Statistik noter der omfatter de fleste emner

Introduction to Statistics (Danmarks Tekniske Universitet)

Grundlæggende

5. maj 2020 12:04

Man bruger statistik til at estimerer parametre udfra data (stikprøver) til at kunne sige noget om pupolationen. Man kan også regne sandsyndelighed eller sammenhænge ud.

Hyppighed
Hvor ofte/mange gange en observation er til stede

Gennemsnit ($\overline{x}$)
Gennemsnittet estimere middelværdien

$$\bar{x} = \frac{obs + obs \dots}{antal}$$

$$\bar{x} = \frac{t_{0,975} - t_{0,025}}{2}$$

Estimeret middelværdi og spredning ($\hat{\mu}$) ($\hat{\sigma}$)
Grundet man ikke kender den rigtige middelværdi eller spredning så bruger man et estimat

Middelværdi (μ)
Det er gennesnittet af populationen. Altså af den totale mulige mængde observationer. Man kan tage alle de mulige udfald og vægte dem med den sandsyndelighed som de forekommer og så får man middeludfaldet. Desto flere observationer desto tættere er estimatet af middelværdien

<u>En terning som eksempel</u>

$$\mu = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} \dots 6 \cdot \frac{1}{6} = 3{,}5$$

Kastes en terning en milliard gange ligger gennemsnittet nok tæt på 3,5.

Median
Medianen er der lige mange observationer under og over.
I nogle tilfælde, f.eks. hvis man har ekstreme værdier, er medianen at foretrække frem for gennemsnittet

Spredning/Standard afvigelse (σ) (s)
Standard afvigelsen er et begreb for hvor meget en stokastisk variabel, fordeler sig ved middelværdien

$$\sigma = \sqrt{Variansen} = \sqrt{(x_1 - \mu)^2 + (x_2 - \mu)^2 \dots}$$

Varians (σ^2) (s^2)
Varians er et begreb inden for sandsynlighedsregning og statistik, der angiver variabiliteten af en stokastisk variabel. Mens middelværdien angiver det niveau, som den stokastiske variabels værdier i gennemsnit ligger på, er variansen et mål for, hvor meget disse værdier i gennemsnit afviger fra middelværdien

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 \dots}{n}$$

Kovarians (s_{xy})
Det er et gennemsnittet af produktet af y og x afvigelser

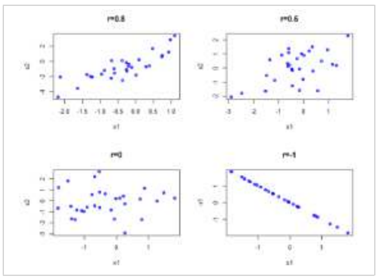
$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Korrelation (r)
Kvantificerer sammenhæng mellem to variabler på
Den måler kun den lineære sammenhæng. Der kan derfor godt være en sammenhæng stadig, selvom der ikke er en lineær sammenhæng

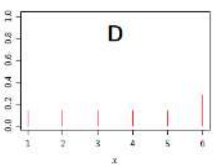
$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{s_{xy}}{s_x \cdot s_y}$$

$$\hat{\rho} = r = \sqrt{R^2} sgn(\hat{\beta}_1)$$

Fx



Tæthedsfunktion/sandsyndelighedsfunktion
Sandsyndligheder for at den stokastiske variabel bliver udfaldet som oberservationen når eksperimentet udføres.
Hvis vi har en observation kan vi så se fordelingen? Nej, men hvis vi har n antal observationer, så har vi en stikprøve og dermed kan man se fordelingen

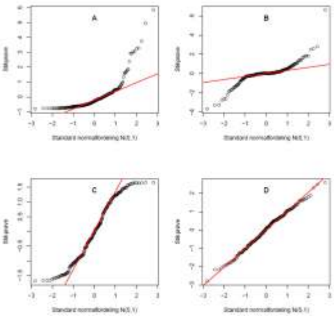


Fordelingsfunktion
Fordelingsfunktion er tæthedsfunktionen akkumuleret

Stokastisk variabel (X)
Værdi afhængig af udfald af endnu ikke udført eksperiment. Stokastisk betyder tilfældig

Observation (x)
Data afhængig af udfald af udført eksperiment.

Qqplot
Ofte vil man være interesseret i at afgøre om en stikprøve kan antages at stamme fra en bestemt fordeling (f.eks. en normalfordeling)



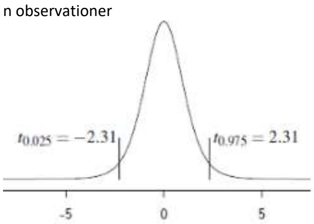
Scatterplot
Et plot der viser værdier, typisk to variabler plottet mod hinanden

Histogram
Et histogram er en måde grafisk at vise et datasæt på, som illustrerer hyppigheden, værdier i datasættet forekommer med.

Konfidens interval (KI)
Til hver stikprøve angives et interval, hvor man ud fra en hvis sikkerhed antager at parameteren fx gennemsnittet ligger indenfor intervallet. Ved formlen fås begge værdier og dermed et interval

$$\bar{x} \pm t_{0.975} \cdot \frac{s}{\sqrt{n}}$$

X er stikprøvegennemsnittet, $t_{0.975}$ er den øvre fraktil, s er estimeret spredning og n observationer



Vi kan bruge konfidens-interval baseret på t-fordelingen i stort set alle situationer, blot n er "stor nok" (n > 30)
Man kan også lave KI for spredningen og variansen (ikke færdiggjort)

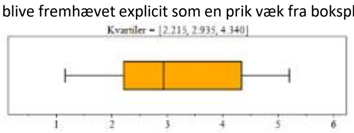
Signifikans niveau (α)
I praksis vælger man forud for den statistiske afprøvning eller test, ved hvilket signifikansniveau man vil forkaste nulhypotesen. Ofte 5% anvendes
$\alpha = 0{,}05$

Evidens
Evidens for noget er at det antages at være indtil "vished grænsende sandsynligt", men ikke er et faktum

Fraktiler

Nedre kvartil	25%	Q_1
Median	50%	m
Øvre kvartil	75%	Q_3

Boksplot
Kassens øvre og nedre grænse viser øvre og nedre kvartil, og kassen indeholder således halvdelen af de observerede værdier. Ved en ekstrem observation vil den blive fremhævet explicit som en prik væk fra boksplottet.

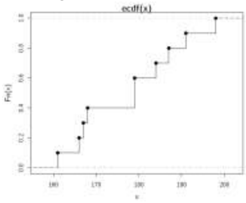


Sandsyndelighedsfordeling
Man skal anvende den sandsyndelighedsfordeling til det fænomen, som man står over for. Altså for at beregne sandsyndeligheden skal man vælge den rigtige fordeling

Teststørrelsen (T_{obs})

$$t = \frac{\bar{x} - \mu_0}{\dots}$$

Fordelingsfunktion
Fordelingsfunktion er tæthedsfunktionen akkumuleret



The central limit theorem
Gennemsnittet af en tilfældig stikprøve følger uanset hvad en normalfordeling
<u>I hvert fald for én sidet test</u>

Chi i anden
Variansestimater opfører sig som en X^2-fordeling

Let
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$
then:
$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$
is a random variable following the χ^2-distribution with $v = n-1$ degrees of freedom.

Konfidensinterval for varians og spredning (chi i anden)
Variansen:

A $100(1-\alpha)\%$ confidence interval for the variance σ^2 is:
$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}};\ \frac{(n-1)s^2}{\chi^2_{\alpha/2}}\right]$$
where the quantiles come from a χ^2-distribution with $v = n-1$ degrees of freedom.

