Statistisk modellering og regressionsanalyse

Claus Thorn Ekstrøm
KU Biostatistik
ekstrom@sund.ku.dk

Marts 18, 2019

Slides @ biostatistics.dk/talks/


166
2497
cph
AU 82 692

Hvad er statistik?

> *Statistics is a science, not a branch of mathematics, but uses mathematical models as an essential tool.*

-- John Tukey

og

> *[…] data analysis is **detective work** - numerical detective work — or counting detective work — or graphical detective work.*

-- John Tukey, 1977

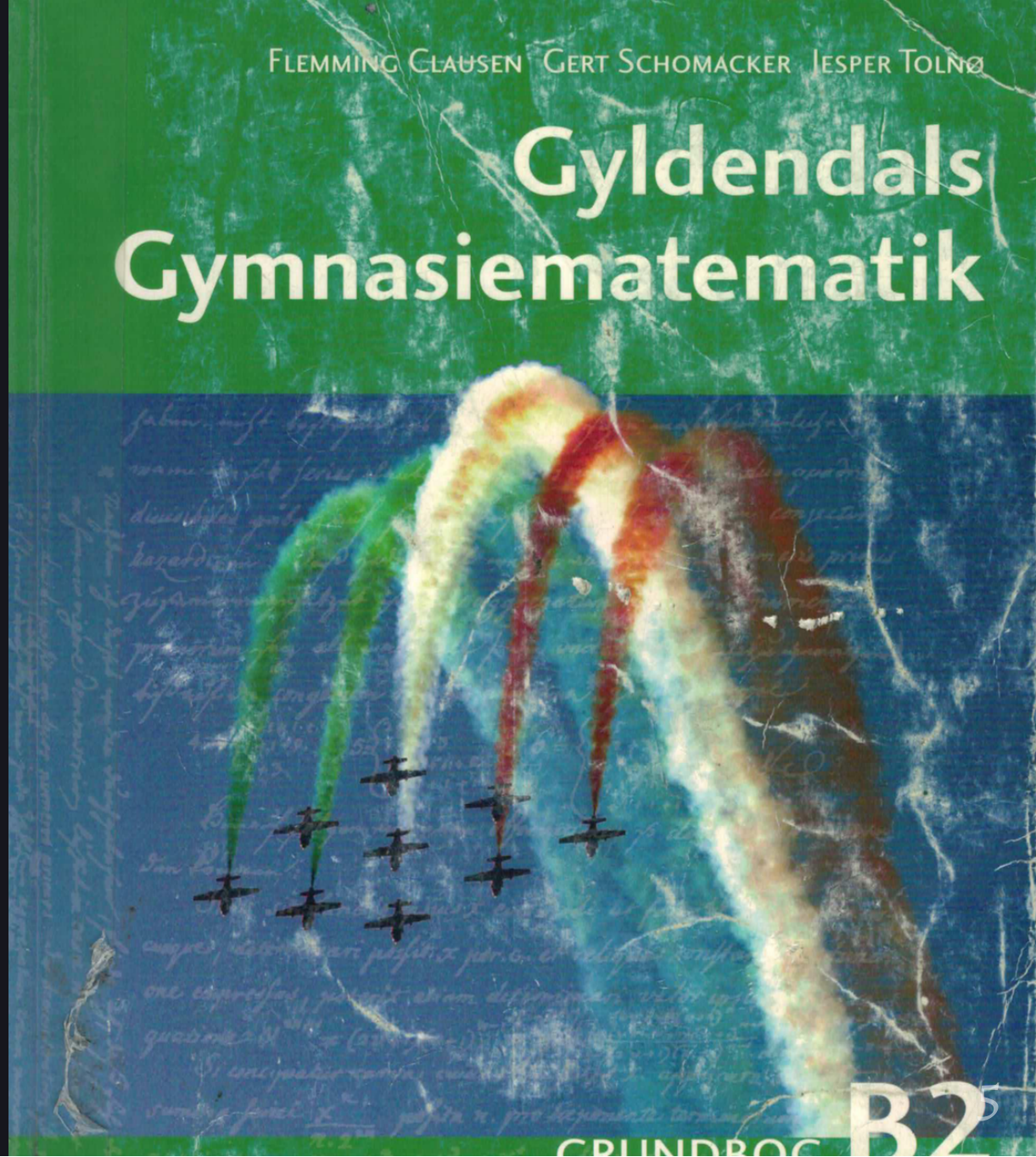
Hvad bruger vi statistik til?

- **Mønstre**.
  Hvad ser vi?
- **Prædiktion**.
  Hvad forventer vi ved ny observation?
- **Kausalitet**.
  Hvorfor?

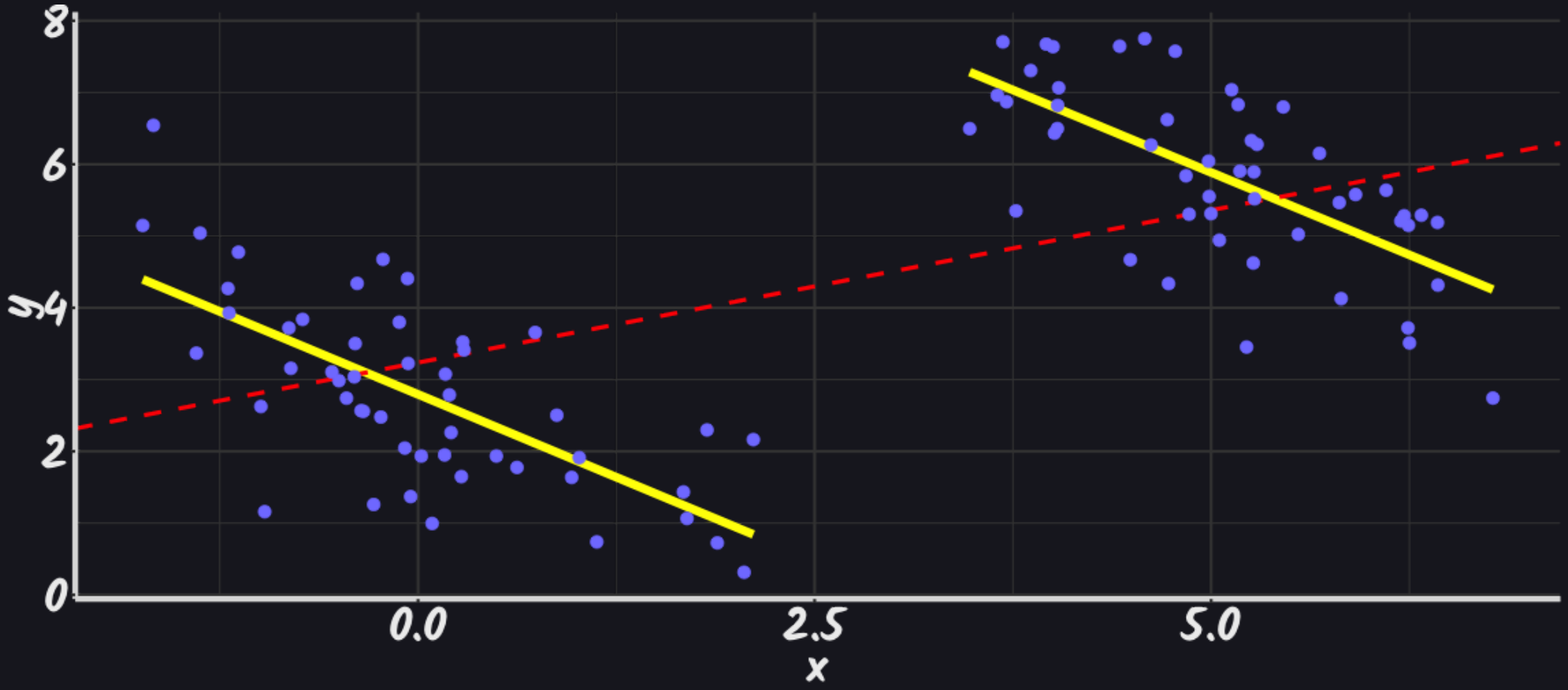
"Konklusionen er tvivlsom selv om graferne viser en tydelig sammenhæng"

