Faculty of Health Sciences

Basal Statistik

Regressionsanalyse.

Lene Theil Skovgaard

2. marts 2020

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

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

Data

Sammenhørende registreringer (x_i, y_i),
for en række individer eller 'units', $i = 1, \cdots, n$:

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)

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

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 *multipel regression*, i form af en *kovariansanalyse* (omtales i næste forelæsning)

Eksempel fra bogen

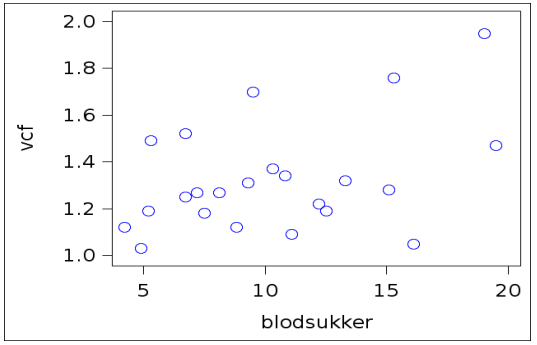
Hvilken sammenhæng er der mellem fastende blodsukkerniveau og sammentrækningsevne for venstre hjertekammer hos diabetikere? (n=23)

Outcome: Y=vcf, %/sec.

Kovariat:
X=blodsukker, mmol/l

OBS	BLODSUKKER	VCF
1	15.3	1.76
2	10.8	1.34
3	8.1	1.27
.	.	.
.	.	.
.	.	.
21	4.9	1.03
22	8.8	1.12
23	9.5	1.70

Scatter plot

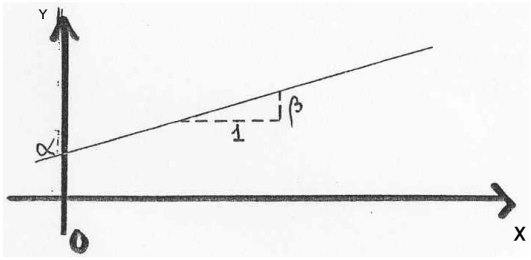



```
data vcf;
infile "http://staff.pubhealth.ku.dk/
    ~lts/basal/data/vcf.txt"
     URL firstobs=2;
input blodsukker vcf;
run;

proc sgplot data=vcf;
  scatter x=blodsukker y=vcf;
run;
```

Er der nogen sammenhæng her?

Matematisk model

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



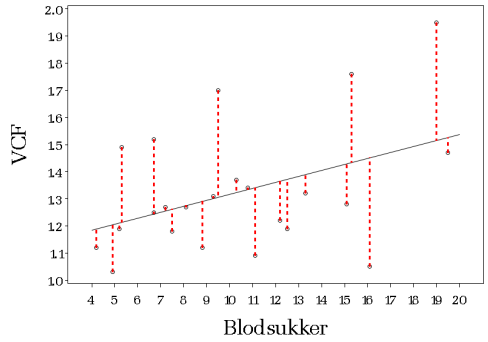
Intercept α: Skæring med Y-akse
Hældning β

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".

Statistisk model

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



(ingen kode til denne figur)

∗ Estimation

Mindste kvadraters metode:
Bestem α og β, så kvadratafvigelsessummen

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

bliver *mindst mulig*. Resultat (estimater):

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

Regressionsanalyse i praksis i SAS

REG:
Den simple, der giver mest automatik:

```
proc reg data=vcf;
  model vcf = blodsukker / clb;
run;
```

GLM:
Den, der kan udvides til generelle lineære modeller,
(og som derfor bruges her):

```
proc glm data=vcf;
  model vcf = blodsukker / clparm;
run;
```

