

Basal Statistik

Regressionsanalyse i R.

Lene Theil Skovgaard

2. marts 2020

1 / 85



Simpel lineær regression

Regression og korrelation

- ▶ Simpel lineær regression
- ▶ Todimensionale normalfordelinger
- ▶ Korrelation vs. regression
- ▶ Modelkontrol
- ▶ Diagnostics

Home pages:

<http://publicifsv.sund.ku.dk/~sr/BasicStatistics>

E-mail: ltsk@sund.ku.dk

*: Siden er lidt teknisk

2 / 85



Simpel lineær regression

Retningsbestemt relation

(men *ikke* nødvendigvis *kausal*)

mellem to kvantitative (kontinuerte) variable:

Y: **Respons** eller **outcome**, afhængig (dependent) variabel

X: **Forklarende variabel**, kovariat

(somme tider *Independent*/uafhængig - *meget uheldigt!*)

Gennemgående eksempel at tænke på:

Y =blodtryk, X =alder

3 / 85



Data

Sammenhørende registreringer (x_i, y_i) ,

for en række individer eller 'units', $i = 1, \dots, n$:

Bemærk: x_i 'erne kan **vælges på forhånd!**

- ▶ Det er **smart**,
fordi man kan designe sig til mere præcise estimater
- ▶ Det er **farligt**,
hvis man har tænkt sig at benytte korrelationer
(mere om det senere)

4 / 85



Eksempel I

Sammenhæng mellem kolinesteraseaktivitet (KE) og tid til opvågning (TID)

- ▶ **Outcome:** TID
- ▶ **Forklarende variabel:** KE
- ▶ **Konklusioner:**
Hvor lang tid forventer vi til opvågning, baseret på en måling af KE?
Hvor stor er usikkerheden på denne prediktion?

Et outcome, en (kvantitativ) kovariat:
Simpel lineær regression

5 / 85



Eksempel II

Sammenligning af lungekapacitet (FEV_1) for rygere og ikke-rygere

Problem: FEV_1 afhænger også af f.eks. højde

- ▶ **Outcome:** FEV_1
- ▶ **Forklarende variable:** højde, rygevaner
- ▶ **Konklusioner:**
Hvor meget dårligere er lungefunktionen hos rygere?

Et outcome, to kovariater:
Her er der tale om *multipl regression*, i form af en *kovariansanalyse* (omtales i næste forelæsning)

6 / 85



Eksempel fra bogen

Hvilken sammenhæng er der mellem **fastende blodsukkerniveau** og **sammentrækningsevne** for venstre hjertekammer hos diabetikere? (n=23)

Outcome: $Y = vcf$, %/sec.

Kovariat:

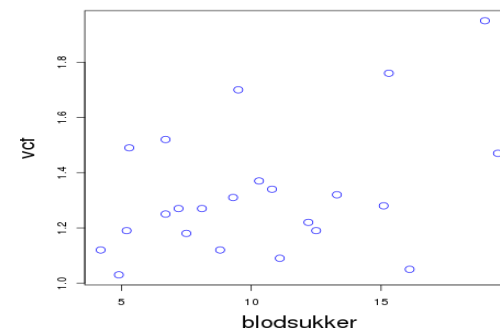
$X = \text{blodsukker}$, mmol/l

	blodsukker	vcf
1	15.3	1.76
2	10.8	1.34
3	8.1	1.27
...		
14	15.1	1.28
15	6.7	1.52
16	8.6	NA
17	4.2	1.12
...		
22	4.9	1.03
23	8.8	1.12
24	9.5	1.70

7 / 85



Scatter plot



```
plot(hjerte$vcf, hjerte$blodsukker,
     ylab="vcf", xlab="blodsukker",
     cex.lab=1.8, cex=1.5, col="blue")
```

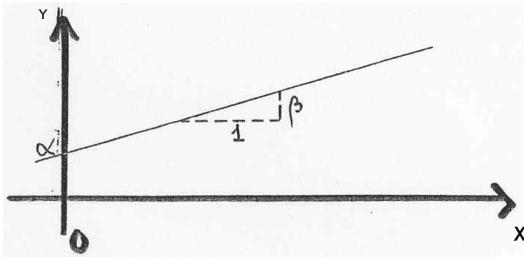
Er der nogen sammenhæng her?

8 / 85



Matematisk model

Ligningen for en ret linie: $Y = \alpha + \beta X$



Intercept α : Skæring med Y-akse
Hældning β

9 / 85



Fortolkning

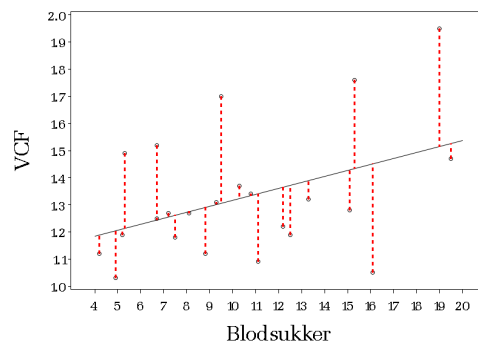
- ▶ α : **intercept**, afskæring (skæring med Y-akse)
Sammentrækningsevnen for en diabetiker med en blodsukkerværdi på 0.
Samme enheder som outcome.
Som regel en utilladelig ekstrapolation!
- ▶ β : **hældning**, regressionskoefficient
Forskellen i sammentrækningsevne hos 2 diabetikere, der afviger i blodsukkerværdi med 1 mmol/l.
Ofte parameteren med størst interesse.
Enheder som “Outcomes enheder” pr. “kovariats enhed”.

10 / 85



Statistisk model

$Y_i = \alpha + \beta X_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ uafh.
 ε_i 'erne er **residualerne**,
dvs. afvigelserne (i Y) fra observeret til forventet.