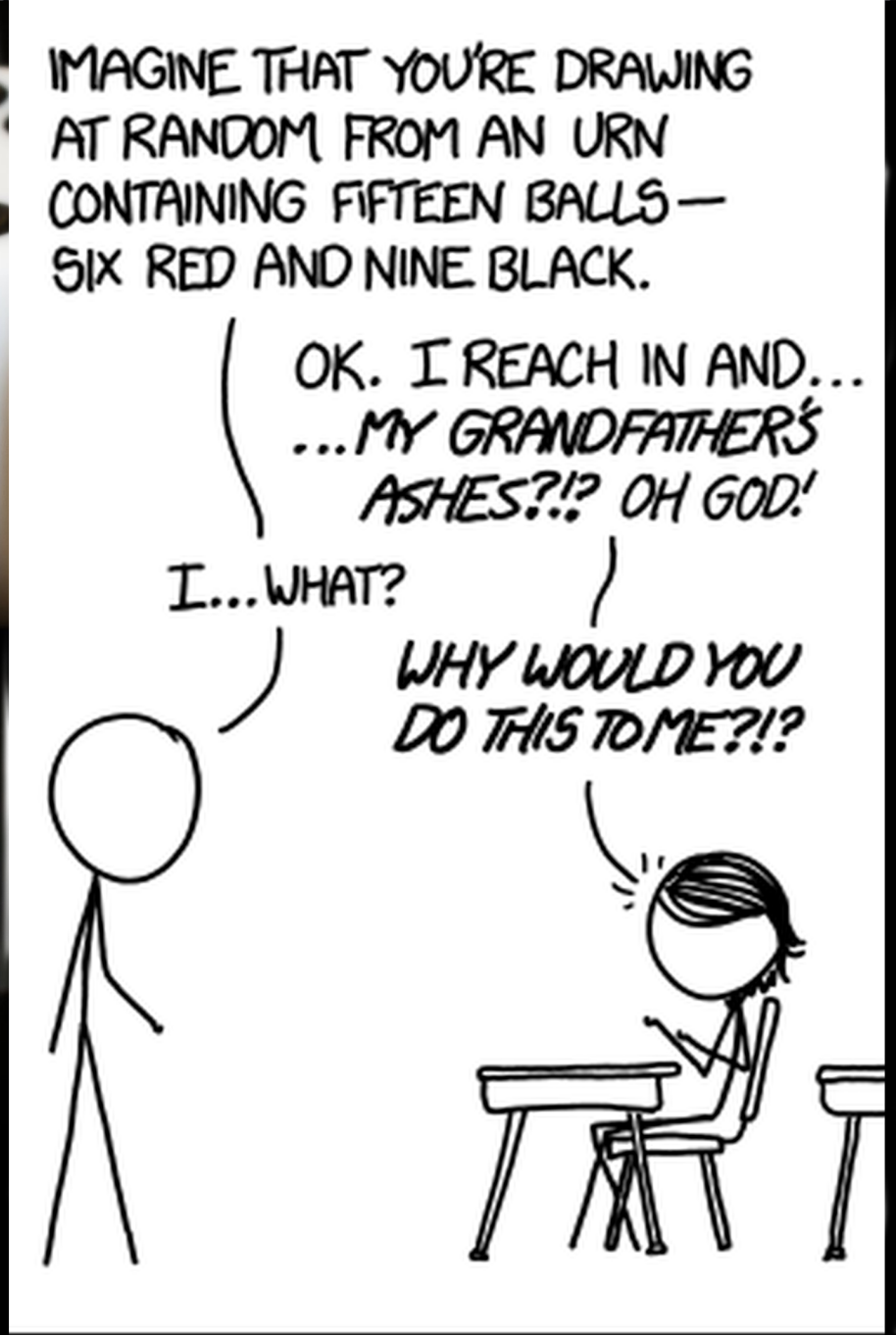
"Når man beder en hjælp fra en ekspert, får man ofte mere at vide, end man ønsker eller kan forstå"

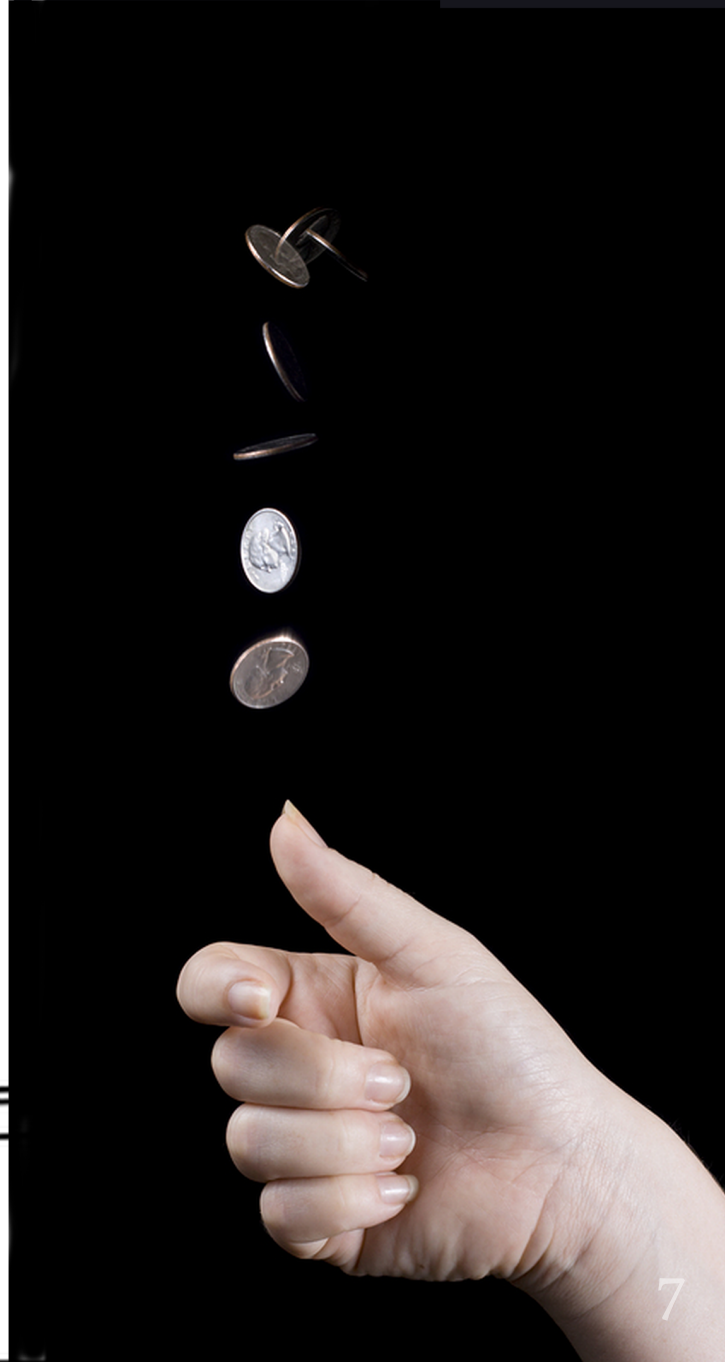
"Der er ikke fejl i udregningerne. Resultatet strider mod enhver sund fornuft, men det er ikke desto mindre rigtigt udregnet, ..."



