



Ensidet variansanalyse (flere grupper) og lineær regression

Anders Tolver
Institut for Matematiske Fag

Statistisk Dataanalyse 1, Kursusuge 3, mandag
Dias 1/27



Opsummering og dagens program

Kursusuge 1 + 2:

- Datatyper og deskriptiv statistik
- Normalfordelingen
- Lineær regression og ensidet ANOVA: Figurer og estimater — men ikke mere
- Én stikprøve: Statistisk model, estimation og standard errors, konfidensintervaller

I dag:

Statistisk model, estimation og SE, konfidensintervaller for

- Ensidet ANOVA, dvs. flere stikprøver
- Lineær regression
- **Repeter selv:** en enkelt stikprøve (fra 13/9-2023)

Statistisk Dataanalyse 1, Kursusuge 3, mandag
Dias 2/27



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	✓	nu	nu	nu			
Lineær regr.	✓	nu	nu	nu			
To stikprøver							
Multipel regr.							
Tosidet ANOVA							

Statistisk Dataanalyse 1, Kursusuge 3, mandag
Dias 3/27



Statistiske begreber

Statistiske grundbegreber indtil videre:

- Population og stikprøve
- Gennemsnit, stikprøvespredning, median, kvartiler
- Statistisk model og parametre
- Estimer og standard error (SE) for estimer
- Konfidensinterval

Statistisk Dataanalyse 1, Kursusuge 3, mandag
Dias 4/27



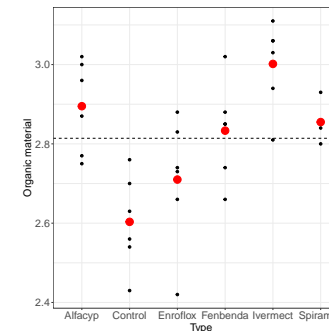
Ensidet ANOVA — flere stikprøver



antibio-datasættet

```
library(isdals)
data(antibio)
head(antibio, n = 7)
```

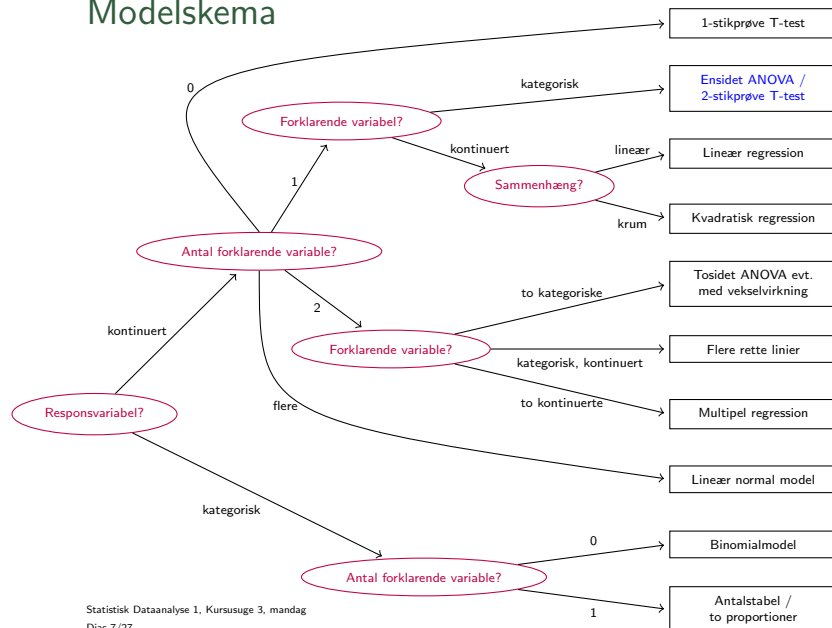
```
##      type org
## 1 Ivermect 3.03
## 2 Ivermect 2.81
## 3 Ivermect 3.06
## 4 Ivermect 3.11
## 5 Ivermect 2.94
## 6 Ivermect 3.06
## 7 Alfacyc 3.00
```



- Respons: Mængden af organisk materiale efter otte uger
- Modelschema: Kont. respons, én kategor. forklarende var.
- Ensidet ANOVA, flere stikprøver



