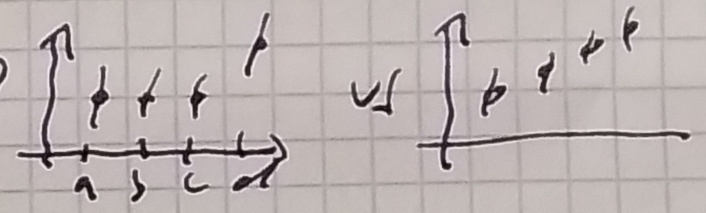
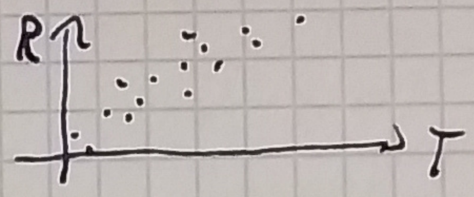
10. Linear Regression

H_0 fails in ANOVA – what now? 

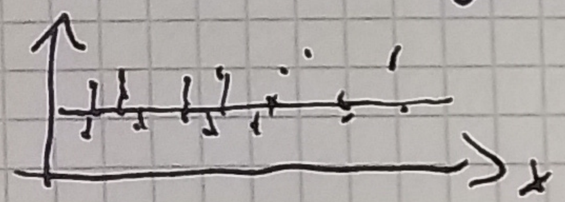
We have (y_i, x_i) instead of groups 

⇒ Regression.

10.1 Constant Function

We have data (y_i, x_i) and want to fit it with a function $g(x) = a$ ← a constant

MODEL:



Our "estimator" for $\hat{y} = \hat{a}$

we define the distance to the estimator $r_i = y_i - \hat{y}(x_i)$

residual ↑

$= y_i - \hat{a}$

minimise square deviation from model:

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_i (y_i - \hat{a})^2$$

$$\frac{\partial SS_E(a)}{\partial a} = \sum_i 2(y_i - a) \cdot (-1) \stackrel{!}{=} 0$$

$$= -2 \sum_i y_i + 2na \stackrel{!}{=} 0$$

$$= -2n\bar{y} + 2na \stackrel{!}{=} 0 \Rightarrow \hat{a} = \bar{y}$$

$\hat{a}$ is an extremum check $\frac{\partial^2 SS_E(a)}{\partial a^2} = +2n > 0$ always...

$\Rightarrow \hat{a} = \bar{y}$ minimizes the residual $\Rightarrow$ "best fit"

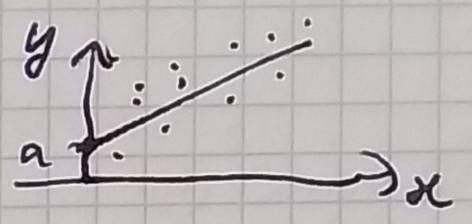
→ property of the arithmetic mean.

10.2 Linear function

$\hat{y} = a + bx$ → find $\hat{a}$ and $\hat{b}$ that minimize the residual^2 / SS_E

intercept, slope

$r_i = y_i - (\hat{a} + \hat{b} x_i)$



$$\frac{\partial SS_E(a,b)}{\partial a} = \frac{\partial}{\partial a}\left(\sum_i (y_i - a - bx_i)^2\right)$$

$$= 2\sum_i (y_i - a - bx_i)(-1) \stackrel{!}{=} 0$$

$$-2(n\bar{y} - na - nb\bar{x}) \stackrel{!}{=} 0$$

$$\boxed{\hat{a} = \bar{y} - \hat{b}\bar{x}} \quad (I) \qquad \text{check } \frac{\partial^2 SS_E}{\partial a^2} = 2n > 0$$

$$\frac{\partial SS_E}{\partial b} = 2\sum (y_i - a - bx_i)(-x_i) \stackrel{!}{=} 0$$

$$\Rightarrow -\sum (x_i y_i) - na\bar{x} - b\sum x_i^2 \stackrel{!}{=} 0$$

$(\bar{x}, \bar{y})$

$\sum \Delta_i \cdot (>0) = >0$

$\rightarrow <0$

$$(I) \Rightarrow \sum x_i y_i - n(\bar{y} - b\bar{x})\bar{x} - b\sum x_i^2 \stackrel{!}{=} 0$$

