# StatData1: forel221031 (første version)

Anders Tolver

## Contents



## Formålet med dette dokument

Formålet med dette dokument er at fremhæve nogle af de centrale statistiske modeller og eksempler, som har været diskuteret i forbindelse med undervisningen i Statistisk Dataanalyse 1 i 2022.

Det er vigtigt, at I opnår erfaring med at kunne afkode, hvilken statistisk metode / modelklasse, som er velegnet til at besvare opgaver / delspørgsmål som knytter sig til et konkret datasæt. Hertil kan modelvalgsdiagrammet være særlig velegnet. Start med at identificere og afgøre om *responsvariablen* er en kvantitativ / kontinuert

variabel eller om der er tale om en antalstabel. Fokuser dernæst på at finde / identificere antallet og datatypen for eventuelle forklarende variable.

Når du har identificeret den model / del af pensum, som er relevant for at besvare en konkret delopgave, så kan du med fordel finde et R-program, der kan hjælpe dig med at huske, hvordan man skriver R-koden, og hvordan man fortolker output. Oversigten nedenfor i dette dokument forsøger at hjælpe dig til at finde og træne centrale begreber og teknikker vha. opgaver og eksempler fra undervisningen i kursusuge 1-7. Vær dog opmærksom på, at jeg ikke kan afgøre, hvor I hver især har jeres største udfordringer. Det anbefales at supplere dette dokument med egne kommentarer, eller at lave en tilsvarende oversigt, som er tilpasset dine personlige udfordringer.

---

# Parrede vs. ikke-parrede stikprøver

## Hvornår

Hvis vi måler flere gange på de *samme individer* og interesserer os for forskellen.

## Hvordan

Udregn forskellene og opfat data som en enkelt stikprøve af forskelle. Brug formlen for en enkelt stikprøve til at udregne konfidensinterval.

Eksemplet om halthed af heste fra R-programmet d. 21/9-2022 viser 3 metoder, hvorpå man kan få R til at udregne estimater / konfidensintervaller.

**Data:** hentes fra `isdals` R-pakken.

```
library(isdals)
data(lameness)
lameness
```

```
##   horse    lame healthy
## 1     1  4.3541 -0.9914
## 2     2  4.7865  1.4710
## 3     3  6.1945  1.2459
## 4     4 10.7383  0.4024
## 5     5  3.3007  0.0325
## 6     6  4.8678 -0.6396
## 7     7  7.8965  0.7246
## 8     8  3.9338  0.0604
```

**Metode 0:** brug formel for en enkelt stikprøve til at beregnes konfidensinterval for middelværdien af forskellene.

```
lameness$diff <- lameness$lame - lameness$healthy
head(lameness)
```

```
##   horse    lame healthy    diff
## 1     1  4.3541 -0.9914  5.3455
## 2     2  4.7865  1.4710  3.3155
## 3     3  6.1945  1.2459  4.9486
## 4     4 10.7383  0.4024 10.3359
## 5     5  3.3007  0.0325  3.2682
## 6     6  4.8678 -0.6396  5.5074
```

```
s <- sd(lameness$diff)
KI_up <- mean(lameness$diff) + qt(0.975, 8 - 1) * s / sqrt(8)
KI_up
```

```
## [1] 7.44163
```

```
KI_low <- mean(lameness$diff) - qt(0.975, 8 - 1) * s / sqrt(8)
KI_low
```

```
## [1] 3.49997
```

**Metode 1:** Få R til at lave KI for middelværdien af forskellene ...

**Bemærk:** Når man (som nedenfor) kun skriver `~ 1` i kaldet til `lm()`, så vil R estimere en model svarende til enstikprøve-problemet. Dvs. en model, hvor alle observationerne har den samme middelværdi.

```
one_sample_mod <- lm(lame - healthy ~ 1, data = lameness)
summary(one_sample_mod)
```

```
##
## Call:
## lm(formula = lame - healthy ~ 1, data = lameness)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2026 -1.7369 -0.3238  0.4527  4.8651
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.4708     0.8335   6.564 0.000315 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.357 on 7 degrees of freedom
```

```
confint(one_sample_mod)
```

```
##               2.5 %  97.5 %
## (Intercept) 3.49997 7.44163
```

