Stx

Matematik i Biologi A

Jette Holmgaard Vestergaard

Dronninglund Gymnasium

Januar 2022

Indholdsfortegnelse

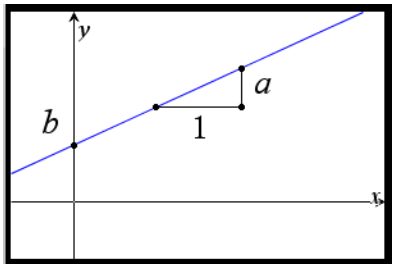
Illustrationer
Forside:
https://pixabay.com/da/photos/sommerfugl-blomster-best%c3%b8ve-6624801/

Forord

Dette hæfte er skrevet til Biologi A på stx. Hæftet er skrevet efter ønske fra Opgavekommissionen og Fagkonsulenten i biologi. Målet med hæftet er at give et overblik over de forskellige emner fra matematik, som eleverne skal kende, for at kunne løse opgaverne til den skriftlige prøve i Biologi A. Alle emnerne er obligatoriske på matematik B, som alle elever har haft.
I matematik er der tre hovedemner: funktioner, statistik og geometri.
I biologiopgaverne anvendes der kun matematik indenfor de to hovedemner: funktioner og statistik.
Efter hvert afsnit er der oplistet en række eksamensopgaver, hvor den pågældende teori har været anvendt. Enkelte opgaver er nævnt flere gange, f.eks. under både Lineær vækst og under Grafisk afbildning.
Hæftet er et lærerhæfte og ikke en lærebog. Hæftet er derfor ikke målrettet elever. Hæftet er målrettet lærerne, for at give dem et overblik over de forskellige emner fra matematik, som eleverne skal kende. Enkelte afsnit kan dog fint udleveres til elever.

1. Funktioner

1.1 Lineær vækst



Forskrift for en lineær funktion:

$$f(x) = a \cdot x + b$$

a er hældningskoefficienten.
b er skæring med y-aksen (startværdien).
Begge værdier skal altid fortolkes i en kontekst og ikke kun generelt.

Eksempel:
I en model kan sammenhængen mellem alder og længde for en population af spækhuggere beskrives ved modellen

$$f(x) = 37{,}5x + 273$$

Hvor $f(x)$ angiver længden (cm), og x angiver alderen (år).
$a = 37{,}5$ fortæller, at spækhuggerne i gennemsnit bliver 37,5 cm længere pr år.
$b = 273$ fortæller, at spækhuggerne i gennemsnit er 273 cm, når de bliver født.

Forklaringsgraden r^2:
r^2 er et mål for, hvor tæt punkterne ligger på den bedste rette linje.
r^2 er *ikke* et udtryk for, hvor god en model er.
En mere detaljeret forklaring ses i Appendiks om Lineær regression og r^2.

Frem og tilbage mellem x og y:
Eleverne skal kunne regne frem og tilbage mellem x og y. Dvs. hvis de får oplyst x skal de kunne beregne y og omvendt.
Også her sker beregningerne i en kontekst, dvs. eleverne skal først kunne afkode teksten. Er det en x- eller en y-værdi, de får oplyst. Derefter skal de kunne beregne den anden variabel.

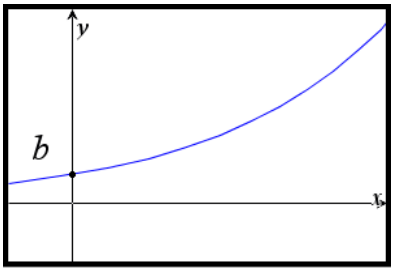
Eksamensopgaver:	2017 - 24/5 opgave 2	2019 - 27/5 opgave 4
2016 - 24/8 opgave 3	2017 - 24/5 opgave 3	2019 - 29/5 opgave 4

1.2 Eksponentiel vækst

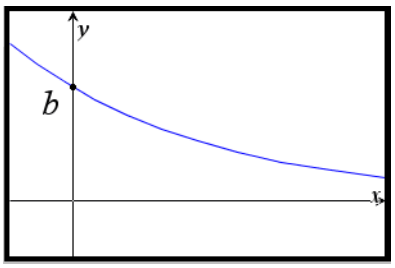
Forskrift for eksponentiel vækst:

$$f(x) = b \cdot a^x \quad \text{eller} \quad f(x) = b \cdot e^{k \cdot x} \quad \text{hvor} \quad e^k \approx a$$

a kaldes fremskrivningsfaktoren. a fortæller, om grafen er voksende eller aftagende.
$a > 1$: grafen er voksende.
$0 < a < 1$: grafen er aftagende.
a fortæller desuden, hvor meget grafen vokser eller aftager.
$a = 1 + r$, hvor r er vækstraten.
b er skæring med y-aksen.



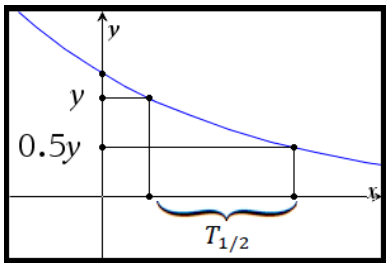
$e \approx 2{,}718281 \ldots$ er Eulers tal.
k fortæller, om grafen er voksende eller aftagende.
$k > 0$: grafen er voksende.
$k < 0$: grafen er aftagende.
b er skæring med y-aksen.



