 3, \ldots \)

\( \Rightarrow \) For instance: \( Y_i \sim B(n_i, p_i) \)
- \( Y_i \) is not Gaussian

Where: \( n_i \) = number of trials used to estimate \( Y_i \)

\[ p_i = \frac{E(Y_i)}{n_i} \] - expected value of count \( Y_i \) over \( n_i \) trials

In this case, we need to identify "\( \eta \)", "\( g(\cdot) \)", and "\( \Theta \). Note this:

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

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

\[ \eta_i = p_i; \ \Theta = [\beta_0, \beta_1]^T \]

\[ g(x_i; \Theta) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \]

\[ Y_i = f_{Y_i}(y|\eta_i) \]

\[ \eta_i = g(x_i; \Theta) \]

\[ \text{regression problem} \]

where:

\[ f_{Y_i}(y|\eta_i) = \text{P}(Y_i = y) = \binom{n_i}{y} \eta_i^y (1-\eta_i)^{n_i-y} \]

\[ \text{definition of } B(n_i, p_i) \]

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

\[ \logit(\eta_i) = [1 x_i] \Theta \]

\[ Y_i \sim f_{Y_i}(y|\eta_i) \]

\[ (*) \]
Similarly one can formulate the

\[ Y_i \sim f_{\eta_i} (y | \eta_i) \]

correspondent multiple regression problem:

\[ \text{logit}(\eta_i) = [1 \ z_{i1} \ldots z_{ik}] \theta \]

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

Note: In both problems, \( f_{\eta_i} (\cdot) \) 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:

\[ Y_i \sim B (n_i, \eta_i) \quad i=1, 2, 3, \ldots, n \Rightarrow \text{log-likelihood function:} \]

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

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

By replacing \( \eta_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \), we have:

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

\[ \left. - (n_i-y_i) \log (1 + e^{\beta_0 + \beta_1 x_i}) \right] \]

\[ \Rightarrow \]

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

\[ \Rightarrow \text{The log-likelihood function is concave} \Rightarrow \text{The MLE of the parameter} \]

\[ \hat{\theta} \text{ is obtained by imposing } l'(\theta) = 0 \text{ and the S.E. of the ML estimation} \]

\[ \hat{\theta} \text{ can be obtained from } 1/l''(\theta) \]
* What are the advantages of Logistic regression vs. Linear regression?

Logistic

\[ Y_i \sim B(n_i, \eta_i) \]

\[ \logit(\eta_i) = \beta_0 + \beta_i x_i \]

Linear

\[ Y_i \sim N(\eta_i, \sigma^2) \]

\[ \eta_i = \beta_0 + \beta_i x_i \]

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

- The function \( g(.) \) needs NOT to be linear ⇒ 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 \((B(\cdot), \logit(\cdot))\) ⇒ We can consider numerous alternatives and still preserve the advantages ⇒ We can define new types of regression problems. For instance:

\[ Y_i \sim B(n_i, \eta_i) \]

\[ \phi^{-1}(\eta_i) = \beta_0 + \beta_i x_i \]

where: \( \phi(z) = \Phi(z) \) and \( z \sim N(0, 1) \)

\[ \phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2} \, du \]

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

\[ p_c = E_{Y_c}(Y_c) \Rightarrow F_{Y_c}(y) = \sum_{k} \binom{n_c}{k} p_c^k (1-p_c)^{n_c-k} \Rightarrow F_{Y_c} \text{ is a combination of } \text{expit functions} \]
where \( \text{expit}(p) = \frac{e^p}{1 + e^p} \Rightarrow F_Y(\cdot) \) 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):

\[
Y = \begin{cases} 
1 & W > c \\
0 & W \leq c 
\end{cases}
\]

where \( c \) given and \( W \sim N(\mu_w, 1) \)

By defining \( Z \triangleq \mu_w - W \) we have: \( \mathbb{P}(Y=1) = \mathbb{P}(W > c) = \mathbb{P}(Z < \mu_w - c) \)

and \( Z \sim N(0, 1) \). Hence, we can describe the process with the model:

\[
\begin{align*}
Y &\sim \text{B}(1, \eta) \\
\phi^*(\eta) &= \beta_0 + \beta_1 x
\end{align*}
\]

Similarly, one can replace the probability function \( \text{B}(n_i; \eta_i) \) 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: \[ \theta = [\beta_0 \ \beta_1]^T \]

\[ \eta_i = \lambda_i \]

\[ f_{\mathcal{G}}(y|\eta) = \frac{e^{-\eta} \eta^y}{y!} \]

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

\[ \ell(\eta_1, \eta_2, \ldots, \eta_n) = \sum_{i=1}^{n} \log(\eta_i^{y_i}) = -\sum_{i=1}^{n} \eta_i + \sum_{i=1}^{n} y_i \log \eta_i \]