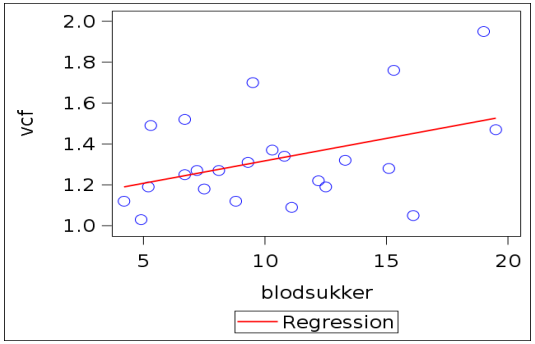

Output fra regressionsanalyse i SAS (GLM)

```
Dependent Variable: vcf

                                 Sum of
Source                     DF       Squares   Mean Square  F Value  Pr > F
Model                       1    0.20726892    0.20726892     4.41  0.0479
Error                      21    0.98609630    0.04695697
Corrected Total            22    1.19336522

R-Square     Coeff Var      Root MSE      vcf Mean
0.173684     16.34634      0.216696      1.325652

Source                     DF     Type I SS   Mean Square  F Value  Pr > F
blodsukker                  1    0.20726892    0.20726892     4.41  0.0479

Source                     DF   Type III SS   Mean Square  F Value  Pr > F
blodsukker                  1    0.20726892    0.20726892     4.41  0.0479

                                    Standard
Parameter          Estimate           Error    t Value    Pr > |t|
Intercept       1.097814878      0.11748118       9.34      <.0001
blodsukker      0.021962522      0.01045358       2.10      0.0479

Parameter        95% Confidence Limits
Intercept       0.853499382  1.342130374
blodsukker      0.000223108  0.043701937
```


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 (Root MSE),
  $s = \hat{\sigma} = 0.216696 \approx 0.217$
  Denne størrelse benyttes til konstruktion af prediktionsgrænser (kommer senere), som er normalområder for given blodsukkerværdi.

Estimeret regressionslinie



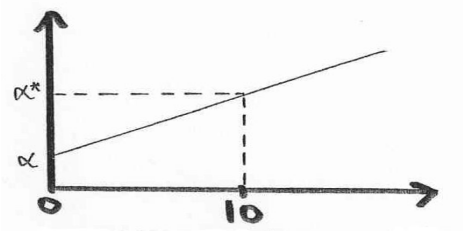

```
proc sgplot data=vcf;
  reg x=blodsukker y=vcf /
      markerattrs=(color=blue)
      lineattrs=(color=red);
run;
```

Estimeret regressionslinie: vcf=1.10 + 0.0220 blodsukker

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 ....

Usikkerheder (standard errors) 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.

Forventet værdi for blodsukker-værdien 10, II

Man kan benytte to muligheder:

- Flytte Y-aksen hen i 10 ved at benytte kovariaten `blodsukker10=blodsukker-10`
- Lade SAS gøre det ved hjælp af en `estimate`-sætning:

```
estimate 'blodsukker=10' intercept 1 blodsukker 10;
```

som giver outputtet:

Parameter	Estimate	Standard Error	t Value	Pr > |t|
blodsukker=10	1.31744010	0.04535290	29.05	<.0001

Parameter	95% Confidence Limits	
blodsukker=10	1.22312358	1.41175662

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

Variansen 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

Spredning omkring regressionslinie

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

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

som er estimat for spredningen omkring regressionslinien

som i SAS benævnes (lidt uheldigt*):
'*Root Mean Square Error*',
forkortet `Root MSE` i output.

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

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

Usikkerhed på estimaterne

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

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

dvs. hældningen er præcist bestemt, hvis

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

Konfidensinterval = Sikkerhedsinterval

Estimeret usikkerhed på $\hat{\beta}$: 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.e.)

Vi bruger det til at konstruere et 95% konfidensinterval

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

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

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), P=0.048

dvs. lige på grænsen af det signifikante

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:

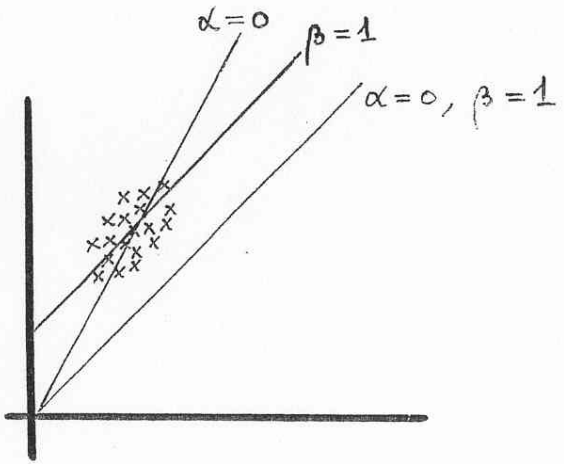
$$1.098 \pm 2.080 \times 0.1175 = (0.854, 1.342)$$

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

Når man har flere hypoteser...