**Metode 2/3:** Brug `t.test()`

```
t.test(lameness$diff)
```

```
##
##  One Sample t-test
##
## data:  lameness$diff
## t = 6.5639, df = 7, p-value = 0.0003147
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  3.49997 7.44163
## sample estimates:
## mean of x
##    5.4708
```

```
t.test(lameness$lame, lameness$healthy, paired=TRUE)
```

```
##
```

```
##  Paired t-test
##
## data:  lameness$lame and lameness$healthy
## t = 6.5639, df = 7, p-value = 0.0003147
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.49997 7.44163
## sample estimates:
## mean of the differences
##                  5.4708
```

### Eksempler på opgaver

Udvalgte delspørgsmål fra

- Januar 2018, opgave 3
- November 2019, opgave 2.
- Februar 2021, opgave 1.1.

---

## Ensidet variansanalyse

### Hvornår

Kvantitativ / kontinuert responsvariabel og en kategorisk forklarende variabel.

### Hvordan

R-programmet fra 19/9-2022 kommer godt rundt i emnet. Du bør fx. arbejde med at lære/forstå følgende:

- kunne aflæse output med og uden intercept
- kunne teste hypotesen om, at middelværdien er ens i alle grupper (ved et F-test)
- i nogle eksempler: teste en ensidet ANOVA mod en lineær regressionsmodel, hvis den forklarende variable kan opfattes både som et kategorisk og en numerisk variabel (se R-programmet fra 3/10-2022 / Opgave HS 22)
- kunne fortolke output, hvis analysen foretages med en log-transformeret variabel som respons

**Aflæsning af output:** Samme model forskellig output/parametrisering (eksempel med gødning og antibiotika)

```
data(antibio)
### Model, parametrisering med de seks gennemsnit
model1 <- lm(org ~ type - 1, data=antibio)
summary(model1)
```

```
##
## 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
```

```
### Model med Alfacyp som referencegruppe
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)   2.89500    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
## Multiple R-squared:  0.5874, Adjusted R-squared:  0.5137
## F-statistic: 7.973 on 5 and 28 DF,  p-value: 8.953e-05
```

```
### Model med Control som referencegruppe
antibio$type <- relevel(antibio$type, ref = "Control")
model3 <- lm(org ~ type, data=antibio)
summary(model3)
```

```
##
## 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)   2.60333    0.04970  52.379  < 2e-16 ***
```

```
## typeAlfacyp   0.29167    0.07029   4.150 0.000281 ***
## typeEnroflox  0.10667    0.07029   1.518 0.140338
## typeFenbenda  0.23000    0.07029   3.272 0.002834 **
## typeIvermect  0.39833    0.07029   5.667  4.5e-06 ***
## typeSpiramyc  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
## Multiple R-squared:  0.5874, Adjusted R-squared:  0.5137
## F-statistic: 7.973 on 5 and 28 DF,  p-value: 8.953e-05
```

**F-test:** Her vises en metode til at få R til at lave F-testet for, om middelværdien kan antages at være ens i alle grupper (her givet ved variablen `type`). Helt ok, hvis du har andre metoder til at producere F-teststørrelsen på.

```
### Test for ens middelværdi i alle grupper kan laves som følger ...
nulmodel <- lm(org ~ 1, data=antibio)
anova(nulmodel, model1)
```

```
## Analysis of Variance Table
##
## Model 1: org ~ 1
## Model 2: org ~ type - 1
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     33 1.0058
## 2     28 0.4150  5   0.59082 7.9726 8.953e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Test af lineær regression mod ensidet ANOVA:** For eksemplet med elektriske ål fra R-programmet til 3/10-2022 er der flere målinger af frekvensen (=respons) for hver vandtemperatur. Derfor kan man ved et F-test undersøge om middelværdien af responsen afhænger lineær af vandtempaturen ved at sammenligne en lineær regressionsmodel mod modellen, hvor vandtemperaturen inddrages som en faktor i modellen.