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

\[ \eta_i = e^{\beta_0 + \beta_1 x_i} \]

\[ y_{\text{sample mean}} \]

Hence, we have:

\[ \frac{\partial \ell}{\partial \beta_0} = 0 \iff \left( -\sum_{i=1}^{n} e^{\beta_1 x_i} \right) e^{\beta_0} + n \bar{y} = 0 \iff \hat{\beta}_0 = \log \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} e^{\beta_1 x_i}} \]

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

\[ \iff \frac{\sum_{i=1}^{n} x_i e^{\beta_1 x_i}}{\sum_{i=1}^{n} e^{\beta_1 x_i}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i} \quad - \text{A solution can be found numerically} \]

Note this: In every example considered thus far, the possibility of using the ML method stems from the fact that (i) the function \( f_{\mathcal{G}}(\cdot) \) 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 \( \eta_i \) is the expected value of the data \( \Rightarrow \) We can generalize:
Model (a) is called "Generalized Linear Model" (GLM) and function \( g(.) \) is called the "link" function, as it links the random and systematic components of the model. In fact:

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

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 $\theta$?

In this, one can consider the likelihood function $L(\theta)$ and exploit the fact that the higher $L(\theta)$, the better the fit $\Rightarrow$ One can choose a set of parameters $\theta_1$ over another set $\theta_2$ based on the ratio:

$$\frac{L(\theta_1)}{L(\theta_2)} \leftrightarrow \text{or, equivalently}, \quad r = \frac{\log L(\theta_1)}{-\log L(\theta_2)}$$

In particular, if $\theta_1 = \beta_0$ and $\theta_2 = [\beta_0 \beta_1]^T$ we have:

- $D_n = -2 \log L(\beta_0)$ - null deviance
- $D_r = -2 \log L(\theta_2)$ - 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

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

Analogously, one can define:

$$\text{SST} \triangleq \sum_{i=1}^{n} (y_i - \bar{y})^2$$ (total sum of squares) $\Rightarrow$ It accounts for the entire variability in the data
The ratio: \( R^2 \equiv \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} \) is the proportion of variability in \( Y \) that is attributable to the regression line.

\[ R^2 \Rightarrow \text{the proportion of variability in } Y \text{ that is explained by } X \]

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

\[ R^2_N \equiv 1 - \left( \frac{L(\hat{\beta}_0)}{L(\hat{\beta})} \right)^{2/n} \]

with the max value given for: \( R^2_{N \text{ max}} = 1 - (L(\hat{\beta}_0))^{2/n} \)

*Non-parametric Regression*

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

\[ \eta_i = g(x_{1i}, x_{2i}, \ldots, x_{pi}) \]

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 \( \theta \) to estimate and a general form for \( g(\cdot) \) must be determined \( \Rightarrow \) It is a "non-parametric regression" problem.

\[ \eta_i \quad \text{Option #1: We can define } g(\cdot) \text{ as the sequence of values obtained for each eval. of the variables } x_i \text{'s. For instance:} \]

\[ g(x) \text{ is defined by } (\hat{g}(x_1), \ldots, \hat{g}(x_n)) \]

where \( \hat{g}(\cdot) \) is an estimation.
The set of estimators \( \hat{\theta}(x_i) \) is a "smoother".

One possible smoother is obtained by assuming that each estimation \( \hat{\theta}(x_i) \) is a linear combination of all the measurements \( y_1, y_2, ..., y_n \) according to some weights \( h_i, h_2, ..., h_n \), i.e.,

\[
\hat{\theta}(x_i) = [h_i, h_2, ..., h_n] [y_1] [y_2] ... [y_n]
\]

Hence, in matrixial form:

\[
[\hat{\theta}(x_1) \hat{\theta}(x_2) \cdots \hat{\theta}(x_n)]^T = H y
\]

\[
y = [y_1 \ y_2 \cdots \ y_n]^T
\]

\[
H = \sum h_i y_i^2 - n \times n \text{ matrix}
\]

Note that: In linear regression, we have \( \eta = [1 \ x_i] \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \Rightarrow \text{Denoted with } \eta = X \theta \)

\[
X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \eta = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{bmatrix}
\]

We have: \( \eta = X \theta \)

\[
\theta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}^T
\]

An estimator \( \hat{\theta} \) of \( \theta \) is computed by using the least-squares method and is given by:

\[
\hat{\theta} = (X^T X)^{-1} X^T y \Rightarrow \text{Hence we have: } \hat{\theta} = [\hat{\theta}(x_1) \cdots \hat{\theta}(x_n)] = X (X^T X)^{-1} X^T y
\]

