

Sommer 2011

Programmering og Anvendt Statistik med SAS

Rettevejledning

1)

Det korte svar er b) fordi output sætningen er uden for do løkken. I a) og c) er outputsætningen inden i do løkken, hvorved der genereres fem observationer. I øvrigt er output sætningen i b) overflødig, da der en imbedded output-statement før run-statementet

2)

Det er et umådeligt kedeligt program

*følgende linje definerer navnet på det datasæt, der skal skrives til.

Eksisterer det i forvejen, vil det blive overskrevet;

data aa;

*følgende er en DO-sætning, der siger, at for værdier for i, der går fra 1 til

50

skal det efterfølgende programstykke (indtil end) udføres;

do i=1 to 50;

*x sættes lig med i. Dermed kommer der en kolonne i datasættet med x-værdier fra 1-50

(i overenstemmelse med værdierne for i/en kopi af data for i);

x=i;

*endnu en kolonne defineres med navnet y. Værdierne for y er simuleret ved udtrykket

$1 + (2 \text{ gange værdien af } x) + \text{en tilfældig værdi fra en uniform distribution over intervallet } [0:1];$

*Tallet 1234 er seed for generatoren af tilfældige tal. den sikrer at der opnås de samme talværdier hver gang programmet afvikles, men da kaldet af ranuni er inde i et stort datastepp bliver de genererede tal i selve datasteppet selvfølgelig "tilfældige";

y=1+2*x+ranuni(1234);

*Følgende sikrer, at alle observationer/beregninger inden for DO-løkken skrives til datasættet

inden DO afsluttes med end;

output;

end;

*Run sikrer at det hidtige programstykke eksekveres;

RUN;

*makroen gives navnet m, og vil efter kørsel kunne kaldes fra makro kataloget "Sasmacr";

%macro m;

*for ii-værdierne 5 til 50 skal det efterfølgende program køres indtil end statement;

```

%do ii=5 %to 50;
*følgende kalder en generel regressionsprocedure, som skal køres for datasættet
aa;
proc reg data=aa;
*regressionsmodellen specificeres;
model y=x;
*forudsætning for udvælgelse af observationer er, at i er mindre end den af
makroen dannede ii,
dvs. i ligger indenfor intervallet [4:49];

*Der udføres altså mange regressioner med fra 4 til 49 observationer;

where i<&ii;
*Run sikrer at det hidtidige programstykke eksekveres;
run;
*makroens DO-løkke afsluttes;
%end;
*makroen afsluttes;
%mend;

```

3)

I opgaven bruges successiv summation i linien k+i; Formlens højreside er et helt elementært regneudtryk.

```

data b;
do i=1 to 10;
k+i;
sum=(i+1)*i/2;
output;
end;
run;
proc print;
run;

```

4)

Det korte svar er, at det skal være c), fordi det kun er i dette print, at en variable hedder k. Output b) er dannet med by i; og a) er dannet helt uden by-statement.

5)

