

First AID STAT

Athena & Nanna

2017

FIRST AID STAT

[Hvilken type data? 2](#_Toc498969996)

[Deskriptiv statestik (SAU1) 2](#_Toc498969997)

[Typer af deskriptiv statestik: 2](#_Toc498969998)

[Numerisk variabel: 3](#_Toc498969999)

[Katekorisk variabel: 3](#_Toc498970000)

[Koder i R til DESKRIPTIV STATESTIK 3](#_Toc498970001)

[Kontinuerte data (SAU 2) 3](#_Toc498970002)

[Simpel lineær regression: 3](#_Toc498970003)

[Estimation 4](#_Toc498970004)

[Prædikation 4](#_Toc498970005)

[Referenceinterval 4](#_Toc498970006)

[Koder i R til Lineær regression 5](#_Toc498970007)

[Standardfejl (SAU 3) 5](#_Toc498970008)

[Konfidensinterval 5](#_Toc498970009)

[Hypoteser 5](#_Toc498970010)

[Nulhypotese 5](#_Toc498970011)

[P-værdi 6](#_Toc498970012)

[Multipel lineær regression (SAU 4) 6](#_Toc498970013)

[Indikatorvariabel 6](#_Toc498970014)

[Matematisk fortolkning 7](#_Toc498970015)

[Interaktionseffekt 7](#_Toc498970016)

[Interaktionsled og interaktionseffekt 8](#_Toc498970017)

[Koder i R til multibel lineær regression 8](#_Toc498970018)

[Logistisk regression (SAU 6) 9](#_Toc498970019)

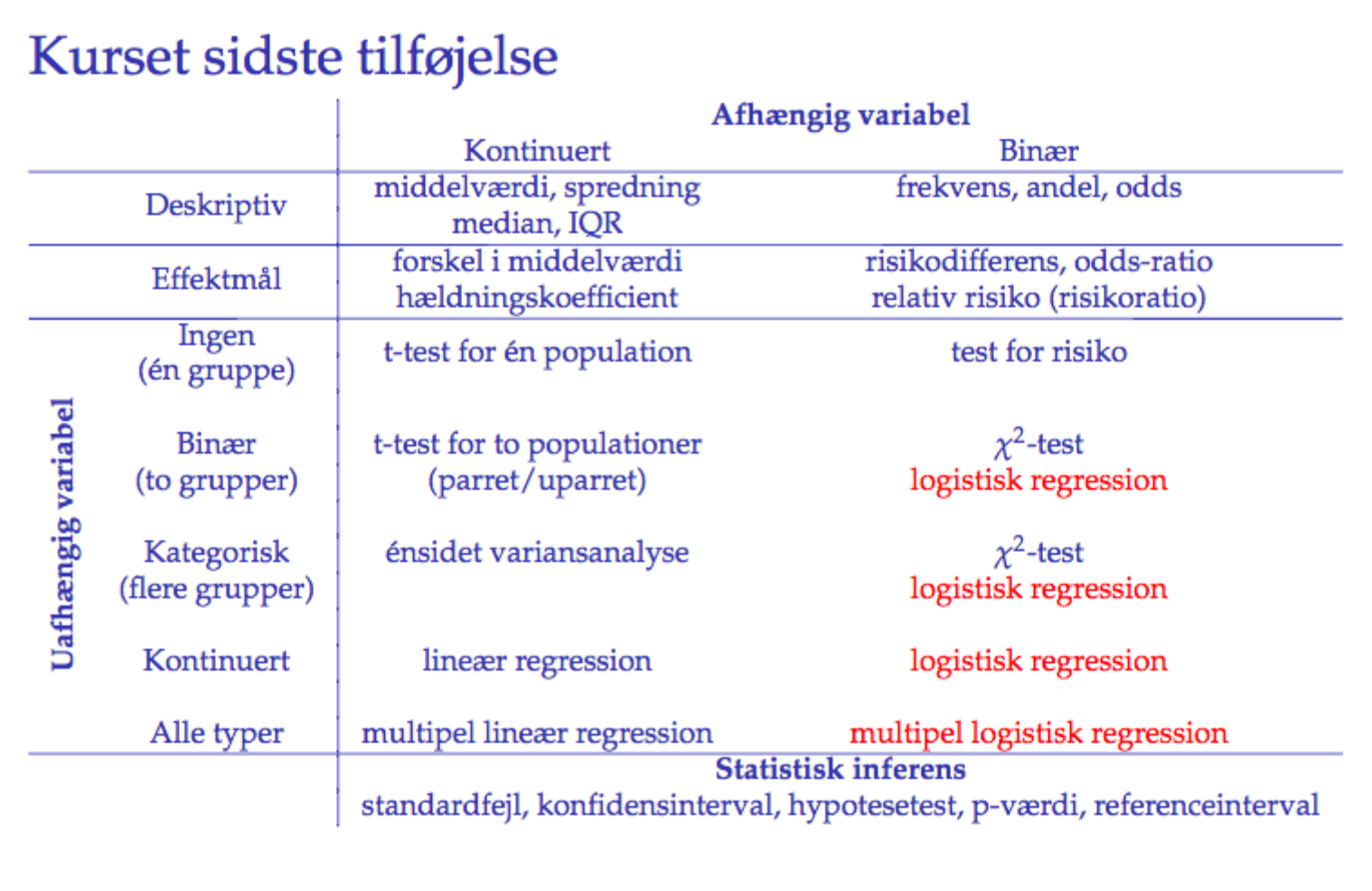
[Regressionsparametre 9](#_Toc498970020)