Standardafvigelsen:

A $100(1-\alpha)\%$ confidence interval for the sample standard deviation $\hat{\sigma}$ is:
$$\left[\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}};\ \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}}\right]$$

<u>For at få X^2 kvartilerne skal man anvende i Rstudio:</u> (tjek efter om det er 0.025 og 0.975)
qchisq(c(0.025, 0.975), df = 10)

Outliers
En ekstrem værdi som kan ligge langt væk fra resten af dataen fx en basketball spiller mellem almindelige mennesker

over for. Altså for at beregne sandsynligheden skal man vælge den rigtige fordeling

Teststørrelsen (T_{obs})
$$t_{obs} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

P-værdi
Pværdi er sandsynligheden for at opnå testresultater mindst lige så ekstreme som de resultater, der faktisk blev observeret under testen, under forudsætning af, at nulhypotesen er korrekt.
$p_{værdi} = 2 \cdot P(T > |t_{obs}|)$

$p < 0.001$	Very strong evidence against H_0
$0.001 \le p < 0.01$	Strong evidence against H_0
$0.01 \le p < 0.05$	Some evidence against H_0
$0.05 \le p < 0.1$	Weak evidence against H_0
$p \ge 0.1$	Little or no evidence against H_0

Tabel for om der er evidens ud fra p værdien

Hypotse
Nulhypotese opstilles
Der ingen forskel mellem de to sovemidler
H_0: $\mu_0 = 0$

$P_{værdi} < \alpha$ = afvis H_0
$|t_{obs}| > t_{1-\frac{\alpha}{2}}$ = afvis H_0

Hypotesetest med alternativer
H_1
Type 1 og 2 fejl
(ikke færdiggjort)

Kritiske værdier
$t_{\frac{\alpha}{2}}$ og $t_{1-\frac{\alpha}{2}}$
Det er de værdier for en t fordeling hvor H_0 afvises hvis de er uden for de kritiske værdier. Ækvivalent med KI

Eftertjek af nulhypotese
Opstil nulhypotsen
↓
Bestem signifikansniveauet
↓
Find t_{obs}
↓
Udregn $p_{værdi}$
↓
$P_{værdi} < \alpha$ = afvis H_0

Diskret

5. maj 2020 13:05

<u>Bemærk at tæthedsfunktion, fordelingsfunktion, kovarians, korrelation og middelværdi har anden betydning for kontinuere tal end diskrete</u>

Disket sandsyndelighed

De kan peges på
Hvor mange der bruger briller herinde
Antal mange flyvere letter den næste time

Konkrete statistiske fordelinger

Binomial
Hypergeometrisk
Poissonfordeling

Binomial fordeling (enten eller)

Den beskriver sandsynligheden for at få X succeser i n uafhængige identiske forsøg. <u>Hvert forsøg har to mulige udfald som kan være plat eller krone, god eller dårlig osv.</u>
Binomialfordelingen anvendes også for at analysere stikprøver <u>med tilbagelægning</u>

<u>X følger binomial fordeling:</u>
$X \sim B(n, p)$
n er antal gentagelser
p er sandynligheden for succes i hver gentagelse

<u>Tæthedsfunktion</u>:
$f(x; n, p) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$

Hypergeometrisk fordeling

Når man vil analysere stikprøver <u>uden tilbagelægning</u> anvendes den hypergeometriske fordeling <u>(tænk på træk fra en hat)</u>
Når noget er udvalgt ændrer sandsynligheden sig

<u>X følger hypergeometrisk fordeling:</u>
$X \sim H(n, a, N)$
n er antallet af trækninger
a er antallet af succeser i populationen
N elementer store population

<u>Tæthedsfunktion</u>:
$f(x; n, a, N) = P(X = x) = \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}}$

Poissonfordelingen

Poissonfordeling er en diskret sandsynlighedsfordeling, som anvendes for at beskrive hændelser, som indtræffer uafhængigt af hinanden
Poissonfordelingen karakteriseres ved en intensitet, dvs. p˚a formen antal/enhed

<u>X følger poissonfordelingen</u>
$X \sim P(\lambda)$
Parameteren l angiver intensiteten

<u>Tæthedsfunktion</u>**:**
$f(x) = P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$

Tilsvarende middelværdi og varians

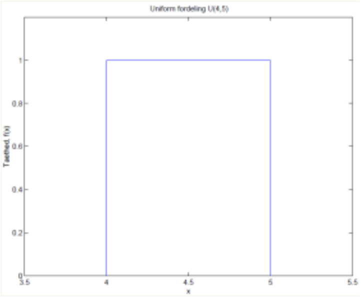
Fordeling	Middelværdi	Varians
Binomialfordelingen	$\mu = n \cdot p$	$\sigma^2 = n \cdot p \cdot (1-p)$
Hypergeometrisk	$\mu = n \cdot \frac{a}{N}$	$\sigma^2 = \frac{na \cdot (N-a) \cdot (N-n)}{N^2 \cdot (N-1)}$
Poissonfordelingen	$\mu = \lambda$	$\sigma^2 = \lambda$

Kontinuer

<u>Bemærk at tæthedsfunktion, fordelingsfunktion, kovarians, korrelation og middelværdi har anden betydning for kontinuere tal end diskrete</u>

Kontinuere fordelinger
Der er uendelige muligheder for tal

Uniform/lige fordeling
Ligefordeling er en sandsynlighedsfordeling, hvor alle udfald har lige stor sandsynlighed



Skrivemåde:

$X \sim U(\alpha,\beta)$ (Læses: X følger en uniform fordeling med parametre α og β)

Tæthedsfunktion:

$f(x) = \frac{1}{\beta-\alpha}$

Middelværdi:

$\mu = \frac{\alpha+\beta}{2}$

Varians:

$\sigma^2 = \frac{1}{12}(\beta-\alpha)^2$

Normalfordeling
Symmetrisk

Skrivemåde:

$X \sim N(\mu,\sigma^2)$

Tæthedsfunktion:

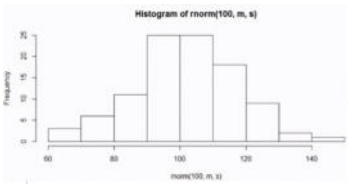
$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

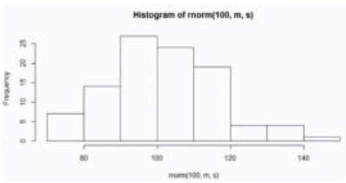
Middelværdi:

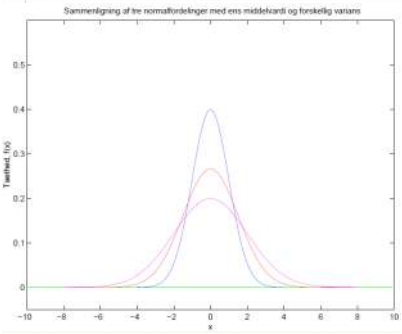
$\mu = \mu$

Varians:

$\sigma^2 = \sigma^2$







Log transformeret normalfordling
Højre skæv normalfordeling

Standard normalfordeling
Spredning er 1 og den ligger omkring 0.
2,5% punktet er 1,96 til højre fra 0

$$Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0{,}1^2)$$

Den følger en normalfordeling med middelværdi 0 og varians 1

Omformning fra normalfordelt til standard normalfordelt

En vilkårlig normalfordelt variabel $X \sim N(\mu,\sigma^2)$ kan standardiseres ved at beregne

$$Z = \frac{X-\mu}{\sigma}$$

En log-normalfordelt variabel $Y \sim LN(\alpha,\beta^2)$, kan transformeres til en standard normalfordelt variabel Z ved

$$Z = \frac{\ln(Y)-\alpha}{\beta}$$

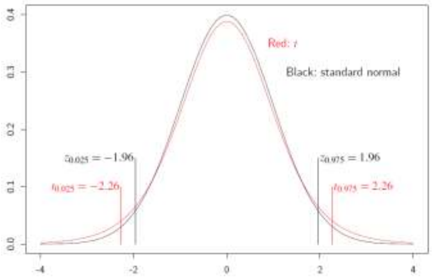
dvs.

