Mixed Models with R:

From Hierarchical to Mixed

Two equivalent views of hierarchical
different and important features

Georges Monette

random@yorku.ca

Hierarchical Models to Mixed Models

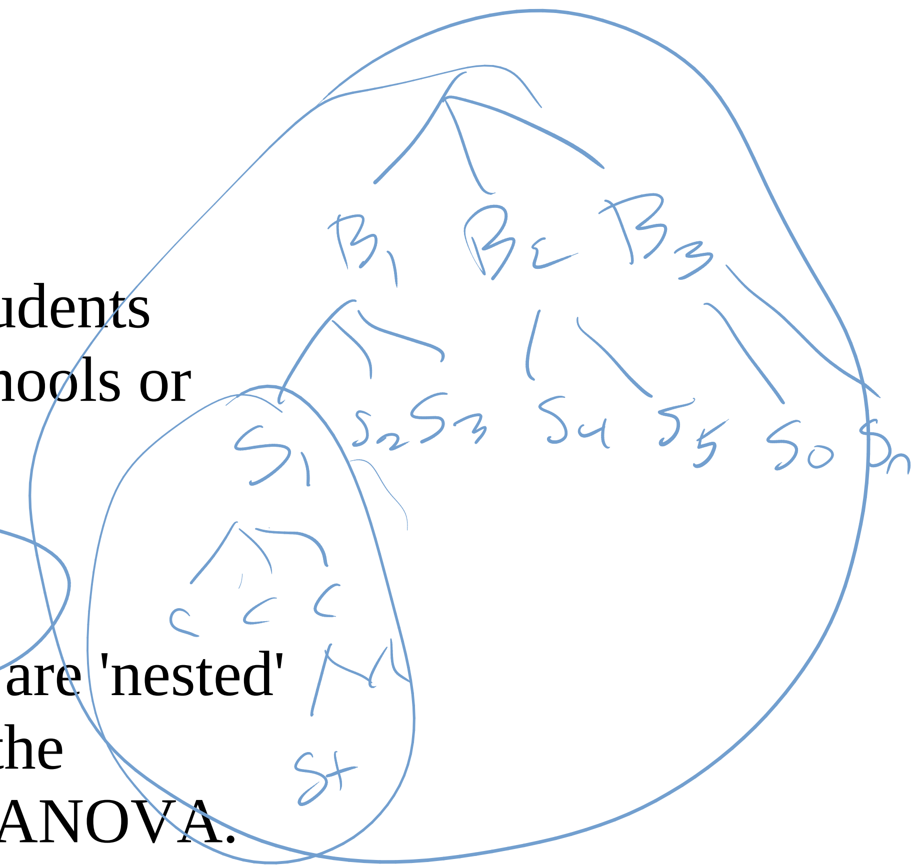
Contents

Notes:

- A version of this document that includes the data use to generate graphs will be available through the course web page.
- Lab 1, also available through the course web page, presents many additional concepts that complement the material in this document.
- Many concepts relevant to Hierarchical and Mixed Models will be seen in the sections on Longitudinal, Non-Linear and Generalized Linear Mixed Models.

The many hierarchies of statistics

Hierarchical Data:
refers to the structure of data with nested sampling levels: e.g. students sampled in schools and schools sampled from a population of schools or patients whose symptoms are measured on a number of visits.



$Y \sim 1$

$Y \sim 1 + X$ $\quad$ $Y \sim 1 + X + Z$

Hierarchical Models:
is often used to refer to a set of models used where some models are 'nested' within each other, i.e. a simpler model is obtained by restricting the parameters of a more complex model. This is the usual basis for ANOVA.

Hierarchical Model:
(the sense in which we use it) a model with hierarchical components intended to analyze hierarchical data. Of course, nothing prevents us from considering hierarchies of hierarchical models in which case we are using both concepts in the same sentence – although they refer to entirely different hierarchies.

Statistical goals for estimation:

We need to keep our goals in mind as we consider various approaches to analyze data. When you want to estimate something, e.g. a treatment effect or a comparison between two groups, you want your procedure to be:

1) ***consistent:*** You want to know that you are estimating the right thing with little bias.   i.e. you are aiming at the right target and, although your aim might be shaky, you won't be consistently off in any direction.

2) ***efficient:*** you want to shake as little as possible. You want to use the 'best' method available with this data and model to minimize the true standard error of estimation (what it really is, not what your procedure reports it to be)

3) ***honest:*** you would like to have an honest estimate of the true standard error. Otherwise, your CIs will have the wrong size and your p-values will be off. You may have more power than you think leading you to commit Type II errors unnecessarily or you may have less power than you think leading you to commit Type I errors too often.

4) ***robust:*** the more a good method remains good when assumptions are violated, the more robust it is. Robustness is more important if you are not confident of assumptions or if you know that the formal assumptions are not satisfied.

We don't necessarily need hierarchical models to analyze hierarchical data so we consider simpler approaches first and we will see how they measure up to our four criteria.

Hierarchical data

High school example:

For multilevel modeling we will use a subset of a data set presented at length in Bryk and Raudenbush (1992). The major variables are math achievement (mathach) and socio-economic status (ses) measured for each student in a sample of high school students in each of 40 selected schools in the U.S. Each school belongs to one of two sectors: 21 are Public schools and 19 are Catholic schools. There are 1,977 students in the sample. The sample size from each school ranges from 29 students to 66 students. The data are available as the data frame 'hs' in the package 'spida'. The full data set is 'hsfull' and two split halves are 'hs1' and 'hs2'. The following is a listing of the first 50 lines of the 'hs' file:

spida2

```
> head(hs,50)
   school mathach    ses    Sex Minority Size   Sector PRACAD DISCLIM
1     1317  12.862  0.882 Female       No  455 Catholic   0.95  -1.694
2     1317   8.961  0.932 Female      Yes  455 Catholic   0.95  -1.694
3     1317   4.756 -0.158 Female      Yes  455 Catholic   0.95  -1.694
4     1317  21.405  0.362 Female      Yes  455 Catholic   0.95  -1.694
5     1317  20.748  1.372 Female       No  455 Catholic   0.95  -1.694
6     1317  18.362  0.132 Female      Yes  455 Catholic   0.95  -1.694
7     1317  14.752  0.132 Female       No  455 Catholic   0.95  -1.694
8     1317  11.290 -0.008 Female      Yes  455 Catholic   0.95  -1.694
9     1317  10.493 -0.108 Female      Yes  455 Catholic   0.95  -1.694
10    1317  10.956  0.612 Female      Yes  455 Catholic   0.95  -1.694
11    1317  21.405  0.482 Female      Yes  455 Catholic   0.95  -1.694
12    1317  23.355  0.502 Female       No  455 Catholic   0.95  -1.694
13    1317  12.283  0.482 Female      Yes  455 Catholic   0.95  -1.694
14    1317   9.257  0.472 Female      Yes  455 Catholic   0.95  -1.694
15    1317  11.502 -0.578 Female       No  455 Catholic   0.95  -1.694
16    1317  20.039  1.152 Female      Yes  455 Catholic   0.95  -1.694
17    1317  21.405 -0.288 Female      Yes  455 Catholic   0.95  -1.694
18    1317  23.736  0.942 Female      Yes  455 Catholic   0.95  -1.694
19    1317  11.027  0.722 Female      Yes  455 Catholic   0.95  -1.694
20    1317  17.203 -0.108 Female      Yes  455 Catholic   0.95  -1.694
21    1317  10.661  1.462 Female      Yes  455 Catholic   0.95  -1.694
22    1317   7.031 -0.028 Female      Yes  455 Catholic   0.95  -1.694
23    1317  13.677  0.702 Female       No  455 Catholic   0.95  -1.694
24    1317  13.373  0.082 Female      Yes  455 Catholic   0.95  -1.694
25    1317  10.121 -0.108 Female      Yes  455 Catholic   0.95  -1.694
```

26	1317	10.394	0.322	Female	Yes	455	Catholic	0.95	-1.694
27	1317	6.973	0.302	Female	Yes	455	Catholic	0.95	-1.694
28	1317	11.064	-0.098	Female	No	455	Catholic	0.95	-1.694
29	1317	11.531	-0.848	Female	Yes	455	Catholic	0.95	-1.694
30	1317	8.253	-1.248	Female	Yes	455	Catholic	0.95	-1.694
31	1317	7.142	0.122	Female	Yes	455	Catholic	0.95	-1.694
32	1317	3.220	0.272	Female	Yes	455	Catholic	0.95	-1.694
33	1317	15.784	0.582	Female	No	455	Catholic	0.95	-1.694
34	1317	17.246	0.642	Female	Yes	455	Catholic	0.95	-1.694
35	1317	9.337	0.952	Female	Yes	455	Catholic	0.95	-1.694
36	1317	15.555	-0.258	Female	Yes	455	Catholic	0.95	-1.694
37	1317	8.382	0.492	Female	Yes	455	Catholic	0.95	-1.694
38	1317	11.621	0.992	Female	No	455	Catholic	0.95	-1.694
39	1317	4.810	0.832	Female	Yes	455	Catholic	0.95	-1.694
40	1317	17.869	-0.068	Female	Yes	455	Catholic	0.95	-1.694
41	1317	8.057	-0.088	Female	Yes	455	Catholic	0.95	-1.694
42	1317	11.794	0.972	Female	Yes	455	Catholic	0.95	-1.694
43	1317	18.939	0.542	Female	No	455	Catholic	0.95	-1.694
44	1317	20.261	0.132	Female	Yes	455	Catholic	0.95	-1.694
45	1317	10.066	-0.008	Female	Yes	455	Catholic	0.95	-1.694
46	1317	20.236	0.812	Female	No	455	Catholic	0.95	-1.694
47	1317	4.508	1.122	Female	No	455	Catholic	0.95	-1.694
48	1317	18.827	0.062	Female	No	455	Catholic	0.95	-1.694
49	1906	14.449	0.132	Female	Yes	400	Catholic	0.87	-0.939
50	1906	20.455	0.382	Female	No	400	Catholic	0.87	-0.939

? hs

The first 48 lines are students belonging to school labelled 1317. The last 2 lines are the first cases of the second school in the sample labelled 1906.

The uniform quantile plot of each variable gives a good snapshot of the data. Think of lining up each variables from shortest to tallest and plotting the result:

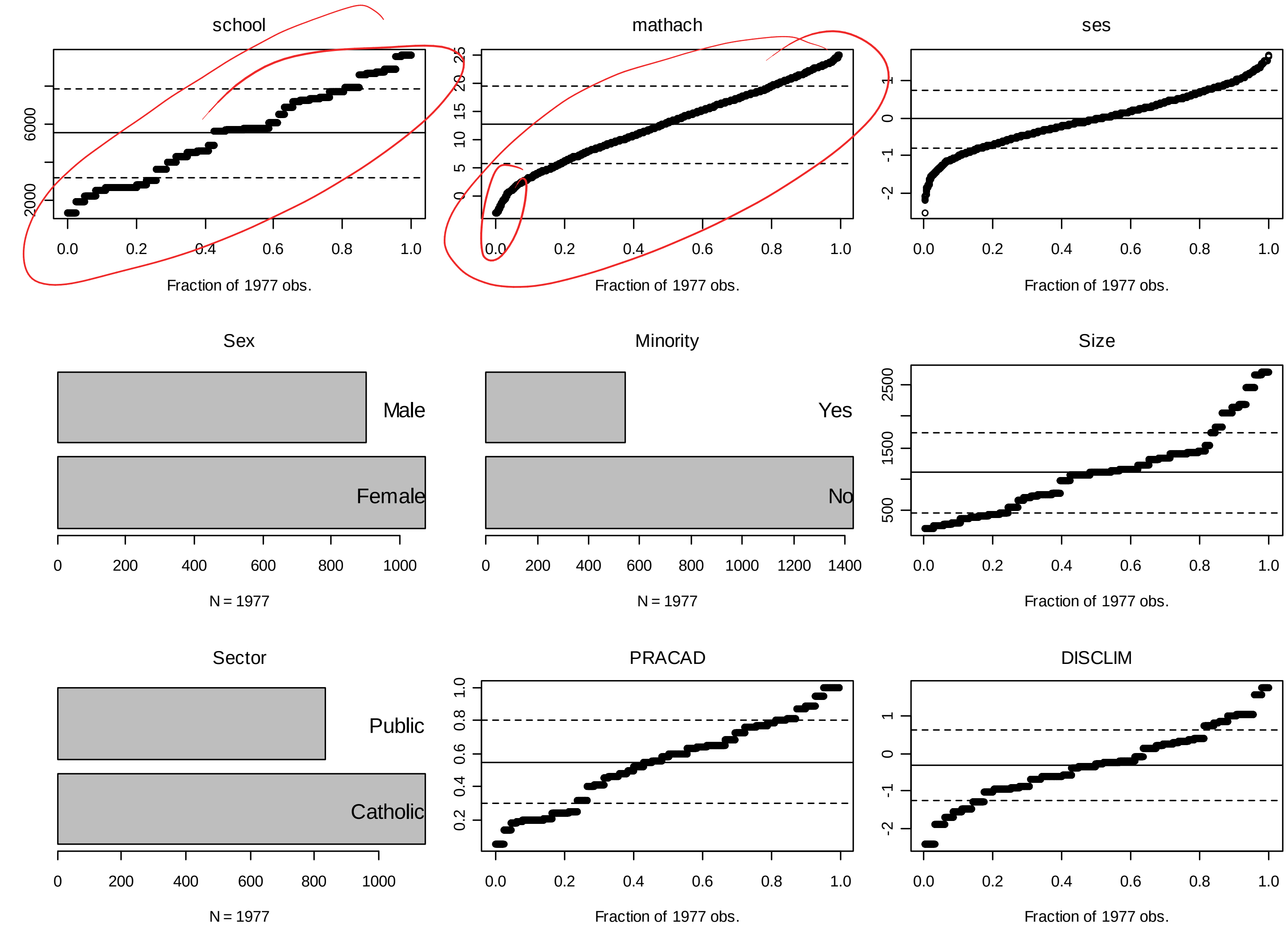
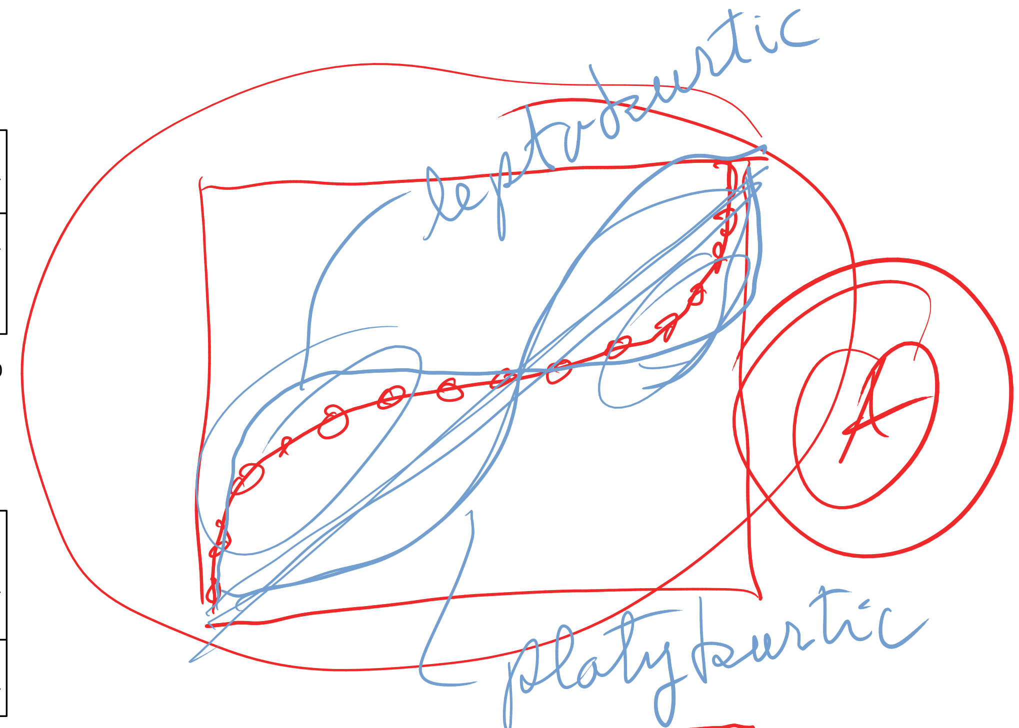
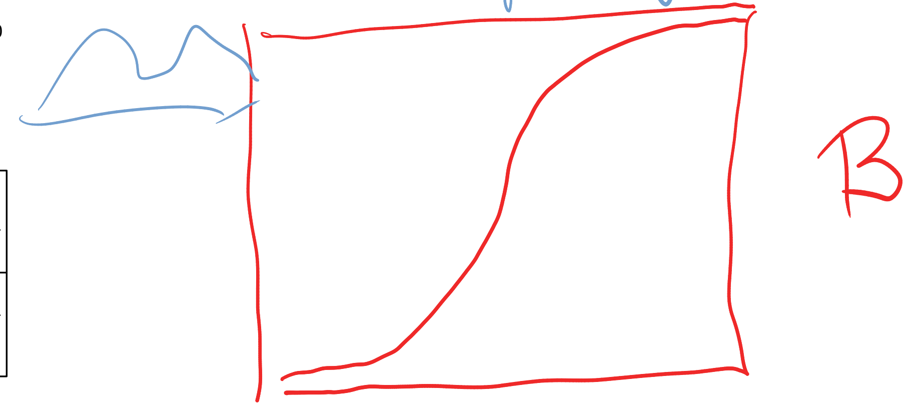


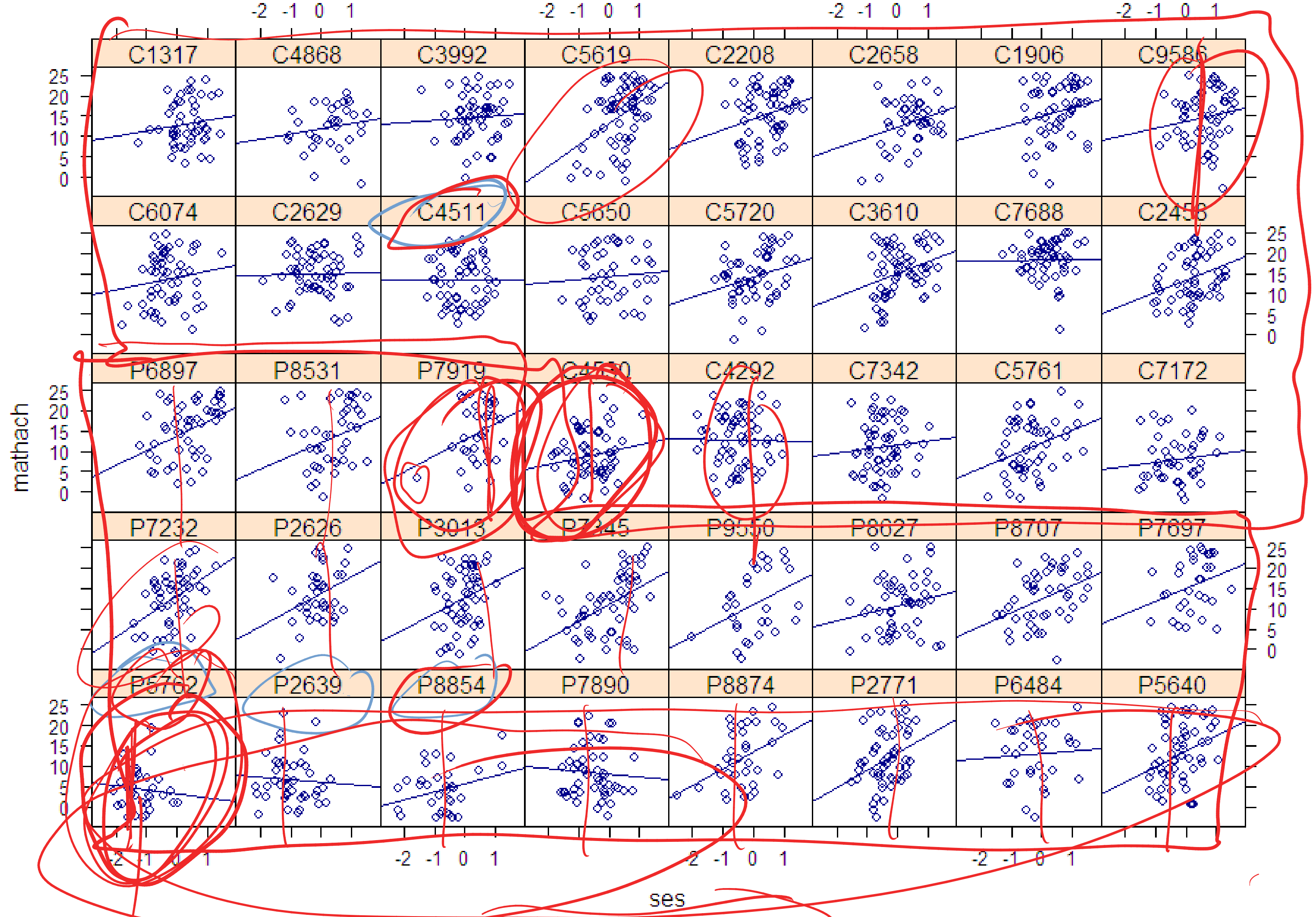
Figure 1: Uniform quantile plots of high school data

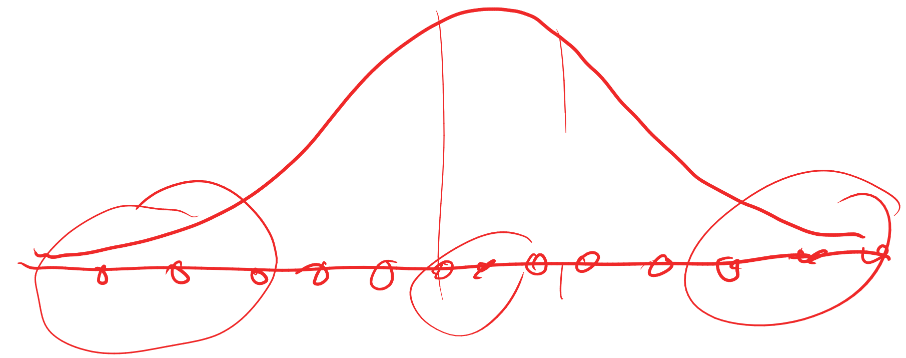


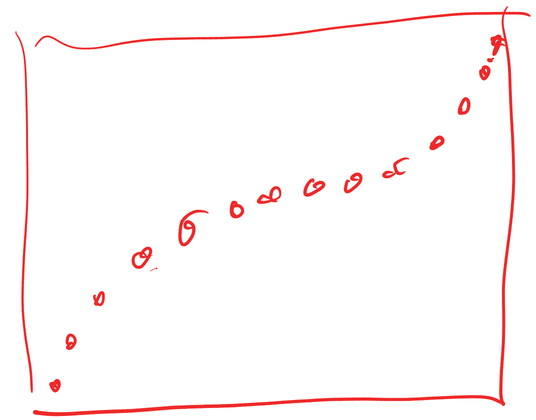


xqplot(hs, type='u')

Normal A or B







Lab 1.R

Figure 2: Trellis plot with least-squares line in each school. Note that the LS line could vary because of randomness in the observations within a school – i.e. they would vary even if all schools had exactly the same relationship between mathach and ses – and because the underlying relationship might vary from school to school.

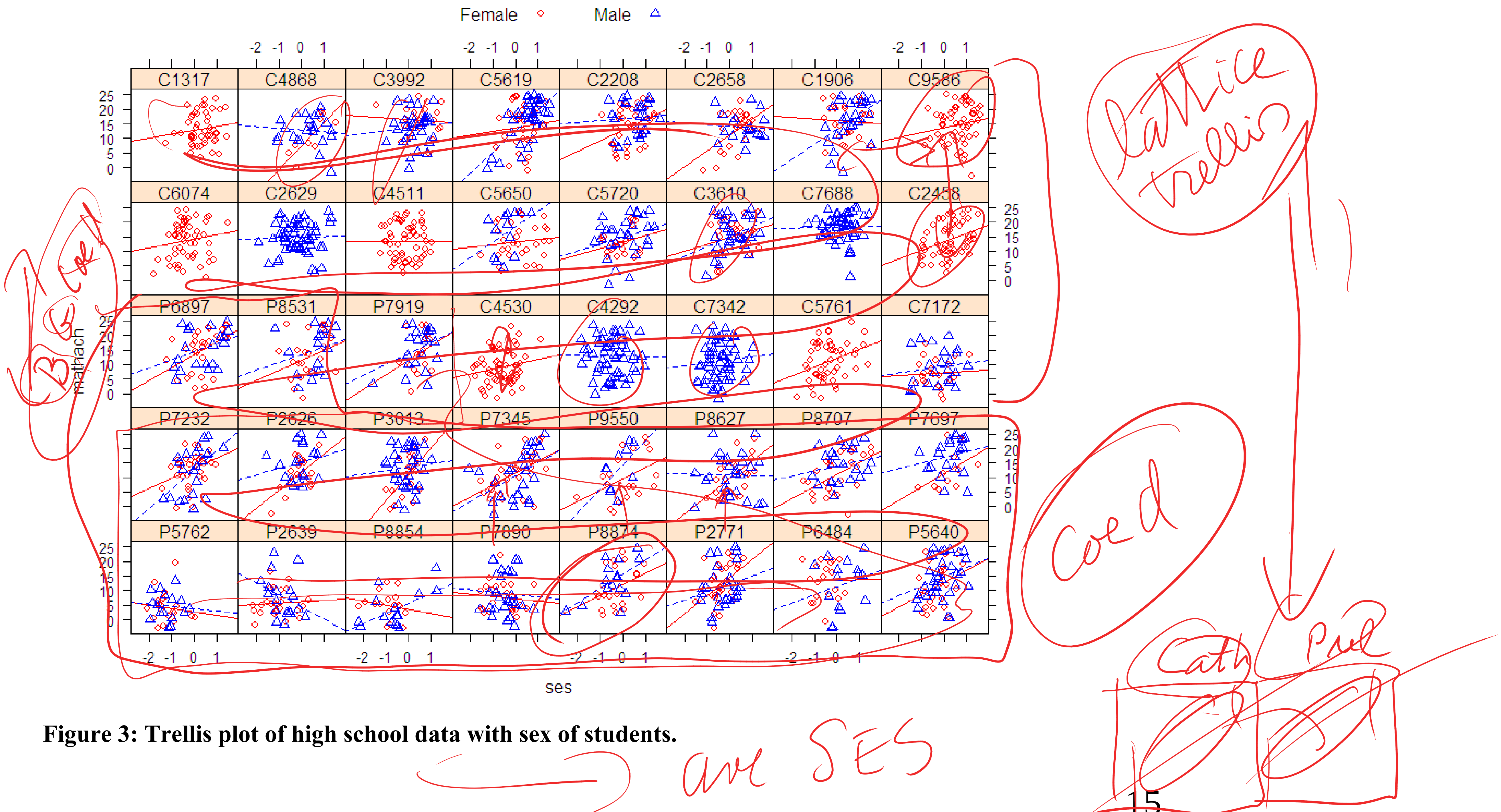


Figure 3: Trellis plot of high school data with sex of students.

ave SES

Comparing mathach and its relationship with ses in the two school sectors.

