



## Tosidet variansanalyse

Anders Tolver  
Institut for Matematiske Fag



## Dagens program

### Tosidet variansanalyse (ANOVA)

- Additive model (uden vekselvirkning)
- Model med vekselvirkning
- Forskel på additive effekter og vekselvirkning
- Test for vekselvirkning
- Forskellige parametriseringer (primært af den additive model)

### Generel info

- Afeveringsopgave til onsdag den 9. oktober
- Gamle eksamensopgaver: Kør selv analyserne hvis der er data
- HS-opgaver minder også om kommende eksamensopgaver



## Overblik

Vi skal have „udfyldt“ følgende skema over modeller (rækker) og statistiske begreber (søjler):

	Intro	Model	Est.+SE	KI	Test	Kontrol	Præd.
En stikprøve	✓	✓	✓	✓	✓	✓	✓
Ensidet ANOVA	✓	✓	✓	✓	✓	✓	✓
Lineær regr.	✓	✓	✓	✓	✓	✓	✓
To stikprøver	✓	✓	✓	✓	✓	✓	✓
Multipel regr.	✓	✓	✓	✓	✓	✓	✓
Tosidet ANOVA	nu	nu	nu	nu	nu	nu	nu
Blandede modeller							



## Tosidet ANOVA uden vekselvirkning



## Eksempel: Højde på studieretninger

Spørgeskema med studerende på Statistisk Dataanalyse 2017:  
bl.a. info om studieretning og højde.

- Svar fra 50 BB + 42 HV + 31 JØ + 31 NR + 2 andre. Skipper de "2 andre".
- Der mangler desuden højde for en mindre antal studerende  $\rightarrow n = 152$

Spørgsmål: Er den gennemsnitlige højde forskellig på studierne?

- Respons: Højde
- Forklarende variabel: Studieretning
- Lægger op til ensidet ANOVA



## Ensidet ANOVA

```
oneway <- lm(hojde ~ studie, data = useData)
onesample <- lm(hojde ~ 1, data = useData)
drop1(oneway, test = "F")

## Single term deletions
##
## Model:
## hojde ~ studie
##      Df Sum of Sq  RSS   AIC F value    Pr(>F)
## <none>                11299 662.91
## studie  3      1185.2 12484 672.07   5.1745 0.001985 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Har vi nu vist at "unge menneskers studievalg har noget med deres højde at gøre"? Eller **er der noget vi har overset?**



## Tosidet ANOVA

Køn påvirker (formentlig) både højde og studievalg.

Vores egentlige spørgsmål er nok snarere: Er der en forskel på højden på de fire studieretninger, selv hvis vi **justerer for køn?**

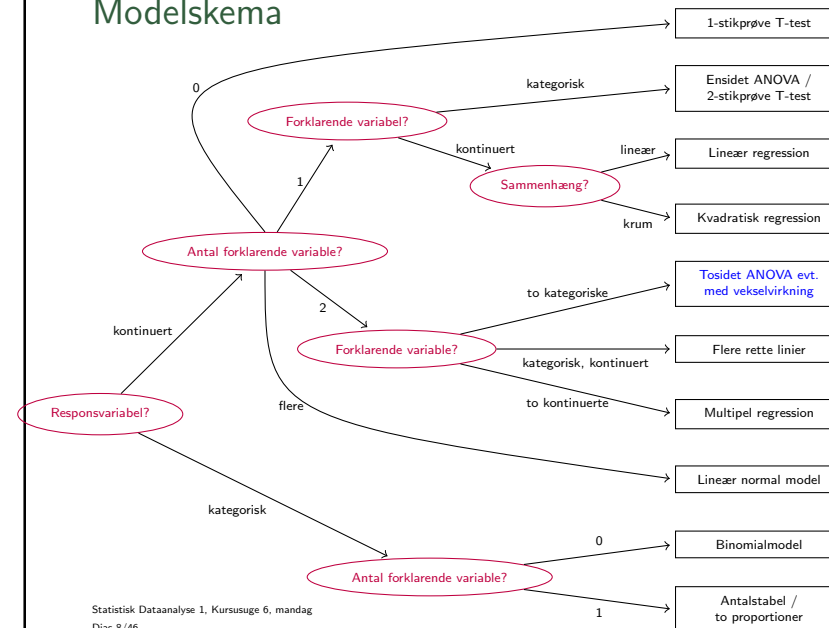
Ny analyse:

- Respons: Højde
- Forklarende var. Studieretning og køn. Begge er **kategoriske**
- **Tosidet ANOVA**

Check modelskemaet.



## Modelskema



## Statistisk model

Model for **tosidet ANOVA uden vekselvirkning**, kaldes også den **additive model** for tosidet ANOVA:

$$\text{højde}_i = \alpha_{\text{studie}_i} + \beta_{\text{kon}_i} + e_i$$

hvor  $e_i$ 'erne som sædvanlig er uafhængige  $N(0, \sigma^2)$

Parametre:

- Et  $\alpha$  per studie:  $\alpha_{J\emptyset}$ ,  $\alpha_{NR}$ ,  $\alpha_{HV}$ ,  $\alpha_{BB}$
- Et  $\beta$  per køn:  $\beta_M$  og  $\beta_K$
- Residualspredning  $\sigma$



## Additiv tosidet ANOVA

Vi **kan allerede det hele**: Estimation, modelkontrol, hypotesetest, konfidens- og prædiktionsintervaller fra uge 3–4.

R: Tilføj leddene til `lm`, med + imellem:

```
twoway.add <- lm(hojde ~ studie + kon, data=useData)
```

NB. Det er lidt sværere at bestemme antal frihedsgrader — men det klarer R heldigvis for os.

Hvad nu?

- **Modelkontrol**: Se dagens R-materiale
- **Fortolkning** af parameterestimater
- **Test** for studieretning når vi justerer for køn



## Fortolkning af parameterestimater

R vælger en **referencegruppe for hver variabel**. Her: BB og kvinder.

Følgende estimater aniges:

- „Intercept“: Estimeret middelværdi gives for **kombinationen** af de to referencer, altså for kvindelige BB-studerende
- Estimerede **forskelle** mellem de andre studieretninger og BB
- Estimeret **forskel** mellem mænd og kvinder



## Spørgsmål

- Estimat for gennemsnitshøjde blandt kvindelige BB-stud.?
- Estimat for gennemsnitshøjde blandt mandlige BB-stud.?
- Estimat for gennemsnitshøjde blandt mandlige JØ-stud.?
- Hvilket studie estimeres til at have de højeste studerende (når der er korigeret for køn)?
- Estimat for  $\sigma$ ?
- Antal frihedsgrader? Er det mærkeligt?
- Hvordan skal  $p$ -værdierne fortolkes?



## Additiv tosidet ANOVA

```
twoway.add <- lm(hojde ~ studie + kon, data=useData)
summary(twoway.add)

##
## Call:
## lm(formula = hojde ~ studie + kon, data = useData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.5701  -4.1354  -0.2316   4.0228  17.1185
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    168.1051     0.9859  170.517  <2e-16 ***
## studieHusdyrvidenskab    1.1211     1.3901   0.806   0.421
## studieJordbrugsøkonomi  -0.5350     1.5086  -0.355   0.723
## studieNaturressourcer    0.2531     1.4892   0.170   0.865
## konMand         14.5233     1.2590  11.535  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.352 on 147 degrees of freedom
## Multiple R-squared:  0.525, Adjusted R-squared:  0.512
## F-statistic: 40.61 on 4 and 147 DF,  p-value: < 2.2e-16
```



## Test for studieretning når vi justerer for køn

Statistisk model:

$$højde_i = \alpha_{studie_i} + \beta_{kon_i} + e_i$$

Hypotese:

$$H_0 : \alpha_{JØ} = \alpha_{NR} = \alpha_{HV} = \alpha_{BB}$$

Testes med *F*-test. Flere metoder i R, men med samme resultat:

- Fit stat. model + model under hypotese og brug anova med de to modeller som argumenter. Hvad er nulmodellen her?
- drop1: Kan vi "droppe" hvert af leddene fra modellen?
- Brug **ikke** anova med kun en model som argument



## Test for studieretning når vi justerer for køn: med drop1

```
twoway.add <- lm(hojde ~ studie + kon, data = useData)
drop1(twoway.add, test = "F")