Modelskema



Problemformulering og gruppegennemsnit

Kan vi generalisere ud fra data og sige, at det forventede indhold af organisk stof afhænger af den anvendte type antibiotika?

```
library(tidyverse)
summarize(group_by(antibio, type), n = n()
           , mean_org = mean(org), sd_org = sd(org))
```

```
## # A tibble: 6 x 4
##   type      n mean_org sd_org
##   <fct> <int>   <dbl> <dbl>
## 1 Alfacyc     6     2.90  0.117
## 2 Control     6     2.60  0.119
## 3 Enroflox    6     2.71  0.162
## 4 Fenbenda    6     2.83  0.124
## 5 Ivermect    6     3.00  0.109
## 6 Spiramyc    4     2.86  0.0545
```

Vi har brug for at kende usikkerheden på gruppegennemsnit!



Statistisk model

Data:

- y_1, \dots, y_n kvantitative, kontinuerte obs. fra k grupper
- $g(i)$ er gruppen hørende til måling i

Statistisk model:

- y_1, \dots, y_n er uafhængige
- y_i er normalfordelt $\sim N(\mu_i, \sigma^2)$
- middelværdien $\mu_i = \alpha_{g(i)}$ afhænger af gruppen $g(i)$

(Ukendte) populationsparametre:

- Hver gruppe antages at have sin egen middelværdi (forventede værdi): $\alpha_1, \dots, \alpha_k$
- Spredning σ ens for alle grupper

Middelværdierne $\alpha_1, \dots, \alpha_k$ og spredningen σ er **parametre** i modellen, som vi vil udtale os om ud fra de givne data.



R-output fra ensidet ANOVA

Lad os se på `summary()` fra en ensidet variansanalysemodel:

```
modell1 <- lm(org ~ type - 1, data = antibio)
summary(modell1)

##
## Call:
## lm(formula = org ~ type - 1, data = antibio)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29000 -0.06000  0.01833  0.07250  0.18667
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## typeAlfacyp    2.89500    0.04970   58.25  <2e-16 ***
## typeControl    2.60333    0.04970   52.38  <2e-16 ***
## typeEnroflox    2.71000    0.04970   54.53  <2e-16 ***
## typeFenbenda    2.83333    0.04970   57.01  <2e-16 ***
## typeIvermect    3.00167    0.04970   60.39  <2e-16 ***
## typeSpiramyc    2.85500    0.06087   46.90  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1217 on 28 degrees of freedom
## Multiple R-squared:  0.9985, Adjusted R-squared:  0.9981
## F-statistic: 3034 on 6 and 28 DF, p-value: < 2.2e-16
```



Estimation

Estimator:

- For middelværdier: $\hat{\alpha}_j = \bar{y}_j$ — gruppegennemsnit
- Den fælles spredning: $\hat{\sigma} = s$ — sammenvejet spredning. Hvordan beregnes denne fælles spredning?

Interesseparameter er ofte **forskelle mellem grupperne**, fx $\alpha_2 - \alpha_1$. Estimeres med $\hat{\alpha}_2 - \hat{\alpha}_1 = \bar{y}_2 - \bar{y}_1$.

Men hvor meget kan vi stole på estimatorne?

- Standard error for $\hat{\alpha}_j$? For $\hat{\alpha}_2 - \hat{\alpha}_1$?
- Konfidensinterval for α_j ? For $\alpha_2 - \alpha_1$?

Repetition (fra onsdag i kursusuge 2):

- Hvad mener vi med **standard error for estimat**?
- Hvordan fandt vi standard error for estimatet for middelværdien i situationen med en enkelt stikprøve?