$Y \sim X + G$

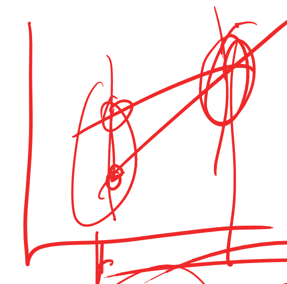
Some possible approaches:

SEC

Cath

Pub.

1) Pool the data from the schools within each sector and analyze with OLS. i.e. completely ignore the individual schools and regress mathach on ses and sector alone.



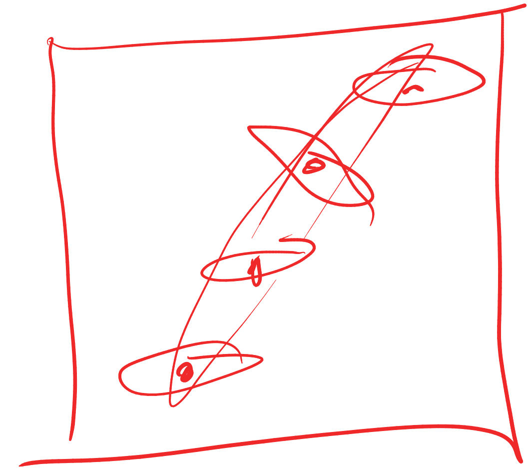
2a) Use a fixed effects model (Allison, 2005) to estimate relationship in each school and then compare the mean level of each sector. Can we just fit a model on SES, School and Sector to estimate the effect of Sector?

$Y \sim X * G$

2b) Use a fixed effects model with varying intercept from school to school but assume same slope within each Sector.

$Y \sim X * Sc$

3) Use a two step approach: fit a regression to each school and then estimate the mean intercept and slope of the schools in each sector with a multivariate analysis of the using the fitted intercepts and slopes as data.



4) Fit a 'between school' model: take the average ses and average mathach from each school and then perform a regression on the resulting means.

5) Use a hierarchical model

6) Use a hierarchical model with a contextual variable to see that we were really estimating two things to begin with.

Looking at hierarchical data

Doug Bates

2000

```
> library( spida )   # also loads nlme, lattice, car, MASS

> data( hs )

> head( hs )
    X school mathach    ses sector female    Sex Minority Size   Sector
1 141   1317  12.862  0.882      1      1 Female       No  455 Catholic
2 142   1317   8.961  0.932      1      1 Female      Yes  455 Catholic
3 143   1317   4.756 -0.158      1      1 Female      Yes  455 Catholic
4 144   1317  21.405  0.362      1      1 Female      Yes  455 Catholic
5 145   1317  20.748  1.372      1      1 Female       No  455 Catholic
6 146   1317  18.362  0.132      1      1 Female      Yes  455 Catholic
  PRACAD DISCLIM HIMINTY
1   0.95  -1.694       1
2   0.95  -1.694       1
3   0.95  -1.694       1
4   0.95  -1.694       1
5   0.95  -1.694       1
6   0.95  -1.694       1
```

nlme

2

1

0 ← Population

> Level 1 variables: vary from student to student within schools
>
> Level 2 variables: characteristics of schools

micro

student

macro

school

```
>     sapply( hs, class)
        X     school   mathach        ses     sector     female        Sex
"integer" "integer" "numeric" "numeric" "integer" "integer"  "factor"
 Minority       Size     Sector     PRACAD    DISCLIM    HIMINTY
 "factor" "integer"  "factor" "numeric" "numeric" "integer"

> tab( hs, ~ Sector + school )
          school
Sector     1317 1906 2208 2458 2626 2629 2639 2658 2771 3013 3610 3992 4292
  Catholic   48   53   60   57    0   57    0   45    0    0   64   53   65
  Public      0    0    0    0   38    0   42    0   55   53    0    0    0
  Total      48   53   60   57   38   57   42   45   55   53   64   53   65
          school
Sector     4511 4530 4868 5619 5640 5650 5720 5761 5762 6074 6484 6897 7172
  Catholic   58   63   34   66    0   45   53   52    0   56    0    0   44
  Public      0    0    0    0   57    0    0    0   37    0   35   49    0
  Total      58   63   34   66   57   45   53   52   37   56   35   49   44
          school
Sector     7232 7342 7345 7688 7697 7890 7919 8531 8627 8707 8854 8874 9550
  Catholic    0   58    0   54    0    0    0    0    0    0    0    0    0
  Public     52    0   56    0   32   51   37   41   53   48   32   36   29
  Total      52   58   56   54   32   51   37   41   53   48   32   36   29
          school
Sector     9586 Total
  Catholic   59  1144
  Public      0   833
  Total      59  1977
```

```
>     table( hs$school)  # number of observations per school

1317 1906 2208 2458 2626 2629 2639 2658 2771 3013 3610 3992 4292 4511 
  48   53   60   57   38   57   42   45   55   53   64   53   65   58 
4530 4868 5619 5640 5650 5720 5761 5762 6074 6484 6897 7172 7232 7342 
  63   34   66   57   45   53   52   37   56   35   49   44   52   58 
7345 7688 7697 7890 7919 8531 8627 8707 8854 8874 9550 9586 
  56   54   32   51   37   41   53   48   32   36   29   59 

>     tab( ~ Sector + school, hs) # each school is in one Sector
          school
Sector      1317 1906 2208 2458 2626 2629 2639 2658 2771 3013 3610 3992
  Catholic    48   53   60   57    0   57    0   45    0    0   64   53
  Public       0    0    0    0   38    0   42    0   55   53    0    0
  Total       48   53   60   57   38   57   42   45   55   53   64   53
          school
Sector      4292 4511 4530 4868 5619 5640 5650 5720 5761 5762 6074 6484
  Catholic    65   58   63   34   66    0   45   53   52    0   56    0
  Public       0    0    0    0    0   57    0    0    0   37    0   35
  Total       65   58   63   34   66   57   45   53   52   37   56   35
          school
Sector      6897 7172 7232 7342 7345 7688 7697 7890 7919 8531 8627 8707
  Catholic     0   44    0   58    0   54    0    0    0    0    0    0
  Public      49    0   52    0   56    0   32   51   37   41   53   48
  Total       49   44   52   58   56   54   32   51   37   41   53   48
```

```
            school
Sector       8854 8874 9550 9586 Total
  Catholic      0    0    0   59  1144
  Public       32   36   29    0   833
  Total        32   36   29   59  1977
```

```
>     tab( ~ Sex + school, hs)
```

```
          school
Sex        1317 1906 2208 2458 2626 2629 2639 2658 2771 3013 3610 3992
  Female     48   27   35   57   18    0   24   27   28   19   29   21
  Male        0   26   25    0   20   57   18   18   27   34   35   32
  Total      48   53   60   57   38   57   42   45   55   53   64   53
          school
Sex        4292 4511 4530 4868 5619 5640 5650 5720 5761 5762 6074 6484
  Female      0   58   63   11   30   24   32   24   52   21   56   20
  Male       65    0    0   23   36   33   13   29    0   16    0   15
  Total      65   58   63   34   66   57   45   53   52   37   56   35
          school
Sex        6897 7172 7232 7342 7345 7688 7697 7890 7919 8531 8627 8707
  Female     29   22   30    0   29    0   11   24   16   23   24   26
  Male       20   22   22   58   27   54   21   27   21   18   29   22
  Total      49   44   52   58   56   54   32   51   37   41   53   48
```

```
        school
Sex       8854 8874 9550 9586 Total
  Female    17   21   19   59  1074
  Male      15   15   10    0   903
  Total     32   36   29   59  1977
```

```
>     by ( hs, hs$Sector, function( dd ) tab( ~ Sex + school, dd))
```

```
hs$Sector: Catholic
        school
Sex       1317 1906 2208 2458 2629 2658 3610 3992 4292 4511 4530 4868
  Female    48   27   35   57    0   27   29   21    0   58   63   11
  Male       0   26   25    0   57   18   35   32   65    0    0   23
  Total     48   53   60   57   57   45   64   53   65   58   63   34
        school
Sex       5619 5650 5720 5761 6074 7172 7342 7688 9586 Total
  Female    30   32   24   52   56   22    0    0   59   651
  Male      36   13   29    0    0   22   58   54    0   493
  Total     66   45   53   52   56   44   58   54   59  1144
------------------------------------------------------
```

```
hs$Sector: Public
        school
Sex       2626 2639 2771 3013 5640 5762 6484 6897 7232 7345 7697 7890
  Female    18   24   28   19   24   21   20   29   30   29   11   24
  Male      20   18   27   34   33   16   15   20   22   27   21   27
  Total     38   42   55   53   57   37   35   49   52   56   32   51
        school
Sex       7919 8531 8627 8707 8854 8874 9550 Total
  Female    16   23   24   26   17   21   19   423
  Male      21   18   29   22   15   15   10   410
  Total     37   41   53   48   32   36   29   833
```

Note: All Public schools are co-educational.

Try 3-way table: `> tab( hs, ~ Sex + school + Sector)`

Summary variables and informative labels

Is a school male, female or co-ed?

```
>      some( hs )
         X school mathach    ses sector female    Sex Minority Size
419  1517   2771  11.226  0.302      0      0   Male       No  415
483  1760   3013  15.741  0.192      0      0   Male       No  760
627  2969   4292   9.255  0.972      1      0   Male      Yes 1328
960  3785   5640  14.699 -0.268      0      0   Male       No 1152
1272 4953   6897  15.885 -0.388      0      0   Male      Yes 1415
1632 5696   7890   3.295 -0.118      0      1 Female       No  311
1709 6132   8531  24.418  0.592      0      0   Male       No 2190
1717 6140   8531  -1.509  0.092      0      1 Female      Yes 2190
        Sector PRACAD DISCLIM HIMINTY
419     Public   0.24   1.048       0
483     Public   0.56  -0.213       0
627   Catholic   0.76  -0.674       1
960     Public   0.41   0.256       0
1272    Public   0.55  -0.361       0
1632    Public   0.21   0.845       0
1709    Public   0.58   0.132       0
1717    Public   0.58   0.132       0
```

Note: 'female' and 'Sex' are individual variable

Generating sex composition as a variable:

group mean variable = derived variable a Level 2 variable

compositional

contextual

```
>     hs$Sex.comp <- capply ( hs$Sex == "Female", hs$school, mean )
>     some(hs)
         X school mathach    ses sector female    Sex Minority Size
662   3004   4292   6.703 -0.138      1      0   Male      Yes 1328
929   3754   5640   9.223 -0.548      0      0   Male       No 1152
1123  4009   5762  -2.252 -1.028      0      0   Male      Yes 1826
1298  5093   7172   5.549  0.462      1      1 Female      Yes  280
1304  5099   7172   9.915 -0.628      1      0   Male      Yes  280
1383  5178   7232  16.278 -0.338      0      1 Female      Yes 1154
1536  5578   7688   9.587  0.612      1      0   Male      Yes 1410
1641  5705   7890  -2.362 -0.048      0      0   Male      Yes  311
1739  6162   8627  11.322  0.272      0      0   Male       No 2452
1922  7130   9586   7.974  0.212      1      1 Female       No  262

        Sector PRACAD DISCLIM HIMINTY  Sex.comp
662   Catholic   0.76  -0.674       1 0.0000000 (Catholic boys school)
929     Public   0.41   0.256       0 0.4210526
1123    Public   0.24   0.364       1 0.5675676
1298  Catholic   0.05   1.013       1 0.5000000 (Catholic coed school)
1304  Catholic   0.05   1.013       1 0.5000000 (Catholic coed school)
1383    Public   0.20   0.975       0 0.5769231
1536  Catholic   0.65  -0.575       0 0.0000000
1641    Public   0.21   0.845       0 0.4705882
1739    Public   0.25   0.742       0 0.4528302
1922  Catholic   1.00  -2.416       0 1.0000000 (Catholic girls school)
```

```
> hs.sch <- up( hs , ~ school)
> dim( hs.sch )
[1] 40  6
> some( hs.sch )
     school Size   Sector PRACAD DISCLIM    id
2771   2771  415   Public   0.24   1.048 P2771
4292   4292 1328 Catholic   0.76  -0.674 C4292
4530   4530  435 Catholic   0.60  -0.245 C4530
5720   5720  381 Catholic   0.65  -0.352 C5720
6074   6074 2051 Catholic   0.32  -1.018 C6074
6897   6897 1415   Public   0.55  -0.361 P6897
7172   7172  280 Catholic   0.05   1.013 C7172
8531   8531 2190   Public   0.58   0.132 P8531
8707   8707 1133   Public   0.48   1.542 P8707
8854   8854  745   Public   0.18  -0.228 P8854
```

Only Level 2 variables – constant within schools

```
> hs.sch.all <- up( hs , ~ school, all = T)
> dim( hs.sch.all )
[1] 40 10
> some( hs.sch.all )
     school  mathach         ses    Sex Minority Size   Sector PRACAD DISCLIM    id
1317   1317 13.177687  0.34533333 Female      Yes  455 Catholic   0.95  -1.694 C1317
2629   2629 14.907772 -0.13764912   Male       No 1314 Catholic   0.81  -0.613 C2629
2658   2658 13.396156  0.43844444 Female       No  780 Catholic   0.79  -0.961 C2658
3992   3992 14.645208  0.36539623   Male       No 1114 Catholic   0.73  -1.534 C3992
5640   5640 13.160105 -0.17659649   Male       No 1152   Public   0.41   0.256 P5640
5650   5650 14.273533  0.02244444 Female      Yes  720 Catholic   0.60  -0.070 C5650
```

Level 2 and Level 2 summaries of Level 1 variables

Means of numeric variables, modes of factors

Types of variables in multilevel models

1. Variables that **vary from student to student** within schools – Level 1

2. Variables that vary between schools and **do not vary within schools** – Level 2
   a. Variables that are characteristics of the school
   b. Variables that are **derived** from within school variables, e.g. **group mean** ses in the sample in the school.

3. (really a version of 1) Variables that are derived by combining 1 and 2: e.g. deviations from the within group mean ses, i.e. **within school variable centered within groups. (CWG)**

Synomyms:

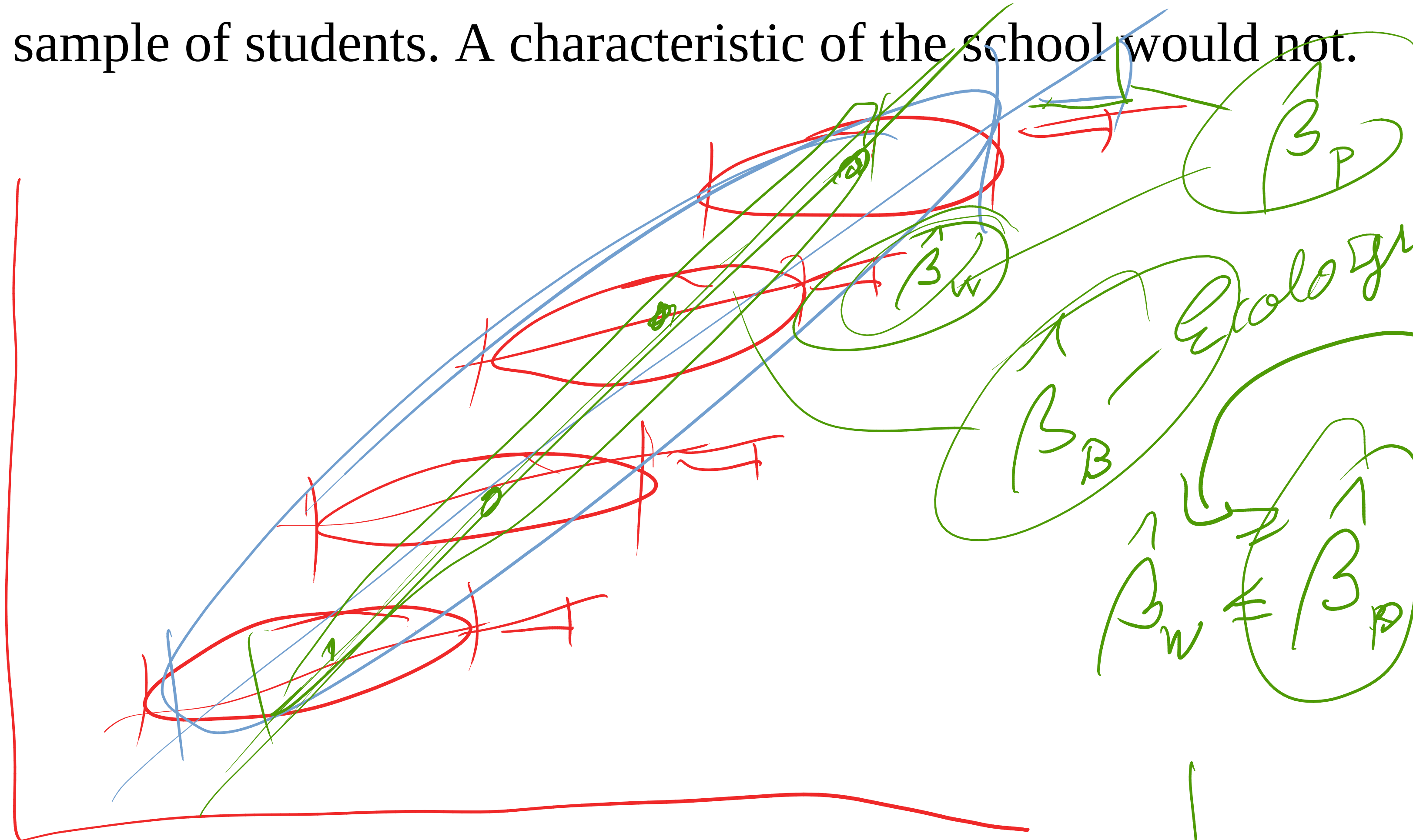
1. Variables that vary within clusters (=groups):
   Level 1[1] variables (if we count from the bottom as in SPSS or HLM), micro variables, within cluster variable, time-varying variables (if X is time, student-level variables

2. Variables that are constant within schools:
   Level 2 variables (in SPSS, HLM), macro variables, between cluster variables, ***contextual variables,*** time-invariant variables (if X is time), school-level variables.

---

[1] I believe that Pinheiro and Bates are alone counting in the opposite direction: Level 0 is the whole population, Level 1 the schools, Level 2 the students. This only matters when predicting from a multilevel model.

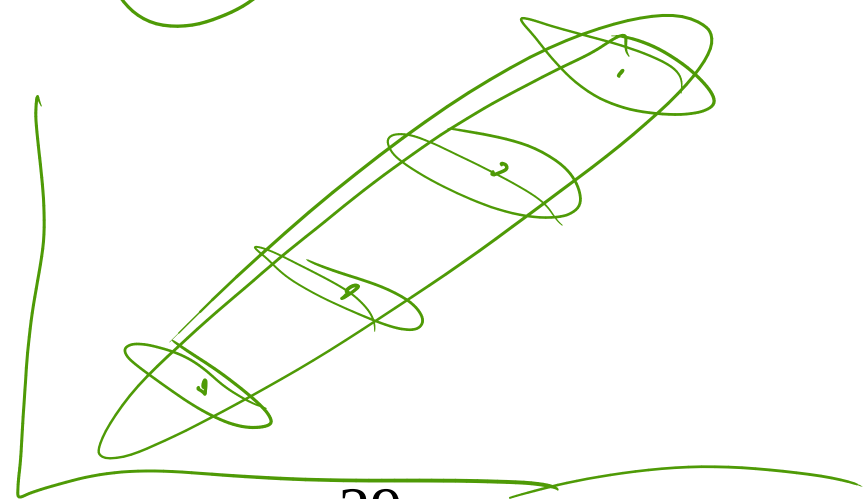
Note: The difference between a characteristic of the school and a 'derived' variable is that a derived variable could have a different value with a different sample of students. A characteristic of the school would not.



Ecological correlation

$\hat\beta_{xxx}$

$\hat\beta_W \le \hat\beta_P \le \hat\beta_B$



```
>     hs$Sex.cat <- factor( ifelse( hs$Sex.comp == 1, "Girls",
                  ifelse( hs$Sex.comp == 0 , "Boys", "Coed")) )
>     some(hs)
        X school mathach    ses sector female    Sex Minority Size
13    153   1317  12.283  0.482      1      1 Female      Yes  455
27    167   1317   6.973  0.302      1      1 Female      Yes  455
526  2284   3610  21.034  1.012      1      0   Male       No 1431
1394 5341   7342  23.271 -0.748      1      0   Male       No 1220
1417 5364   7342  12.821 -0.248      1      0   Male       No 1220
1441 5388   7342  11.664  0.862      1      0   Male       No 1220
1658 5722   7919  13.184 -0.038      0      0   Male       No 1451
1876 6504   8874  20.879  0.732      0      0   Male      Yes 2650
1884 6512   8874  24.479  0.652      0      0   Male       No 2650
1898 7106   9550  20.149  0.472      0      1 Female       No 1532

       Sector PRACAD DISCLIM HIMINTY  Sex.comp Sex.cat
13   Catholic   0.95  -1.694       1 1.0000000   Girls
27   Catholic   0.95  -1.694       1 1.0000000   Girls
526  Catholic   0.80  -0.621       0 0.4531250    Coed
1394 Catholic   0.46   0.380       1 0.0000000    Boys
1417 Catholic   0.46   0.380       1 0.0000000    Boys
1441 Catholic   0.46   0.380       1 0.0000000    Boys
1658   Public   0.50  -0.402       0 0.4324324    Coed
1876   Public   0.20   1.742       0 0.5833333    Coed
1884   Public   0.20   1.742       0 0.5833333    Coed
1898   Public   0.45   0.791       0 0.6551724    Coed
```

Creating a more informative school id

```
>  hs$sid <- factor( paste( substr( hs$Sector, 1,1),
                hs$school, substr( hs$Sex.cat, 1,1), sep = ''))
# Keep each sector together, within sector order by mean ses:
>  hs$sid <- reorder( hs$sid, hs$ses + 1000 * (hs$Sector == "Catholic"))
>  some(hs)

        X school mathach    ses sector female    Sex Minority Size
137   789   2208  14.150  0.482      1      1 Female       No 1061
165   992   2458   7.814 -1.058      1      1 Female      Yes  545
231  1167   2626  10.350 -0.448      0      1 Female       No 2142
265  1201   2629  20.891 -0.278      1      0   Male       No 1314
358  1384   2658   9.459  0.702      1      1 Female       No  780
407  1505   2771  17.129 -0.328      0      1 Female       No  415
450  1548   2771  21.020 -1.098      0      1 Female       No  415
687  3029   4292  19.030 -0.498      1      0   Male      Yes 1328
953  3778   5640  16.212 -0.308      0      0   Male       No 1152
1617 5681   7890   0.930 -1.038      0      0   Male       No  311
       Sector PRACAD DISCLIM HIMINTY  Sex.comp Sex.cat    sid
137  Catholic   0.68  -0.864       0 0.5833333    Coed C2208C
165  Catholic   0.89  -1.484       1 1.0000000   Girls C2458G
231    Public   0.40   0.142       0 0.4736842    Coed P2626C
265  Catholic   0.81  -0.613       0 0.0000000    Boys C2629B
358  Catholic   0.79  -0.961       0 0.6000000    Coed C2658C
407    Public   0.24   1.048       0 0.5090909    Coed P2771C
450    Public   0.24   1.048       0 0.5090909    Coed P2771C
687  Catholic   0.76  -0.674       1 0.0000000    Boys C4292B
953    Public   0.41   0.256       0 0.4210526    Coed P5640C
```

Easy manipulation of multilevel data

Creating a multilevel data set:

1. Create a data set for each level, e.g. school and students. Or board, school and student with 3 levels.

2. Include an index variable for each level – a variable that has a unique value for each row of its data set. In each data set include the values of the index for the data set immediately above it.

3. Make sure all variable names are unique across all data sets except for the index variables that need to have the same name in a data set and the data immediately below.

How? You can use Excel and save as '.csv' file. Then read into R.

```
> schoolfile <- read.csv("schoolfile.csv")
> studentfile <- read.csv("studentfile.csv")
```

Merge files into a single combined file (often called a 'long' file) for analysis:

```
> combfile <- merge( schoolfile, studentfile )
```

Note: `hs` is already a long file in which Level 2 variables were entered directly in a Level 1 file. You can also do this but there are slightly higher chances of errors if Level 2 variables are entered inconsistently.