Eksempel:
$a = 1{,}12$: $r = a - 1 = 1{,}12 - 1 = 0{,}12$, dvs grafen vokser 12 % pr x-enhed.
$a = 0{,}94$: $r = a - 1 = 0{,}94 - 1 = -0{,}06$, dvs grafen aftager 6 % pr x-enhed.

Halveringstid:
For eksponentiel vækst findes der en konstant, så y-værdien halveres, hver gang x vokser med denne konstant. Denne konstant kaldes halveringskonstanten og betegnes $T_{1/2}$. Halveringskonstanten beregnes ved formlerne:

$$T_{1/2} = \frac{\log(0{,}5)}{\log(a)} \quad \text{eller} \quad T_{1/2} = \frac{\ln(0{,}5)}{k}$$



Frem og tilbage mellem x og y:
Som ved lineær vækst, skal eleverne kunne regne frem og tilbage mellem x og y og fortolke resultaterne i en kontekst.

Eksamensopgaver:
2016 - 30/5 opgave 2
2020 - 29/5 opgave 1
2021 - 1/6 opgave 1
2021 - 1/6 opgave 3

1.3 Grafisk afbildning

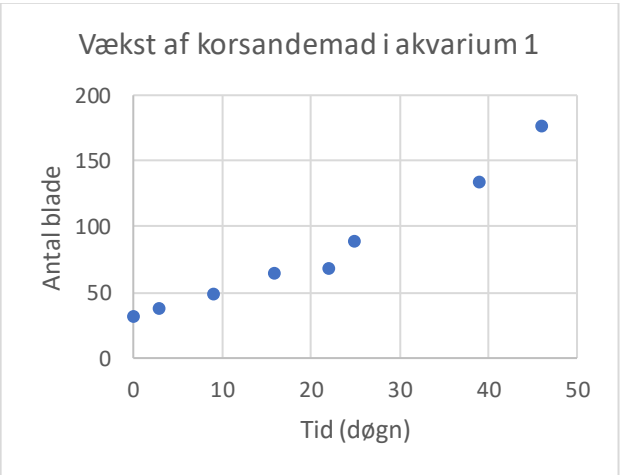
I opgaver, hvor eleverne bliver præsenteret for et datasæt, skal eleverne kunne lave en grafisk afbildning af data. Det vil ofte være et xy-plot, men kan f.eks. også være et søjlediagram.
I et xy-plot skal punkterne afbildes korrekt i et koordinatsystem. Den uafhængige variabel skal være på x-aksen og den afhængige variabel skal være på y-aksen. Der skal være enheder på akserne og aksetitler. Det vil desuden ofte være en god ide med en diagramtitel.
Punkterne skal afbildes som xy-plot, hvor punkterne er synlige. Eleverne skal desuden kunne indtegne tendenslinjer vha. lineær eller eksponentiel regression.

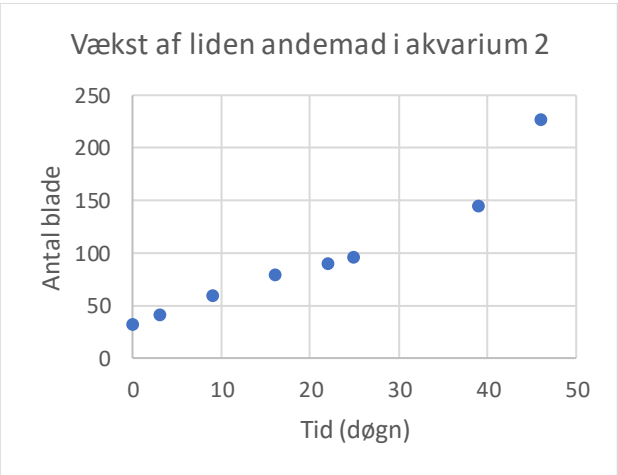
Eksempel (2016 – 30/5 opgave 2):
Tabellen viser resultatet fra et forsøg med korsandemad og liden andemad.

Tid (døgn)	Akvarium 1 Korsandemad (antal blade)	Akvarium 2 Liden andemad (antal blade)
0	30	30
3	37	40
9	48	58
16	64	77
22	67	89
25	88	95
39	133	143
46	176	225

Spørgsmål 2: Afbild resultaterne for akvarium 1 og 2 med antal blade som funktion af tiden.