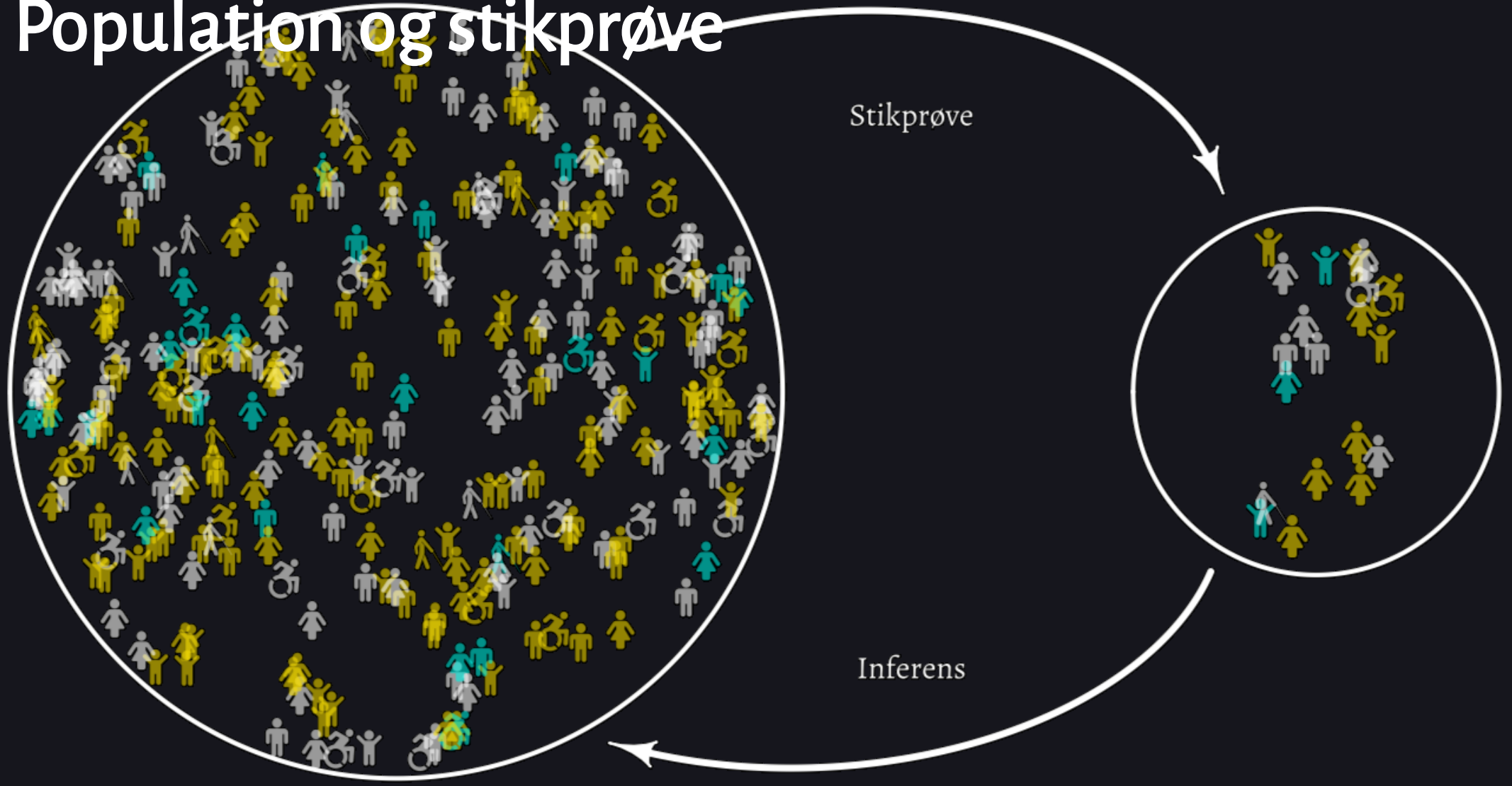
"Strider mod enhver sund fornuft"




IMAGINE THAT YOU'RE DRAWING AT RANDOM FROM AN URN CONTAINING FIFTEEN BALLS— SIX RED AND NINE BLACK.
OK. I REACH IN AND...
...MY GRANDFATHER'S ASHES?!? OH GOD!
I...WHAT?
WHY WOULD YOU DO THIS TO ME?!?

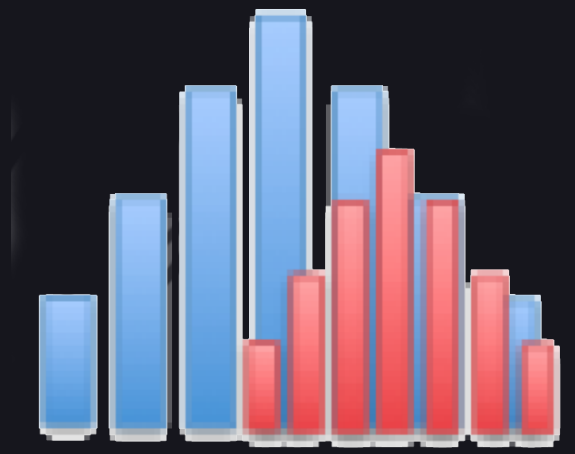


Population og stikprøve



Data-genererende proces

Sandsynlighedsregning
Givet model, forudsig data
Deduktiv



Statistik
Givet data og modelklasse - find optimale parametre
Induktiv

En model er en klasse af funktioner

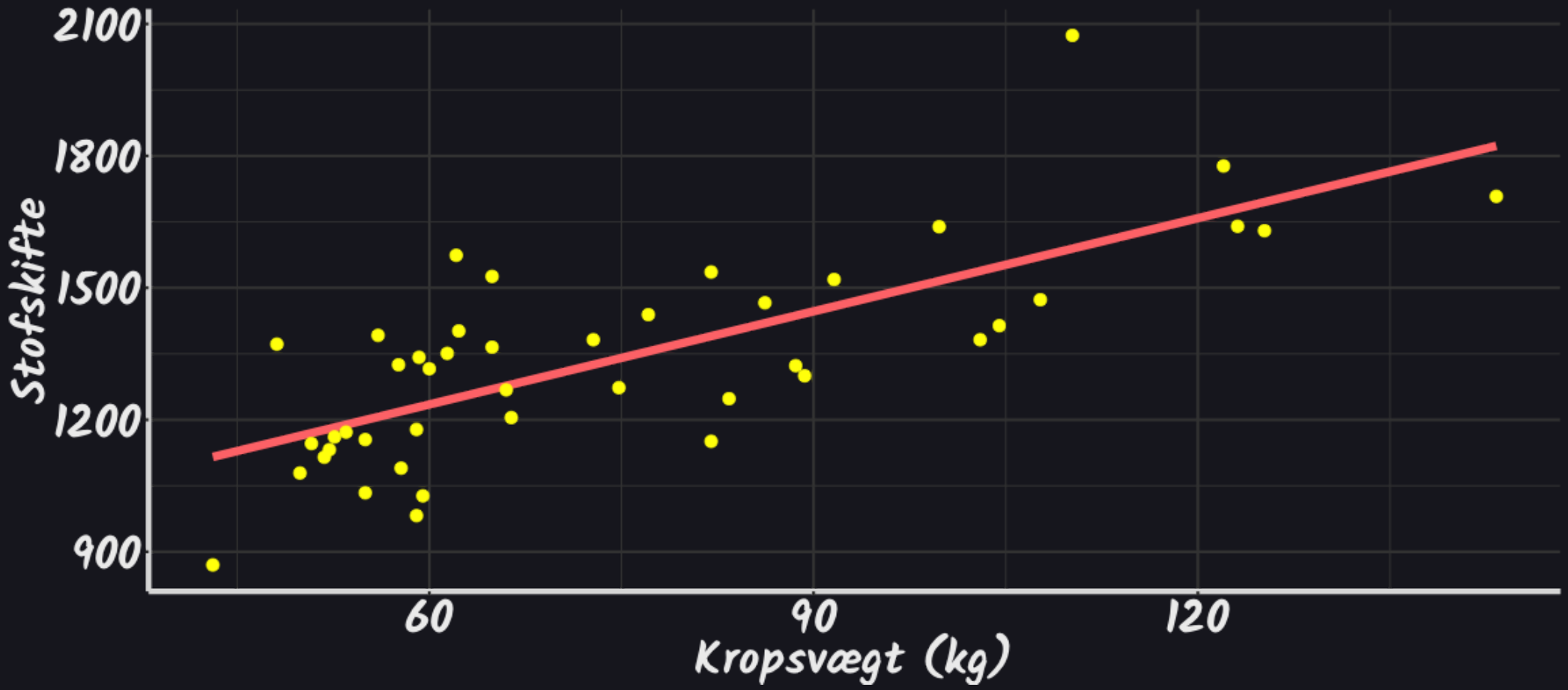
- Y_i - udfaldet
- $f_\beta(x_i)$ - den forventede værdi

$$\text{Model fx: } Y_i = \underbrace{\beta_0 + \beta_1 \cdot x_i}_{f_\beta(x_i)} + \underbrace{\varepsilon_i}_{\text{Støj med middelværdi } 0}$$