$$Z \sim N(0,1^2)$$

T fordeling
Bemærk, der kan være mange t fordelinger med mange forskellige måder at udregne frihedsgrader på
<u>En t fordeling er et estimat af standard normalfordeling.</u>
Det følger en t fordeling når vi bruger vores estimat af spredning i stedet for spredning for populationen. Altså en t fordeling anvendes når man beregner gennemsnittet af en normalt fordelt population i situationer hvor stikprøvestørrelsen er lille og populationsstandardafvigelsen er ukendt

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t$$

Frihedsgrader for t fordelingen er n - 1



Her er der n=10 observationer, men med n-1=9 frihedsgrader

Skrivemåde:

$X \sim LN(\alpha, \beta^2)$ (Hvis X følger log-normal så følger $\ln(X)$ normal)

Tæthedsfunktion:

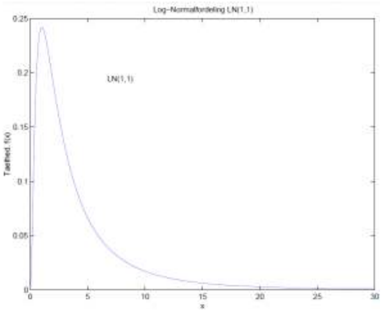
$$f(x) = \begin{cases} \frac{1}{x\sqrt{2\pi}\beta} e^{-(\ln(x)-\alpha)^2/2\beta^2} & x > 0,\ \beta > 0 \\ 0 & \text{ellers} \end{cases}$$

Middelværdi:

$\mu = e^{\alpha + \beta^2/2}$

Varians:

$\sigma^2 = e^{2\alpha+\beta^2}(e^{\beta^2} - 1)$



Eksponentiel fordeling

Skrivemåde:

$X \sim Exp(\lambda)$

Tæthedsfunktionen

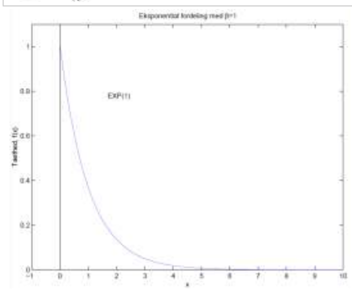
$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{ellers} \end{cases}$$

Middelværdi

$\mu = \frac{1}{\lambda}$

Varians

$\sigma^2 = \frac{1}{\lambda^2}$



Two sample test

10. maj 2020 15:49

To sidet test
Det er det mest almindelige at have to sidet test. Stik prøve sammenlignet med en anden stikprøve

Teststørrelse (T_{obs}) (Welch t-test)
Beregning af teststørrelsen (t_{obs})

When considering the null hypothesis about the difference between the means of two *independent* samples

$$\delta = \mu_2 - \mu_1 \quad \textit{(delta er forskellen i middelværdi)}$$
$$H_0: \ \delta = \delta_0 \quad \textit{(typisk er } \delta_0 = 0\textit{)}$$

the (Welch) two-sample t-test statistic is

$$t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Med v som frihedsgrader

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Fremgangsmetode til at afprøve nulhypotesen for to sidet

1. Compute the test statistic using Equation (3-48) and ν from Equation (3-50)

$$t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \quad \text{and} \quad \nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

2. Compute the evidence against the *null hypothesis*

$$H_0: \quad \mu_1 - \mu_2 = \delta_0,$$

vs. the *alternative hypothesis*

$$H_1: \quad \mu_1 - \mu_2 \neq \delta_0,$$

by the

$$p\text{-value} = 2 \cdot P(T > |t_{\text{obs}}|),$$

where the t-distribution with ν degrees of freedom is used

3. If p-value $< \alpha$: we reject H_0, otherwise we accept H_0,
or
The rejection/acceptance conclusion can equivalently be based on the critical value(s) $\pm t_{1-\alpha/2}$:
if $|t_{\text{obs}}| > t_{1-\alpha/2}$ we reject H_0, otherwise we accept H_0

Konfidensinterval (Welch t-test)

$$\bar{x} - \bar{y} \pm t_{1-\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Følger en t fordeling og har v antal friehdsgrader
Konfidens intervaller kan også overlappe for en to sidet test. Hvis de overlapper kan man ikke konkluderer noget som helst.

Parrede sample test
En parret t test er når to ting bliver bliver målt imod én ting som er konstant.
Fx en person prøver to forskellige sovemedicin eller når temperatur bliver målt vinter og sommer
Det er en fordel at anvende oftest, da det kan have stor effekt på spredning og varians at sammenligne to individer som afprøves to gange end fire indivder afprøves en enkel gang

Lineær regression med flere variabler

3. april 2020 09:05

Dette laves for at lave modeller som har sammenhænge
Altså, vi er interesseret i at modellere Y's afhængighed af de forklarende eller uafhængige variabler (explanatory eller independent variables) x1, x2,..., xp

Simpel Lineær regressionsmodel

Lineær sammenhæng mellem Y og x1, x2,..., xp, ved en regressionsmodel på formen

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, \sigma^2) \text{ og i.i.d.}$$

Y_i er en stokastisk afhængig variabel (responsvariabel), ε_i er afvigelsen som er stokastiske variabel, x er forklarende variabel og β er regressionskoefficient
β_0 er skæringen og β_1 er hældningen

Multipel lineær regression

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, \sigma^2) \text{ og i.i.d.}$$

Y_i er en stokastisk afhængig variabel (responsvariabel), ε_i er afvigelsen som er stokastiske variabel, x er forklarende variabel og β er regressionskoefficient
β_0 er skæringen og β_1 er hældningen

Antagelse til lineær regressions

$\varepsilon_i \sim N(0, \sigma^2)$ og i.i.d.
Der er den antagelse, at afvigelsen (ε) er normalfordelt med gennemsnittet 0 og med konstant varians. Observationerne er uafhængige og identiske fordelte, hvilket betyder, at observationerne er tilfældigt valgt af populationen.

Backward / forward selection

Udvide modellen ved at anvende backward forward selection
At kun medtage de signifikante sammenhænge for at ende op med en model. Model udvælgelsen kan så tolkes ud fra den model man har valgt.

Residualer

Afvigelserne
Residual betyder det ''resterende'' eller ''det, der er til overs''.

Standard error / Residual error / Residual varians ($\hat{\sigma}_{\beta,i}$)

Residual varians angiver hvor usikkert en regressionskoefficient er bestemt
Det er afstanden fra observationen til linjen hvor meget hældningen kan varierer fra gang til gang

Ikke lineær sammenhænge (kurveliniær) (polyomial regression)

Man kan også lave sammenhænge hvis man forsøger at sætte variablerne i anden x^2
Dette gøres hvis modelkontrolle afvises

Hvis vi ønsker at estimere en model af typen

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

kan vi benytte multipel lineær regression i modellen

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i$$

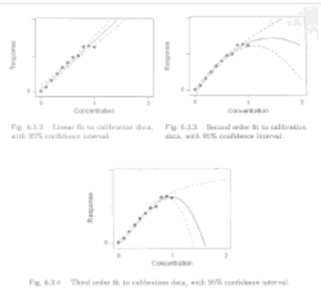


Fig. 6.3.2. Linear fit to calibration data, with 95% confidence interval.
Fig. 6.3.3. Second order fit to calibration data, with 95% confidence interval.
Fig. 6.3.4. Third order fit to calibration data, with 95% confidence interval.

