Multivariate analysis of educational data

Contents

1. Introduction

The data set used in this section forms part of the data library of the Multilevel Project at the University of London and comes from the Junior School Project (Mortimer, *et al.*, 1988).

Mathematics and language tests were administered in three consecutive years to more than 1000 students from 50 primary schools, which were randomly selected from primary schools maintained by the Inner London Education Authority.

The following variables are available in the data file **JSP.PSF**:

SCHOOL	School code (1 to 50)
CLASS	Class code (1 to 4)
STUDENT	Student ID (1 to 1402)
GENDER	Gender (boy=1; girl=0)
RAVENS	Ravens test score in year 1(score 4–36)
MATH1	Score on mathematics test in year 1 (score 1–40)
MATH2	Score on mathematics test in year 2 (score 1–40)
MATH3	Score on mathematics test in year 3 (score 1–40)
ENG1	Score on language test in year 1 (score 0–98)
ENG2	Score on language test in year 2 (score 0–98)
ENG3	Score on language test in year 3 (score 0–98)
CONSTANT	Intercept, value=1 throughout

The school number (SCHOOL) is used as the level-3 identification variable, with Student ID (STUDENT) as the level-2 identification variable. The level1 units are the scores obtained by a student for the mathematics and language tests, as represented by MATH1 to MATH3 and ENG1 to ENG3.

The aim of this analysis is to examine the variation in test scores over pupils. It would also be interesting to determine the extent to which schools vary with respect to the response variable(s). One of the main benefits of analyzing different responses simultaneously in one multivariate analysis is that the way in which measurements relate to the explanatory variables can be directly explored. The gender and Ravens test score (GENDER and RAVENS) will be used as explanatory variables in the second of the models described in this section.

As the residual covariance matrices for the second and third levels of the hierarchy are also obtained from this analysis, differences between coefficients of explanatory variables for different responses can be studied.

Finally, each respondent does not have to be measured on each response, as is the case for the data set we consider here, where scores for all three years are not available for all students. This type of model is one in which the MISSING_DEP and MISSING_DAT commands can be used to good advantage, as will be shown. In the case of missing responses, a multilevel multivariate analysis of the responses that are available for respondents can be used to provide information in the estimation of those that are missing.

The first twenty observations in the **JSP.LSF** file are shown below.

JSP.lsf

	school	student	gender	ravens	math1	math2	math3	eng1	eng2	eng3	constant
1	1.00	1.00	0.00	23.00	23.00	24.00	23.00	72.00	80.00	39.00	1.00
2	1.00	2.00	1.00	15.00	14.00	11.00	-9.00	7.00	17.00	-9.00	1.00
3	1.00	3.00	1.00	22.00	36.00	32.00	39.00	88.00	89.00	83.00	1.00
4	1.00	4.00	1.00	14.00	24.00	26.00	32.00	12.00	25.00	12.00	1.00
5	1.00	5.00	0.00	19.00	22.00	23.00	-9.00	67.00	78.00	-9.00	1.00
6	1.00	6.00	0.00	16.00	19.00	23.00	11.00	52.00	76.00	19.00	1.00
7	1.00	7.00	1.00	17.00	22.00	22.00	26.00	37.00	68.00	31.00	1.00
8	1.00	8.00	0.00	21.00	18.00	29.00	28.00	57.00	86.00	40.00	1.00
9	1.00	9.00	1.00	30.00	30.00	31.00	-9.00	42.00	59.00	-9.00	1.00
10	1.00	10.00	0.00	25.00	29.00	29.00	-9.00	46.00	79.00	-9.00	1.00
11	1.00	11.00	0.00	32.00	31.00	28.00	32.00	69.00	84.00	50.00	1.00
12	1.00	12.00	0.00	15.00	18.00	26.00	-9.00	54.00	74.00	-9.00	1.00
13	1.00	13.00	0.00	25.00	23.00	-9.00	27.00	63.00	-9.00	39.00	1.00
14	1.00	14.00	0.00	29.00	39.00	35.00	36.00	83.00	88.00	80.00	1.00
15	1.00	15.00	0.00	34.00	24.00	30.00	33.00	37.00	44.00	37.00	1.00
16	1.00	16.00	1.00	27.00	25.00	29.00	33.00	15.00	20.00	4.00	1.00
17	1.00	17.00	1.00	21.00	27.00	30.00	30.00	43.00	55.00	29.00	1.00
18	1.00	19.00	0.00	17.00	14.00	10.00	25.00	52.00	84.00	46.00	1.00
19	1.00	20.00	1.00	22.00	14.00	20.00	17.00	11.00	23.00	9.00	1.00
20	1.00	21.00	0.00	15.00	16.00	15.00	18.00	27.00	50.00	15.00	1.00
	1.00	22.00	1.00	25.00	30.00	29.00	33.00	79.00	87.00	42.00	1.00