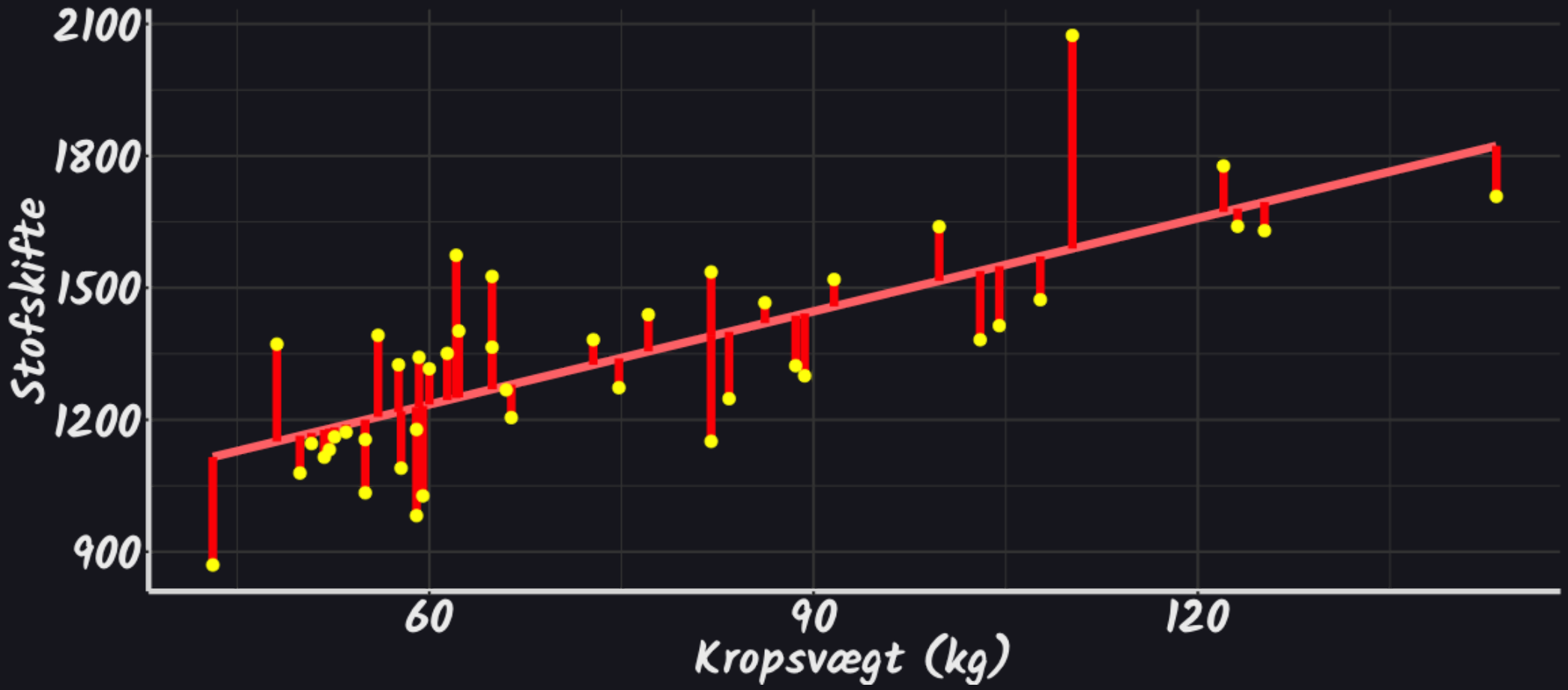
Interessante spørgsmål:

- Hvilke værdier af β vil gøre data mest sandsynlige (for fast f)?
- Hvordan vælger man f?

Hvilende stofskifte og kropsvægt



Residualer



Mindste kvadraters metode

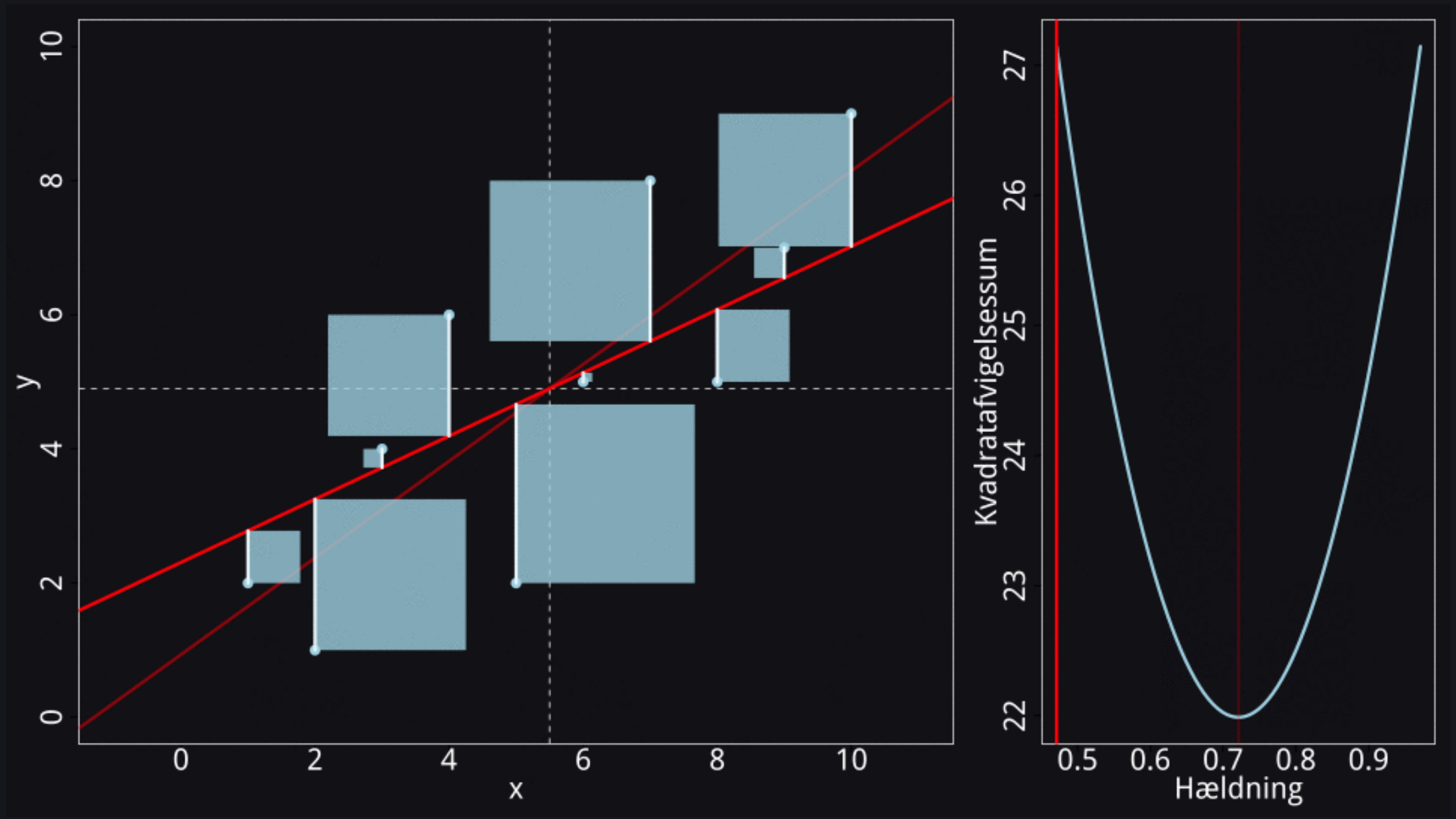
Minimér de kvadratiske residualer

$$r_i = y_i - f_\beta(x_i)$$

$$\arg\min \sum_{i=1}^{N} (y_i - f_\beta(x_i))^2$$

Hvor god er modellen? Prædiktionen?