## Single term deletions
##
## Model:
## hojde ~ studie + kon
##      Df Sum of Sq    RSS   AIC  F value Pr(>F)
## <none>                 5930.7 566.93
## studie  3      44.4  5975.1 562.06   0.3666 0.7772
## kon     1     5368.6 11299.3 662.91  133.0654 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



## Test for studieretning når vi justerer for køn: med anova

```
oneway.kon <- lm(hojde ~ kon, data=useData)
anova(oneway.kon, twoway.add)

## Analysis of Variance Table
##
## Model 1: hojde ~ kon
## Model 2: hojde ~ studie + kon
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1      150 5975.1
## 2      147 5930.7  3     44.371 0.3666 0.7772
```



## Konklusion

Der er **ikke** signifikant forskel på højden af studerende på de fire studieretninger når vi korrigerer for køn ( $p = 0.78$ ).

I denne situation var vi mest interesseret i den ene variabel (studieretning), men vi **kunne også have undersøgt den anden**:

- Hypotese,  $H_0 : \beta_M = \beta_K$
- Testes med  $F$ -test eller  $t$ -test. Begge giver  $p \approx 0$
- Konklusion: Gennemsnitshøjden **er** forskellig for mænd og kvinder, også når vi korrigerer for studieretning

Uden vekselvirkning: Vi lavede implicit en antagelse ...



## Antagelsen om additivitet

I eksemplet brugte vi den additive model, modellen uden vekselvirkning:

- Vi antog implicit at der var samme højdeforskel i mænd og kvinder for alle studieretninger (og omvendt).
- R: Vi skrev + mellem leddene i  $lm$ .

Men sådan er det jo ikke nødvendigvis: Det kan være at effekten af en variabel afhænger af den anden variabel  $\rightarrow$  **vekselvirkning**



## Additive effekter vs. vekselvirkning



## Prisskilt fra isbod

- 1 kugle ..... 15
- 2 kugler ..... 20
- 3 kugler ..... 23
- 1 kugle med guf ..... 19
- 2 kugler med guf ..... 24
- 3 kugler med guf ..... 27



## To ækvivalente prisskilte

### Prisskilt 1:

- 1 kugle ..... 15
- 2 kugler ..... 20
- 3 kugler ..... 23
- 1 kugle med guf ..... 19
- 2 kugler med guf ..... 24
- 3 kugler med guf ..... 27

### Prisskilt 2:

- 1 kugle, uden guf ..... 15
- 2 kugler ..... +5
- 3 kugler ..... +8
- med guf ..... +4

Seks forskellige is at vælge imellem, men **"effekterne" af guf og størrelse indgår additivt**. Guf koster altid 4 kr ekstra.

Dermed kan priserne beskrives med kun **fire parametre** ( $1 + 2 + 1$ )



## Eksempel med højdedata

Tilsvarende for den additive model for højdedata

- Der er otte kombinationer af studieretning og køn
- Men kun  $1 + 3 + 1 = 5$  parametre i den additive model: En for ref-gruppen, tre for studieretningsforskelle, en for kønsforskel.



## Vekselvirkning

Når effekten af én variabel af niveauet af en anden variabel, så siger man at der er **vekselvirkning** mellem de to variable.

Engelsk: **Interaction**

- Is: Ingen vekselvirkning mellem guf og kugler: Guf kostede 4 kr uanset antal kugler.  
Ækvivalent: Prisen for ekstra kugler er den samme uanset om der skal guf på eller ej.
- Højde: Antog at kønsforskellen er den samme på alle studier.  
Ækvivalent: Forskel ml. studier er den samme for begge køn.



