Lecture FYS-STK3155/4155, September 21, 2023

$$\mathbb{E}[y_i] = \sum_j x_{ij} \beta_j = x_i * \beta$$

$$\text{var}[y_i] = \sigma^2$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

$$y_i = f(x_i) + \varepsilon_i$$

$$y_i \in (-\infty, +\infty)$$

$$x_i \in (-\infty, +\infty)$$

assumption: y_i are iid.

$$y_i \sim N(x_i * \beta, \sigma^2)$$

$$P(y_i, x_i \mid \beta)$$

$$D = \{ (x_0, y_0), (x_1, y_1) \ldots (x_{n-1}, y_{n-1}) \}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ \frac{-(y_i - x_i * \beta)^2}{2\sigma^2} \right\}$$

$$P(D \mid \beta) = \prod_{i=0}^{n-1} p(y_i, x_i \mid \beta)$$

$$\hat{\beta} = \arg\max_{\beta \in \mathbb{R}^p} P(D \mid \beta)$$

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\max} \log P(D|\beta)$$

$\leftarrow$ MLE
maximum likelihood estimator

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left[-\log P(D|\beta)\right]$$

$$\log \prod_{i=1}^{n} a_i = a_1 a_2 a_3 \dots a_n$$

$$\sum_{i=1}^{n} \log a_i$$

$$-\log(P(D|\beta)) =$$

$$-\log\left[\prod_{i=0}^{n-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - x_i*\beta)^2}{2\sigma^2}\right]\right]$$

$$= \overbrace{\frac{n}{2}\log(2\pi\sigma^2)}^{\propto}$$

$$+ \sum_{i=0}^{n-1} \frac{(y_i - x_i*\beta)^2}{2\sigma^2} = C(\beta)$$

$$\Bigg[\quad \sum_{i=0}^{n-1} \log(2\pi\sigma^2)^{1/2}$$

$$C(\beta) = \alpha + \frac{1}{2\sigma^2}\|(y - X\beta)\|_2^2$$

$$\frac{\partial C(\beta)}{\partial \beta} = 0 \implies$$

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

Logistic Regression:

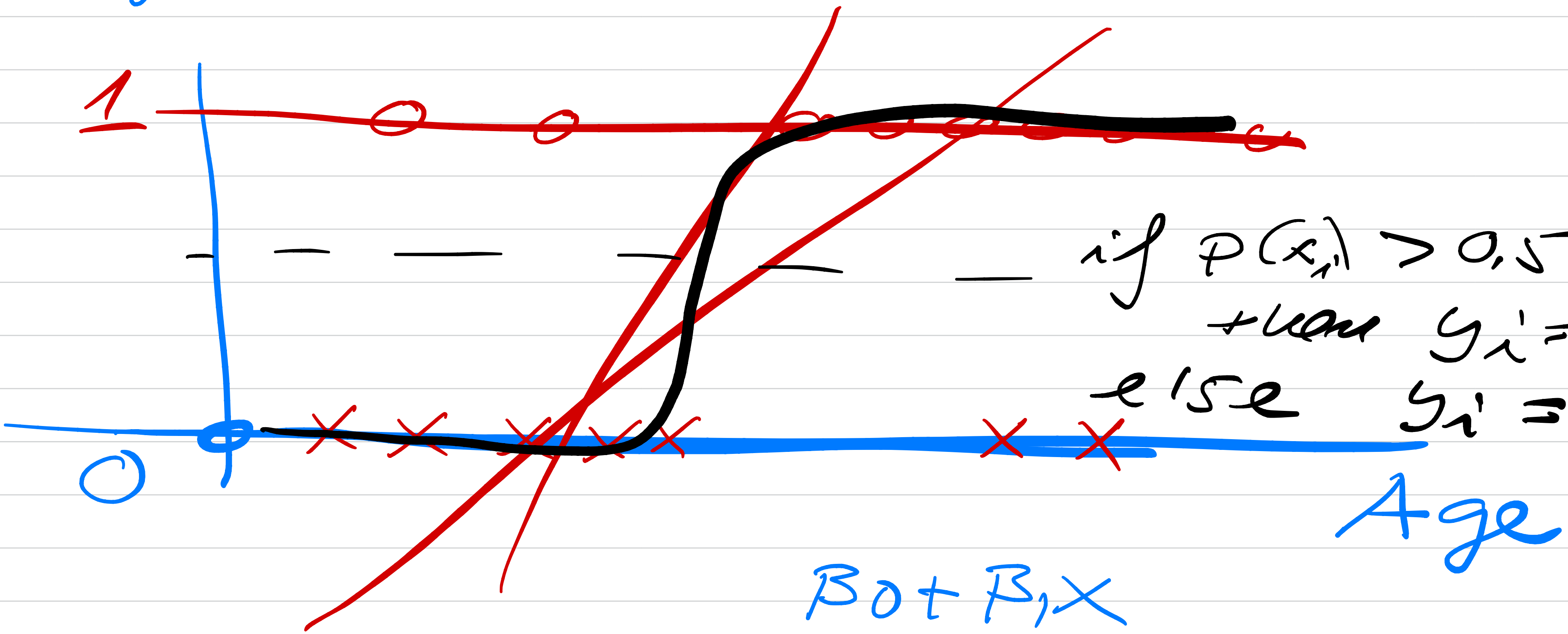
Classification problem

Discrete output; binary

$$y_i = \{0, 1\}$$

$$y_i = p(x_i) + \varepsilon_i$$

$$f(x) \Rightarrow p(x) \text{ , probability}$$



if $p(x_i) > 0{,}5$ then $y_i = 1$
else $y_i = 0$

$$p(x) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \quad \text{(sigmoid)}$$

assumption

$$y(x) = p(x) + \varepsilon$$

$$p(x) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

$$\Rightarrow P(y_i, x_i \mid \beta) = p_i$$

$$D = \{(x_0, y_0), (x_1, y_1), \ldots, (x_{n-1}, y_{n-1})\}$$

$$P(D \mid \beta) = \prod_{i=0}^{n-1} p_i$$

$$C(\beta) = -\log P(D|\beta)$$

$$\frac{\partial C}{\partial \beta} = 0$$

$y_i = 1$, probability P_i

$y_i = 1 = P_i + \varepsilon_i \Rightarrow$

$\varepsilon_i = 1 - P_i$

$y_i = 0 = P_i + \varepsilon_i \Rightarrow \varepsilon_i = -P_i$

what is the probability of ε?

$$\mathbb{E}[\varepsilon] = \sum_{i=0}^{n-1} P_i \varepsilon_i$$

$$P_i = P$$

$$= (1-p)P + (-p)(1-p) = 0$$

$y_i = 0$ $\quad P_{y_i=0} = 1 - P_i$ $\quad P_i$ is the probability for $y_i = 1$

$$\mathrm{var}[\varepsilon] = (1-p)^2 P + (-p)^2 (1-p)$$

Binomial!

$$= P(1-P)$$

$$P(D/\beta) = \prod_{i=0}^{n-1} P_i^{y_i} (1-P_i)^{1-y_i}$$

$$P_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

$$f(x_i) \simeq \tilde{y}_i = X_{i*} \beta$$

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left[-\log P(D/\beta) \right]$$

$$C(\beta) = -\sum_{i=0}^{n-1} \Big[y_i \log p_i + (1-y_i) \log(1-p_i) \Big]$$

$$\left[\begin{aligned} y_i \log p_i &= y_i \log\left[\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \right] \\ &= y_i(\beta_0+\beta_1 x_i) - y_i \log(1+e^{\beta_0+\beta_1 x_i}) \end{aligned} \right.$$

$$C(\beta) = -\sum_{i=0}^{n-1} \Big[y_i(\beta_0+\beta_1 x_i) - \log(1+e^{\beta_0+\beta_1 x_i}) \Big]$$

$$\frac{\partial C(\beta)}{\partial \beta_0} = -\sum_{i=0}^{n-1}(y_i - p_i) = -g_0$$

$$= 0$$

$$\frac{\partial C(\beta)}{\partial \beta_1} = 0 = -\sum_{i=0}^{n-1} x_i(y_i - p_i)$$

$$= -g_1$$

for the more general case

$$\frac{\partial C(\beta)}{\partial \beta} = 0 \Rightarrow X^T(p-y) = g$$

$$X^T \in \mathbb{R}^{p \times n} \quad y, p \in \mathbb{R}^n \quad g \in \mathbb{R}^p$$

$$\frac{\partial^2 C}{\partial \beta \, \partial \beta^T} = H = X^T W X$$

$$W = \begin{cases} w_{ii} = p_i(1-p_i) \\ 0 \quad \text{if } i \neq j \end{cases}$$

β is in $p_i = \dfrac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$

$$\frac{\partial C}{\partial \beta} = 0 \Rightarrow X^T(p-y) = 0$$

non-linear equation i.e. β

$$g = X^T(p - y) = g(\beta) \text{ through } p(\beta)$$

$$H = X^T W X = H(\beta)$$

$W(\beta)$

Taylor expansion of $C(\beta)$ around $\hat{\beta}$ $\hat{\beta} - \beta^{(n)}$

n refers to iteration - n -

$$\frac{\partial C}{\partial \beta} = \vec{\nabla}_\beta C = 0$$

$$C(\hat{\beta}) = C(\beta^{(n)}) + (g^{(n)})^T(\hat{\beta}-\beta^{(n)})$$

known

$$+ \frac{1}{2}(\hat{\beta}-\beta^{(n)})^T H(\beta^{(n)})(\hat{\beta}-\beta^{(n)})$$

$$+ \ldots \qquad \left.\frac{\partial^2 C}{\partial \beta \partial \beta^T}\right|_{\beta=\beta^{(n)}} = H(\beta^{(n)})$$