Observeret - forventet


y
x
0
2
4
6
8
10
Kvadratafvigelsessum
22
23
24
25
26
27
0.5
0.6
0.7
0.8
0.9
Hældning

Fortolkning for lineær regression

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i \quad \mathbb{E}(\varepsilon_i) = 0, \mathbb{V}(\varepsilon_i) = \sigma^2$$

- $\hat{\beta}_0$ skæring
- $\hat{\beta}_1$ hældning
- $\hat{\sigma}$ spredning af residualerne

β_1 er den interessante parameter. Gennemsnitlig *relevante* effekt. β_0 kun sjældent relevant.

σ - hvor tæt er modellen og observationerne på hinanden

Spredningen

$\hat{\sigma}$ er kvadratroden af gennemsnittet af de kvadrerede afvigelser

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - f_\beta(x_i))^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N} r_i^2}{N}}$$

Men dette er i forhold til den *sande* model, hvor β er kendt.

Spredningen

Hvis $\hat{y}_i$ er den estimerede værdi for en måling i ud fra en statistisk model, så er residualspredningen

$$s = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N-2}} = \sqrt{\frac{\sum_{i=1}^{N}\hat{r}_i^2}{N-2}}$$

hvor $\hat{r}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$.

Løst sagt: "Gennemsnitlig" afvigelse mellem observationerne og modellen.

Spredningen (the dirty business)

LS estimaterne kan skrives som $\hat{\beta} = (X^\intercal X)^{-1} X^\intercal y$

Sæt $H = \underbrace{X(X^\intercal X)^{-1} X^\intercal}_{\text{Projektionen på søjlerne udspændt af X}}$. Så er

$$\begin{aligned}
\mathbb{V}(y - \hat{y}) = \mathbb{V}(y - X\hat{\beta}) &= \mathbb{V}((I - H)y) \\
&= (I - H)\mathbb{V}(y)(I - H)^\intercal \\
&= \sigma^2 (I - H)^2 \\
&= \sigma^2 (I - H)
\end{aligned}$$

$(I - H)$ idempotent så rank=sporet, som er $n - p$.

Hvad fortæller en statistisk model?

Hvis modellen er rimelig så fortæller den noget, om den *overordnede sammenhæng* i data.

Fortæller ikke nødvendigvis noget om *årsagssammenhæng*. Det kan man nogle gange ud fra designet (RCT) og modellen (kausal inferens).

Lineær regression fungerer på samme måde uanset om X er fast eller X observeret.

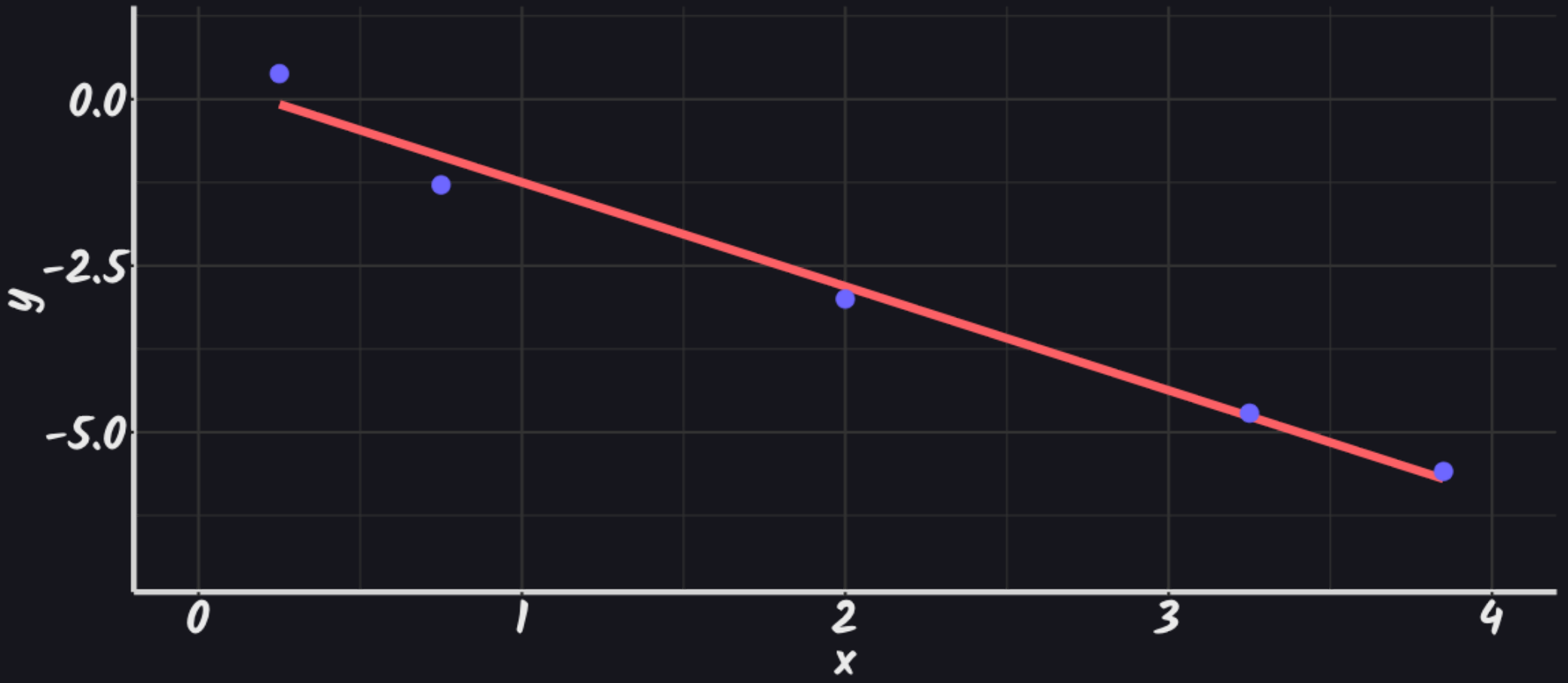
Hvornår kan man løbe ind i problemer?

- Hvis modellen er markant forkert.
- x på y eller y på x?
- Outliers.
- Interpolation og ekstrapolation.
- Fortolkning af skæringen.
- Sammenhæng vs årsagssammenhæng.

