Simple linear Reg. model

X and Y

$$Y = \beta_0 + \beta_1 X.$$ ✓✓

— If the two (random) variables are probabilistically related, then for a fixed value of X, there is uncertainty in the value of the second.

$Y = \beta_0 + \beta_1 X + \varepsilon$ where ε is a random variable. ↳ there is random error

Y	X
4	1
10	3

(1.5) (3)

⇒ There are parameters β_0, β_1 and δ^2 such that for any fixed value of the independent variable X the dependent variable is a random variable related to X through the model equation

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

The quantity ε in the model equation is the error — a random variable assumed to be symmetrically distributed with

$$E(\varepsilon) = 0 \text{ and } Var(\varepsilon) = \delta_\varepsilon^2 = \delta^2$$

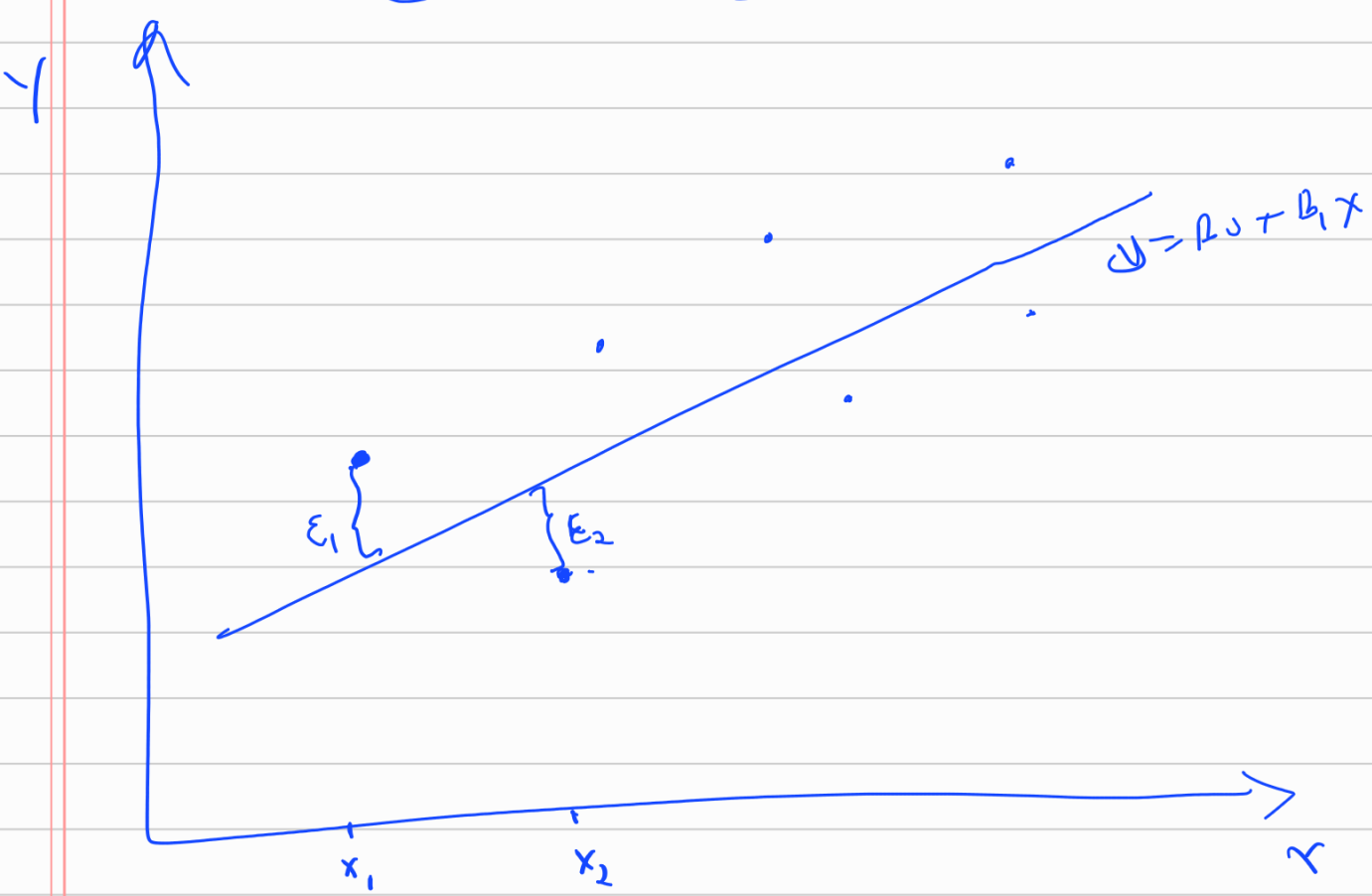
$$\varepsilon \sim N(0, \sigma^2) \simeq \varepsilon \sim N(0, 1).$$

