

## Opsummering på kursusuge 1-4: (Ch. 1-7 i lærebogen)

Anders Tolver  
Institut for Matematiske Fag

Statistik Dataanalyse 1, Kursusuge 5, mandag  
Dias 1/33



## Dagens program

Vi repeterer de fleste ting fra kursusuge 1-4 (slides 1-17+R).

- **Lineær regression og ensidet ANOVA:**  
Eksempel delvist beskrevet i lærebogens afsnit 14.1 (s. 388)
- **R programmet:**  
fokus på fortolkning af output og kobling til generelle begreber

Øvrigt materiale (læses primært selv):

- (Video om) **Prædiktion:**  
Slide 28-35 + R-program fra 29/9-2021
- Lidt om nogle **Generelle begreber:**  
Modelspecifikation, modelfit, eksperimentelle vs. observationelle data mm. (slides 18-26)
- **Mere om hypotestest** (slides 27-33)

Statistik Dataanalyse 1, Kursusuge 5, mandag  
Dias 2/33



## Lineær regression og ensidet ANOVA: Ikke altid enten-eller

Statistik Dataanalyse 1, Kursusuge 5, mandag  
Dias 3/33



## Elektrisk ål, lever i Syamerika



Statistik Dataanalyse 1, Kursusuge 5, mandag  
Dias 4/33



## Lineær regression vs ensidet ANOVA

Vi har foreløbig skelnet ret skarpt mellem situationer hvor man kan bruge lineær regression og ensidet ANOVA. Men...

**Afsnit 14.1 (side 388):** Elektriske ål udsender elektriske signaler, men afhænger frekvensen af vandtemperaturen?

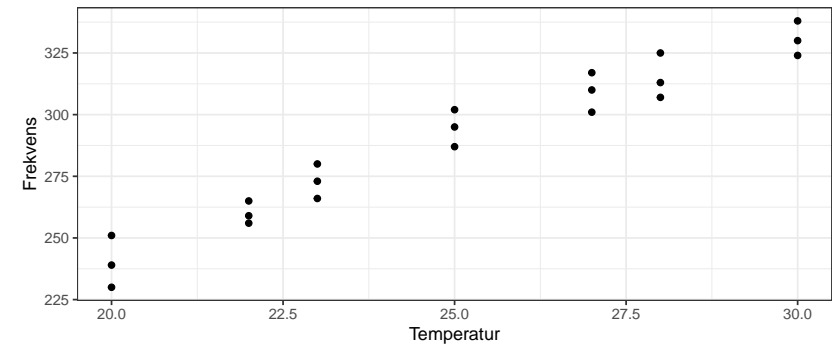
Data, ligger som *eels* i *isdals*:

- Syv vandtemperaturer, tre elektriske ål for hver temperatur
- Signalfrekvensen målt for hver af de 21 ål
- To variable: temp, freq

Datatyper? Tegn data! Hvilke modeller byder sig til?



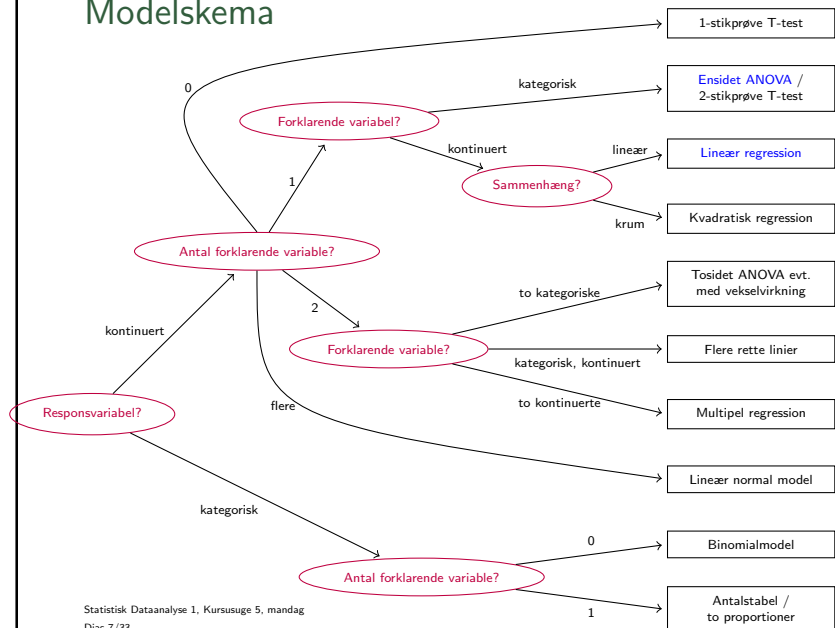
## Elektriske ål: Dataplot



- Ser rimeligt lineært ud → **lineær regression** er en mulighed
- Kan også tænke på temperatur som en inddeling af obs. i syv grupper → **ensidet ANOVA** er en mulighed



