



Ensidet variansanalyse (flere grupper) og lineær regression

Anders Tolver
Institut for Matematiske Fag

Statistisk Dataanalyse 1, Kursusuge 3, mandag
Dias 1/35



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 (s. 18-37 fra 11/9-2019)

Statistisk Dataanalyse 1, Kursusuge 3, mandag
Dias 2/35



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 3/35



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 4/35



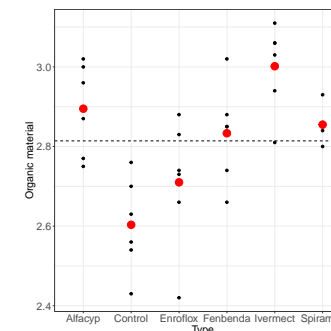
Ensidet ANOVA — flere stikprøver



antibio-datasættet

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

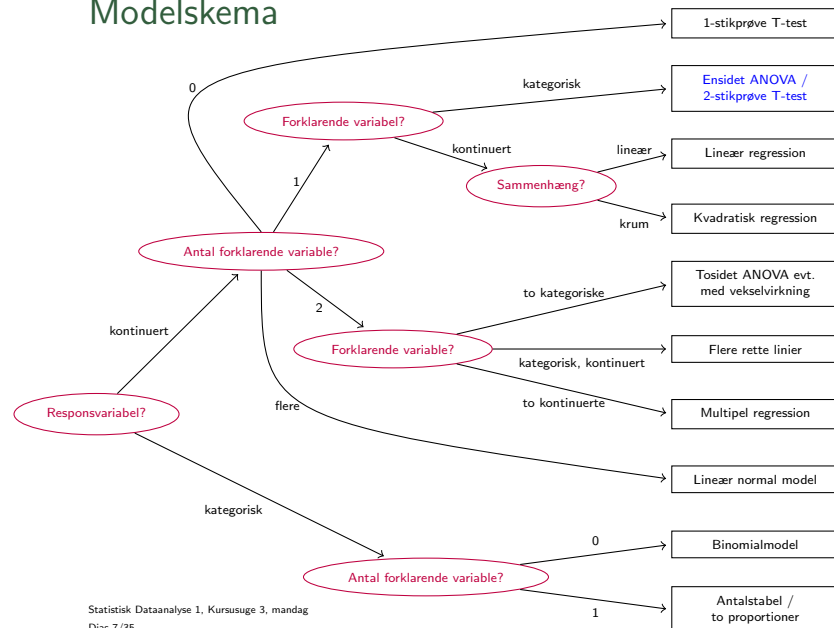
##      type org
## 1 Ivermect 3.03
## 2 Ivermect 2.81
## 3 Ivermect 3.06
## 4 Ivermect 3.11
## 5 Ivermect 2.94
## 6 Ivermect 3.06
## 7 Alfacyc 3.00
```



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



Modelskema



Statistisk model

Data: y_1, \dots, y_n fra k grupper med n_j observationer i gruppe j .

Hver gruppe antages at have sin egen middelværdi (forventede værdi): $\alpha_1, \dots, \alpha_k$.

Statistisk model: Uafhængighed + alle obs. er normalfordelte med den relevante gruppemiddelværdi og samme spredning. **Tegn!**

Formelt:

- y_1, \dots, y_n uafhængige
- y_i normalfordelte med middelværdi $\alpha_{g(i)}$ og spredning σ , hvor $g(i)$ angiver gruppen for observation i .

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



Estimation

Estimerer:

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

Ofte særligt interesseret i **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 estimererne?

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



Fælles/sammenvejet spredning

Gruppe j : n_j observationer, gruppegennemsnit \bar{y}_j , stikprøvespredning s_j .

Fælles varians og spredning:

$$\begin{aligned} s^2 &= \frac{1}{n-k} \left((n_1-1) \cdot s_1^2 + \dots + (n_k-1) \cdot s_k^2 \right) \\ &= \frac{1}{n-k} \left((y_1 - \bar{y}_{g(1)})^2 + \dots + (y_n - \bar{y}_{g(n)})^2 \right) \\ s &= \sqrt{s^2} \end{aligned}$$

Bemærk: Division med $n-k$ = antal obs. – antal grupper. Dette er antallet af **frihedsgrader** for denne model.



Fælles/sammenvejet spredning

Behandling	n_j	\bar{y}_j	s_j
Control	6	2.603	0.119
α -cyperm.	6	2.895	0.117
Enrofloxacin	6	2.710	0.162
Fenbendaz.	6	2.833	0.124
Ivermectin	6	3.002	0.109
Spiramycin	4	2.855	0.054

Fælles spredning,

$$s = \sqrt{\frac{1}{34-6} \left(5 \cdot 0.119^2 + \dots + 3 \cdot 0.054^2 \right)} = 0.1217$$



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 **estimerer**



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**?

- $df = n - k = \text{antal obs. minus antal middelværdiparametre}$
- Det samme som der stod i nævneren i beregningen af s



R

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



Version med gruppegennemsnit

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

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



Version med gruppegennemsnit

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

                2.5 %    97.5 %
typeAlfacyp  2.793191  2.996809
typeControl  2.501524  2.705142
typeEnroflox 2.608191  2.811809
typeFenbenda 2.731524  2.935142
typeIvermect 2.899858  3.103476
typeSpiramyc 2.730310  2.979690
```