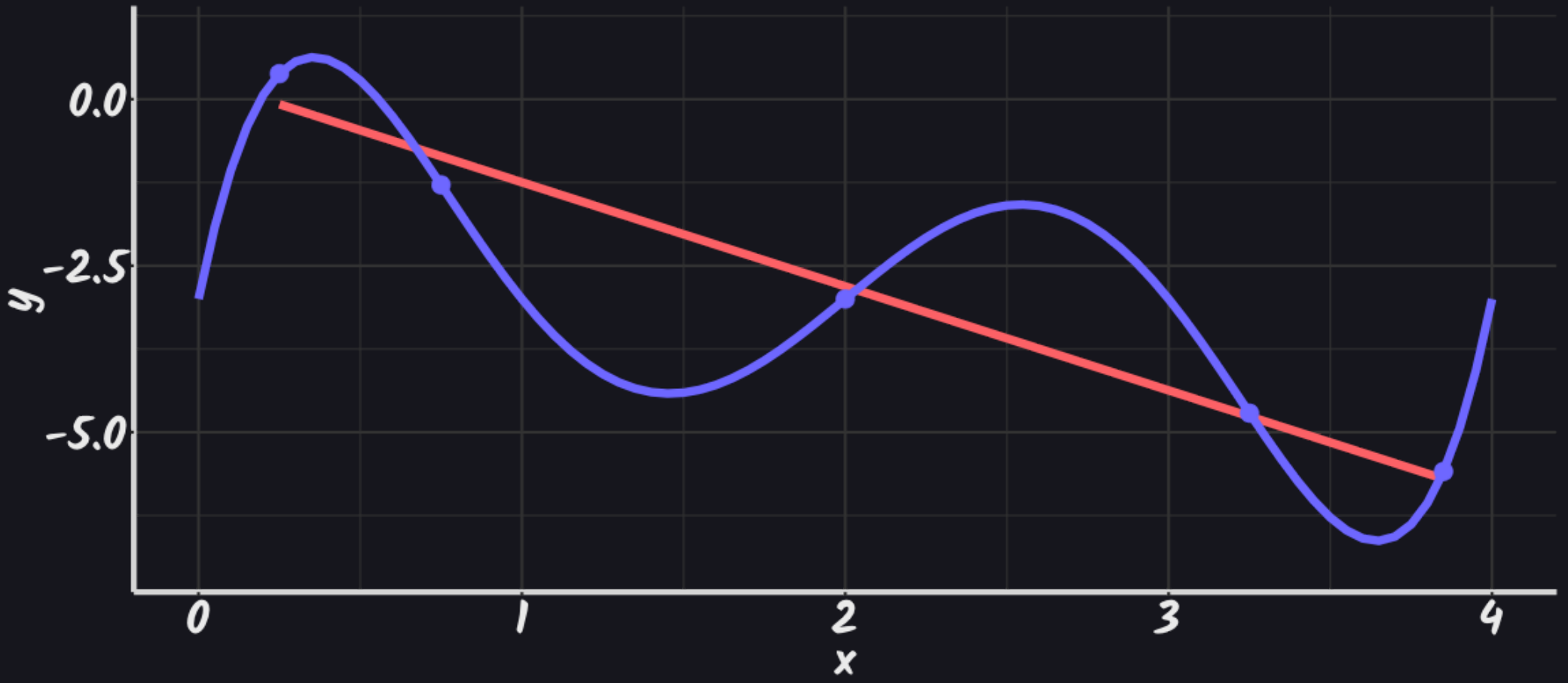
Hvornår er en model god?

Hvornår er den rigtig?

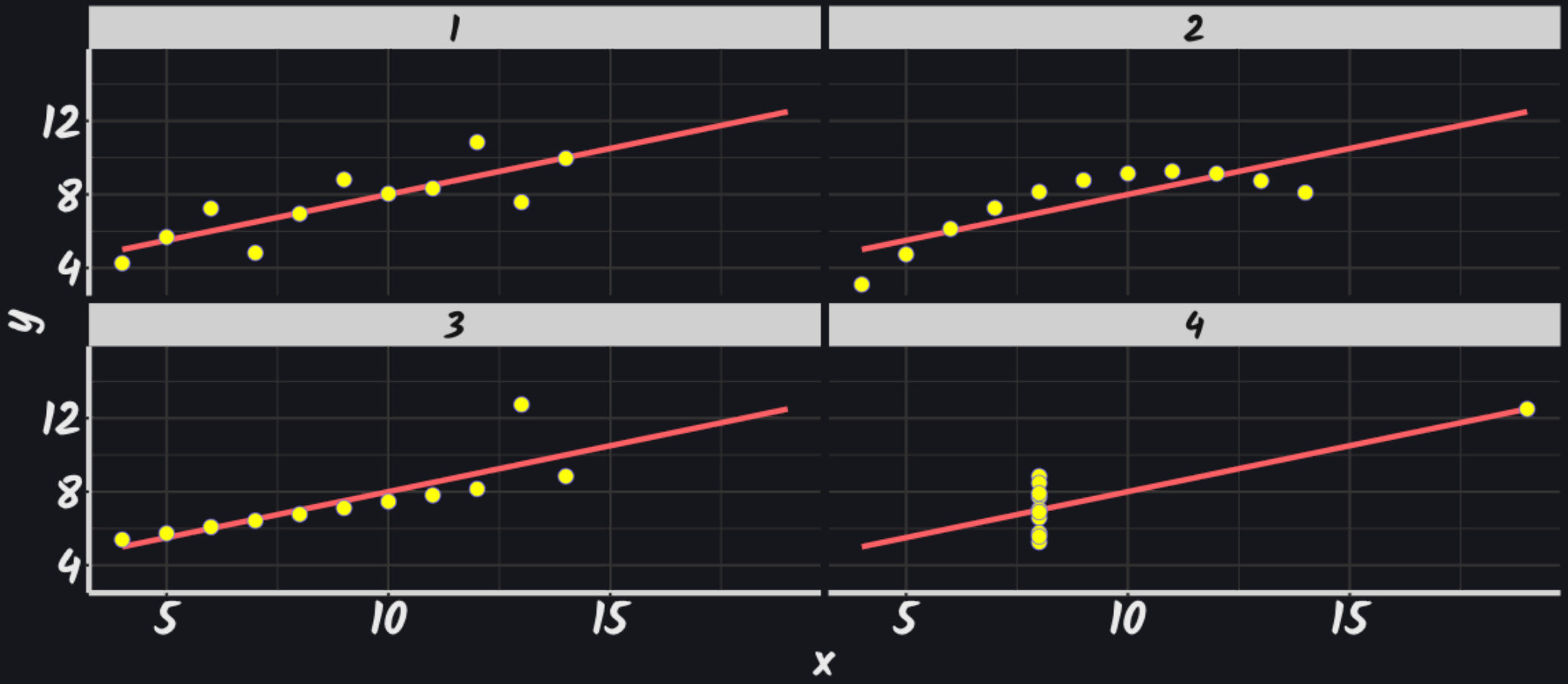
Er lineær regression en god model her?



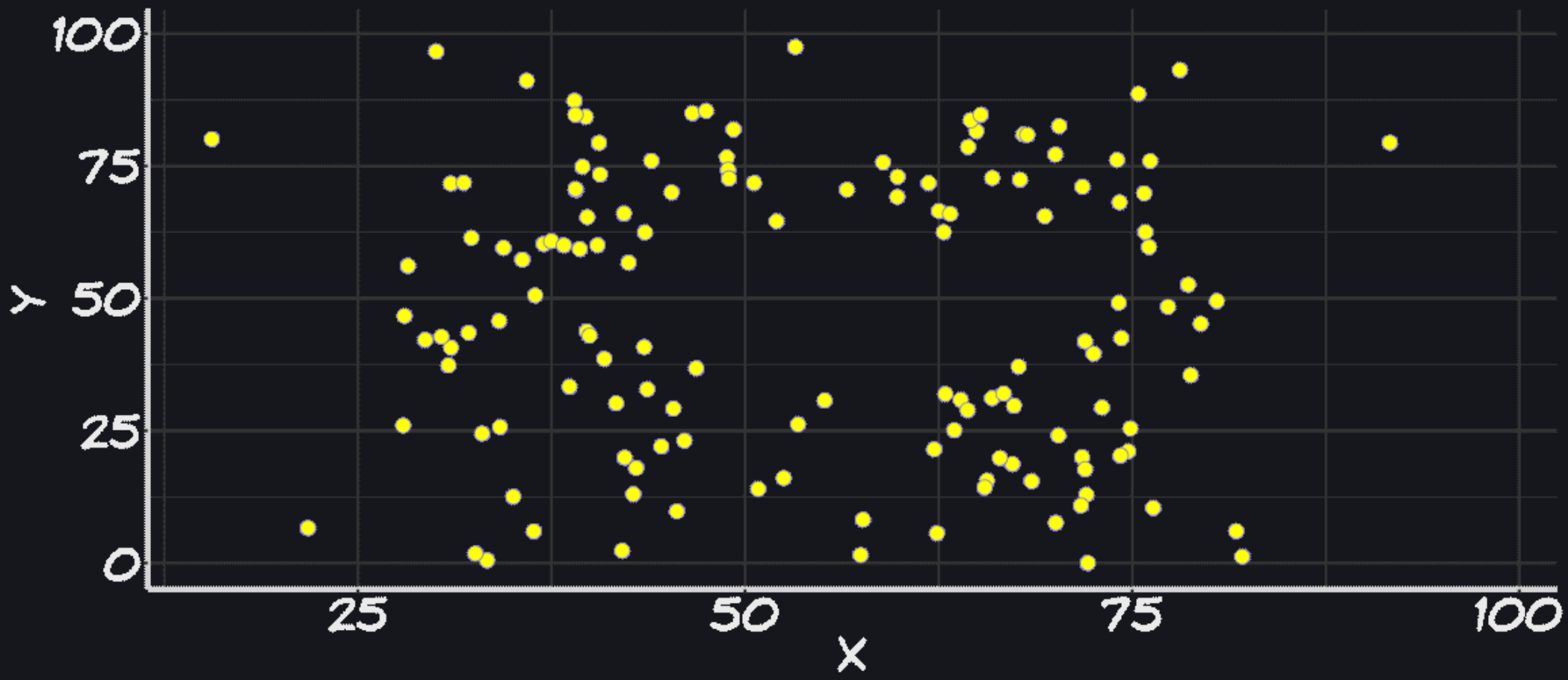
Er lineær regression en god model her?