[Skalaer til betragtning af logistisk regressionsmodel 9](#_Toc498970021)

[Koder i R til logistisk regression 9](#_Toc498970022)

# Hvilken type data?

Brug oversigten til at vælge hvilken model, der egner sig bedst til beskrivelse:



# Deskriptiv statestik (SAU1)

Giver overblik over data.

Illustreres gennem:

* Scatter plot
* Histogram
* Boxplot
* Søjlediagram

## Typer af deskriptiv statestik:

* Kvantetative data: *NUMERISKE VARIABLE*
* Kvalitativ data: *KATEGORISK VARIABLE*

### Numerisk variabel:

* Værdi: gennemsnit/median
* Varians: spredning (standardafvigelse/SD)

### Katekorisk variabel:

* Frekvens
* Andel

## Koder i R til DESKRIPTIV STATESTIK

|  |  |
| --- | --- |
| **Hvad vil du lave?** | **R-kode** |
| Søjlediagram/bar plot | barplot |
| Histogram | hist |
| Median | median |
| Inter-quartile range (De midterste 50 %) | IQR |
| Boksplot | boxplot |
| Middelværdi | mean |
| Kvartiler | quantile |
| Tabelopsummering | table |
| Punktdiagram | plot (tilføj: type=”l” for at forbinde punkter) |

For at få udskrevet én bestemt vektor, fx x, bruges navnet på datasæt samt $. (eks: myData$x)

# Kontinuerte data (SAU 2)

Man ønsker at beskrive sammenhæng mellem kontinuerte varible.

## Simpel lineær regression:

y = afhængig variabel

x = uafhængig varibel

a = værdien af y, når x = 0 (skæring med y-aksen)

b = linjens hældning (ændring af y, når x ændres +1)

e = fejlled, naturlig variiation, måleusikkerhed, støj (e ̴N(0,s). e er normalfordelt med middelværdi 0 og spredning, s – den gennemsnitlige afstand mellem punkter og linje.