Teststørrelse

$$t_{\text{obs}} = \frac{\hat{\beta}_0}{\hat{\sigma}_{\beta_0}}$$

Konfidensinterval

$$\hat{\beta}_0 \pm t_{1-\alpha/2}\, \hat{\sigma}_{\beta_0}$$
$$\hat{\beta}_1 \pm t_{1-\alpha/2}\, \hat{\sigma}_{\beta_1}$$

n-2 frihedsgrader
i R kan $\hat{\sigma}_{\beta_0}$ og $\hat{\sigma}_{\beta_1}$ aflæses ved "Std. Error" ved "summary(fit)"

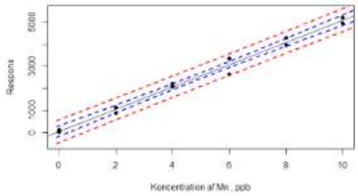
Fittede værdier ($\hat{y}$)

Model kontrol

Studer residualer for at undersøge om der er noget galt med den lineære model man prøver at opstille. Ellers kan man jo ikke regne på det, hvis modellen ikke passer til vores lineære antagelser
Man laver model kontrollen efter analysen grundet man vil kigge på residualerne af analysen og det kan man først efter

<u>Residual mod fittede værdier</u>
Kigges efter trumpeter = Ikke homogent så = ikke varians konstant
Skulle meget gerne være en vandret linje (ikke perfekt) = varians er konstant så

<u>Residual mod x variabel</u>
Kigges efter linearitet og krumninger = problem med linearitets antagelsen
Fx



Den blå er konfidens interval og den røde er prædiktions interval
Konfidens er usikkerheden på middelværdien
Prædiktions er usikkerheden på en enkel observation
Det er smallere inde på midten pga begge intervaller indeholder usikkerhed på hældningen og afskæringen som bidrager til usikkerheden

Opsummering fra slides

Model kontrol: Analyser residualerne for at checke at forudsætningerne er opfyldt

$\varepsilon_i \sim N(0, \sigma^2)$ og er independent and identically distributed (i.i.d.)
- Husk: ε_i er afvigelsen (en stokastisk variabel)
- Husk: $e_i = \hat{\varepsilon}_i$ er residualet (realisationen eller observationen af afvigelsen)

Samme som for simpel lineær model, dog også plot med residualer vs. inputs

Forudsætninger
- Uafhængige observationer (tænk)
- Normalfordelte fejl (plot).
- Konstant varians σ^2 (plot).
- En lineær sammenhæng (plot).

Spørgsmål til eksamen

Hvilket udsagn nedenfor repræsenterer <u>ikke</u> en nødvendig antagelse for en simpel lineær regressionsmodel?

1 ☐ Fejlene ε_i er uafhængige.
2 ☐ Fejlene ε_i er identisk fordelte.
3 ☒ De afhængige variable Y_i er identisk fordelte.
4 ☐ De afhængige variable Y_i er uafhængige.
5 ☐ De afhængige variable Y_i og fejlene ε_i har samme varians.

Summary i Rstudio

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.571 -0.673  0.132  0.745  2.190
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.109      0.482   -0.23     0.82
## size            0.127      0.023    5.50  1.2e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.979 on 51 degrees of freedom
## Multiple R-squared:  0.372,Adjusted R-squared:  0.36
## F-statistic: 30.2 on 1 and 51 DF,  p-value: 1.25e-06
```

- `Residuals:      Min      1Q   Median      3Q      Max`
  Residualernes: Minimum, 1. kvartil, Median, 3. kvartil, Maximum
- `Coefficients:`
  `Estimate Std. Error t value Pr(>|t|) "stjerner"`
  Koefficienternes:
  Estimat $\hat{\sigma}_{\beta_i}$ t_{obs} p-værdi
  - Testen er $H_{0,i}: \beta_i = 0$ vs. $H_{1,i}: \beta_i \neq 0$
  - Stjernerne er sat efter p-værdien
- `Residual standard error: XXX on XXX degrees of freedom`
  $\varepsilon_i \sim N(0, \sigma^2)$: Udskrevet er $\hat{\sigma}$ og ν frihedsgrader (brug til hypotesetesten)
- `Multiple R-squared:  XXX`
  Forklaret varians r^2

<u>Multiple R-squared</u> er er andelen af den totale variation som er forklaret af modellen (forklarende variabel/korrelation)
<u>Std. Error</u> er regressionskoefficienternes varians
<u>DF</u> er graderne af frihed som er udregnet udfra variablerne altså DF = n observationer - x antal variabler

BMI Projek
2 aflevering

Eksempel

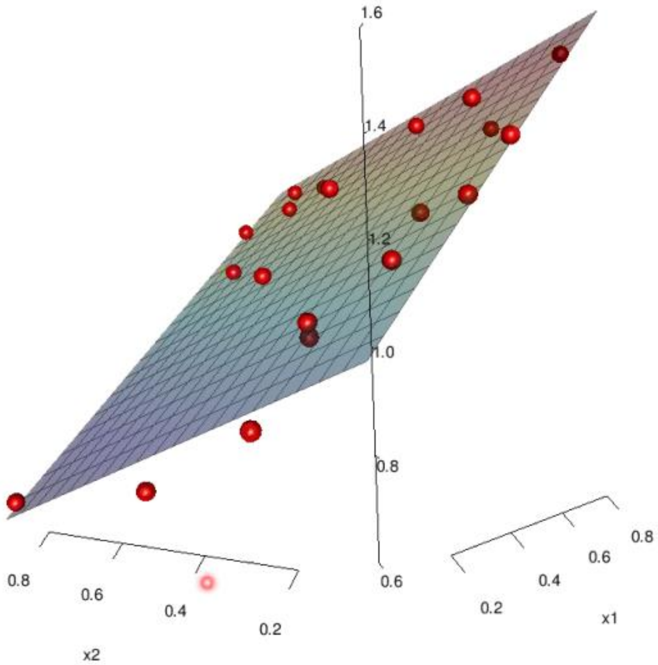
14. februar 2020 08:34

Backward / forward selection
At kun medtage de signifikante sammenhænge for at ende op med en model. Model udvælgelsen kan så tolkes ud fra den model man har valgt

fx
Model for ozon
Temperatur + wind + noise
Da noise ikke er signifikans (pga det ikke har lavere end 5%) så medtages det ikke i modellen

B0 skæringen
B1 hælding x1
B2 hældning x2

Her er planen
C: $\hat{\beta}_0 > 0$, $\hat{\beta}_1 > 0$ og $\hat{\beta}_2 < 0$



Fx for backward selection

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9195 -0.1555  0.0104  0.1465  0.6304
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.0528     0.2285   -0.23   0.8201
## x1           -0.7357     0.3034   -2.42   0.0275 *
## x2            0.2618     0.2937    0.89   0.3859
## x3            1.1817     0.3553    3.33   0.0043 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.37 on 16 degrees of freedom
## Multiple R-squared:  0.507,Adjusted R-squared:  0.414
## F-statistic: 5.48 on 3 and 16 DF,  p-value: 0.00878
```

Skal modellen reduceres i backward selection step?