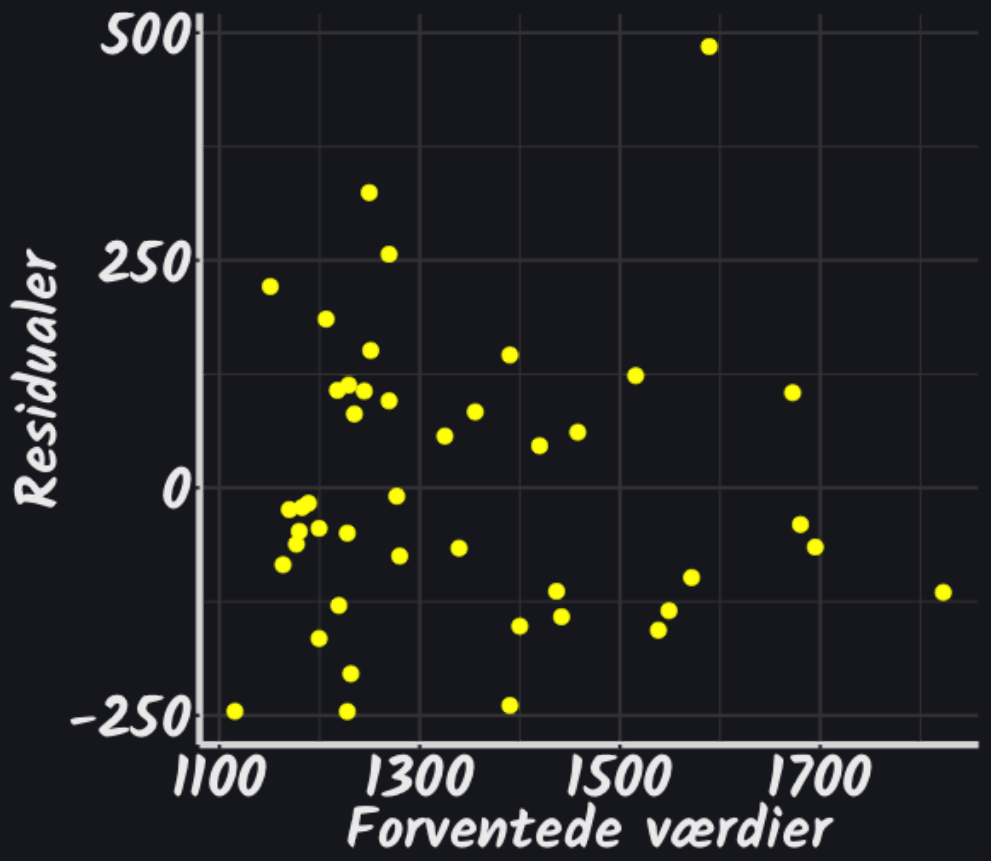
Anscombes data



Datasaurus



Residualplot - er antagelserne opfyldte?

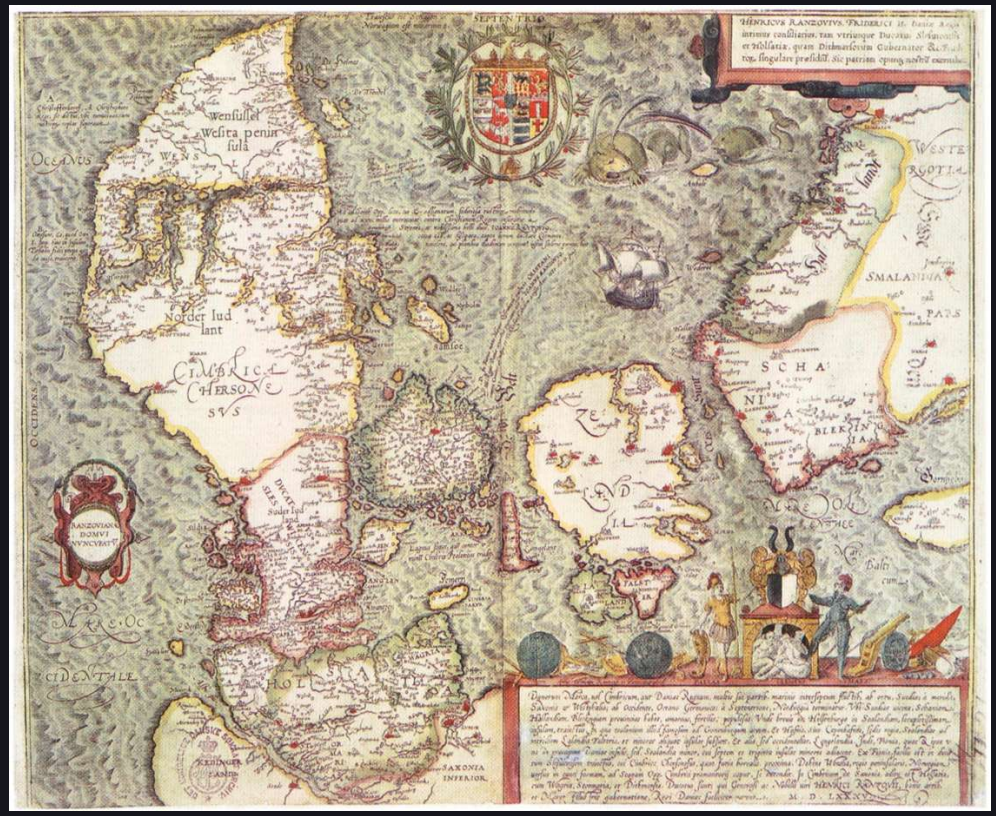


1. Middelværdi ca. 0
2. Ingen systematiske afvigelser
3. Check for outliers
4. Varianshomogenitet
5. (Uafhængighed - kan ikke nødvendigvis ses)

Er der så ikke noget rigtigt svar?

Næh. Eller jo … Viden tilegnes ved, at en påstand stadig holder vand efter gentagne efterprøvninger.

> *All models are wrong but some are useful*-- George Box



Brugen af R^2 i gymnasiet

Per Bruun Brockhoff, DTU Compute, Ernst Hansen, KU Matematik og Claus Thorn Ekstrøm, KU Biostatistik

Der lader til at være en vis forvirring blandt og uenighed mellem forskellige faggrupper omkring R^2–værdien, også kaldet "forklaringsgraden" eller "determinationskoefficienten". Uenigheden omkring brugen og nytten af R^2 som et mål til at bekrive en statistisk model optræder ikke kun i gymnasiet: globalt set skaber brugen af R^2 tilsvarende gnidninger. Den anvendes rigtig meget i visse miljøer. Man kan imidlertid finde en del fagstatistikere, der vil tænde advarselslampen overfor forskellige over– og fejlfortolkninger af R^2–værdien, som det er let at lade sig besnære af, og som mange miljøer uden tvivl gør sig skyldige i engang imellem.