We saw above how to create a Level 2 derived variable from Level 1 data with `capply`

Going from the long file to the short file with 'school invariant' variables only:

```
>     hs.sid <- up ( hs, ~ sid )
>     some( hs.sid )

       school sector Size   Sector PRACAD DISCLIM HIMINTY  Sex.comp
C2208C   2208      1 1061 Catholic   0.68  -0.864       0 0.5833333
C2658C   2658      1  780 Catholic   0.79  -0.961       0 0.6000000
C3610C   3610      1 1431 Catholic   0.80  -0.621       0 0.4531250
C4530G   4530      1  435 Catholic   0.60  -0.245       1 1.0000000
       Sex.cat    sid
C2208C    Coed C2208C
C2658C    Coed C2658C
C3610C    Coed C3610C
C4530G   Girls C4530G
```

Looking at Hierarchical Data

Look at relationships (mathach ~ ses) in hierarchical data

3 main tools

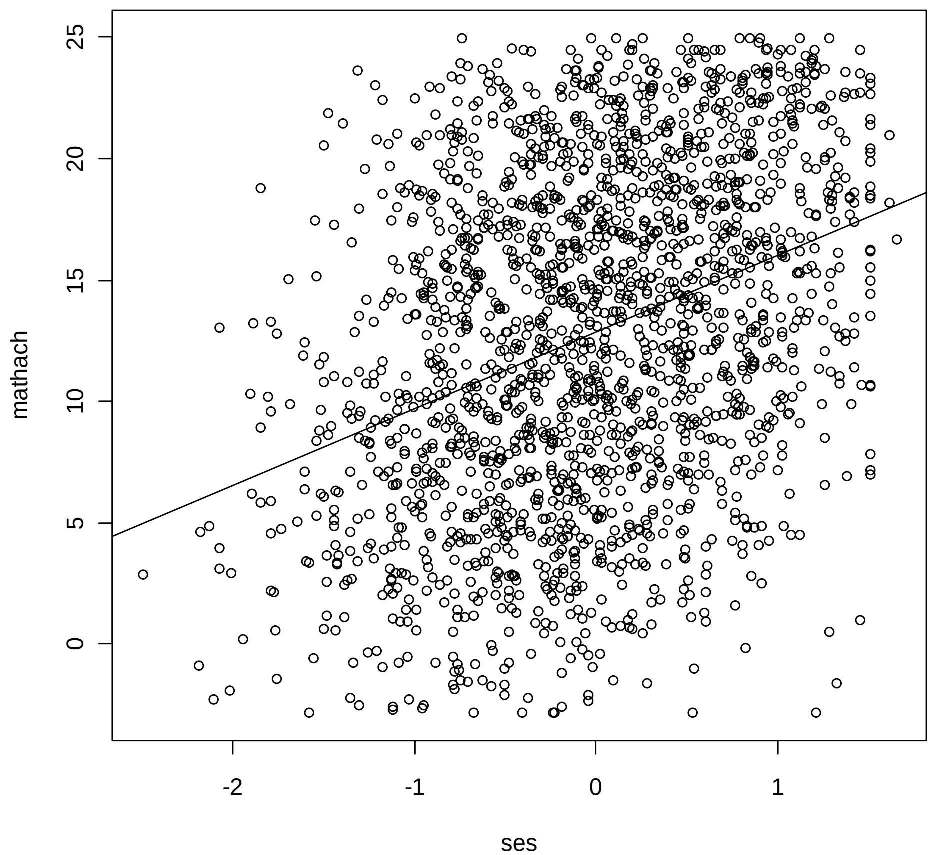
1) Traditional graphics

2) Lattice (=trellis) graphics

3) 3D graphics

Traditional graphics:

```
>  fit <- lm( mathach ~ ses, hs)
>  plot( mathach ~ ses, hs)
>  abline( fit )
```



Advantage:

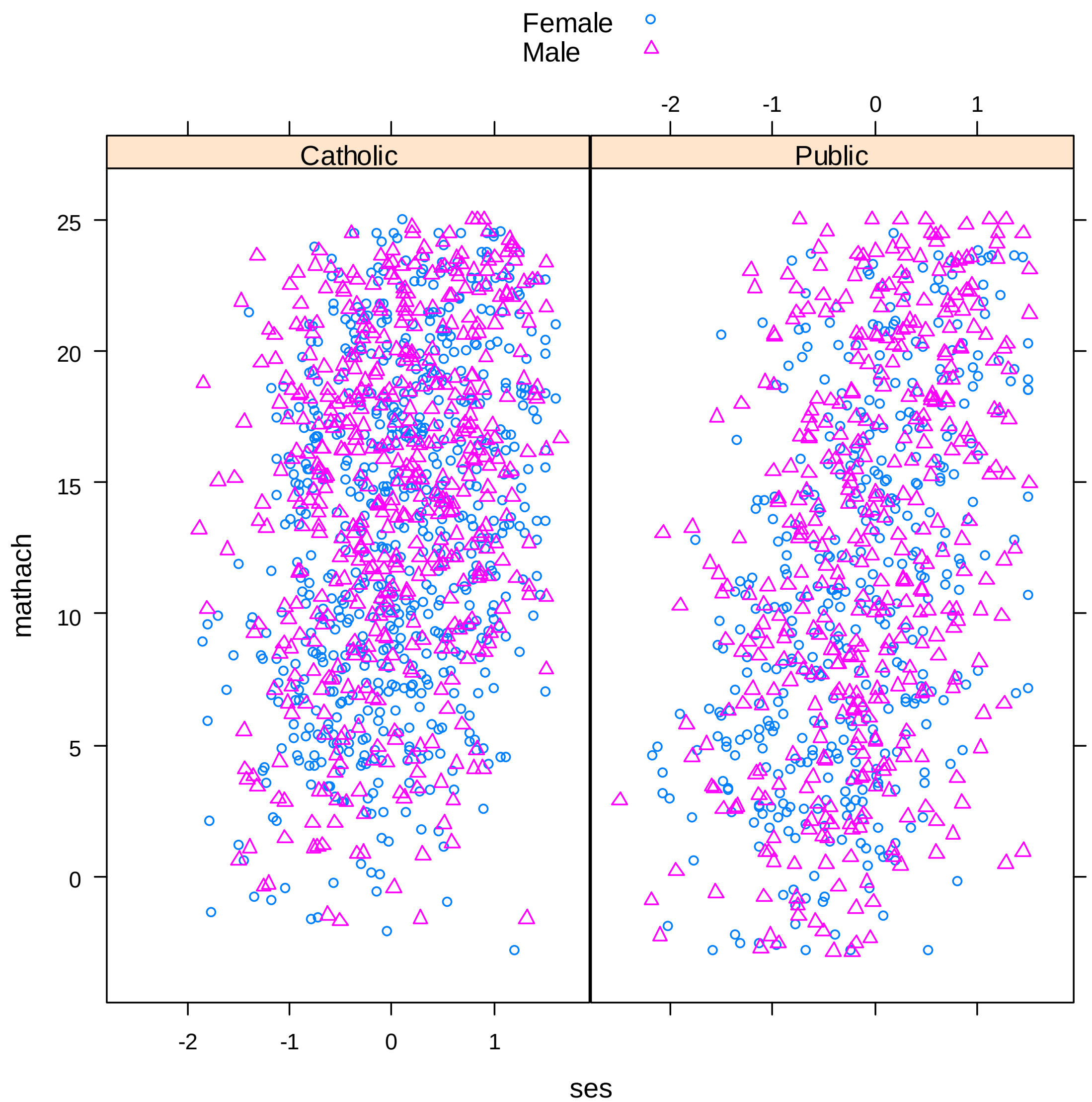
* Easy to add new objects
* Intuitive
* Somewhat interactive

Lattice graphics

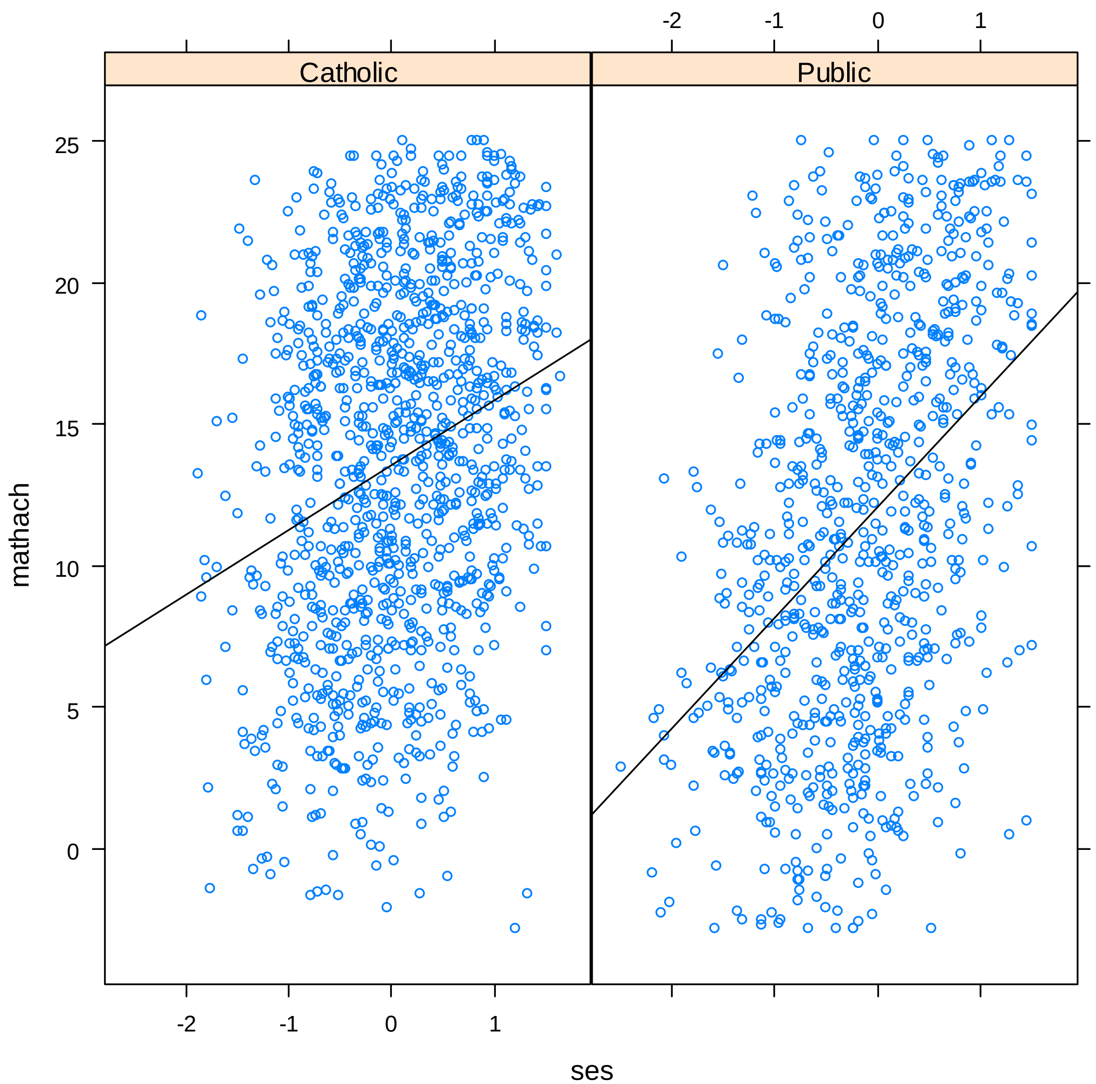
Easy to create panels and groups within panels

```
> library( lattice )

> xyplot( mathach ~ ses
   | Sector, hs,
   groups = Sex,
   auto.key = T)
```

But it's more difficult to add extra elements to the graph. This must be done in 'panel' functions that are called to generate each panel or with the 'trellis.focus' interface.



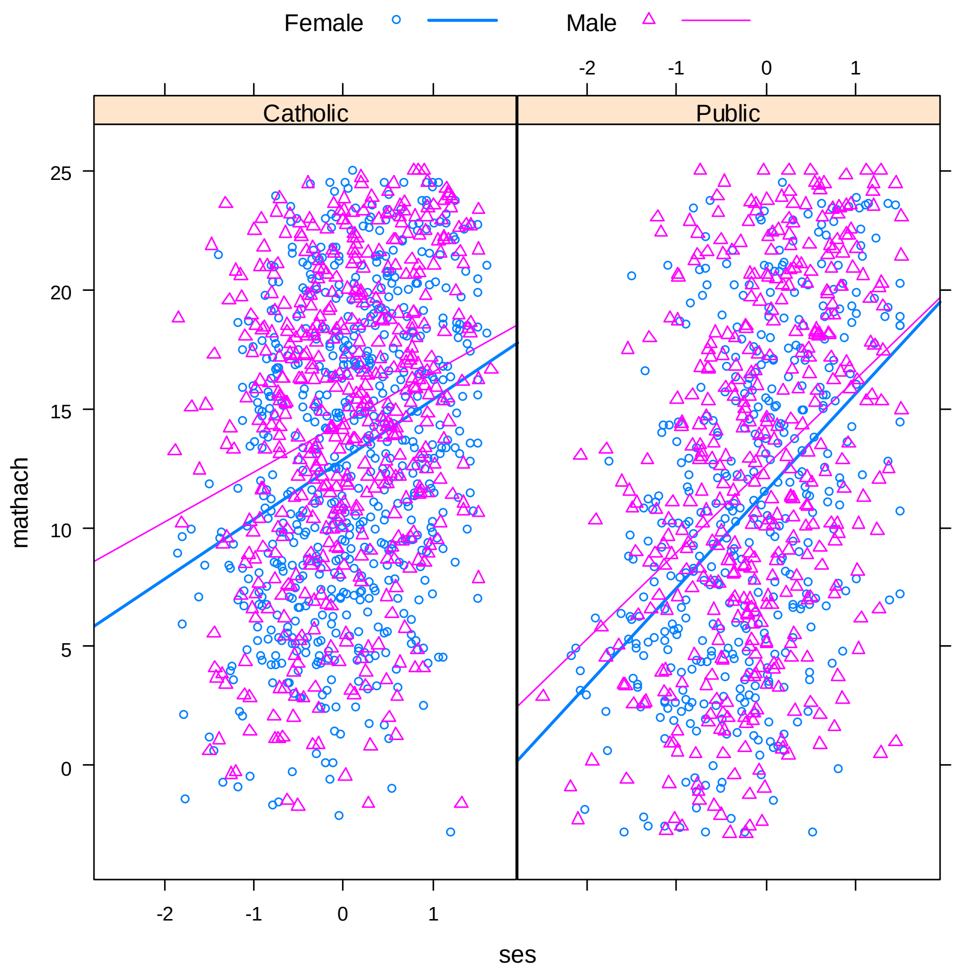
```
> xyplot(
   mathach ~ ses | Sector,
   hs,
   panel =
    function(x, y, ...) {
     panel.xyplot( x, y,
              ...)
     panel.lmline( x, y,
               ...) } )
```



The 'panel' function is defined on the fly. It uses arguments that will be passed to it automatically when it is called within xyplot to draw the panels. It uses convenience functions `'panel.xyplot'` and `'panel.lmline'` that are designed to work well within panels. Try **`??panel`**

```
xyplot( mathach ~ ses | Sector , hs,  groups = Sex,
        auto.key = list( columns = 2, lines = T, points = T),
        panel = panel.superpose.2,
        panel.groups = function(x, y, ...) {
            panel.xyplot( x, y, ...)
            panel.lmline( x, y ,...) })
```

A more complex example using groups and `panel.groups` that is called for each group within each panel.



Exploring the relationship between mathach and ses

We want to explore how mathach and its relationship with ses differ between sectors.

As mentioned previously there are a number of plausible approaches:

1) Pooling the data: ignore schools, just pool all the data in each sector together and do an OLS regression.

```
lm( matach ~ ses * Sector, hs)
```

2 a) Use a fixed effects model (version 1) to estimate relationship in each school and then compare the mean level of each sector.

```
lml <- lm( matach ~ factor(school)/ses -1,
              hs)
ddu <- up( hs, ~ factor(school))
ind <- ddu$Sector == "Catholic"
L <- rbind( "Catholic" = ind,
```

```
                "Public" = 1-ind)
L <- L/apply(L,1, sum)
L <- cbind( rbind( L, 0,0), rbind( 0,0,L))
rownames( L ) <- c("Cath Int", "Pub Int",
                   "Cath Slope", "Pub Slope")
wald (lml, L)
diffmat <- rbind( "Int" = c( -1, 1, 0, 0),
                  Slope = c( 0 , 0, -1, 1))

wald (lml, diffmat %*% L)
 numDF denDF  F.value p.value
     1     2  1897 21.03533 <.00001
      Estimate Std.Error   DF   t-value  p-val
Int   -2.027255  0.351992 1897 -5.759378 <.00001
Slope  1.109995  0.454114 1897  2.444309 0.0146
```

Difference of averages that give equal weight to each school. Uses only within-school variability except for pooled estimate of σ^2.

Question: Why is this so complicated? Can't we just fit a model regressing on SES, School and Sector to estimate the effect of Sector?

2 b) Fixed effects model (version 2): OLS regression with different intercepts in each school but common slopes in each sector.

```
fit2 <- lm( mathach ~ ses + factor(school)
    + ses:Sector, hs)
wald( fit2, "ses:Sector")
```

```
Coefficients Estimate Std.Error   DF  t-value p-value
  ses:Sector 1.204719  0.436134 1935 2.762267 0.00579
```

Here we assume all slopes are the same within each sector. The average Sector slope gives more weight to schools with larger samples and more spread in ses. Between school variability in levels plays no role in SEs.

3) MANOVA approach: Get individual school intercepts and slopes as in 2a but then do a MANOVA to compare the two sectors.

```
lcoefs <- coef( lmList( mathach ~ ses |
           factor(school), hs))
lm.mult <- lm( as.matrix(lcoefs) ~ Sector,
           up( hs, ~ factor(school)))
summary(lm.mult)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.6672     0.3506   4.756 2.84e-05 ***
SectorPublic   1.1100     0.5086   2.182   0.0353 *
```

SE is measured from between school variability not within school variability. The fact that the precision of estimates varies from school to school is ignored. However inferences to generalize to the larger population. Note the larger p-value

4) Ecological or between school model: Summarize the data from each school with the mean ses and the mean mathach from each school. Do an OLS regression on the resulting data.

```
fit.eco <- lm( mathach ~ ses,
               up( hs, ~ factor(school), all = T))
summary( fit.eco )
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.7926     0.2898  44.150  < 2e-16 ***
ses           5.7734     0.6984   8.267  5.1e-10 ***
```

This is estimating something totally different: the difference in between school slopes, not within school slopes

5) Use a Hierarchical Linear Model

The HLM uses both between school variation and within school variation to estimate the standard error of estimates. Inference generalizes to the larger population. Some estimates in the HLM rely on the assumption that between school and withing school effects are the same.

6) Use a Hierarchical Linear Model with appropriate contextual variables.

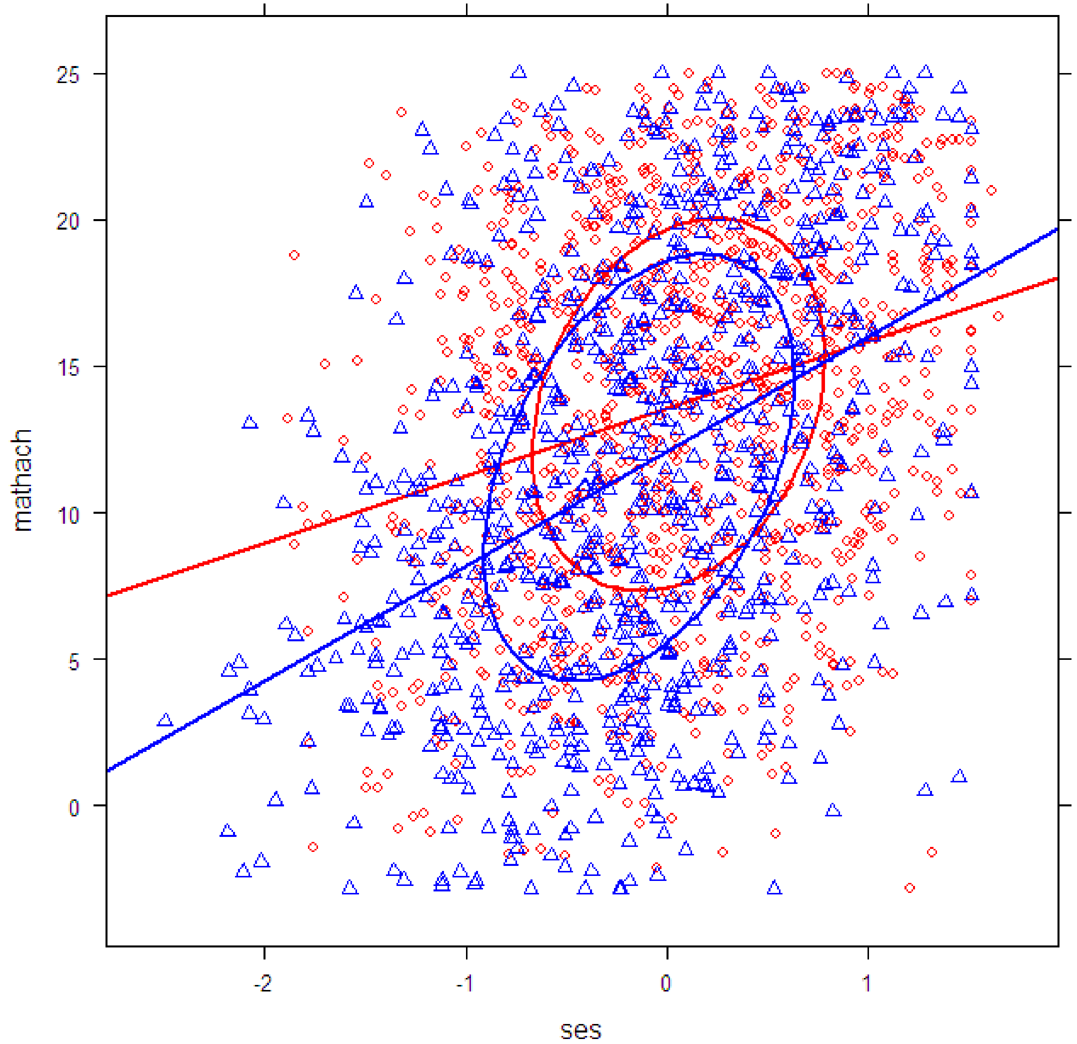
Using a derived contextual variable for ses (group mean ses in each school) as well as raw (or centered within school) ses allows separate unbiased estimation of both within and between school effects.

Method 1: Pooling of data – ignore schools

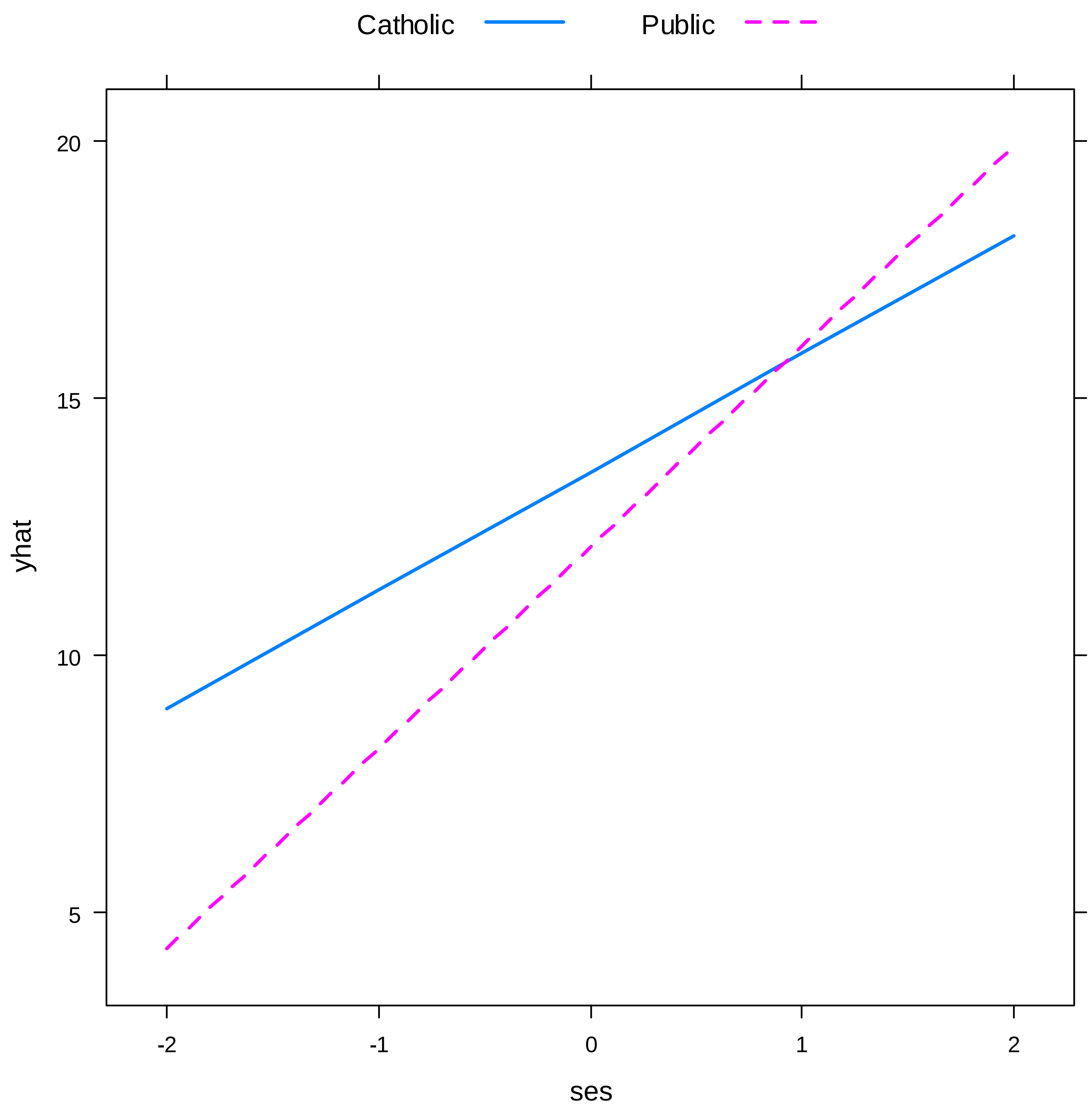
```
>     fit.pooled <- lm( mathach ~ ses * Sector, hs)
>     summary(fit.pooled)
Call:
lm(formula = mathach ~ ses * Sector, data = hs)
Residuals:
     Min       1Q   Median       3Q      Max 
-19.1774  -4.8286   0.2949   4.9595  15.7836 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     13.5579     0.1881  72.067  < 2e-16 ***
ses              2.2999     0.2582   8.908  < 2e-16 *** [eff. of ses|Cath]
SectorPublic    -1.4666     0.2921  -5.021 5.60e-07 *** [Pub-Cath|ses=0]
ses:SectorPublic 1.6051     0.3845   4.174 3.12e-05 *** [diff. of slopes]

Residual standard error: 6.344 on 1973 degrees of freedom
Multiple R-squared: 0.1404,     Adjusted R-squared: 0.1391 
F-statistic: 107.4 on 3 and 1973 DF,  p-value: < 2.2e-16
```

Coefficients in blue are 'marginal' to the interaction and should be interpreted – if at all – with care. The coefficient for "ses" (2.2999) is NOT "the estimated effect of ses" – it is the estimated "effect" of ses when SectorPublic = 0, i.e. in Catholic schools

Method 1: Fitted lines



The code to produce this and following graphs is contained in the on-line appendix