$$\Rightarrow \boxed{\hat{b} = \frac{\sum (x_i y_i) - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}} \quad \text{w/o } SS_T$$

check: $\frac{\partial^2 SS_E}{\partial b} = 2\sum x_i^2 \geq 0$

10.3 ANOVA for (one-slope) linear regression:

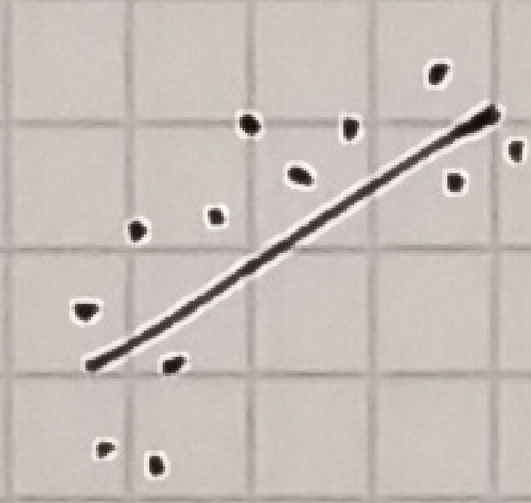
	ν	SS	MS	F
Model	1	$SS_M = \sum (\hat{y}(x_i) - \bar{y})^2$	SS_M/ν_M	MS_M/MS_E
Error/ Residuum	$n-2$	$SS_E = \sum (y_i - \hat{y}(x_i))^2$	SS_E/ν_E	
Total	$n-1$	$SS_T = \sum (y_i - \bar{y})^2$		

F test: if $F_{measured} > F_{\nu_M, \nu_E}(P)$ ← e.g. 95%

⇒ reject H_0: "y data does not depend on x"

(all x "groups" are the same)

⇒ conclude there is some correlation.

Goodness of fit  vs 

$$R^2 = \frac{SS_M}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

$0 \leq R^2 \leq 1$

(0 ← meaningless, 1 ← perfect fit)

→ Slides

10.4 What is "linear"?

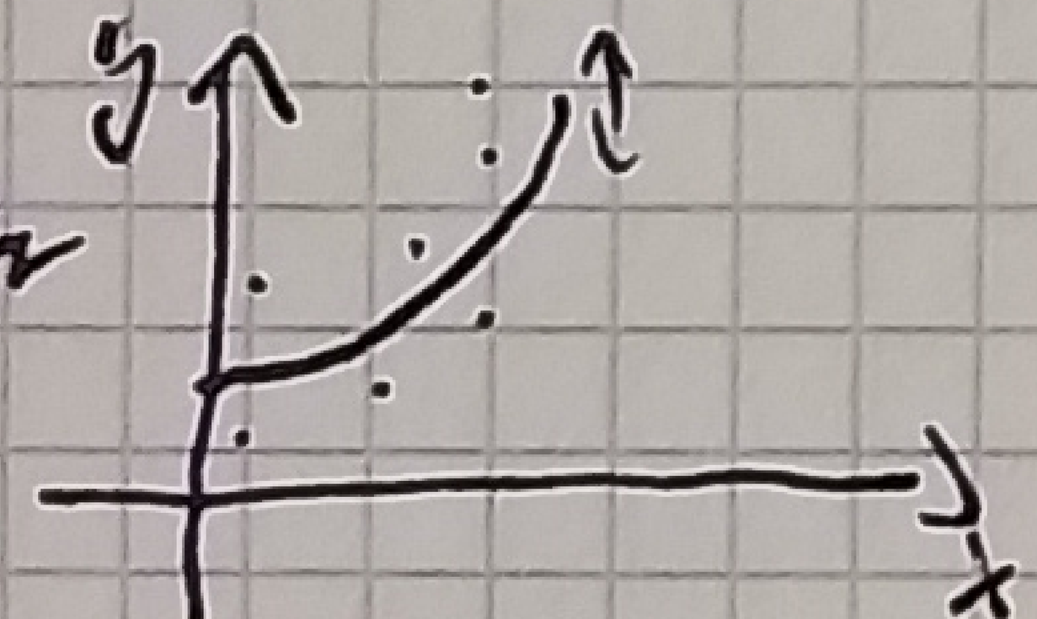
→ the parameters a, b.

