Modelkontrol og prædiktion

Claus Ekstrøm
E-mail: ekstrom@life.ku.dk

Program

- Test af hypotese i ensidet variansanalyse
  - F-tests og F-fordelingen.
- Multiple sammenligninger. Bonferroni-korrektion
- Opsummering af statistiske modeller/eksempler
- Modelkontrol
- Prædiktion

Multiple sammenligninger

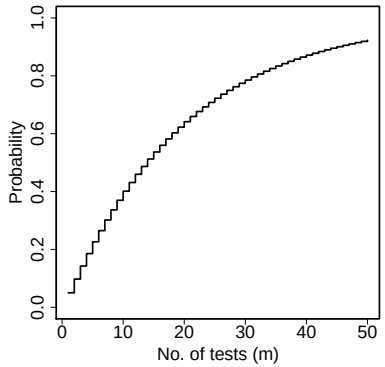
Hver gang vil laver et test er der risiko for at lave en fejl af type I.

Risikoen afhænger af signifikansniveauet — ofte 5%.

Ved et test: risiko for fejl: 5%

Ved m tests:

$$1 - 0.95^m$$



Opgave 5.2: fosforkoncentration

Lineær regression

Uge	1	2	3	4	5	6	7	8	9
Forfor	0.51	0.48	0.44	0.44	0.39	0.35	0.28	0.24	0.19

Statistisk model:

$$\text{fosfor}_i = \alpha + \beta \cdot \text{uge}_i + e_i, \quad e_1, \ldots, e_n \sim N(0, \sigma^2) \text{ uafhængige}$$

R: `model1 <- lm(fosfor ~uge)`

Opgave 6.7: vægttilvækst hos kyllinger

Ensidet ANOVA

Feed type	Weight gain				
1	55	49	42	21	52
2	61	112	30	89	63
3	42	97	81	95	92
4	169	137	169	85	154

Statistisk model:

$$\text{gain}_i = \alpha_{\text{feed}(i)} + e_i, \quad e_1, \ldots, e_n \sim N(0, \sigma^2) \text{ uafhængige}$$

R: `model2 <- lm(gain~factor(feed))`
Bemærk `factor(feed)`!

Opgave 6.1: drægtighed for heste

En enkelt stikprøve

Drægtighedstider for 13 heste:

339 339 339 340 341 340 343 348 341 346 342 339 337

Statistisk model:

$$\text{gest}_i \sim N(\mu, \sigma^2) \text{ uafhængige}$$

Modellen kan også skrives:

$$\text{gest}_i = \mu + e_i, \quad e_1, \ldots, e_n \sim N(0, \sigma^2) \text{ uafhængige}$$

R: `model3 <- lm(gest~1)`

Alle modeller

$$\text{fosfor}_i = \alpha + \beta \cdot \text{uge}_i + e_i, \quad e_1, \ldots, e_n \sim N(0, \sigma^2) \text{ uafhængige}$$

$$\text{gain}_i = \alpha_{\text{feed}(i)} + e_i, \quad e_1, \ldots, e_n \sim N(0, \sigma^2) \text{ uafhængige}$$

$$\text{gest}_i = \mu + e_i, \quad e_1, \ldots, e_n \sim N(0, \sigma^2) \text{ uafhængige}$$

Variabeltyper:

- Responsvariabel, y: fosfor, gain, gest
- Forklarende variabel: age (kvantitativ), feed (faktor/kategorisk)

Antagelser:

- Alle e_i (eller y_i) er normalfordelte
- Middelværdien af y_i afhænger evt. af en forklarende variabel
- Alle e_i (eller y_i) har samme spredning
- $e_1, \ldots, e_n$ (eller $y_1, \ldots, y_n$) uafhængige

Resumé 1: statistiske modeller og inferens

Modellerne for lineær regression, ensidet variansanalyse og en enkelt stikprøve er i virkeligheden meget ens!