A: Nej B: Ja, x_1 skal væk C: Ja, x_2 skal væk D: Ja, x_3 skal væk

Svar C: Ja, x_2 skal væk, den er ikke signifikant forskellig fra 0 og mest insignifikant

Varians regneregler

12. maj 2020 16:49

Theorem 2.54 Mean and variance of linear functions

Let $Y = aX + b$ then

$$\mathrm{E}(Y) = \mathrm{E}(aX + b) = a\,\mathrm{E}(X) + b, \tag{2-71}$$

and

$$\mathrm{V}(Y) = \mathrm{V}(aX + b) = a^2\,\mathrm{V}(X). \tag{2-72}$$

Theorem 2.56 Mean and variance of linear combinations

The mean of a linear combination of independent random variables is

$$\mathrm{E}(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1\,\mathrm{E}(X_1) + a_2\,\mathrm{E}(X_2) + \cdots + a_n\,\mathrm{E}(X_n), \tag{2-73}$$

and the variance

$$\mathrm{V}(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1^2\,\mathrm{V}(X_1) + a_2^2\,\mathrm{V}(X_2) + \cdots + a_n^2\,\mathrm{V}(X_n). \tag{2-74}$$

https://02323.compute.dtu.dk/enotes/chapter2-ProbabilitySimulation

Regneregler eksempel

Assume that X is normally distributed with mean 10 and variance 4, Y is normally distributed with mean 20 and variance 25, and X and Y are independent.

Question IV.1 (8)

Then $2Y - 2X + 4$ has the variance:

1 ☐ 36

2 ☐ 58

3 ☐ 84

4* ☐ 116

5 ☐ None of the values above.

------------------------------------ FACIT-BEGIN ------------------------------------

Use the variance identities in Theorem 2.54 and 2.56 to get

$$V(2Y - 2X + 4) = 4\,V(Y) + 4\,V(X) = 4 \cdot 4 + 4 \cdot 25 = 116.$$

Or you can simulate it:

```
k <- 100000
x <- rnorm(k, 10, sqrt(4))
y <- rnorm(k, 20, sqrt(25))
z <- 2*y - 2*x + 4
var(z)
```

```
## [1] 116.0609
```

Eksamen

16. maj 2020 13:50

Hvad er sandsynligheden for at det vejer mere end 25 kg (maj 18

$X_1 \sim N(20, 5^2)$ kg

```
1 - pnorm(q = 25, mean = 20, sd = 5)

## [1] 0.1586553
```

1-pnorm er når det er over middel men fjern 1- når der er under fx hvis det havde været 19 kg

Hvor meget vejer det når der er 20% chance (maj 18

$X_2 \sim N(50, 10^2)$ kg.

```
qnorm(0.8, mean = 50, sd = 10)

## [1] 58.41621
```

Hvad er 90% konfidensintervallet for variansen (chi i anden)

Middelværdi = 6

```
(6-1)*var(x)/qchisq(c(0.95, 0.05), df = 6-1)

## [1]  3.695257 35.712948
```

Bestem IQR

The IQR is the difference between the 0.25 and 0.75 sample quantiles, here computed using Definition 1.7:

```
quantile(x, 0.75, type = 2) - quantile(x, 0.25, type = 2)

##  75%
## 26.1
```

Hvad er sandsynligheden for binomial

```
dbinom(0, size = 20, p = 0.2)

## [1] 0.01152922

pbinom(0, size = 20, p = 0.2)

## [1] 0.01152922
```

For at få X^2 kvartilerne skal man anvende i Rstudio:

qchisq(c(0.025, 0.975), df = 10)

Bestem konfidensintervallerne for simulationen ved brug af R kode ved parametrisk bootstrapping (maj 18

```
set.seed(7643)
k <- 10000
R1 <- rnorm(k, mean = 2, sd = 0.2)
R2 <- rnorm(k, mean = 3, sd = 0.5)
R <- 1/(1/R1 + 1/R2)
```

Anvend:

```
quantile(R, c(0.025,0.975))

##      2.5%     97.5%
## 0.9647361 1.4016874
```

Prop.test (maj 18

Undersøger hypotese

```
2* □  prop.test(x = c(7297, 6691), c(30817, 30193), correct = FALSE)
```

Frihedsgrader ved søjle og række data (18 maj

	Y2012	Y2016	Y2017	*Sum*
Hum	7966	7297	6691	*21954*
Samf	10173	10253	10006	*30432*
Sund	2789	3137	3157	*9083*
TekNat	8551	10130	10339	*29020*
Sum	*29479*	*30817*	*30193*	*90489*

$(række - 1) * (søjle - 1) = frihedsgrad$

Forventede antal i teknat 2017 (maj 18

$$\frac{\text{column total} \times \text{row total}}{\text{grand total}}$$

Udregn p værdi og teststørrelsen

```
tobs<-(2.38)
tobs

pvalue <- 2 * (1-pt(abs(tobs), df=19))
pvalue
```

Lav vektor og t.test (datasæt)

```
x <- c(-6.5, -6.5, -5.2, -4.9, -4.8, -2.9, -2.6, -2.0, -1.9, -1.8, -1.5,
       -0.1, 1.9, 2.1, 2.6, 5.2, 8.0)

t.test(x)

##
##  One Sample t-test
##
## data:  x
## t = -1.24, df = 16, p-value = A
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -3.330  0.871
## sample estimates:
## mean of x
##      B
```

Test af nulhypotese med en given my

t.test(x, mu=0)

22. maj 2020 16:02

Ikke 'metrisk metoder

22. maj 2020 13:55

Ikke parametrisk test

Der er ikke behov for at følge en fordeling. En god idé ved lav data
Ikke-parametriske metoder fokuserer på p-værdier og ikke konfidens interval eller estimater
Ofte bruges ikke-parametriske metoder, når fordelingerne er meget skæve. Så er median mere sigende for centret i fordelingen end middelværdien.
Man bliver ikke nød til at antage noget

Sign test/fortegns test

Man kigger på differensen og fortegnet så

p=P(positiv differens)

ikke at forveksle med en 'p-værdi', men er fortegnet på forskellen

Wilcoxon eller Mann-Whitney test (to sidet)

Ikke-parametrisk alternativ til to-stikprøve t-test
Vi vil gerne sammenligne summen af rangen i de to grupper
"Wilcoxon- angtest er en ikke-parametrisk hypotestest, der bruges til at sammenligne to prøver på en enkelt prøve for at vurdere, om deres gennemsnitlige antal er forskellige"

$$Z = \frac{U_1 - E(U_1)}{\sigma_{U1}} \sim N(0,1)$$

<u>Middelværdi og varians</u>

$\mu_{U1} = E(U_1) = \frac{n_1 n_2}{2}$ og $\sigma_{U1}{}^2 = var(U_1) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$

Rang

Den laveste observation får rang=1, den næste rang=2 uafhængigt af om det er fra den ene eller anden stikprøve. Ekstreme observationer har altså mindre indflydelse ved at kigge på rang

Kruskal-Wallis test (fler sidet test)

Man vil lave varians analyse

$$H = \frac{12}{n \cdot (n+1)} \left(\sum_{i=1}^{k} \frac{R_i^2}{n_i} \right) - 3 \cdot (n+1) \quad H \sim \chi^2_{(k-1)}$$

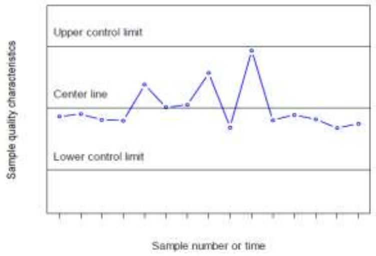
R er rangsummen

Simpel Kontrolkort

22. maj 2020 16:03

Kontrol
Ved kontrol kigger man efter variation som ikke er naturlige

Shewhart kontrolkort
Prikkerne er middelværdien som skal være inden for kontrolgrænserne