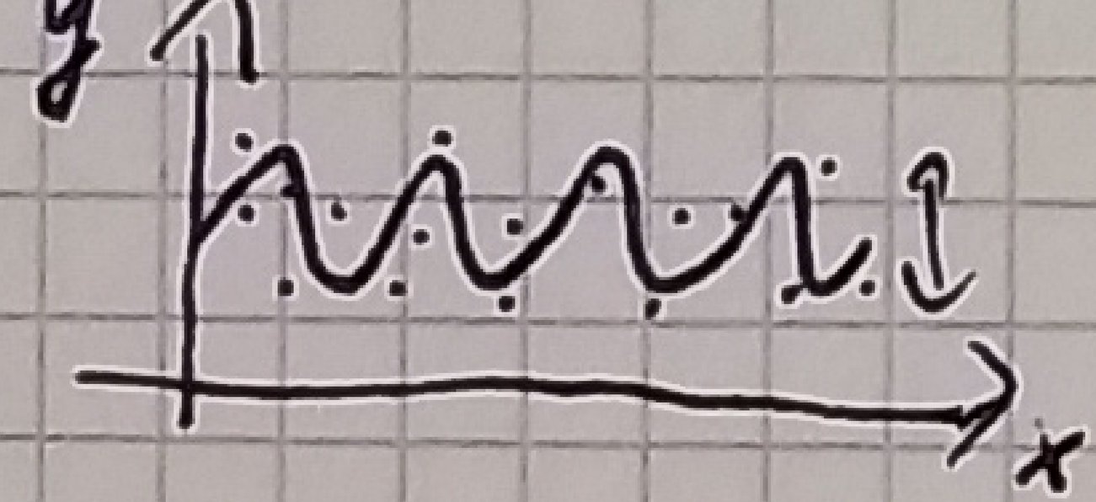
$y = a + bx$ ✓ $y = a + b\underbrace{x^2}_{z}$ ✓ $y = a + b\underbrace{\sin(x)}_{z}$ ✓

$y = a + \sin(bx)$ ✗ → Taylor expan $y = a + b_1 x + b_2 x + \dots$

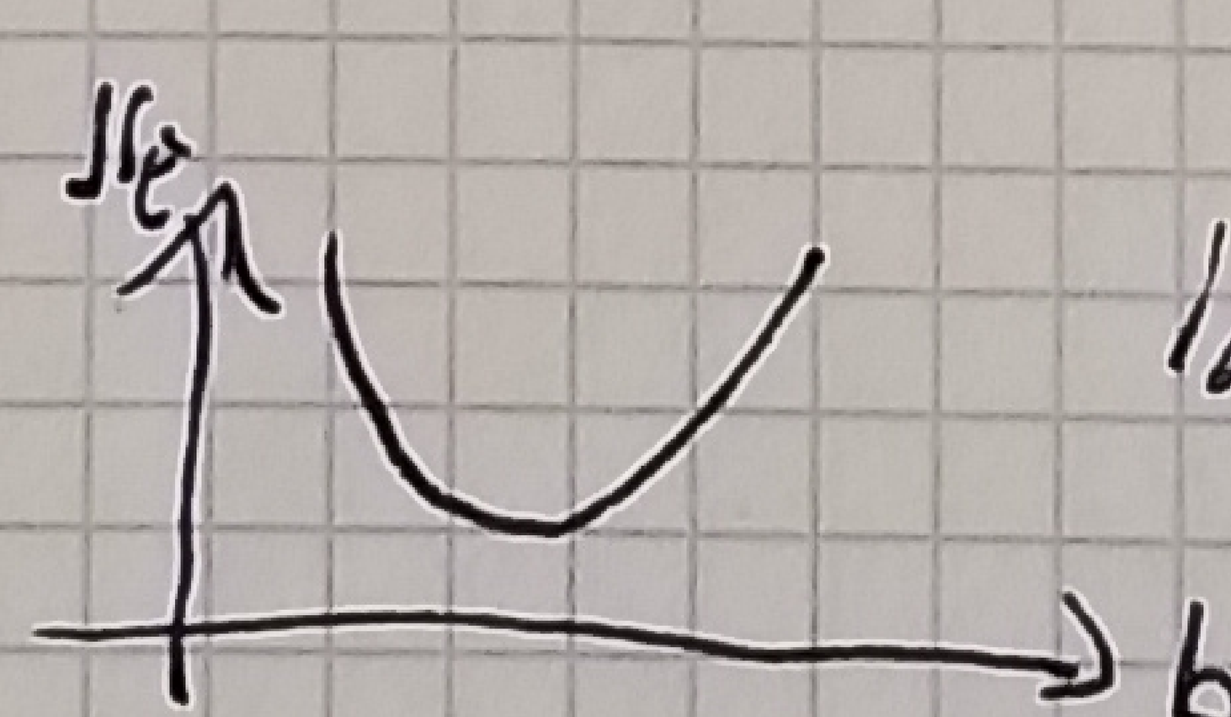
→ more parameters.