Version med referencegruppe

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

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.89500	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



Version med referencegruppe

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

	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



Version med selvvalgt referencegruppe

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

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.60333	0.04970	52.379	< 2e-16 ***
myTypeAlfacyp	0.29167	0.07029	4.150	0.000281 ***
myTypeEnroflox	0.10667	0.07029	1.518	0.140338
myTypeFenbenda	0.23000	0.07029	3.272	0.002834 **
myTypeIvermect	0.39833	0.07029	5.667	4.5e-06 ***
myTypeSpiramyc	0.25167	0.07858	3.202	0.003384 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1217 on 28 degrees of freedom



Version med selvvalgt referencegruppe

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

```
> confint(model3)
```

	2.5 %	97.5 %
(Intercept)	2.50152445	2.7051422
myTypeAlfacyp	0.14768716	0.4356462
myTypeEnroflox	-0.03731284	0.2506462
myTypeFenbenda	0.08602049	0.3739795
myTypeIvermect	0.25435382	0.5423128
myTypeSpiramyc	0.09069268	0.4126407



R: Diverse

Vigtigt:

- Tre versioner af **samme model**. Forskellige parametriseringer
- Nyttige til forskellige ting, skal kunne forstå alle tre output og benytte det mest hensigtsnæssige

Til den tekniske side:

- Kategoriske variable kaldes også **faktorer**.
- En variabel kan laves til en faktor med funktionen `factor`
- Hvis data er indlæst med Excel, så skal man bruge `factor` før man kan bruge `relevel`



Spørgsmål

- Estimat for forventet værdi for Alfacyl? For Fenbenda?
- Estimat for residualspredningen σ ? Samme i alle versioner?
- Hvordan fremkommer tallet 0.04970 (SE for intercept)?
- $SE(\hat{\alpha}_j)$ større for Spiramyc end for de andre grupper. Hvorfor?
- For `model2`: Hvorfor er SE for Intercept (svarede til Alfacyl) og Control helt forskellige selvom begge har $n_j = 6$?
- For `model2`: Hvorfor er det interessant om nul ligger i KI?
- Er det også interessant for Intercept? For `model1`?
- Konklusion vedr. effekt af antibiotikum på nedbrydning?



Ensidede ANOVA: Opsummering

- Flere stikprøver
- Antagelser: Uafhængighed, normalfordeling, samme spredning, (potentielt) forskellige middelværdier
- Ofte mest interesseret i forskelle mellem middelværdier
- Estimerer, standard errors, konfidensintervaller
- `lm(y ~ gruppe)`: R vælger referencegruppe og angiver estimat for gruppen og forskelle til denne gruppe

Output fra `summary`:

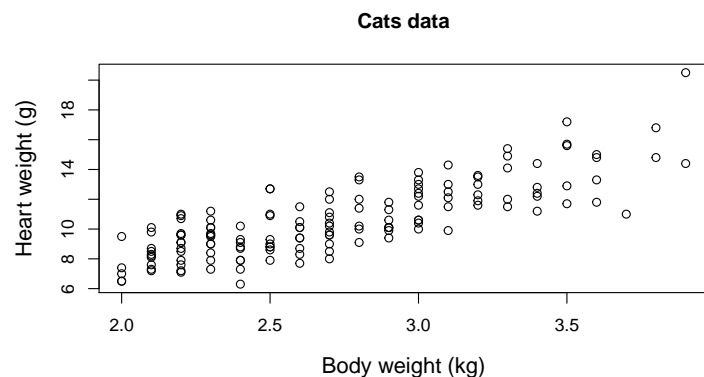
- **Hele linien hører til samme parameter**
- Fx: Hvis det er en forskel der estimeres, så hører Std. Error til denne forskel. Tilsvarende med konfidensintervaller.