## Prisskilte uden/med vekselvirkning

Nye priser giver rabat på guf hvis man køber store is:

### Gamle priser:

- 1 kugle ..... 15
- 2 kugler ..... 20
- 3 kugler ..... 23
- 1 kugle med guf ..... 19
- 2 kugler med guf ..... 24
- 3 kugler med guf ..... 27

### Nye priser:

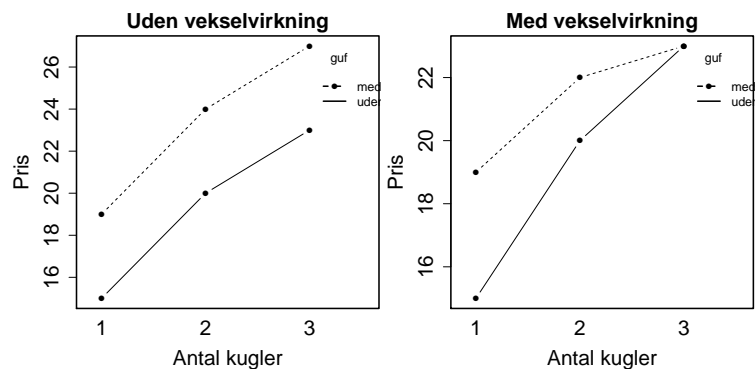
- 1 kugle ..... 15
- 2 kugler ..... 20
- 3 kugler ..... 23
- 1 kugle med guf ..... 19
- 2 kugler med guf ..... 22
- 3 kugler med guf ..... 23

**Nu er der vekselvirkning/interaktion!** Prisen for guf afhænger af antal kugler: 4/2/0 kr ved 1/2/3 kugler.

Det kræver **seks parametre** at beskrive den nye prisstruktur.



## Vekselvirkningsgraf/interaktionsplot

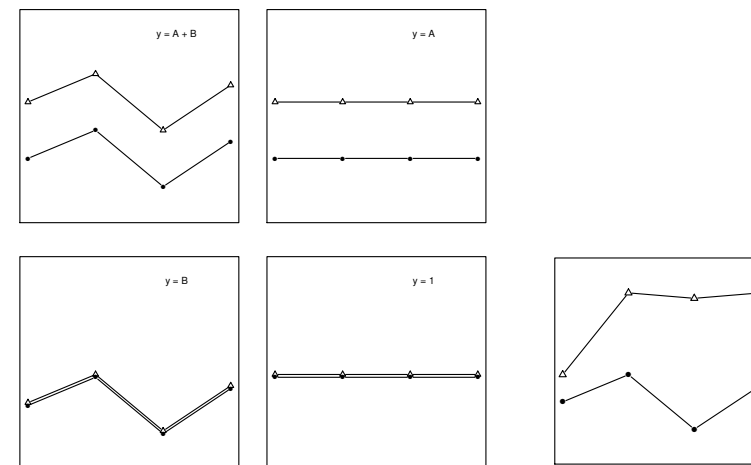


Plottet visualiserer vekselvirkning. Kig efter **parallelitet**:

- Parallelle profiler  $\leftrightarrow$  Ingen vekselvirkning
- Ikke-parallelle profiler  $\leftrightarrow$  Vekselvirkning



## Vekselvirkningsgraf/interaktionsplot, forventede værdier



## Tosidet ANOVA med/uden vekselvirkning



## Model uden vekselvirkning

Modellen uden vekselvirkning:

$$\text{højde}_i = \alpha_{\text{studie}_i} + \beta_{\text{kon}_i} + e_i$$

Modellen angiver middelværdien for alle otte kombinationer af studie og køn — men lægger **restriktioner** på dem.

Model uden vekselvirkning = additiv model.



## Model med vekselvirkning

Modellen med vekselvirkning lægger **ingen restriktioner** på de otte middelværdier. Vi skriver

$$\text{højde}_i = \alpha_{\text{studie}_i} + \beta_{\text{kon}_i} + \gamma_{\text{studie}_i, \text{kon}_i} + e_i$$

eller blot

$$\text{højde}_i = \gamma_{\text{studie}_i, \text{kon}_i} + e_i$$

Dette svarer faktisk til en ensidet ANOVA efter den variabel der inddeler obs. i otte grupper.

Opskrivningen med græske bogstaver ikke så vigtig. Vigtigt:

- at forstå den konceptuelle forskel mellem de to modeller
- at kunne fortolke output/estimer fra R



## Eksempel: Højde efter studieretning og køn

Ingen mandlige HV-studerende i datasættet:

- Lidt bøvlet når vi skal have vekselvirkning med  $\rightarrow$  vi dropper HV-studerende (selvom det faktisk ikke er nødvendigt)
- Datasættet useData2 indeholder data fra 110 studerende med højderegistreringer: 49 BB, 30 JØ, 31 NR.