(ingen kode til denne figur)

11 / 85



* Estimation

Mindste kvadraters metode:

Bestem α og β , så kvadratafgivelsessummen

$$\sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2 = \sum_{i=1}^n \varepsilon_i^2$$

bliver *mindst mulig*. Resultat (**estimer**):

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

12 / 85



Regressionsanalyse i praksis

```
lm(vcf ~ blodsukker, na.action=na.exclude, data=hjerte)
```

Det er dog smartere at kalde modellen noget, så man kan bruge diverse indbyggede funktioner på den efterfølgende.

```
model1 = lm(vcf ~ blodsukker, data=hjerte)
```

På næste side har vi derefter benyttet

```
summary(model1)
confint(model1)
```

13 / 85



Output fra regressionsanalyse i R

```
>summary(model1)
```

Call:

```
lm(formula = vcf ~ blodsukker, data = hjerte)
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.09781    0.11748   9.345 6.26e-09 ***
blodsukker    0.02196    0.01045   2.101  0.0479 *
---

```

Residual standard error: 0.2167 on 21 degrees of freedom

(1 observation deleted due to missingness)

Multiple R-squared: 0.1737, Adjusted R-squared: 0.1343

F-statistic: 4.414 on 1 and 21 DF, p-value: 0.0479

```
> confint(model1)
```

```
              2.5 %      97.5 %
(Intercept) 0.8534993816 1.34213037
blodsukker   0.0002231077 0.04370194

```

14 / 85



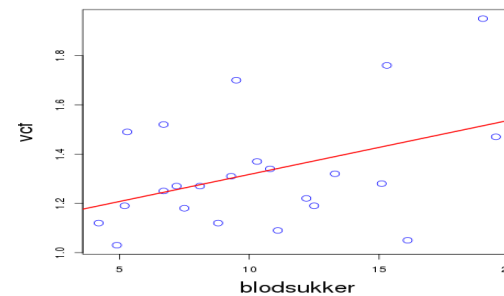
Vigtige informationer fra output

- ▶ **Hældning** (slope, vises i output under betegnelsen 'blodsukker', fordi det er koefficienten til denne),
 $\hat{\beta} = 0.02196 \approx 0.022$,
 med tilhørende spredning (**standard error**)
- ▶ **Spredningen omkring linien** (Residual standard error),
 $s = \hat{\sigma} = 0.2167 \approx 0.217$
 Denne størrelse benyttes til konstruktion af **prediktionsgrænser** (kommer senere), som er **normalområder for given blodsukkerværdi**.

15 / 85



Estimeret regressionslinje



```
plot(hjerte$blodsukker, hjerte$vcf,
     ylab="vcf", xlab="blodsukker",
     cex.lab=1.8, cex=1.5, col="blue")
```

```
abline(model1, col="red", lwd=2)
```

Estimeret regressionslinje: $vcf = 1.10 + 0.0220 \text{ blodsukker}$

16 / 85



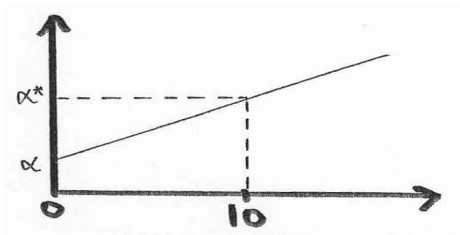
Forventet værdi for specifikke værdier af kovariaten

Fittede (predikterede, forventede) værdier:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$$

Forventet værdi af vcf for en diabetiker med blodsukker 10mmol/l:

$$1.10 + 0.0220 \times 10 = 1.32$$



men hvad med **usikkerheden (s.e.)** for denne?

17 / 85



Forventet værdi for specifikke værdier af kovariaten, II

Usikkerheder på disse kan *ikke* udregnes ved at kombinere s.e. på hhv. **intercept og hældning!**

Dette skyldes, at intercept og hældning er (negativt) korrelerede: Højt intercept giver lav hældning - og omvendt.

Lad R gøre det ved at skrive:

```
nyedata<-data.frame(blodsukker=10)
predict(model1, nyedata, interval = "confidence")
```

som giver outputtet:

	fit	lwr	upr
1	1.31744	1.223124	1.411757

18 / 85



Estimaterne fra regressionsanalysen

Regressionsanalysen indeholder **3 parametre**:

- ▶ 2 hørende til linien (intercept og hældning)
- ▶ 1, som er **spredningen omkring linien** (σ): den biologiske variation af vcf for folk med samme blodsukker-værdi

Varansen omkring regressionslinien (σ^2) estimeres ved

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

ca. gennemsnitlig kvadratisk afstand, blot er n (antallet af observationer) erstattet af $n-2$ (antallet af frihedsgrader), her 21

19 / 85



Spredning omkring regressionslinie

Da man ikke kan *forstå* eller *fortolke* varianser direkte, tager vi straks kvadratrods:

$$s = \sqrt{s^2}$$

som er **estimat** for **spredningen omkring regressionslinien**

som benævnes

- ▶ i SAS (lidt uheldigt *): 'Root Mean Square Error', forkortet Root MSE i output.
- ▶ i R (ligeså uheldigt *): 'Residual Standard Error'

Estimatet er 0.2167, med de **samme enheder** som outcome vcf

* uheldigt, fordi det *ikke* er en standard error, men en standard deviation

20 / 85



Usikkerhed på estimererne

Hvor gode er skønnene over de ukendte parametre α og β ?

Hvor meget anderledes resultater kunne vi forvente at finde ved en ny undersøgelse?

Det kan vises, at

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$$

dvs. **hældningen er præcist bestemt, hvis**

- ▶ observationerne ligger tæt på linien (σ^2 lille)
- ▶ variationen i x -værdier (S_{xx}) er stor

21 / 85



Konfidensinterval = Sikkerhedsinterval

Estimeret usikkerhed på $\hat{\beta}$: $\text{s.e.}(\hat{\beta}) = \frac{s}{\sqrt{S_{xx}}}$

Dette estimat kaldes **standard error** for $\hat{\beta}$,
eller generelt **standard error of the estimate (s.e.)**

Vi bruger det til at konstruere et **95% konfidensinterval**

$$\hat{\beta} \pm t_{97.5\%}(n-2) \times \text{s.e.}(\hat{\beta}) = \hat{\beta} \pm \text{ca.} 2 \times \text{s.e.}(\hat{\beta})$$

$$= 0.0220 \pm 2.080 \times 0.0105 = (0.0002, 0.0438)$$

22 / 85



T-test for "parameter=0"

Vi kan også teste, typisk ' $H_0 : \beta = 0$ ' ved **t-testet**

$$t = \frac{\hat{\beta}}{\text{s.e.}(\hat{\beta})} \sim t(n-2)$$

som her giver

$$t = \frac{0.0220}{0.0105} = 2.10 \sim t(21), \quad P=0.048$$

dvs. lige på grænsen af det signifikante

23 / 85



Og hvad med interceptet....?