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

- Estimaterne 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

Variationsopspaltning

Forklaring til de øverste linier i output s. 14:

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

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

x er en god forklarende variable,
hvis SS_{model} er *stor* i forhold til SS_{error}

Jeg foretrækker at kalde SS_{error} for SS_{residual}...

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: $R^2 = 0.17$, dvs. vi kan forklare 17% af variationen i `vcf` v.hj.a. variablen `blodsukker`.

Determinationskoefficienten er
kvadratet på korrelationskoefficienten
mellem `vcf` og `blodsukker` (som dermed er $r = \sqrt{0.17} = 0.42$)

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?

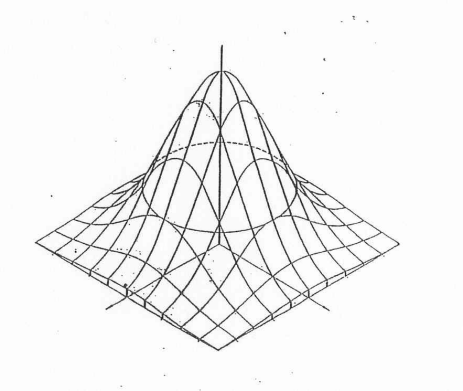
Korrelationskoefficienten estimeres ved:

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

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

Todimensional normalfordelingstæthed, r=0

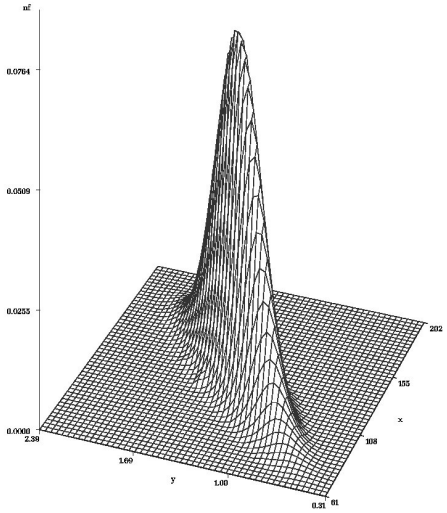
Korrelation 0



Alle lodrette snit giver normalfordelinger

- med samme middelværdi
- og samme varians

Todimensional normalfordelingstæthed, r=0.9

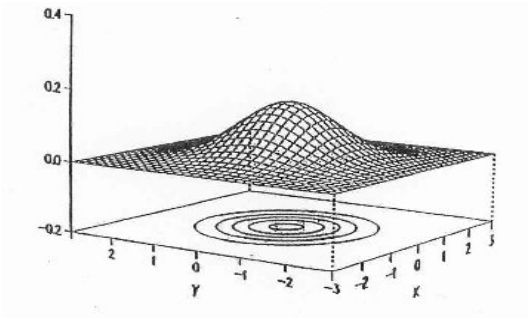


Korrelation 0.9

Udpræget retning i figuren

Konturkurverne

for en normalfordeling bliver ellipser



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

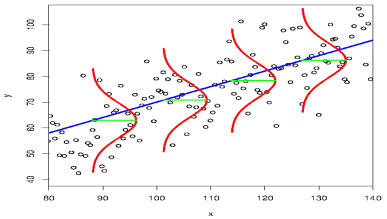
- Giver det indtryk af noget elliptisk?
- med flest observationer i de "inderste" ellipser

Regression kontra korrelation

Regressionsmodellen

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

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.



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.