```
### Eksempel med elektriske ål
library(isdals)
data(eels)
head(eels)
```

```
##   temp freq
## 1   20  230
## 2   20  239
## 3   20  251
## 4   22  256
## 5   22  259
## 6   22  265
```

```
eels <- transform(eels, tempFac = factor(temp))
ensidet <- lm(freq ~ tempFac, data=eels)
linreg <- lm(freq ~ temp, data=eels)
anova(linreg, ensidet)
```

```
## 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
```

**Hvad er det egentlig vi konkluderer på baggrund af testet?** Den lineære regressionsmodel skal opfattes som en mere restriktiv model end den ensidede ANOVA. Dette skyldes at en lineær regressionsmodel her tvinger middelværdierne hørende til de 7 forskellige vandtemperaturer, som er afprøvet i forsøget, til at være givet som en lineær funktion af vandtemperaturen. Vores nulhypotese er den mest restriktive model (her: den lineære regressionsmodel), så når vi får en stor p-værdi på 0.6292 kan vi ikke forkaste nulhypotesen. Der er mao. *ikke* på baggrund af data grund til at afvise, at middelværdien af frekvensen af de udsendte signaler for ålene kunne være givet ved en lineær funktion af vandtemperaturen.

**Log-transformeret respons:** Regn opgave 1 fra eksamen i November 2019. Det vigtigste budskab er, at fortolkningen af estimaterne udtaler sig om medianer, når de regnes tilbage til oprindelig skala vha. eksponentialfunktionen.

**Tænk eventuelt over følgende spørgsmål:**

Med udgangspunkt i eksemplet med antibiotika og gødning ovenfor:

- hvordan kan man teste om middelværdien af responsen (dvs. indholdet af organisk stof) er ens for `Control`-gruppen og for `Enroflox`-gruppen?

**Svar:** Hvis man kigger på output under `summary(model3)` ovenfor, hvor gruppen `Control` er valgt som reference, så kan svaret umiddelbart aflæses. Forskellen melllem middelværdien i `Enroflox`- og `Control`-gruppen estimeres til 0.10667, og i samme linje aflæses t-teststørrelsen 1.518 og den tilhørende p-værdi på 0.140338. Benyttes et 5 % - signifikans niveau, så kan vi altså *ikke* afvise, at forskellen mellem middelværdien i `Enroflox` og `Control`-grupperne kunne være lig med 0.

- hvordan kan man teste om middelværdien af responsen (dvs. indholdet af organisk stof) er ens for alle antibiotikagrupper (dvs bortset fra `Control`-gruppen)?

**Svar:** Dette test kan udføres som et F-test. Alle detaljerne findes i R-programmet til forelæsningen d. 26/9-2022.

- hvad er antallet af frihedsgrader, og hvordan er dette antal fremkommet?

**Svar:** For en ensidet ANOVA er antallet af frihedsgrader (df = degrees of freedom) givet som antallet af observationer (her: n = 34) minus antallet af grupper (her: k = 6). Formlen er: df = n - k = 34 - 6 = 28. For alle andre modeller fittet med `lm()`-funktionen i R, der gælder samme formel df = n - k, hvor man kan finde k ved at optælle antallet af estimater der optræder under `coefficients`, når man ser på et `summary()` af modellen.

### Eksempler på opgaver

Der er rigtig mange tidligere eksamensopgaver om ensidet ANOVA fx.

- November 2019, opgave 1
- Januar 2020, opgave 1
- November 2020, opgave 1
- November 2020, opgave 3.5 (quiz-opgave)

---

## Lineær regression

### Hvornår

Kvantitativ / kontinuert responsvariabel og en kvantitativ / kontinuert forklarende variabel.

## Hvordan

R-programmet fra 19/9-2022 viser, hvordan man fitter og aflæser estimater fra modellen, samt hvordan man konstruerer et konfidensinterval for et punkt på regressionslinjen.

**Sammenhæng mellem hjertevægt og kropsvægt for katte:** Forstår du både R-kode og R-output?

