

Regression

$y_t \in \mathbb{R}$ response, dependent variable

$w_t \in \mathbb{R}^d$ covariates, regressors, independent variables

$g(w_t) = E[y_t | w_t]$ conditional mean of y_t given w_t .

◦ This course will be about modelling and estimating this function.

$$\Rightarrow y_t = g(w_t) + \varepsilon_t, \text{ where } E[\varepsilon_t | w_t] = E[y_t - g(w_t) | w_t] \\ = E[y_t | w_t] - g(w_t) \\ = g(w_t) - g(w_t) = 0$$

In this course, we will:

1] approximate $E[y_t | w_t] \approx \Sigma_t' \beta$, where $\Sigma_t = f(w_t) \in \mathbb{R}^d$
(power approximations, spline approximations, Chebyshev polynomials, etc.)

2] Estimate β reasonably well. (Ordinary least squares.)

3] Make finite (small) sample and large sample inference. will use central limit theorems.

◦ In finite case, we will typically make normality assumptions.

◦ will also explore Monte Carlo methods under non-normality.

Example 1: Engel (≈ 1870 's)

- $y_t =$ food expenditure
- $w_t =$ income

$$E[y_t | w_t] \approx \beta_0 + \beta_1 w_t = \underbrace{[1 \quad w_t]}_{\equiv X_t'} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}}_{\equiv \beta} = X_t' \beta$$

alternatively,

$$\begin{aligned} E[y_t | w_t] &\approx \beta_0 + \beta_1 1\{w_t \in [w_0, w_1)\} + \beta_2 1\{w_t \in [w_1, w_2)\} \\ &\quad + \dots \\ &\approx \underbrace{[1 \quad 1\{w_t \in [w_0, w_1)\} \quad \dots]}_{\equiv X_t'} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \end{bmatrix}}_{\equiv \beta} = X_t' \beta \end{aligned}$$

- The second functional form is more flexible, but we cannot estimate it as precisely.
- The first approximation is called the power expansion. The second approximation is referred to as the Haar expansion.
- $X_t' = [1 \quad w_t \quad w_t^2 \quad \dots \quad w_t^N] =$ power basis functions
- $X_t' = [1 \quad 1\{w_t \in [w_0, w_1)\} \quad 1\{w_t \in [w_1, w_2)\} \quad \dots] =$ dummy basis functions.
- Splines combine Haar and power expansion.

- Why do we care about $E[y_t | w_t]$? (Mean regression)
- Why not care about conditional α -quantile of y_t given w_t ? (Quantile Regression)

Example 2:

- y_t^* = infant birthweight
- w_t = smoking or quality of medical care
- $E[y_t^* | w_t] \approx \mathbf{X}_t' \beta$

$$\Rightarrow \frac{\partial E[y_t^* | w_t]}{\partial w_t} \approx \frac{\partial \mathbf{X}_t'}{\partial w_t} \beta < 0$$

- How does smoking affect very low birthrates?
 - Quantile regression

$$y_t \equiv \mathbb{1}\{y_t^* < c\}$$

$$E[y_t | w_t] = \Pr[y_t^* < c | w_t]$$

◦ conditional distribution function.

$$\approx \mathbf{X}_t' \beta \quad (\text{linear probability model})$$

Suppose we recover $E[y_t | w_t]$. What do we make of it?

- 1] Descriptive - uncover interesting stylized facts
- 2] Treatment effect - (y_t, w_t) come from randomized experiment
 ↳ can then examine causality
- 3] Structural effect
 ↳ mapping estimates back into economic (causal) models.

• Fights occur between the treatment effect school and the structural effect school