Man *kan* forestille sig situationer, hvor det er rimeligt at teste
f.eks. ' $H_0 : \alpha = \alpha_0$ '

I så fald benyttes det tilsvarende t-test

$$t = \frac{\hat{\alpha} - \alpha_0}{\text{s.e.}(\hat{\alpha})} \sim t(n-2)$$

eller man udregner et 95% konfidensinterval for α :

$$1.098 \pm 2.080 \times 0.1175 = (0.854, 1.342)$$

Dette er bare ikke altid særligt interessant
– fordi interceptet i sig selv ikke er særligt interessant.

24 / 85

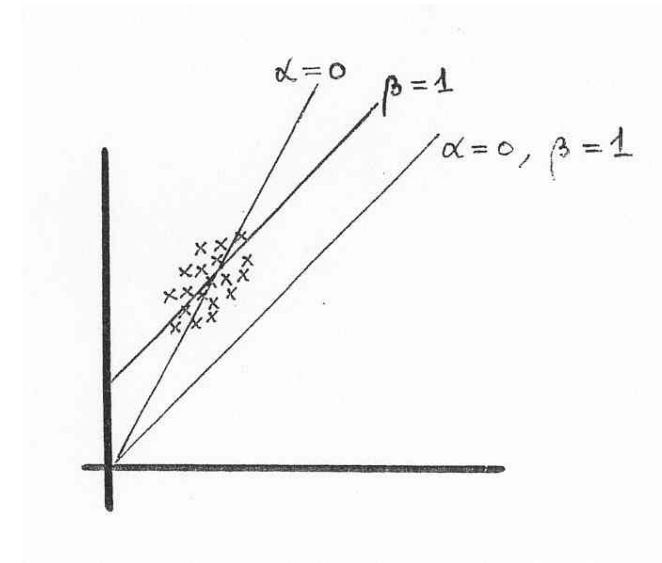


Når man har flere hypoteser...

Vi kan altså teste hypoteser om *såvel* α som β , **men**:

- ▶ Estimerne for intercept og hældning er (negativt) korrelerede (her -0.92)
- ▶ Accepter ikke to 'sideordnede' test
Selv om vi kan acceptere test vedr. både α (f.eks. intercept=0) og β (f.eks. hældning 1)
hver for sig,
kan vi **ikke nødvendigvis acceptere begge samtidig**

Hypotetisk eksempel



25 / 85

26 / 85

Variationsopspaltning

Variationsanalyse for model1 fra s. 13:

```
> anova(model1)
Analysis of Variance Table

Response: vcf
      Df Sum Sq Mean Sq F value Pr(>F)
blodsukker  1  0.20727  0.207269   4.414  0.0479 *
Residuals 21  0.98610  0.046957
```

kan udtrykkes ved

$$SS_{\text{total}} = \sum_{i=1}^n (y_i - \bar{y})^2 = SS_{\text{blodsukker}} + SS_{\text{Residuals}}$$

Total variation = variation, som **kan** forklares
+ variation, som **ikke kan** forklares

x er en **god** forklarende variable,

hvis $SS_{\text{blodsukker}}$ er *stor* i forhold til SS_{Residual}

27 / 85

Determinationskoefficient, R^2

Den andel af variationen (i y), der kan forklares ved modellen:

$$R^2 = \frac{SS_{\text{model}}}{SS_{\text{total}}}$$

Her finder vi determinationskoefficienten: $R^2 = 0.17$ (se s. 14),
dvs. vi kan forklare 17% af variationen i vcf v.h.j.a. variabelen
blodsukker.

Determinationskoefficienten er
kvadratet på **korrelationskoefficienten**

mellem vcf og blodsukker (som dermed er $r = \sqrt{0.17} = 0.42$)

28 / 85

Korrelationen

Korrelationen r mellem to variable måler

- I hvor høj grad ligner scatter plottet en ret linie?
- **Ikke:** Hvor nær ligger punkterne ved den rette linie?

Korrelationskoefficienten estimeres ved:

$$r = r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

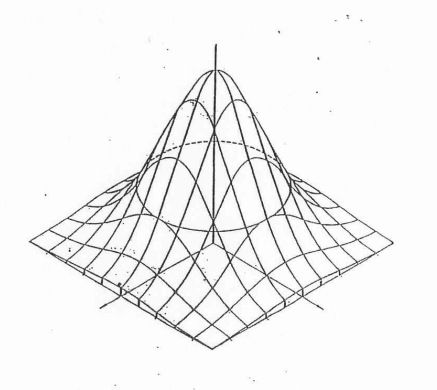
- antager værdier mellem -1 og 1 (0 = uafhængighed)
- +1 og -1 svarer til perfekt lineær sammenhæng, hhv. positiv og negativ

29 / 85



Todimensional normalfordelingstæthed, $r=0$

Korrelation 0



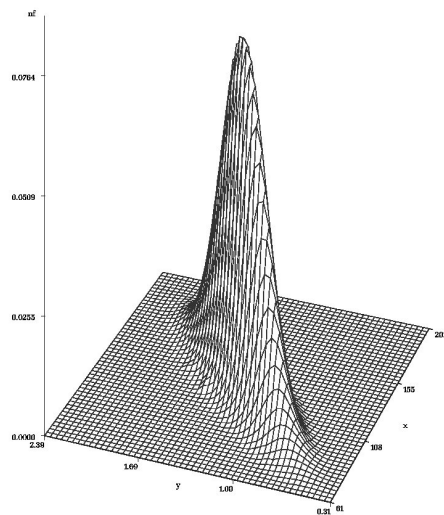
Alle lodrette snit giver normalfordelinger

- med samme middelværdi
- og samme varians

30 / 85



Todimensional normalfordelingstæthed, $r=0.9$



Korrelation 0.9

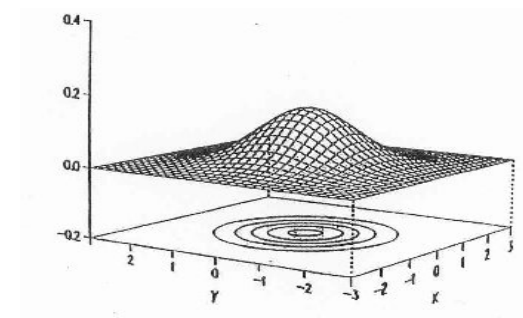
Udpræget **retning** i figuren

31 / 85



Konturkurverne

for en normalfordeling bliver **ellipser**



så check af todimensional normalfordeling kan foretages som visuelt check af scatterplottet:

- Giver det indtryk af noget elliptisk?
- med flest observationer i de "inderste" ellipser

32 / 85

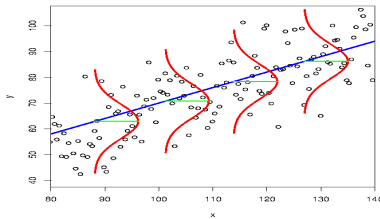


Regression kontra korrelation

Regressionsmodellen

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

siger, at de **betingede fordelinger** af Y , givet X , er normalfordelinger, med samme varians σ^2 (samme spredning σ) og med middelværdier, der afhænger lineært af X .



33 / 85



Regression kontra korrelation, II

Antagelserne til brug for fortolkning af en korrelation er **skrappere** end for regressionsanalysen, fordi:

Her er der også et krav om **normalfordeling på x_i 'erne!!**
Hvis ikke dette krav er opfyldt, kan man ikke fortolke korrelationskoefficienten!

...men man kan godt teste, om den er 0.

34 / 85



Test af hældning vs test af korrelation

Det er en og samme sag

De to estimater (for korrelation og hældning) er 0 på samme tid

$$r_{xy} = \hat{\beta} \sqrt{\frac{S_{xx}}{S_{yy}}}$$

Test for $\beta = 0$ er **identisk** med test for $\rho_{xy} = 0$

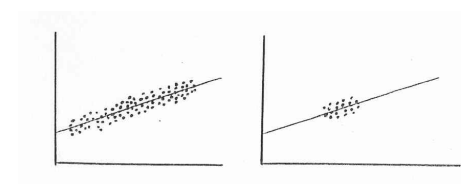
men **fortolkningen** af de to størrelser er vidt forskellig:

- ▶ $\hat{\beta}$ fortolkes i substans-termer, med forståelige enheder
- ▶ r_{xy} er skala-uafhængig - og meget vanskelig at tillægge nogen virkelig mening

35 / 85



Pas på med fortolkning af korrelation



$$1 - r_{xy}^2 = \frac{s^2}{s^2 + \hat{\beta}^2 \frac{S_{xx}}{n-2}}$$

Hold $\hat{\beta}$ og s^2 fast:

S_{xx} stor $\Rightarrow 1 - r_{xy}^2$ tæt på 0 $\Rightarrow r_{xy}^2$ tæt på 1

r_{xy}^2 kan gøres **vilkårlig tæt på 1** ved at sprede x 'erne
– hvordan mon?

36 / 85



Forskellige typer af korrelationer

Pearson: Det er den type, vi netop har beskrevet, som altså bygger på en antagelse om tilfældig udvælgelse fra en todimensional normalfordeling

Spearman: Et **non-parametrisk** alternativ, som *kun* kræver tilfældig udvælgelse fra en todimensional *fordeling* (der altså ikke behøver at være en normalfordeling)

I tilfælde af et selekteret sample bliver begge korrelationer meningsløse

– og under alle omstændigheder giver de bare et ret ufortolkeligt tal.....

37 / 85



Korrelationer i praksis

I R skrives

```
> with(hjerte, cor.test(blodsukker, vcf))
```

der giver **Pearson korrelationen** som default:

Pearson's product-moment correlation

```
data: blodsukker and vcf
t = 2.101, df = 21, p-value = 0.0479
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.005496682 0.707429479
sample estimates:
      cor
0.4167546
```

Bemærk, at P-værdien for Pearson-korrelationen er nøjagtig den samme som ved test af hældning=0 i regressionsanalysen.

Dette vil altid være tilfældet

38 / 85



Korrelationer i praksis, II

Spearman korrelationen skal specificeres som option:

```
> with(hjerte, cor.test(blodsukker, vcf, method="spearman"))
```

Spearman's rank correlation rho

```
data: blodsukker and vcf
S = 1380.4, p-value = 0.1392
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.318002
```

Korrelationerne er altså:

Pearson: 0.42, P=0.048

Spearman: 0.32, P=0.14

39 / 85



Hvornår bruger vi så korrelationen

- ▶ Til at teste om der er en sammenhæng
Brug gerne Spearman korrelation og slip for så mange antagelser, **men så får man også kun en P-værdi**
- ▶ Til at sammenligne styrken af sammenhænge mellem mange variable, målt på de samme individer
- ▶ I observationelle studier uden selektion:
Her kan korrelationen fortolkes, hvis der er tale om en **todimensional normalfordeling** men spørgsmålet er stadig, om det giver den information, der er ønskelig/brugbar

40 / 85



Eksempler

Spørgsmål: Hvor meget stiger blodtrykket med alderen?

Svar: Korrelationen er 0.42.....

hellere: 5mmHg per år, med CI=(3.7, 6.3)

Spørgsmål: Er rygning mere skadeligt for kvinder end for mænd?

Svar: Næh, for korrelationen mellem pakkeår og FEV₁ er 0.62 for mænd og kun 0.49 for kvinder....

Pas på: Dette kan skyldes en større spredning på variabelen rygning blandt mænd, og behøver altså *ikke* at have noget at gøre med *effekten* af rygning

41 / 85



Sammenligning af målemetoder

I en sådan situation giver det **ingen mening** at benytte korrelation!