```
library(MASS)
data(cats)
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
```

Et 95 % - konfidensinterval hørende til en kropsvægt på `Bwt = 2.5` kg vil med 95 % sandsynlighed indeholde *gennemsnittet af (uendelig mange) nye observationer* af hjertevægten, for katte med en kropsvægt på 2.5 kg.

```
newData = data.frame(Bwt = 2.5)
predict(linreg, newData, interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 9.728494 9.464902 9.992087
```

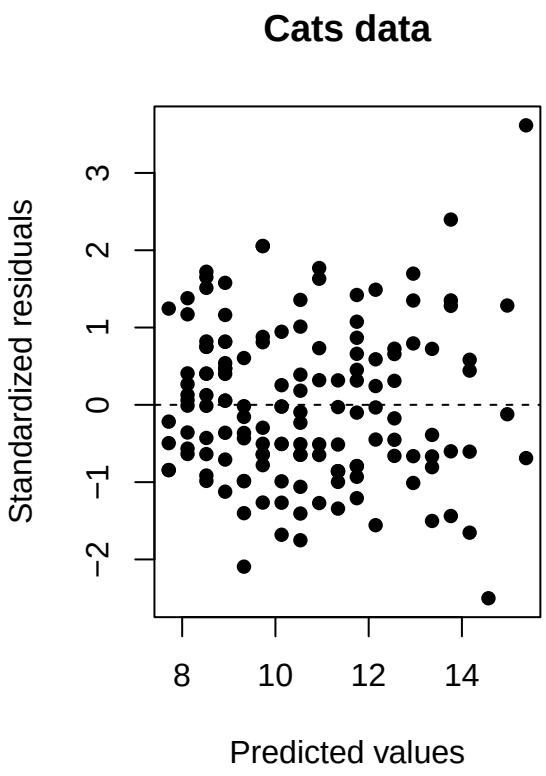
**Spørgsmål til R-kode ovenfor:** Er der en grund til, at du skriver `newData = data.frame(Bwt = 2.5)` i stedet for `newData <- data.frame(Bwt = 2.5)` i R-koden til beregning af konfidensintervallet ovenfor? ATs svar: Nej, jeg anbefaler at man bruger `<-` i stedet for `=`. Dette er også gjort ved beregning af prædiktionsintervallet nedenfor.

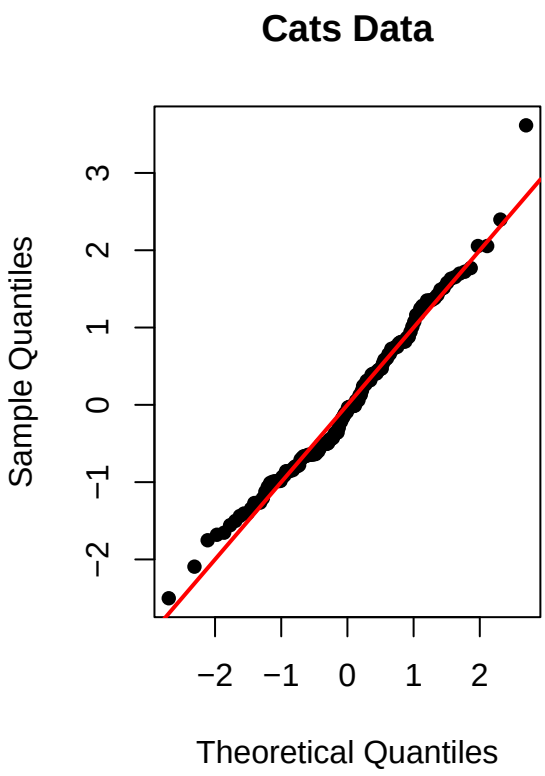
**Modelkontrol og prædiktionsintervaller:**

R-programmet fra 28/9-2022 gennemgår modelkontrol og prædiktionsintervaller for en lineær regressionsmodel.

Hvad bør man se efter og kommentere på baggrund af et residualplot og et QQ-plot?