One line of information is given for each student. Note that, for the second and fourth students, no mathematics score was available in the third year of the study (MATH3). A missing data code of −9 was assigned to all missing values on both explanatory and response variables in the data set.

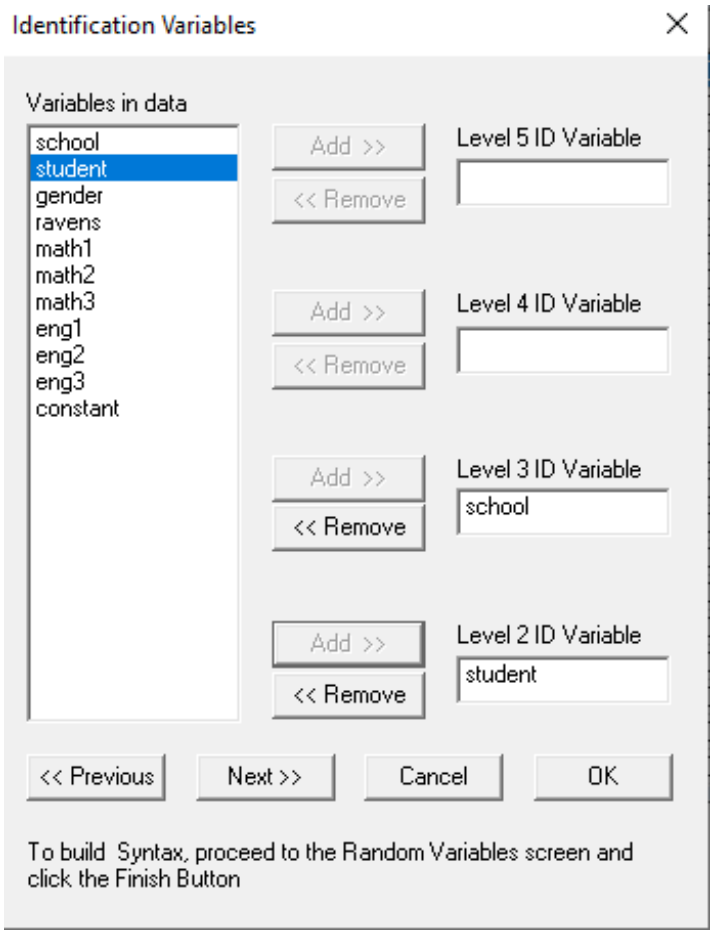
In order to perform a multilevel analysis, we need one line of information for each level-1 unit, in this case each of the six test scores. The data manipulation required in creating this revised data file format will be performed automatically by the program in the case of a multivariate model. Six dummy variables are created for each of the explanatory variables used. For the explanatory variable GENDER, for example, the dummy variables GENDER1, GENDER2, ..., GENDER6 will denote the gender effect for each of the response variables.

