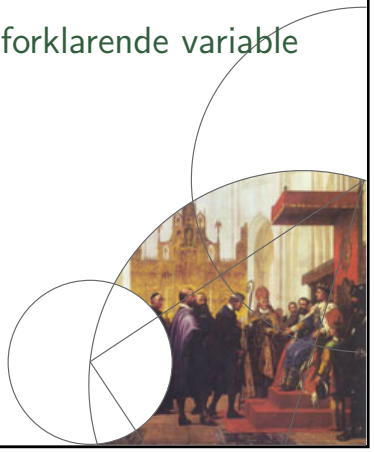
Blandede modeller:
- både kontinuerte og kategoriske forklarende variable

Anders Tolver
Institut for Matematiske Fag



Dagens program

Formiddag:

- Tosidet ANOVA: repetition
- Blandede modeller: både kategoriske og kvantitative forklarende variable
  - Eksempel: løbetider på DHL-stafetten
- Lineære modeller
  - Eksempel: lungefunktionsmålinger (FEV)

Eftermiddag:

- Diskussion af Quiz 6
- Hængepartier

Afleveringsopgave 3:
Opgave 1 fra eksamen i jan. 2019 kan afleveres ved øvelserne mandag d. 21/10

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	✓	✓	✓	✓	✓	✓	✓
Blandede modeller	nu	nu	nu	nu	nu	nu	nu

Tosidet ANOVA

Hvornår og hvordan?

Kontinuert respons og to kategoriske forklarende variable.

Vekselvirkning: Effekten af den ene variabel afhænger af den anden variabel, og vice versa.

Typisk work flow:

- Fit model med vekselvirkning — hvis det giver faglig mening og hvis der er gentagelser
- Modelkontrol (skal der fx transformeres?)
- Test for vekselvirkning
- Hvis vekselvirkning ikke er signifikant: Test for hovedeffekter
- Afrapportering af estimater og konfidensintervaller: Fra model med/uden vekselvirkning afhængig af konklusionen af testet.

Samme model - forskellige parametriseringer

Tosidet ANOVA med vekselvirkning:
Der er forskellige måder at afrapportere estimaterne på ...

```
twoway.int <- lm(hojde ~ studie + kon + studie*kon, data=useData2)
```

```
##                                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                    167.7647     1.0921 153.6144   0.0000
## studieJordbrugsøkonomi          -0.4570     2.0766  -0.2201   0.8262
## studieNaturressourcer            1.6639     2.0222   0.8228   0.4125
## konMand                         15.6353     1.9739   7.9211   0.0000
## studieJordbrugsøkonomi:konMand  -0.6489     3.0661  -0.2116   0.8328
## studieNaturressourcer:konMand   -3.0639     3.0296  -1.0113   0.3142
```

```
twoway.int2 <- lm(hojde ~ studie:kon - 1, data=useData2)
```

```
##                                    Estimate Std. Error  t value Pr(>|t|)
## studieBiologi-Bioteknologi:konKvinde 167.7647     1.0921 153.6144        0
## studieJordbrugsøkonomi:konKvinde     167.3077     1.7662  94.7283        0
## studieNaturressourcer:konKvinde      169.4286     1.7019  99.5503        0
## studieBiologi-Bioteknologi:konMand   183.4000     1.6442 111.5416        0
## studieJordbrugsøkonomi:konMand       182.2941     1.5445 118.0291        0
## studieNaturressourcer:konMand        182.0000     1.5445 117.8387        0
```

Den additive model

Test for vekselvirkning

```
twoway.add2 <- lm(hojde ~ studie + kon, data = useData2)
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
```

Estimater fra den tosidede ANOVA uden vekselvirkning

```
##                       Estimate Std. Error   t value Pr(>|t|)
## (Intercept)           168.10511    0.98408 170.82536  0.00000
## studieJordbrugsøkonomi  -0.53498    1.50585  -0.35527  0.72309
## studieNaturressourcer    0.25308    1.48656   0.17024  0.86514
## konMand                 14.52331    1.25674  11.55629  0.00000
```

Fortolkning af estimater? Hvorfor kun 4 parametre?

Blandede modeller

Hvad er blandede modeller?

Modeller der både indeholder kategoriske og kvantitative forklarende variable; stadig kontinuert respons.

