Videregående Statistik
Niels Eriksen (dvp957), Casper Eneqvist (xfj248)

Hjemmeopgave 7

**1) Brug Oaxaca dekomponeringsteknikken til at vurdere forskellen mellem national ikke national. (der er taget ln til scoren, men det er ikke nødvendigt at bruge ln, du kan også bare bruge scoren). I første omgang skal du alene bruge ESCS som forklarende variabel.**

Vi kører følgende kode for at få gennemsnittene at etniske og ikke etniske danskere:

```
proc import
datafile="/home/caspereneqvist0/my_courses/anders.milhoj/Sommerskole
2018/Uge 2/min_PISA_renset.sav" out=ud15 dbms = sav replace;
data dk2015;
set ud15;
where cnt='DNK';
national=.;
if IMMIG=1 then national=1;
if 2<=IMMIG<=3 then national=2;
ln_r=log(score_n);
run;
proc means data=dk2015 n mean maxdec=2;
class national;
var score_n ln_r escs ST013Q01TA;
run;
```

Dette giver samme tabel som angivet i opgaven, hvorfor denne blot er kopieret som output:

| PISA DK 2015 | national | antal | score_n | LN-R | ESCS | antal bøger |
|---|---|---|---|---|---|---|
| etnisk dansker | 1 | 5281 | 507,07 | 6,21 | 0,63 | 3,40 |
| anden etnisk herkomst | 2 | 1686 | 429,10 | 6,04 | -0,12 | 2,36 |
| i alt | | 6967 | | | | |
| differens | | | 77,97 | 0,17 | 0,75 | 1,04 |

Vi kører to regressioner via følgende kode:

```
proc reg  data=dk2015;
model score_n=ESCS;
where  national=1;
run;
proc reg  data=dk2015;
model score_n=ESCS;
```

```
where  national=2;
run;
```

Som giver føglende output:

| Parameter Estimates | | | | | | |
|---|---|---|---|---|---|---|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | Intercept | 1 | 486.42873 | 1.36387 | 356.65 | <.0001 |
| ESCS | Index of economic, social and cultural status (WLE) | 1 | 33.61964 | 1.30426 | 25.78 | <.0001 |

| Parameter Estimates | | | | | | |
|---|---|---|---|---|---|---|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | Intercept | 1 | 431.47251 | 2.06961 | 208.48 | <.0001 |
| ESCS | Index of economic, social and cultural status (WLE) | 1 | 12.84425 | 1.95447 | 6.57 | <.0001 |

Dernæst skriver vi gennemsnittene op

National=1: 507,07=486,43+33,62*0,63

National=2: 429,10=431,47+12,84*(-0,12)

Gappet mellem etniske danskerne (national=1) og anden etnicitet (national=2) er

429,10-507,07=-77,97

Dernæst udregner vi gappet mellem intercept og parameter i

Gab= -77,97= (431,47-486,43)+(0,63)*(12,84-33,62) + 12,84(-0,12-0,63) =77,57

-68,051 = uforklaret + -9,630=forklaret =77,57

Dette giver en forklaringsgrad på 12,4%

**2) Inddrag en eller flere variable til yderligere at forklare forskellen og forhåbentligt reducere gabet.**

Vi tilføjer den ekstra variable “antal bøger i huset” som variable i regression via følgende kode:

```
proc reg  data=dk2015;
model score_n=ESCS st013q01ta;
where  national=1;
run;
proc reg  data=dk2015;
model score_n=ESCS st013q01ta;
where  national=2;
Run;
Og finder gennemsnittet af antal bøger i huset
proc means mean data=dk2015;
var st013q01ta;
where national=1;
run;
```

```
proc means mean data=dk2015;
var st013q01ta;
where national=2;
run;
```

Dette giver følgende output:
national=1

| Parameter Estimates | | | | | | |
|---|---|---|---|---|---|---|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | Intercept | 1 | 438.98023 | 2.77771 | 158.04 | <.0001 |
| ESCS | Index of economic, social and cultural status (WLE) | 1 | 20.10662 | 1.43603 | 14.00 | <.0001 |
| ST013Q01TA | How many books are there in your home? | 1 | 16.55143 | 0.84672 | 19.55 | <.0001 |

| Analysis Variable : ST013Q01TA How many books are there in your home? |
|---|
| Mean |
| 3.3963810 |

national=2

| Parameter Estimates | | | | | | |
|---|---|---|---|---|---|---|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | Intercept | 1 | 401.46263 | 4.52993 | 88.62 | <.0001 |
| ESCS | Index of economic, social and cultural status (WLE) | 1 | 6.76821 | 2.08163 | 3.25 | 0.0012 |
| ST013Q01TA | How many books are there in your home? | 1 | 12.71665 | 1.66896 | 7.62 | <.0001 |

| Analysis Variable : ST013Q01TA How many books are there in your home? |
|---|
| Mean |
| 2.3640747 |