## Modelskema



## Spørgsmål

Opskrive modellerne.

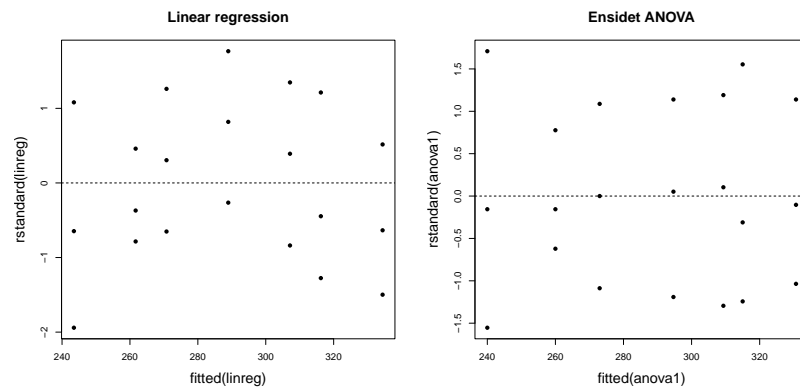
Kan vi estimere følgende størrelser vha. begge modeller?

- Forventet frekvens ved 22 grader
- Forventet frekvens ved 26 grader
- Forskel i forventet frekvens ved 20 og 30 grader
- Forskel i forventet frekvens ved temperaturstigning på 1 grad

Dette taler umiddelbart for den lineære regression. Men hvis nu sammenhængen ikke havde været approksimativt lineær, så...



## Residualplots for de to modeller



Kommentarer?



## Modellerne er nestede

Antagelserne om uafhængighed, normalfordelte restled og ens spredninger for alle observationer er helt ens for de to modeller.

Kun middelværdierne er forskellige for de to modeller:

- **Ensidet ANOVA:** Forventet frekvens er ens indenfor grupper, men der er ingen restriktioner på de syv middelværdier
- **Lineær regression:** Forventet frekvens er en lineær funktion af temperaturen

Lineariteten er en ekstra restriktion → lineær regression er et specialtilfælde af ensidet ANOVA → de to modeller er **nestede**.



## Test for linearitet: Mulighed 1

Vi kan teste den simple model mod den mere komplekse model.

I dette tilfælde svarer dette til at teste hypotesen om at middelværdien vokser lineært med temperaturen:

$$H_0 : \alpha_j = \gamma + \beta t_j$$

hvor  $\alpha_j$  er middelværdien i temperaturgruppe  $j$ , og  $\gamma$  og  $\beta$  er skæring og hældning.

- Vi får  $F_{\text{obs}} = 0.70$  der skal vurderes i  $F$ -fordelingen (5, 14) frihedsgrader. Dette giver  $p$ -værdien 0.63.
- Vi afviser ikke hypotesen; data passer fint med linearitet.

Dette test dur kun fordi vi har **gentagelser** for hver temperatur.



## Test for linearitet: Mulighed 2

Et andet test for linearitet: Sammenlign en lineær og en kvadratisk regression.

- Lineær regression:  $y_i = \gamma + \beta \cdot t_i$
- Kvadratisk regression:  $y_i = \gamma + \beta \cdot t_i + \delta \cdot t_i^2$

Den lineære regressionsmodel er et specialtilfælde af den kvadratiske regressionsmodel svarende til at  $\delta = 0$ .

To muligheder:

- Fit den kvadratiske regression og se på  $t$ -testet for hypotesen  $H_0 : \delta = 0$  vha. summary.
- Fit begge modeller, sammenlign dem med  $F$ -test vha. anova

Vi får  $p = 0.08$ , så heller ikke her kan vi ikke afvise linearitet.