Two models will be fitted and discussed:

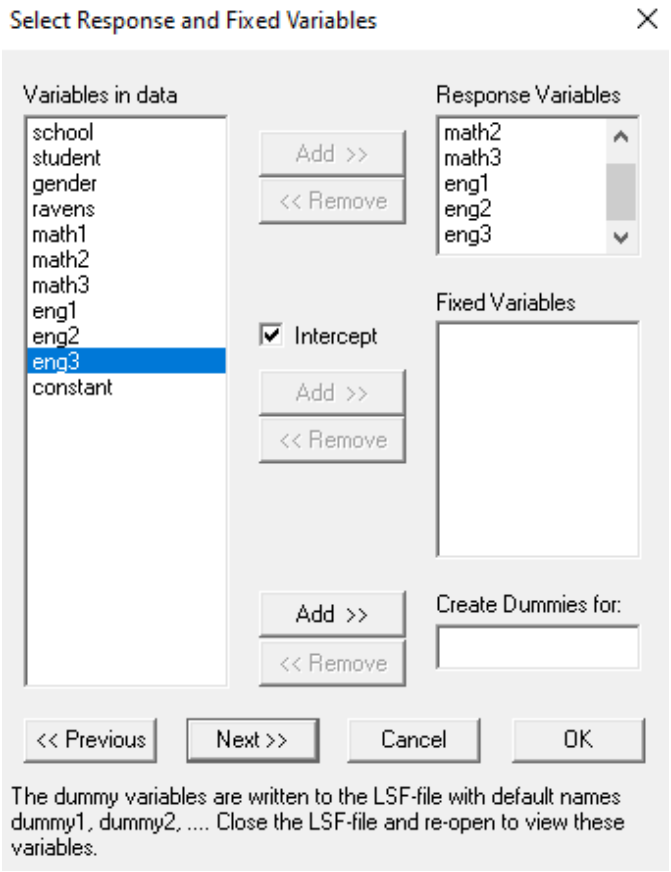
- A variance decomposition model
- Adding explanatory variables to the model

2. A variance decomposition model

In the case of a multivariate model, only levels 2 and 3 of the hierarchy have to be identified, because the actual measurements for each level-2 unit serve as level-1 units. In this case, we select SCHOOL as the level-3 identification variable. The student ID (STUDENT) is selected as the level-2 identification variable.



We choose the six variables MATH1 to MATH3 and ENG1 to ENG3 as response variables.



In this model, we wish to start our analysis of the data with a look at the differences in intercepts over students and schools. Note that in this analysis, no RANDOM commands should be included. These commands are only required when running a multivariate model when more variables than just the intercept is required at higher levels of the hierarchy. If no RANDOM commands are included, the inclusion of an intercept term on higher levels is automatically assumed by the program and dummy variables for the explanatory predictors given in the FIXED command are generated.

The resulting input file (**JSP1.PRL**) is given below.

Note, that the optional MISSING_DEP command is used to identify -9 as the missing data code for all response variables.

```
OPTIONS Maxiter=30 Conv=0.0001 CovBW=Yes ACM=Yes;
TITLE=Multivariate Analysis of Education Data;
SY=JSP.LSF;
ID2=student;
ID3=school;
RESPONSE=math1 math2 math3 eng1 eng2 eng3;
FIXED=intcept;
MISSING_DEP=-9;
```

The output for this model is written to the default output file **JSP1.OUT**. Partial output is given below.

```
                        +--------------+
                        | DATA SUMMARY |
                        +--------------+

 NUMBER OF LEVEL 3 UNITS :       49
 NUMBER OF LEVEL 2 UNITS :     1192
 NUMBER OF LEVEL 1 UNITS :     6472
```

ID3	:	1	2	3	4	5	6	7	8
N2	:	34	13	21	24	29	24	15	31
N1	:	184	72	96	144	166	120	78	174
ID3	:	9	10	11	12	13	14	15	16
N2	:	22	14	12	28	25	15	24	19
N1	:	130	48	70	154	138	86	106	106
ID3	:	17	18	19	20	21	22	23	24
N2	:	7	20	17	15	32	16	20	21
N1	:	40	110	88	84	184	92	116	126
ID3	:	25	26	27	28	29	30	31	32
N2	:	18	29	27	18	25	37	36	26
N1	:	94	156	146	96	124	182	214	148
ID3	:	33	34	35	36	37	38	39	40
N2	:	46	34	19	32	22	14	16	13
N1	:	262	184	106	178	118	78	90	68
ID3	:	41	42	43	44	45	46	47	48
N2	:	13	47	6	17	19	37	72	25
N1	:	70	262	32	88	102	204	412	144
ID3	:	49							
N2	:	46							
N1	:	202							