Dernæst skriver vi gennemsnittene op

National=1: 507,07=438,98+20,106*0,63+16,551*3,396

National=2: 429,10=401,46+6,77*(-0,12)+12,72*2,364

Gappet mellem etniske danskerne (national=1) og anden etnicitet (national=2) er

429,10-507,07=-77,97

Dernæst udregner vi gappet mellem intercept og parameter i

Gab= -77,97= (401,46-438,98)+(0,63)*(6,77-20,106)+3,396*(12,72-16,551) +

6,77(-0,12-0,63)+12,72(2,364-3,396) =77,57

-58.932 = uforklaret + -18,64 =forklaret =77,57

Dette giver en forklaringsgrad på 24,02%

Det ses at forklaringsgrader steget markant ved tilføjelsen af en ekstra variabel.

**3) Brug en imputeringsteknik til ESCS i første omgang og derefter til evt. andre variable. Giver dette anledning til ændring i din konklusion?**

Vi starter med at give observationer uden en ESCS værdi en værdi vha. SURVEYIMPUTE. Der gøres ved følgende kode hvor vi også bruger proc standard til at fastholde middelværdier:

```
proc surveyimpute seed=100 data=dk2015
method=hotdeck(selection=srswor);
var ESCS;
output out=dk2015a;
run;
proc standard data=dk2015a replace out=dk2015a;
var escs;
Run;
```

Dette giver følgende output:

| Imputation Summary | | |
|---|---|---|
| Observation Status | Number of Observations | Sum of Weights |
| Nonmissing | 6985 | 6985 |
| Missing | 176 | 176 |
| Missing, Imputed | 176 | 176 |
| Missing, Not Imputed | 0 | 0 |

Det ses at 176 observationer manglede en værdi og har fået tildelt en værdi. Dernæst køres samme regression som i opgave 2 blot med det nye datasæt dk2015a

```
proc reg  data=dk2015a;
model score_n=ESCS;
where  national=1;
run;
proc reg  data=dk2015a;
model score_n=ESCS;
where  national=2;
run;
```

Dette giver følgende output:

| Parameter Estimates | | | | | | |
|---|---|---|---|---|---|---|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | Intercept | 1 | 486.08913 | 1.36705 | 355.57 | <.0001 |
| ESCS | Index of economic, social and cultural status (WLE) | 1 | 33.54856 | 1.30738 | 25.66 | <.0001 |

| Parameter Estimates | | | | | | |
|---|---|---|---|---|---|---|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | Intercept | 1 | 430.42425 | 2.03650 | 211.36 | <.0001 |
| ESCS | Index of economic, social and cultural status (WLE) | 1 | 12.12063 | 1.92503 | 6.30 | <.0001 |

Udregningerne er de samme som i opgave 1, hvorfor kun resultatet rapporteres. Den forklarede del af gappet på -78,256 er -9,090 og den uforklarede del er -69,166. Dette giver en forklaringsgrad på 11,6% hvilket er stort set det samme som 12,4% som vi fandt i opgave 1. Dette giver derfor ikke basis for ændring i konklusionen. Det samme gælder hvis man kører flere variable, hvis kigger på parameterværdierne heller ikke her har ændret sig meget.

Nu bruges European Social Survey data.
Brug følgende programstump:

```
proc import datafile='XXXXXXXXX\ESS7e02_1'out=ud7 dbms=sav replace;

Proc means data=ud7 ;
Class cntry=;
Var trstlgl trstplc trstplt trstprl trstprt trstep trstun;
run;
```

De 7 variable betegnes trust variable

**5) Undersøg vha. Cronbach Alpha i hvilke lande om de syv trust variable danner en skala.**

Vi kører følgende kode, som danner Cronbach-Alpha værdier for de syv trust variable pr land.

```
proc corr alpha data=ud7 nocorr nomiss noprob nosimple;
by cntry;
var trstlgl trstplc trstplt trstprl trstprt trstep trstun;
run;
```

Nedenfor er indsat et eksempel for landet Østrig. Vi ser, at Cronbach-Alpha værdien er højere end 0,7 hvilket indikerer at vi kan slå de syv trust variable sammen.