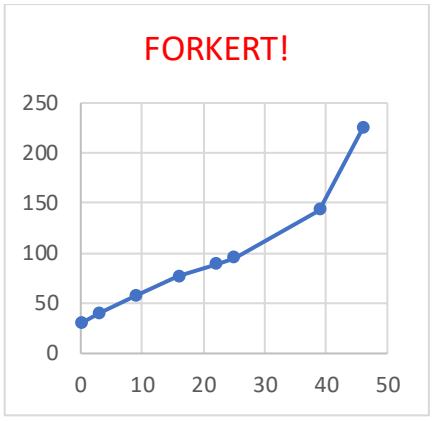
I dette tilfælde kunne korrekte afbildninger se således ud:

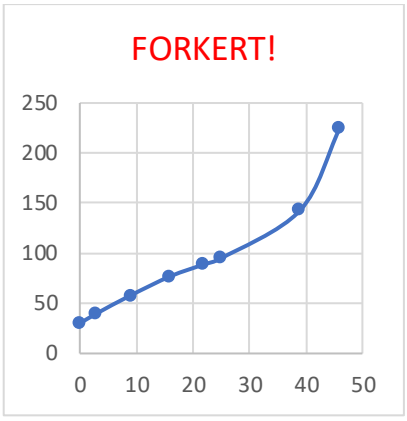


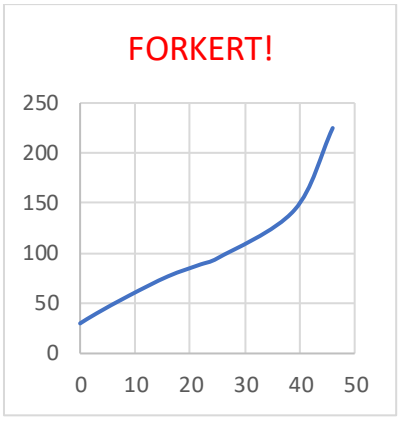


Man skal generelt *ikke* forbinde punkterne med linjestykke, da man som regel ikke har belæg for at antage, at der er forskellige lineære sammenhænge mellem de to variabler mellem hvert par af datapunkter.
Man skal heller *ikke* indtegne "bløde" grafer gennem punkterne.
Man må heller *ikke* skjule datapunkterne.
Endelig skal man passe på med at komme til at vælge kurver, hvor der er lige langt mellem datapunkterne på x-aksen. Denne type diagram anvendes kun, når tallene på x-aksen skal opfattes som tekst.

Nedenstående 3 diagramtyper for akvarium 2 er derfor alle *forkerte* afbildninger i den givne situation.







I opgaven skal eleverne desuden vurdere, om andemad vokser eksponentielt. De skal derfor kunne indtegne en eksponentiel tendenslinje og derefter udtale sig om modellens anvendelighed.

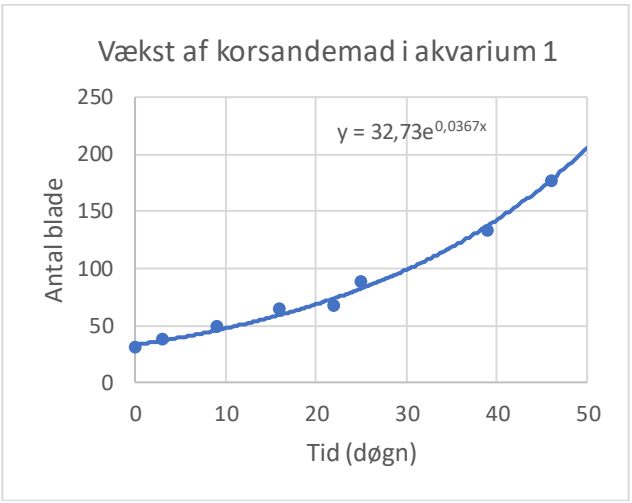
For akvarium 1 indtegnes den eksponentielle tendenslinje. Punkterne ligger tilfældigt rundt om grafen, så en eksponentiel model er en god model. Forskriften bliver

$$y = 32{,}73 \cdot e^{0{,}0367 \cdot x}$$

Vækstraten bliver

$$r = a - 1 = e^{k} - 1 = e^{0{,}0367} - 1 = 0{,}037$$

Dvs at antal blade i korsandemad vokser med ca 3,7 % pr døgn.



Eksamensopgaver:
2016 - 30/5 opgave 2
2016 - 24/8 opgave 3
2017 - 30/5 opgave 3
2017 - 24/8 opgave 3
2018 - 24/8 opgave 1
2019 - 27/5 opgave 4
2019 - 29/5 opgave 4
2020 - 29/5 opgave 4
2021 - 1/6 opgave 1

2. Statistik

2.1 Procentregning (andele og R-værdier)

Eleverne skal kunne udføre helt simple procentberegninger. De skal f.eks. kunne beregne og fortolke andele.

Eksempel:
I en klasse er der x piger og y drenge.

1) Beregn andelen af piger:
Andelen af piger er $\frac{x}{x+y}$.

2) Giv en fortolkning af, at forholdet $\frac{x}{y} = 0{,}75$:
Forholdet $\frac{x}{y}$ er lig med $0{,}75$. Dvs. at der er færre piger end drenge. Helt præcist betyder det, at antallet af piger svarer til 75 % af antallet af drenge.

