Lecture 22 22 Nov

CATEGORICAL Variables in the model

$$Y_i = \mu(x_i, G_i) + \varepsilon_i$$

$$G_i = \begin{cases} 1, & \text{Young} \\ 2, & \text{Young Adult} \\ 3, & \text{Middle} \\ 4, & \text{Senior} \end{cases} \qquad \begin{cases} 1 & \text{Native Americans} \\ 2 & \text{Caucasians} \\ 3 & \text{African} \\ 4 & \text{Other} \end{cases}$$

$$\mu(x_i, G_i) = \beta_0 + \beta_1 x_i + \beta_2 G_i$$

Young group: $\mu(x_i, G_i = 1) = \beta_0 + \beta_2 \cdot (1) + \beta_1 x_i$

$$= (\beta_0 + \beta_2) + \beta_1 x_i$$

Young Adult $\mu(x_i, G_i = 2) = \beta_0 + \beta_2 (2) + \beta_1 x_i$

$$= \beta_0 + 2\beta_2 + \beta_1 x_i$$

Define indicator function

$$G_{1i} = \begin{cases} 1, & \text{Native American} \\ 0, & \text{o/w} \end{cases}$$

$$G_{2i} = \begin{cases} 1, & \text{Caucasian} \\ 0, & \text{o/w} \end{cases}$$

G_{3i} = African

G_{4i} = others

<u>One model</u>

$$\mu(x_i, G_i) = \mu(x_i, G_{1i}, G_{2i}, G_{3i}, G_{4i})$$

$$= \sum_{g=1}^{4} (\beta_0^g + \beta_1^g x_i) G_{gi}$$

<u>Another model</u>

$$\mu(x_i, G_i) = \mu(x_i, G_{2i}, G_{3i}, G_{4i})$$

$$= (\beta_0^1 + \beta_1^1 x_i) + \sum_{g=2}^{4} (\delta_0^g + \delta_1^g x_i) G_{gi}$$

Native American

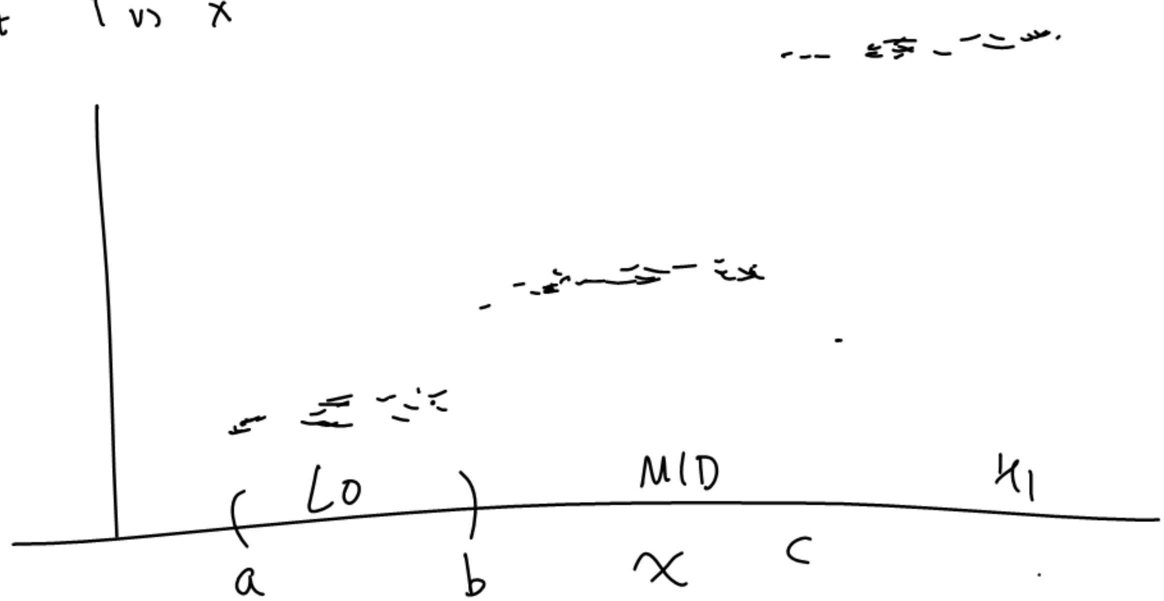
$$\mu(x_i, G_i = 1) = \mu(x_i, G_{2i} = 0, G_{3i} = 0, G_{4i} = 0)$$
$$= \beta_0^1 + \beta_1^1 x_i$$

"others"

$$\mu(x_i, G_i = 4) = \mu(x_i, G_{2i} = 0, G_{3i} = 0, G_{4i} = 1)$$
$$= (\beta_0^1 + \delta_0^4) + (\beta_1^1 + \delta_1^4) x_i$$