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

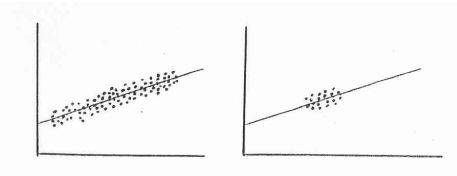
$$r_{xy} = \hat{\beta}\sqrt{\frac{S_{xx}}{S_{yy}}}$$

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

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?

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.....

Korrelationer i praksis

I SAS benyttes PROC CORR:

```
proc corr pearson spearman;
    var vcf blodsukker;
    run;
```

med oplagte options (`pearson` er default).

og vi får output som vist på næste side.

Output med korrelationer

```
The CORR Procedure
   2  Variables:    vcf        blodsukker

Pearson Correlation Coefficients, N = 23
       Prob > |r| under H0: Rho=0

                     vcf       blodsukker
vcf              1.00000         0.41675
                                 0.0479

blodsukker       0.41675         1.00000
                  0.0479
```

```
The CORR Procedure
   2  Variables:    vcf        blodsukker

Spearman Correlation Coefficients, N = 23
       Prob > |r| under H0: Rho=0

                     vcf       blodsukker
vcf              1.00000         0.31800
                                 0.1392

blodsukker       0.31800         1.00000
                  0.1392
```

Her aflæses korrelationerne:

Pearson: 0.42, P=0.048

Spearman: 0.32, P=0.14

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

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

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 FEV_1 er 0.62 for mænd og kun 0.49 for kvinder....

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

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

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!

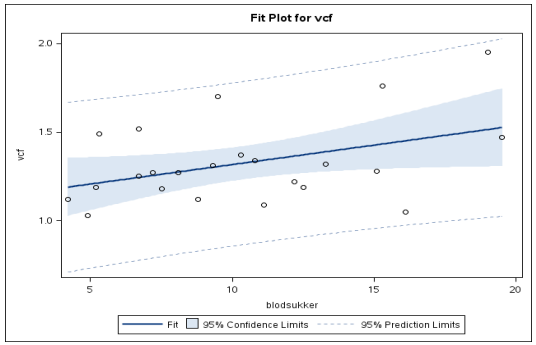
Prediktionsgrænser for enkeltindivider

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.

Konfidens- og prediktionsgrænser i praksis

Figuren dannes enten automatisk, eller ved brug af `ods graphics`, se s. 81



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 (`Root MSE`)

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.

Modelkontrol

Den statistiske model var

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

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!!

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Grafisk modelkontrol

Fokus er her på

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

eller en modifikation af disse:

- Normerede/Standardiserede/Studentized: `STUDENT`
- Leave-one-out: `PRESS`
- Studentized Press: `RSTUDENT`

Alle disse kan fås i nyt datasæt ved at skrive (se også s. 82):

```
output out=ny p=yhat r=resid student=stresid cookd=cook
             rstudent=uresid press=press;
```

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*)

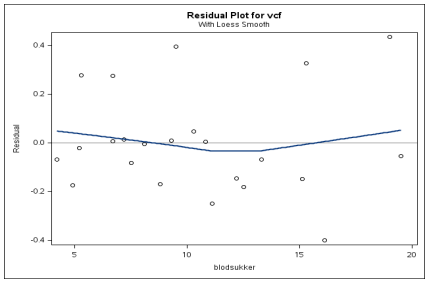
Disse plots ses på s. 51 og 52, og kommenteres s. 51 og 53

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Residualplots i praksis

Check af linearitet Residualer plottet mod kovariaten,
med overlejret udglatning (hele koden s. 82):

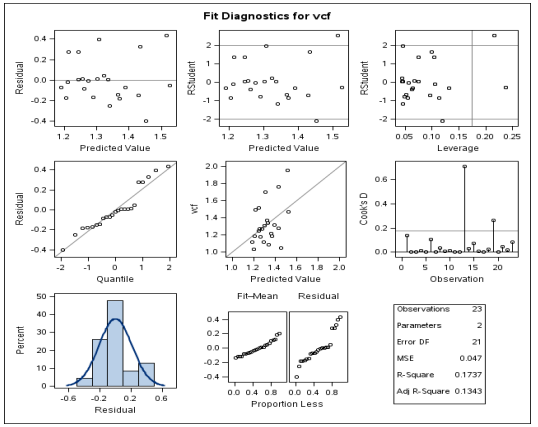
```
proc glm PLOTS=(DIAGNOSTICS RESIDUALS(smooth)) data=vcf;
```