Y 🡨 X, Afhængig variabel 🡨 Uafhængig variabel

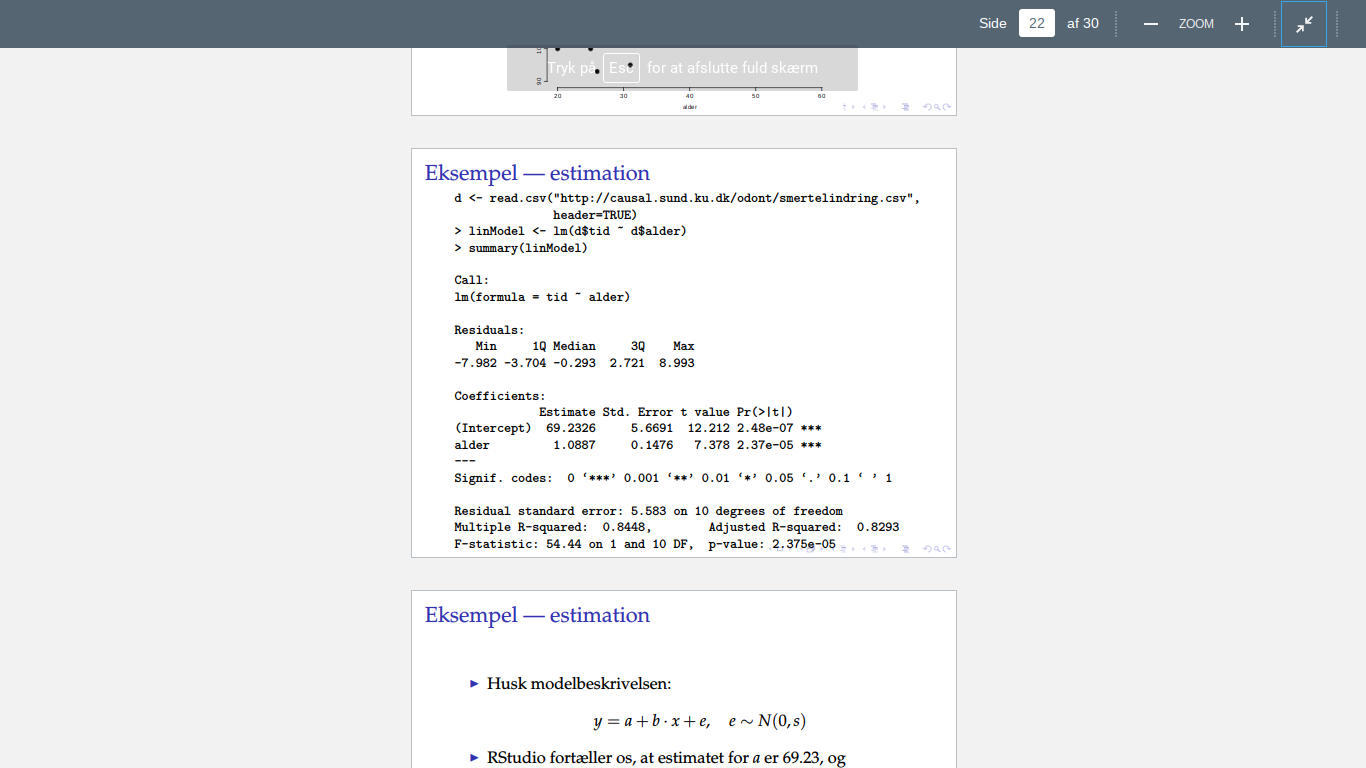
y = a + b ∙ x + e

*HUSK: tag stilling til, hvilken variabel, der er den afhængige (y).*

### Estimation

For at finde a og b laver vi *estimation*. Vi finder de værdier, der har den bedste tilpasning til linjen.

Fra nedenstående eksempelt ses det, at a kan aflæses som (intercept) og b som alder.



### Prædikation

Efterfølgende kan man lave en prædikation, ved at sætte en værdi ind på x plads. Altså den uafhængige varibel, fx år.

