Faculty of Health Sciences

Basal statistik

Lille SAS Manual

Lene Theil Skovgaard

3. februar 2020

Selve sproget

Siderne 9-18

- Indlæsning (s. 9-12)
- Definition af nye variable (s. 13)
- Missing values / Manglende værdier (s. 14)
- Transformation (s. 15)
- Sammensætning af datasæt (`set` og `merge`, s. 16-17)
- Opdeling af datasæt (`by`, `where`, s. 18)

Basale procedurer

- `PROC MEANS` (s. 19):
  Summary statistics
- `PROC UNIVARIATE` (s. 26,31):
  Summary statistics, grafik og tests
- `PROC SORT` (s. 17):
  Sortering efter variable

Grafik

Siderne 20-29

- Scatter plot, plot med linier (s. 21-23)
- Histogrammer (s. 24)
- Box plot (s. 25)
- Fraktildiagram, qqplot (s. 26)
- Kaplan-Meier plot (s. 27)
- Opdelte plots (panels) (s. 28)
- Plots i forbindelse med statistiske analyser (s. 29)

Statistiske analyser, I

Se oversigten på s. 7

- Parret sammenligning (s. 31)
  - Parret T-test
  - Wilcoxon signed rank test
- Sammenligning af to grupper (s. 32)
  - Uparret T-test
  - Mann-Whitney test
- Sammenligning af mere end to grupper (s. 33)
  - Variansanalyse (ANOVA)
  - Kruskal-Wallis test
- Lineær regression (s. 34)
- Multipel lineær regression (s. 35-36)

Statistiske analyser, II

Se oversigten på s. 7

- Tabeller/χ^2-test (s. 37-38)
- Logistisk regression (s. 39)
- Proportional odds modeller (s. 40)
- Overlevelsesdata (s. 41)
  - Kaplan-Meier kurver
  - Log-rank test
  - Cox regression
- Gentagne målinger, mixed models (s. 42)
  - Varianskomponentmodeller
  - Gentagne målinger over tid

Forklarende variable = Kovariater

Outcome	**Dikotom**	**Kategorisk**	**Kvantitativ**	**Kategoriske og kvantitative**
Dikotom	2*2-tabeller	χ^2-test	Logistisk regression	
Kategorisk	Tabeller/χ^2-test		Generaliseret logistisk regression	
Ordinale	svært, f.eks. proportional odds modeller			
Kvantitativ	Mann-Whitney Wilcoxon signed rank	Kruskal-Wallis Friedman	Robust multipel regression	
Normalfordelt	T-test parret/uparret	Variansanalyse ensidet/tosidet	Kovariansanalyse Multipel regression	
Censureret	Log-rank test		Cox regression	
Korrelerede normalfordelte	Varianskomponent-modeller		Modeller for gentagne målinger	

Notation

På de følgende sider benyttes forskellige variabelnavne.
Nogle af disse vil være let gennemskuelige, f.eks. `alder`, `bmi`, mens andre er mere generiske og forklares herunder:

- Y er et kvantitativt outcome (f.eks. blodtryk eller fødselsvægt)
- GRP er en gruppering af individer (f.eks. køn med to værdier, eller behandling med tre værdier)
- X1 og X2 er kvantitative kovariater (f.eks. alder og bmi)

Datasæt kaldes her ofte blot a1 eller a2, men i praksis bør man benytte mere sigende navne (helst korte).

Indlæsning fra nettet

Hent tekst-filen fil.txt fra nettet

```
DATA a1;
INFILE "http://staff.pubhealth.ku.dk/~lts/basal/data/fil.txt" URL FIRSTOBS=2;
INPUT grp$ y x1 x2;
RUN;
```

eller sådan her

```
FILENAME navn URL "http://staff.pubhealth.ku.dk/~lts/basal/data/fil.txt";

DATA a1;
INFILE navn FIRSTOBS=2;
INPUT grp$ y x1 x2;
RUN;
```

Indlæsning fra eget drev

Hent tekst-filen fil.txt fra mappen "minfolder"

```
DATA a1;
INFILE "C:\minfolder\fil.txt" FIRSTOBS=2;
INPUT grp$ y x1 x2;
RUN;
```

Hentning af allerede eksisterende SAS-data

Hent SAS-data "fil" fra mappen "minfolder"

```
LIBNAME sas "C:\minfolder";

DATA a1;
SET sas.fil;
RUN;
```