Det er derfor den statistiske inferens også er den samme i de tre slags modeller (p er antallet af middelværdiparametre):

- middelværdiparametre estimeres med LS
- spredningen σ estimeres på "samme måde"
- Konfidensintervaller: $\text{estimat} \pm t_{0.975,n-p} \cdot \text{SE}(\text{estimat})$
- Hypotesetest udføres som t-test eller F-test

Modellerne kan udvides til at omfatte flere forklarende variable — kvantitative variable og/eller faktorer. Lineære normale modeller:

$$y_i = \text{middelværdi}_i + e_i \quad e_1, \ldots, e_n \sim N(0, \sigma^2) \text{ uafhængige}$$

Residualer

Forventet værdi eller fittet værdi eller prædikteret værdi, $\hat{y}_i$:

- $\hat{y}_i = \widehat{\text{fosfor}}_i = \hat{\alpha} + \hat{\beta} \cdot x_i$
- $\hat{y}_i = \widehat{\text{gain}}_i = \hat{\alpha}_{g(i)}$
- $\hat{y}_i = \widehat{\text{gest}}_i = \hat{\mu}$

Residualer:

$$r_i = y_i - \hat{y}_i = \text{observeret} - \text{fittet}$$

Residualerne er vores bedste gæt på e'erne! Så

- $\hat{\sigma} = s = \sqrt{\frac{1}{n-p}\sum_{i=1}^{n} r_i^2}$ hvor p er antal middelværdiparametre (2, k, 1)
- residualerne kan bruges til modelkontrol!

Residualerne kan standardiseres så de har spredning 1:
$\tilde{r}_i = r_i / \text{SE}(r_i)$.

Residualer i R

```
> model1 <- lm(fosfor~uge)      ## Lineær regression
> fit1 <- fitted(model1)        ## Fittede værdier
> res1 <- residuals(model1)     ## Rå residualer
> stdres1 <- rstandard(model1)  ## Standard. residualer

> model2 <- lm(gain~factor(feed))
> fit2 <- fitted(model2)
> res2 <- residuals(model2)
> stdres2 <- rstandard(model2)
```

Og hvis du har `isdals` installeret

```
> residualplot(model1)
```

Modelkontrol: hvorfor?

Modelkontrol består i at kontrollere om modelantagelserne er rimelige for vores data.

Hvorfor skal vi lave modelkontrol?

- Hvis antagelserne er ok, så indeholder 95%-CI populationsværdien med 95% sandsynlighed, og p-værdierne er korrekte.

  Vi kan stole på vores resultater!
- Hvis antagelserne ikke er ok, så ved vi ikke om vi kan stole på vores resultater!

Antagelserne om $e_1, \ldots, e_n$ kontrolleres vha. de standardiserede residualer $\tilde{r}_1, \ldots, \tilde{r}_n$.

Modelkontrol: hvordan?

Antagelser:

1. e_i er normalfordelt
2. e_i har middelværdi 0 — uanset de forklarende variable
3. e_i har samme spredning — uanset de forklarende variable
4. $e_1, \ldots, e_n$ uafhængige

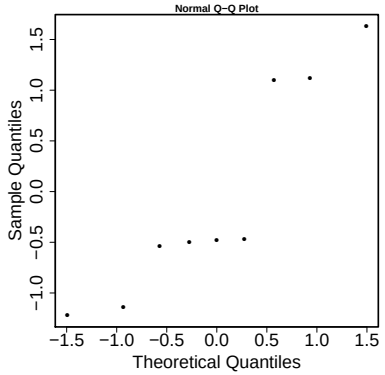
Hvordan?

- Uafhængighed er snarere et spørgsmål om eksperimentielt design
- Kontrollerer de tre første antagelser om $e_1, \ldots, e_n$ vha. de standardiserede residualer $\tilde{r}_1, \ldots, \tilde{r}_n$

Thorvald Nicolai Thiele, 1838–1910

Man skal tegne før man kan regne

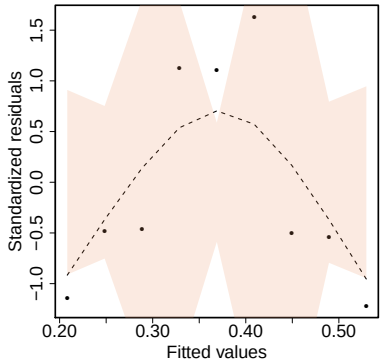
Plantevækst og fosfor: antagelse 1



Antagelse 1. e_i er normalfordelt:

- QQ-plot over $\tilde{r}_1, \ldots, \tilde{r}_n$
- Sammenlign med ret linie, med skæring 0 og hældning 1

Plantevækst og fosfor: antagelse 2 og 3



Antagelse 2. og 3. e_i har middelværdi 0 og samme spredning

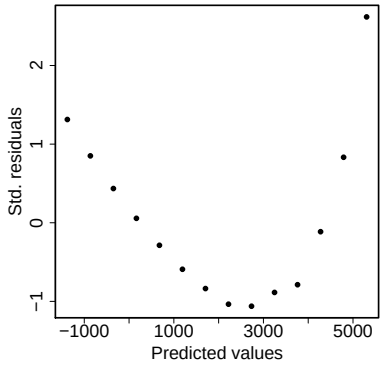
- Residualplot, $\tilde{r}_i$ mod $\hat{y}_i$
- Ingen systematik i den lodrette variation
- Er det outliers, dvs. ekstreme observationer?

($\tilde{r}_1, \ldots, \tilde{r}_n$ har spredning 1)

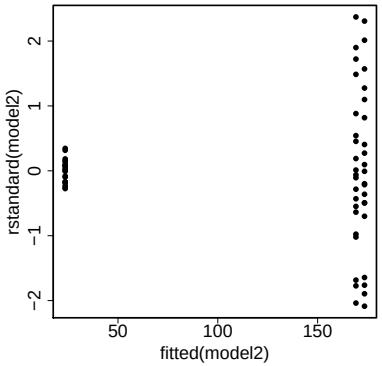
Hvis man har installeret `isdals` kan man bruge `residualplot(model)`

Residualplot for to andre datasæt

Andemad (eks. 2.4 og 6.2)

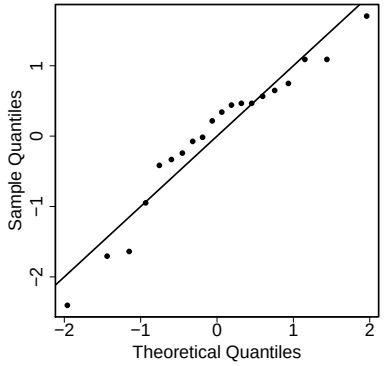


Pillbugs/bænkebidere (case 2)

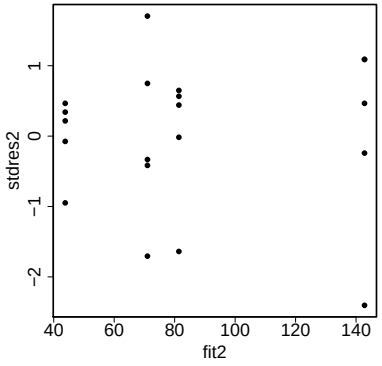


Residualanalyse for kyllingedata

QQ-plot



Residualplot



Antagelse 4: uafhængighed

Kan som regel ikke testes vha. data (residualer), men er snarere et spørgsmål om eksperimentets design.

Observationerne må ikke "dele information".

Hvis en observation ligger højere end forventet, ændrer det så vores viden om hvorvidt de nogle af de øvrige observationer ligger højere/lavere end forventet?

Eksempler på afhængige data:

- Data fra samme marker, samme personer, samme planter, etc.
- Data fra søskende, kuld, ...

Sommetider vil vi gerne have afhængighed — men så skal der tages højde for det i modellen.

Resumé 2: modelkontrol

Det er ekstremt vigtigt at at lave modelkontrol, for ellers ved vi ikke om vi kan stole på konfidensintervaller, p-værdier osv.

Modelkontrol udføres først og fremmest grafisk, vha. residualplot og QQ-plot for standardiserede residualer. Især residualplottet er vigtigt!