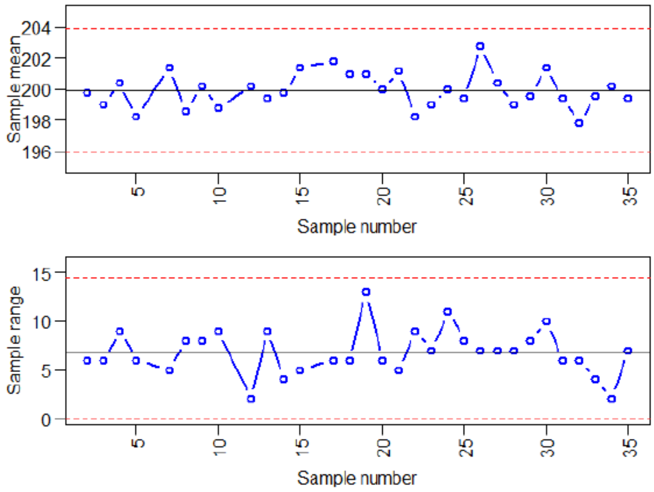


Actioin limits
Sættes til 3

Warning limits
Sættes til 2

R- og X-kontrol kort
Man kigger først på R kort før X kort

Her er A_2, D_3, D_4 konstanter



<u>X-kontrolkort</u>

Kigges på om middelværdien er skæv. Det øverste

$$UCL = \bar{\bar{X}} + A_2\,\bar{R}$$
$$Center = \bar{\bar{X}}$$
$$LCL = \bar{\bar{X}} - A_2\,\bar{R}$$

$$LCL = \bar{X} - 3 \cdot \frac{\overline{MR}}{d_2}$$
$$Center = \bar{X}$$
$$UCL = \bar{X} + 3 \cdot \frac{\overline{MR}}{d_2}$$

<u>R-kontrolkort</u>

Kigges på variansen/spredningen. Det nederste

$$UCL = D_4\,\bar{R}$$
$$Center = \bar{R}$$
$$LCL = D_3\,\bar{R}$$

Fase I
Skabe en process i kontrol og finde grænserne man vil arbejde med

Fra hendes slides (dovenskan)

- For at sætte kontrolkortet i gang har man brug for en periode hvor processen er i kontrol.
- **Fase I er en retrospektiv** analyse af nogle indsamlede data (gerne 20-25 stikprøver).
- De bruges til at konstruere "prøve" grænser.
- Hvis processen er i kontrol, bruges grænserne **fremadrettet til at overvåge processen (Fase II).**
- Hvis ikke Fase I er i kontrol, gås alle ude af kontrol punkter igennem for at se om man kan finde årsagen.
  - Findes årsagen udelukkes punktet, og man genberegner.
  - Hvis ikke overvejes om man vil beholde/ ikke beholde punktet.

Fase II
Processen kører og man tager stikprøver og ser om målingerne ligger inde for grænserne

Fra hendes slides (dovenskan)

- **Fase II er fremadrettet.**
- Man antager at processen er nogenlunde stabil.
- Fokus er på at **overvåge** processen, ikke på at opnå kontrol.
- Kontrolgrænserne fra Fase I bruges.
- Når kortet har kørt i kontrol et stykke tid vil man revidere det. F.eks. Hver måned, hvert år, efter 25 nye samples, 50 nye samples.
- Der skal helst være mindst 25 samples til at beregne de nye kontrolgrænser.

Moving range

Kigger på spredningen ved kun få observationerne

$$MR_i = |X_i - X_{i-1}|$$

Hvor n = 2

$$LCL = D_3\,\overline{MR} = 0 \cdot \overline{MR}$$
$$Center = \overline{MR}$$
$$UCL = D_4 \cdot \overline{MR} = 3.267 \cdot \overline{MR}$$

Kapabilitet

En yderligerer grænse som er en specifikations grænse. Det er nogen man selv fastsætter

$$\widehat{C_p} = \frac{USL-LSL}{6\hat{\sigma}}$$

"Upper specific limit"
Gerne være større end 1

Kapabilitet hvis processen ikke er centreret

$$C_{pk} = \min(C_{pu}, C_{pl})$$

$$C_{pu} = \frac{USL-\mu}{3\sigma} \text{ og } C_{pl} = \frac{\mu-LSL}{3\sigma}$$

- Hvis $C_p = C_{pk}$ så er processen centreret midt i specifikationsintervallet
- Hvis $C_{pk} < C_p$ så er processen forskubbet.
- **Både C_p og C_{pk} bygger på en antagelse om normalfordelingen**

Kalibrering /simpel regressions

23. maj 2020 13:24

Kalibrering

Man anvender lineær regression med y mod x og derved aflæser x ved at komme med et y

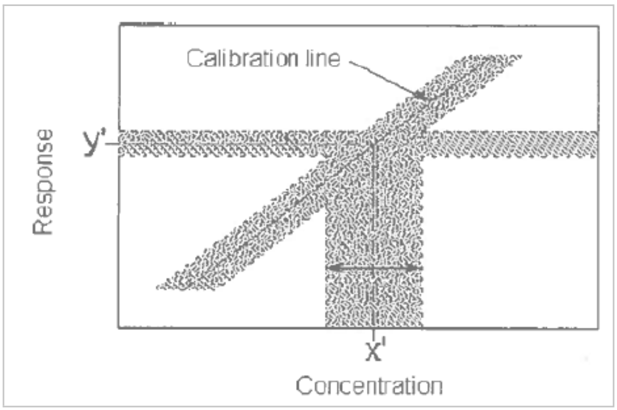
$$x_0 = \frac{y_0 - \hat{\beta}_0}{\hat{\beta}_1}$$

Spredningen

$$s_{x0} = \frac{\hat{\sigma}}{|\hat{\beta}_1|}\sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(y_0 - \bar{y})^2}{\hat{\beta}_1^{\,2} \sum (x_i - \bar{x})^2}} = \frac{\hat{\sigma}}{|\hat{\beta}_1|}\sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

M er antal gange man har målt samme værdi. N er antal data punkter

Der er usikkerhed og det kan ses her



Konfidens grænser (hvor usikker x er bestemt)

$$x_0 \pm t_{\frac{\alpha}{2}, n+m-3} s_{x0}$$

Detektionsgrænse

Detektionsgrænsen (C_L) er den største koncentration hvor 0 er med i Konfidensintervalle. Hvis 0 er med i konfidensintervallet, kan man ikke afvise at x0 er 0

Ikke-lineær regression

…

Mutipel regressionsanalyse

25. maj 2020 12:07

Mutipel regressionsanalyse

Man kigger på sammenhængen
Ikke længere en linje men nu en plan (3 dimensioner)

$$Y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \beta_3 \cdot x_{i3} + \ldots\ldots + \beta_p \cdot x_{ip} + e_i$$

$$e_i \sim N(0, \sigma^2) \quad i.u.s.v. \quad i = 1, 2, \ldots, n$$

e er epsilon

Fx

<u>Opstillet som ligning system</u>

$$y_1 = \beta_0 + \beta_1 \cdot x_{11} + \beta_2 \cdot x_{12} + \beta_3 \cdot x_{13} + e_1$$
$$y_2 = \beta_0 + \beta_1 \cdot x_{21} + \beta_2 \cdot x_{22} + \beta_3 \cdot x_{23} + e_2$$
$$y_3 = \beta_0 + \beta_1 \cdot x_{31} + \beta_2 \cdot x_{32} + \beta_3 \cdot x_{33} + e_3$$
$$y_4 = \beta_0 + \beta_1 \cdot x_{41} + \beta_2 \cdot x_{42} + \beta_3 \cdot x_{43} + e_4$$
$$y_5 = \beta_0 + \beta_1 \cdot x_{51} + \beta_2 \cdot x_{52} + \beta_3 \cdot x_{53} + e_5$$
$$y_6 = \beta_0 + \beta_1 \cdot x_{61} + \beta_2 \cdot x_{62} + \beta_3 \cdot x_{63} + e_6$$

<u>På matrix form</u>

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \boxed{\begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ 1 & x_{31} & x_{32} & x_{33} \\ 1 & x_{41} & x_{42} & x_{43} \\ 1 & x_{51} & x_{52} & x_{53} \\ 1 & x_{61} & x_{62} & x_{63} \end{bmatrix}} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Mindstre kvadraters modetode

Mindste kvadraters metode er en standard fremgangsmåde til at finde den bedste løsning for et overbestemt system. Altså den mindste afstand til planen
Den bedste, skal her forstås som dén løsning der giver den mindste sum af kvadraterne på fejlene i hver enkelt ligning.