Eksamensopgaver:
2015 - 29/5 opgave 4
2019 - 27/5 opgave 4
2020 - 29/5 opgave 3

2.2 Deskriptiv statistik

Eleverne skal kende til de mest simple statistiske deskriptorer til beskrivelse af et datasæt.
Når eleverne bliver præsenteret for et datasæt, som består af en række gentagelser af et forsøg, vil middelværdi og spredning være relevante deskriptorer. Det vil ofte være relevant at tegne et boksplot, da det både viser middelværdien og variationen i datasættet.
Boksplot anvendes især, når man vil sammenligne to eller flere grupper, hvor der i hver gruppe er blevet foretaget en række gentagelser af et forsøg.
En detaljeret beskrivelse af de forskellige størrelser, eksempler på anvendelse i biologiopgaver samt en tastevejledning til Excel findes i hæftet *Deskriptiv statistik i Biologi A – 2. udgave*.

Eksamensopgaver:
2019 - vejl sæt 2 opgave 2
2019 - vejl sæt 2 opgave 4
2020 - 28/5 opgave 2

2.3 Statistiske modeller

Hardy-Weinberg ligevægt

Hvis en population er i Hardy-Weinberg ligevægt mht to alleler, A og a, af et gen, er frekvenserne af de tre genotyper, AA, Aa og aa, følgende:

Genotype	AA	Aa	aa
Frekvens	p^2	$2pq$	q^2

hvor p = frekvensen af A og q = frekvensen af a.
Summen af genotypefrekvenserne er 1, dvs

$$p^2 + 2pq + q^2 = 1$$

Og summen af allelfrekvenserne er også 1, dvs

$$p + q = 1$$

I opgaver om Hardy-Weinberg ligevægt skelnes der mellem 2 situationer.
Enten kendes én af allelfrekvenserne eller også kendes én af genotypefrekvenserne.

Eksempel 1:
Én af allelfrekvenserne er kendt.
I en population i Hardy-Weinberg ligevægt er frekvensen af A lig med $p = 0{,}25$.
Bestem genotypefrekvenserne.
Da summen af de to allelfrekvenser er 1, er allelfrekvensen af a lig med

a: $q = 1 - p = 1 - 0{,}25 = 0{,}75$

De tre genotypefrekvenser bliver så

AA: $p^2 = 0{,}25^2 = 0{,}0625$
Aa: $2pq = 2 \cdot 0{,}25 \cdot 0{,}75 = 0{,}375$
aa: $q^2 = 0{,}75^2 = 0{,}5625$

Eksempel 2:
Én af genotypefrekvenserne er kendt.
I en population i Hardy-Weinberg ligevægt, er frekvensen af aa lig med $0{,}36$.
Bestem allelfrekvenserne.
Frekvensen af allelen a kan bestemmes ved at løse ligningen

a: $q^2 = 0{,}36$
a: $q = \sqrt{0{,}36} = 0{,}60$

Da summen af de to allelfrekvenser er 1, er allelfrekvensen af A lig med

A: $p = 1 - q = 1 - 0{,}60 = 0{,}40$

Eksamensopgaver:
2018 - 24/8 opgave 2
2019 - 27/5 opgave 3

Allelfrekvenser og genotypefrekvenser, når der ikke er Hardy-Weinberg ligevægt

Hvis en population *ikke* er i Hardy-Weinberg ligevægt, er der ikke et særligt system i frekvenserne af de tre genotyper, AA, Aa og aa. Der gælder blot, at summen af frekvenserne er 1. Hvis frekvenserne af de tre genotyper kaldes u, v og w, dvs

Genotype	AA	Aa	aa
Frekvens	u	v	w

gælder der at

$$u + v + w = 1$$

Da der er lige mange A'er og a'er i genotype Aa, findes allelfrekvenserne for A og a ved at halvere frekvensen af Aa og derefter lægge hver halvdel til frekvenserne af henholdsvis AA og aa.
Hvis p = frekvensen af A og q = frekvensen af a, bliver de to allelfrekvenser

A: $p = u + \frac{1}{2} \cdot v$

a: $q = w + \frac{1}{2} \cdot v$

Eksempel:
I en population er der observeret følgende antal af tre genotyper AA, Aa og aa.

Genotype	AA	Aa	aa	Sum
Antal	125	40	35	200

Genotypefrekvenserne beregnes som andele.

AA: $u = \frac{125}{200} = 0{,}625$

Aa: $v = \frac{40}{200} = 0{,}200$

aa: $w = \frac{35}{200} = 0{,}175$

Allelfrekvenserne beregnes vha formlerne for p og q.

A: $p = u + \frac{1}{2} \cdot v = 0{,}625 + \frac{1}{2} \cdot 0{,}200 = 0{,}725$

a: $q = w + \frac{1}{2} \cdot v = 0{,}175 + \frac{1}{2} \cdot 0{,}200 = 0{,}275$

Eksamensopgaver:
2019 - 29/5 opgave 2
2019 - 26/8 opgave 2

2.4 Statistiske tests

Binomialtest

Eleverne skal kunne gennemføre beregningerne og fortolke resultaterne i et binomialtest.
En detaljeret beskrivelse af teorien bag et binomialtest, eksempler på anvendelse i biologiopgaver samt en tastevejledning til Excel findes i hæftet *Binomialtest i Biologi A – 2. udgave*.