## Modellen uden vekselvirkning

```
useData2 <- filter(useData, !(studie == "Husdyrvidenskab"))
twoway.add2 <- lm(højde ~ studie + kon, data = useData2)
summary(twoway.add2)

##
## Call:
## lm(formula = højde ~ studie + kon, data = useData2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.5701  -3.1051  -0.1051   3.8949  17.1185
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    168.1051     0.9841  170.825  <2e-16 ***
## studieJordbrugsøkonomi -0.5350     1.5059   -0.355    0.723
## studieNaturressourcer  0.2531     1.4866    0.170    0.865
## konMand         14.5233     1.2567   11.556  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.34 on 106 degrees of freedom
## Multiple R-squared:  0.5726, Adjusted R-squared:  0.5605
## F-statistic: 47.33 on 3 and 106 DF, p-value: < 2.2e-16
```



## Med vekselvirkning

```
twoway.int <- lm(højde ~ studie + kon + studie*kon, data=useData2)
round(summary(twoway.int)$coef, digits = 5)

##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    167.76471     1.09212  153.61443  0.00000
## studieJordbrugsøkonomi -0.45701     2.07657   -0.22008  0.82624
## studieNaturressourcer  1.66387     2.02220    0.82280  0.41251
## konMand         15.63529     1.97388    7.92109  0.00000
## studieJordbrugsøkonomi:konMand -0.64887     3.06611   -0.21163  0.83281
## studieNaturressourcer:konMand -3.06387     3.02956   -1.01132  0.31421
```





## Modelfit og fortolkning af estimater

Modelfit:

- Uden vekselvirkning: `lm(hojde ~ studie + kon, data=useData2)`
- Med vekselvirkning:  
`lm(hojde ~ studie + kon + studie*kon, data=useData2)`

Modellen uden vekselvirkning: Estimater læses som før, men er lidt anderledes da HV ikke er med mere.

Modellen med vekselvirkning:

- Hvorfor netop seks linier med estimater?
- Estimat for BB, kvinder? For JØ, kvinder? For JØ, mænd?



## Opsummering

Tosidet ANOVA efter to kategoriske variable, A og B:

- Model uden vekselvirkning:  $A+B$
- Model med vekselvirkning:  $A+B+A*B$
- Faktisk mange versioner af modellen med vekselvirkning:  $A+B+A:B$  eller  $A*B$  eller  $A:B$ . Prøv selv!

Estimater:

- R vælger referencegrupper for A og B (i de fleste versioner). Så er interceptet estimatet for referencekombinationen.
- Estimat for andre kombinationer: Interceptestimatet plus de relevante estimater.



## Test for vekselvirkning



## Er der faktisk vekselvirkning?

- Uformelt: Vekselvirkningsgraf/interaktionsplot
- Formelt: Hypotesetest

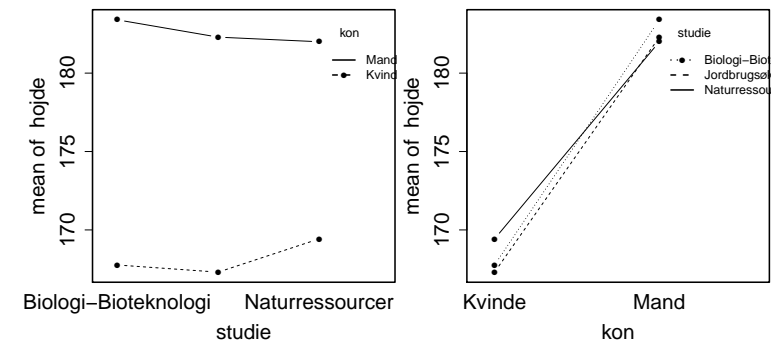


## Vekselvirkningsgraf/interaktionsplot

- Gennemsnit plottes med profiler med den ene variabel på x-aksen og med profiler for niveauerne af den anden var.
- Er profilerne **parallelle, på nær tilfældig variation?**
- Parallelle → tegn på at der ikke er vekselvirkning. Ikke-parallelle → tegn på at der er vekselvirkning.  
Under alle omstændigheder nyttig til at forstå samspillet.
- Svært at vurdere om ikke-parallelitet faktisk skyldes vekselvirkning eller blot tilfældig variation
- R: `interaction.plot` (se dagens R-kode)



## Vekselvirkningsgraf/interaktionsplot



- Profiler ser ganske parallelle ud, så næppe vekselvirkning
- Helt parallelle profiler på "den ene graf" ⇔ Helt parallelle profiler på "den anden graf"



## Hypotesetest

Model uden vekselvirkning er et **specialtilfælde** af model med vekselvirkning → de to modeller er nestede →  $F$ -test.