Standard errors for estimator

Standard error for estimat = (estimeret) spredning for estimatet

Husk at $\hat{\alpha}_j = \bar{y}_j$ er gennemsnit af n_j observationer. Derfor:

$$SE(\hat{\alpha}_j) = \frac{s}{\sqrt{n_j}}$$

Desuden: $SE(\hat{\alpha}_2 - \hat{\alpha}_1)^2 = SE(\hat{\alpha}_2)^2 + SE(\hat{\alpha}_1)^2$, så

$$SE(\hat{\alpha}_2 - \hat{\alpha}_1) = s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Igen vigtigt at skelne mellem s og $SE(\hat{\alpha}_j)$:

- s : spredning på **enkeltobs**. Residual standard error.
- $SE(\hat{\alpha}_j)$ og $SE(\hat{\alpha}_2 - \hat{\alpha}_1)$: spredning på **estimator**



Konfidensintervaller

Vil gerne have **konfidensintervaller** for middelværdier og deres forskelle. Har ingredienserne!

$$95\% \text{ KI : } \text{estimat} \pm t_{0.975, df} \cdot \text{SE}(\text{estimat})$$

Hvor mange **frihedsgrader**?

- $df = n - k = \text{antal obs. minus antal middelværdiparametre}$

I **R-programmet** bør du have fokus på:

- Hvordan benyttes `qt()`-funktionen til beregning af $t_{0.975, df}$?
- Hvor (og hvornår) kan man aflæse $\text{SE}(\text{estimat})$ direkte i R-output?
- Hvordan (og hvornår) kan man bruge `confint()`-funktionen til beregning af konfidensintervaller?



Quiz: R-output fra ensidet ANOVA

Lad os se lidt mere på output fra `model1`:

```
summary(model1)$coefficients
```

```
##              Estimate Std. Error  t value    Pr(>|t|)
## typeAlfacyp  2.895000  0.04970149  58.24775 9.090203e-31
## typeControl  2.603333  0.04970149  52.37938 1.731916e-29
## typeEnroflox 2.710000  0.04970149  54.52553 5.685971e-30
## typeFenbenda 2.833333  0.04970149  57.00701 1.653035e-30
## typeIvermect 3.001667  0.04970149  60.39390 3.326166e-31
## typeSpiramyc 2.855000  0.06087164  46.90197 3.689834e-28
```

- Hvordan er tallene 2.895 og 2.710 i søjlen Estimate udregnet fra datasættet?
- Hvad er fortolkningen af tallet 0.0497 i søjlen Std.Error?
- Hvorfor er Std.Error for Spiramyc større end for andre grupper?
- Find et 95 % - konfidensinterval for det forventede indhold af organisk stof i en prøve hørende til `type = Alfacyp`.



Samme model kan fittes på flere måder i R

Med gruppemiddelværdierne som parametre:

```
model1 <- lm(org ~ type - 1, data=antibio)
```

Med en referencegruppe valgt af R:

```
model2 <- lm(org ~ type, data=antibio)
```

Med en selvvalgt referencegruppe:

```
antibio$myType <- relevel(antibio$type, ref="Control")
model3 <- lm(org ~ myType, data=antibio)
```

Selvvalgt ref-gruppe, hvis data er indlæst med `read_excel`:

```
antibio$myType <- relevel(factor(antibio$type),
                          , ref="Control")
model3 <- lm(org ~ myType, data=antibio)
```



Version med referencegruppe

Hvilke forskelle ses i forhold til `summary(model1)`?

```
model2 <- lm(org ~ type, data = antibio)
summary(model2)

##
## Call:
## lm(formula = org ~ type, data = antibio)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29000 -0.06000  0.01833  0.07250  0.18667
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.895000    0.04970   58.248 < 2e-16 ***
## typeControl  -0.29167    0.07029  -4.150 0.000281 ***
## typeEnroflox -0.18500    0.07029  -2.632 0.013653 *
## typeFenbenda -0.06167    0.07029  -0.877 0.387770
## typeIvermect  0.10667    0.07029   1.518 0.140338
## typeSpiramyc -0.04000    0.07858  -0.509 0.614738
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1217 on 28 degrees of freedom
## Multiple R-squared:  0.5874, Adjusted R-squared:  0.5137
## F-statistic: 7.973 on 5 and 28 DF,  p-value: 8.953e-05
```