Catholic
Public
yhat
20
15
10
5
-2
-1
0
1
2
ses

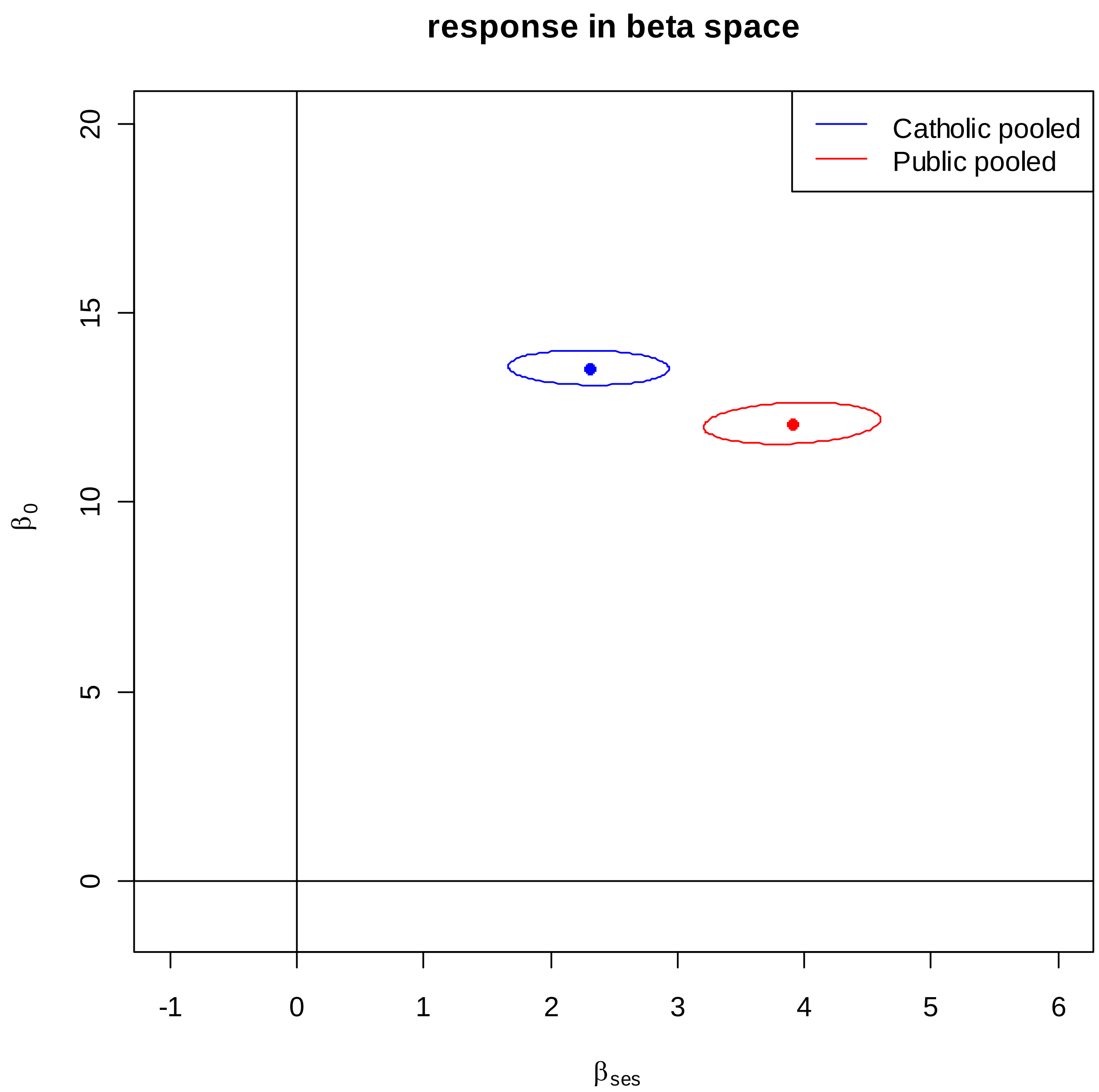
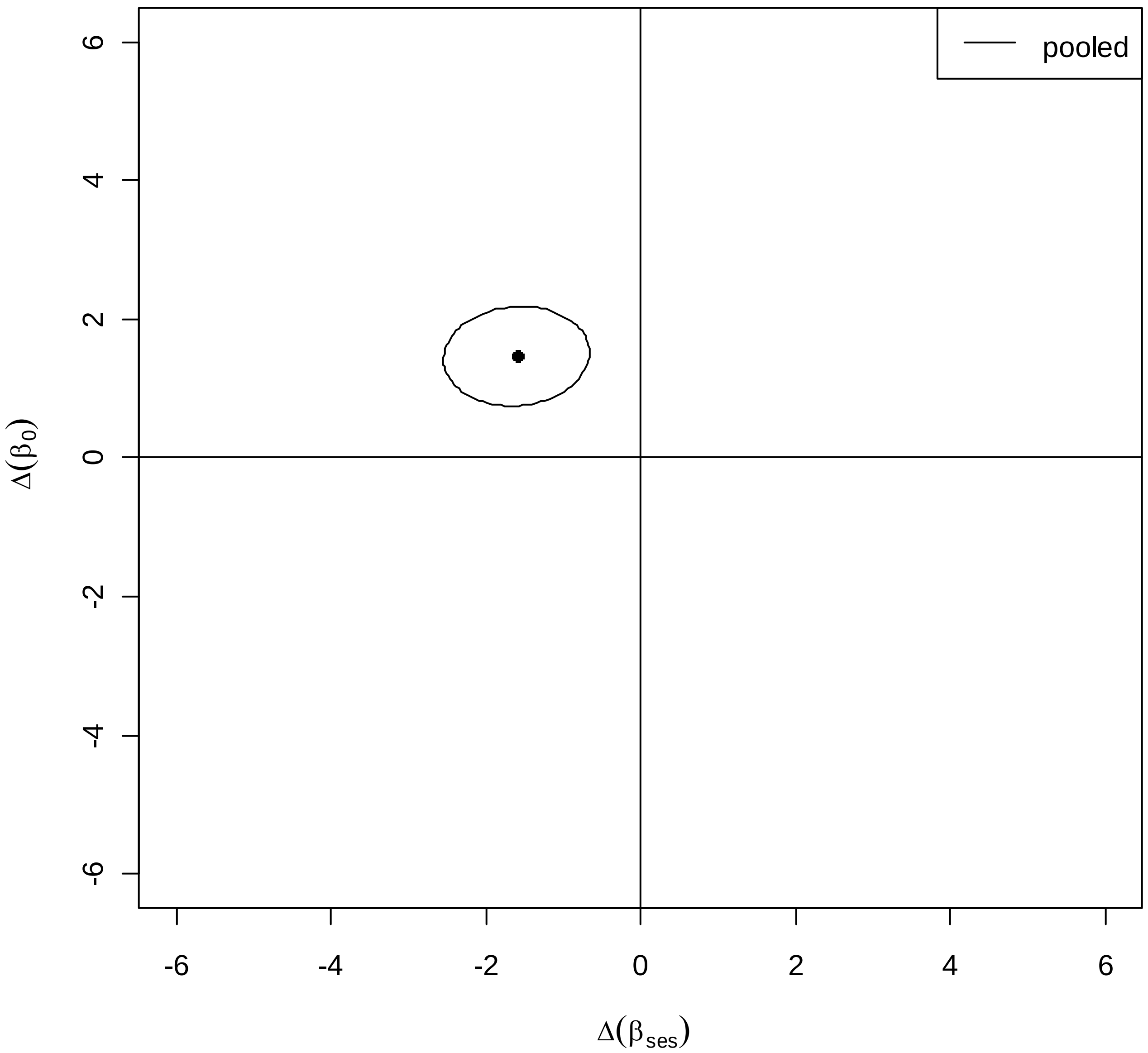
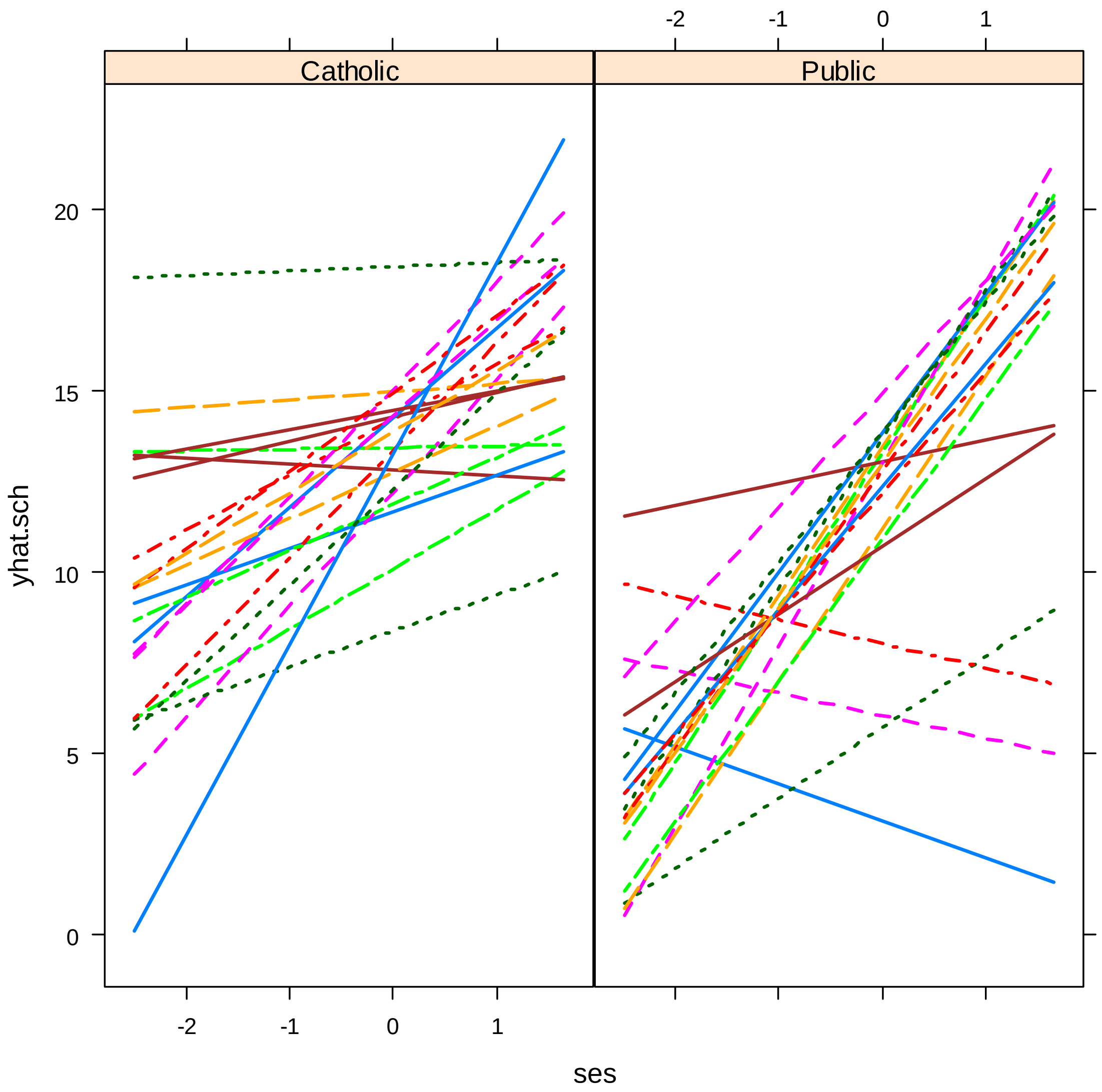


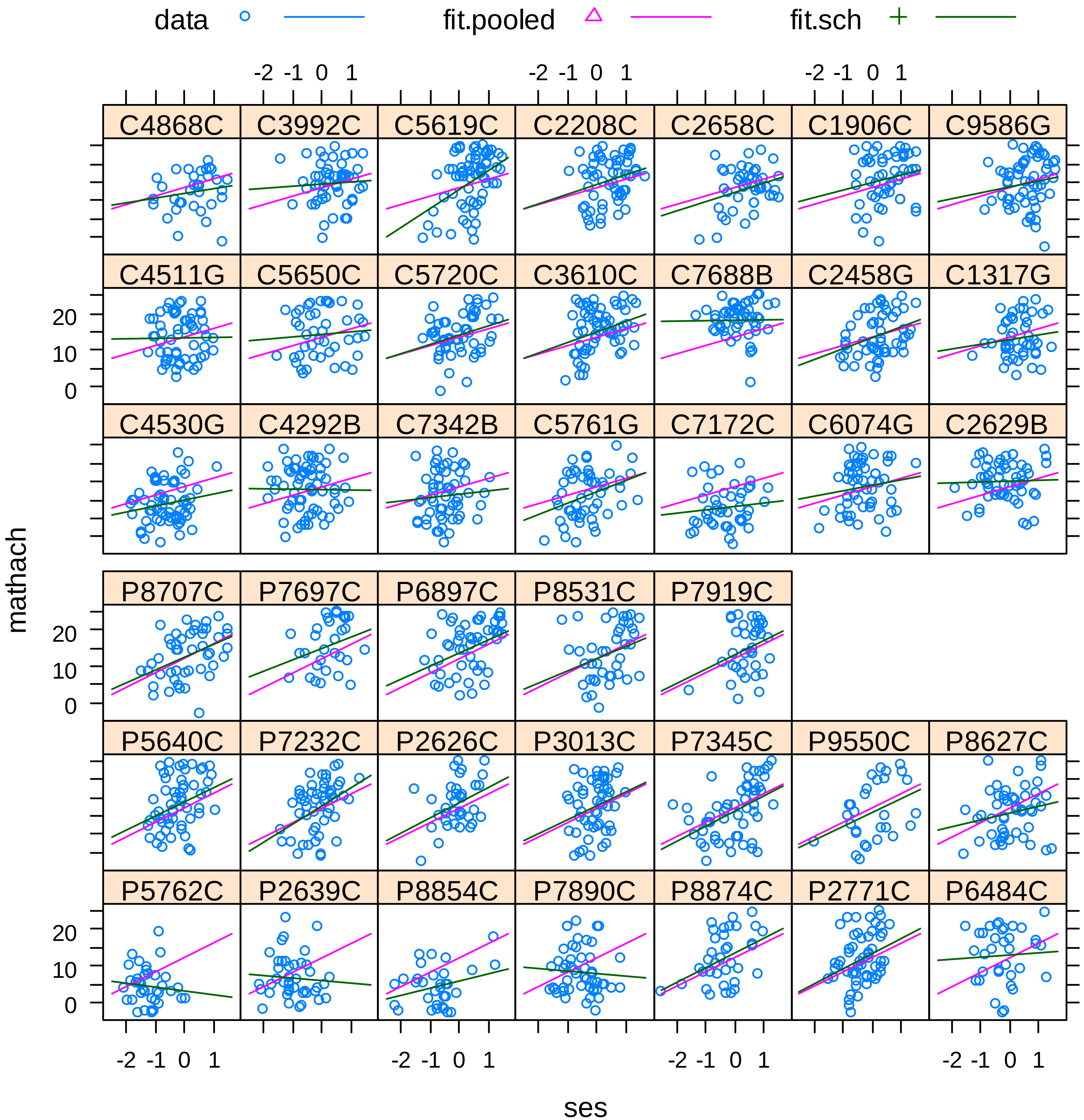
Figure 4: 95% confidence ellipse for intercept and slope in each Sector

difference between sectors in beta space



Method 2: Fit each school then average slopes and intercepts in each sector




data
fit.pooled
fit.sch
C4868C
C3992C
C5619C
C2208C
C2658C
C1906C
C9586G
C4511G
C5650C
C5720C
C3610C
C7688B
C2458G
C1317G
C4530G
C4292B
C7342B
C5761G
C7172C
C6074G
C2629B
P8707C
P7697C
P6897C
P8531C
P7919C
P5640C
P7232C
P2626C
P3013C
P7345C
P9550C
P8627C
P5762C
P2639C
P8854C
P7890C
P8874C
P2771C
P6484C
mathach
ses

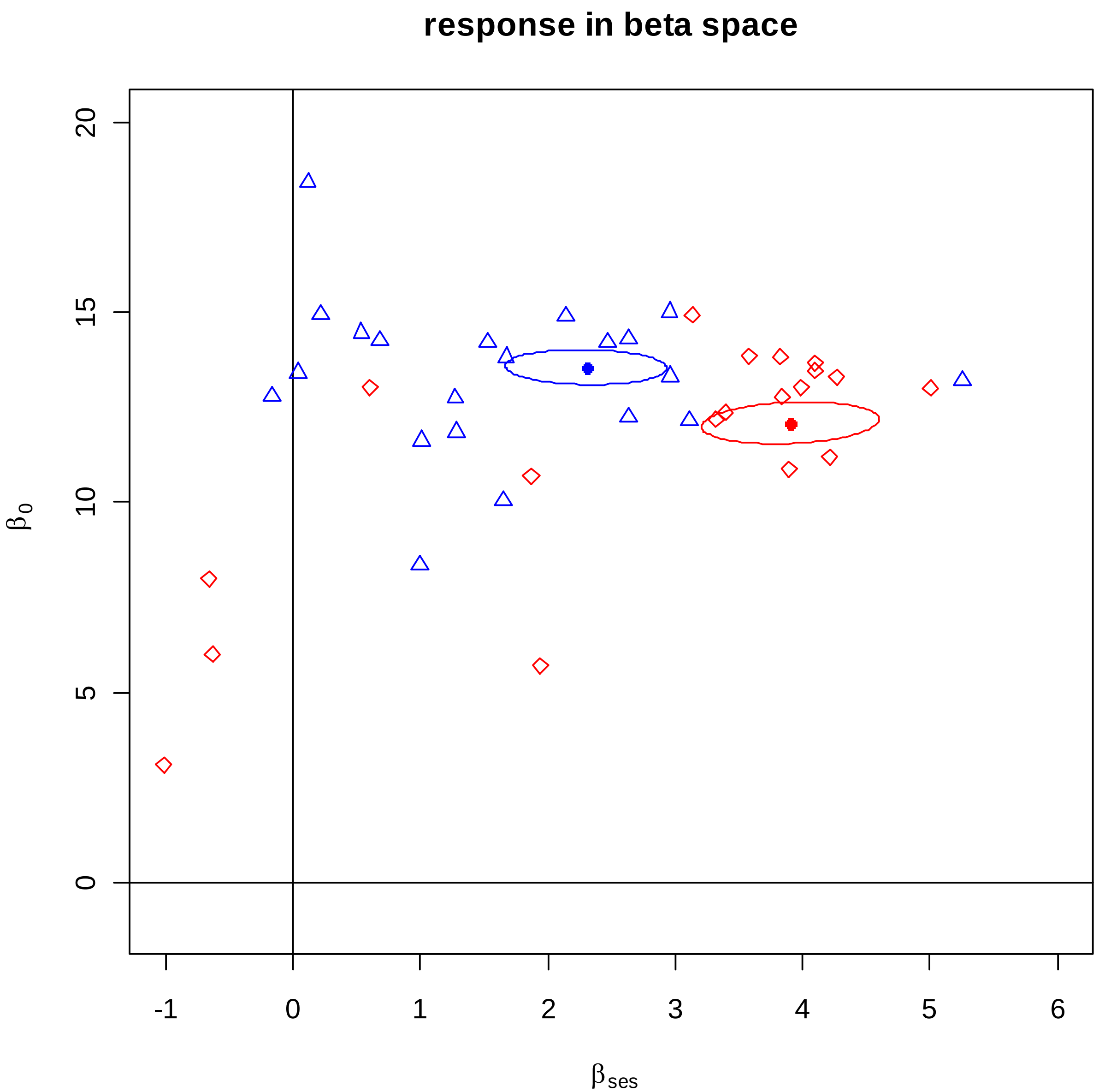


Figure 5: Pooled data estimates with CE plus estimated line for each school
Estimated lines for each sector using pooled data + estimated line for each school

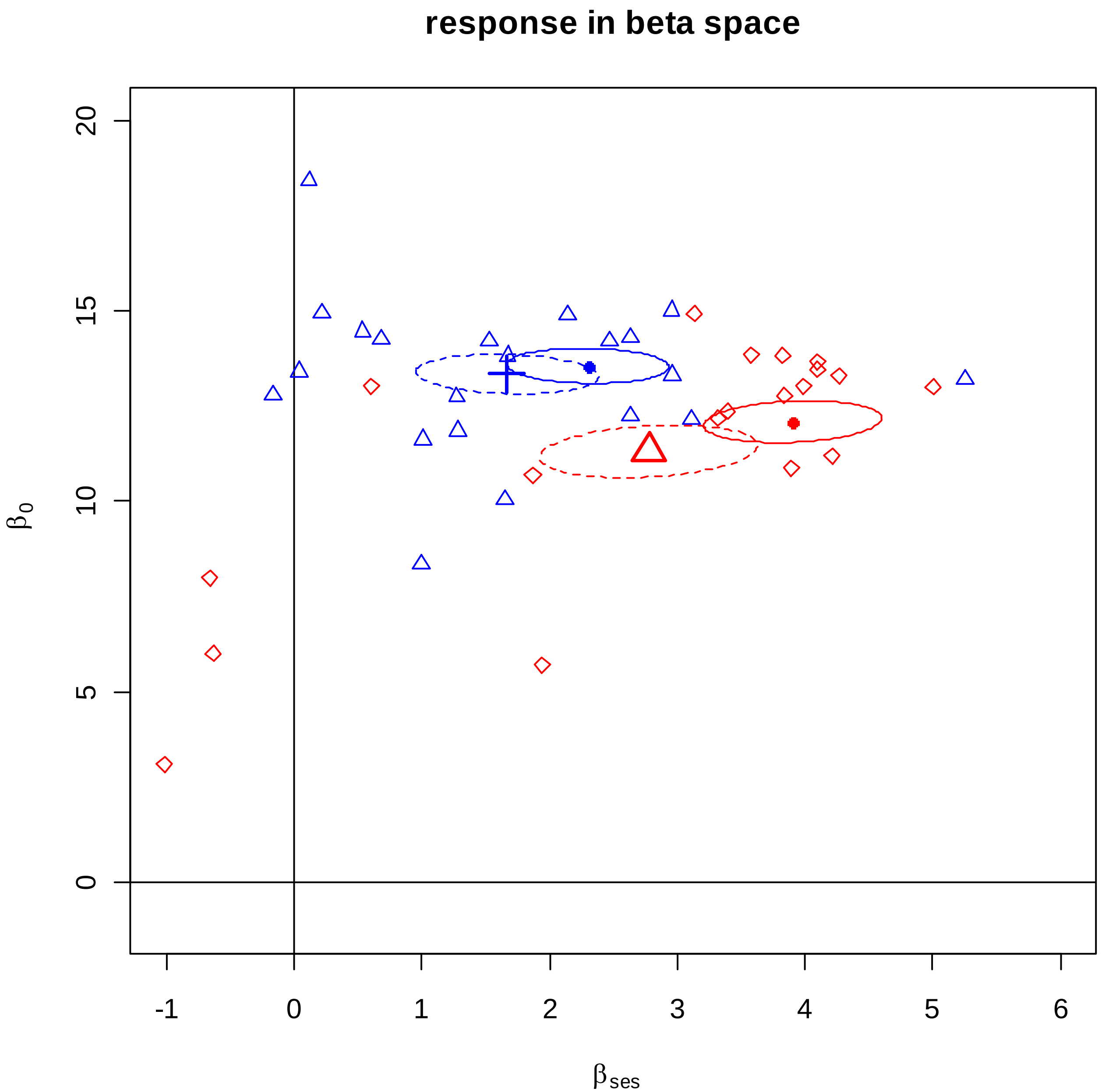


Figure 6: Adding Sector means and CEs based on averaging the estimate for each school

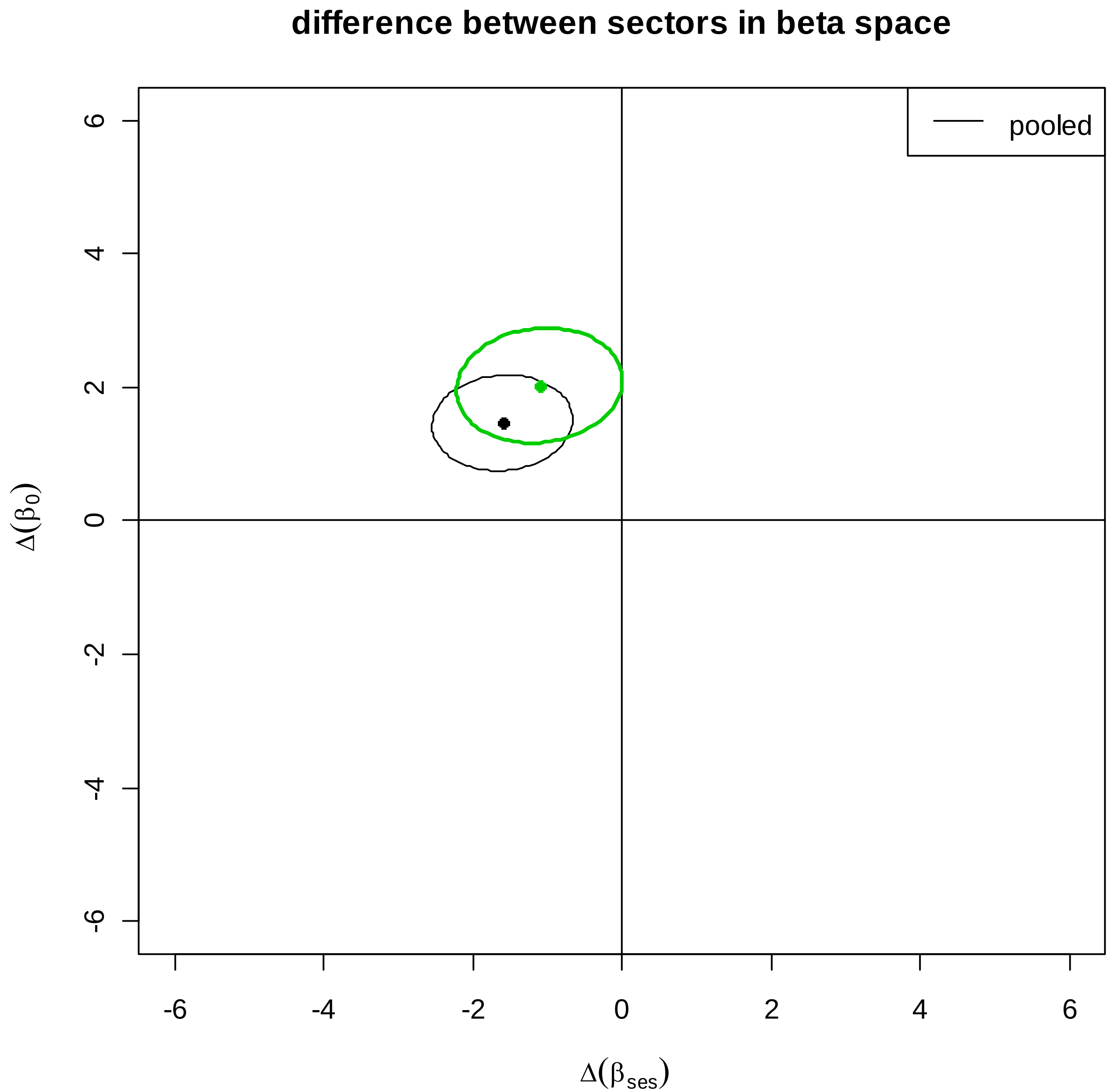


Figure 7: Adding CE based on average of schools

What is the problem with this?
The estimated std. error depends ONLY on within school variability

In other words if we moved the individual school arbitrarily far apart we would still have the same CE for the Sector effect.