- ▶ Korrelationen udtrykker **sammenhæng**, *ikke* overensstemmelse (der er f.eks. en sammenhæng mellem alder og blodtryk, men der er naturligvis ikke overensstemmelse)
- ▶ Naturligvis er der sammenhæng mellem to målemetoder, der foregiver at måle det samme, så det behøver man ikke teste (ellers er der i hvert fald noget helt galt)
- ▶ Man skal i stedet **kvantificere differenserne** mellem de to metoder, og udregne **limits of agreement** (på relevant skala)

– og nu tilbage til regressionsanalyse

42 / 85



Konfidensgrænser for linien

For hver værdi af kovariaten (her x =blodsukker):

Fittede (predikterede, forventede) værdier er selve linien:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$$

Konfidensgrænser for denne linie (smalle grænser)

- ▶ benyttes til sammenligning med andre grupper af personer
- ▶ man benytter spredningen $s_{\text{konf}} = s \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}}$
- ▶ Disse grænser bliver **vilkårligt snævre**, når antallet af observationer øges.
- ▶ **De kan ikke bruges til diagnostik!**

43 / 85



Prediktionsgrænser for enkeltindivider

Normalområder for sammentrækningsevne (y), for givet

x =blodsukker

(brede grænser):

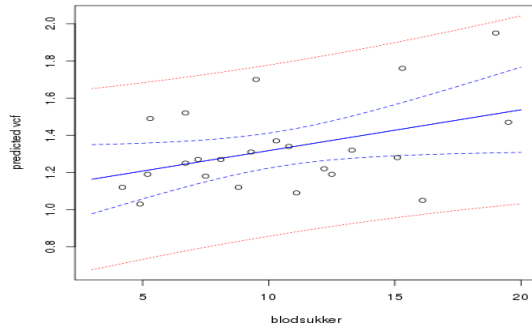
- ▶ De benyttes til at afgøre, om en ny person er *atypisk* i forhold til normen (diagnostik), idet de omslutter ca. 95% af fremtidige observationer, også for store n .
- ▶ man benytter spredningen $s_{\text{pred}} = s \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}}$
- ▶ Disse grænser bliver **ikke nævneværdigt snævrere**, når antallet af observationer øges.

44 / 85



Konfidens- og prediktionsgrænser i praksis

Se kode s. 81



45 / 85



Summa summarum om grænser

De smalle grænser, **konfidensgrænser**:

- ▶ svarer til standard error
- ▶ bruges til at vurdere sikkerheden i estimatet
- ▶ afhænger kraftigt af værdien af kovariaten

De brede grænser, **prediktionsgrænser**:

- ▶ svarer til standard deviation
- ▶ kaldes også normalområder eller referenceområder
- ▶ bruges til at diagnosticere individuelle patienter
- ▶ udregnes approksimativt ved ± 2 spredninger (Residual standard error)

46 / 85



Har vi gjort det godt nok?

Modellens konklusioner er kun rimelige, hvis modellen selv er rimelig.

Modelkontrol: Passer modellen rimeligt til data?

Diagnostics: Passer data til modellen?
Eller er der **indflydelsesrige observationer** eller **outliers**?

Check af disse to forhold *burde* naturligvis foretages fra begyndelsen, men da de kræver fit af modellen, *kan* de først foretages efterfølgende.

47 / 85



Modelkontrol

Den statistiske model var

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \text{ uafhængige}$$

Hvilke antagelser skal vi checke her?

- ▶ Uafhængighed: **Tænk:** Er der flere observationer på hvert individ, søskende el.lign?
- ▶ Linearitet
- ▶ Varianshomogenitet (ε_i 'erne har samme spredning)
- ▶ Normalfordelte residualer (ε_i 'erne)

Obs: Intet krav om normalfordeling på x_i 'erne!!

48 / 85



Grafisk modelkontrol

Fokus er her på

residualerne = modelafvigelse

= **observeret** værdi - **fittet** værdi: $\hat{\varepsilon}_i = y_i - \hat{y}_i$

eller en modifikation af disse, der fås som alternative funktioner i R

- ▶ Sædvanlige residualer: `resid(model1)`
- ▶ Normerede/Standardiserede/Studentized: `rstandard(model1)`
- ▶ Studentized Press: `rstudent(model1)`

49 / 85



Residualplots til modelkontrol

Residualer (af passende type) plottes mod

1. den forklarende variabel x_i
 - for at checke **linearitet** (se efter *krumninger, buer*)
2. de fittede værdier $\hat{y}_i$
 - for at checke **varianshomogenitet** (se efter *trompeter*)
3. fraktildiagram eller histogram
 - for at checke **normalfordelingsantagelsen** (se efter *afvigelse fra en ret linie*)

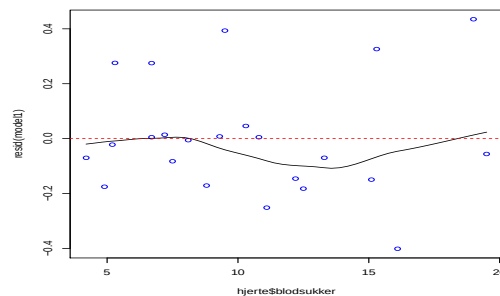
Disse plots ses på s. 51 og 52, og kommenteres s. 51 og 53

50 / 85



Residualplots i praksis

Check af lineariteten: Residualer plottet mod kovariaten, med overlejret udglætning (kode s. 82):

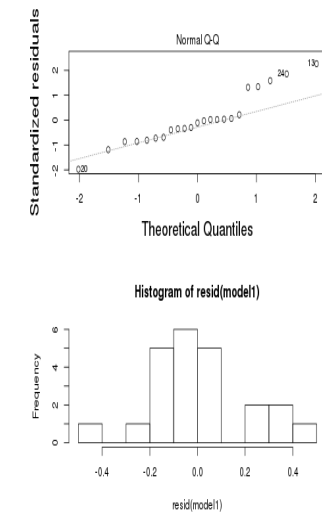
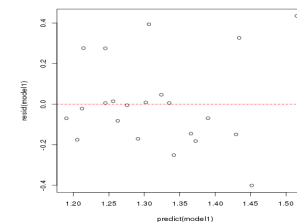


Her ses ingen oplagte (store) buer

51 / 85



Diagnostics Panel - kode s. 82



52 / 85



Kommentarer til de diagnostiske plots s. 52

Til venstre: Plot af residualer mod predikterede værdier, dvs. plot af type 2 fra s. 50:

Her anes muligvis lidt trompetfacon

Øverst til højre: Fraktildiagram til check af normalfordelte residualer.