VIGTIGT: I januar 2020 kom der en ændring til vejledningen i Matematik. Det 1-sidede binomialtest, som er det sværeste af de to test, blev fjernet fra kernestoffet i matematik, så eleverne efter januar 2020 kun bliver stillet opgaver i matematik med 2-sidede binomialtest. I biologi forventes det derfor ikke mere, at eleverne mestrer det 1-sidede binomialtest. 2. udgave af hæftet til biologi er derfor en redigeret udgave, hvor det 1-sidede test er fjernet.

Eksamensopgaver:
2018 - vejl sæt 1 opgave 4
2019 - vejl sæt 2 opgave 2

Signifikant sandsynlighed

I forbindelse med statistiske tests, skal eleverne kunne fortolke p-værdier.
Statistiske tests er bygget op på følgende måde:
Der opstilles en nulhypotese, H_0. Den vil ofte være det modsatte af dét, man vil undersøge.
Der opstilles også en alternativ hypotese, H_1. Den er det modsatte af nulhypotesen, og den gælder, hvis nulhypotesen forkastes.
Derefter udtages der en repræsentativ stikprøve af hele populationen for at teste nulhypotesen.
Der beregnes en p-værdi, der angiver sandsynligheden for at få en "lige så skæv eller endnu mere skæv" stikprøve under nulhypotesen, hvis man udtager en ny stikprøve.
Hvis p-værdien er lille, skyldes det enten 1) eller 2):
1) Man har observeret noget, der sjældent forekommer under denne nulhypotese.
2) Nulhypotesen er forkert.
Som udgangspunkt tror man altid på sin nulhypotese. Men hvis p-værdien er meget lille, begynder man at tvivle på nulhypotesen, da man jo har observeret noget, der sjældent forekommer under nulhypotese.
Spørgsmålet er så, hvor lille p-værdien skal være, før man ikke længere tror på nulhypotesen.
Ved de fleste statistiske test vælger man at sige, at hvis p-værdien er under 5 %, så har man observeret noget, der er meget usandsynligt under denne antagelse, og så må antagelsen, altså nulhypotese, være forkert.
Denne grænse på 5 % kaldes signifikansniveauet.
Man siger, at en p-værdi er ***signifikant***, hvis den er mindre end signifikansniveauet.

Eksempel:
I 2016 var 18 % af Danmarks gymnasieelever rygere. Man ønsker at undersøge, om dette har ændret sig.
Nulhypotese, H_0: Andelen af rygere blandt Danmarks gymnasieelever er uændret, altså stadig 18 %.
Alternativ hypotese, H_1: Andelen af rygere blandt Danmarks gymnasieelever har ændret sig.
Der udtages en stikprøve på 500 gymnasieelever.
I stikprøven er der 70 rygere. Dvs at der er 14 % rygere. Spørgsmålet er, om denne forskel fra tidligere er signifikant.
Der laves her et binomialtest, hvilket resulterer i en p-værdi på 2 %. Det vil sige, at sandsynligheden for at observere 70 rygere ud af 500 eller noget der er endnu mere "skævt" under nulhypotesen, kun er 2 %. Altså

har man observeret noget meget sjældent under denne nulhypotese. Vi tror derfor ikke på, at nulhypotesen er sand. Da p-værdien er under 5 % forkastes nulhypotesen.
Ved et signifikansniveau på 5 % kan man konkludere, at andelen af rygere blandt Danmarks gymnasieelever har ændret sig.

Opsummering:
En p-værdi er *ikke* sandsynligheden for, at nulhypotesen er sand.
p-værdien angiver sandsynligheden for at få en "lige så skæv eller endnu mere skæv" stikprøve under nulhypotesen, hvis man udtager en ny stikprøve.
Hvis p-værdien er under 5 % forkastes nulhypotesen og det konkluderes, at der er en signifikant forskel på de grupper, der undersøges.

Eksamensopgaver:
2018 - 4/6 opgave 1

Matematisk hypotese vs biologisk hypotese

Ordet ***hypotese*** optræder i mange biologiopgaver uden at det forventes, at der udføres et matematisk test. Formuleringen vil ofte være noget i stil med: "Forskerne havde en hypotese om at….". I disse opgaver forventes der grafer eller deskriptiv statistik eller biologiske argumenter for, om man tror på hypotesen eller ej.
Ordet hypotese anvendes derfor ikke på samme måde i de to fag.
Når man undersøger en hypotese i biologi, undersøger man en stikprøve og konkluderer derefter noget om stikprøven.
Når man tester en hypotese i matematik, undersøger man en repræsentativ stikprøve af en population og konkluderer derefter noget om hele populationen. Hvis der forventes et matematisk test, står det eksplicit i opgaven.

Eksamensopgaver:
2019 - vejl sæt 2 opgave 4
2019 - 27/5 opgave 1
2019 - 29/5 opgave 2
2020 - 28/5 opgave 1
2020 - 29/5 opgave 1
2021 - 31/5 opgave 1
2021 - 31/5 opgave 2
2021 - 1/6 opgave 2

3. ”Foretag relevant databehandling…”

I forbindelse med gymnasiereformen i 2017 indførtes en ny type formulering i biologiopgaver. I det vejledende opgavesæt 1 fra 2018 optræder formuleringen ”Foretag relevant databehandling, der kan vise, at…” for første gang.
Her lægges der op til en større selvstændighed hos eleverne. De skal selv kunne vurdere, om det er relevant med en grafisk afbildning (noget med funktioner, ofte en afbildning af et eller andet over tid), eller om det er mere relevant med noget deskriptiv statistik.