```
### Residualplot med vandret linie i 0
par(mfrow = c(1, 2)) ### arrangerer de to figurer ved siden af hinanden
plot(fitted(linreg), rstandard(linreg), pch = 16, xlab = "Predicted values", ylab = "Standardized resid
     , main = "Cats data")
abline(h = 0, lty = 2)
### QQ-plot 0/1-linien
qqnorm(rstandard(linreg), pch = 16, main = "Cats Data")
abline(0, 1, lwd = 2, col = "red")
```




Et 95 % - prædiktionsinterval hørende til en kropsvægt på `Bwt = 2.5` kg vil med 95 % sandsynlighed indeholde *en enkelt ny observation* af hjertevægten, for en kat med en kropsvægt på 2.5 kg.

```
newData <- data.frame(Bwt = 2.5)
### praediktion og praediktionsinterval
predict(linreg, newData, interval = "p")
```

```
##        fit      lwr      upr
## 1 9.728494 6.845352 12.61164
```

**Tænk gerne over følgende:**

- hvordan man tester om der er sammenhæng mellem respons og forklarende variabel i lineær regression (= om hældningen er nul!)

**Svar:** Når man kigger på outputtet fra `summary(linreg)` ovenfor, så kan man direkte finde et t-test for hypotesen om, at hældningen er nul (i datalinjen hørende til `Bwt`).

- hvordan man tester om hældningen har en ganske bestemt værdi (R-program fra 21/9-2022)

**Svar:** Her kan du bruge den generelle formel for, hvordan man kan udføre et t-test, når man kender både standard error (SE) på estimatet og antallet af frihedsgrader. Se detaljerne i R-programmet til 22/9-2022.

- hvordan man tester en lineær regressionsmodel mod fx. en kvadratisk regressionsmodel (R-program fra 3/10-2022)

**Svar:** Denne teknik er vigtig at kende til. Man kan altid teste en lineær regressionsmodel mod en mere fleksibel model (fx. en kvadratisk model). Det er *kun* hvis man har flere målinger for hver værdi af den forklarende variable, at man kan teste en lineær regressionsmodel mod en ensidet ANOVA (som vi gjorde ovenfor i eksemplet med de elektriske ål).

## Eksempler på opgaver

- November 2018, Opgave 2.1

- Januar 2019, Opgave 2.1-2.4
- November 2020, Opgave 2
- Opgaverne fra januar 2019 og November 2020 er ekstra udfordrende, fordi både respons og forklarende variabel indgår log-transformerede i den lineære regressionsmodel.

---

# Multiple lineær regression

## Hvornår

Kvantitativ / kontinuert responsvariabel og (mindst) to kvantitative / kontinuerte forklarende variable.

## Hvordan

R-programmet fra 5/10-2022 om sammenhæng mellem volumen og højde+diameter af kirsebærtræer indeholder flere eksempler på multiple lineære regressionsmodeller.

**Kirsebærtræer:** Kan du opskrive den tilhørende statistiske model og fortolke output nedenfor?

```
library(isdals)
data(trees)
multipel1 <- lm(Volume ~ Height + Girth, data = trees)
summary(multipel1)
```

```
##
## Call:
## lm(formula = Volume ~ Height + Girth, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4065 -2.6493 -0.2876  2.2003  8.4847
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
## Height        0.3393     0.1302   2.607   0.0145 *
## Girth         4.7082     0.2643  17.816  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

Den statistiske model udtrykkes som

$$V_i = \alpha + \beta \cdot H_i + \gamma \cdot G_i + e_i,$$

hvor restleddene $e_i$ er uafhængige og normalfordelte $\sim N(0, \sigma^2)$. Estimatet for parameteren hørende til variablen `Height` (H) er 0.3393. Fortolkningen er at det forventede volumen (V) øges med 0.3393 enheder, når højden (H) øget med 1 enhed.