Det ser rimeligt ud, men der anes ligesom et *hul* i fordelingen (type 3 fra s. 50)

Nederst til højre: Histogram over residualerne, ligeledes til check af normalfordelingen. (type 3 fra s. 50)

Det *hul*, der nævnes ovenfor ses som minimummet lige over midten af fordelingen

Man kan med fordel supplere med plots på siderne 51 og 54

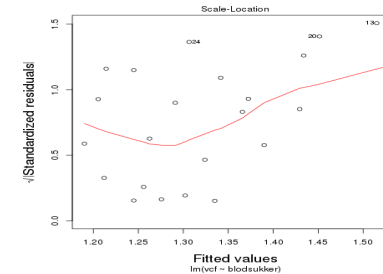
53 / 85



Bedre check af varianshomogeniteten

Plot af kvadratrods af numerisk værdi af normerede residualer, mod de predikterede værdier:

```
plot(model1, which=3, cex.lab=1.5)
```



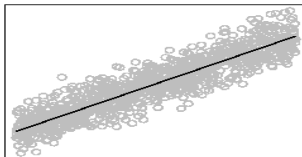
Der er en tendens til højere spredning for de høje predikterede værdier. **Måske konstant relativ spredning?**

54 / 85

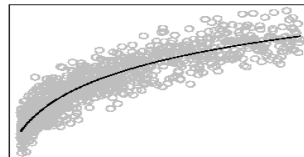


Afvigelser fra modellen

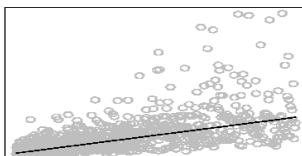
assumptions OK



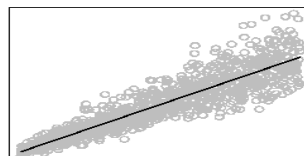
non-linearity



non-normality



increasing variance



55 / 85



Afhjælpning af problemer

Linearitet:

Hvis lineariteten ikke holder i rimelig grad, bliver modellen ufortolkelig.

Hvad gør man så?

- ▶ transformerer variablene med
 - ▶ **logaritmer** - ligegyldigt hvilken, se s. 61-63
 - ▶ kvadratrods, invers
 - ▶ benytter lineære splines ("knæk-linier", kommer senere)
- ▶ tilføjer flere kovariater, f.eks.
 - ▶ alder, køn, medicin etc., eller $\log(x)$
- ▶ foretager **ikke-lineær regression**

56 / 85



Afhjælpning af problemer, II

Varianshomogenitet:

Hvis varianshomogeniteten ikke holder i rimelig grad, mister vi styrke, og **prediktionsgrænser bliver upålidelige!**

Hvad gør man så?

- ▶ I tilfælde af **trompetfacon**: Transformation af Y med **logaritmer**
- ▶ Vægtet regression...
- ▶ (Non-parametriske metoder... – men så får vi ingen kvantificeringer)
- ▶ Robuste metoder: (r1m i pakken MASS)

57 / 85



Varianshomogenitet, fortsat

Trompetfacon betyder, at residualernes variation (og dermed størrelse) er større for højere niveauer af outcome - ofte i form af

Konstant relativ spredning

= **konstant variationskoefficient**

Variationskoefficient = $\frac{\text{spredning}}{\text{middelværdi}}$

Dette sker ofte, når man måler på små positive størrelser, og løsningen er (*sædvanligvis*) at transformere outcome Y med en **logaritme**

58 / 85



Afhjælpning af problemer, III

Normalfordelte residualer:

Hvis normalfordelingen ikke holder i rimelig grad, mister vi styrke, og **prediktionsgrænser bliver upålidelige!**

Hvad gør man så?

- ▶ I tilfælde af **hale mod højre**: Transformation med **en logaritme**
- ▶ Non-parametriske metoder...

59 / 85



Normalfordeling (og varianshomogenitet)

Antagelsen om normalfordeling (og til en vis grad varianshomogenitet) er **ikke kritisk** for selve fittet:

- ▶ Man får stadig "gode" estimater
- ▶ Pålidelige tests og konfidensintervaller

fordi normalfordelingen som regel passer godt for estimatet $\hat{\beta}$ pga **Den centrale grænseværdisætning**, der siger at summer og andre funktioner af *mange* observationer bliver 'mere og mere' normalfordelt.

Men prediktionsgrænserne bliver misvisende og ufortolkelige!!

60 / 85



Lidt om logaritmer – måske genopfriskning...?

Alle logaritmefunktioner har et **grundtal**

- ▶ 10-tals logaritmen ($\log_{10}$) har grundtal 10
- ▶ Den såkaldt *naturlige* logaritme ($\log$, før i tiden ofte kaldet $\ln$) har grundtal $e = 2.71828$
- ▶ 2-tals logaritmen ($\log_2$) har grundtal 2

Alle logaritmer er proportionale, f.eks.

$$\log_2(x) = \frac{\log(x)}{\log(2)} = \frac{\log_{10}(x)}{\log_{10}(2)}$$

og det betyder derfor intet for resultatet, hvilken en, man benytter fordi man altid får det samme, når man tilbagetransformerer.

Alligevel er der visse *fif*....:

61 / 85



Logaritmetransformation af outcome

- ▶ for at opnå linearitet
- ▶ for at opnå ens spredninger (varianshomogenitet)
- ▶ for at opnå normalitet af residualerne

Her er kun et enkelt (ret ubetydeligt) fif:

Hvis fokus er på estimation af variationskoefficienten (se s. 58), kan man med fordel benytte den **naturlige** logaritme), idet

$$\text{Spredning}(\log(y)) \approx \frac{\text{Spredning}(y)}{y} = \text{CV}$$

dvs. en konstant variationskoefficient (CV) på Y betyder konstant spredning på $\log(Y)$. Når Y **logaritmetransformerer**, bliver effekten af kovariater (når de tilbagetransformerer) til **faktorer**, der skal ganges på.

62 / 85



Logaritmetransformation af kovariat

Den forklarende variabel (x) transformeres altid for at opnå linearitet.

Her kan med fordel anvendes **2-tals logaritmer**, idet 1 enhed i $\log_2(x)$ så svarer til en *fordobling af* x , der har konstant effekt.

Hældningen β udtrykker altså effekten af en fordobling af kovariaten, og successive fordoblinger antages at have samme effekt.