Eksamensopgaver:
2018 – vejl sæt 1 opgave 1
2021 - 31/5 opgave 1
2021 - 31/5 opgave 2

4. Appendiks om lineær regression og r^2

Lineær regression.

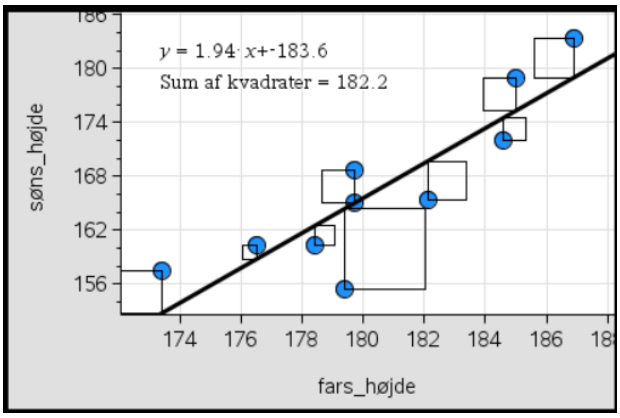
Lineær regression betyder "at finde den bedste rette linje gennem en række datapunkter". Den bedste rette linje defineres som den linje, der giver "de mindste kvadrater på de lodrette afstande".

Eksempel:
I en undersøgelse af 10 jævnaldrende drenge og deres fædre har man målt højden af far og søn.

Søn nr	1	2	3	4	5	6	7	8	9	10
Fars højde (cm)	178,4	182,1	179,4	179,7	186,9	184,6	173,4	176,5	179,7	185,0
Søns højde (cm)	160,3	165,4	155,4	168,7	183,4	172,0	157,5	160,3	165,0	179,0

Kilde: Bo Markussen: "Lineær regression A-niveau", Københavns Universitet, 2018.

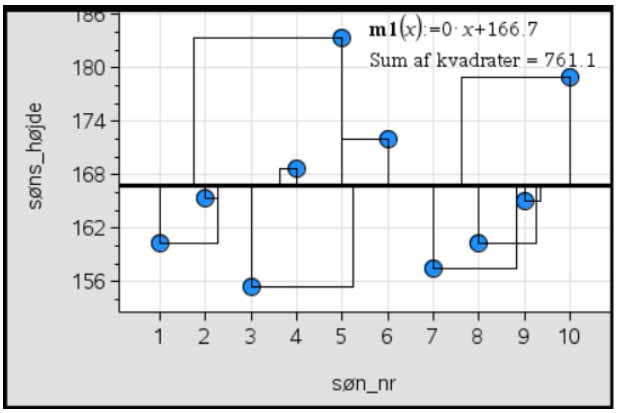
Den ***bedste rette linje*** er den linje, der gør summen af arealerne af kvadraterne på de lodrette afstande mellem punkterne og linjen mindst mulig.
I dette tilfælde bliver arealet af kvadraterne 182,2. Dette tal er imidlertid ikke vigtigt i sig selv og bliver normalt heller ikke angivet, når man laver lineær regression.
Summen af arealerne skal blot bruges af værktøjsprogrammet til at vurdere den rette linje i forhold til andre rette linjer.



Forklaringsgraden, r^2.

Når man laver lineær regression, beregnes desuden en størrelse, r^2, der kaldes ***forklaringsgraden***. Denne størrelse fortolkes ofte som et mål for, hvor god en model er. Det er imidlertid forkert. r^2 er et mål for, hvor meget af variationen i y-værdierne, der matematisk set kan forklares vha. x-variablen.

Man forestiller sig, at x-variablen, altså fædrenes højde er ukendt. Man kender altså kun sønnernes højde. Hvis man i den situation skal lave en model for sønners højde, vil det bedste bud være en vandret lineær model, $f(x) = b$, hvor b er gennemsnittet af sønnernes højde. I dette tilfælde bliver den vandrette model $f(x) = 166{,}7$, og summen af arealerne af kvadraterne bliver 761,1.

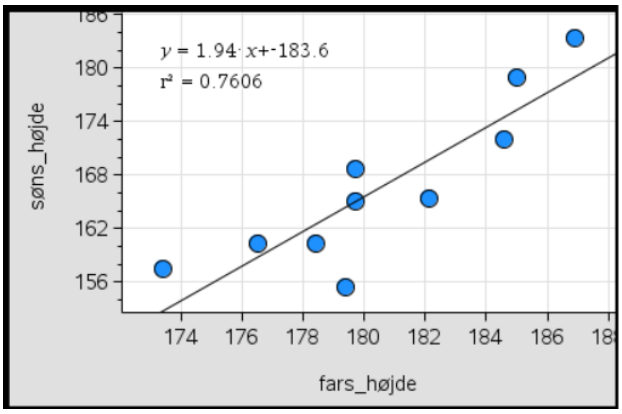


Forklaringsgraden, r^2, er et mål, der viser, hvor stor en procentdel af det samlede areal der forsvinder, når x-variablen inddrages i den lineære model.