- Respons y
- Kontinuert variabel x, kategorisk variabel `grp`

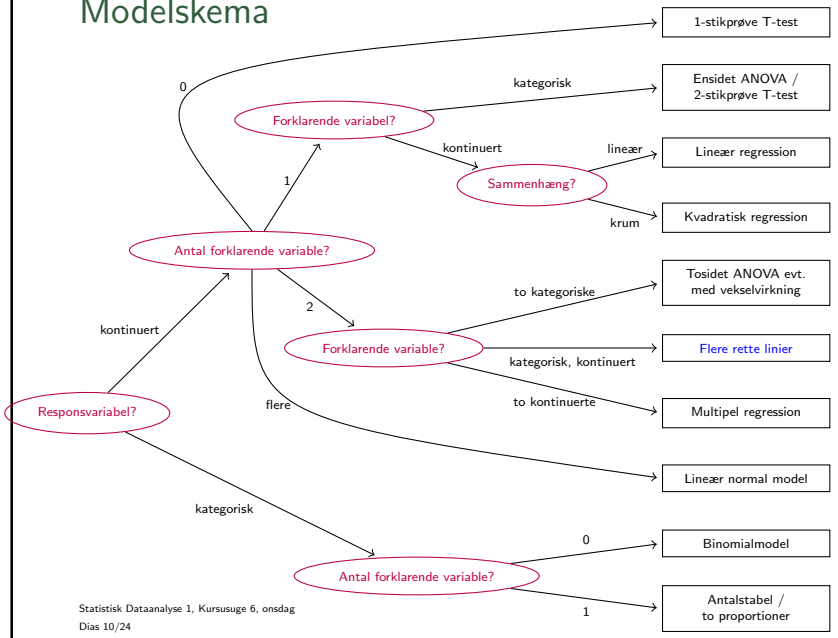
Statistisk model svarende til flere parallelle linier:

$$y_i = \alpha_{\texttt{grp}_i} + \beta \cdot x_i + e_i, \quad e_1, \ldots e_n \text{ iid. } N(0, \sigma^2)$$

Additiv model = model uden vekselvirkning.

Vekselvirkning, dvs. at effekten af den x afhænger af `grp`, svarer til ikke-parellelle linier. Det ser vi kun en lille smule på.

Modelskema



Analyse

R-syntaks:

```
lm(y ~ x + grp, data=...)
```

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

Eksempel: DHL-stafet

Data fra DHL-stafetten i 2006:

- Der blev løbet mandag–torsdag, 5000 hold hver dag.
- Hvert hold består af fem personer, frit sammensat af mænd og kvinder. Der kan altså være 0–5 kvinder på et hold.
- Alle personer løber 5 km; altså 25 km for hvert hold

Data ligger som **dhl** i *isdals*. Variable: `day`, `men`, `women`, `hours`, `minutes`, `seconds`.

Spørgsmål: Hvor meget langsommere løber kvinder end mænd? Er der forskel på dagene?

Analyse

De første overvejelser:

- Hvordan laver vi en fornuftig responsvariabel?
- Hvad skal vi bruge som forklarende variable? Hvilke typer?
- Hvilken figur kunne vi tænke os at lave?

Analyse:

- Statistisk model, modelkontrol
- Estimér hvor meget langsommere kvinder løber per km
- Hvilken dag var hurtigst/langsomst? Er der signifikant forskel på dagene?

Resultater:

- Kortfattede slides i dag - se i stedet dagens R program

Flere sjove spørgsmål

1. Ruten blev ændret mellem tirsdag og onsdag pga. kraftig regn.

- Undersøg om tiderne er ens mandag og tirsdag (en rute) og ens onsdag+torsdag (den anden rute)
- Bestem ét estimat for forskellen i løbstid mellem de to ruter

2. Vi har antaget at den forventede løbstid vokser lineært med antal kvinder på holdet. Kan vi undersøge om det faktisk er OK?

3. Vi har antaget hældningerne er ens de fire dage. Kan vi undersøge om det faktisk er OK?

DHL: Konklusion

Vi brugte tid som respons, antal kvinder (kvantitativ) og dag (kategorisk) som forklarende variable.

- Kvinders pace (kilometertid) estimeres til at være 0.81 minutter langsommere end mænds. 95% KI: $(0.74, 0.88)$.
- Der var signifikant forskel på dagene ($p = 0.000013$).

  Nærmere undersøgelser viste at der ikke var forskel mellem mandag og tirsdag, og mellem onsdag og torsdag ($p = 0.62$).
- Undersøgelser viste desuden ingen tegn på ikke-parellelitet ($p = 0.88$) ikke-linearitet ($p = 0.07$).