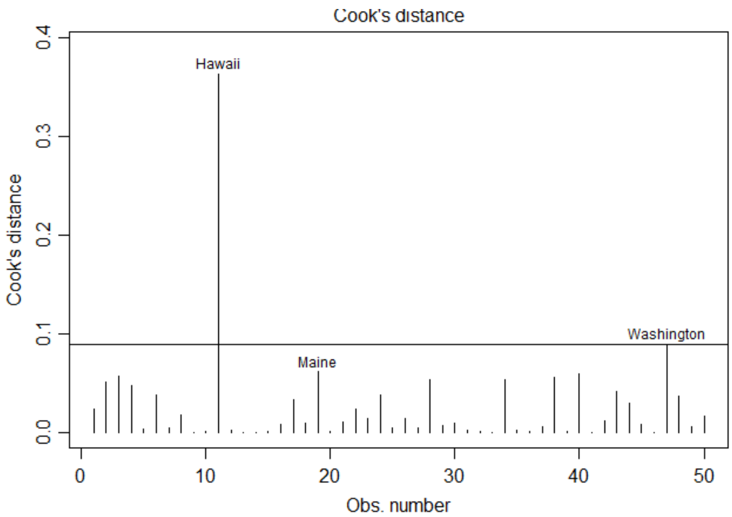
generalized additive models

Cooks afstand

Beregningen på om en måling skal udelukkes. De kan have indflydelse på beta parametrene

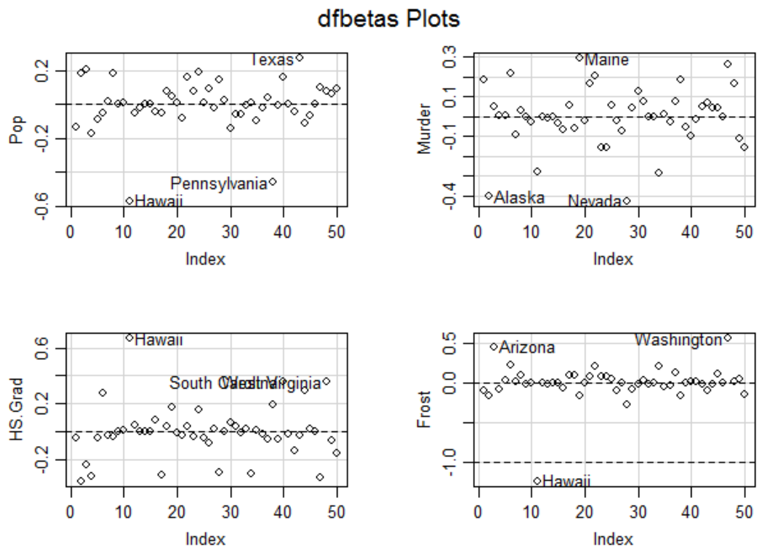
$$D_i = \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}},$$

$$h_{ii} = \boldsymbol{x_i}^T\ (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x_i}$$



DFBETA

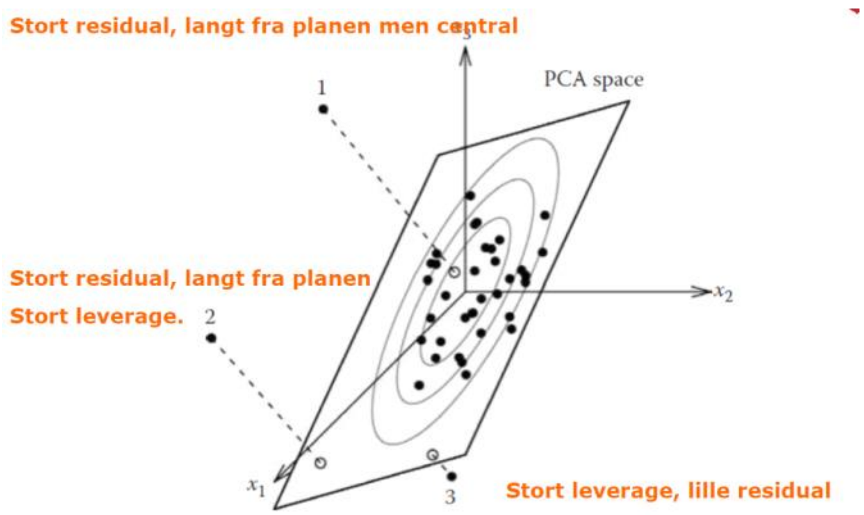
Cooks afstand kan spaltes ud på koordinater som måler hvor meget det ændrer sig hvis man udelader den enkelte måling



Leverage (ikke pensum)

Leverage er et mål for hvor langt en observations forklarende variable ligger fra gennemsnittet af de forklarende variable

(ligger fra gennemsnittet af de forklarende variable



Total test

Undersøgelse om de foraklrende variable bidrager til beskrivelsen af responsen

Ofte nulhypotese test

Alle hældningerne er lig med 0

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$
$$H_1 : Mindst\ et\ \beta \neq 0$$

Teststørrelse

$$F = \frac{R^2/p}{(1-R^2)/n-(p+1)} = \frac{(\sum(\hat{y}_i - \bar{y})^2)/p}{(\sum(y_i - \hat{y}_i)^2)/(n-p-1)}$$

F-fordelt F(p, n-p-1)

Modelreduktion - kvadratsummer

- I den generelle lineære model bruges variationsopspaltningen:

$$SS_{total} = SS_{model} + SS_{residual}$$

Modelkvadratsum $SS_{model} = \sum_i (\hat{Y}_i - \bar{Y})^2$
- Forklaret variation:
- Hvor meget varierer de prædikterede værdier?
- (stort er godt)

Residualkvadratsum $S_{residual} = \sum_i (Y_i - \hat{Y}_i)^2$
- Tilbageblevet variation: Hvor store er modelafvigelserne?
- (småt er godt)

Prædiktion

Principel komponent analyse

25. maj 2020 20:41

PCA

At finde hensigtsmæssige linearkombinationer af de oprindelige variable og reducerer antallet af forklarende variable til et mindre antal principal komponenter.
Formålet er at forklare store dele af variationen.
De originale variable kan være korrelerede så derfor anvender man PCA

Den første principel komboent skal være den mest forklarende variabel.
Den næste skal så være den næst mest foraklarende (her skal den være ortogonal)

$$t_1 = p_{11}X_1 + p_{21}X_2 + p_{31}X_3$$

$$t_2 = p_{12}X_1 + p_{22}X_2 + p_{32}X_3$$

T er scores og p er loadings
De er ukorreleret

Husk

Man skal helst have observationer antal som antal variable.
Man kan sagtens tilføje flere og flere principale komponenter
Summen af de oprindelige variables varians er lig summen af de principale komponenters varians
Man arbejde i matricer

Modellen

Man bestemmer de principale komponenter T som en linearkombination udfra de originale X variable.