Adjusted between cluster sample size= 24
Within cluster sample size= 1192

ITERATION NUMBER 8

+-----------------------+
| FIXED PART OF MODEL |
+-----------------------+

COEFFICIENTS	BETA-HAT	STD.ERR.	Z-VALUE	PR > |Z|
intcept1	24.90370	0.33546	74.23792	0.00000
intcept2	24.87234	0.40108	62.01311	0.00000
intcept3	30.04909	0.37761	79.57736	0.00000
intcept4	47.15338	1.32158	35.67945	0.00000
intcept5	64.96594	1.22017	53.24352	0.00000
intcept6	40.71988	1.38296	29.44399	0.00000

+-----------------------+
| -2 LOG-LIKELIHOOD |
+-----------------------+

DEVIANCE= -2*LOG(LIKELIHOOD) = 45991.2578958991
NUMBER OF FREE PARAMETERS = 48

+----------------------+
| RANDOM PART OF MODEL |
+----------------------+

LEVEL 3		TAU-HAT	STD.ERR.	Z-VALUE	PR > |Z|
math1	/math1	3.31028	1.10053	3.00791	0.00263
math2	/math1	2.29203	1.09364	2.09579	0.03610
math2	/math2	5.23047	1.57789	3.31486	0.00092
math3	/math1	2.36877	1.02957	2.30075	0.02141
math3	/math2	3.13485	1.25421	2.49945	0.01244
math3	/math3	4.79640	1.39721	3.43285	0.00060
eng1	/math1	9.95162	3.74223	2.65927	0.00783
eng1	/math2	9.49602	4.22124	2.24958	0.02448
eng1	/math3	9.95900	4.00265	2.48810	0.01284
eng1	/eng1	59.32680	17.15964	3.45735	0.00055
eng2	/math1	9.93290	3.47650	2.85716	0.00427
eng2	/math2	11.41833	4.08446	2.79555	0.00518
eng2	/math3	11.59906	3.87949	2.98984	0.00279
eng2	/eng1	42.49228	14.15024	3.00294	0.00267
eng2	/eng2	53.08595	14.64628	3.62453	0.00029
eng3	/math1	10.16830	3.82941	2.65532	0.00792
eng3	/math2	10.71620	4.43202	2.41790	0.01561
eng3	/math3	13.71315	4.43882	3.08937	0.00201
eng3	/eng1	45.00410	15.55552	2.89313	0.00381
eng3	/eng2	51.61419	15.11775	3.41415	0.00064
eng3	/eng3	71.08796	18.81475	3.77831	0.00016

LEVEL 2		TAU-HAT	STD.ERR.	Z-VALUE	PR > |Z|
math1	/math1	47.17671	1.99173	23.68631	0.00000
math2	/math1	38.63798	1.91420	20.18492	0.00000
math2	/math2	55.45831	2.35660	23.53314	0.00000
math3	/math1	31.21276	1.65633	18.84448	0.00000
math3	/math2	36.54802	1.84653	19.79278	0.00000
math3	/math3	41.33874	1.86018	22.22293	0.00000
eng1	/math1	109.03175	5.79022	18.83032	0.00000
eng1	/math2	109.44780	6.15532	17.78100	0.00000
eng1	/math3	88.00030	5.36198	16.41191	0.00000
eng1	/eng1	549.41232	23.15761	23.72492	0.00000
eng2	/math1	88.73168	4.92512	18.01613	0.00000
eng2	/math2	94.99391	5.32980	17.82315	0.00000
eng2	/math3	79.52971	4.69229	16.94901	0.00000
eng2	/eng1	388.36174	18.26625	21.26117	0.00000
eng2	/eng2	409.27297	17.36258	23.57213	0.00000
eng3	/math1	94.37662	5.19417	18.16973	0.00000
eng3	/math2	99.80683	5.60600	17.80356	0.00000
eng3	/math3	86.38268	4.95170	17.44505	0.00000
eng3	/eng1	382.76743	18.73683	20.42861	0.00000
eng3	/eng2	317.45309	15.98496	19.85949	0.00000
eng3	/eng3	425.49404	18.97401	22.42510	0.00000