- Hypotese,  $H_0$ : Ingen vekselvirkning mellem studie og køn (dvs. kønseffekt den samme for alle studier, eller omvendt).
- Beskriver modellen med vekselv. faktisk data bedre end modellen uden vekselvirkning?
- Brug **anova** med de to modeller som argumenter, eller **drop1** på model med vekselvirkning.



## Konklusion

Der er ikke signifikant vekselv. mellem studie og køn ( $p = 0.59$ )

Vi ser defor nærmere på modellen uden vekselvirkning:

- Der er en sign. kønseffekt ( $p \approx 0$ ), men ikke en signifikant effekt af studieretning ( $p = 0.88$ ).
- Mænd estimeres til at være 14.5 cm (SE 1.26) højere end kvinder; 95% konfidensinterval (12.0, 17.0)



## R: Hypotesetest ved brug af anova

```
anova(twoway.add2, twoway.int)

## Analysis of Variance Table
##
## Model 1: hojde ~ studie + kon
## Model 2: hojde ~ studie + kon + studie * kon
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1      106 4261.1
## 2      104 4217.4  2      43.7 0.5388 0.5851
```



## R: Hypotesetest ved brug af drop1

```
drop1(twoway.int, test="F")

## Single term deletions
##
## Model:
## hojde ~ studie + kon + studie * kon
##           Df Sum of Sq    RSS    AIC F value Pr(>F)
## <none>                                4217.4 413.12
## studie:kon  2      43.7 4261.1 410.25  0.5388 0.5851
```



## Modellen uden vekselvirkning (- studerende på HV)

```
drop1(twoway.add2, test="F")

## Single term deletions
##
## Model:
## hojde ~ studie + kon
##           Df Sum of Sq    RSS    AIC F value Pr(>F)
## <none>                                4261.1 410.25
## studie  2      9.9 4271.1 406.50  0.1233 0.8841
## kon     1    5368.6 9629.7 497.93 133.5478 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(twoway.add2)$coef

##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    168.1051102   0.9840758  170.8253621 3.135005e-131
## studieJordbrugsøkonomi -0.5349840   1.5058537  -0.3552696  7.230936e-01
## studieNaturressourcer  0.2530765   1.4865576  0.1702433  8.651433e-01
## konMand         14.5233067   1.2567448  11.5562892  1.759654e-20
```



## Diverse om vekselvirkning

Vekselvirkning ml.  $A$  og  $B$  siger ikke at der er sammenhæng mellem  $A$  og  $B$ , men at effekten af  $A$  på  $y$  afhænger af  $B$ .

Vi taler om **hovedeffekter** og **vekselvirkning** af de to variable:

- Ofte ligger den primære interesse i hovedeffekterne, men sommetider er vekselvirkningen det primære
- Inddrag kun vekselvirkning hvis det giver faglig mening

Vekselvirkningsmodellen kræver **gentagelser**: Kan ikke fittes hvis der kun er en obs. for hver kombination af de to variable.



## Diverse om vekselvirkning

Det giver ikke mening af tale om effekten (bestemt form) af en variabel hvis den indgår i vekselvirkning med en anden:

- Fx kan man ikke bestemme estimatet for kønseffekten i modellen hvor studie og køn indgår med vekselvirkning
- Fx kan man ikke teste hovedeffekten af køn i modellen hvor studie og køn indgår med vekselvirkning



## Opsummering — til eget brug

- Hvornår kan man bruge tosidet ANOVA?
- Hvad betyder det at der vekselvirkning mellem to variable?
- Hvordan fitter du en tosidet ANOVA (med/uden vekselvirkning) i R, og hvordan bruger du estimerne?
- Hvordan undersøger man om de er vekselvirkning?