$$b = \hat{\beta} - \beta^{(n)} \implies$$

$$C(\hat{\beta}) = C(\beta^{(n)}) + (g^{(n)})^T b$$

$$+ \frac{1}{2} b^T H^{(n)} b + \ldots$$

Keep to second order in b only

$$\frac{\partial C}{\partial b} = g^{(n)} - H^{(n)} \cdot b = 0$$

$$b = \hat{\beta} - \beta^{(n)} \quad \Longrightarrow$$

$$\hat{\beta} = \beta^{(n)} - (H^{(n)})^{-1} g^{(n)}$$

$$= \beta^{(n+1)} \equiv \hat{\beta}$$

For logistic regression

$$H^{(n)} = H(\beta^{(n)}) = X^T W(\beta^{(n)}) X$$

$$g^{(n)} = X^T (P(\beta^{(n)}) - y)$$

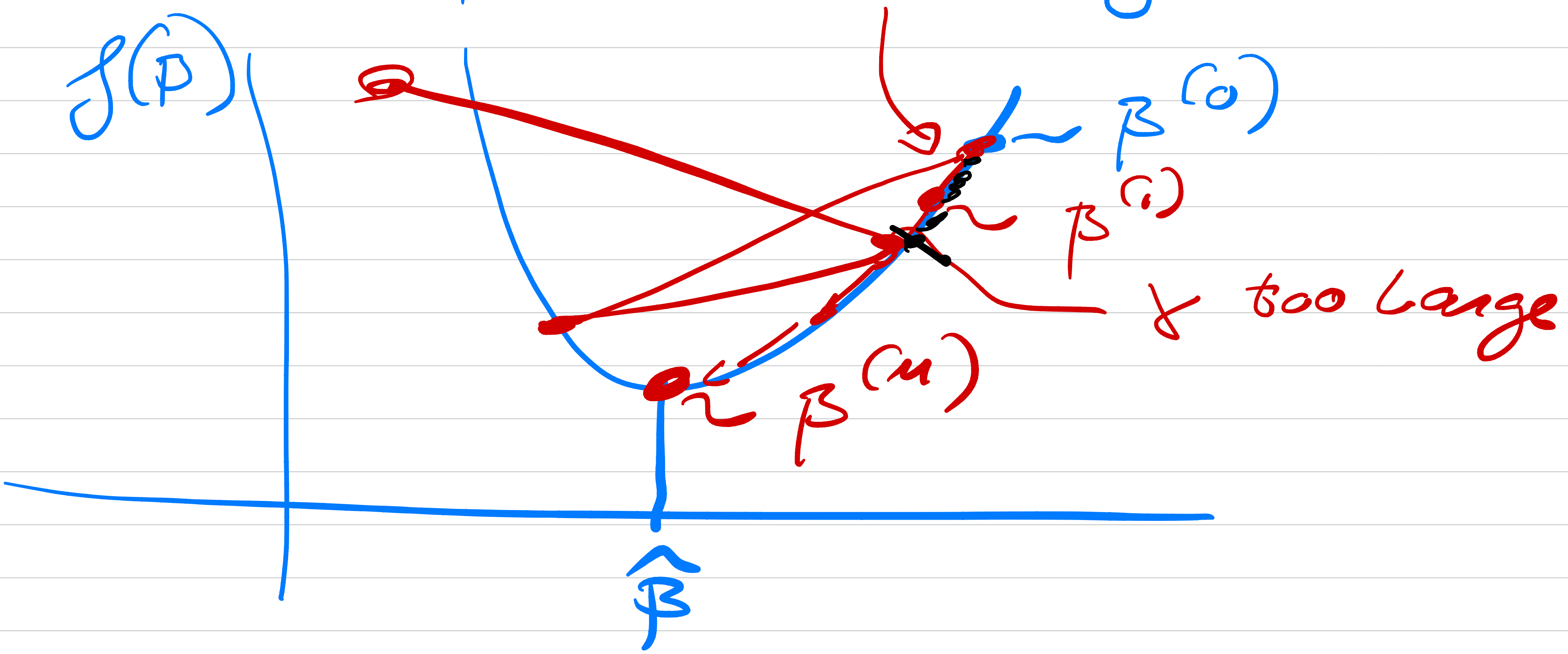
$$\beta^{(n+1)} = \beta^{(n)} - (X^T W(\beta^{(n)}) X)^{-1} \times X^T (P(\beta^{(n)}) - y)$$

Newton-Raphson's method

start with a guess $\beta^{(0)}$ and iterate n-times till $g = 0$

$H^{-1} \longrightarrow \gamma^{(n)}$ = learning rate

$$\beta^{(n+1)} = \beta^{(n)} - \gamma^{(n)} \cdot g^{(n)}$$



optimal rate $\gamma < \dfrac{2}{\lambda_{max}}$

λ_{max} is the largest eigenvalue of H