Version med referencegruppe

Hvordan skal vi fortolke konfidensintervallerne her?

```
confint(model2)
```

##	2.5 %	97.5 %
## (Intercept)	2.79319111	2.99680889
## typeControl	-0.43564618	-0.14768716
## typeEnroflox	-0.32897951	-0.04102049
## typeFenbenda	-0.20564618	0.08231284
## typeIvermect	-0.03731284	0.25064618
## typeSpiramyc	-0.20097398	0.12097398

Manuel beregning af konfidensintervaller (eksempler):

```
# (Intercept) / Alfacyc
2.895 + c(-1, 1) * qt(0.975, df = 28) * 0.04970
## [1] 2.793194 2.996806
# typeControl / forskel: Control - Alfacyc
-0.29167 + c(-1, 1) * qt(0.975, 28) * 0.07029
## [1] -0.4356525 -0.1476875
```

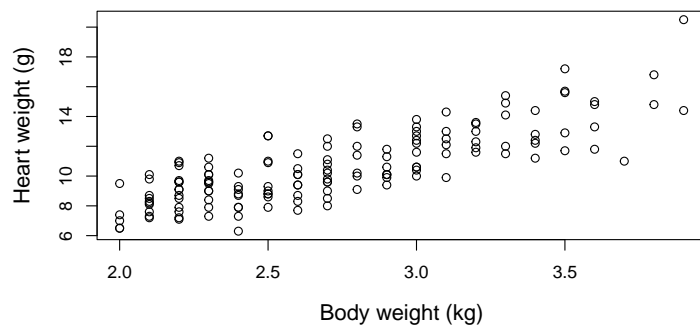


Lineær regression



Data: Kattes hjerte- og kropsvægt

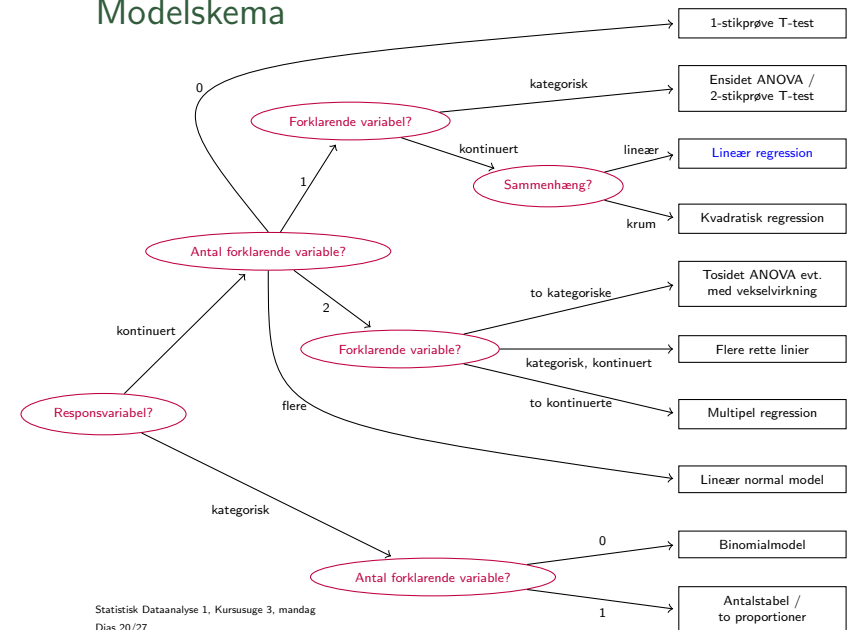
Cats data



Tilnærmelsesvis lineær sammenhæng, pånær tilfældig variation.



Modelskema



Statistisk model

Data: Par $(x_1, y_1), \dots, (x_n, y_n)$, kvantitativ, kontinuert

Statistisk model: Uafhængighed + alle obs. normalfordelt med middelværdi givet ved ret linie og samme spredning omkring linie

Formelt:

- Tænker på x_i 'erne som givne
- y_1, \dots, y_n uafhængige
- y_i normalfordelt med middelværdi $\mu_i = \alpha + \beta x_i$ og spredning σ .