Man kan også gøre det *endnu mere fortolkeligt* ved at benytte logaritmen med grundtal f.eks. 1.1, svarende til en faktor 1.1, altså en 10% forøgelse af kovariat-værdien.

$$\log_{1.1}(x) = \frac{\log_{10}(x)}{\log_{10}(1.1)}$$

63 / 85



Numeriske check af antagelserne

er mulig, men bør kun bruges som en rettesnor

- ▶ **Linearitet:**
Tilføj f.eks. $\log(x)$ sammen med x selv, og test derefter, om det resulterende fit er væsentlig bedre end linien
- ▶ Benyt lineære splines (knæk-linier, kommer senere) og test, om der overhovedet er et knæk
- ▶ **Varianshomogenitet:**
De findes, men i R ved jeg ikke lige...??
- ▶ **Normalfordelingen:**
Der findes test, men de kan ikke generelt anbefales

64 / 85



Regression diagnostics

Understøttes konklusionerne generelt af *hele* materialet?
Eller er der observationer med meget stor indflydelse på resultaterne?

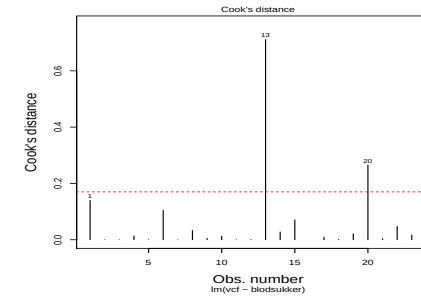
- ▶ Udelad den i^{te} person og bestem nye estimater for samtlige parametre.
- ▶ Udregn **Cook's afstand**, et mål for ændringen i parameterestimater.
- ▶ Spalt evt. Cook's afstand ud i separate dele, som måler f.eks.: Hvor mange s.e. ($\hat{\beta}_1$) ændrer $\hat{\beta}_1$ sig, hvis den i^{te} person udelades?

65 / 85



Cooks afstand

```
plot(model1, which=4, cex.lab=1.5)
abline(0.17, 0, col="red", lty=2)
```



En enkelt observation (måske to) skiller sig ud i forhold til de øvrige
Tommelfingerregel: Sammenlign med $\frac{4}{n} = \frac{4}{23} \approx 0.17$ (rød stiplet linie)

66 / 85



Cooks afstand spaltet i komponenter

Benyt f.eks. `dfbetas(model1)`

og få svar på:

Hvor mange s.e. ($\hat{\beta}_1$)'er ændres f.eks. $\hat{\beta}$,
når den i^{te} person udelades?

```
> dfbetas(model1)
      (Intercept)      blodsukker
1  -0.2420842618  0.413388799
2   0.0014389950  0.0004831275
3  -0.0049416085  0.0029873821
4   0.1082483336 -0.1460999486
5   0.0150296323 -0.0103878021
.      .      .
11  0.0059772286 -0.0043452705
12 -0.0344536319  0.0276872374
13 -0.8744415583  1.1972730268
14  0.0981585950 -0.1718078324
15  0.3408595759 -0.2477949502
.      .      .
```

En enkelt observation (nr. 13, tilfældigvis) kan ændre
hældningsestimatet med mere end 1 standard error.
Tommelfingerregel: Sammenlign med $\frac{2}{\sqrt{n}} = \frac{2}{\sqrt{23}} \approx 0.42$

67 / 85



Udeladelse af observation nr. 13

Estimeret linie fra tidligere: $y = 1.098 + 0.022x$

$$\hat{\beta} = 0.02196(0.01045), \quad t = \frac{0.02196}{0.01045} = 2.1, P = 0.048$$

Regressionsanalyse **uden observation nr. 13**

Estimeret linie: $y = 1.189 + 0.011x$

$$\hat{\beta} = 0.01082(0.01029), \quad t = \frac{0.01082}{0.01029} = 1.05, P = 0.31$$

68 / 85



Outliers

Observationer, der *ikke passer* ind i sammenhængen

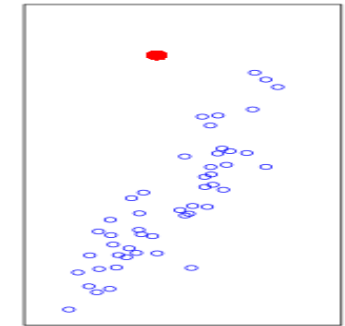
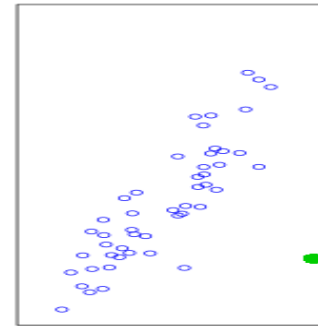
- ▶ de er ikke nødvendigvis indflydelsesrige
- ▶ de har ikke nødvendigvis et stort residual

Press-residualer

Residualer, der fremkommer efter at den pågældende observation har været udelukket fra estimationen, i R tillige normeret til `rstudent(model1)`.

To forskellige eksempler

på mulige *outliers*:



69 / 85



Hvad gør vi i hver af disse situationer?

70 / 85



Kan vi udelade disse observationer?

Lad os forestille os, at figuren på forrige side viser blodtrykket som funktion af alderen, og at den grønne person er **110 år**, mens den røde person er **50 år**

- ▶ **Vi kan godt udelade den grønne person**
faktisk *bør* vi gøre det
for ikke at ødelægge beskrivelsen af det store flertal.
Men husk at skrive det som inklusionskriterium!
- ▶ **Vi kan ikke udelade den røde person**, fordi vi ikke kan konstruere et tilsvarende inklusionskriterium.
Forestil jer sætningen: "*Her beskriver vi blodtrykket for personer, hvis blodtryk ligger pænt i forhold til den beskrivelse, vi er ved at lave....*"

71 / 85



Udeladelse af enkeltobservationer?

Hvad gør vi ved indflydelsesrige observationer og outliers?

- ▶ ser nærmere på dem, de er tit ganske interessante
- ▶ anfører et mål for deres indflydelse

Hvornår kan vi **udelade** dem?

- ▶ hvis de ligger meget *yderligt i kovariat-værdier*
 - ▶ husk at afgrænse konklusionerne tilsvarende!
- ▶ hvis man kan finde årsagen
 - ▶ og da skal **alle sådanne** udelades!