## Opsummering

- Når der er gentagelser for hver værdi af  $x$ -variablen, kan vi både køre ensidet ANOVA og lineær regression
- Forskellen er om middelværdien antages at være lineær i  $x$  eller må være hvad som helst.
- Vi kan teste for linearitet ved at sammenligne de to modeller
- Man kan også teste for linearitet ved at sammenligne med kvadratisk regression—men kun hvis kvadratisk model er OK.



## R: lineær regressionsmodel

```
library(iscals)
data(eels)
linreg <- lm(freq ~ temp, data = eels)
summary(linreg)

##
## Call:
## lm(formula = freq ~ temp, data = eels)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.492  -4.768  -1.952   6.048  13.048
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  61.6497    12.6431   4.876 0.000105 ***
## temp         9.0921     0.5014  18.134 1.87e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.571 on 19 degrees of freedom
## Multiple R-squared:  0.9454, Adjusted R-squared:  0.9425
## F-statistic: 328.8 on 1 and 19 DF,  p-value: 1.875e-13
```



## R: ensidet ANOVA

```
eels <- transform(eels, tempFac = factor(temp))
anova1 <- lm(freq ~ tempFac, data=eels)
summary(anova1)

##
## Call:
## lm(formula = freq ~ tempFac, data = eels)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0000  -6.6667  -0.6667   7.0000  11.0000
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  240.000    4.551   52.732 < 2e-16 ***
## tempFac22    20.000    6.437   3.107 0.007720 **
## tempFac23    33.000    6.437   5.127 0.000154 ***
## tempFac25    54.667    6.437   8.493 6.78e-07 ***
## tempFac27    69.333    6.437  10.772 3.69e-08 ***
## tempFac28    75.000    6.437  11.652 1.36e-08 ***
## tempFac30    90.667    6.437  14.086 1.17e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.883 on 14 degrees of freedom
## Multiple R-squared:  0.9564, Adjusted R-squared:  0.9377
## F-statistic: 51.14 on 6 and 14 DF,  p-value: 1.004e-08
```



## R: test for linearitet mod ensidet ANOVA

```
anova(linreg, anova1)

## Analysis of Variance Table
##
## Model 1: freq ~ temp
## Model 2: freq ~ tempFac
##      Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1         19 1089
## 2         14  870    5    219.02 0.7049 0.6292
```