LEVEL 3 COVARIANCE MATRIX

	math1	math2	math3	eng1	eng2
math1	3.31028				
math2	2.29203	5.23047			
math3	2.36877	3.13485	4.79640		
eng1	9.95162	9.49602	9.95900	59.32680	
eng2	9.93290	11.41833	11.59906	42.49228	53.08595
eng3	10.16830	10.71620	13.71315	45.00410	51.61419

	eng3
eng3	71.08796

LEVEL 3 CORRELATION MATRIX

	math1	math2	math3	eng1	eng2	eng3
math1	1.0000					
math2	0.5508	1.0000				
math3	0.5945	0.6259	1.0000			
eng1	0.7101	0.5391	0.5904	1.0000		
eng2	0.7493	0.6852	0.7269	0.7572	1.0000	
eng3	0.6629	0.5557	0.7426	0.6930	0.8402	1.0000

LEVEL 2 COVARIANCE MATRIX

	math1	math2	math3	eng1	eng2
math1	47.17671				
math2	38.63798	55.45831			
math3	31.21276	36.54802	41.33874		
eng1	109.03175	109.44780	88.00030	549.41232	
eng2	88.73168	94.99391	79.52971	388.36174	409.27297
eng3	94.37662	99.80683	86.38268	382.76743	317.45309

	eng3
eng3	425.49404

LEVEL 2 CORRELATION MATRIX

	math1	math2	math3	eng1	eng2	eng3
math1	1.0000					
math2	0.7554	1.0000				
math3	0.7068	0.7633	1.0000			
eng1	0.6772	0.6270	0.5839	1.0000		
eng2	0.6386	0.6305	0.6114	0.8190	1.0000	
eng3	0.6661	0.6497	0.6513	0.7917	0.7607	1.0000

- From the Data Summary, we see that data from 1192 students from 49 schools were used in the analysis. The number of level-1 units (*i.e.*, measurements) per school ranged from 32 in the case of school number 43 to 412 for school number 47.
- From the output for the fixed part of the model, it can be seen that all six fixed effects are highly significant, indicating significant differences in the six measurements over the students.
- There is significant variation in the mean effects of the six response variables over both schools and students. The variation over schools (level 3) is higher than over students. At the student level, the largest variation is in the mean effects for the language tests, ranging between 59.326 (for measurements from the first year), to 71.088 (for measurements from the third year of the study). The same tendency is observed for the mean effects of language test scores over schools, with the highest variation recorded for the language test score from the third year of the study.
- The $-2\ln L$ value for this analysis at convergence after 7 iterations was 45991.2579.

3. Adding explanatory variables to the model

Using the model discussed in the previous section as point of departure, we now proceed to add fixed effects to the model. The variables GENDER and RAVENS, indicating the gender of a student and the student's score on the Ravens test in the first year of the study, respectively, are added to the FIXED command.