Prædikation er et punkt på regressionslinjen og udtryk for en gennemsnitlig værdi af den afhængige variabel. Man kan regne et referenceinterval for prædikationen.

### Referenceinterval

- 2\*2,5 / + 2\*2,5. Dette siger noget om, værdien for 95% af befolkningen.

## Koder i R til Lineær regression

|  |  |
| --- | --- |
| **Hvad vil du lave?** | **R-kode** |
| Lineær model | linModel |
| Opsummering på model | summary(linModel) |
| Tegne den estimerede linje | abline |

# Standardfejl (SAU 3)

Spredningen på stikprøvemiddelværdien. Vi kan bruge vores standardfejl til at beregne konfidensinterval; Hvor sikre er vi på at stikprøven har ramt rigtigt?

## Konfidensinterval

Endepunkterne for konfidensintervallet er vores sikkerhedsgrænse.

Et 95% konfidensinterval er sandsynligheden for at 95% af værdierne fra en ny stikprøve ville ligge inden for vores grænser. Altså, at intervallet indeholder den sande værdi.

## Hypoteser

”Uskyldig til det modsatte er bevist”.

Vi tester en hypotese, når vi laver statestik.

### Nulhypotese

Beskriver fraværet af en effekt, fx, man får ikke lavere blodtryk af blodtryksnedsættende præparat. Ved en lineær regressionsmodel, vil det betyde at vores linje har en hældning på 0.

Stemmer data ikke overens med 0-hypotesen forkastes den, *SIGNIFIKANS*. Hvis data stemmer overens, kan den ikke forkastes og vi må konkludere, at der ikke er tilstrækkeligt bevis for effekt, *INSIGNIFIKANS*.

Størrelsen af ’beviset’ kaldes P-værdien.

## P-værdi

Definition: sandsynligheden for at få et estimat, der ligger lige så langt fra eller længere fra nulhypotesen, end den værdi, vi faktisk fik.

ALTSÅ.. vi vil gerne langt væk fra 0-hypotesen, så vi kan forkaste den.

P-værdien:

< 0,05 🡪 vi forkaster 0-hypotesen.

≥ 0,05 🡪 vi forkaster ikke 0-hypotesen.

|  |  |
| --- | --- |
| **Hvad vil du lave?** | **R-kode** |
| 95% konfidensinterval i linModel | confint |

# Multipel lineær regression (SAU 4)

Én kontinuert variebel og flere uafhængige variable – både kontinuerte og kategoriske.

Fx kan effekten af alder(uafhængig variabel) på højde (afhængig variable) være forskellige i forhold til køn (uafhængig, kategorisk variabel).

y = a + b1 ∙ x1 + b2 ∙ x2.... + e

b’erne er de tilhørende regressionkoefficienter.

