

REGRESSION ANALYSIS

Linear model

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \quad y = \beta x$$

$$\text{ex. } y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_k x^k$$

Nonlinear:

$$y = f(x, \theta)$$

Parameter estimation

True model:

$$\vec{y} = \beta_0 + \beta_1 \vec{x} + \varepsilon$$

True parameters:

$$\beta \quad \sigma_\varepsilon \quad \sigma_\varepsilon^2$$

Regression:

$$\vec{y} = b_0 + b_1 \vec{x}$$

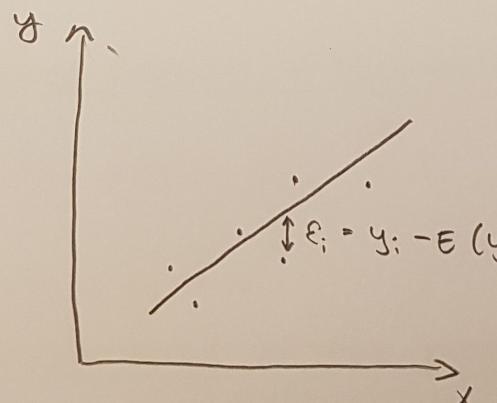
Estimated:

$$b \quad s_\varepsilon \quad s_\varepsilon^2$$

ex: flipping a coin - do I get H or T?

true value: 50%

experiment: maybe 2 T, 8 H in 10 experiments bc I've done very few experiments



$$SSE = \sum_{i=1}^n \varepsilon_i^2$$

$$SE = \sqrt{\frac{SSE}{n-p}}$$

regression line: $\left\{ \begin{array}{l} SE = 0 \\ \frac{\partial SSE}{\partial b_i} = 0 \end{array} \right.$

$$S_\varepsilon^2 = \frac{SSE}{n-p}$$

p = # parameters

n = # points

$$\vec{b} = (\bar{x}' \bar{x})^{-1} \bar{x}' \vec{y}$$

\bar{x} = matrix

$$SSE = (\vec{y} - \bar{x}\vec{b})' (\vec{y} - \bar{x}\vec{b})$$

x	y
20	66,1
40	66,12
60	66,21
80	66,18
100	66,3

$$\vec{y} = \begin{pmatrix} 66,1 \\ 66,12 \\ 66,21 \\ 66,18 \\ 66,3 \end{pmatrix} \quad \bar{x} = \begin{pmatrix} b_0 & b_1 \\ 1 & 20 \\ 1 & 40 \\ 1 & 60 \\ 1 & 80 \\ 1 & 100 \end{pmatrix}$$

$$\Rightarrow \vec{b} = \begin{pmatrix} 66,044 \\ 0,0023 \end{pmatrix} \leftarrow b_0 \quad (\bar{x}' \bar{x})^{-1} = \begin{pmatrix} 1,1 & -0,015 \\ -0,015 & 0,002 \end{pmatrix} \leftarrow b_1$$

$$SSE = 0,0041 \quad S_{\epsilon}^2 = \frac{0,0041}{5-2} = 0,0014$$

Estimation of confidence intervals & bands

Confidence interval of b_i (example with known σ_{ϵ})

The model: $\vec{y} = \beta_0 + \beta_1 x + \epsilon$

Regression: $\vec{y} = b_0 + b_1 x$

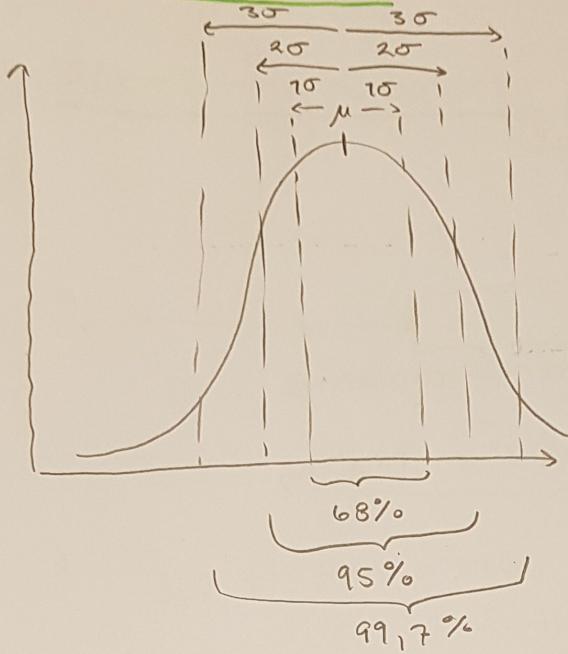
$$\beta_0 = 0,3, \quad \beta_1 = 0,7, \quad \sigma_{\epsilon} = 0,1$$

$$\boxed{\sigma_{b_j}^2 = \sigma_{\epsilon}^2 \left\{ (x' x)^{-1} \right\}_{jj}}$$

$$x. \quad S_{b_0}^2 = 0,0014 \cdot 1,1$$

$$S_{b_1}^2 = 0,0014 \cdot 0,002$$

68-, 95-, 99,7% rule



$$b_0 = 0,29967 \pm 2 \cdot s_{b_0}$$

$\underbrace{\phantom{0,29967 \pm 2 \cdot s_{b_0}}}_{95\%}$

Confidence interval of b_i (example with unknown σ_ε)

$$b_1 = \underbrace{0,91723 \pm 2 \cdot 0,16}_{\text{not including the real value } (\beta_1 = 0,7)}$$

it's because the s_i has a lot of uncertainty



we need to use t-distribution!

use the t-value (from table) instead of the "2"
when calculating b_0, b_1, \dots

⇒ the confidence interval is

$$\boxed{b_i \pm t_{(n-p, \alpha/2)} \cdot s_{b_i}}$$

Confidence interval of the mean

Variance of the mean: $s_{\hat{y}(x_0)}^2 = s_\varepsilon^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$

→ 0 if $n \rightarrow \infty$

Confidence interval of the prediction

$$s_{\hat{y}(x_0)}^2 = s_\varepsilon^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

does not → 0 if $n \rightarrow \infty$

Pointwise conf. interval

$$y \pm t_{(n-p, \alpha/2)} s_y$$

Simultaneous conf. interval

$$y \pm \sqrt{P F_{(p, n-p, \alpha)}} \cdot s_y$$

Identify and correct problems

- See "handout of regression analysis" on PingPong

model assessment

Assumptions of regression analysis:

1. The mean of the error is 0.
2. Errors are normally distributed
3. The variance of the error σ^2_ϵ is constant
4. Errors are independent (uncorrelated)

} has to be satisfied
for the regression
analysis to have
a meaning