Her ses ingen oplagte (store) buer

Diagnostics Panel - kode s. 82

fås som regel automatisk, ellers brug `ods graphics`



Kommentarer til Diagnostics Panel

Vi ser foreløbig på den venstre kolonne på s. 52:

Øverst: Plot af residualer mod predikterede værdier, dvs. plot af type 2 fra s. 50:
Her anes muligvis lidt trompetfacon

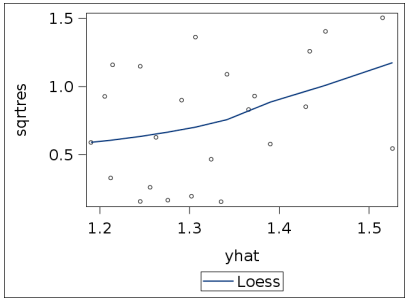
Mellemste: Fraktildiagram til check af normalfordelte residualer.
Det ser rimeligt ud, men der anes ligesom et *hul* i fordelingen (type 3 fra s. 50)

Nederst: 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

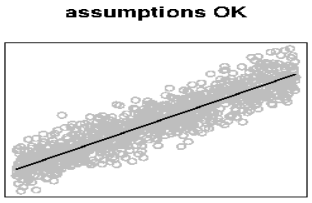
Bedre check af varianshomogeniteten

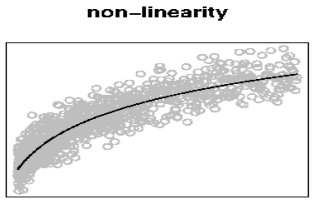
Plot af kvadratrod af numerisk værdi af normerede residualer, mod de predikterede værdier (se kode s. 83):

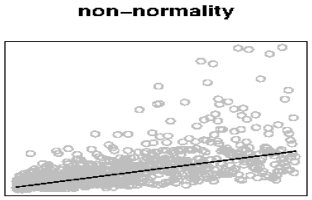


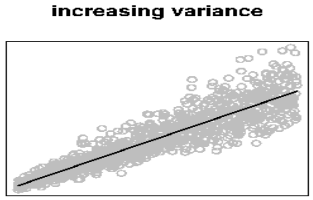
Der er en tendens til højere spredning for de høje predikterede værdier. Måske konstant *relativ* spredning?

Afvigelser fra modellen









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
  - kvadratrod, invers
  - benytter lineære splines ("knæk-linier", kommer senere)
- tilføjer flere kovariater, f.eks.
  - `alder`, køn, medicin etc., eller $\log(x)$
- foretager ikke-lineær regression

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: (`quantreg`)

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

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

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

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...

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!!

Lidt om logaritmer – måske genopfriskning...?

Alle logaritmefunktioner har et grundtal

- 10-tals logaritmen (`log10`) har grundtal 10
- Den såkaldt *naturlige* logaritme (`log`, før i tiden ofte kaldet `ln`) har grundtal $e = 2.71828$
- 2-tals logaritmen (`log2`) 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...*:

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 variationskoefficenten (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 logaritmetransformeres, bliver effekten af kovariater (når de tilbagetransformeres) til faktorer, der skal ganges på.

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.

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

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 ikke i GLM
- Normalfordelingen:
  Der findes test, men de kan ikke generelt anbefales

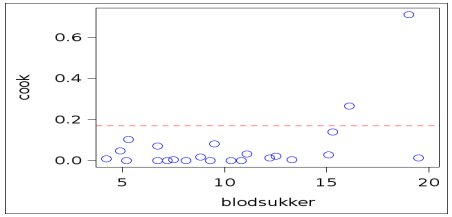
Regression diagnostics

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?