(Ukendte) **populationsparametre:**

- Skæring/intercept α , hældning β
- Spredningen σ

Parametrene i modellen er α, β og spredningen σ , som vi vil udtale os ud fra de givne data.



Quiz: R-output fra lineær regression

Lad os se på `summary()` fra en lineær regressionsmodel:

```
linreg <- lm(Hwt ~ Bwt, data = cats)
summary(linreg)

##
## Call:
## lm(formula = Hwt ~ Bwt, data = cats)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.5694 -0.9634 -0.0921  1.0426  5.1238
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.3567      0.6923  -0.515   0.607
## Bwt           4.0341      0.2503  16.119 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.452 on 142 degrees of freedom
## Multiple R-squared:  0.6466, Adjusted R-squared:  0.6441
## F-statistic: 259.8 on 1 and 142 DF,  p-value: < 2.2e-16
```

- Angiv estimater for regressionslinjens hældning og skæring?
- Hvad er fortolkningen af **Residual standard error**
- Angiv et 95 % - konfidensinterval for parameteren β ?



Hvordan udregnes estimater?

Estimater for α og β via mindste kvadraters metode: $\hat{\alpha}, \hat{\beta}$.

Estimeret regressionslinie:

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

Estimat for σ :

$$\hat{\sigma} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2}$$

Men hvor meget kan vi stole på estimaterne?

- Standard error for $\hat{\alpha}, \hat{\beta}, \hat{y}$
- Konfidensinterval for $\alpha, \beta, \alpha + \beta x$



Standard errors for estimator

Formler:

$$SE(\hat{\beta}) = \frac{s}{\sqrt{SS_x}}, \quad SE(\hat{\alpha}) = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}},$$

$$SE(\hat{y}) = s \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{SS_x}}$$

hvor $SS_x = \sum (x_i - \bar{x})^2$.

Formlerne er stort set uinteressante, men:

- Husk at SE er udtryk for præcisionen af estimaterne
- Er det bedst at samle x 'erne eller at sprede dem?
- For hvilken værdi er \hat{y} mest præcist estimeret (mindst SE)?



Konfidensintervaller

Vil gerne have **konfidensintervaller** for parametre og estimeret regressionslinie:

$$95\% \text{ KI : } \text{estimat} \pm t_{0.975, df} \cdot \text{SE}(\text{estimat})$$

Hvor mange frihedsgrader?

- $df = n - 2 = \text{antal obs. minus antal middelværdiparametre}$

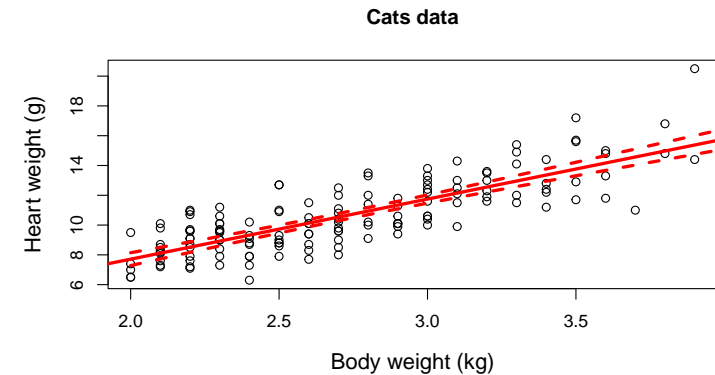
Eksempler på beregning af konfidensintervaller findes i dagens R-program.

I **R-programmet** bør du have fokus på:

- Hvordan man finder konfidensintervaller for α og β ved brug af `confint()`.
- Hvordan man finder konfidensintervallet for et punkt $\hat{y}\hat{\alpha} + \hat{\beta}x$ på linjen.



Estimeret linje med 95 % - konfidensinterval



R-kode til at lave figuren kan ses i dagens R-program.



Opsummering — til eget brug

- Hvad er fortolkningen af standard error (SE)?
- Hvilke 'ingredienser' skal bruges for at lave et konfidensinterval?
- Hvordan skal værdierne i et konfidensinterval fortolkes?
- Hvad mener vi med at R bruger en referencegruppe i ensidet ANOVA?