- I residualplottet skal den lodrette variation være tilfældig. Må ikke være systematisk forskellige "fra venstre til højre".
- Meget store standardiserede residualer svarer til ekstreme observationer eller outliers. Bør undersøges nærmere.
- I QQ-plottet skal punkterne som sædvanlig være spredt tilfældigt om en ret linie, her linien med skæring 0 og hældning 1.
- Er det rimeligt at antage uafhængighed?

Plantevækst og fosfor: prædiktion

Plante bliver fulgt i 7 uger. Forventet forforkoncentration er $\alpha + \beta \cdot 7$ der estimeres til

$$\hat{y} = \hat{\alpha} + \hat{\beta} \cdot 7 = 0.56972 - 0.04017 \cdot 7 = 0.28853$$

med estimeret spredning (side 110)

$$\mathrm{SE}(\hat{y}_0) = s\sqrt{\frac{1}{n} + \frac{(7-\bar{x})^2}{\mathrm{SS}_x}} = 0.02031 \cdot \sqrt{0.4216} = 0.00856$$

95%-konfidensinterval:

$$0.28853 \pm 2.306 \cdot 0.00856 = (0.2688; 0.3083)$$

En plante på 7 uger får målt forfoskoncentrationen til 0.25. Hvorfor kan vi ikke bruge konfidensintervallet til at afgøre om det er usædvanligt?

Plantevækst og fosfor: prædiktion

Konfidensintervallet udtaler sig om den forventede værdi — ikke en ny observation.

Konfidensintervallet tager kun hensyn til estimationfejlen — ikke observationsfejlen.

95%-prædiktionsinterval:

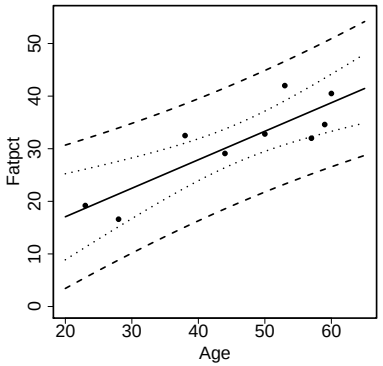
$$\hat{y} \pm t_{0.975,n-2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SS}_x}}$$

Plante på 7 uger:

$$\hat{y} \pm 2.306 \cdot 0.02031 \cdot \sqrt{1 + 0.4216} = (0.2377, 0.3394)$$

Er en fosforkoncentration på 0.25 usædvanlig for en plante på 7 uger?

Konfidensinterval vs. prædiktionsinterval



CI vs. PI:

- Fortolkning: forventet værdi eller ny observation
- PI altid bredere end CI
- CI kan gøres vilkåligt smalt ved at øge n, PI kan ikke

Prædiktion i ensidet ANOVA og en enkelt stikprøve: se afsnit 7.2.3!

Resumé 3: prædiktion

Prædiktion handler om at "forudsige" nye observationer.

- 95%-prædiktionsintervallet indeholder med sandsynlighed 95% en ny observation for en given værdi af de(n) forklarende variabel.
- Et prædiktionsintervaller er altid bredere end det tilsvarende konfidensinterval fordi det også tager hensyn til "observationsfejlen"
- Kan ikke gøres vilkårligt smalle ved at øge n.

Resumé 1–3

Modellerne for lineær regression, ensidet ANOVA og en enkelt stikprøve er "samme suppe".

- Samme antagelser — på nær specifikationen af middelværdien
- To typer forklarende variable: kvantitative og faktorer
- Statistisk inferens "ens": LS-estimation, konfidensintervaller, test, prædiktion, modelkontrol
- Flere forklarende variable kan kobles på — stadig samme modeltype og samme måde at lave statistisk inferens

Modellerne er baseret på normalfordelingen — pga. den centrale grænseværdisætning!

Dagens hovedpunkter

- Multiple sammenligninger — hvorfor er det et problem, og hvad kan vi gøre ved det?
- Modelkontrol
  - Analyse af standardiserede residualer — hvad skal vi se efter?
- Prædiktion. Forskel på konfidens- og prædiktionsintervaller. Udregning.

Ordliste

Engelsk	Dansk
explanatory variable	forklarende variabel
independence	uafhængighed
response variable	responsvariabel
standardized residual	standardiseret residual
outlier	ekstrem observation