linear





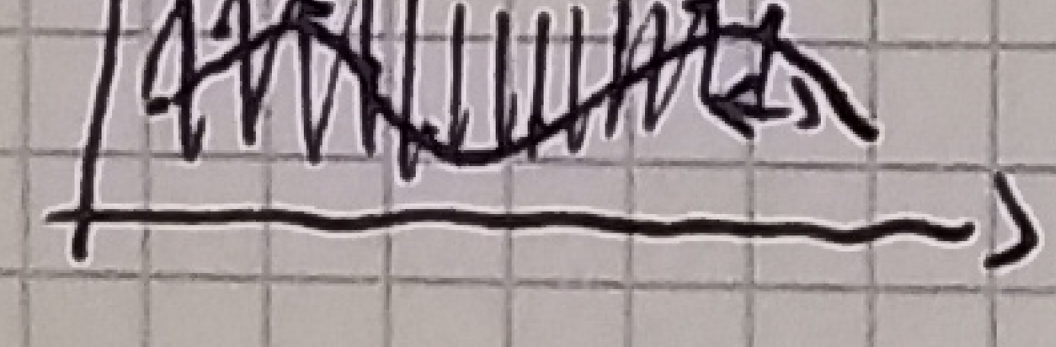
then SS_E

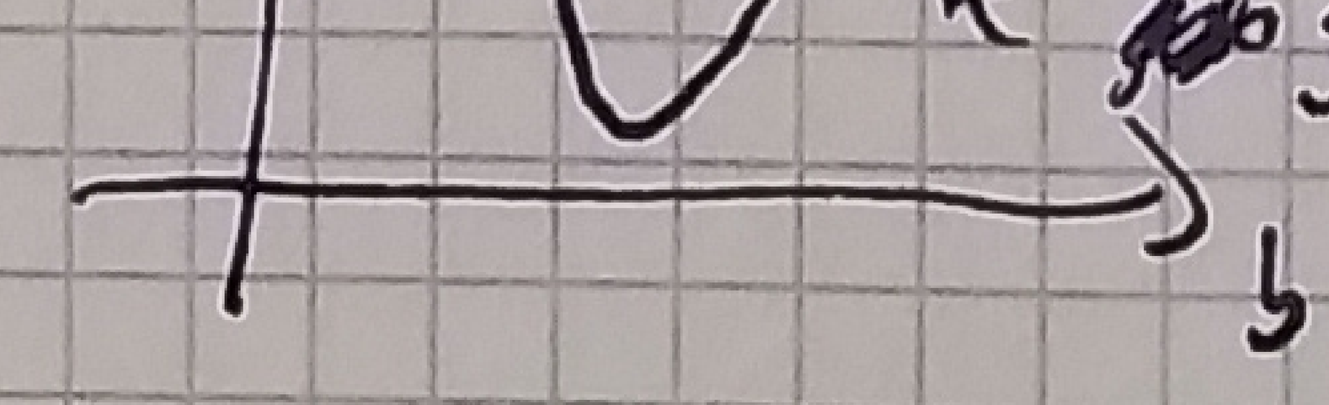


local optimum = global optimum.

→ deterministic

non-linear





get stuck...?