

Bivariate Data

- Data will involve two variables

x	y
categorical	numerical
numerical	numerical
categorical	categorical

- Descriptive Analysis of Bivariate data in the lab.

Examples

- Maternal smoking

vs
child weight

- Mother's height

vs
child height

Questions

X - independent variable

Y - dependent variable

What is the relationship between X and Y?

Towards Answer

- Many natural phenomena

there is an intrinsic randomness associated to both X and Y

• Measurement data - The error occurring during measurement is assumed random.

Models

$$Y = f(X)$$

Try find f ?
to

$$Y = g(X) + \epsilon$$

↑
deterministic
↓

randomness
(independent of X)

Idea:

$$Y|X=x = g(x) + \epsilon$$

Basic Model (Simple Linear Regression)

- let $g(x) = \beta_0 + \beta_1 x$ for some $\beta_0, \beta_1 \in \mathbb{R}$

Model:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Assumption 1

$$E[\varepsilon] = 0$$

$$\text{Var}[\varepsilon] = \sigma^2$$

ε - independent of X

Observation

$$E[Y | X=x] = \beta_0 + \beta_1 x + 0$$

$$\Rightarrow \boxed{E[Y | X=x] = \beta_0 + \beta_1 x} \stackrel{!}{=} M_{Y|X=x}$$

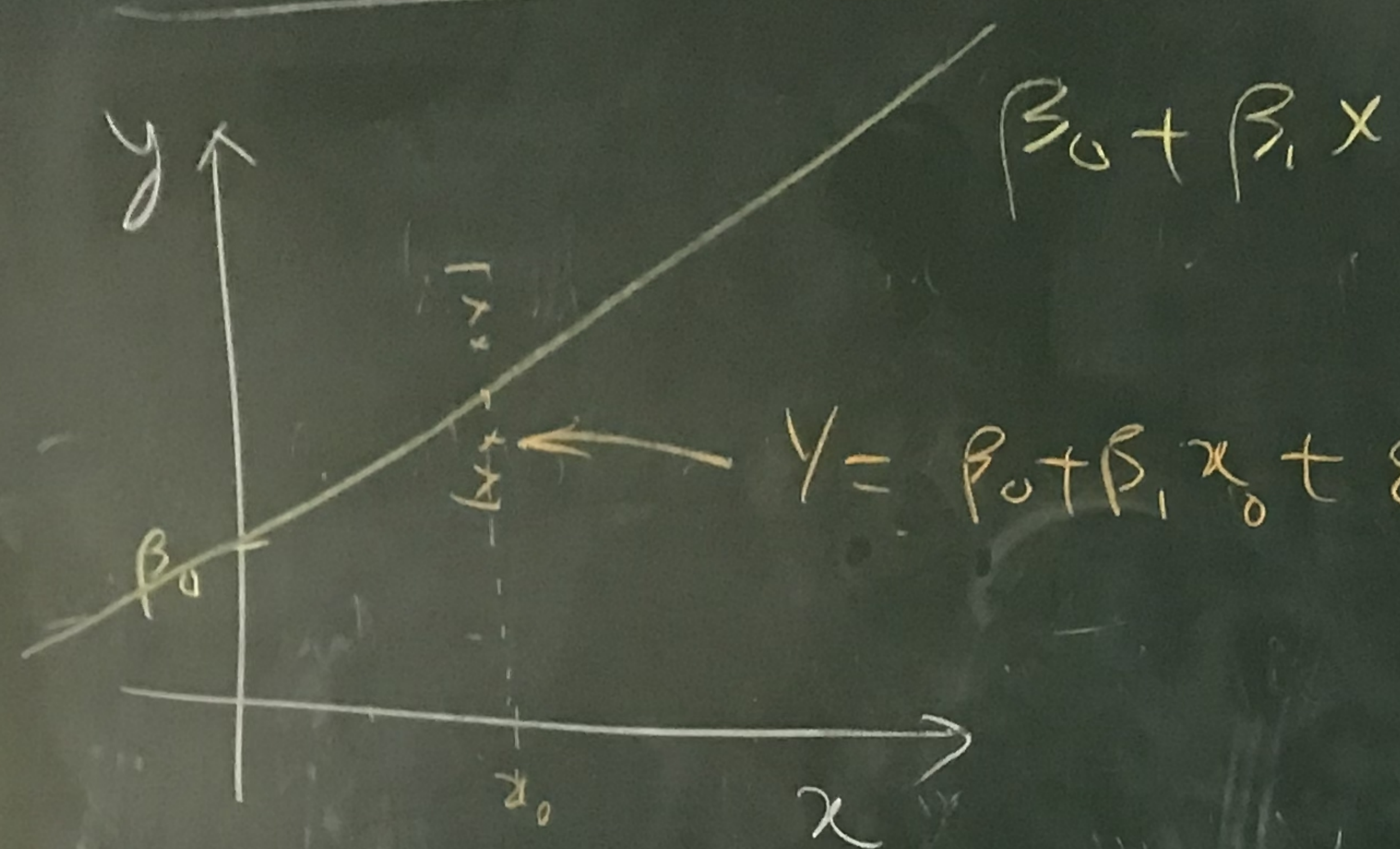
Questions:

- ① Is the relationship linear?
- ② How to estimate β_0, β_1 ?
- ③ Can we provide confidence intervals for β_0 & β_1 ?