X: the independent, predictor or explanatory variable → N'of random.

Y: Dependent / response variable. for fixed X, Y will be random variable.

ε: the random deviation or random error term for fixed x, ε will be random variable.

What exactly does ε do?



Constants → β_0 → Vertical intercept of the model

Constants → β_1 → Slope / Gradient

Estimating the model parameters

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{for } i = 1, 2, \dots, n.$$

Least method: the sum of squared vertical deviations from the points $(x_1, y_1), \dots (x_n, y_n)$ to the line such that

$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

- The pair estimates of β_0 and β_1, denoted by b_0 and b_1, are called the least squares estimates - they are those values that minimize

$$f(b_0, b_1)$$

$$\Rightarrow f'(b) = 0$$

$$\frac{\partial f(b_0, b_1)}{\partial b_0} = \sum 2(y_i - (b_0 + b_1 x_i))(-1) = 0$$

(1) $$\frac{\partial f(b_0, b_1)}{\partial b_0} = \sum 2(y_i - (b_0 + b_1 x_i))(-1) = 0$$

③

$$\frac{\partial f(b_0, b_1)}{\partial b_1} = \sum 2\left(y_i - (b_0 + b_1 x_i)\right)(-x_i) = 0$$

We need to solve

$$\sum 2\left(y_i - (b_0 + b_i x_i)\right)(-1) = 0 \quad \text{--- (i)}$$

$$\sum 2\left(y_i - (b_0 + b_i x_i)\right)(-x_i) = 0 \quad \text{--- (ii)}$$

$$\Rightarrow \sum \left(y_i - (b_0 + b_1 x_i)\right) = 0 \quad \text{--- (i)}$$

$$\Rightarrow \sum x_i \left(y_i - (b_0 + b_i x_i)\right) = 0 \quad \text{--- (ii)}$$

$$\Rightarrow \boxed{\begin{aligned} \sum y_i - n b_0 - b_i \sum x_i &= 0 \\ \sum x_i y_i - b_0 \sum x_i - b_i \sum x_i^2 &= 0 \end{aligned}}$$

$$b_1 = \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \checkmark$$

$$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$b_0 = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

In some

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{Cov(x,y)}{Var(x)}$$

$$= \frac{S_{xy}}{S_{xx}}$$

$$y = \beta_0 + \beta_1 x$$

$\hat{\beta}_0$ => remain.

$$Cov(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Regression assumptions

a) Linearity of the data. The relationship between the predictor (x) of the outcome (y) is assumed to be linear.

b) Normality of residuals: The residual errors are assumed to be normally distributed

$$\varepsilon \sim N(0, \delta^2)$$

c) Homogeneity of residual variance: The residuals are assumed to have a constant variance (homoscedasticity)

d) Independence of the residual errors

Hypothesis tests for linear models

- Hypothesis? A claim about pop parameters - Test the claim to establish whether it is true or not true.

Types

{ Null hypothesis (H_0) ✓ maintains the status quo
Alternative " (H_a, H_A, H_1) challenges

- Which one do we test?

H_0: ! ! ! ! ! ! !
1 2

Decision: We reject H_0 (false) / Fail to reject (True)

⇒ $H_0: \mu = \mu_0$ vs $H_a: \mu \neq \mu_0$ } Two-tailed ✓

⇒ $H_0: \mu = \mu_0$ vs $H_a: \mu > \mu_0$ } one tailed to right
$H_a: \mu < \mu_0$ } one tail to the left

↳ We fail to reject / We do not reject H_0.