Principle of marginality $\subset$ Principle of invariance:
Things that shouldn't matter, shouldn't matter!

'Principle of variance':
Things that should matter, should matter!

If this method gives us exactly the same answer regardless of the between school variability that signals that the SE can not generalize to the population of schools -- only to the putative population of new student samples within these PARTICULAR schools.

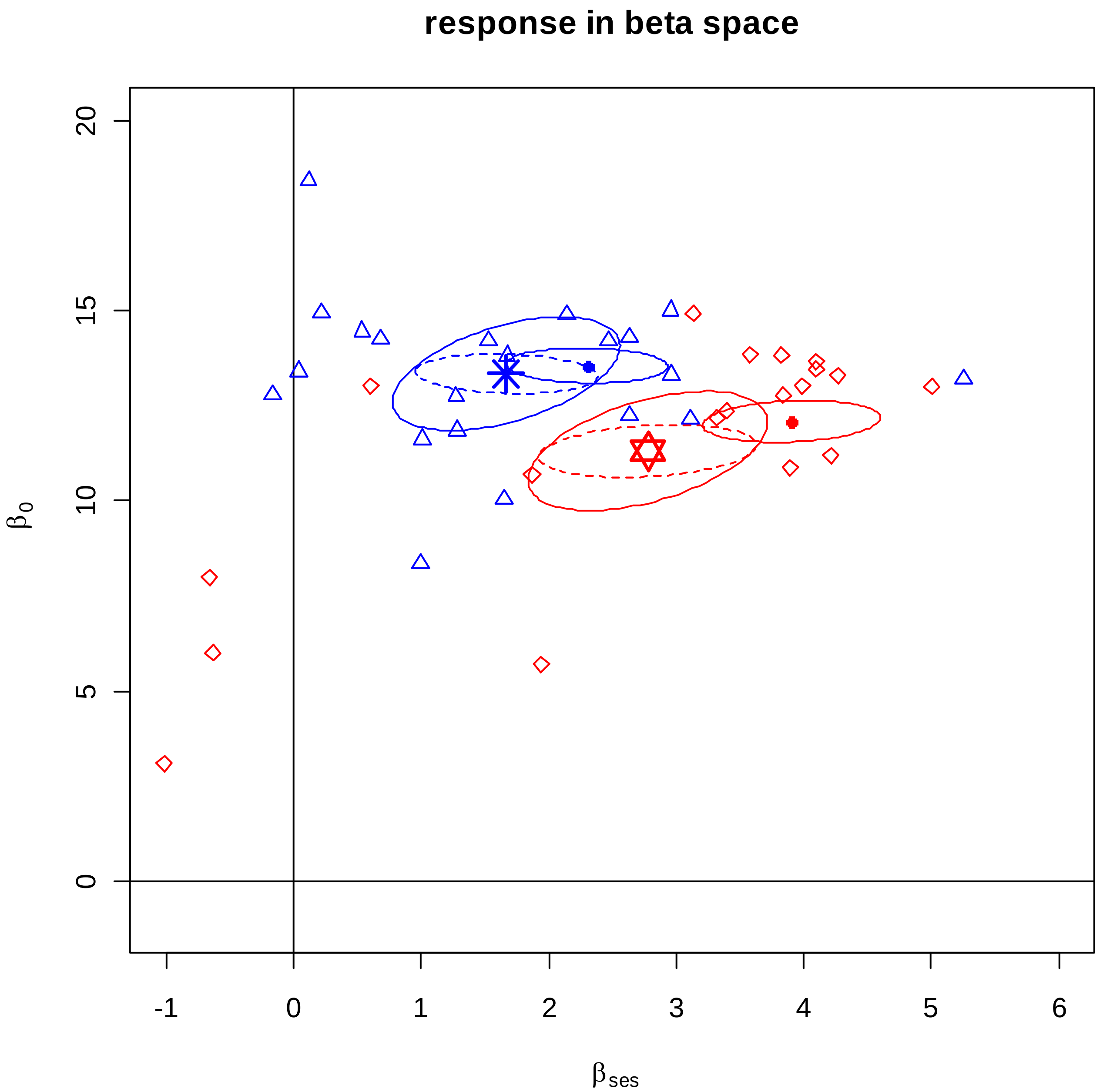
We ignored that we shouldn't ignore? The between school variation.

Method 3: Two-stage approach or 'derived variables' approach

Idea:

First: Estimate slope and intercept within each school as we did in Method 2.

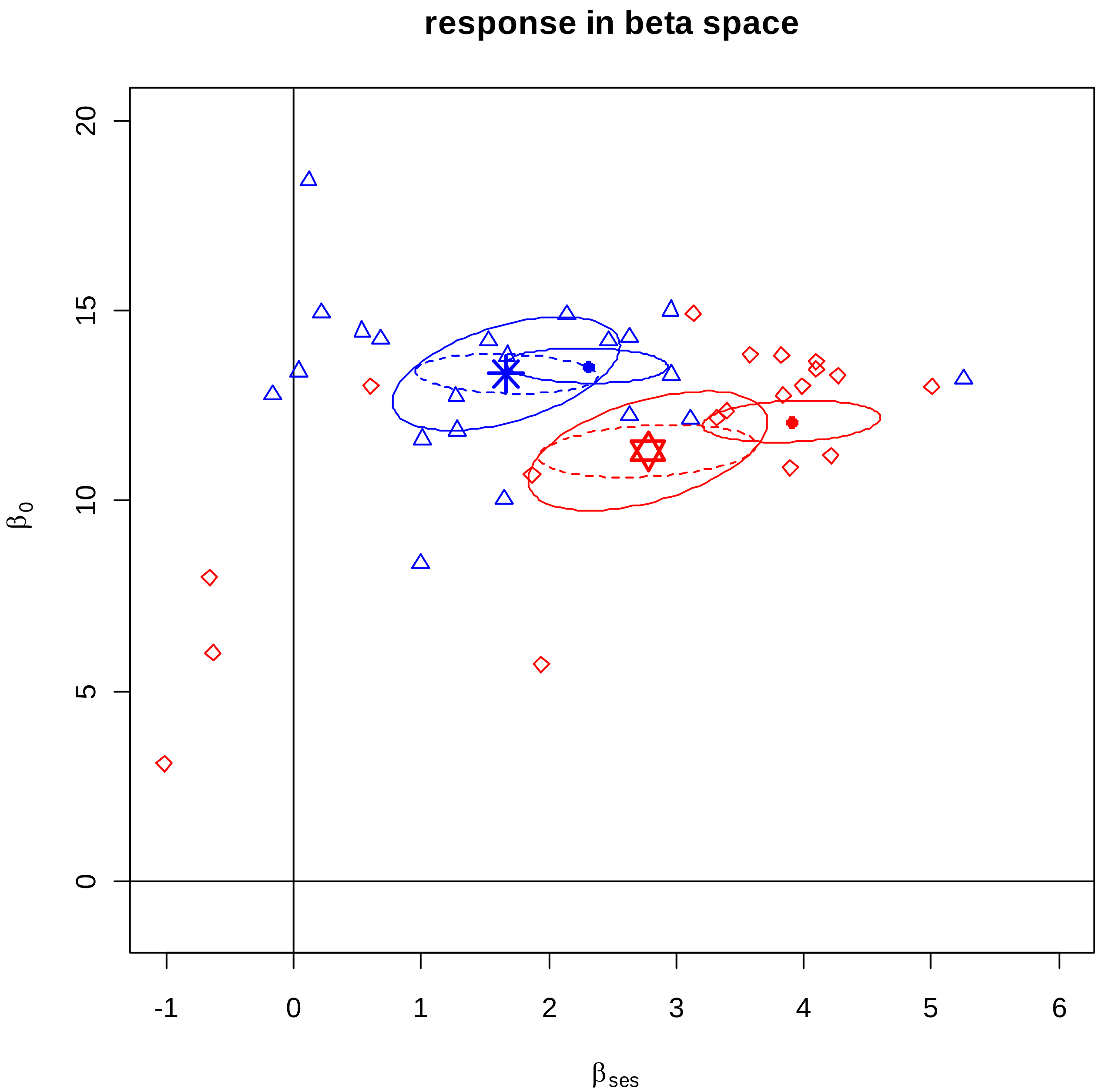
Second: Use the estimated slopes and intercepts as a multivariate sample and do a MANOVA test of equality of the two sector means.



Dashed ellipse was obtained from fixed effects model.

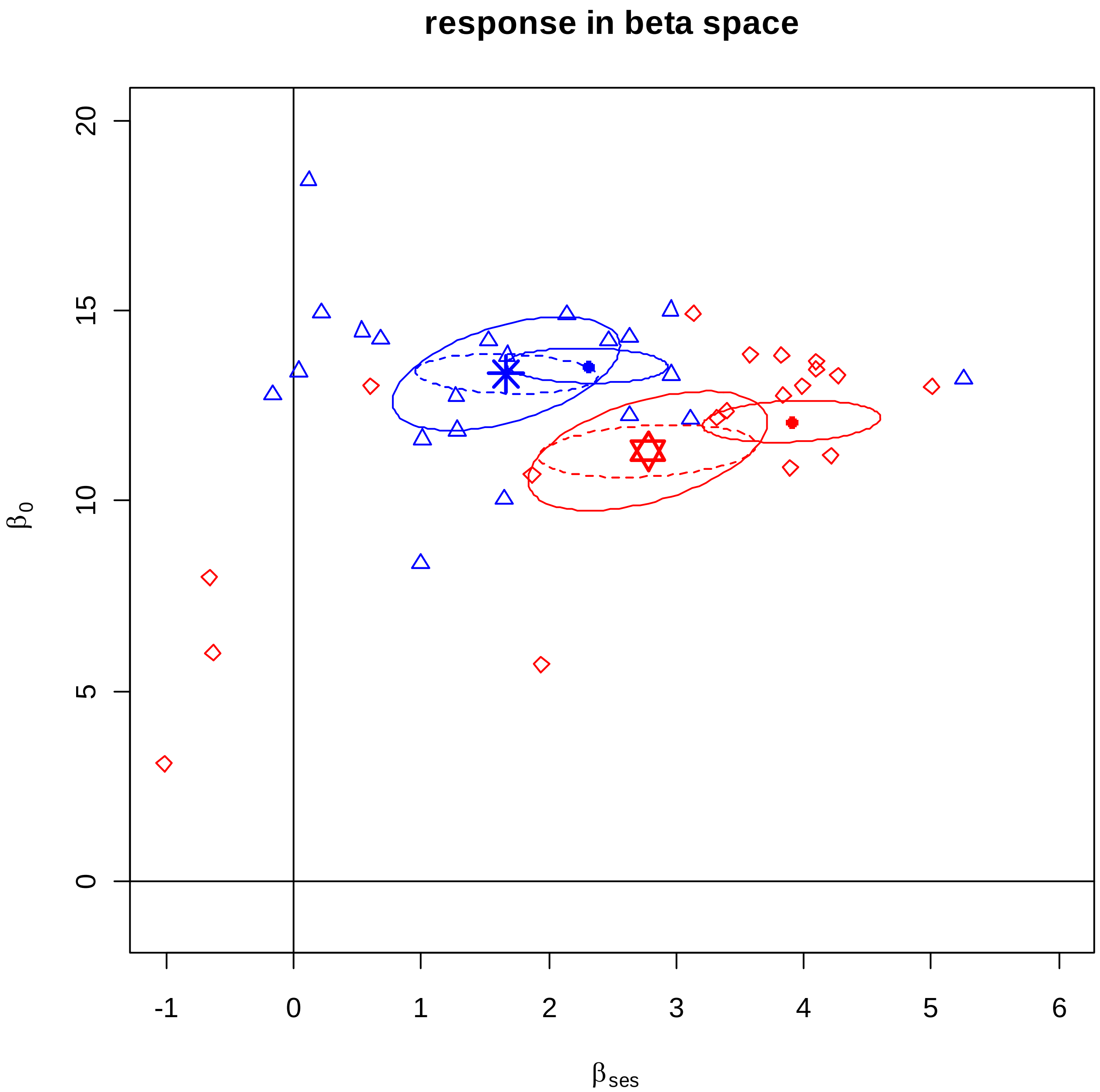
Solid ellipse with same center, from Manova model.

Note:
1) both have the same centre
2) the latter is larger because it generalizes to new samples with MORE variability.



The fixed effects model 95% CE is valid for new samples of students from the same schools. It does not generalize to the population of schools.

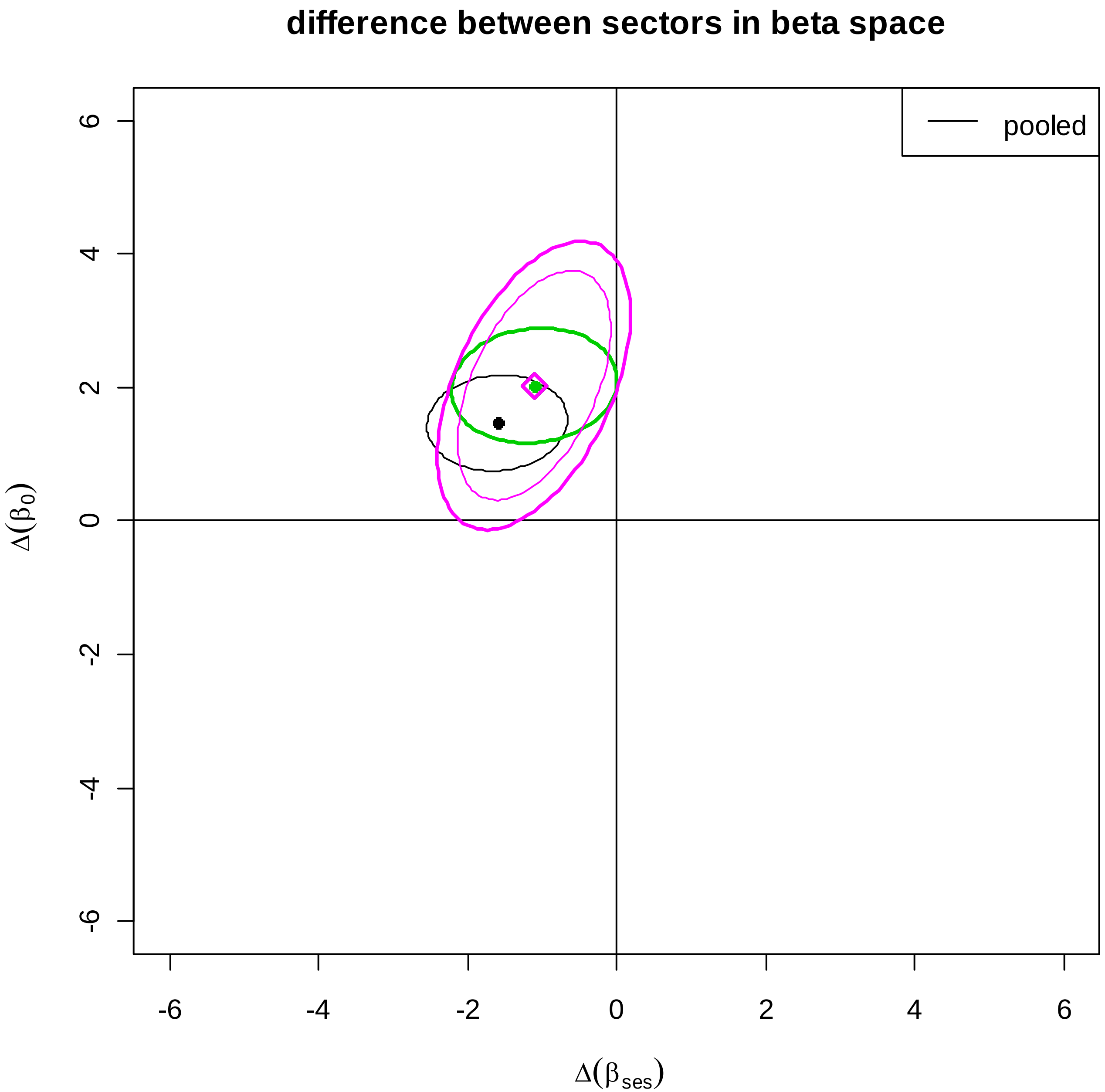
The Manova model **does** generalize to new schools



Disadvantages
(often small):

- gives equal weight to all schools regardless of information in sample (n, spread of ses

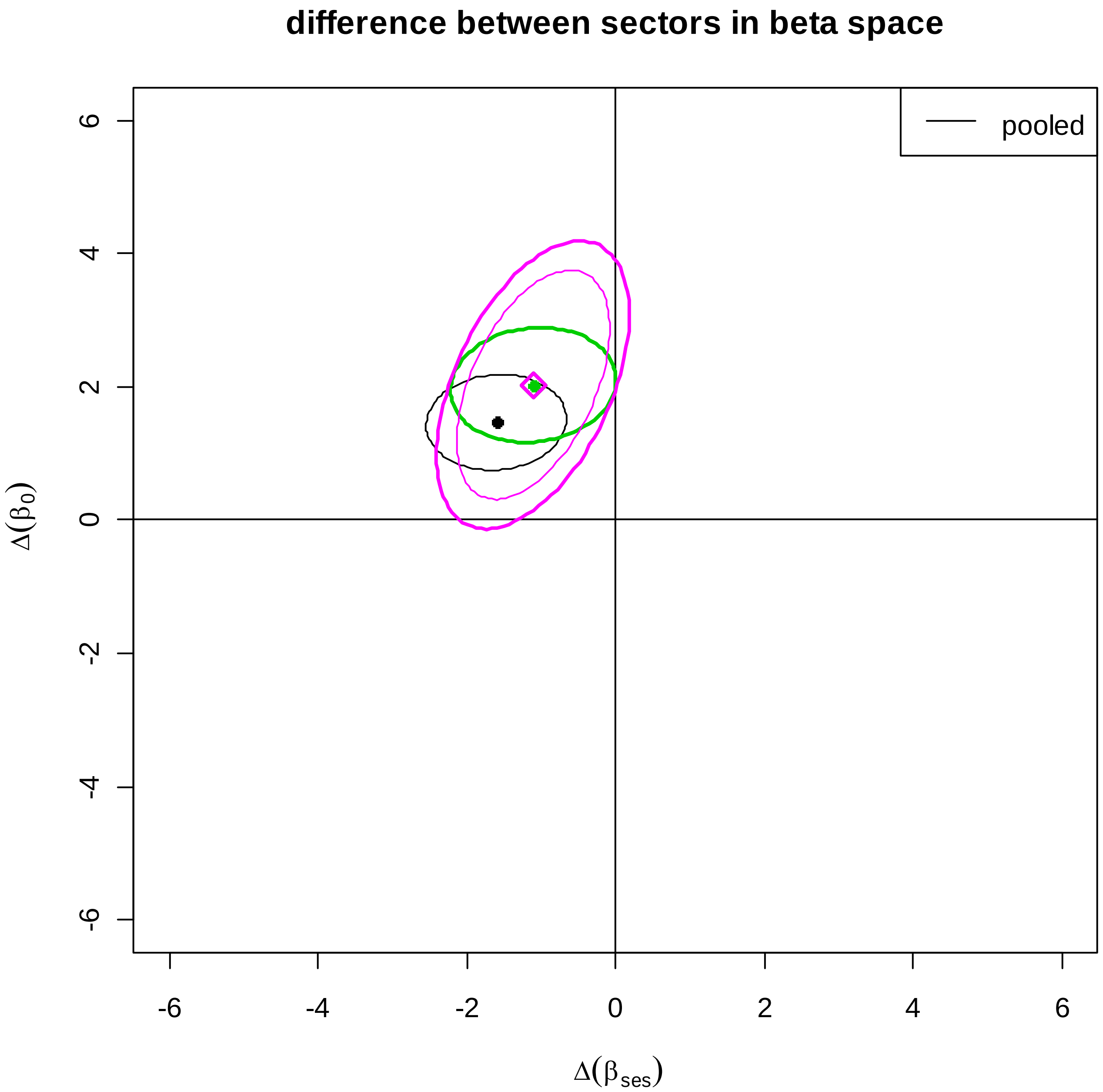
- need to discard data from schools where there are too few points to fit a model (here if n=1)



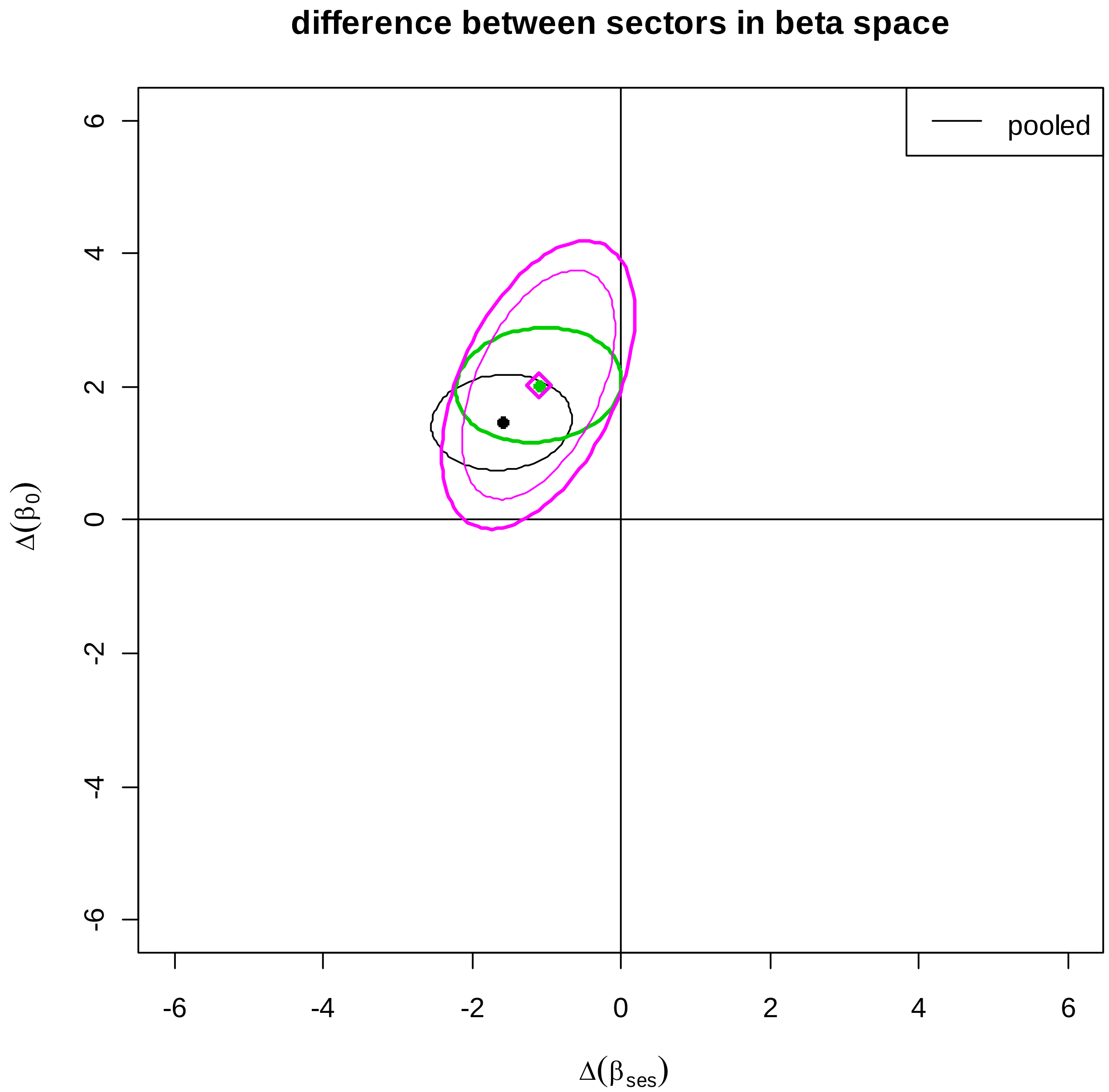
The magenta ellipses are based on the Manova model.

The large magenta ellipse has approximate 95% coverage.

The smaller ellipse has 95% shadows. Thus the p-value for the difference in the effect of ses in the two sectors would be just below 0.05.