## R: test for linearitet mod kvadratisk model

```
kvreg <- lm(freq ~ temp + I(temp^2), data=eels)
summary(kvreg)$coefficients

##              Estimate Std. Error   t value    Pr(>|t|)
## (Intercept) -126.2885503 102.3259298  -1.234179  0.233007469
## temp        24.3929671   8.2876571   2.943289  0.008692145
## I(temp^2)    -0.3060172   0.1654839  -1.849227  0.080916950

anova(linreg, kvreg)

## Analysis of Variance Table
##
## Model 1: freq ~ temp
## Model 2: freq ~ temp + I(temp^2)
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1      19 1089.02
## 2      18  915.16  1    173.86 3.4196 0.08092 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



## Modelspecifikation, modelfit, eksperimentelle/observationelle data



## Hvad betyder det at angive/specificere en model?

Flere muligheder men **fortæl altid hvad de indgående variable dækker over.**

1. Opskriv en ligning for observation  $i$  og angiv antagelserne for restleddene:

$$y_i = \text{middelværdi}_i + e_i$$

hvor  $e_1, \dots, e_n$  er iid.  $N(0, \sigma^2)$ .

Det relevante udtryk for middelværdien skal indsættes!

2. Skriv at  $y_i$ 'er er uafhængige og at  $y_i$  er normalfordelt med middelværdi \*\*\* og spredning  $\sigma$ .

Det relevante udtryk for middelværdien skal indsættes!

3. Fortæl hvilken slags analyse der er tale om, hvilken variabel der bruges som responsvariabel og hvilken/hvilke variable der bruges som forklarende variabel/variable.



## Eksempel: Lineær regression - elektriske ål

1. Hvis  $\text{freq}_i$  og  $\text{temp}_i$  er frekvens og vandtemperatur for målinger på ål  $i$ , så antager vi at

$$\text{freq}_i = \alpha + \beta \cdot \text{temp}_i + e_i$$

hvor  $e_1, \dots, e_n$  er iid.  $N(0, \sigma^2)$ .

2. Frekvensen ( $\text{freq}$ ) for ålene er uafhængige og frekvensen for ål  $i$  er normalfordelt med middelværdi  $\alpha + \beta \cdot \text{temp}_i$  og spredning  $\sigma$ . Her angiver  $\text{temp}$  vandtemperaturen.
3. Vi bruger en lineær regressionsmodel med frekvens som responsvariabel og vandtemp. som forklarende variabel.



## Hvad betyder det at angive/specificere en model?

- Det skal, på den ene eller den anden måde, være klart hvad du bruger som responsvariabel og forklarende variabel/variable.
- Hvis du bliver bedt eksplicit om at redegøre for forudsætningerne i modellen, så er mulighed 3 ikke godt nok.
- Hvis du senere refererer til parametre, fx  $\mu$ ,  $\alpha$ ,  $\beta$ ,  $\alpha_1, \dots, \alpha_k$ :  
Det skal være klart fra modelopskrivning eller anden tekst hvad parameteren dækker over.
- Ensidede ANOVA:  $y_i = \alpha_{g(i)} + e_i$ . Her er  $g$  den funktion som tager obs.nummer og fortæller hvilken gruppe obs. kommer fra.



## Hvad betyder det at fitte en model?

- At **fitte en model** betyder bare at køre den relevante `lm`-kommando
- Kig altid på et **summary** af modellen og check at det ser ud som du forventer, fx at antal parametre er korrekt.
- Bagefter kan du lave modelkontrol og overveje hvad der er relevant at kigge nærmere på



## Hvad betyder det at udføre en ensidet ANOVA?

At lave/udføre en ensidet ANOVA dækker over hele analysen:

- Overvejelser om model (hvilke variable)
- Modelfit + modelkontrol
- Typisk en overall sammenligning af alle grupperne
- Angivelse af relevante estimater, SE'er, konfidensintervaller
- Evt. relevante parvise sammenligninger

Har indtil videre været lidt sløset med terminologien og brugt *ensidet ANOVA* til også at dække data-setup, modellen og andre enkeltdele.

Ved eksamen: Jeg kunne i princippet fortsætte og blot sige *Analyser data* — men det gør jeg ikke. I skal svare på specifikke spørgsmål.



## Eksperimentielle data

Vi har mest snakket om **data fra eksperimenter** eller på anden måde kontrolleret dataindsamling:

- (Nogenlunde) veldefinerede populationer
- Randomisering for at undgå utilsigtet sammenblanding af effekter (konfundering)
- Vi kan lave kausale konklusioner, altså konklusioner om årsagssammenhænge:  $y$  ændrer sig **fordi** vi har ændret på ...



## Observationelle data

Vi har ikke kigget på **observationelle data**:

- Ingen kontrol over den forklarende variabel