En kort besvarelse til første deler programmet

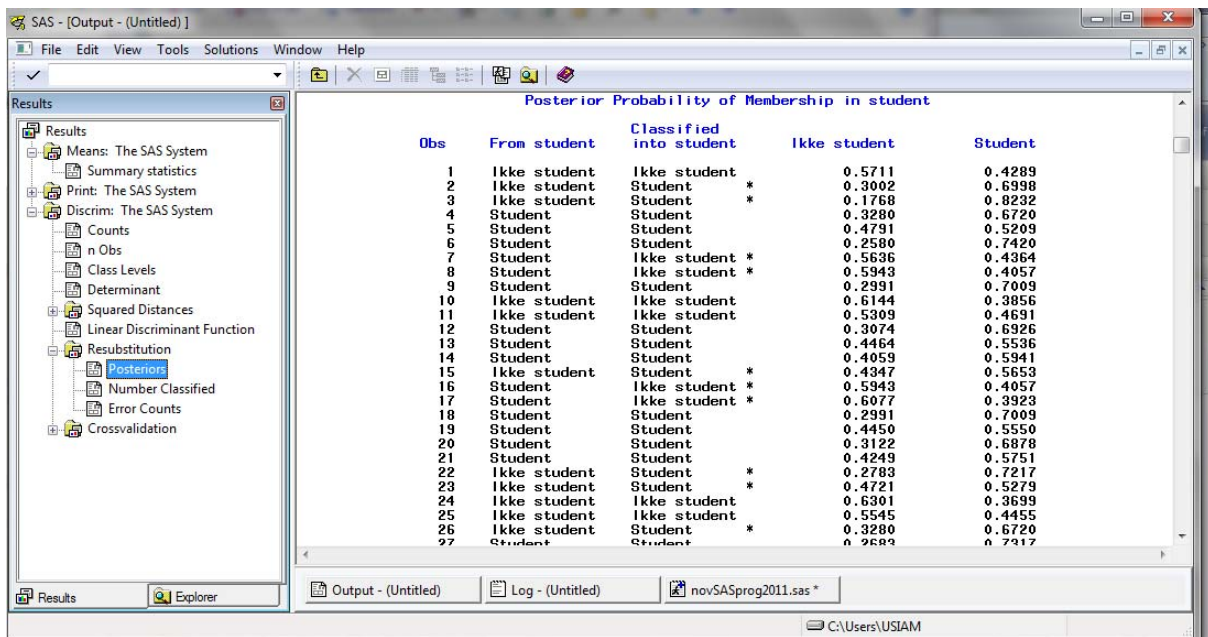
```

proc discrim data=sasprog.stamdata outstat=bstat method=normal pool=yes
list crossvalidate;
class student;
*priors prop;
var V204 V179 V154 V131 V57;;

```

run;

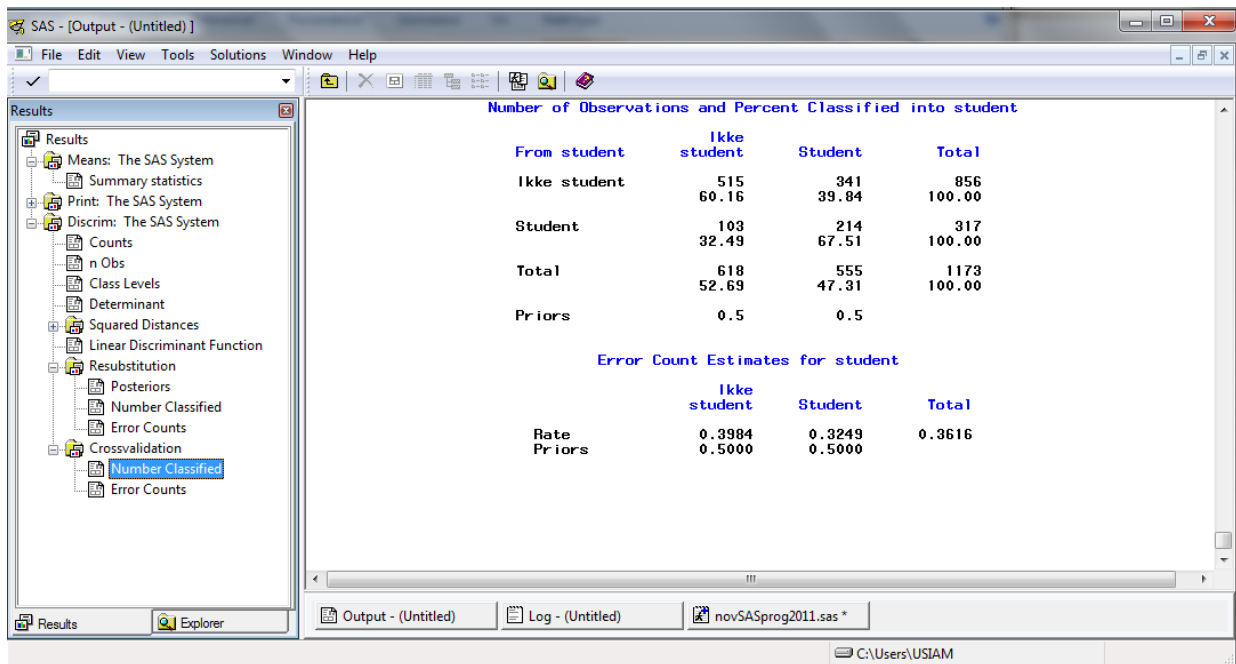
Outputtet indeholder forskellige tabeller, der viser resultatet af diskriminansanalysen. I tabellen over "posteriors" angives sandsynligheden for at respondenter har studentereksamen for hver enkelt observation.