$$X = \underbrace{TP^T}_{Struktur} + \underbrace{E}_{Støj}$$

$$\underset{n \times p}{X} = \underset{n \times A}{T} \; \underset{A \times p}{P'} + \underset{n \times p}{E}$$

X variabel, T er scores og P er loadings

Korrelation og kovarians matrice

Analysen kan laves på kovarians matricen eller korrelations matricen (standariserer det oprindelige variable)
fx

	A1	A2	A3	A4	A5	A6
A1	1.00	0.94	0.84	0.15	-0.29	-0.35
A2	0.94	1.00	0.94	0.42	0.00	-0.09
A3	0.84	0.94	1.00	0.60	0.23	0.16
A4	0.15	0.42	0.60	1.00	0.86	0.77
A5	-0.29	0.00	0.23	0.86	1.00	0.96
A6	-0.35	-0.09	0.16	0.77	0.96	1.00

A er erstattes med X i stedet. Her kan det ses, at de markeret gule er kolleret hvilket ikke er godt. Derfor skal man beregne PC

Egenværdi

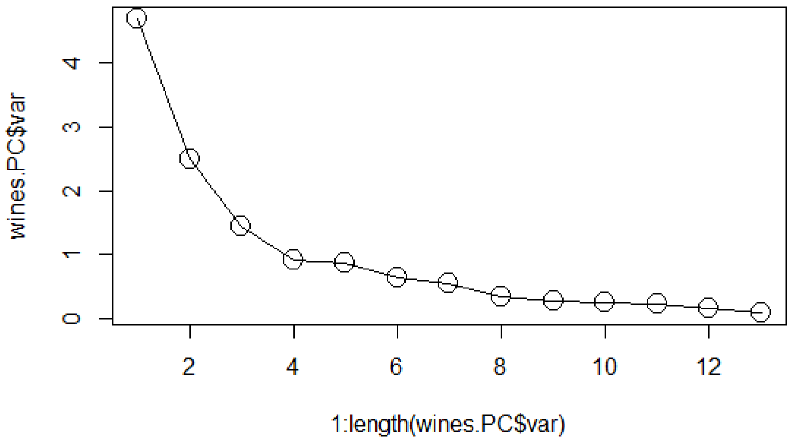
Dens størrelse udtrykker hvor meget af variationen af den PC forklarer. Bruges til at udregne PC

Scree plots

Scree plottet viser, hvor stor en andel af variansen den enkelte principal komponent forklarer

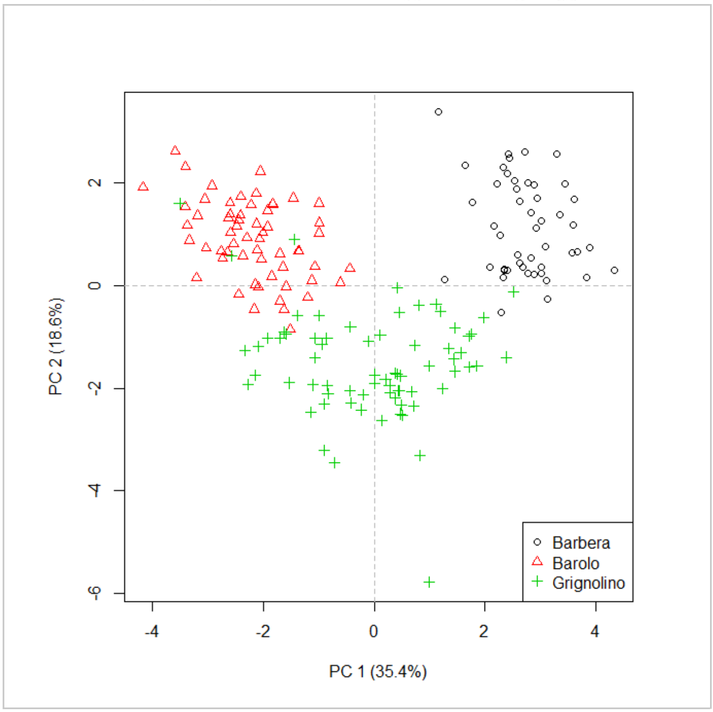
Man vælger antal PC, der hvor der er en "albue". Inden "albuen" forklarer hver PC hver en del af variationen, efter er det er der ikke flere klare retninger i data

Fx



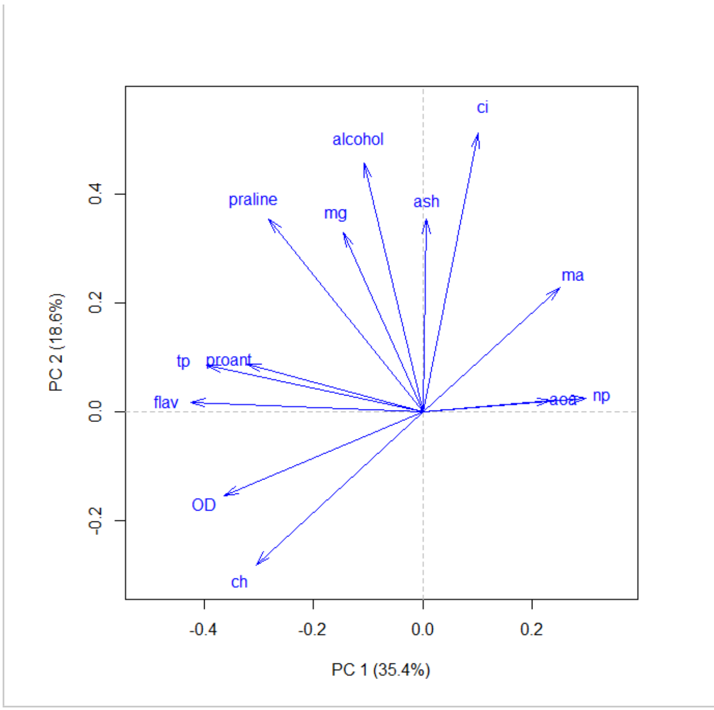
Score plot

Score plot af to scorevektorer (vores "nye" observationer) mod hinanden viser, hvordan observationerne er relateret til hinanden. Ofte kan man ved hjælp af disse plot identificerer hvilke komponenter, der kan diskriminere/ klassificere/ kategorisere mellem typer eller grupper af observationer.



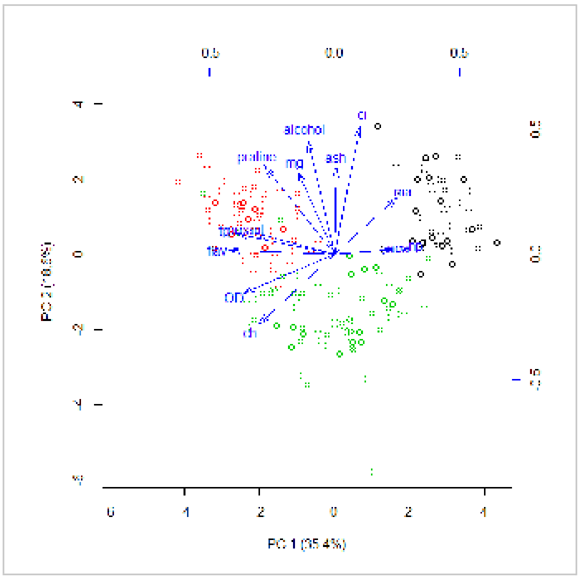
Loading plot

Loading plot af to loadingvektorer mod hinanden viser, hvordan variablene er relateret til hinanden. Nogle variable hører mere til på en PC end på en anden. Desuden bruges loading plot ofte til at fortolke principal komponenterne – at give den et navn.



Biplot

Biplot er score og loading plots på samme plot.



Principel komponent regression

26. maj 2020 12:46

Noter

Noter