(A2) $\varepsilon \stackrel{d}{=} \text{Normal}(0, \sigma^2)$ independent of x

Along with (A2) $y \mid E(Y \mid X=x) = \beta_0 + \beta_1 x$

Scatter plot of X and Y



$$Y = \beta_0 + \beta_1 x_0 + \epsilon \quad \stackrel{d}{=} \text{Normal}(\beta_0 + \beta_1 x_0, \sigma^2)$$

$$\beta_0 = E[Y | X=0]$$

$\beta_1 \equiv$ slope of line.

$$\sigma^2 = \text{Var}(Y | X=x) \rightarrow \begin{matrix} \text{Variability} \\ \text{around} \\ \text{conditional mean} \end{matrix}$$

Key Assumptions

- linear relation
 $Y \sim X$
 $E[Y | X=x] = \beta_0 + \beta_1 x$
- independent errors
- mean & variance are constant
- Normally distributed

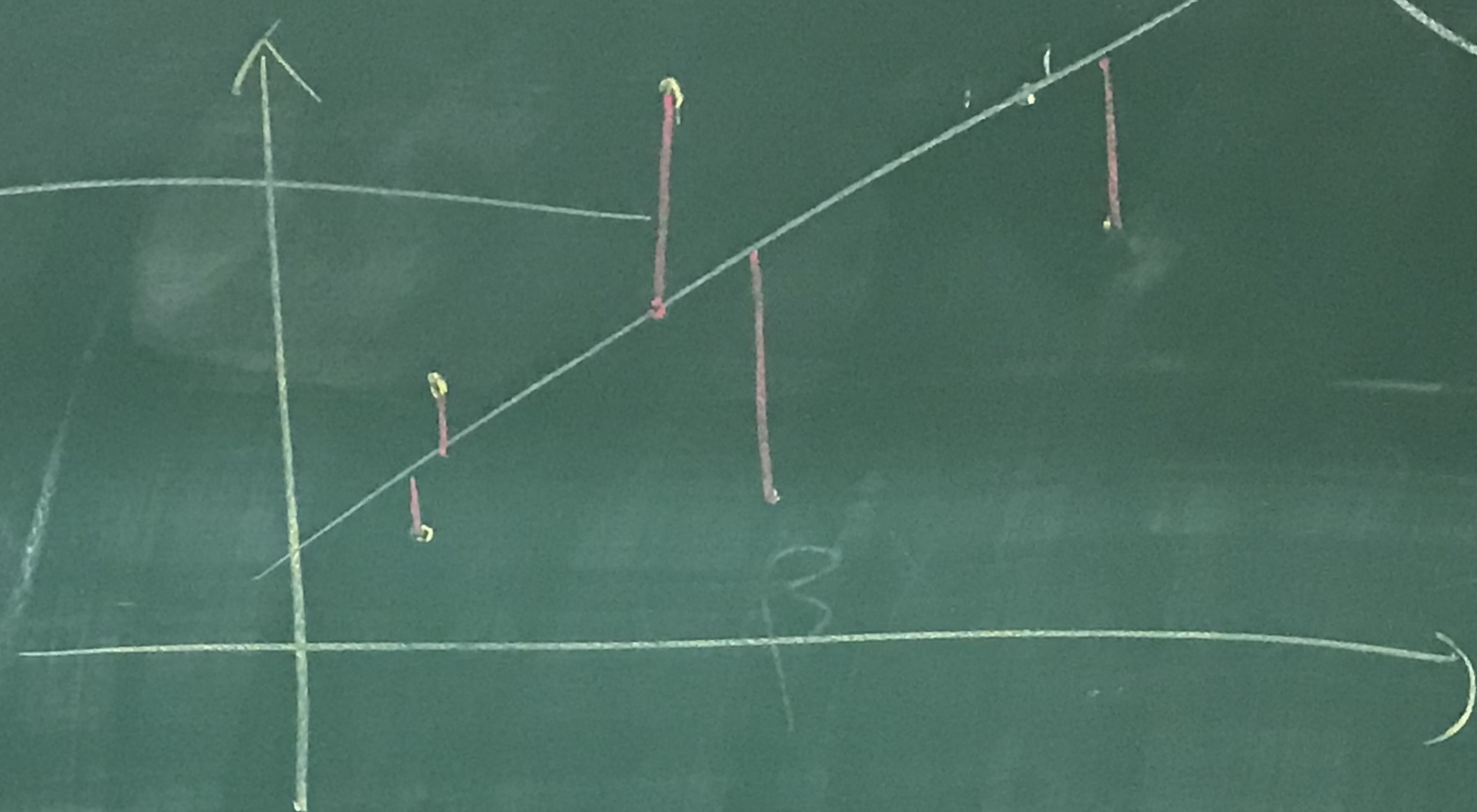
How does one estimate β_0 and β_1 ?

Given

$$(y_i, x_i)$$

$$1 \leq i \leq n$$

Squared
vertical
distance



$$\beta_0 + \beta_1 x$$

$$d_i = y_i - (\beta_0 + \beta_1 x_i)$$

See -
W. OM-18
for Motivation

Find β_0, β_1 such that $\sum_{i=1}^n d_i^2$ is minimized

Def: $h(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$

cf. -minimizing
of h

Calculus / otherwise - get a solution

$$\hat{\beta}_0 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

①

$$\hat{\beta}_1 \text{ (Slope)} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{--- (2)}$$

the "best fit line" passes through $(\bar{x}, \bar{y})$

Terminology

"best" fit line in the above procedure

{

Given β_0, β_1 then $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$e_i = y_i - \hat{y}_i$ (Residual Error)

Error: $\sum_{i=1}^n e_i^2$ Residual sum of squares

ϕ (can we estimate σ^2 ?)

Simple linear Regression - line Population