Hent SAS-data "fil" fra mappen "sasuser"

```
DATA a1;
SET sasuser.fil;
RUN;
```

Indlæsning fra Excel

Hent fil.xls ind fra mappen "minfolder" på eget drev

```
PROC IMPORT OUT= WORK.a1 DATAFILE="C:\minfolder\fil.xls"
            DBMS=xls REPLACE;
     GETNAMES=YES;
RUN;
```

Se også Birthes mere udførlige vejledning på linket:

`http://publicifsv.sund.ku.dk/~lts/basal/programmer/IndlaesExcel.sas`

Man kan altid erstatte navnet "a1" med "sasuser.a1",
hvis man vil lave en permanent sasuser-fil

Definition af nye variable

skal skrives efter `data a1;` og inden det første `run;`
dvs. i det, vi kalder et DATA step

Eksempel:

```
DATA a1;
INFILE "C:\article2\vitaminD.txt" FIRSTOBS=2;
INPUT country$ category$ vitd age height weight sunexp$ vitdintake;

logvitd=log10(vitd);
bmi=weight/(height/100)**2;

IF bmi>25 THEN fat=1;
IF bmi>0 AND bmi<=25 THEN fat=0;
RUN;
```

Bemærk:
Det er vigtigt, at der kræves `bmi>0` ved definition af `fat=0`, idet missing values ellers ville blive til `fat=0`, se s. 14

Missing values = manglende/uoplyste værdier

- Numeriske variable (tal, der kan regnes på):
  Benyt punktum, og aldrig -9, 999 etc.
- Karaktervariable (f.eks. `mand`, `kvinde` etc):
  Benyt `NA` (Not Available) eller punktum
  aldrig blanke

Bemærk: Et punktum for en numerisk variabel anses for at være mindre end alle tal (dvs. minus uendelig), så man skal passe på med logiske sammenligninger, såsom definitionen af `fat` på s. 13

Transformation

typisk (langt overvejende) logaritmetransformation:

Her er i virkeligheden tale om definition af en ny variabel, idet man aldrig bør redefinere en allerede eksisterende variabel.

Se derfor s. 13, hvor variablen `vitd` bliver logaritmetransformeret og lagt i den nye variabel `logvitd`.

Vaiabelnavne skal starte med et bogstav og
må ikke indeholde mellemrum!

Brug evt. underscore til at adskille bogstaver.

Sammensætning af data, I

Sammensæt to datasæt, et med kvinderne (`female`) og et med mændene (`male`):

```
DATA alle_observationer;
SET male female;
RUN;
```

Disse datasæt sættes under hinanden, så der altså kommer flere observationer i det fælles datasæt.

Sammensætning af data, II

Sammensæt to datasæt, et med alder, højde, vægt etc. (`basalt`) og et med blodprøvesvar (`blod`) *for de samme individer* (`id`):

```
PROC SORT DATA=basalt; BY id;
RUN;
PROC SORT DATA=blod; BY id;
RUN;

DATA alle_variable;
MERGE basalt blod; BY id;
RUN;
```

Disse datasæt sættes ved siden af hinanden, så der altså kommer flere variable i det fælles datasæt.
Det kræver forudgående sortering af begge datasæt (PROC SORT).

Opdelte analyser

Foretag analyser på f.eks. mænd og kvinder hver for sig:

```
PROC SORT DATA=a1, BY gender;
RUN;

PROC REG DATA=a1; BY gender;
MODEL y=x1;
RUN;
```

Foretag kun analysen på f.eks. kvinderne:

```
PROC REG DATA=a1; WHERE gender="female";
MODEL y=x1;
RUN;
```

Summary statistics

såsom gennemsnit, median, spredning etc.

- Den *skrabede* kode:

  ```
  PROC MEANS DATA=a1;
  RUN;
  ```

- Koden med lidt mere selvbestemmelse,
  i form af summary statistics, variable samt opdeling efter gruppe (`gender`):

  ```
  PROC MEANS N MEAN MEDIAN STDERR DATA=a1;
  CLASS gender;
  VAR x1 x2;
  RUN;
  ```

Grafik

Der er grundlæggende 3 metoder/systemer til grafik i SAS:

1. `PROC GPLOT`:
   Den *gamle* procedure, som kan nærmest alt, men som giver ret grimme plots, hvis ikke man tilføjer en del options.
2. `PROC SGPLOT` (og `SGSCATTER`, `SGPANEL`):
   De nyere procedurer, som meget let giver pæne tegninger, men som er sværere at ændre på.
3. `ODS`-systemet i forbindelse med statistiske procedurer, som giver relevante tegninger af prediktioner, residualer mv. i en pæn udgave.