Arealet uden fædrenes højde i modellen var 761,1.
Arealet med fædrenes højde i modellen var 182,2.
Procentdelen af det samlede areal, der forsvinder, når fædrenes højde inddrages i modellen, er:

$$\frac{761{,}1 - 182{,}2}{761{,}1} = 0{,}7606$$

Dvs at arealet er blevet 76 % mindre ved at medtage fædrenes højde i modellen. Altså forklarer fædrenes højde matematisk set 76 % af variationen i sønnernes højde. Og det er dét, forklaringsgraden fortæller.



Man skal imidlertid aldrig blindt lave den fortolkning, at x-variablen forklarer r^2 % af variationen i y-variablen. Der kan være skjulte variabler, der påvirker både x- og y-variablen og som i virkeligheden er den reelle forklaring på variationen i y-variablen.

Eksempel:
Tabellen viser antal storkepar og antal levendefødte børn i Danmark i perioden 1973-1983.

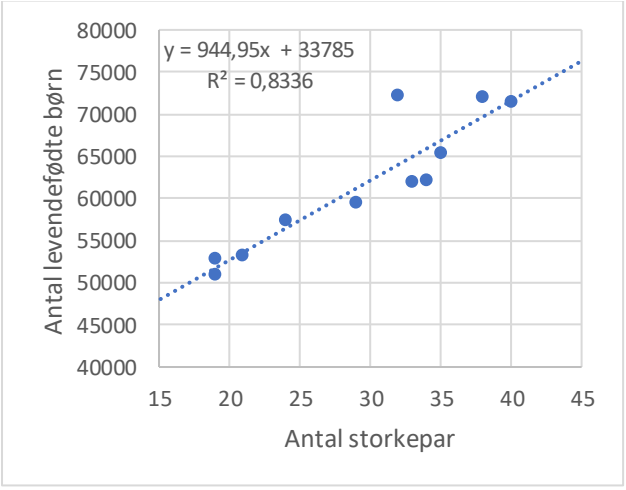
Årstal	Antal storkepar	Antal levendefødte børn
1973	38	71.895
1974	40	71.327
1975	32	72.071
1976	35	65.267
1977	33	61.878
1978	34	62.036
1979	29	59.464
1980	24	57.293
1981	21	53.089
1982	19	52.658
1983	19	50.822

Kilde: Brian Krog Christensen, Peter Limkilde: "Ind i Naturvidenskab", Gyldendal, 2007.
Kilde: https://www.statistikbanken.dk/statbank5a/default.asp?w=1536

Når man plotter antal levendefødte børn som funktion af antal storke, ses det, at jo flere storke desto flere levendefødte børn, så ordsproget ser ud til at passe.

"Storken kommer med børnene."

Der er imidlertid ingen sammenhæng mellem de to variabler. Den bagvedliggende variabel er tid. Begge variabler er tilfældigvis faldet i samme tidsperiode, men har ikke noget med hinanden at gøre.



Hvornår er en lineær model en god model?

Når man skal vurdere, om en model er en god model, skal man som nævnt *ikke* kigge på forklaringsgraden. r^2 fortæller kun, hvor meget af variationen i y-variablen som x-variablen rent matematisk forklarer, men ikke noget om, om modellen er god. Man kan heller ikke sætte en grænse for, hvor høj r^2 skal være, før en model er god.

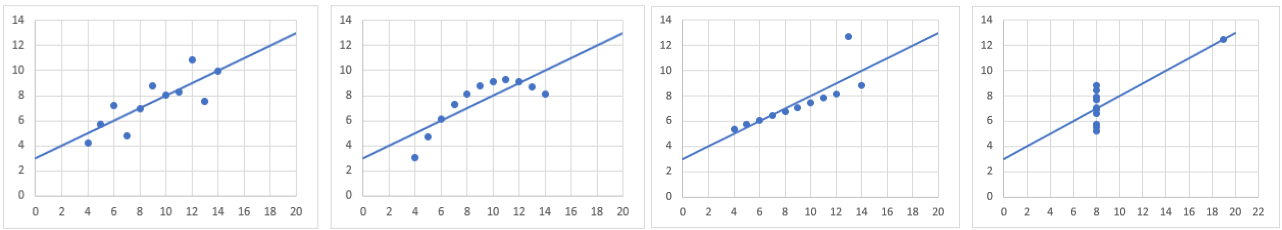
Et klassisk eksempel, der illustrerer dette, er Anscombes Quartet.

I 1973 konstruerede F. J. Anscombe 4 datasæt, som havde samme forklaringsgrad, men som var meget forskellige. Alle 4 datasæt havde en forklaringsgrad på 0,67.

Tabellen viser Anscombes Quartet.