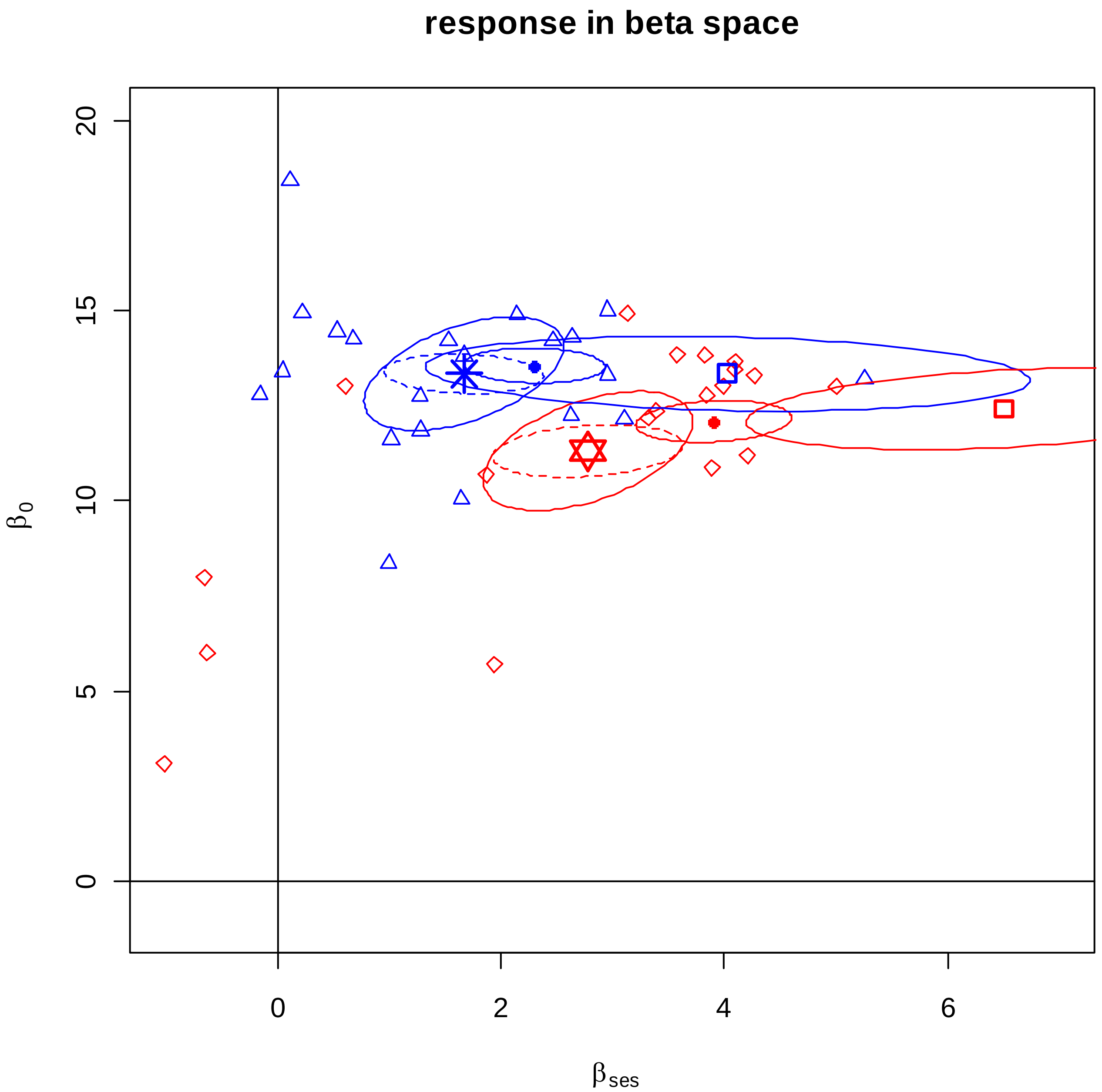
The magenta ellipse generalizes to the sectors, the green ellipse only to new students from the same set of schools.



The magenta ellipses are based on the Manova model.

The large magenta ellipse has approximate 95% coverage.

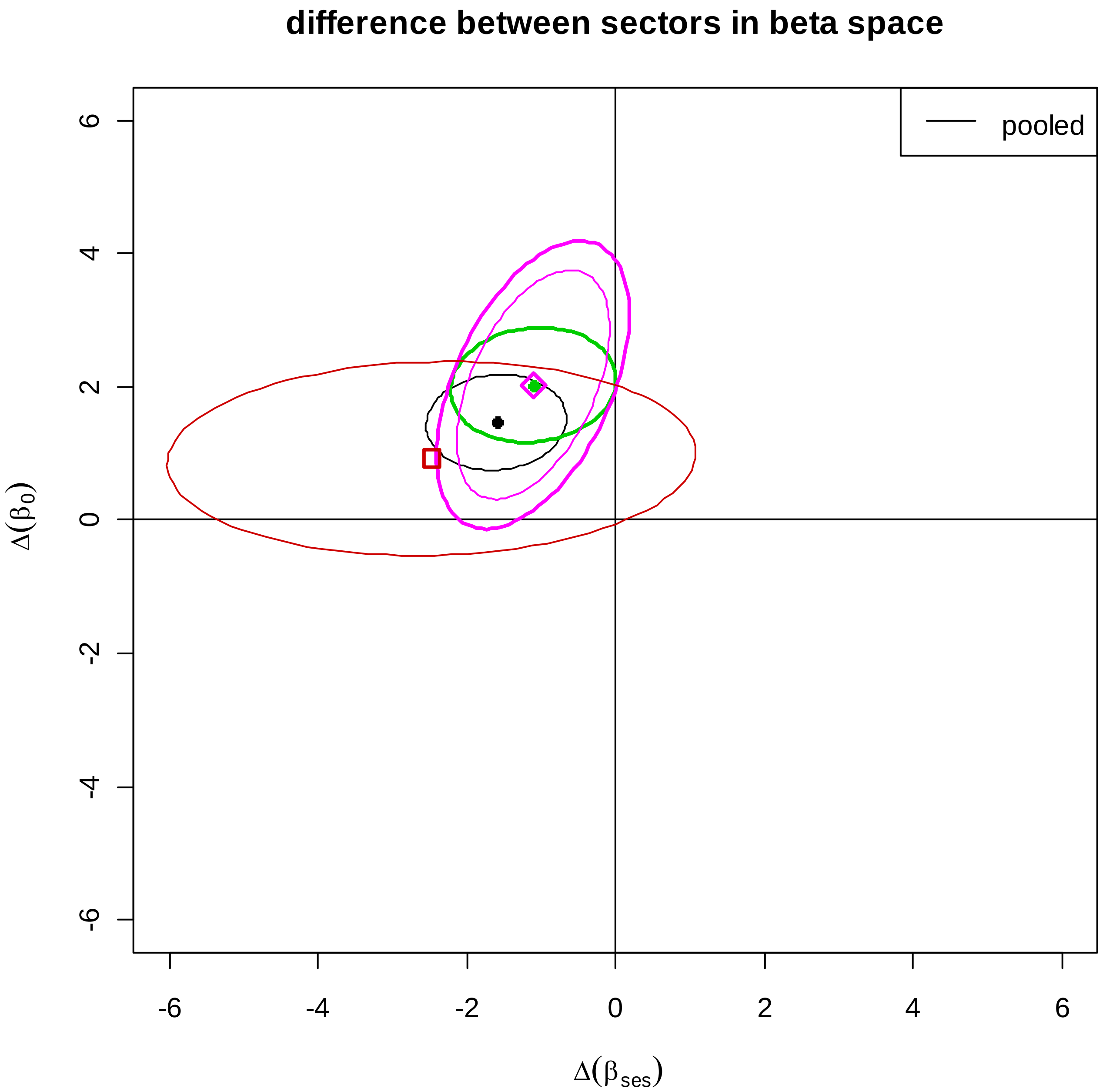
The smaller ellipse has 95% shadows. Thus the p-value for the difference in the effect of ses in the two sectors would be just below 0.05.



Adding the between school model

Note that the pooled estimates are somewhere between the 'within school estimates' and the 'between school estimates'

Method 4: The between-school model



Note that the pooled estimate of differences also lies on an arc between the 'within estimate' and the 'between school' estimate

Why do we get three estimates?

Because there are three effects of ses:

1. **between schools: ecological association**
2. **within schools: conditional association**
3. **across schools: marginal association**

Within school effect

Between school effect

Pooled effect (across schools)

Interesting fact:

Paradoxes of Regression:

Robinson's Paradox refers to the fact that

β_W and β_B can have different signs.

Simpson's Paradox refers to the fact that

β_W and β_P can have different signs.

Some Fallacies of Regression:

Ecological fallacy consists in estimating β_B and believing you have estimated β_W.

Atomistic fallacy consists in estimating β_W and believing you have estimated β_B.

Summary of methods

Method	***Consistent?***	***Efficient?***	***Honest?***
Pooled data	Estimates a combination of β_W and β_B in each sector.	For what?	No. Does not take clustering into account. You might have far fewer independent pieces of information than you think.
Fixed effects	Estimates β_W in each sector.	Yes	Only generalizes to new students from the same fixed set of schools. Does not generalize to the population of schools in each sector, i.e. to the sectors themselves. Reported SE likely to be too small to generalize to new schools

Method	***Consistent?***	***Efficient?***	***Honest?***
2-step method: derived variables, **regress then average**	Estimates β_W in each sector.	No[2] – unless size and spread of ses is similar in each cluster. Does not give more weight to schools with more information (n or spread of ses)	Yes. Generalizes to the population of schools.

[2] Although the estimate may be similar to the fixed-effects estimate because they both estimate the same thing, it is not, in general, equal because the two estimates give different weight to each school's estimated slope.

Method	*Consistent?*	*Efficient?*	*Honest?*
Ecological or Between School analysis: **average then regress**	Estimates β_B in each sector. Note that we are generally really interested in β_W	No – does not take differences in sample size and spread of data into account but it would be easy to do so.	Yes.

Hierarchical Models

Method	***Consistent?***	***Efficient?***	***Honest?***
HLM	Yes under common tacit but unrealistic supposition that $\beta_B = \beta_W$ Otherwise the estimate is, like the pooled estimate, a combination of β_W and β_B in each sector. But will be closer – generally much closer – to β_W than the pooled estimate. It is consitent for β_w as the cluster size increases – not as the number of clusters increases	Yes	Yes
HLM + contextual variable	Gives separate consistent separate estimates of β_w and β_B.	Yes	Yes

Review of the matrix formulation of regression

You don't need to understand this in depth to use HLMs but it's useful to know where many of the results come from. If you already know regression formulated with matrices, then it's easier to see how to make the jump from OLS regression to HLM regression.

$Y = X\beta + \varepsilon$ is such a universal and convenient shorthand that we need to spell out what it means and how it is used.

Here's the equation for a single observation assuming 2 X variables:

$$Y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i \qquad j = 1, \cdots, N$$

with ε_i iid $N(0, \sigma^2)$.

We pile these equations one on top of the other:

$$
\begin{aligned}
Y_1 &= \beta_0 + x_{11}\beta_1 + x_{21}\beta_2 + \varepsilon_1 \\
Y_2 &= \beta_0 + x_{12}\beta_1 + x_{22}\beta_2 + \varepsilon_2 \\
&\vdots \\
Y_j &= \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i \\
&\vdots \\
Y_N &= \beta_0 + x_{1N}\beta_1 + x_{2N}\beta_2 + \varepsilon_N
\end{aligned}
$$

Note that the βs remain the same from line to line but Ys, xs and εs change. Using vectors and matrices and exploiting the rules for multiplying matrices:

$$
\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & & \\ 1 & x_{1N} & x_{2N} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}
$$

or, in short-hand:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

In multilevel models with, say J schools indexed by $j = 1, \dots, J$ and with the jth school having n_j students, we block students of the same school together. We just add js to show that this is the jth school. The big difference is that the βs might change from school to school and that the sample size can change from one school to the next. So we use n_j to denote the sample size for the jth school:

$$\begin{bmatrix} Y_{1j} \\ Y_{2j} \\ \vdots \\ Y_{n_i j} \end{bmatrix} = \begin{bmatrix} 1 & x_{11j} & x_{21j} \\ 1 & x_{12j} & x_{22j} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n_j j} & x_{2n_j j} \end{bmatrix} \begin{bmatrix} \beta_{0j} \\ \beta_{1j} \\ \beta_{2j} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1j} \\ \varepsilon_{2j} \\ \vdots \\ \varepsilon_{n_j j} \end{bmatrix}$$

or, in short hand:

$$\mathbf{Y}_j = \mathbf{X}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j$$

We can stack schools on top of each other. If all schools are assumed to have the same value for $\boldsymbol{\beta}_j = \boldsymbol{\beta}$, then we can stack the **Xs** vertically:

$$\begin{bmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_j \\ \vdots \\ \mathbf{Y}_J \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_j \\ \vdots \\ \mathbf{X}_J \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_j \\ \vdots \\ \boldsymbol{\varepsilon}_J \end{bmatrix}$$

or, in shorter form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

If the $\beta_j s$ are different we can stack the $X_j s$ diagonally:

$$\begin{bmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_j \\ \vdots \\ \mathbf{Y}_J \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \cdots & \vdots \\ 0 & \cdots & \mathbf{X}_j & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \vdots & 0 & \cdots & \mathbf{X}_J \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_j \\ \vdots \\ \boldsymbol{\beta}_J \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_j \\ \vdots \\ \boldsymbol{\varepsilon}_J \end{bmatrix}$$

or, in shorter form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

again!

Something that gets used over and over again is the fact that if $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I})$, i.e. all ε's are independent and normal with the same variance then the best estimator of $\boldsymbol{\beta}$ is the OLS (ordinary least-squares) estimator:

$$\widehat{\boldsymbol{\beta}}^{\mathbf{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

with variance

$$\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

If the components of ε are not iid but $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a known variance matrix (or, at least, known up to a proportional factor) then the GLS (generalized least-squares) estimator is:

$$\widehat{\boldsymbol{\beta}}^{GLS} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$$

with variance

$$(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}.$$

The Hierarchical Model

We develop the ideas for mixed and multilevel modeling in two stages:

1. Multilevel models as presented in Bryk and Raudenbush (1992) in which the unobserved parameters at the lower level are modeled at the higher level. This is the representation used in HLM, the software developed by Bryk and Raudenbush and, to a limited extent in MLwiN.

2. Mixed models in which the levels are combined into a combined equation with two parts: one for ‘fixed effects’ and the other for ‘random effects.’ This is the form used in R, SAS and in many other packages.

Although the former is more complex, it is more natural and and intuitive. It also gives us important insights into the structure of these models.

We will use the high school Math Achievement data for an extensive example. We think of our data as structured in two levels: **students within schools** and **between schools**.

We also have two types of predictor variables:

1. **within-school Level 1variables:** Individual student variables: SES, Sex, individual minority status. These variables are also known by many other names, e.g. inner variables, micro variables, level-1 variables[3], time-varying variables in the longitudinal context.

2. **between-school Level 2 variables:** Sector: Catholic or Public, school meanses, size, mean ses of sample, sample size. These variables are also known as outer variables, macro variables, level-2 variables, or time-invariant variables in a longitudinal context. A between-school variable can be created from a within-school variable by taking the

[3] In some hierarchical modeling traditions, e.g. R, the numbering of levels is reversed going from the top down instead of going from the bottom up. One needs to check which approach an author or package is using.

average of the within-school variable within each school. Such a derived between-school variable is known as a ‘contextual’ variable. These variables are useful only if the average differs from school to school. Balanced data in which the set of values of within-school variables is be the same in each school does not give rise to contextual variables.

Basic structure of the model:

1. Each school has a true regression line that is not directly observed

2. The observations from each school are generated by taking random observations generated with the school’s true regression line

3. The true regression lines for each school come from a population or populations of regression lines

Within School model:

For school i: (For now we suppose all schools come from the same population, e.g. only one Sector)

1) True but unknown $\boldsymbol{\beta}_j = \begin{bmatrix} \beta_{0j} \\ \beta_{\text{SES}\,j} \end{bmatrix} = \begin{bmatrix} \beta_{0j} \\ \beta_{1j} \end{bmatrix}$ for each school

2) The data are generated as

$$Y_{ij} = \beta_{0j} + \beta_{0j} X_{ij} + \varepsilon_{ij}$$

$$\varepsilon_{ij} \sim N(0, \sigma^2) \text{ independent of } \boldsymbol{\beta}_j\text{'s}$$

Between School model:

We start by supposing that the $\boldsymbol{\beta}_j = \begin{bmatrix} \beta_{0j} \\ \beta_{\mathrm{SES}\,j} \end{bmatrix} = \begin{bmatrix} \beta_{0j} \\ \beta_{1j} \end{bmatrix}$ are sampled from a single population of schools. In vector notation:

$$\boldsymbol{\beta}_j = \boldsymbol{\gamma} + \mathbf{u}_j \quad \mathbf{u}_j \sim N(\mathbf{0}, \mathbf{G})$$

where

$$\mathbf{G} = \begin{bmatrix} g_{00} & g_{10} \\ g_{10} & g_{11} \end{bmatrix}$$

is a variance matrix.
Writing out the elements of the vectors:

$$\boldsymbol{\beta}_j = \begin{bmatrix} \beta_{0j} \\ \beta_{1j} \end{bmatrix} = \begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix} + \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix}, \qquad \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} g_{0} \\ g_{1} \end{bmatrix} \right.$$

Note:

$$\mathrm{Var}(\beta_{0i}) = g_{00}$$
$$\mathrm{Var}(\beta_{1i}) = g_{11}$$
$$\mathrm{Cov}(\beta_{0i}, \beta_{1i}) = g_{10} = g_{01}$$

A simulated example

To generate an example we need to do something with SES although its distribution is not part of the model. In the model the values of SES are taken as given constants.

We will take:

$$\boldsymbol{\gamma} = \begin{bmatrix} 12 \\ 2 \end{bmatrix}, \; \mathbf{G} = \begin{bmatrix} 16 & 8 \\ 8 & 25 \end{bmatrix}, \; \sigma^2 = 20$$

Once we have generated $\boldsymbol{\beta}_j$ we generate $N_j \sim Poisson(30)$ and $SES \sim N(0,1)$

Here’s our first simulated school in detail:

For $j=1$:

SES :

```
-1.05 -0.78  1.05 -1.01  0.77  1.85  0.87 -1.18  0.18  2.08 -1.14 -1.71
-0.64 -0.41  0.86  1.29  0.04  0.23  0.90  0.50 -2.10 -1.89  0.38
```

$\boldsymbol{\varepsilon}_j$:

```
4.46 -0.73  0.30  7.63 -7.03  1.20 -6.23 -4.66  6.17  0.75 -1.43  0.46
3.64 -2.39  2.24  2.60  3.96  0.71 -3.74  3.30  4.42 -4.59 -3.61
```

$Y_{ij} = \beta_{0j} + \beta_{1j} SES_{ij} + \varepsilon_{ij}$:

```
14.53   8.09   0.70  17.56  -5.34  -2.10  -4.99   6.03  10.58  -3.59
 9.09  13.57  11.83   4.75   3.51   1.88   9.00   4.91  -2.66   6.24
19.37   9.38  -0.13
```

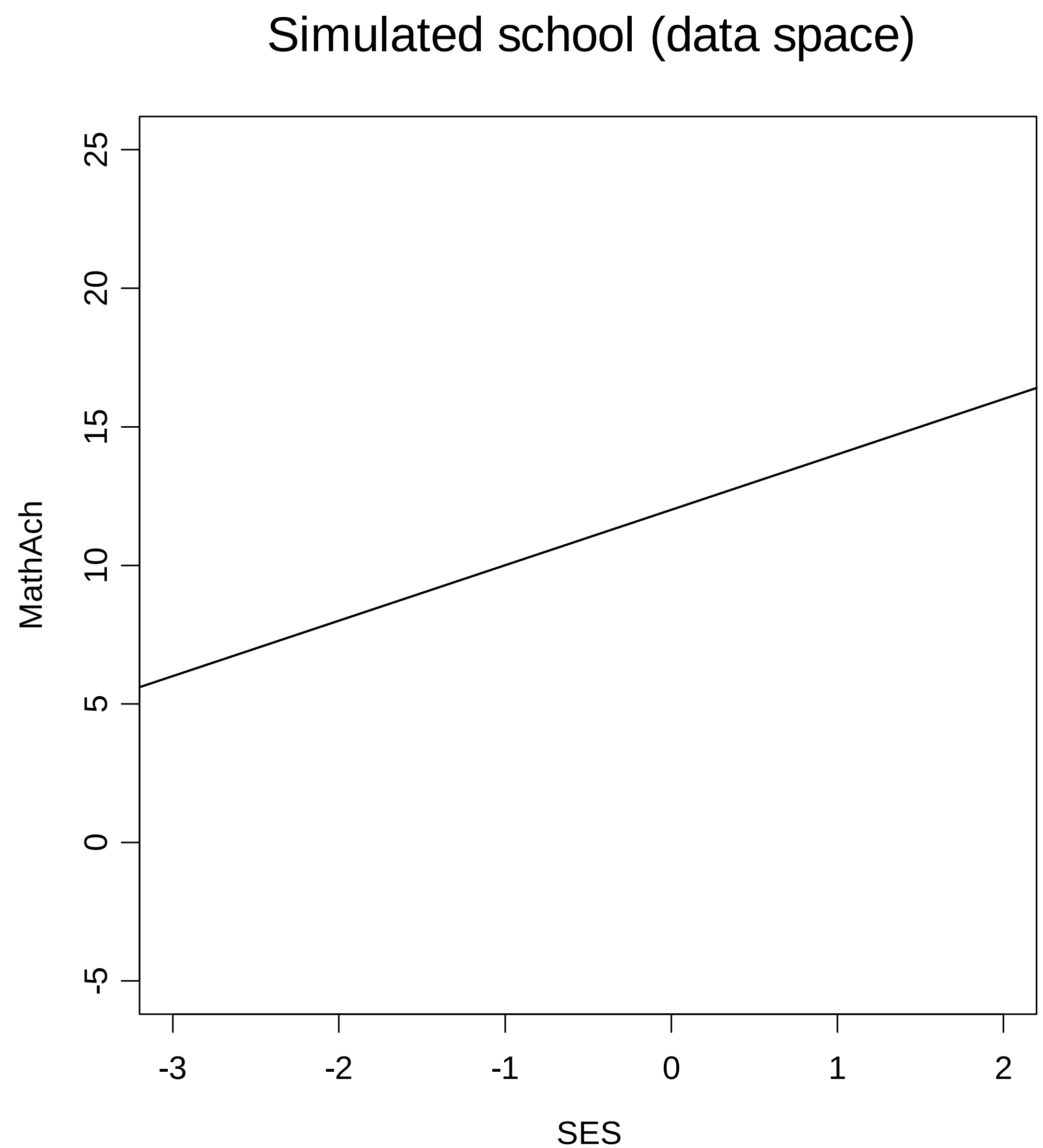


Figure 8: Simulation: mean population regression line γ

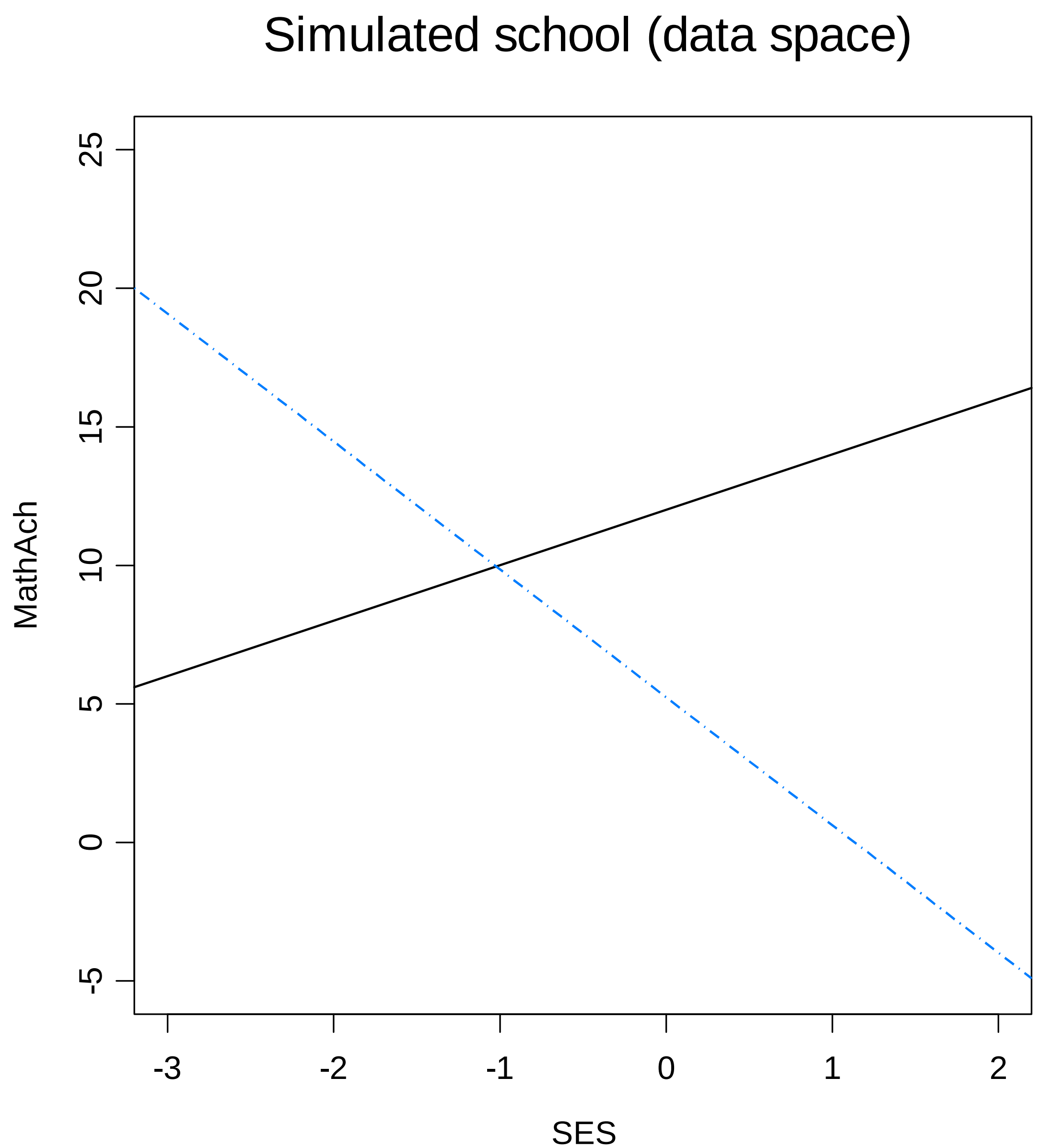


Figure 9: Simulated school: True regression line in School 1: $\boldsymbol{\beta}_j = \boldsymbol{\gamma} + \mathbf{u}_j$