This is how the FIXED command should look:

```
FIXED=intcept gender ravens;
```

To make this change, the input file previously used (**JSP1.PRL**) can be edited directly, and then saved as the new input file **JSP2.PRL**:

```
OPTIONS Maxiter=30 Conv=0.0001 Summary=None Output=RAWDATA;
TITLE=Multivariate Analysis of Education Data, added fixed effects;
SY=JSP.LSF;
ID2=student;
ID3=school;
RESPONSE=math1 math2 math3 eng1 eng2 eng3;
FIXED=intcept gender ravens;
MISSING_DEP=-9;
```

For this model the following output was obtained.

```
ITERATION NUMBER     7


                      +----------------------+
                      | FIXED PART OF MODEL  |
                      +----------------------+


-------------------------------------------------------------------------------
  COEFFICIENTS             BETA-HAT      STD.ERR.      Z-VALUE       PR > |Z|
-------------------------------------------------------------------------------
  intcept1                  7.68926       0.80046       9.60607       0.00000
  intcept2                  6.48252       0.90269       7.18136       0.00000
  intcept3                 14.84541       0.84586      17.55061       0.00000
  intcept4                  2.64745       2.90410       0.91163       0.36197
  intcept5                 28.95062       2.57309      11.25128       0.00000
  intcept6                 -2.58204       2.74794      -0.93963       0.34741
  gender1                  -0.48513       0.33815      -1.43466       0.15138
```

gender2	-0.79986	0.37167	-2.15205	0.03139
gender3	-0.45790	0.34277	-1.33588	0.18159
gender4	-10.57862	1.20875	-8.75172	0.00000
gender5	-9.21809	1.06819	-8.62960	0.00000
gender6	-7.02657	1.10494	-6.35926	0.00000
ravens1	0.69639	0.02943	23.66040	0.00000
ravens2	0.74945	0.03248	23.07353	0.00000
ravens3	0.61505	0.03029	20.30821	0.00000
ravens4	1.97944	0.10546	18.77035	0.00000
ravens5	1.61390	0.09328	17.30225	0.00000
ravens6	1.86293	0.09754	19.09957	0.00000

+----------------------+
| -2 LOG-LIKELIHOOD |
+----------------------+

DEVIANCE= -2*LOG(LIKELIHOOD) = 45321.6134304450
NUMBER OF FREE PARAMETERS = 60

+----------------------+
| RANDOM PART OF MODEL |
+----------------------+

LEVEL 3		TAU-HAT	STD.ERR.	Z-VALUE	PR > |Z|
math1	/math1	2.35970	0.77231	3.05538	0.00225
math2	/math1	1.47571	0.79018	1.86756	0.06182
math2	/math2	4.59531	1.29317	3.55353	0.00038
math3	/math1	1.45284	0.73222	1.98414	0.04724
math3	/math2	2.34726	0.97705	2.40240	0.01629
math3	/math3	3.94231	1.11612	3.53215	0.00041
eng1	/math1	5.89266	2.50131	2.35583	0.01848
eng1	/math2	5.10887	3.00832	1.69825	0.08946
eng1	/math3	5.75844	2.83649	2.03013	0.04234
eng1	/eng1	40.46361	11.99950	3.37211	0.00075
eng2	/math1	5.42374	2.23795	2.42353	0.01537
eng2	/math2	6.60379	2.82875	2.33453	0.01957
eng2	/math3	7.26912	2.69805	2.69422	0.00706
eng2	/eng1	22.52145	9.10418	2.47375	0.01337
eng2	/eng2	33.03809	9.66632	3.41786	0.00063
eng3	/math1	5.14461	2.49143	2.06492	0.03893
eng3	/math2	5.67697	3.12932	1.81413	0.06966
eng3	/math3	9.01133	3.15762	2.85384	0.00432
eng3	/eng1	22.93497	10.12809	2.26449	0.02354
eng3	/eng2	30.00480	9.77962	3.06810	0.00215
eng3	/eng3	47.71171	12.97810	3.67632	0.00024