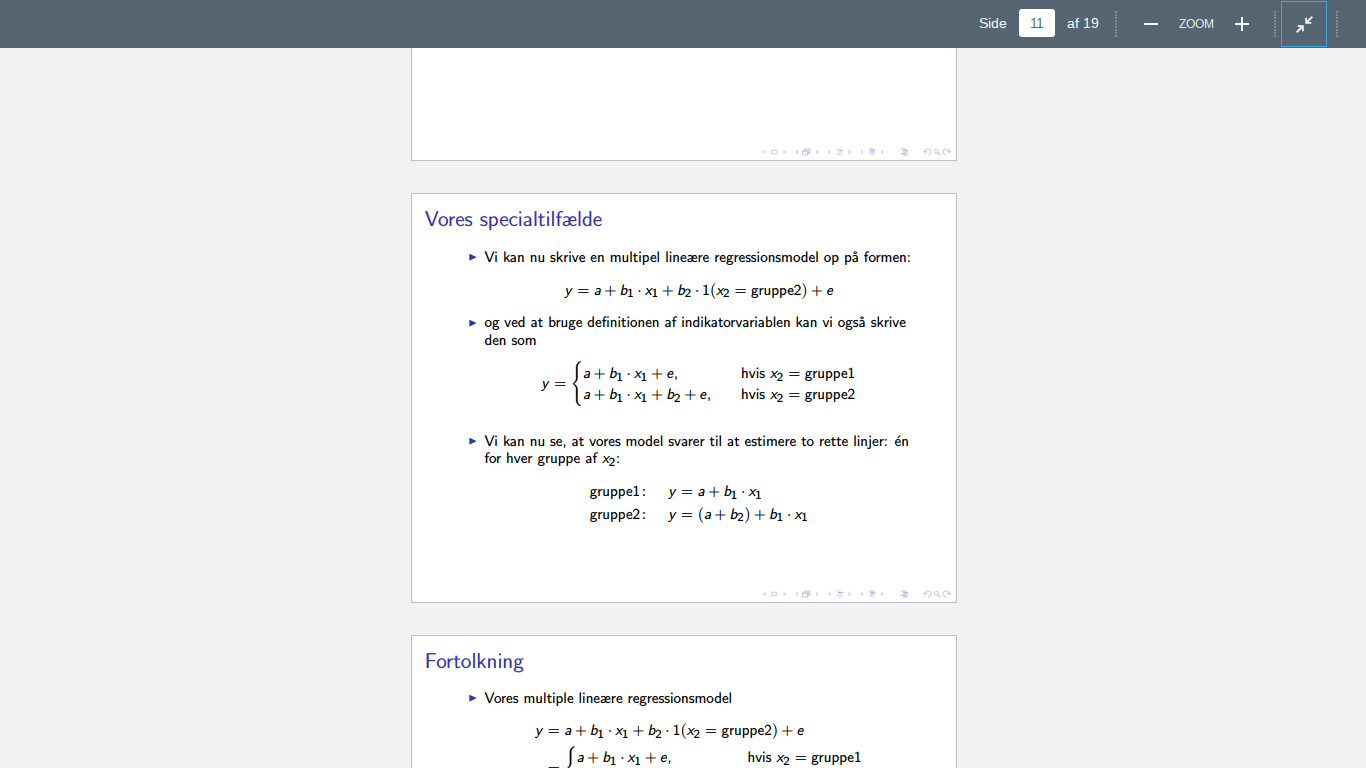
Vores kategoriske variable skal tildeles talværdier, for at vi kan sætte de, ind i en statistisk model, fx 1 og 2 (ja/nej el. Mand/kvinde). Dette kaldes en indikatorvariabel.

## Indikatorvariabel

y =

y = a + b1 ∙ x1 + e, hvis x2 = gruppe 1

y = a + b1 ∙ x1 + b2 ∙ x2 + e, hvis x2 = gruppe 2



### Matematisk fortolkning

Vi vil få to paralle linjer;

Gruppe1: a = skæring med y-aksen og b1 er hældningen.

Gruppe2: a + b2 = skæring med y-aksen og b1 er hældningen.