I		II		III		IV	
x	y	x	y	x	y	x	y
10	8,04	10	9,14	10	7,46	8	6,58
8	6,95	8	8,14	8	6,77	8	5,76
13	7,58	13	8,74	13	12,74	8	7,71
9	8,81	9	8,77	9	7,11	8	8,84
11	8,33	11	9,26	11	7,81	8	8,47
14	9,96	14	8,10	14	8,84	8	7,04
6	7,24	6	6,13	6	6,08	8	5,25
4	4,26	4	3,10	4	5,39	19	12,50
12	10,84	12	9,13	12	8,15	8	5,56
7	4,82	7	7,26	7	6,42	8	7,91
5	5,68	5	4,74	5	5,73	8	6,89

Kilde: https://en.wikipedia.org/wiki/File:Anscombe%27s_quartet_3.svg



De 4 afbildninger i diagrammet viser de 4 datasæt sammen med den bedste rette linje.

Det ses tydeligt, at det kun er i det første tilfælde, at en lineær model, er en god model.

I det andet tilfælde, vil en polynomiel model være bedre.

I det tredje tilfælde, ser der ud til, at der er en fejlmåling. Så man vil kunne lave en god lineær model, hvis man fjernede det næstsidste punkt. Man skal imidlertid altid være varsom med at slette data.

I det fjerde tilfælde giver det slet ikke mening at forsøge at lave en model overhovedet.

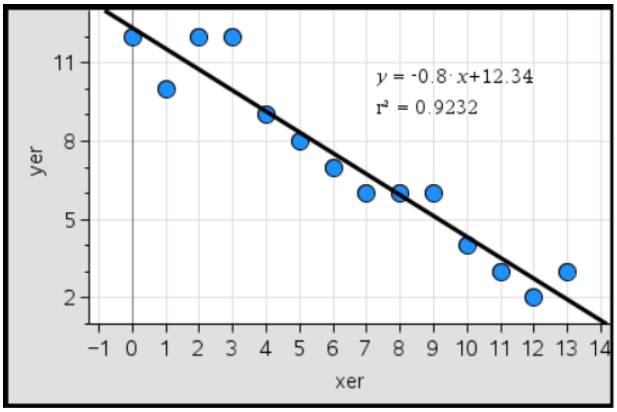
Det ses altså tydeligt her, at en forklaringsgrad på 0,67 ikke er en garanti for en god model.

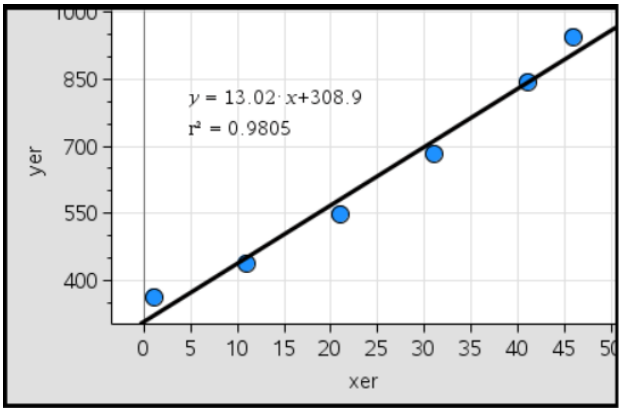
Når man skal vurdere, om en model er god, skal man derfor se på punkternes beliggenhed i forhold til tendenslinjen.

En tendenslinje er en god model, hvis punkterne ligger tilfældigt uden systematiske afvigelser rundt om linjen.

En tendenslinje er en dårlig model (eller i hvert fald ikke den bedste model), hvis punkterne ligger systematisk rundt om linjen, dvs. hvis der kan tegnes en anden graf, der passer bedre på punkterne.

Eksempel:





I diagrammet til venstre ligger punkterne tilfældigt uden systematiske afvigelser rundt om linjen, så den rette linje er en god model, på trods af, at punkterne ikke alle ligger tæt på linjen. Her er $r^2 = 0{,}92$.
I diagrammet til højre ligger punkterne alle tæt på den rette linje, men der er systematik i afvigelserne, og det ser ud til, at en "buet" graf vil passe bedre på punkterne, så den rette linje er en dårlig model eller i hvert fald ikke den bedste model, på trods af, at $r^2 = 0{,}98$. En eksponentiel model vil være en bedre model i dette tilfælde.
Den lineære model vil dog godt kunne bruges til nogenlunde af forudsige y-værdier i det interval, hvori der er foretaget målinger, dvs. i x-intervallet [0;45], da punkterne jo ligger tæt på grafen i dette interval, men den lineære model vil være dårlig til at forudsige y-værdier for x-værdier større end ca. 45.
Den eksponentielle model vil imidlertid være bedre til at forudsige y-værdier, både i intervallet og udenfor intervallet.

Opsummering:
r^2 er et udtryk for, hvor mange procent af variationen i y-variablen, der rent matematisk kan forklares af x-variablen.
r^2 kan *ikke* bruges til at vurdere, om en lineær model er god.
En lineær model er god, hvis punkterne ligger tilfældigt uden systematiske afvigelser rundt om linjen.

Litteratur

Brockhoff, Hansen & Ekstrøm, *Brugen af R^2 i gymnasiet*, Københavns Universitet, 2017

Markussen & Rønn-Nielsen, *Lineær Regression A-niveau*, Københavns Universitet, 2018

Undervisningsministeriet, *Matematik Screening, Vejledende opgavesæt 2*, 2017