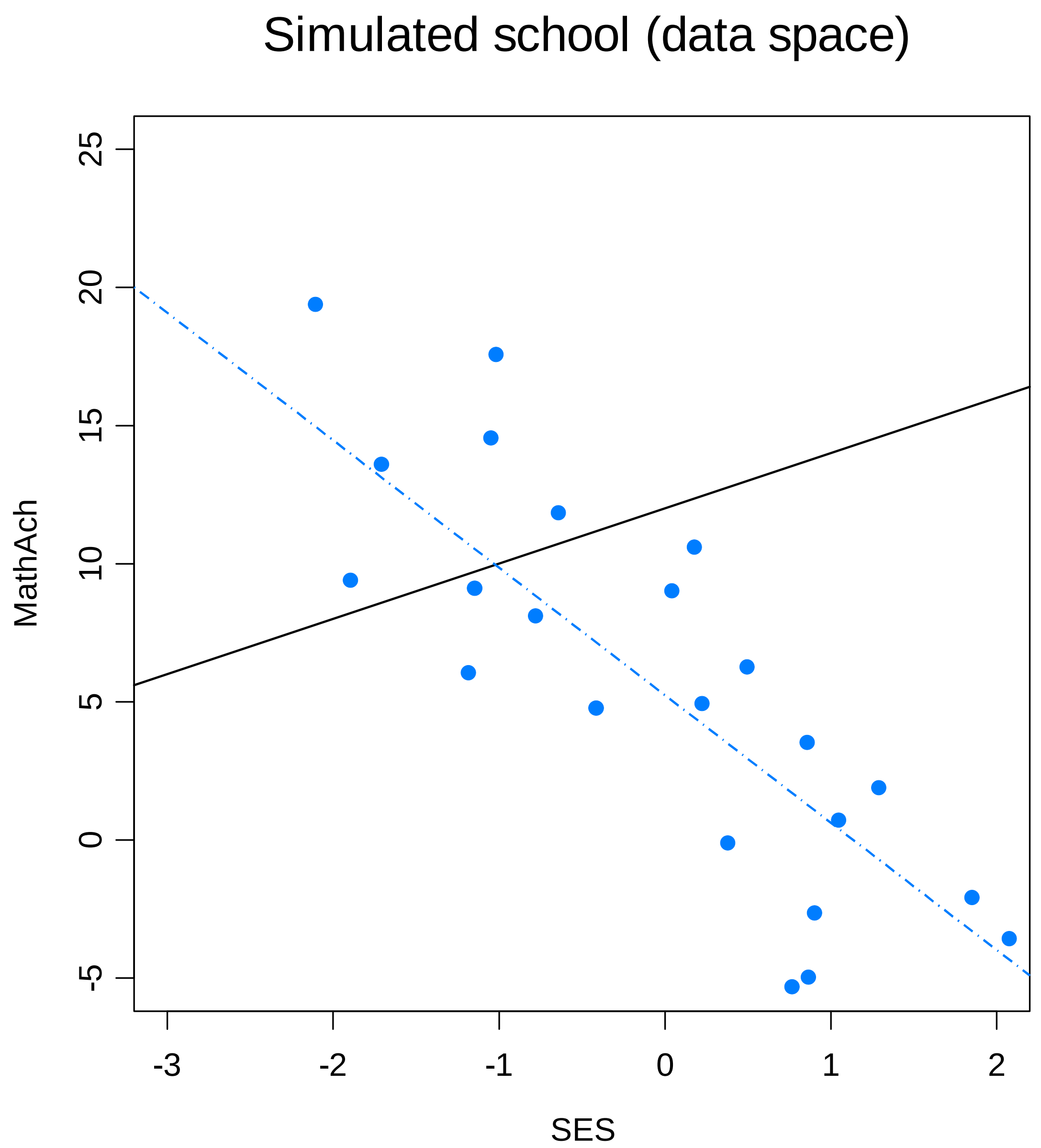


Figure 10: School 1 regression line with data generated by
$Y_{ij} = \beta_{0i} + \beta_{1i} SES_{ij} + \varepsilon_{ij}$

Simulated school (data space)

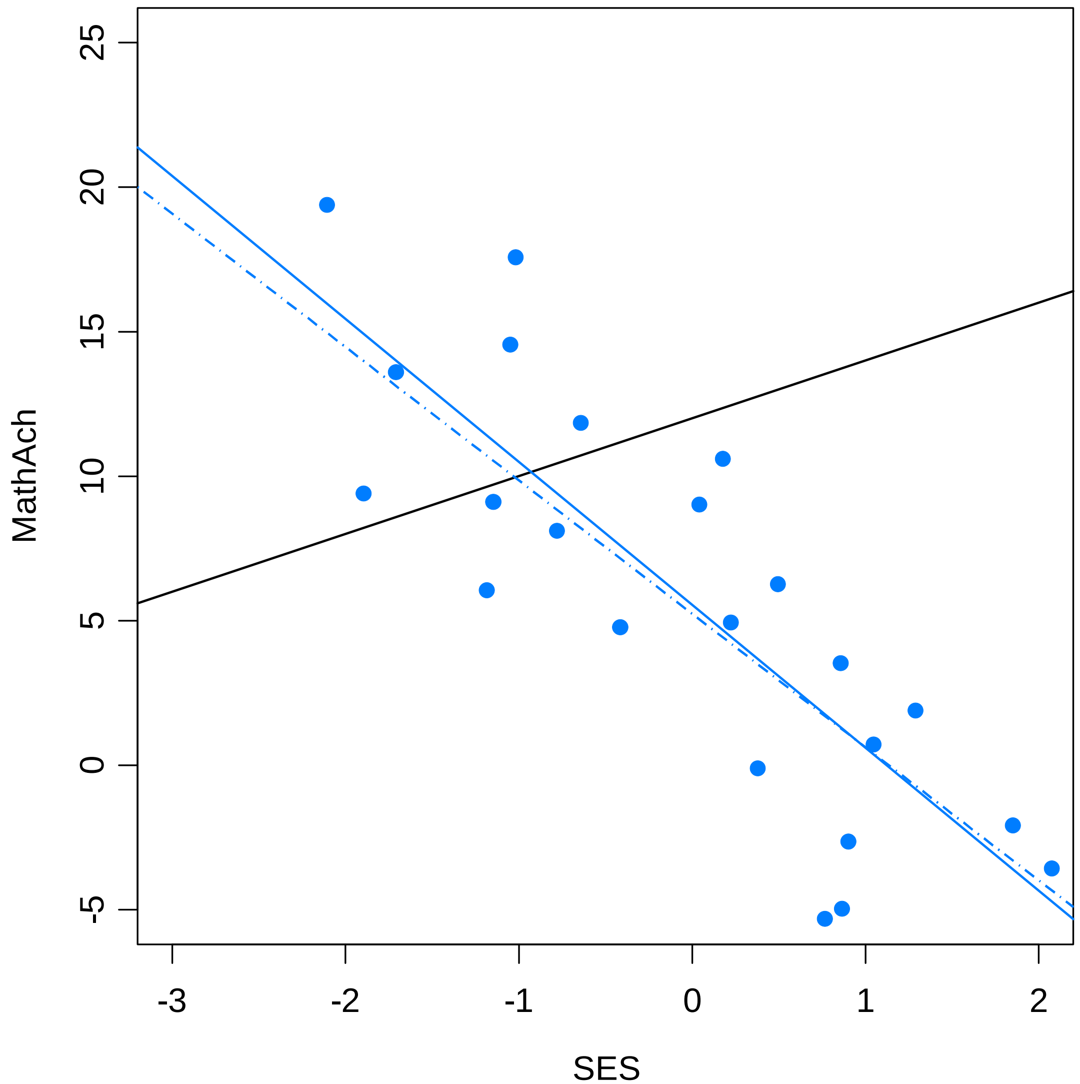


Figure 11: Simulated school: True regression line β_i, data, and least-squares line $\hat{\beta}_i$

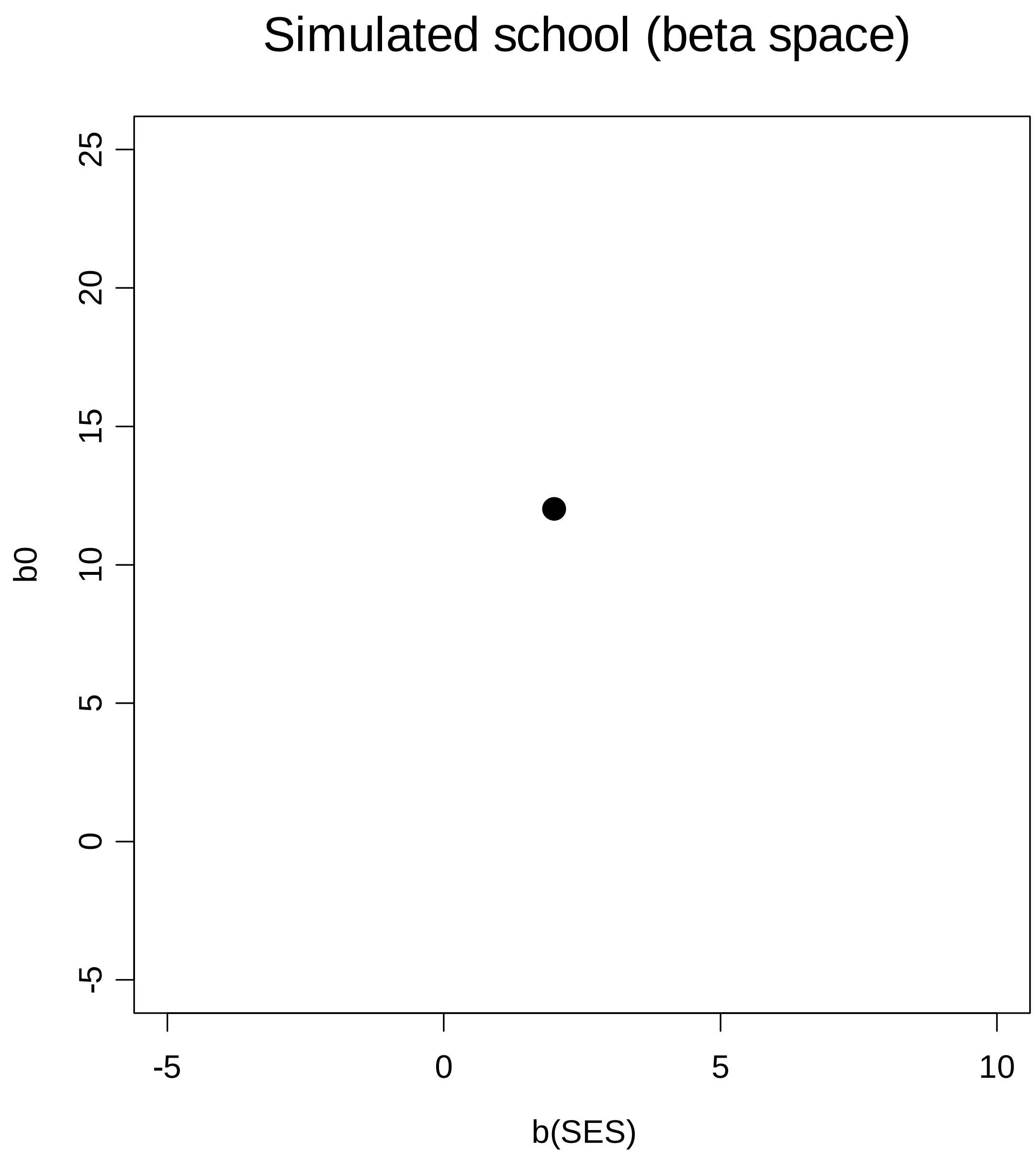


Figure 12: Simulated school in beta space with true mean line represented by a point.

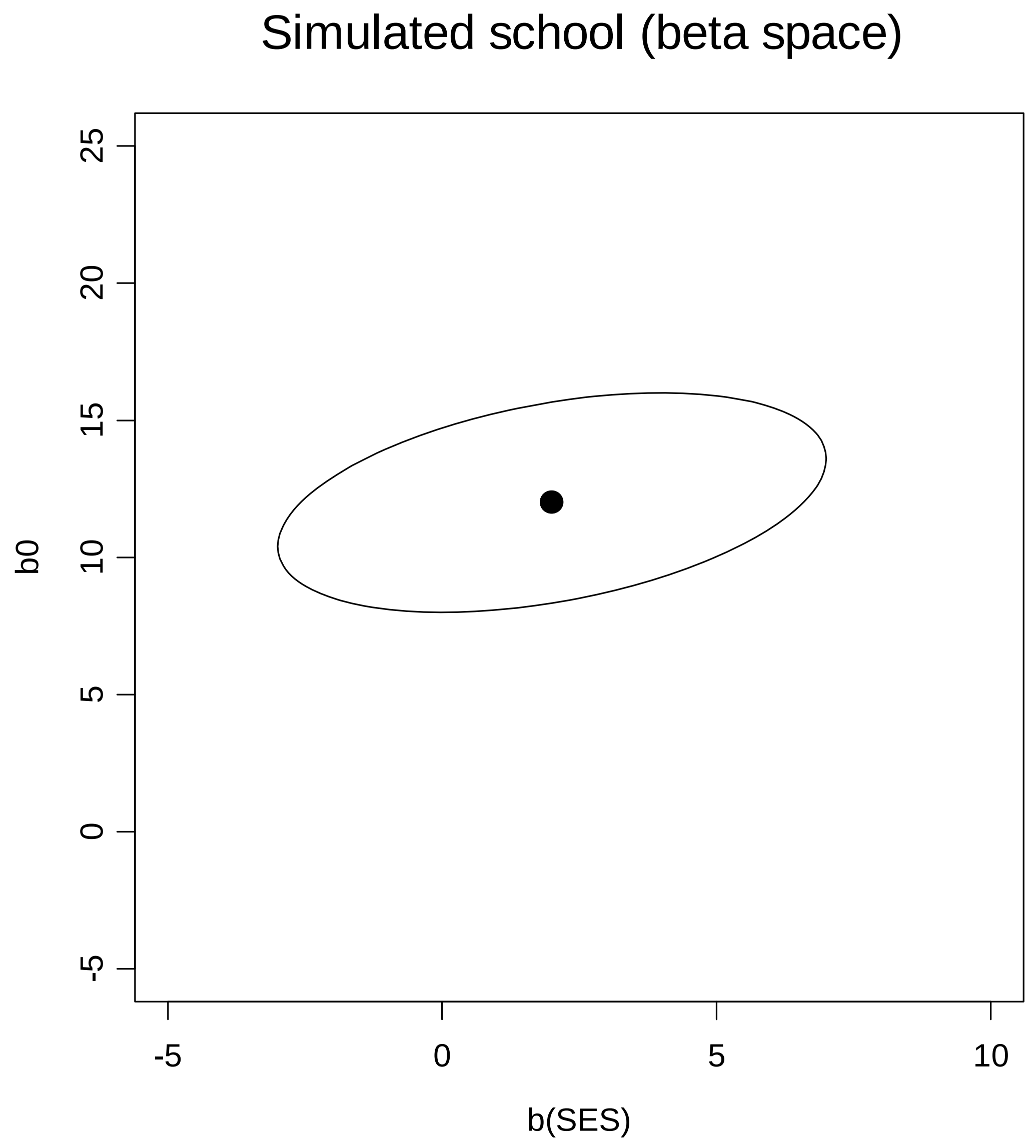


Figure 13: Simulated school: population mean line in beta space with dispersion ellipse with matrix G for random slopes and intercepts. Note that shadows of the ellipse yield the mean plus or minus 1 standard deviation

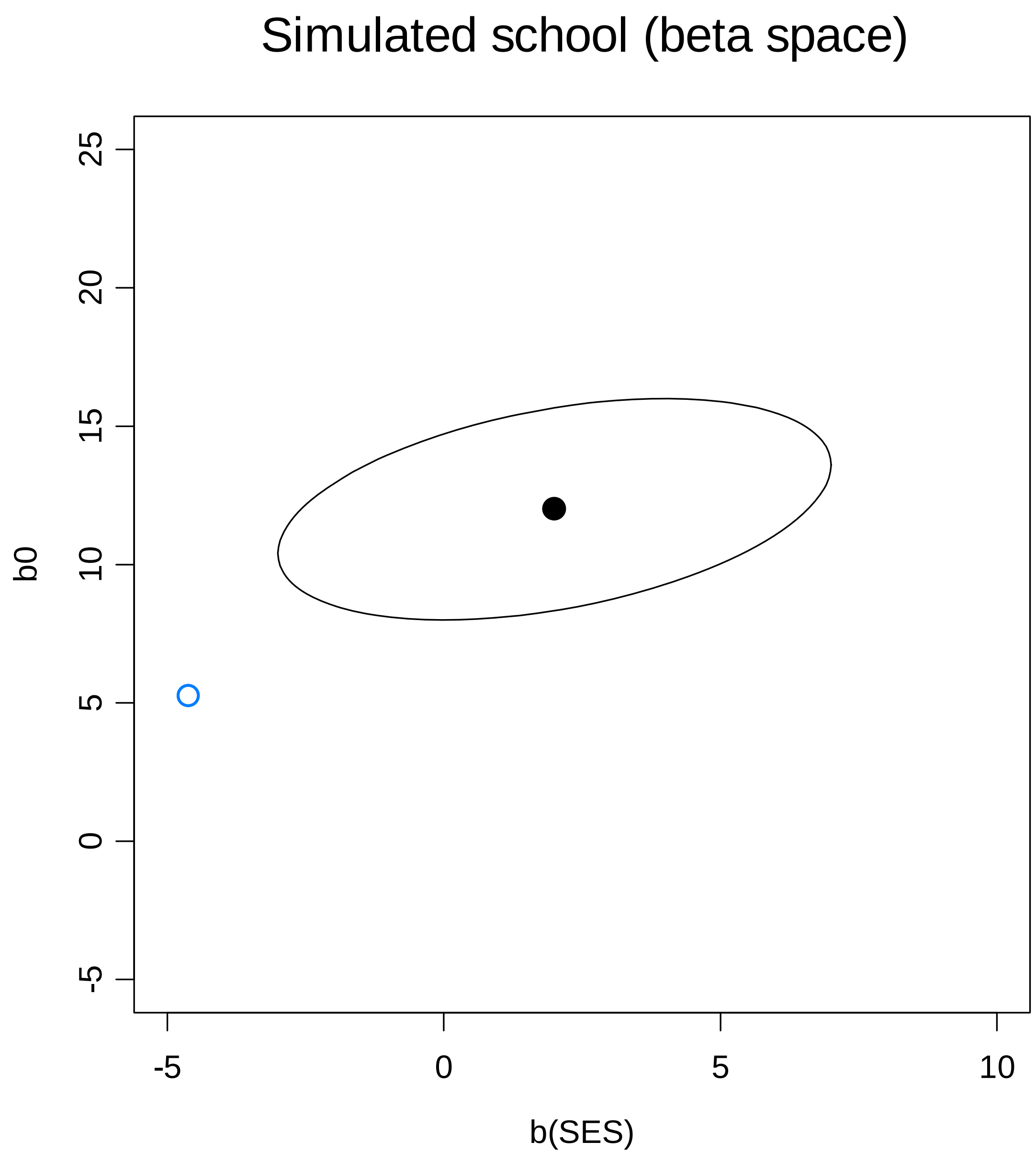


Figure 14: A random ‘true’ intercept and slope from the population. This one happens to be somewhat atypical but not wholly implausible.

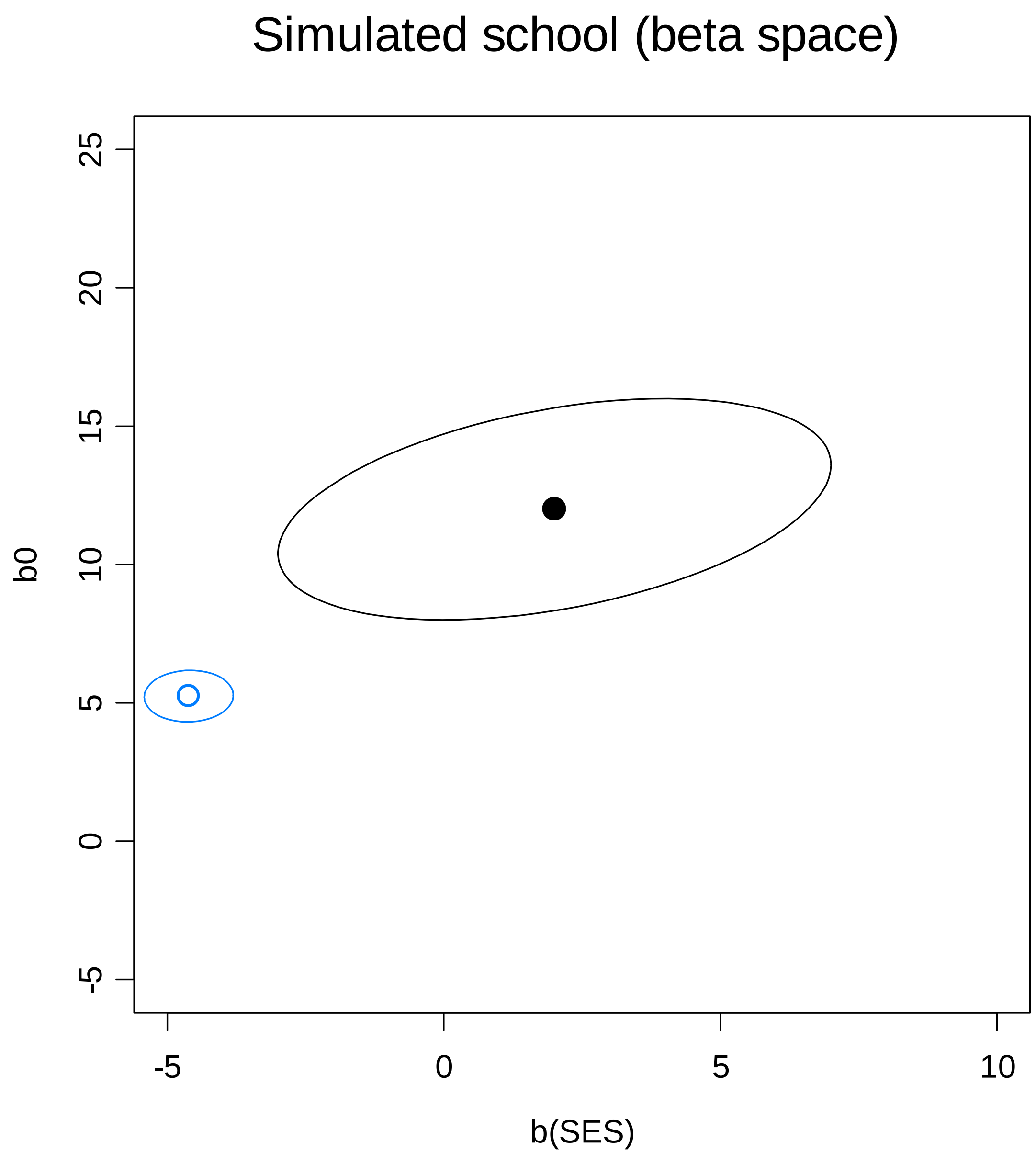


Figure 15: 'True' intercept and slope with dispersion ellipse with matrix $\sigma^2\left(\mathbf{X}_j'\mathbf{X}_j\right)^{-1}$ for $\hat{\boldsymbol{\beta}}_j$.

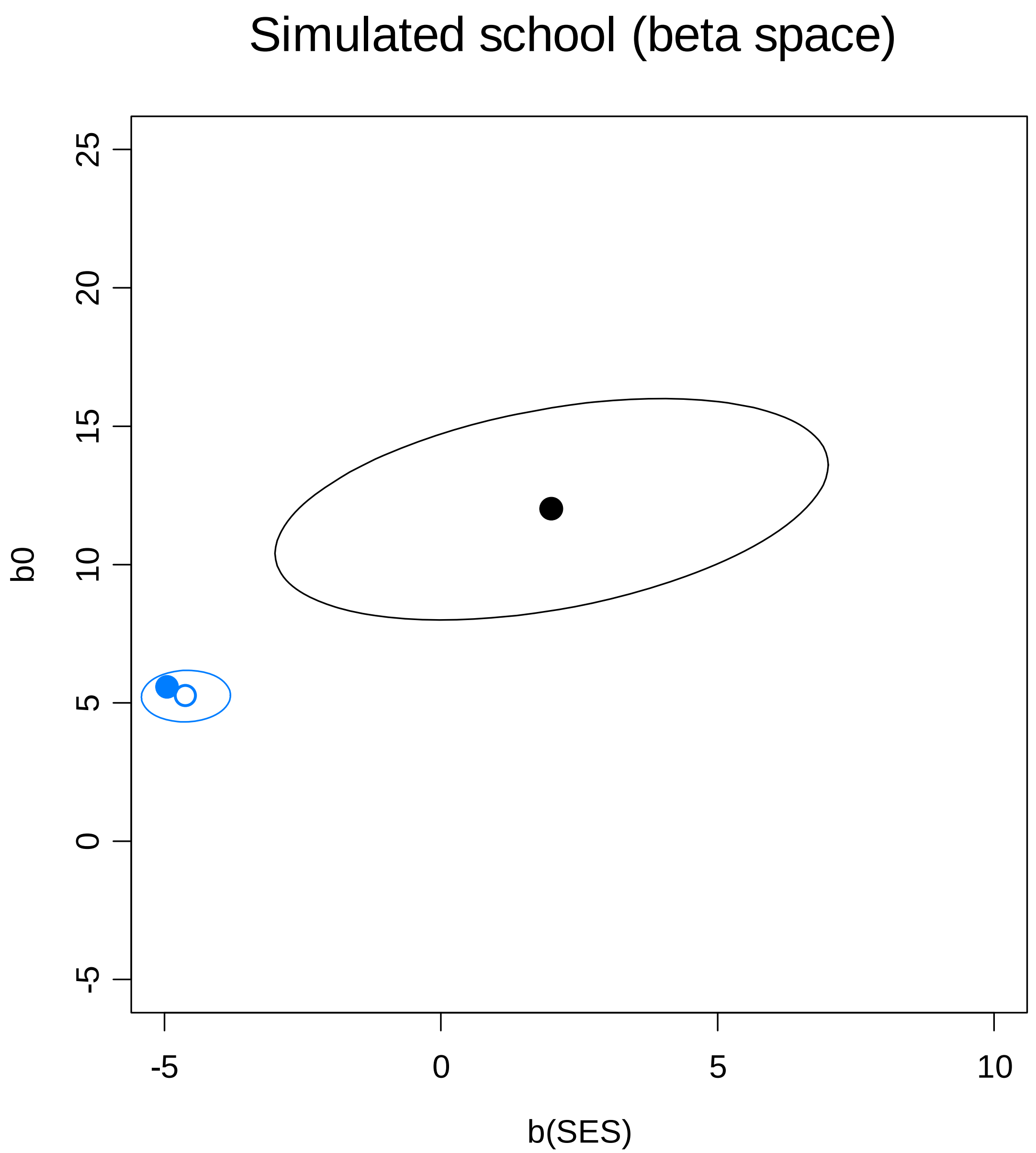


Figure 16: Observed value of $\hat{\beta}_j$.