Lineære modeller

Modeller

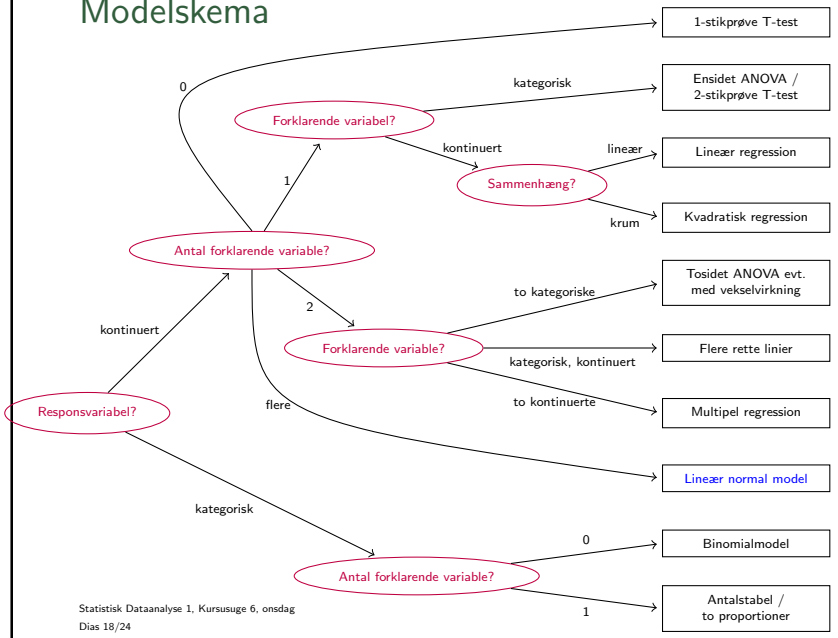
Vi har diskuteret dataanalyser med følgende karakteristika:

- 0/1/2 forklarende variable. Kategoriske/kontinuerte, med/uden vekselvirkning
- Kontinuert responsvariabel der antages at være normalfordelt givet den/de forklarende variable
- Uafhængige

Klassen af modeller kan udvides til flere forklarende variable og evt. vekselvirkninger mellem ≥ 2 variable.

Uafhængighed + normalfordeling + visse antagelser om middelværdierne $\rightarrow$ **lineær normal model**

Modelskema



Eksempel: Case 7 side 440

Data fra 654 børn.

- Respons: Lungefunktionsmåling (`FEV`)
- Forklarende variable: Alder (`Age`), Højde i tommer (`Ht`), Køn (`Gender`), rygning i hjemmet (0/1, `Smoke`)

Særligt interesseret i hvordan rygning i hjemmet påvirker lungefunktionen.

Modelovervejelser

Er de forklarende variable kategoriske eller kvantitative?

Lægefaglige overvejelser:

- Det er tænkeligt at rygning i hjemmet påvirker FEV forskelligt for drenge og piger.
- Man mener at FEV vokser med alderen, men at effekten muligvis aftager med alderen
- Man mener at barnets fysiske størrelse muligvis har en effekt på FEV, også selvom man tager alderen i betragtning

Hvordan kan/bør/skal de forklarende variable indgå i modellen?

Forslag til model

Forslag til modelformel:

```
FEV ~ Age + I(Age^2) + Ht + GenderFac + SmokeFac
          + SmokeFac*GenderFac
```

- Modelkontrol
- Estimater for effekt af rygning i hjemmet, hvert køn for sig
- Fælles estimat for effekt af rygning i hjemmet
- Er der signifikant effekt af rygning (sammenlign analyserne)

Kortfattede slides i dag - se i stedet dagens R program!

Hvad kan vi (ikke)?

Hvad kan vi?

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

Hvad kan vi, og hvad kan vi ikke?

Hvad har vi gjort:

- Har kun beskæftiget os med kontinuerte responsvariable, og kun modeller baseret på normalfordelingen
- Har kun beskæftiget os med uafhængige data
- De statistiske begreber er de samme uanset datatyperne

Sidste uge af kurset: Lidt om kategoriske responsvariable, men slet ikke så avancerede modeller som for kontinuerte data.

Snakker slet ikke om data med afhængighed; fx blokforsøg, tidsrækker og gentagne målinger → **kommer på StatData2**