LEVEL 2		TAU-HAT	STD.ERR.	Z-VALUE	PR > |Z|
math1	/math1	31.84368	1.34833	23.61706	0.00000
math2	/math1	22.14572	1.23554	17.92388	0.00000
math2	/math2	37.70830	1.61030	23.41693	0.00000
math3	/math1	17.66778	1.10211	16.03093	0.00000

math3	/math2	21.92750	1.24318	17.63828	0.00000
math3	/math3	29.34575	1.34391	21.83617	0.00000
eng1	/math1	65.59655	3.92248	16.72322	0.00000
eng1	/math2	62.26680	4.16530	14.94895	0.00000
eng1	/math3	49.65248	3.75896	13.20909	0.00000
eng1	/eng1	406.75188	17.17856	23.67788	0.00000
eng2	/math1	53.71682	3.40837	15.76025	0.00000
eng2	/math2	56.85501	3.69230	15.39826	0.00000
eng2	/math3	48.64973	3.37484	14.41540	0.00000
eng2	/eng1	272.02299	13.42728	20.25897	0.00000
eng2	/eng2	314.33951	13.37559	23.50099	0.00000
eng3	/math1	53.74137	3.50517	15.33203	0.00000
eng3	/math2	55.52744	3.78879	14.65570	0.00000
eng3	/math3	50.44039	3.44613	14.63681	0.00000
eng3	/eng1	256.17082	13.47180	19.01534	0.00000
eng3	/eng2	214.35205	11.70207	18.31744	0.00000
eng3	/eng3	310.90414	14.08313	22.07635	0.00000

LEVEL 3 COVARIANCE MATRIX

	math1	math2	math3	eng1	eng2
math1	2.35970				
math2	1.47571	4.59531			
math3	1.45284	2.34726	3.94231		
eng1	5.89266	5.10887	5.75844	40.46361	
eng2	5.42374	6.60379	7.26912	22.52145	33.03809
eng3	5.14461	5.67697	9.01133	22.93497	30.00480

	eng3
eng3	47.71171

LEVEL 3 CORRELATION MATRIX

	math1	math2	math3	eng1	eng2	eng3
math1	1.0000					
math2	0.4481	1.0000				
math3	0.4763	0.5515	1.0000			
eng1	0.6030	0.3747	0.4559	1.0000		
eng2	0.6143	0.5360	0.6369	0.6160	1.0000	
eng3	0.4849	0.3834	0.6571	0.5220	0.7557	1.0000

LEVEL 2 COVARIANCE MATRIX

	math1	math2	math3	eng1	eng2
math1	31.84368				
math2	22.14572	37.70830			
math3	17.66778	21.92750	29.34575		
eng1	65.59655	62.26680	49.65248	406.75188	

```
eng2           53.71682     56.85501     48.64973    272.02299    314.33951
eng3           53.74137     55.52744     50.44039    256.17082    214.35205


                    eng3

eng3           310.90414


       LEVEL 2 CORRELATION MATRIX

                 math1      math2      math3       eng1       eng2       eng3

math1          1.0000
math2          0.6391     1.0000
math3          0.5780     0.6592     1.0000
eng1           0.5764     0.5028     0.4545     1.0000
eng2           0.5369     0.5222     0.5065     0.7607     1.0000
eng3           0.5401     0.5128     0.5281     0.7204     0.6857     1.0000
```

Results for the fixed part of the model show that the expected score of girls (GENDER = 0) for all test scores is higher than the expected score for boys (GENDER = 1). For boys (GENDER = 1), the coefficients GENDER4 to GENDER6 are negative and highly significant. In the case of the mathematics test scores (MATH1 to MATH3 as represented by GENDER1 to GENDER3) the effects are smaller. Keep in mind that the range of scores differed between the mathematics and language tests.

The effects of the RAVENS test are positive and highly significant for all six response variables, with the largest effects for the mathematics tests. An increase of one unit in the RAVENS test score implies an expected increase in the third year language test of 0.61 and an expected increase of 1.86 for the expected third year mathematics test score.

Results for the random part of the model are consistent with the results for the previous model fitted, with larger variation for the three language tests and, in general, more variation over schools than over students.

The $-2\ln L$ recorded for this model is 45321.61343. When compared to the $-2\ln L$ of 45991.2579 obtained previously, a marked decrease is observed. In the first model, 48 parameters were estimated, compared to 60 parameters for the model discussed here. The introduction of the GENDER and RAVENS variables have contributed significantly to the explanation of variance in the response variables.