Moreover, since \( \eta_1, \eta_2, ..., \eta_n \) are predictions, one can also compute the variance of the estimation:

\[
\text{var} \left( [\hat{\theta}(x_1) \hat{\theta}(x_2) \cdots \hat{\theta}(x_n)]^T \right) = \text{var} (H y) = H \text{var} (y) H^T = \sigma^2 H H^T
\]

If the variance is the same \( \forall i = 1, 2, ..., n \)
Option #2: We can approximate \( g(x) \) with a polynomial function of adequate order \( s \):

\[
g(x) = b_0 + b_1 x + b_2 x^2 + \ldots + b_s x^s
\]

In this case, we can define \( w_i = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \), \( w_2 = \begin{bmatrix} x_2^2 \\ x_2^3 \\ \vdots \\ x_n^2 \end{bmatrix} \), \ldots, \( w_5 = \begin{bmatrix} x_1^5 \\ x_2^5 \\ \vdots \\ x_n^5 \end{bmatrix} \) and have:

\[
\left[ \hat{g}(x_1), \hat{g}(x_2), \ldots, \hat{g}(x_n) \right] = \left[ 1 \ w_1 \ w_2 \ldots \ w_s \right] \left[ \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_s \end{bmatrix} \right]
\]

\( w_1, w_2, \ldots, w_s \) are called "basis functions." We can use the least-squares method to estimate \( [b_0 \ b_1 \ldots \ b_s]^T \) from data \( [y_1 \ y_2 \ldots \ y_n]^T \) 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 \( [b_0 \ b_1 \ldots \ b_s]^T \) and a continuity constraint for the extremes of adjacent intervals, i.e.:

\[
\ell_i(x) = \left\{ \begin{array}{ll}
(x - \xi_i)^+ & \text{if } x \geq \xi_i \\
0 & \text{otherwise}
\end{array} \right.
\]

\( \forall x \in [x_{\min}, x_{\max}] \), \( i = 1, 2, \ldots, p \)

\[
g(x) \equiv \left[ b_0 \ b_1 \ldots \ b_s \right] \begin{bmatrix} x \\ x^2 \\ \vdots \\ x^s \end{bmatrix}^T + \sum_{i=1}^{p} \left[ b_{s+i(\xi_{i-1})} \ldots \ b_{s+i(\xi_{i+1})} \right] \begin{bmatrix} (x - \xi_i)^+ \\ \vdots \\ (x - \xi_i)^s \end{bmatrix}
\]
The function $g(x)$ is polynomial on each segment $[\xi_i, \xi_{i+1}]$ and, defined the variables:

$$x_1 \equiv x; \quad x_2 \equiv x^2; \quad \ldots; \quad x_s \equiv x^s; \quad x_{s+1} \equiv (x-\xi_1)^s; \quad \ldots;$$

$$x_{2s} \equiv (x-\xi_1)^s; \quad \ldots; \quad x_{ps+1} \equiv (x-\xi_p)^s; \quad \ldots; \quad x_{ps+s} \equiv (x-\xi_p)^s,$$

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 $\sigma^2$ 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:

$$Y_i \sim f_{\eta_i}(y_i | \eta_i),$$

where $W$ is a matrix of basic functions:

$$W = \begin{bmatrix}
        1 & w_{11} & w_{12} & \ldots & w_{1s} \\
        1 & w_{21} & w_{22} & \ldots & w_{2s} \\
        \vdots & \vdots & \vdots & \ddots & \vdots \\
        1 & w_{ps1} & w_{ps2} & \ldots & w_{ps}\end{bmatrix}$$

$$w_{ij} = x_{i}^j$$

Option #2: In the GLM template, $\eta_i$ is the expected value of the RV $Y_i \Rightarrow$ We can consider:

$$\eta_i = E(Y_i | \eta_i) \left\{ \Rightarrow \eta_i = E(Y_i | \xi_i) \right\}$$

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:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \text{ is an estimator for } E(Y_i | x = x) \Rightarrow \text{We can consider a neighborhood of } x \text{ and average the values } y_i \text{ that are obtained for } x_i \text{ in that neighborhood} \Rightarrow$$

$$g(x) \text{ can be estimated by using the local averages } \bar{y}. \text{ Also, in a more general form, one can use all } y_i \text{ to compute } \bar{y}, \text{ but different weights are given to the}$$
values $y_i$ depending on the distance of $x_i$ from $x$ (i.e., the farther $x_i$, the lower the weight $w_i$):

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

Formula (c) is particularly useful when $w_i = k\left(\frac{x-x_i}{h}\right)$, $i=1,2,...,n$, where $k(u)$ is a smooth function (e.g., a polf) 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:

Mc Cullagh P. & Nelder J.A. "Generalized Linear Models", 2nd ed., CRC, 1990: read ch. 8 (a copy is on HuskyCT)