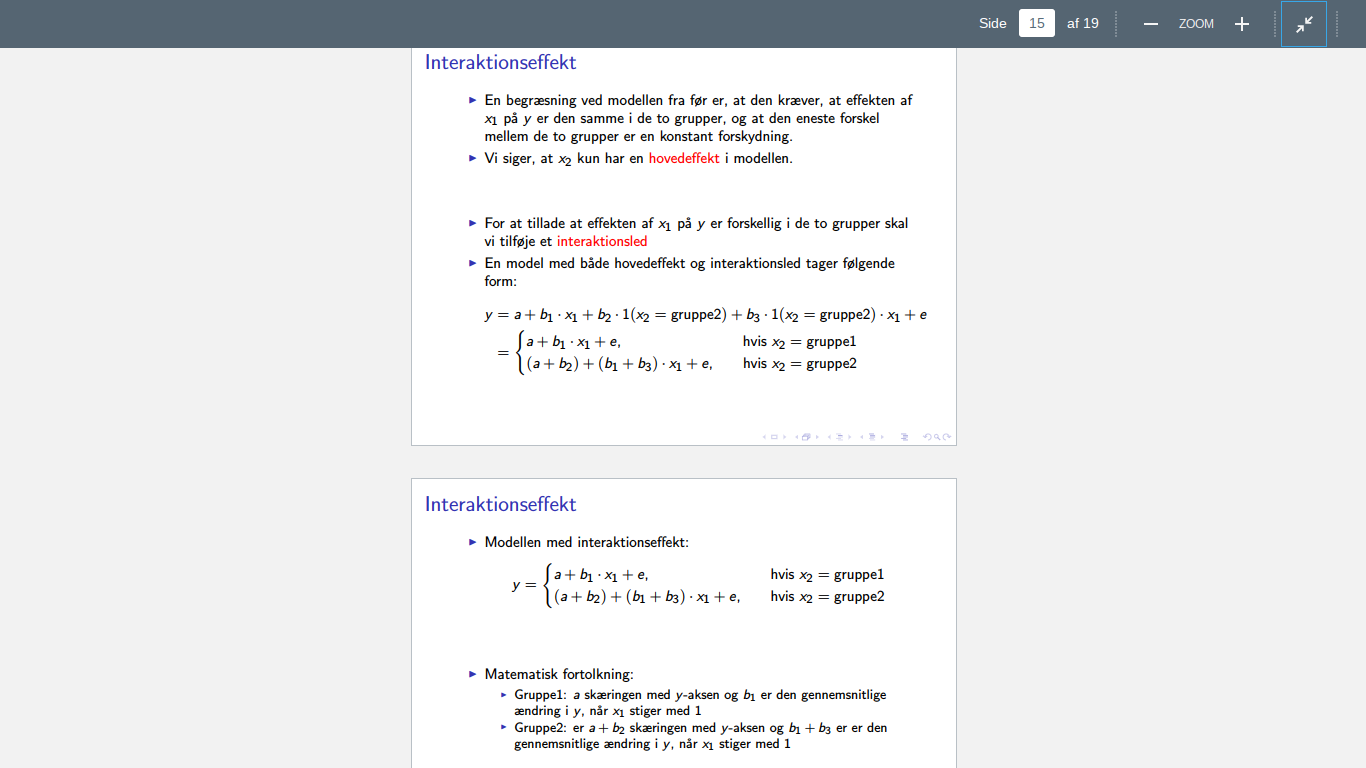
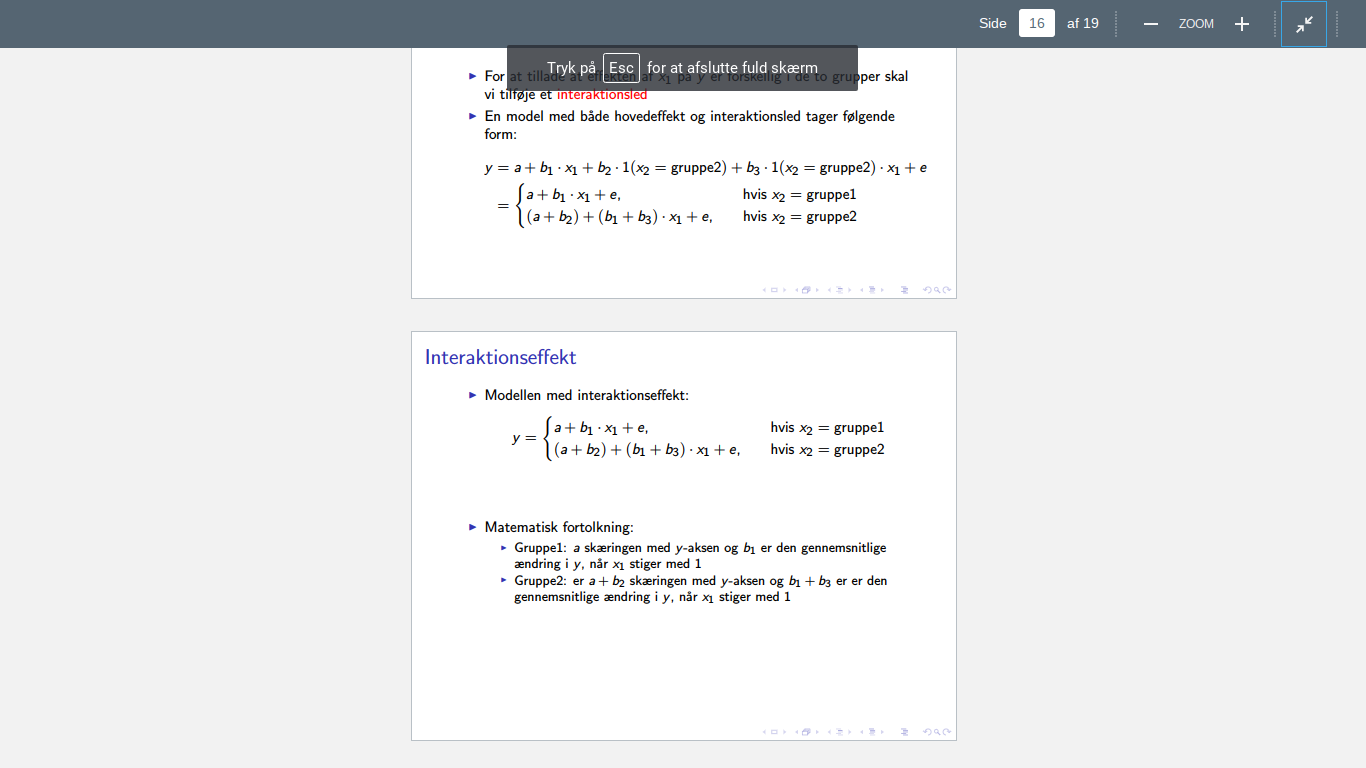
Altså, y afhænger af x på samme måde i begge grupper, MEN gruppe 2 ligger konstant højere/lavere i gennemsnitlig værdi.