X continuous valued $\Longrightarrow$ categorical

Plot Y vs x

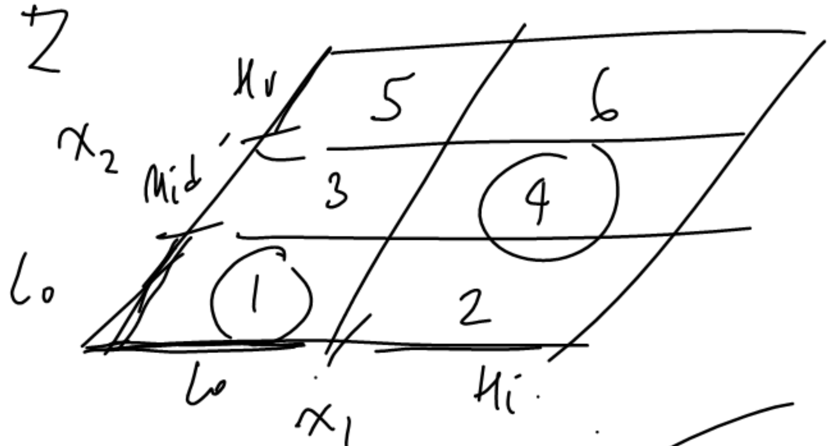


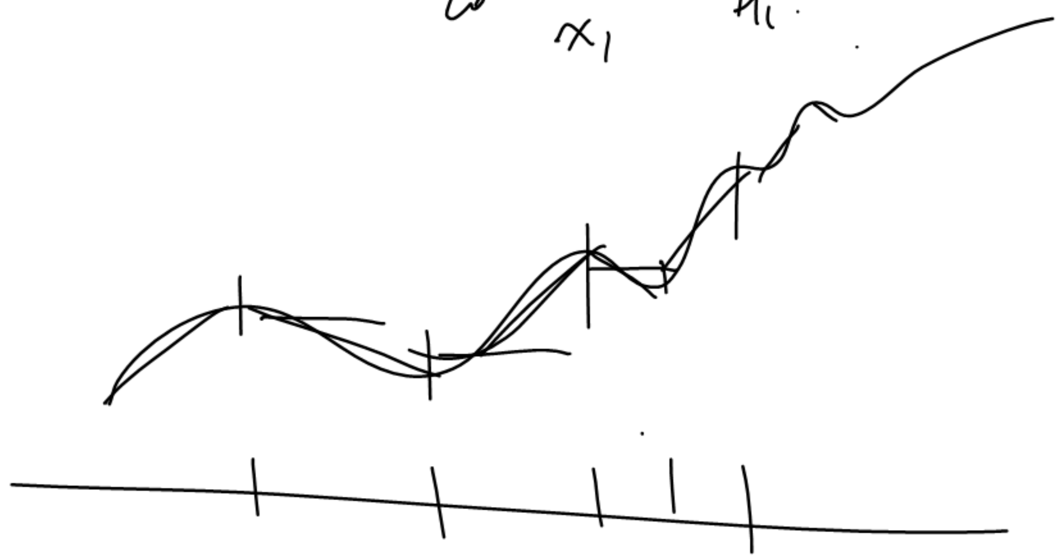
$$X \text{ continuous} \rightarrow Z = h(X)$$
$$= \begin{cases} 1, & \text{if } a < X < b \\ 2, & \text{if } b \le X < c \\ 3, & \text{if } c \le X \end{cases}$$

Define indicators

- $Z_{1i} = \begin{cases} 1, & Z_i = 1 \\ 0, & \text{o/w} \end{cases}$
- Z_{2i}
- Z_{3i}

$(X_1, X_2) \rightarrow Z$





Non parametric Function Estimation

$$Y_i = \mu(x_i) + \varepsilon_i$$

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

$$\mu(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$ PARAMETRIC

$$= \sum_{p=1}^{P} \beta_p x_i^p$$

Higher order polynomial

Taylor Series Expansion.

$\mu(x_i)$ is a "smooth" function

$$\left| \mu(x_i + L) - \mu(x_i) \right| \leq M L^{\gamma}$$ Lipschitz

$\mu(x_i)$ no parametric form

$$\mu_A(x_i) = \sum_{k=1}^{K} \beta_K \phi_k(x_i)$$

where $\{ \phi_1(x) \ldots \phi_K(x) \}$ basis functions

e.g. Fourier (sins + cosines), Spline, wavelets

Estimate $\beta = (\beta_1 \ldots \beta_K)'$ via least squares

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \left(\begin{array}{c|c|c} \phi_1(x_1) & \phi_2(x_1) & \\ \vdots & \vdots & \cdots \\ \phi_1(x_n) & \phi_2(x_n) & \end{array} \right) \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_K \end{pmatrix} + \underline{\varepsilon}$$

$$\underline{Y} = X \beta + \underline{\varepsilon}$$

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

<u>Note</u>: if $\{\phi_1(x) \ldots \phi_K(x)\}$ ONB (ortho-normal basis)

$$\Rightarrow (X'X) = I$$

$$\hat{\beta} = X'Y$$

$$\begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_K \end{pmatrix} = \begin{pmatrix} \underline{\phi}_1' \\ \vdots \\ \underline{\phi}_K' \end{pmatrix} \underline{Y}$$

$$\hat{\beta}_k = \langle \underline{\phi}_k, \underline{Y} \rangle = \underline{\phi}_k' \underline{Y}$$