Cooks afstand

Vi har allerede set en figur af disse i DiagnosticsPanel tidligere, se s. 52 (højre kolonne i midten).
Alternativt kan man benytte en `output`-sætning og selv tegne en passende figur (se s. 82 og 84):



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

Cooks afstand spaltet i komponenter

Benyt option `influence` i `model`-sætningen (kun proceduren REG, se appendix s. 85), og få svar på:
Hvor mange s.e.($\hat{\beta}_1$)'er ændres f.eks. $\hat{\beta}$,
når den i'te person udelades?

```
Dependent Variable: vcf
      Output Statistics

        -------DFBETAS-------
     Obs Intercept     blodsukker

       1   -0.2421      0.4134
       2    0.0014      0.0005
       3   -0.0049      0.0030
       4    0.1082     -0.1461
       5    0.0150     -0.0104
       .      .           .
       .      .           .

      11    0.0060     -0.0043
      12   -0.0345      0.0277
      13   -0.8744      1.1973    <------------------
      14    0.0982     -0.1718
      15    0.3409     -0.2478
       .      .           .
       .      .           .
```

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

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: $y = 1.189 + 0.011x$

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

Outliers

Observationer, der *ikke passer* ind i sammenhængen

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

Press-residualer

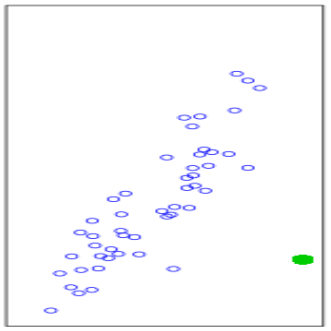
Residualer, der fremkommer efter at den pågældende observation har været udelukket fra estimationen.

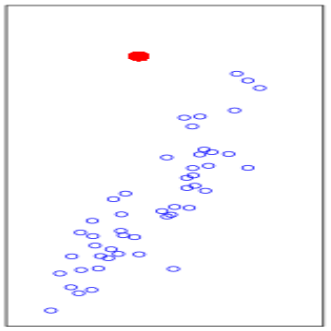
(residualer `without current observation`)

Man kan med fordel danne et nyt datasæt med predikterede værdier og residualer, se kode i appendix s. 82

To forskellige eksempler

på mulige *outliers*:





Hvad gør vi i hver af disse situationer?

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....*

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?

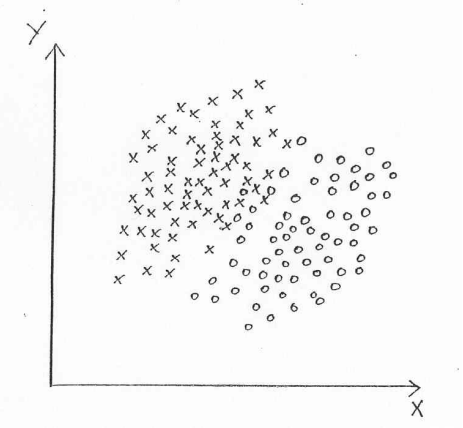
- hvis de ligger meget *yderligt i kovariat-værdier*
  - husk at afgrænse konklusionerne tilsvarende!
- hvis man kan finde årsagen
  - og da skal alle sådanne udelades!

– mere om dette senere (ved øvelserne)

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Confounding, bare lige et smugkig

meget mere om dette senere

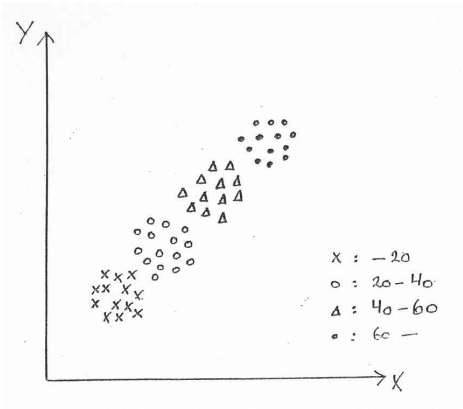


(Kor)relationen er:

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

Eks: Kolesterol vs. chokoladeindtag

Confounding, II



(Kor)relationen er:

- tilsyneladende positiv
- 0 for hver aldersgruppe

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

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

APPENDIX

med SAS-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

Indlæsning og scatter plot

Slide 8

```
data vcf;
infile "http://staff.pubhealth.ku.dk/~lts/basal/data/vcf.txt" URL firstobs=2;
input blodsukker vcf;
run;

proc sgplot data=vcf;
  scatter x=blodsukker y=vcf;
run;
```

eller evt. med den *gamle* kode:

```
proc gplot data=vcf;
  plot vcf*blodsukker;
run;
```

Regressionsanalyse

Slide 13

Med GLM:

```
proc glm data=vcf;
  model vcf = blodsukker / clparm;
run;
```

eller med REG:

```
proc reg data=vcf;
  model vcf = blodsukker / clb
run;
```

Udbygget kode s. 79 og 82

Regressionslinie

Slide 16

For at tegne linien skrives f.eks.

```
proc sgplot data=vcf;
  reg x=blodsukker y=vcf /
      markerattrs=(color=blue) lineattrs=(color=red);
run;
```

eller med den lidt *gamle* kode

```
proc gplot data=vcf;
  plot vcf*blodsukker;
symbol v=circle c=red i=rl h=2 l=1;
run;
```

Forventet værdi for specifikke værdier af kovariaten

Slide 18

I SAS benyttes en estimate-sætning,

Her for en blodsukkerværdi på 10:

```
proc glm data=vcf;
  model vcf=blodsukker / clparm;
  estimate 'blodsukker=10' intercept 1 blodsukker 10;
run;
```

Korrelationer

Slide 38-39

```
proc corr pearson spearman;
    var vcf blodsukker;
    run;
```

Pearson: betegner den *sædvanlige* korrelation baseret på en todimensional normalfordeling

Spearman: betegner den nonparametriske korrelation

Konfidens- og prediktionsgrænser

Slide 45

Den pæne figur s. 45 fås formentlig automatisk, når der udføres en lineær regression.

Ellers må man benytte `ods graphics on;`
eller den mere gammeldags procedure `gplot` med en af nedenstående `symbol`-sætninger:

```
proc gplot data=vcf;
  plot vcf*blodsukker;
*  symbol v=circle c=blue i=rlclm95; /* konfidens */
  symbol v=circle c=blue i=rlcli95; /* prediktion */
run;
```

Modelkontrol

Slide 51 og 52

```
ods graphics on;
proc glm PLOTS=(DIAGNOSTICS RESIDUALS(smooth)) data=vcf;
  model vcf=blodsukker / clparm;
  output out=ny p=yhat r=resid student=stresid
               rstudent=uresid press=press
               cookd=cook;
run;
ods graphics off;
```

danner `DiagnosticsPanel` med mange brugbare figurer, samt en figur til vurdering af lineariteten

Hvis man vil have fuld kontrol over hele modelkontrollen kan man udregne predikterede værdier og residualer, samt gemme disse i et nyt datasæt (her kaldet `ny`) benyttes ovenstående `output`-linie

Check af varianshomogenitet

Slide 54

Her afbildes kvadratroden af de numeriske residualer overfor de predikterede værdier, eller evt. kovariaten
(datasættet `ny` er defineret på s. 82):

```
data tegn;
set ny;

sqrtres=sqrt(abs(stresid));
run;

proc sgplot data=tegn;
    loess Y=sqrtres X=yhat; run;

proc sgplot data=tegn;
    loess Y=sqrtres X=blodsukker; run;
```

Regression diagnostics

Slide 66

Plot af Cooks afstand mod observationsnummer ses i `DiagnosticsPanel`, se s. 52

Vil man mere end det, kan man få dem ud i et nyt datasæt ved at benytte `output`-sætningen (se s. 82) med `cookd=cook;`

Herefter kan vi lave tegninger direkte:

```
proc sgplot data=ny;
    scatter x=blodsukker y=cook /
            markerattrs=(symbol=circle color=blue size=15);
run;
```

Regression diagnostics i REG

Slide 67

I proceduren REG kan Cooks afstand spaltes ud i koordinater:

```
proc reg data=vcf;
  model vcf = blodsukker / influence;
run;
```