B2 er altså den ekstra effekt.

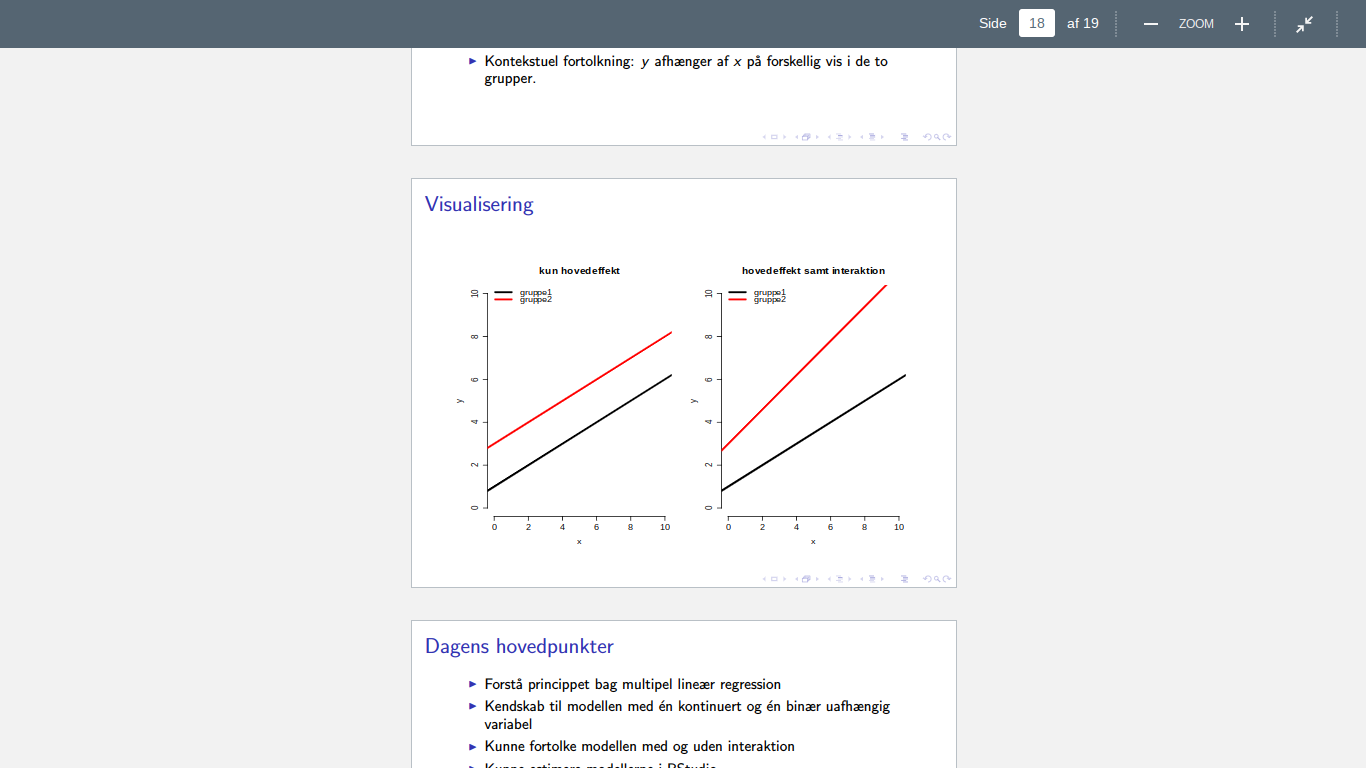
## Interaktionseffekt

Ved vores indikatorvariabel ser vi en begrænsning: effekten af x1 på y er den samme i de to grupper. Der ses kun en konstant forskydning. X2 har kun en hovedeffekt. Derfor indfører man et interaktionsled.

### Interaktionsled og interaktionseffekt



Nu har vi altså tilført en værdi, b3, som vil ændre hældningen på linjen.



## Koder i R til multibel lineær regression

|  |  |
| --- | --- |
| **Hvad vil du lave?** | **R-kode** |
| Lineærregression med multibel variabel, hoveffekt | lm(d$’afhængig’ ̴ d$’uafhængig’ + d$’kategorisk’ |

# Logistisk regression (SAU 6)

Bruges til at undersøge, hvordan forskellige uafhængige variable påvirker en binær afhængig variabel, fx. Caries ift køn. Det er en model for, hvordan sandsynligheden for en given hændelse afhænger af forskellige uafhængige variable.

## Regressionsparametre

Kan betragtes som odds-ratioer.

Transformation: Fra logOdds 🡪 Odds-ratio (den naturlige logaritme til logOdds)

## Skalaer til betragtning af logistisk regressionsmodel

Vi skal betragte den logistiske regressionsmodel på tre forskellige skalaer:

1. log Odds-skala – hvor modellen formuleres

2. Odds-skala – hvor vi beskriver effekterne af de uafhængige variable

3. Sandsynlighedsskala – hvor vi kan prædiktere absolutte risikoer

## Koder i R til logistisk regression

|  |  |
| --- | --- |
| **Hvad vil du lave?** | **R-kode** |
| Generealiseret lineær model | Data <- glm(d$afhængig ~ d$binær/uafhængig, family = binomial) (+ summary) |
| Transformation logOdds 🡪 odds-ratio | Exp(hældningen) |
| Prædikerede risiko/sandsynlighedskala | Exp(a + b) / (1+ exp (a+b)) |