The screenshot shows the SAS Output window with a tree view on the left and a table of results on the right. The table is titled "Posterior Probability of Membership in student" and contains 27 rows of data. The columns are: Obs, From student, Classified into student, Ikke student, and Student. The 'Student' column shows the posterior probability for each observation.

Obs	From student	Classified into student	Ikke student	Student
1	Ikke student	Ikke student	0.5711	0.4289
2	Ikke student	Student *	0.3002	0.6998
3	Ikke student	Student *	0.1768	0.8232
4	Student	Student	0.3280	0.6720
5	Student	Student	0.4791	0.5209
6	Student	Student	0.2580	0.7420
7	Student	Ikke student *	0.5836	0.4364
8	Student	Ikke student *	0.5943	0.4057
9	Student	Student	0.2991	0.7009
10	Ikke student	Ikke student	0.6144	0.3856
11	Ikke student	Ikke student	0.5309	0.4691
12	Student	Student	0.3074	0.6926
13	Student	Student	0.4464	0.5536
14	Student	Student	0.4059	0.5941
15	Ikke student	Student *	0.4347	0.5653
16	Student	Ikke student *	0.5943	0.4057
17	Student	Ikke student *	0.6077	0.3923
18	Student	Student	0.2991	0.7009
19	Student	Student	0.4450	0.5550
20	Student	Student	0.3122	0.6878
21	Student	Student	0.4249	0.5751
22	Ikke student	Student *	0.2783	0.7217
23	Ikke student	Student *	0.4721	0.5279
24	Ikke student	Ikke student	0.6301	0.3699
25	Ikke student	Ikke student	0.5545	0.4455
26	Ikke student	Student *	0.3280	0.6720
27	Student	Student	0.2683	0.7317

En optælling af cross classifications giver



Der er for mange, der klassificeres som student i forhold til hvor mange, der i virkeligheden er i datamaterialet.

I programmet ovenfor gemmes klassifikationsreglen i datasættet bstat. Ved hjælp af disse klassifikationsregler kan respondenterne i datasættet testdata klassificeres ved programmet

```
proc discrim data=bstat testdata=sasprog.testdata testout=tout testlist;
class student;
var V204 V179 V154 V131 V57;
run;
```

Det er svært at sige noget fornuftigt om resultaterne. Der er for mange, der klassificeres som studenter. Diskriminansreglerne er ikke en del af pensum.

The screenshot shows the SAS Results window with a tree view on the left containing 'Results', 'Discrim: The SAS System', 'TEST= Classification', 'Posteriors', 'n Obs', 'Number Classified', and 'Error Counts'. The main window displays two tables:

Number of Observations and Percent Classified into student

	Ikke student	Student	Total
From student			
Ikke student	172	105	277
	62.09	37.91	100.00
Student	37	77	114
	32.46	67.54	100.00
Total	209	182	391
	53.45	46.55	100.00
Priors	0.5	0.5	

Error Count Estimates for student

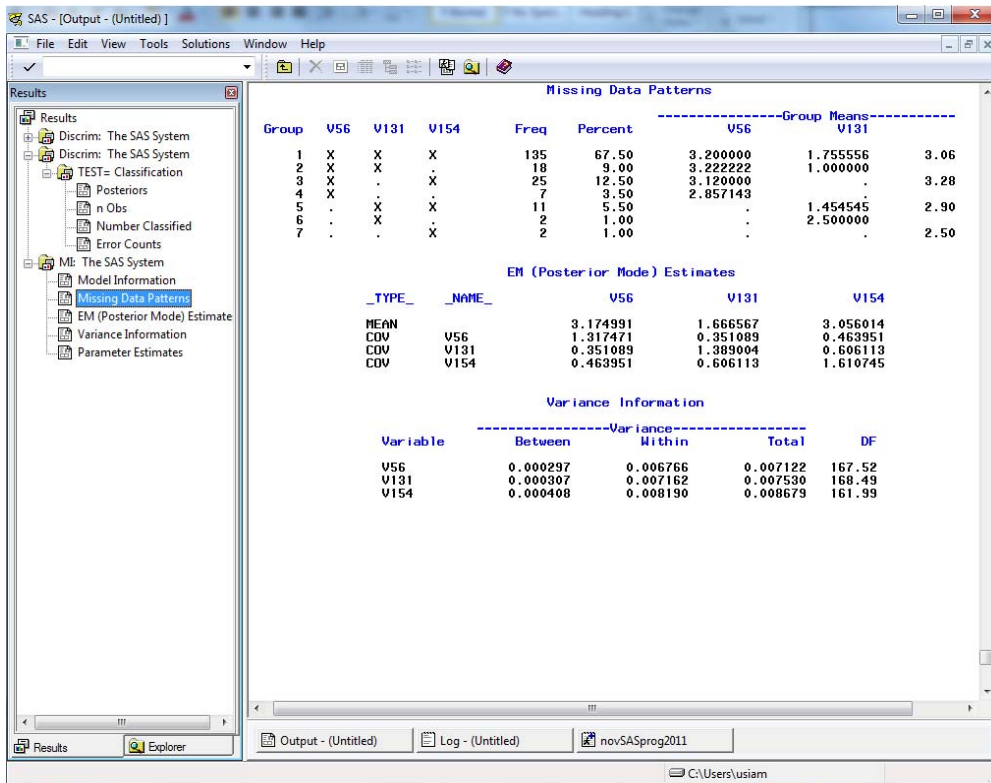
	Ikke student	Student	Total
Rate	0.3791	0.3246	0.3518
Priors	0.5000	0.5000	