PROC GPLOT

Vi vil tegne blodtryk (bp) op mod alderen (alder), med forskellige symboler for køn (variablen gender):

- Den *skrabede* kode:

```
PROC GPLOT DATA=a1;
PLOT bp*alder=gender;
RUN;
```

- Koden til det *pænere* plot:

```
PROC GPLOT DATA=a1;
PLOT bp*alder=gender
  / HAXIS=axis1 VAXIS=axis2 frame;
AXIS1 ORDER=(20 to 80 by 5) VALUE=(H=2) MINOR=none LABEL=(H=3);
AXIS2 ORDER=(100 to 160 by 10) VALUE=(H=2) MINOR=none
      LABEL=(A=90 R=0 H=3);
SYMBOL1 V=circle I=none C=red H=2;
SYMBOL2 V=triangle I=none C=blue H=2;
RUN;
```

SYMBOL-sætninger i GPLOT

- Plots med regressionslinier:
  I=rl for regressionslinie, L= for stipling, og W= for tykkelse:

```
SYMBOL1 V=circle I=rl C=red H=2 L=1 W=2;
SYMBOL2 V=triangle I=rl C=blue H=2 L=2 W=2;
```

- Regressionslinier med konfidens- eller prediktionsgrænser:
  I=rlclm95 hhv I=rlcli95
- Udglattede kurver (for at afgøre linearitet):
  I=sm75s, hvor 75 blot skal være et tale mellem 1 og 99, idet stort tal svarer til meget udglatning

PROC SGPLOT

Plottet fra s. 21 kan fremstilles således:

```
PROC SGPLOT DATA=a1;
SCATTER Y=bp X=alder / group=gender;
RUN;
```

og hvis man selv vil styre noget mere, kan man skrive:

```
PROC SGPLOT DATA=a1;
SCATTER Y=bp X=alder / GROUP=gender
 MARKERATTRS=(SYMBOL=circlefilled);
RUN;
```

Histogram

med overlejret udglattet kurve:

```
PROC SGPLOT DATA=a1;
    HISTOGRAM bp;
    DENSITY bp;
RUN;
```

og hvis vi skal opdele efter køn:

```
PROC SGPANEL DATA=a1;
    PANELBY gender / ROWS=1;
    HISTOGRAM bp;
    DENSITY bp;
RUN;
```

Se også s. 26, hvor histogrammer laves vha UNIVARIATE-proceduren

Box Plots

bruges næsten altid kun med flere grupper,
som f.eks opdelt efter køn:

```
PROC SGPLOT DATA=a1;
    VBOX bp / CATEGORY=gender;
RUN;
```

eller med den direkte procedure:

```
PROC SORT DATA=a1;
BY gender;
RUN;

PROC BOXPLOT DATA=a1;
PLOT bp*gender;
RUN;
```

Fraktildiagrammer

kan enten laves ved hjælp af proceduren UNIVARIATE
(se mere om denne s. 31)
eller i forbindelse med modelkontrol, hvor der laves fraktildiagram af residualer, ved brug af ODS-systemet.

```
PROC UNIVARIATE DATA=a1;
   QQPLOT y;
RUN;
```

Her kan også laves histogrammer:

```
PROC UNIVARIATE DATA=a1;
   HISTOGRAM y;
RUN;
```

Kaplan-Meier overlevelseskurver

Her betegner `eventtime` det tidspunkt, hvor der sker noget for personen. Det kan være enten et event (`censur`$\neq 0$) eller en censurering (`censur`=0).
Der opdeles i grupper, efter variablen `grp`.

```
ODS GRAPHICS ON;
   PROC LIFETEST DATA=a1 PLOTS=(s);
      TIME eventtime*status(0);
      STRATA grp;
   RUN;
ODS GRAPHICS OFF;
```

Flere figurer samlet på en side

kaldes PANELS:

Ved brug af SGPANEL, f.eks. med 4 forskellige behandlinger (`grp`), hvor der i hvert plot tegnes tidsudviklinger af blodtrykket (`bp`) for hvert individ (`id`):

```
PROC SGPANEL DATA=a1;
PANELBY grp / rows=2 columns=2;
SERIES Y=bp X=tid / GROUP=id;
RUN;
```