**Kirsebærtræer (log-transformation):** Kan du opskrive den tilhørende statistiske model og fortolke output nedenfor?

```r
multipel2 <- lm(log(Volume) ~ log(Height) + log(Girth)
, data = trees)
summary(multipel2)
```

```
##
## Call:
## lm(formula = log(Volume) ~ log(Height) + log(Girth), data = trees)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.168561 -0.048488  0.002431  0.063637  0.129223
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.63162    0.79979  -8.292 5.06e-09 ***
## log(Height)  1.11712    0.20444   5.464 7.81e-06 ***
## log(Girth)   1.98265    0.07501  26.432  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08139 on 28 degrees of freedom
## Multiple R-squared:  0.9777, Adjusted R-squared:  0.9761
## F-statistic: 613.2 on 2 and 28 DF,  p-value: < 2.2e-16
```

Der findes en detaljeret analyse af dette eksempel i R-programmet fra 5/10-2022. Selvom den matematiske fortolkning af modellen her (hvor både respons og forklarene variable log-transformeres) er lidt mere kompleks, så giver eksemplet et godt indblik i fleksibiliteten af klassen af multiple lineære regressionsmodeller.

### Eksempler på opgaver

- November 2018, opgave 2
- januar 2019, opgave 2.5 (afleveringsopgave 2)

---

## Tosidet variansanalyse

### Hvornår

Kvantitativ / kontinuert responsvariabel og to kategoriske forklarende variable.

### Hvordan

Vores hovedeksempel vedr. højder for studerende på StatData1 i 2017 søgt forklaret ved køn og studieretning (-se R-programmet fra 10/10-2022).

Du bør arbejde på at forstå følgende:

- forskellen på den additive model og modellen med vekselvirkning
- teste den additive model med modellem med vekselvirkning
- kunne finde estimatet for middelværiden hørende til en gruppe givet som kombination af de to forklarende variable (fx. mandlige jordbrugsøkonomi-studerende) ... både for den additive model og for vekselvirkningsmodellen
- teste om modellen kan reduceres ved at fjerne en af de to forklarende variable fra modellen (ved et F-test)

**Additive model / model med vekselvirkning:** Indlæser først data fra eksamen januar 2020 opgave 2.

```
pestgolf <- read.table(file = "../data/pestgolf.txt", header = T)
head(pestgolf)
```

```
##   Treat Lokation    Kd
## 1   T04      KNY 0.347
## 2   T04      KNY 0.689
## 3   T04      KNY 0.652
## 4   T05      KNY 0.638
## 5   T05      KNY 0.542
## 6   T05      KNY 0.660
```

```
modelVeksel <- lm(Kd ~ Treat * Lokation, data = pestgolf)
modelAdd <- lm(Kd ~ Treat + Lokation, data = pestgolf)
```

*Model med vekselvirkning:* de 6 middelværdier hørende til kombinationer af `Treat` og `Lokation` estimeres helt frit.

*Additive model:* middelværdier hørende til kombinationer af `Treat` og `Lokation` tvinges til at have en *additiv struktur*. Forskellen på estimaterne hørende til `Treat = T04` og `Treat = T05` er den samme, uanset hvilken `Lokation` vi ser på. (Tænk eventuelt på eksempel med isbutik: prisen på guf er den samme uanset antallet af kugler).