6)

```
Proc mi data=sasprog.miss mu0=3 3 3 out=outmi;;
var v56 v131 v154;
run;
```

Missing data pattern viser, at der ikke er tale om monotont manglende værdier, de ruden videre kan fyldes ud ved successive regressioner. Der anvendes derfor MCMC metoden pr default. Det synes meningsløst at fylde op til monotont missing med MCMC og så fortsætte med regressionsmetoder. Der mangler simpelthen for meget.

Ud fra varianserne kan man se at within variansen er ca 25 gange større end between varianser. det betyder at langt hovedparten af den statistiske variation skyldes variationen i data og IKKE variationen mellem de fem imputationer, so pr default dannes. Antal imputationer kan sættes op, men praksis siger, at det er uden betydning. Desuden er det tidsspilde på de langsomme servere i eksamenslokalet.



Testene for hypoteserne om at middelværdierne er 3 dannes ved optionen $\mu_0=3$ i procedurekaldet. Det ser ud som om der ses mindre video (klart mindre end 3) end der haves gæster (en anelse over 3) og hører musik (stort set lig med 3) (eller også er skalaen vendt om, men det er uden betydning for opgaven)

Parameter Estimates

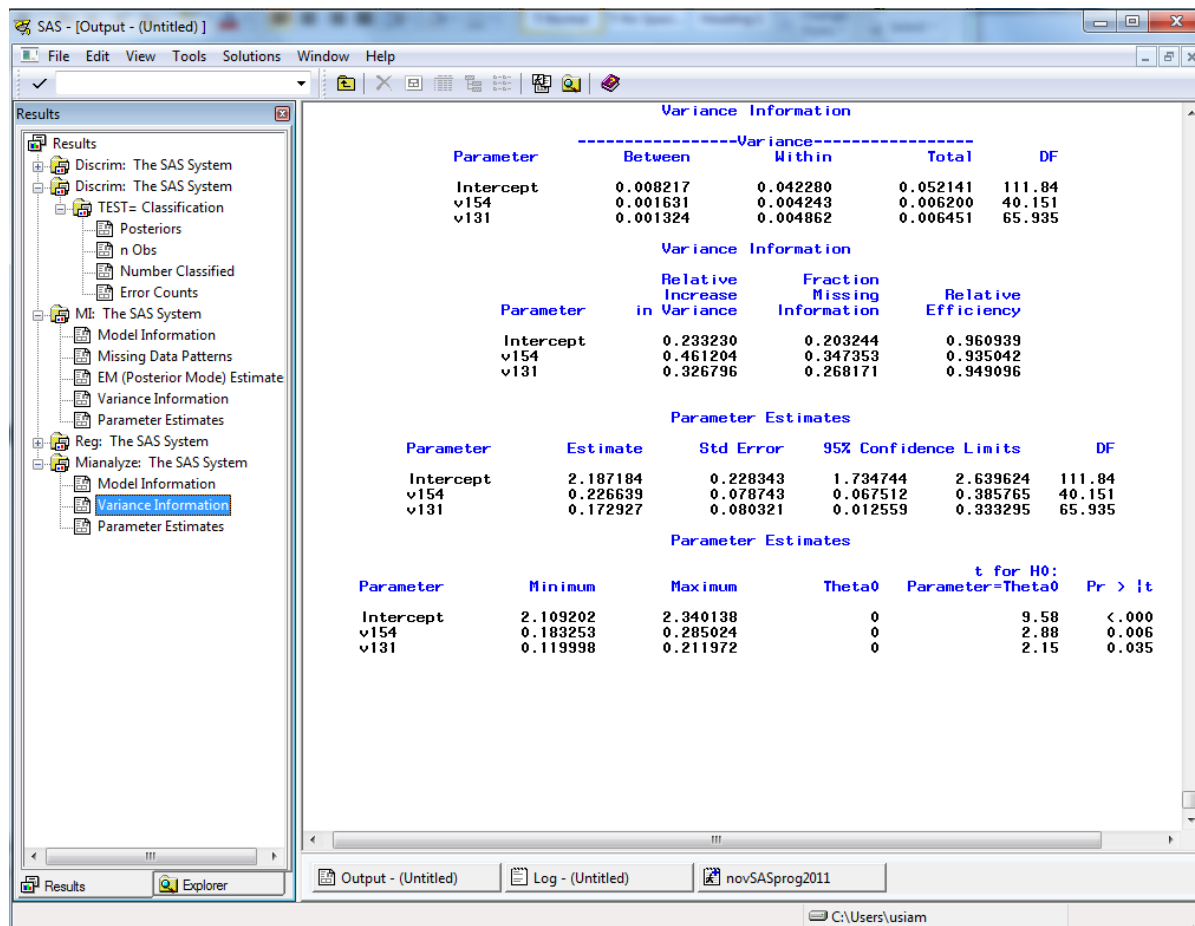
Variable	Minimum	Maximum	Mu0	t for H0:	
				Mean=Mu0	Pr > t
V56	3.143628	3.187296	3.000000	2.03	0.0437
V131	1.629444	1.675840	3.000000	-15.47	<.0001
V154	3.053419	3.108022	3.000000	0.83	0.4065

Det sidste delspørgsmål besvares ved først et kald af proc reg (her er det meget vigtigt at linien by `_imputation_;` er med for ellers regnes på et datasæt med fem gang esp mange observationer, som det oprindelige). Dernæst kaldes proc MIanalyze.

```
proc reg data=outmi outest=outreg covout;