Lineær regression



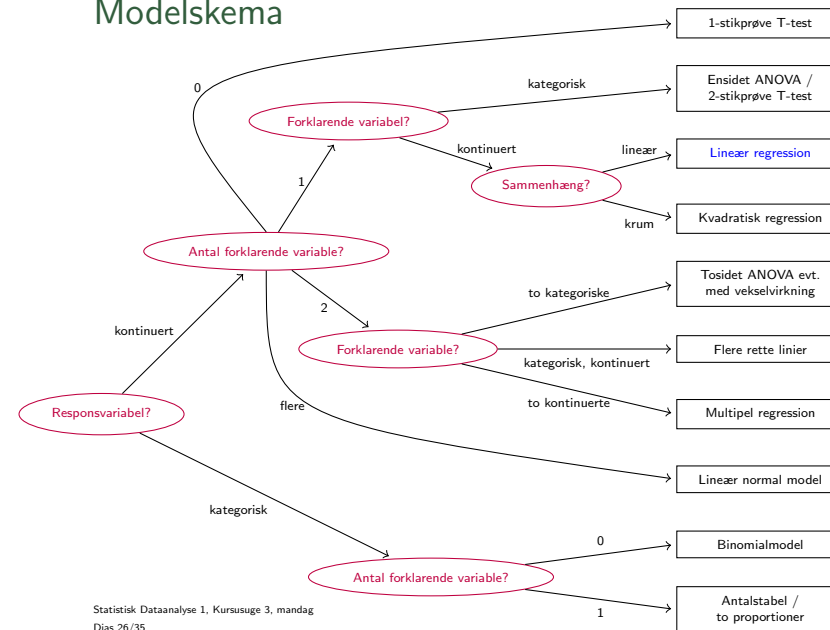
Data: Kattes hjerte- og kropsvægt



Tilnærmelsesvis lineær sammenhæng, på nær tilfældig variation.



Modelskema



Statistisk model

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

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 $\alpha + \beta x_i$ og spredning σ .

Skæring/intercept α , hældning β og spredningen σ er **parametre** i modellen, som vi vil udtale om os ud fra de givne data.



Estimerer

Estimerer 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 estimerterne?

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



Standard errors

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:

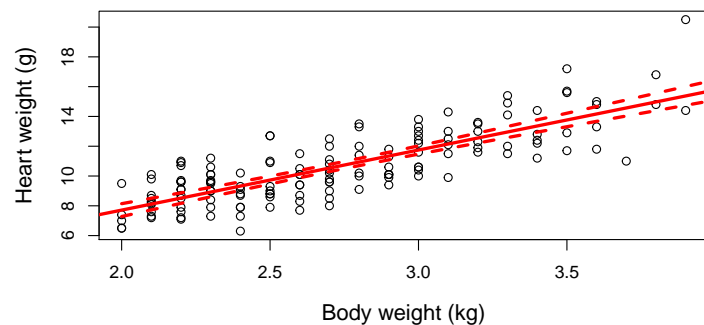
$$95\% \text{ KI : } \text{estimat} \pm t_{0.975, df} \cdot SE(\text{estimat})$$

Hvor mange frihedsgrader?

- $df = n - 2 = \text{antal obs. minus antal middelværdiparametre}$
- Det samme som der stod i nævneren i beregningen af s



Cats data



R

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

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



R

```
> confint(linreg)
              2.5 %    97.5 %
(Intercept) -1.725163 1.011838
Bwt          3.539343 4.528782

> newData <- data.frame(Bwt=2.5)
> newData
  Bwt
1 2.5
> predict(linreg, newData)
      1
9.728494
> predict(linreg, newData, interval="c")
      fit      lwr      upr
1 9.728494 9.464902 9.992087
```



Resultater og opsummering

Resultater:

- $\hat{\beta} = 4.034$ (SE 0.250) med 95% KI (3.539 , 4.529) for β
- For $x = 2.5$ er $\hat{y} = 9.73$ med 95% KI (9.46 , 9.99).
- $\hat{\sigma} = 1.45$

Fortolkning?

Opsummering:

- Antagelser: Uafhængighed, normalfordeling, lineær middelværdi, samme spredning om linien.
- Ofte mest interesset i hældning, y -værdi for givet x
- Estimer, standard errors, konfidensintervaller



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?