**Test af additiv model mod model med vekselvirkning:** Kan udføres som et F-test af additive model (=nulmodel) mod vekselvirkningsmodel (=fuld model).

```
anova(modelAdd, modelVeksel)
```

```
## Analysis of Variance Table
##
## Model 1: Kd ~ Treat + Lokation
## Model 2: Kd ~ Treat * Lokation
##   Res.Df     RSS Df Sum of Sq     F Pr(>F)
## 1     14 0.18774
## 2     12 0.12000  2  0.067739 3.387 0.0682 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypotesen (dvs. den additive model) kan ikke forkaste på et 5 % - niveau, da p-værdien er 0.0682.

**Aflæse / fortolke estimaterne:** Lad os aflæse estimaterne for kombinationen `Treat = T05, Lokation = KNY` både for den additive model og for modellen med vekselvirkning.

*Model med vekselvirkning:*

```
summary(modelVeksel)
```

```
##
## Call:
## lm(formula = Kd ~ Treat * Lokation, data = pestgolf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21567 -0.03192  0.02250  0.04475  0.12633
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.848667   0.057735  14.699  4.9e-09 ***
```

```
## TreatT05               0.315000   0.081649   3.858 0.002277 **
## LokationHONE          -0.459333   0.081649  -5.626 0.000111 ***
## LokationKNY           -0.286000   0.081649  -3.503 0.004359 **
## TreatT05:LokationHONE -0.008333   0.115470  -0.072 0.943656
## TreatT05:LokationKNY  -0.264333   0.115470  -2.289 0.040991 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1 on 12 degrees of freedom
## Multiple R-squared:  0.8998, Adjusted R-squared:  0.858
## F-statistic: 21.55 on 5 and 12 DF,  p-value: 1.299e-05
```

Estimatet bliver: 0.848667 + 0.315000 - 0.286000 - 0.264333.

*Additive model:*

```
summary(modelAdd)
```

```
##
## Call:
## lm(formula = Kd ~ Treat + Lokation, data = pestgolf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.15806 -0.06435 -0.02086  0.06978  0.21306
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.89411    0.05459  16.379 1.58e-10 ***
## TreatT05      0.22411    0.05459   4.105  0.00107 **
## LokationHONE -0.46350    0.06686  -6.933 6.95e-06 ***
## LokationKNY  -0.41817    0.06686  -6.255 2.11e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1158 on 14 degrees of freedom
## Multiple R-squared:  0.8432, Adjusted R-squared:  0.8096
## F-statistic: 25.09 on 3 and 14 DF,  p-value: 6.803e-06
```

Estimatet bliver: 0.89411 + 0.22411 -0.41817.

**Yderligere ting der kan aflæses af outputtet fra den additive model:** Vi aflæser desuden at

- den estimerede forskel på sorptionen for `Treat = T05` og `Treat = T04` (indeholdt i reference/Intercept) er 0.22411
- den estimerede forskel på sorptionen for `Lokation = HONE` og `Lokaltion = DYR` (indeholdt i reference/Intercept) er -0.46350
- den estimerede forskel på sorptionen for `Lokation = KNY` og `Lokaltion = DYR` (indeholdt i reference/Intercept) er -0.41817

**Test for effekt af `Treat`/ `Lokation`:** Test for hver af de to indgående forklarende variable bør kun foretages, hvis der ikke er vekselvirking (dvs. hvis vi ikke kan forkaste den additive model). Den additive model benyttes som *fuld model* og modellen, hvor en af de to forklarende variable fjernes benyttes som *nulmodel* i et F-test med `anova()`.

*Test for effekt af `Treat`:*

```r
modelLokation <- lm(Kd ~ Lokation, data = pestgolf)
anova(modelLokation, modelAdd)
```

```
## Analysis of Variance Table
##
## Model 1: Kd ~ Lokation
## Model 2: Kd ~ Treat + Lokation
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1     15 0.41375
## 2     14 0.18774  1   0.22602 16.855 0.001071 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vi forkaster hypotesen om, at vi kan se bort fra `Treat`: observeret F-teststørrelse = 16.855, P-værdi = 0.001071.