Et eksempel: Anscombes data

Et klassisk eksempel, der viser, hvorfor R^2 i sig selv er problematisk, er Anscombes fire datasæt vist nedenfor (Anscombe 1973). Det er den samme bedste rette linje, der går gennem punkterne i alle fire figurer (hældning 0,5 og skæring 3). Desuden har alle 4 datasæt samme R^2 = 0,667 = 66,7 %, men det er klart, at de modeller, der er givet ved de fire rette linjer ikke beskriver data lige godt. I den øverste højre figur er sammenhængen mellem x og y åbenlyst ikke-lineær, og sammenhængen i figuren i nederste højre hjørne giver det slet ikke mening at modellere som en ret linje.

gel af en bedre betegnelse, det efterfølgende referere ti situationer som "relevante ti

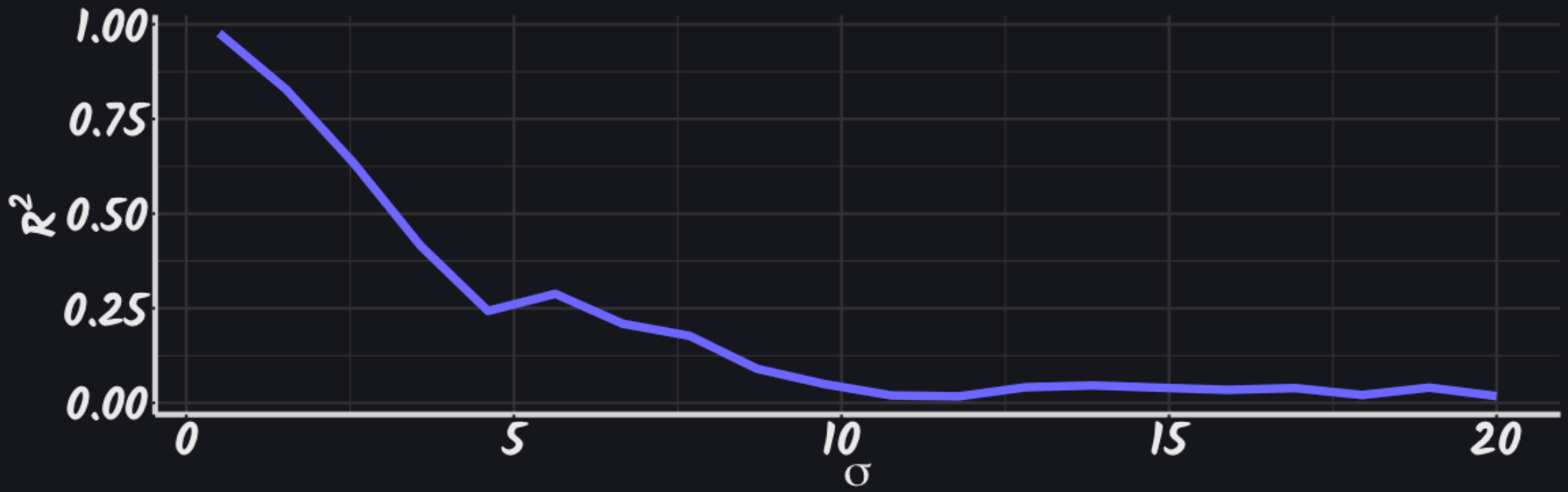
- Der er vigtige og centrale beg ger i hvad man kan uddrage a R^2–værdi, selv inden for de situationer

- En R^2–værdi bør *aldrig* stå h kombiner altid med visualis af data. Man kan huske og sig mantraen: *"Man skal tegn må regne"*

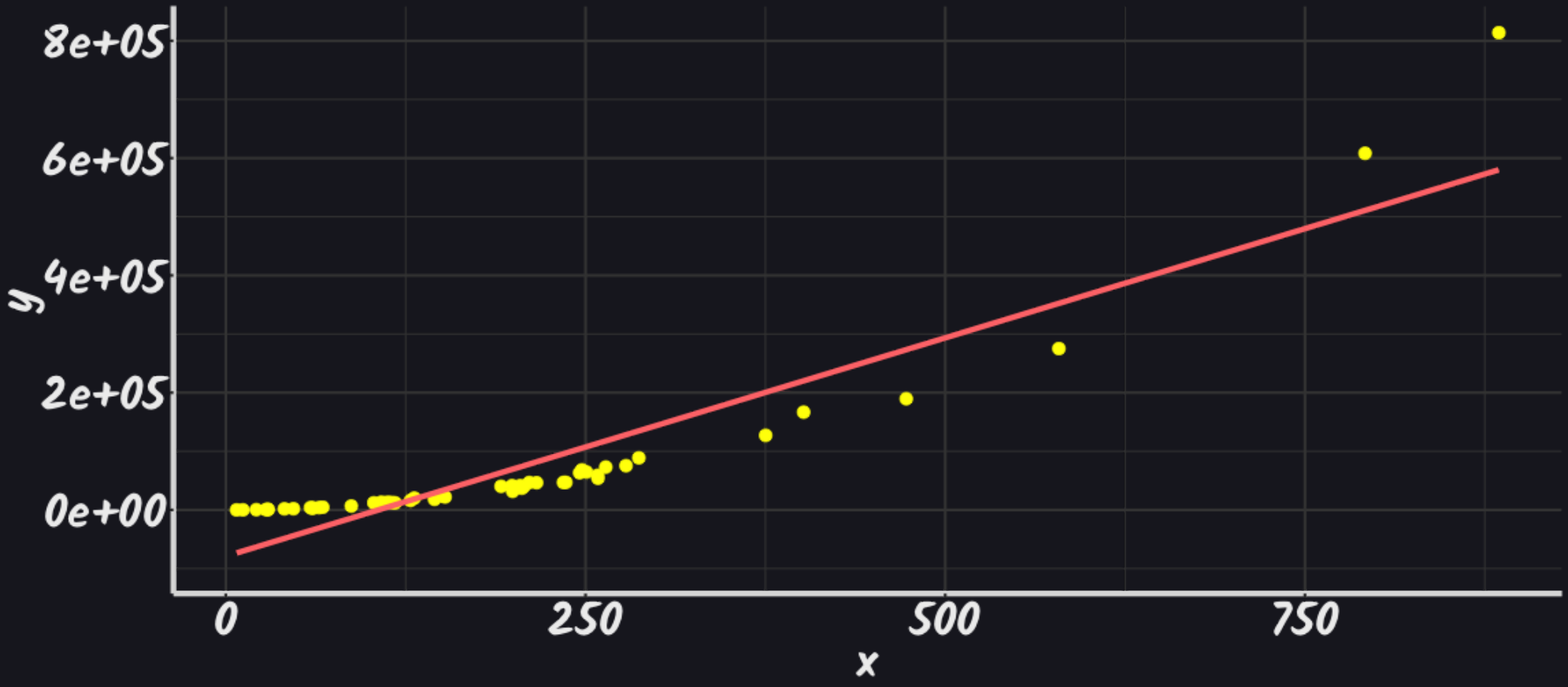
Det sidste punkt er måske det Hvis man vælger at bruge R^2 s

R^2 fortæller ikke, hvor korrekt modellen er

$$R^2 = \frac{\beta_1^2 \mathbb{V}(X)}{\beta_1^2 \mathbb{V}(X) + \sigma^2}.$$



En forkert model kan have R^2 tæt på 1



Siger ikke noget om prædiktionsfejl

$R^2 = 0.65$

Sammenligninger

R^2 kan ikke bruges til at sammenligne modeller med utransformeret Y med en model med transformerede Y

Kan bruges til at sige noget om forskellige modeller (med samme kompleksitet) med *samme* udfald. Men det kan spredningen også.

sandsynlig**vis**.dk

Beskatning og happiness