- Registerdata og andre situationer hvor vi ikke kan, eller det ikke er etisk OK at lave interventioner (fx effekt af rygning)
- Data vedr. økonomi er oftest observationelle(!)
- Vi kan snakke om associationer mellem variable og lave prædiktioner, men **ikke** umiddelbart lave kausale konklusioner

Det er et hot emne i statistiskforskning, hvilke konklusioner man faktisk kan komme med når man har observationelle data (og hvordan man bør gøre det)



## Grænselandet

Vi har bevæget os lidt på grænsen mellem de to i lineær regression, fx i eksemplet med kattes krops- og hjertevægt.

- Vi har målt begge dele på kattene, uden kontrol over kropsvægten. Variable indgår symmetrisk i dataindsamlingen.
- Vi kan ikke sige at hjertevægten stiger **fordi** kropsvægten stiger.
- Sagen er snarere at der er en ikke-målbar variabel, *størrelse*, der påvirker begge dele. **Confounder**.
- Vi kan dog godt tale om sammenhæng/association, lave prædiktioner, og estimere forskelle i Hwt ved forskelle i Bwt.

Hjemmeopgave:

Hvordan kunne man lave et (uetisk?) eksperiment for at undersøge, om hjertevægten stiger som en konsekvens af øget kropsvægt?



## Mere om hypotesetest



## Hypotesetest

Ingredienser i et hypotesetest:

- **Signifikansniveau**, som regel 0.05
- **Nylhypotesen**,  $H_0$
- **Teststørrelse**:  $T$  eller  $F$  (der findes andre...)
- **$p$ -værdi**: Sandsynligheden for at få data der passer lige så dårligt eller dårligere med  $H_0$  hvis  $H_0$  faktisk er sand
- **Konklusion**: Forkaster  $H_0$  hvis  $p < \text{signifikansniveau}$ , ellers ikke



## Den alternative hypotese, $H_A$

Faktisk er der en ingrediens mere, **den alternative hypotese  $H_A$** .

$H_A$  angiver det der er sandt hvis  $H_0$  er falsk, altså det vi konkluderer hvis vi afviser hypotesen.

Eksempler:

- $H_0 : \mu = 0$  og  $H_A : \mu \neq 0$
- $H_0 : \alpha_1 = \dots = \alpha_k$  og  $H_A$  : mindst to  $\alpha_j$ 'er er forskellige

Sommetider bruges mere **specifikke alternativer**, fx  $H_A : \mu > 0$ .

Så taler negative værdier af  $\hat{\mu}$  for hypotesen, ikke imod hypotesen, uanset hvor langt væk fra 0 den ligger.

Hvis man ikke specificerer  $H_A$  eksplicit, mener man blot "det modsatte af  $H_0$ , sådan som vi har gjort indtil nu.



## Hypoteser er restriktioner på modellen

En hypotese består af en eller flere restriktioner på den statistiske model.  
Eksempler:

- En parameter har en specifik, præspecificeret værdi
- Flere parametre er ens (uden at den fælles værdi angives)

I normalfordelingsmodellerne kan alle sådanne hypoteser testes med  $F$ -test.

Hvis hypotesen kan beskrives med en et enkelt ligestegn, fx  $\mu = 0$  eller  $\alpha_3 = \alpha_4$ , så kan man også benytte et  $t$ -test.

I tilfælde, hvor man kan begge dele gælder at

**$p$ -værdien er altid den samme for  $t$ -testet of  $F$ -testet.**

Se eksempel om elektriske ål!



## Nestede modeller

Den statistiske model + hypotesen beskriver en ny statistisk model. Denne model kaldes nulmodellen (null model).

Nulmodellen er altså et specialtilfælde af den oprindelige model, eller en undermodel (sub model). Modellerne er **nestede**.

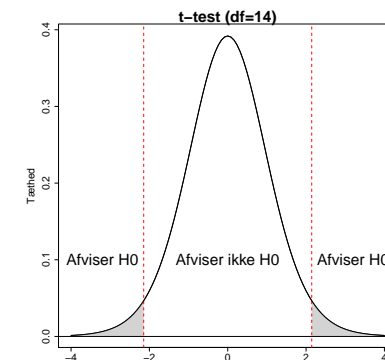
Vi kan **teste den simple model mod den mere komplekse**:

- Udgangspunktet er at den oprindelige model er OK
- Hypotesen er at den simple model kan beskrive data lige så godt som den mere komplekse
- Hvis hypotesen forkastes, konkluderer vi at den komplekse model faktisk beskriver data bedre end den simple

Hvordan? Fit begge modeller med `lm`, og udfør testet med `anova`. Se eksempel med elektriske ål.



## Hvornår afviser vi hypotesen ved et $t$ -test?

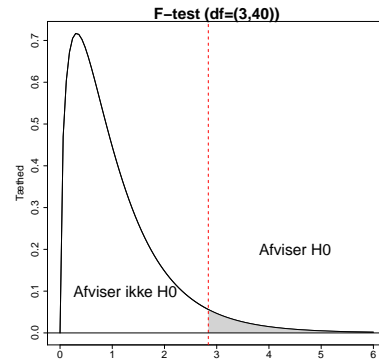


- Store og små værdier **kritiske** (fører til afvisning af  $H_0$ )
- Streger: 2.5% og 97.5% fraktilen i den relevante  $t$ -fordeling
- Afviser  $H_0$  hvis  $T_{\text{obs}}$  er længere fra 0 end disse værdier





## Hvornår afviser vi hypotesen ved et $F$ -test?



- Store værdier **kritiske** (fører til afvisning af  $H_0$ )
- Streg: 95% fraktilen i den relevante  $F$ -fordeling
- Afviser  $H_0$  hvis  $F_{\text{obs}}$  er større end denne værdi