The CORR Procedure

Country=Austria

| 7 Variables: | trstlgl trstplc trstplt trstprl trstprt trstep trstun |
|---|---|

| Cronbach Coefficient Alpha | |
|---|---|
| Variables | Alpha |
| Raw | 0.916897 |
| Standardized | 0.917297 |

Dette mønster gentager sig for alle lande, hvor Cronbach Alpha alle er over 0,7, hvorfor vi kan behandle dem som en og slå dem sammen til en skala, da de ligner hinanden/opfører sig på samme måde.
De forklarer derfor samme mængde som en som spredt ud.

**6) For Danmark. Udfør en principal komponentanalyse disse 7 variable og kommenter konstruktionen af de tre første principiale komponenter.**

Vi kører følgende kode for at få egenværdierne og egenvektorerne

```
proc princomp data=ud7  plots=score(ellipse ncomp=2) out=ud7a
outstat=ud7b;
where cntry='DK';
var trstlgl trstplc trstplt trstprl trstprt trstep trstun;
run;
```

Dette giver følgende output:

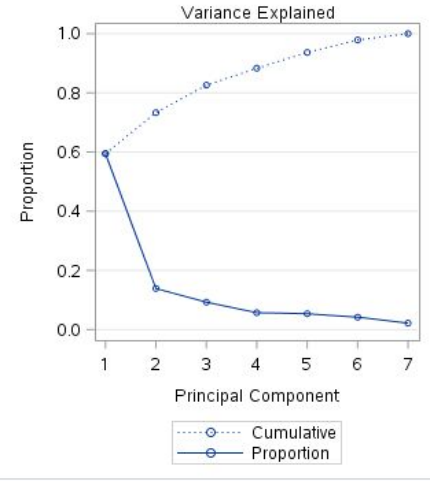
**Eigenvalues of the Correlation Matrix**

| | Eigenvalue | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| 1 | 4.16110564 | 3.19031548 | 0.5944 | 0.5944 |
| 2 | 0.97079016 | 0.32217564 | 0.1387 | 0.7331 |
| 3 | 0.64861452 | 0.24857535 | 0.0927 | 0.8258 |
| 4 | 0.40003917 | 0.02516811 | 0.0571 | 0.8829 |
| 5 | 0.37487105 | 0.08103377 | 0.0536 | 0.9365 |
| 6 | 0.29383728 | 0.14309511 | 0.0420 | 0.9785 |
| 7 | 0.15074217 | | 0.0215 | 1.0000 |

**Eigenvectors**

| | | Prin1 | Prin2 | Prin3 | Prin4 | Prin5 | Prin6 | Prin7 |
|---|---|---|---|---|---|---|---|---|
| trstlgl | Trust in the legal system | 0.356919 | 0.513072 | -.089066 | -.477864 | -.321285 | -.518617 | 0.029864 |
| trstplc | Trust in the police | 0.306546 | 0.649576 | -.256827 | 0.553124 | 0.123078 | 0.311421 | -.006584 |
| trstplt | Trust in politicians | 0.424682 | -.309424 | -.246654 | 0.079705 | 0.201289 | -.213128 | -.755493 |
| trstprl | Trust in country's parliament | 0.406512 | -.165192 | -.249122 | -.514069 | -.067312 | 0.681016 | 0.113215 |
| trstprt | Trust in political parties | 0.412461 | -.337890 | -.229406 | 0.194011 | 0.306188 | -.342397 | 0.643780 |
| trstep | Trust in the European Parliament | 0.380654 | -.241604 | 0.375572 | 0.373455 | -.715699 | 0.059008 | 0.022377 |
| trstun | Trust in the United Nations | 0.343526 | 0.138690 | 0.780721 | -.134059 | 0.480212 | 0.064258 | -.022914 |

Af figuren til højre ses det, at de første tre principale komponenter forklarer mere end 80 pct. Af variationen i y-værdierne.  Dette ses ligeledes af de tre første egenværdier da disse har en høj værdi, som er et mål for deres forklaringskraft.




Det ses af den første principale komponent, at den vægter alle variable cirka lige højt i dannelsen af den nye variabel (Prin1).
Af den anden principale komponent (Prin2) ses det at, denne nye variable vægter primært tillid til retssystemet og politiet højt og tillid til FN lavt. Resten vægtes negativt af varierende grad.