ODS-systemet

Otput **D**elivery **S**ystem giver

- plots af *modellen* (de predikterede værdier)
- relevante modelkontrol-tegninger

i forbindelse med statistiske analyser.
Hvilke plots, der produceres, afhænger helt af den anvendte procedure, men fælles er opsætningen

```
ODS GRAPHICS ON;
PROC et-eller-andet PLOTS=all DATA=a1;
........
RUN;
ODS GRAPHICS OFF;
```

I stedet for `PLOTS=all` kan man skrive navnet på specifikke plots.

Statistiske analyser

På de følgende sider benyttes disse betegnelser:

- Y er et kvantitativt outcome (f.eks. blodtryk eller fødselsvægt)
- GRP er en gruppering af individer (f.eks. køn med to værdier, eller behandling med tre værdier)
- X1 og X2 er kvantitative kovariater (f.eks. alder og bmi)

De nonparametriske metoder benyttes, hvis man ikke med nogenlunde rimelighed kan antage, at *residualerne* er normalfordelte,

i særdeleshed hvis man ikke har store mængder af data.

Parret sammenligning

Her angiver y1 og y2 to målinger på samme unit (f.eks. målt i hvile kontra bevægelse, eller med to forskellige apparater):

- Parametrisk, dvs. parret T-test:

```
PROC TTEST DATA=a1;
PAIRED y1*y2;
RUN;
```

- Nonparametrisk, dvs. et Wilcoxon signed rank test:

```
DATA a2;
SET a1;
diff=y1-y2;
RUN;
PROC UNIVARIATE DATA=a2;
VAR diff;
RUN;
```

Sammenligning af to grupper

- Parametrisk, dvs. T-test:

```
PROC TTEST DATA=a1;
CLASS grp;
VAR y;
RUN;
```

- Nonparametrisk, dvs. et Mann-Whitney test, også kaldet et Wilcoxon rank-sum test:

```
PROC NPAR1WAY DATA=a1 WILCOXON;
CLASS grp;
VAR y;
*Exact;
RUN;
```

Bemærk, at linien `EXACT;` er udkommenteret, idet den ofte bevirker, at programmet går helt i stå
(selv hvis data kun har moderat størrelse).

Sammenligning af mere end to grupper

- Parametrisk, dvs. en ensidet ANOVA:

```
PROC GLM DATA=a1;
CLASS grp;
MODEL y=grp / SOLUTION CLPARM;
RUN;
```

- Nonparametrisk, dvs. et Kruskal-Wallis test:

```
PROC NPAR1WAY DATA=a1 WILCOXON;
CLASS grp;
VAR y;
RUN;
```

Tosidet ANOVA: Se s. 36

Simpel lineær regression, I

- Ved brug af REG (simplest output):

```
PROC REG DATA=a1;
MODEL y=x1 / CLB;
RUN;
```

- Ved brug af GLM (generaliserer til andre typer analyser):

```
PROC GLM DATA=a1;
MODEL y=x1 / SOLUTION CLPARM;
RUN;
```

Multipel lineær regression, I

- Fit to parallelle regressionslinier (kovariansanalyse)

```
PROC GLM DATA=a1;
CLASS grp;
MODEL y = grp x1 / SOLUTION CLPARM;
RUN;
```

- Fit to ikke-parallelle regressionslinier
  (med interaktion=vekselvirkning= effektmodifikation)

```
PROC GLM DATA=a1;
CLASS grp;
MODEL y = grp x1 grp*x1 / SOLUTION CLPARM;
RUN;
```

- Fit en plan (to kvantitative kovariater)

```
PROC GLM DATA=a1;
MODEL y = x1 x2 / SOLUTION CLPARM;
RUN;
```

Multipel lineær regression, II

med kategoriske kovariater, her kaldet `grp` og `gender`:

- Fit en tosidet ANOVA

```
PROC GLM DATA=a1;
CLASS grp gender;
MODEL y = grp gender / SOLUTION CLPARM;
RUN;
```

- Tosidet ANOVA med interaktion,
  og med estimation af `grp`-effekt for hver `gender`:

```
PROC GLM DATA=a1;
CLASS grp gender;
MODEL y = gender grp*gender grp / SOLUTION CLPARM;
RUN;
```