We commit errors: How many errors do we commit while carrying test of hypothesis

⇒ (a) Type I error - Reject H_0 | H_0 true }

5) Type II error – Fail to reject H_0 | H_0 false

α = P(Type I error)

= P(Reject H_0 | H_0 true)

⇒ The significance level or size of the test

$\alpha = 0.05, \quad 0.02, \quad 0.01,$

⇒ Power of the test → $\beta = 0.25$

Power of the test = $1 - \beta = 99.25$

0.25 →

linear model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_{(p-1)} X_{(p-1)i} + \varepsilon_i$$

Where $\varepsilon_i \sim N(0, \delta^2)$

Test on a single parameter

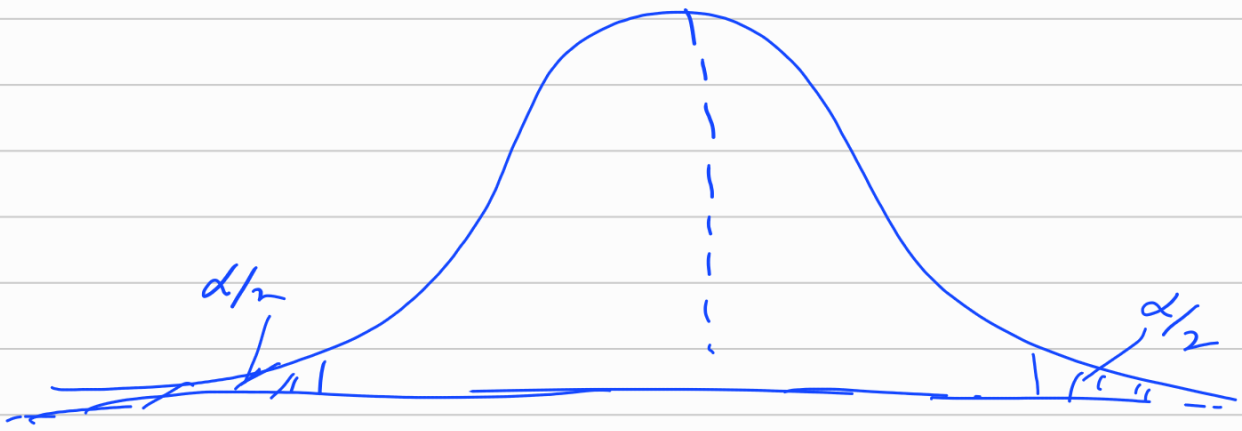
- $H_0: \beta_j = b$ vs $H_1: \beta_j \neq b$
- Significance level $100\alpha\%$, b is calculated $\alpha = 0.05$ or 0.01 or 0.02.

$$|T| = \left| \frac{\hat{\beta}_j - b}{SE(\beta_j)} \right| > t_{(n-p), \frac{\alpha}{2}}$$

$\therefore SE(\vec{\beta_j}) = \sqrt{Var(\vec{\beta_j})}$ → Standard error

$t_{(n-p),\frac{\alpha}{2}}$ → Two-tailed test.

↳ Degrees of freedom



Test Statistics

$H_0: \beta_j = 0 \quad vs \quad H_1: \beta_j \neq 0$

$$|T| = \left| \frac{\vec{\beta_j}}{SE(\vec{\beta_j})} \right| \sim t_{(n-p)}$$

a $100(1-\alpha)\%$ Confidence Interval for β_j

$$\left(\vec{\beta_j} - t_{(n-p),\frac{\alpha}{2}} SE(\vec{\beta_j}),\ \vec{\beta_j} + t_{(n-p),\frac{\alpha}{2}} SE(\vec{\beta_j}) \right)$$

( )

Constant = B_0.

- $H_0: B_0 = 0$ vs $B_0 \neq 0$.
- $\alpha = 0.05$ or $\frac{1}{20}$

g) P-value:
Reject H_0: when p-value is less than α (level of significance).

P-value = 0.000

Decision:

Conclusion:

- Confidence interval

Test for the existence of regression model