Fra den tredje principale komponent (Prin 3) kan det konkluderes, at tillid til EU og FN vægtes, særligt FN, mens resten vægtes negativt af forskellig grad.

**7) vælg et nyt land og gentag analyserne. Sammenlign analyserne for de to lande.**

Vi kører en lignende kode fra før dog med England:

```
proc princomp data=ud7  plots=score(ellipse ncomp=2) out=ud7a
outstat=ud7b;
where cntry='GB';
var trstlgl trstplc trstplt trstprl trstprt trstep trstun;
run;
```

Dette giver følgende output:

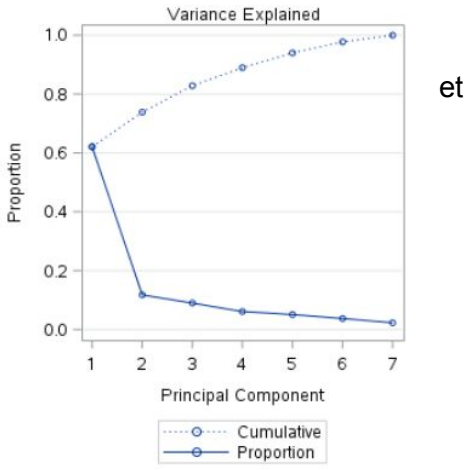
Eigenvalues of the Correlation Matrix

| | Eigenvalue | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| 1 | 4.34437895 | 3.51992282 | 0.6206 | 0.6206 |
| 2 | 0.82445613 | 0.19575818 | 0.1178 | 0.7384 |
| 3 | 0.62869795 | 0.20134836 | 0.0898 | 0.8282 |
| 4 | 0.42734959 | 0.07169248 | 0.0610 | 0.8893 |
| 5 | 0.35565712 | 0.09396332 | 0.0508 | 0.9401 |
| 6 | 0.26169380 | 0.10392733 | 0.0374 | 0.9775 |
| 7 | 0.15776646 | | 0.0225 | 1.0000 |

Eigenvectors

| | | Prin1 | Prin2 | Prin3 | Prin4 | Prin5 | Prin6 | Prin7 |
|---|---|---|---|---|---|---|---|---|
| trstlgl | Trust in the legal system | 0.370507 | 0.472689 | 0.028002 | -.040976 | -.741120 | -.294911 | 0.024405 |
| trstplc | Trust in the police | 0.320937 | 0.698413 | 0.044705 | 0.316990 | 0.527617 | 0.161004 | 0.049355 |
| trstplt | Trust in politicians | 0.422319 | -.188569 | -.358451 | 0.033093 | 0.150482 | -.250994 | -.755555 |
| trstprl | Trust in country's parliament | 0.403225 | -.064840 | -.346198 | -.416310 | -.110211 | 0.715191 | 0.128039 |
| trstprt | Trust in political parties | 0.413636 | -.288595 | -.288377 | 0.036589 | 0.213312 | -.458751 | 0.636524 |
| trstep | Trust in the European Parliament | 0.357129 | -.396028 | 0.377494 | 0.662577 | -.214317 | 0.296866 | 0.007084 |
| trstun | Trust in the United Nations | 0.346620 | -.094321 | 0.723336 | -.532055 | 0.214795 | -.118437 | -.067060 |

Af figuren til højre ses det, at de første tre principale komponenter i dette tilfælde også forklarer mere end 82 pct. af variationen i y-værdierne. Dette ses ligeledes af de tre første egenværdier da disse har en høj værdi, som er et mål for deres forklaringskraft.




Det ses af den første principale komponent, at den vægter alle variable cirka lige højt i dannelsen af den nye variabel (Prin1).
Af den anden principale komponent (Prin2) ses det at, denne nye variable vægter primært tillid til retssystemet

og politiet højt og tillid til FN lavt. Resten vægtes negativt af varierende grad.
Fra den tredje principale komponent (Prin3) kan det konkluderes, at tillid til EU og FN vægtes, særligt FN, mens resten vægtes negativt af forskellig grad, på nær tillid til retssystemet og politiet næsten ikke vægtes

De to analyser (DK og GB) er nærmest identiske, da de foreslåede principale komponenter varierer på næsten samme måde. Dette er ikke overraskende da de to lande har ens værdisæt og ens geografisk placering. Dette kunne tænkes at analysen blev anderledes hvis man valgte et land med markant anderledes værdisæt og geografisk placering end Danmark.