*Test for effekt af `Lokation`:*

```r
modelTreat <- lm(Kd ~ Treat, data = pestgolf)
anova(modelTreat, modelAdd)
```

```
## Analysis of Variance Table
##
## Model 1: Kd ~ Treat
## Model 2: Kd ~ Treat + Lokation
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     16 0.97124
## 2     14 0.18774  2    0.7835 29.214 1.008e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vi forkaster hypotesen om, at vi kan se bort fra `Lokation`: observeret F-teststørrelse = 29.214, P-værdi = 0.000010.

### Eksempler på opgaver

- Januar 2019, opgave 1 kommer godt rundt i pensum her (afleveringsopgave 3)
- Januar 2020, opgave 2 (fra kursusuge 6)
- November 2020, opgave 3.7 (quiz-spørgsmål)
- Februar 2021, opgave 2.4-2.5

---

## Blandede modeller

### Hvornår

Kvantitativ / kontinuert responsvariabel og både en kategoriske og en kvantitativ / kontinuert forklarende variable.

### Hvordan

R-programmet fra 12/10-2022 viser, hvordan man fitter og aflæser estimaterne fra en blandet model (eksempel med løbetider på DHL-stafetten).

Vær særligt opmærksom på

- hvordan man opskriver modellen korrekt

- hvordan man fortolker estimaterne fra output (igen kan man fitte modellen med og uden intercept!)

**Løbetider på DHL-stafetten:** Brug tid på at lære at opskrive modellen og på at kunne fortolke estimaterne. En særlig egenskab med den *blandede model* her er, at forskellene i løbetiderne mellem de forskellige dage er ens, uanset holdsammensætningen (dvs. antal kvinder på holdet).

```
data(dhl)
dhl <- transform(dhl, time = 60*60*hours + 60*minutes + seconds)
dhl <- transform(dhl, group = factor(women))
head(dhl)
```

```
##      day men women hours minutes seconds time group
## 1 Monday   5     0     2      12      10 7930     0
## 2 Monday   4     1     2      13      39 8019     1
## 3 Monday   3     2     2      17      33 8253     2
## 4 Monday   2     3     2      21      57 8517     3
## 5 Monday   1     4     2      26      33 8793     4
## 6 Monday   0     5     2      30      35 9035     5
```

```
model1 <- lm(time ~ day + women, data=dhl)
summary(model1)
```

```
##
## Call:
## lm(formula = time ~ day + women, data = dhl)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -180.451  -40.479    2.638   15.372  207.313
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7818.911     41.030 190.564  < 2e-16 ***
## dayThursday  -248.500     46.824  -5.307 4.02e-05 ***
## dayTuesday     15.833     46.824   0.338 0.738960
## dayWednesday -205.167     46.824  -4.382 0.000321 ***
## women         242.236      9.693  24.990 5.37e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 81.1 on 19 degrees of freedom
## Multiple R-squared:  0.9727, Adjusted R-squared:  0.9669
## F-statistic: 168.9 on 4 and 19 DF,  p-value: 1.45e-14
```

### Eksempler på opgaver

- November 2019, opgave 2.3-2.5.
- Februar 2021, opgave 1.5

---

## Kursusuge 7 / metoder til analyse af antalstabeller

Det er essentielt at forstå, hvordan man udregner sandsynligheder i en binomialfordeling (klassisk quiz-spørgmål i multiple choice opgaver).

Desuden snakkede vi om, hvordan man

- i) estimerer og beregner konfidensintervaller for en andel (dvs. antal / total)
- ii) tester hypoteser om at andelen har en bestemt værdi (fx. 0.5)
- iii) sammenligner to binomialfordelinger
- iv) laver test for homogenitet og test for uafhængighed baseret på antalstabeller.

Der er mange emner og næppe nogen vej uden om at nærstudere slides og R-programmer fra denne uge, hvis du gerne vil være ordentlig forberedt. Husk især at øve på quiz-spørgsmål fra eksamen og quizzer i Absalon, som vedrører denne del af pensum.