Tabeller, χ^2-test

Nu betegner grp en gruppering af individer, f.eks. en behandling, og udfald betegner et dikotomt outcome (altså et, der kun kan antage to værdier, f.eks. 0/1 eller ja/nej).
Vi tester om sandsynligheden for udfald=1 (eller ja) er den samme i de to grupper:

- χ^2-test for uafhængighed

```
PROC FREQ DATA=a1;
TABLES grp*udfald / NOCOL NOPERCENT RISKDIFF RELRISK;
RUN;
```

Bemærk: Rækkefølgen af variablene i TABLES-sætningen er vigtig, man skal have grupperingen stående først!
Det er tilladt med mere end to værdier for såvel grp som udfald, men så kan RISKDIFF RELRISK ikke anvendes.
Vigtigt: Se også s. 38.

Tabeller, Fishers eksakte test

Hvis de forventede antal i tabellen er små,
dvs. hvis de ikke opfylder

- Mindst 80% er over 5
- Alle er mindst 1

så skal man anvende Fishers eksakte test i stedet for χ^2-testet

- Fishers eksakte test for uafhængighed

```
PROC FREQ DATA=a1;
TABLES grp*udfald
    / NOCOL NOPERCENT EXPECTED RISKDIFF RELRISK EXACT;
RUN;
```

Bemærk: De forventede antal kan fås ved at benytte option EXPECTED i TABLES-sætningen.

Logistisk regression

Her betegner udfald et dikotomt outcome (altså et, der kun kan antage to værdier, f.eks. 0/1).
Vi undersøger sammenhængen mellem sandsynligheden for udfald=1, i forhold til en eller flere kovariater, enten kategoriske (grp=0 eller 1) eller kvantitative (x1 og x2):

-
```
PROC GENMOD DESCENDING DATA=a1;
CLASS grp;
MODEL UDFALD = grp x1 x2 / DIST=bin LINK=logit;
RUN;
```
-
```
PROC LOGISTIC DATA=a1;
CLASS grp(ref="0") / PARAM=glm;
MODEL UDFALD(EVENT="1") = grp x1 x2 / LACKFIT LINK=logit;
RUN;
```

Proportional odds modeller

Nu betegner udfald et ordinalt outcome (altså et, der kan antage mere end to værdier, på en ordinal skala, f.eks. smerte i 4 kategorier: ingen, let, moderat og svær).

Vi undersøger sammenhængen mellem sandsynlighederne for de enkelte niveauer, i forhold til en eller flere kovariater, enten kategoriske (grp) eller kvantitative (x1 og x2):

-
```
PROC LOGISTIC DATA=a1 DESCENDING;
  CLASS grp(ref="0") / PARAM=glm;
  MODEL udfald = grp x1 x2;
run;
```

Overlevelsesdata

Her betegner `eventtime` det tidspunkt, hvor der sker noget for personen. Det kan være enten et event (`censur`$\neq 0$) eller en censurering (`censur=0`).
Vi undersøger sammenhængen mellem hazard for den pågældende event, og en eller flere kovariater, enten kategoriske (`grp`) eller kvantitative (`x1` og `x2`):

- Log-rank test (non-parametrisk):

```
PROC PHREG DATA=a1;
CLASS grp / PARAM=GLM;
MODEL eventtime*censur(0) = grp / TIES=DISCRETE;
RUN;
```

- Cox-regression (proportionale intensiteter):

```
PROC PHREG DATA=a1;
CLASS grp;
MODEL eventtime*censur(0) = grp x1 x2 / RL;
RUN;
```

Korrelerede data

typisk flere observationer for hvert individ (`patient`) over tid (`tid`). Patienterne tænkes inddelt i grupper, som skal sammenlignes (`grp`).

- Med en simpel kovarians/korrelations-struktur (CS):

```
PROC MIXED DATA=a1;
CLASS grp;
MODEL y = grp tid grp*tid / DDFM=kr SOLUTION CL;
RANDOM INTERCEPT / SUBJECT=patient;
RUN;
```

- Med en seriel kovarians/korrelations-struktur (AR1):

```
PROC MIXED DATA=a1;
CLASS grp;
MODEL y = grp tid grp*tid / DDFM=kr SOLUTION CL;
REPEATED tid / SUBJECT=patient TYPE=ar1;
RUN;
```