model v56=v154 v131;
by _Imputation_;
run;

proc mianalyze data=outreg;
modeleffects Intercept v154 v131;
test v154=v131=0/mult;
run;
```

I outputtet fra Mianalyze ses på de tre regressionsestimater. Koefficienten til de to hældningskoefficienter er klart forskellige fra nul, så et simultant test synes overflødig. Det udføres dog alligevel med test statementet med en klar forcastelse af, at de to forklarende variable kan udelades på en gang til følge. Om man har gæster afhænger altså positivt af, om man ser video og hører musik (man hører altså musik og ser video med sine gæster - eller hvad?)



SAS - [Output - (Untitled)]

File Edit View Tools Solutions Window Help

10:10 Tuesday, May 31, 2011

The SAS System

The MIANALYZE Procedure

Test: Test 1

Multivariate Inference

Assuming Proportionality of Between/Within Covariance Matrices

Avg Relative Increase in Variance	Num DF	Den DF	F for H0: Parameter=Theta0	Pr > F
0.344686	2	44.345	10.64	0.0002

Results

- Results
 - Discrim: The SAS System
 - Discrim: The SAS System
 - TEST= Classification
 - Posteriors
 - n Obs
 - Number Classified
 - Error Counts
 - MI: The SAS System
 - Model Information
 - Missing Data Patterns
 - EM (Posterior Mode) Estimate
 - Variance Information
 - Parameter Estimates
 - Reg: The SAS System
 - Mianalyze: The SAS System
 - Model Information
 - Variance Information
 - Parameter Estimates
 - Mianalyze: The SAS System
 - Model Information
 - Variance Information
 - Parameter Estimates
 - Test 1
 - Test Specification
 - Variance Information
 - Parameter Estimates
 - Multivariate Inference

Output - (Untitled) Log - (Untitled) novSASprog2011 *

C:\Users\usiam