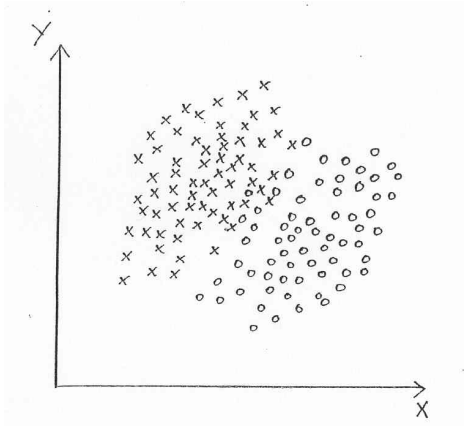
– mere om dette senere (ved øvelserne)

72 / 85



Confounding, bare lige et smuglig

meget mere om dette senere



(Kor)relationen er:

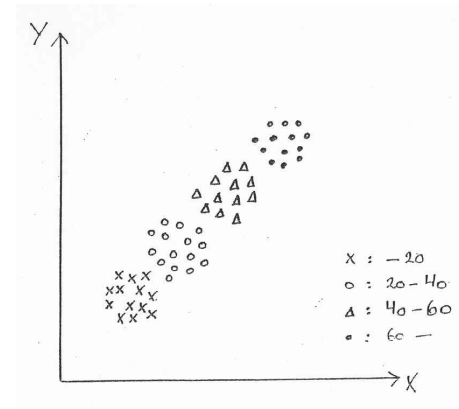
- ▶ positiv for mænd
- ▶ positiv for kvinder
- ▶ negativ for mennesker

Eks: Kolesterol vs. chokoladeindtag

73 / 85



Confounding, II



(Kor)relationen er:

- ▶ tilsyneladende positiv
- ▶ 0 for hver aldersgruppe

X og Y vokser begge med alderen
(f.eks. kolesterol og blodtryk)

74 / 85



APPENDIX

med R-programbidder svarende til diverse slides:

- ▶ Indlæsning og scatter plot, s. 76
- ▶ Regressionsanalyse, s. 77-79
- ▶ Korrelation, s. 80
- ▶ Konfidens- og prediktionsgrænser, s. 81
- ▶ Modelkontrol, s. 82-83
- ▶ Diagnostics, s. 84-85

75 / 85



Indlæsning og scatter plot

Slide8

```
hjerte<-read.table("vcf.dat", header=T)

plot(hjerte$blodsukker, hjerte$vcf,
      ylab="vcf", xlab="blodsukker",
      cex.lab=1.8, cex=1.5, col="blue")
```

eller med modelformel notation

```
plot(hjerte$vcf ~hjerte$blodsukker,
      ylab="vcf", xlab="blodsukker",
      cex.lab=1.8, cex=1.5, col="blue")
```

76 / 85



Regressionsanalyse

Slide 13 og 14

```
model1 = lm(vcf ~ blodsukker,
            na.action=na.exclude,
            data=hjerte)
```

Man kan efterfølgende anvende forskellige funktioner på objektet model1, typisk:

```
summary(model1)
confint(model1)
predict(model1)
```

og evt.

```
anova(model1)
77 / 85
```



Regressionslinie

Slide 16

For at tegne linien skrives først koden for selve plottet:

```
plot(hjerte$blodsukker, hjerte$vcf,
     ylab="vcf", xlab="blodsukker",
     cex.lab=1.8, cex=1.5, col="blue")
```

og derefter tilføjes linien:

```
abline(model1, col="red", lwd=2)
```



Forventet værdi for specifikke værdier af kovariaten

Slide 17-18

I R benyttes en ny data frame, der indeholder de værdier, der ønskes at prediktere for,

her for en blodsukkerværdi på 10:

```
nyedata<-data.frame(blodsukker=10)
predict(model1, nyedata, interval = "confidence")
```



Korrelationer

Slide 38-39

```
with(hjerte,cor.test(blodsukker, vcf))
```

```
with(hjerte,cor.test(blodsukker, vcf, method="spearman"))
```

Pearson: betegner den *sædvanlige* korrelation
baseret på en todimensional normalfordeling
Når intet angives, er denne underforstået (default)

Spearman: betegner den nonparametriske korrelation



Konfidens- og prediktionsgrænser

Slide 45

Den pæne figur fra SAS er vanskelig at reproducere i R, men man kan lave noget, der ligner ved at prediktere for en masse værdier i en ny data frame:

```
nyedata<-data.frame(blodsukker=(30:200)/10)
```

```
predict(model1, nyedata, se.fit = TRUE)
pred.plim <- predict(model1, nyedata, interval = "prediction")
pred.clim <- predict(model1, nyedata, interval = "confidence")
```

og derefter tegne disse

```
matplot(nyedata$blodsukker, cbind(pred.clim, pred.plim),
       lty = c(1,2,2,1,3,3), col = c("blue","blue","blue","blue","red","red"),
       type = "l", ylab = "predicted vcf",xlab="blodsukker")
points(hjerte$blodsukker, hjerte$vcf)
```

81 / 85



Modelkontrol

Slide 51

```
scatter.smooth(hjerte$blodsukker, resid(model1),col="blue")
loess.smooth(hjerte$blodsukker, resid(model1),col="red",lwd=2)
abline(0,0,col="red",lty=2)
```

Slide 52

```
plot(predict(model1), resid(model1))
abline(0,0,col="red",lty=2)
```

```
par(mfrow=c(2,1))
plot(model1,which=2, cex.lab=1.5)
hist(resid(model1))
```

82 / 85



Check af varianshomogenitet

Slide 54

Her afbildes kvadratroden af de numeriske residualer overfor de predikterede værdier:

```
plot(model1,which=3, cex.lab=1.5)
```

83 / 85



Regression diagnostics

Slide 66

Halvautomatisk figur i R:

```
plot(model1,which=4, cex.lab=1.5)
abline(0.17,0,col="red",lty=2)
```

Vil man mere end det, kan man få Cooks afstand ud som en ny variabel:

```
cook = cooks.distance(model1)
```

og derefter anvende disse i de plots, man måtte ønske.

84 / 85



Regression diagnostics i REG

Slide 67

Cooks afstand spaltes ud i koordinater:

```
dfbetas(model1)
```

og analysen kan laves uden en evt. “*forkert*” observation:

```
model1 = lm(vcf[-13] ~ blodsukker[-13],  
            na.action=na.exclude, data=hjerte)
```