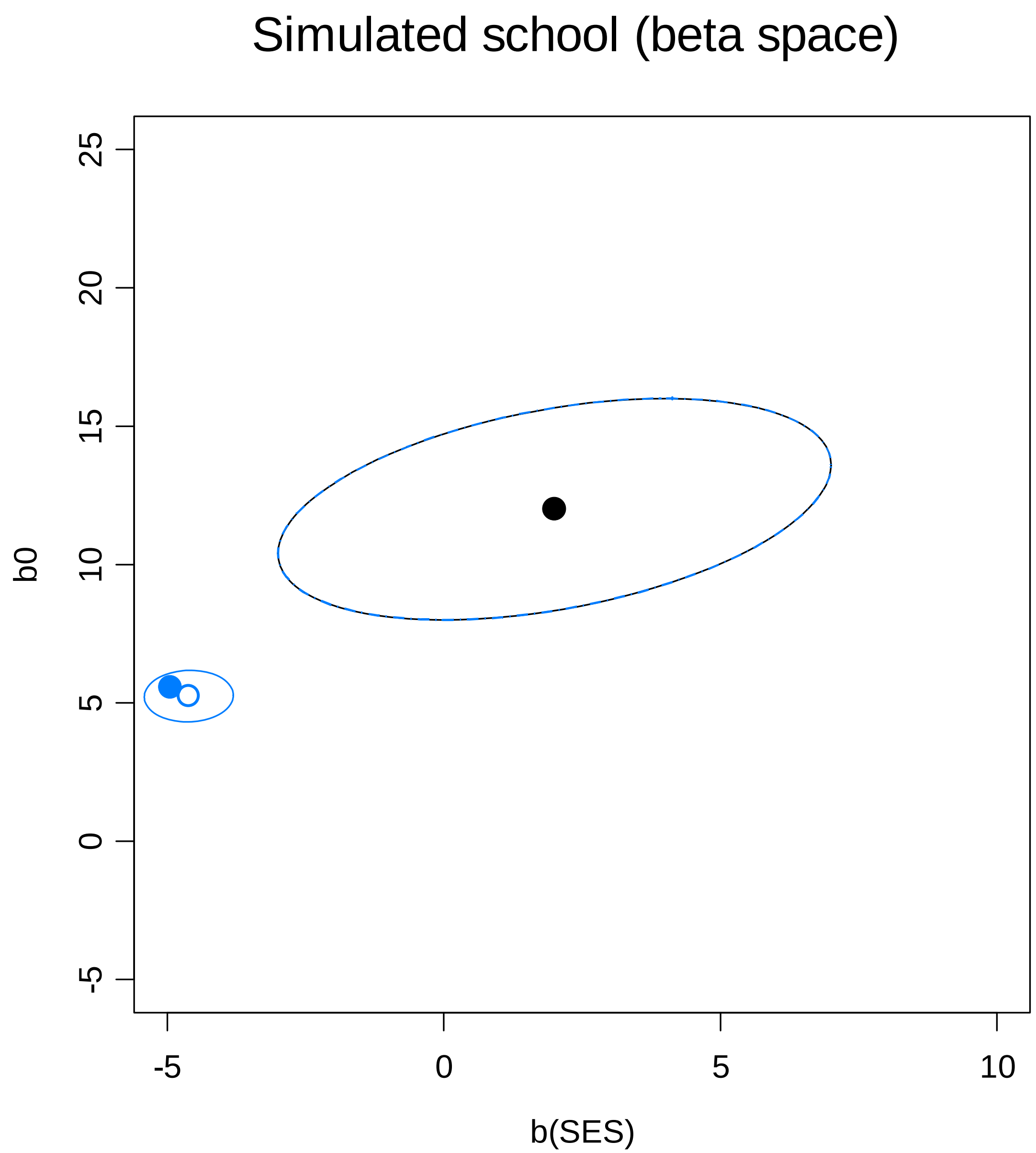


Figure 17: The blue dispersion ellipse with matrix $\mathbf{V}_j = \mathbf{G} + \sigma^2 \left(\mathbf{X}_j ' \mathbf{X}_j \right)^{-1}$ is almost coincident with the dispersion ellipse with matrix T.

Note that with smaller N, larger σ^2 or smaller dispersion for SES, these dispersion ellipse for the true $\boldsymbol{\beta}_j$ (with matrix $\mathbf{T}$) and the dispersion ellipse for $\hat{\boldsymbol{\beta}}_j$ as an estimate of $\boldsymbol{\gamma}$(with matrix $\mathbf{V}_j = \mathbf{G} + \sigma^2 \left(\mathbf{X}_j{}'\mathbf{X}_j\right)^{-1}$) could differ much more than they do here. Also note that the statistical design of the study can make $\sigma^2 \left(\mathbf{X}_j{}'\mathbf{X}_j\right)^{-1}$ smaller but, typically, not **G.**

Between-School Model: What γ means

Instead of supposing that we have a single population of schools we now add the between-school model that will allow us to suppose that there are two populations of schools: Catholic and Public and that the population mean slope and intercept may be different in the two sectors. Let *W* represent the between-school variable sector variable that is the indicator variable for Catholic schools: W_j is equal to 1 if school *j* is Catholic and 0 if it is public.[4]

We have two regression models, one for intercepts and one for the slopes:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{oj}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j}$$

We can work out the following interpretation of the γ_{ij} coefficients by setting w_j to 0 for Public schools and then to 1 for Catholic schools. The interpretation is analogous to that of the ordinary regression to compare two schools except that we are now comparing the two sectors.

---

[4] Between-school variables are not limited to indicator variables. Any variables suitable as a predictor in a linear model could be used as long as it is a function of schools, i.e. has the same value for every subject within each school.

In Public schools:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} \times 0 + u_{0j} = \gamma_{00} + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11} \times 0 + u_{1j} = \gamma_{10} + u_{1j}$$

In Catholic schools:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} \times 1 + u_{0j} = \gamma_{00} + \gamma_{01} + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11} \times 1 + u_{1j} = \gamma_{10} + \gamma_{11} + u_{1j}$$

Thus:

1. γ_{00} is the mean achievement intercept for Public schools, i.e. the mean achievement when SES is 0.
2. $\gamma_{00} + \gamma_{01}$ is the mean achievement intercept for Catholic schools so that γ_{01} is the difference in mean intercepts between Catholic and Public schools.
3. γ_{10} is the mean slope in Public schools.

4. $\gamma_{10}+\gamma_{11}$ is the mean slope in Catholic schools so that γ_{11} is the mean difference in (or difference in mean) slopes between Catholic and Public schools.
5. u_{0j} is the unique ''effect'' of school j on the achievement intercept, conditional given W.
6. u_{1j} is the unique ''effect'' of school j on the slope, conditional given W.

Now, u_{0j} and u_{1j} are Level 2 random variables (random effects) which we assume to have 0 mean and variance-covariance matrix:

$$\mathbf{G}=\begin{pmatrix} g_{00} & g_{01} \\ g_{10} & g_{11} \end{pmatrix}$$

This is a multivariate model with the complication that the dependent variables, β_{0j}, β_{1j} are not directly observable.

As mentioned above, one way to proceed would be to use a two-stage process:

1. Estimate β_{0j}, β_{1j} with least-squares within each school, and
2. use the estimated values in a Level-2 analysis with the model above.

Some problems with this approach are:

1. Each $\hat{\beta}_{0i}, \hat{\beta}_{1i}$ might have a different variance due to differing n_js and different predictor matrices X_i in each school. A Level 2 analysis that uses OLS will not take these factors in consideration.

2. Even if X_i (thus n_j) is the same for each school, we might be interested in getting information on T itself, not on

$$\mathrm{var}(\hat{\boldsymbol{\beta}}_i) = \mathbf{G} + \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

3. $\hat{\beta}_{0i}$, $\hat{\beta}_{1i}$ might be reasonable estimates of the ‘parameters’ β_{0i} and β_{1i} but, as ‘estimators’ of the random variables β_{0i} and β_{1i} they ignore the information contained in the distribution of β_{0i} and β_{1i}.
4. Some level 1 models might not be estimable, so information from these schools would be entirely lost.

Mixed or Combined or Composite model

From the multilevel model to the mixed model

Since

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j}$$

Between School Model

We combine the models by substituting the *between school model* above into the *within school model*:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij}$$

Within School Model

Substituting, we get

$$Y_{ij} = \left(\beta_{0j}\right) + \left(\beta_{1j}\right) X_{ij} + r_{ij}$$

$$= \left(\gamma_{00} + \gamma_{01} W_j + u_{0j}\right)$$

$$+ (\gamma_{00} + \tau_{11} W_j + u_{1j}) X_{ij} + r_{ij}$$

We then rearrange the term to separate fixed parameters from random coefficients:

Same as previous page

$$Y_{ij} = \left(\beta_{0j}\right) + \left(\beta_{1j}\right) X_{ij} + r_{ij}$$

$$= \left(\gamma_{00} + \gamma_{01} W_j + u_{0j}\right)$$

$$+ (\gamma_{00} + \gamma_{11} W_j + u_{1j}) X_{ij} + r_{ij}$$

Grouping fixed and random parts together

$$= \gamma_{00} + \gamma_{01} W_j + \gamma_{10} X_{ij} + \gamma_{11} W_j X_{ij}$$

$$+ u_{0j} + u_{1j} X_{ij} + r_{ij}$$

The last two lines looks like the sum of two linear models:

1) an ordinary linear model with coefficients that are ***fixed*** parameters:

$$\gamma_{00} + \gamma_{01} W_j + \gamma_{10} X_{ij} + \gamma_{11} W_j X_{ij}$$

with fixed parameters $\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11}$, and

2) a linear model with ***random*** coefficients and an error term:

$$u_{0j} + u_{1j} X_{ij} + r_{ij}$$

with random 'parameters' u_{0j} and u_{1j}.

Note the following:

1. **the fixed model contains both outer variables and inner variables as well as an interaction between inner and outer variables. This kind of interaction is called a 'cross-level' interaction. It allows the effect of X to be different in each Sector.**

2. **the random effects model only contains an intercept and an inner variable. There are *very arcane* situations in which it might make sense to include an outer variable in the random effects portion of the model which we will consider briefly later.**

Understanding the connection between the multilevel model and the combined model is useful because some packages require the model to be specified in its multilevel form (e.g. MLWin) while others require the model to be specified in its combined form as two models: the fixed effects model and the random effects model (e.g. SAS PROC MIXED, R and S-Plus lme() and nlme()).

GLS form of the model

Another way of looking at this model is to see it as a linear model with a complex form of error. Let δ_{ij} represent the combined error term – also known as the composite error term:

$$\delta_{ij} = u_{0j} + u_{1j} X_{ij} + r_{ij}$$

We can then write the model as:

$$Y_{ij} = \gamma_{00} + \gamma_{01} W_j + \gamma_{10} X_{ij} + \gamma_{11} W_j X_{ij} + \delta_{ij}$$

This looks like an ordinary linear model except that the δ_{ij}s are **not** identically $N(0,\sigma^2)$ and are **not** independent since the same u_{0j} and u_{1j} contribute to the random error for all δ_{ij}s in the *j*th school. If we let $\boldsymbol{\delta}_j$ be the vector of errors in the *j*th school we can express the distribution of the combined errors as follows:

$$\boldsymbol{\delta}_j \sim N(0, \mathbf{G} + \sigma^2 \left(\mathbf{X}_i{}'\mathbf{X}_i\right)^{-1}), \quad \boldsymbol{\delta}_j \text{ and } \boldsymbol{\delta}_k \text{ are independent for } j \neq k.$$

If $\mathbf{T}$ and σ^2 were known then the variance-covariance matrix of the random errors could be computed and the model fitted with Generalized Least-Squares (GLS).

With $\mathbf{T}$ and σ^2 unknown, we can iteratively estimate them and use the estimated values to fit the linear parameters, γ_{st} by GLS. There are variants depending on the way in which $\mathbf{T}$ and σ^2 are estimated. Using full likelihood yields what is often called ‘‘IGLS,’’ ‘‘ML,’’ or ‘‘FIML.’’ Using the conditional likelihood of residuals given $\hat{Y}$ yields ‘‘RIGLS’’ or ‘‘REML’’ (R for restricted or reduced).

Matrix form

Take all observations in school *j* and assemble them into vectors and matrices: (this is called the Laird-Ware formulation of the model from Laird and Ware (1982))

$$\mathbf{Y}_j = \mathbf{X}_j \boldsymbol{\gamma} + \mathbf{Z}_j \mathbf{u}_j + \mathbf{r}_j$$

where

$$\mathbf{Y}_j = \begin{bmatrix} Y_{1j} \\ \vdots \\ Y_{n_j j} \end{bmatrix}, \quad \mathbf{X}_j = \begin{bmatrix} 1 & W_j & X_{1j} & W_j X_{1j} \\ 1 & W_j & X_{2j} & W_j X_{2j} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & W_j & X_{n_j j} & W_j X_{n_j j} \end{bmatrix}, \quad \mathbf{Z}_j = \begin{bmatrix} 1 & X_{1j} \\ 1 & X_{2j} \\ \vdots & \vdots \\ 1 & X_{n_j j} \end{bmatrix}$$

$$\mathbf{u}_j = \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix}, \quad \boldsymbol{\gamma} = \begin{pmatrix} \gamma_{00} \\ \gamma_{01} \\ \gamma_{10} \\ \gamma_{11} \end{pmatrix}, \quad \mathbf{r}_j = \begin{pmatrix} r_{1j} \\ r_{2j} \\ \vdots \\ r_{n_j j} \end{pmatrix}, \quad j = 1, \cdots, J$$

The distribution of the random elements is: $\mathbf{u}_j \sim N(0, \mathbf{G}), \quad \mathbf{r}_j \sim N(0, \sigma^2 \mathbf{I})$ with u_j independent of r_j.

Now we put the school matrices together into big matrices:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{Z}\mathbf{u} + \mathbf{r}$$

where

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y_1} \\ \vdots \\ \mathbf{Y_J} \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_J \end{bmatrix}, \ \mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_J \end{bmatrix}, \ \mathbf{r} = \begin{bmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_J \end{bmatrix}$$

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{Z}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{Z}_J \end{bmatrix}$$

with

$$\mathbf{u} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{G} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{G} \end{bmatrix} \right)$$

and

$$\mathbf{r} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

which might be deceptive because the ‘‘**I**” is now much larger than before. The new block diagonal matrix for the variance of **u** is often with the same symbol as the variance of $\mathbf{u}_j$. To avoid confusion we can use $\ddot{\mathbf{G}}$.

Notational Babel

Mixed models were simultaneously and semi independently developed by researchers in many different disciplines, each developing its own notation. The notation we are using here is that of Bryk and Raudenbush (1992) which has been very influential in social research. Many publications use this notation. It differs from the notation used in SAS documentation whose development was more influenced by seminal statistical work in animal husbandry. It is, of course, perfectly normal to fit models in SAS but to report findings using the notation in common use in the subject matter area. A short bilingual dictionary follows. Fortunately, **Y**, **X** and **Z** are used with the same meaning.

	Bryk and Raudenbush	SAS help files	Pinheiro and Bates	My current preference
Fixed effects parameters	$\boldsymbol{\gamma}$	$\boldsymbol{\beta}$	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$
Cluster random effect	$\boldsymbol{\beta}$	$\mathbf{b}$		$\boldsymbol{\beta}$
Cluster random effect (centered)	$\mathbf{u}$	$\boldsymbol{\gamma}$	$\mathbf{b}$	$\mathbf{u}$
Variance of random effects	$\mathbf{T}$	$\mathbf{G}$	$\boldsymbol{\Psi}$	$\mathbf{G}$
Within cluster error variance	$\boldsymbol{\Sigma}$	$\mathbf{R}$	$\sigma^2\boldsymbol{\Lambda}$	$\mathbf{R}$

For example in Bryk and Raudenbush the Mixed Model is:

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\gamma} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i$$

$$\mathbf{u}_i \sim N(\mathbf{0}, \mathbf{T}) \quad \boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_i)$$

In Pinheiro and Bates:

$$\mathbf{y}_i = \mathbf{X}_i \beta + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i$$

$$\mathbf{b}_i \sim N(\mathbf{0}, \boldsymbol{\Psi}); \quad \boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2 \boldsymbol{\Lambda}_i)$$

The GLS fit

With the matrix formulation of the model, it is easy to Express the GLS estimator of γ. First denote:

$$\mathbf{V} = \mathrm{Var}(\boldsymbol{\delta}) = \mathbf{Z}\ddot{\mathbf{G}}\mathbf{Z}' + \sigma^2 \mathbf{I}$$

Then the GLS estimator is:

$$\hat{\boldsymbol{\gamma}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$$

We will see that the presence of $\mathbf{V}^{-1}$ can result in an estimate that is very different from its OLS analogue[5]

---

[5] One ironic twist concerns small estimated values of σ^2. Normally this would a cause for rejoicing; however it can result in a nearly singular $\mathbf{V}$.

The model we just derived has every important component we want:

1. a within-cluster variable X with a fixed effect
2. a between cluster variable W with a fixed effect
3. a cross-level interaction X*W with a fixed effect
4. a random intercept varying from cluster to cluster
5. a random slope varying from cluster to cluster.

Fitting this model in R:

```
> library(nlme)
> fit.mixed <- lme( Y ~ X * W, dd,
             random = ~ 1 + X | school)
```

in long form

```
> fit.mixed <- lme( Y ~ 1 + X + W + X:W, dd,
              random = ~ 1 + X | school)
```

Although this need not imply that $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ is nearly singular. Algorithms do not yet take advantage of this. .

From the simple to the complex

Traditional name	fixed part	random part
One way ANOVA with random effects	`Y ~ 1`	`~ 1 | school`
Means as outcomes	`Y ~ 1 + W`	`~ 1 | school`
One way ANCOVA	`Y ~ 1 + X`	`~ 1 | school`
Random coefficients	`Y ~ 1 + X`	`~ 1 + X | school`
Intercepts and slopes as outcomes	`Y ~ 1 + X + W + X:W`	`~ 1 + X | school`
Non random slopes	`Y ~ 1 + X + W + X:W`	`~ 1 | school`
Parallel mean slopes	`Y ~ 1 + X + W`	`~ 1 + X | school`

Contextual cluster mean variable with CWG variable and random CWG slopes	`Y ~ 1 + cvar(X,school) + dvar(X,school)`	`~ 1 + dvar(X,school) | school`
Contextual cluster mean variable with raw variable and random CWG slopes	`Y ~ 1 + cvar(X,school) + X`	`~ 1 + dvar(X,school) | school`
Intercepts and slopes as outcomes with contextual cluster mean variable with CWG variable and random CWG effect	`Y ~ 1 + ( cvar(X,school) + dvar(X,school) ) * W`	`~ 1 + dvar(X,school) | school`

The simplest models

We have now built up the notation and some theory for a fairly general form of the linear mixed model with both Level 1 and Level 2 variables and a random effects model with a random intercept and a random slope. We will now consider the interpretation of simpler models in which we keep only some components of the more general model. Even when we are interested in the larger model, it is important to understand the simple 'sub-models' because they are used for hypothesis testing in the larger model. We will also consider some extensions of the concepts we have seen so far in the context of some of these simpler models.

One-way ANOVA with random effects

This is the simplest random effects models and provides a good starting point to illustrate the special characteristics of these models.

Level 1 model:

$$Y_{ij} = \beta_{0j} + r_{ij}$$

Level 2 model:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

Combined model:

$$Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}$$

$$\mathrm{Var}(Y_{ij}) = \mathrm{Var}(u_{0j} + r_{ij}) = g_{00} + \sigma^2$$

Note the intraclass correlation coefficient:

$$\rho = g_{00} / (g_{00} + \sigma^2)$$

Also note that within each school:

$$\mathrm{E}(\bar{Y}_{.j}) = \gamma_{0j}$$

$$\mathrm{Var}(\bar{Y}_{.j} \,|\, \beta_{0j}) = \frac{\sigma^2}{n_j}$$

but across the population:

$$\mathrm{E}(\bar{Y}_{.j}) = \gamma_{0j}$$

$$\mathrm{Var}(\bar{Y}_{.j} \,|\, \beta_{0j}) = g_{00} + \frac{\sigma^2}{n_j}$$

This is an example of two very useful facts:

1. the unconditional (sometimes called 'marginal' but not by economists) mean is equal to the **mean conditional mean**,

2. the unconditional variance is equal to the **mean of the conditional variance** plus the **variance of the conditional mean**, i.e.:

$$\begin{aligned}\mathrm{Var}(\bar{Y}_{.j}) &= \mathrm{E}(\mathrm{Var}(\bar{Y}_{.j} \mid \beta_{0j}) + \mathrm{Var}(\mathrm{E}(\bar{Y}_{.j} \mid \beta_{0j})) \\ &= \sigma^2 + \mathrm{Var}(\beta_{0j}) \\ &= \sigma^2 + g_{00}\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(Y_{.j}) &= \mathrm{E}(\mathrm{Var}(Y_{.j} \mid \beta_{0j}) + \mathrm{Var}(\mathrm{E}(Y_{.j} \mid \beta_{0j})) \\ &= \sigma^2 + \mathrm{Var}(\beta_{0j}) \\ &= \sigma^2 + g_{00}\end{aligned}$$

Estimating the one-way ANOVA model

There are three kinds of parameters that need to be estimated:

1. **fixed effect parameters** : in this case there is only one: γ_{00},
2. **variance-covariance components**: g_{00} and σ^2,
3. **random effects**: β_{0j} or, equivalently, combined with τ_{00}: u_{0j}.

We use a different approach for each type of parameter.

The **fixed effects parameters** are like linear regression parameters except that they are estimated from observations that are not independent. Instead of using OLS (ordinary least-squares) we use **GLS (generalized least-squares)** using the estimates of the variance-covariance components as the variance matrix in the GLS procedure.

The **variance-covariance parameters** are estimated using **ML (maximum likelihood) or REML (restricted maximum likelihood)**.

Note that each step above assumes that the other one has been completed. What really happens is that estimation goes back and forth between the two steps until convergence.

The **random effects** are not just parameters. They are realizations of random variables. This means that we have two sources of information about them: we can 'estimate' them from the observed data and we can 'guess' them from their distribution. Putting these two sources of information together is the essence of Bayesian estimation, or empirical Bayesian estimation because the distribution of the random effects, determined by $\mathbf{G} = [g_{00}]$, is estimated from the data and model. The random effects are **predicted** (in contrast with 'estimated') using **EBLUPs (Empirical Best Linear Unbiased Predictors)** with the empirical **posterior expectation**:

$$E(\beta_{01}, \cdots, \beta_{0J} | Y_1, \cdots, Y_n)$$

i.e. the expected value of what is unknown given what is known.

We will look at the estimation of the three types of parameters in detail in this example.

First we consider the analysis of the data using OLS in which we treat $\beta_{01}, \cdots, \beta_{0J}$ as non-random parameters. **The coding of the school effect**

determines what is estimated by the intercept term. It is a weighted linear combination of the β_{0j}s:

$$\psi_W = \sum_{j=1}^{J} w_j \beta_{0j}$$

If the coding uses ‘‘true’’ contrasts (each column of the **coding matrix** sums to 0) the weights are all equal to *1/J* and ψ_w is the ordinary mean of β_{0j} s:

$$\psi_w = \frac{1}{J}\sum_{1}^{J}\beta_{0j}$$

In this case

$$\hat{\psi}_w = \frac{1}{J}\sum_{1}^{J}\bar{Y}_j = \bar{Y}_{Schools}$$

With ‘‘sample size’’ coding, e.g.

$$
\begin{array}{cccccc}
 & V_1 & V_2 & V_3 & \cdots & V_{J-1} \\
School_1 & n_J & 0 & 0 & \cdots & 0 \\
School_2 & 0 & n_J & 0 & \cdots & 0 \\
School_3 & 0 & 0 & n_J & \cdots & 0 \\
School_4 & 0 & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
School_{J-1} & 0 & 0 & 0 & \vdots & n_J \\
School_J & -n_1 & -n_2 & -n_3 & \cdots & -n_{J-1}
\end{array}
$$

each column of the design matrix sums to 0 and the intercept will estimate:

$$\psi_w = \frac{\sum_{j=1}^{J} n_j \beta_{0j}}{\sum_{j=1}^{J} n_j}$$

which weights each school according to its sample size. This can be thought of as the mean of the population of **students** instead of the population of **schools**. The estimator would be the overall average of Y:

$$\psi_w = \frac{\sum_{j=1}^{J} n_j \bar{Y}_j}{\sum_{j=1}^{J} n_j} = \bar{Y}_{..} = \bar{Y}_{Students}$$

We are not limited to these two obvious choices. A more appropriate set of weights could be school size, with coding:

$$\begin{array}{cccccc}
 & V_1 & V_2 & V_3 & \cdots & V_{J-1} \\
School_1 & s_J & 0 & 0 & \cdots & 0 \\
School_2 & 0 & s_J & 0 & \cdots & 0 \\
School_3 & 0 & 0 & s_J & \cdots & 0 \\
School_4 & 0 & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
School_{J-1} & 0 & 0 & 0 & \vdots & s_J \\
School_J & -s_1 & -s_2 & -s_3 & \cdots & -s_{J-1}
\end{array}$$

the intercept would estimate:

$$\psi_s = \frac{\sum_{j=1}^{J} s_j \beta_{0j}}{\sum_{j=1}^{J} s_j}$$

In each case the form of the estimate is a weighted mean of the individual school averages:

$$\hat{\psi}_w = \sum_{j=1}^{J} w_j \bar{Y}_j$$

with variance:

$$\mathrm{Var}(\hat{\psi}_w | \beta_{01}, \cdots, \beta_{0J}) = \sum_{j=1}^{J} w_j^2 \frac{\sigma^2}{n_j}$$

where the weights, w_j, sum to 1. Note that the variance is minimized when the weights are proportional to n_j, i.e. $w_j = n_j / n$ where n is the total sample size: $n = \sum_j n_j$. In this case the variance is σ^2 / n. Thus, the **student mean** is the parameter estimated with the least variance.

Mixed model approach

With a mixed model we want to estimate γ_{00} instead of a particular linear combination of β_{0j}s. Any weighted mean $\hat{\psi}_w = \sum_j w_j \bar{Y}_j$ of $\bar{Y}_j$s will be unbiased for γ_{00} because

$$\begin{aligned} \mathrm{E}(\hat{\psi}_w) &= \mathrm{E}(\sum_j w_j \bar{Y}_j) \\ &= \sum_j w_j \mathrm{E}(\beta_{0j}) \\ &= \sum_j w_j \gamma_{00} \\ &= \gamma_{00} \end{aligned}$$

if the w_j s are weights with $\sum_j w_j = 1$.

Now, to calculate the variance of $\hat{\psi}_w$ as an estimator of γ_{00}, we first need the variance of $\bar{Y}_j$ as an estimator of γ_{00} with β_{0j} random:

$$\mathrm{Var}(\bar{Y}_j) = g_{00} + \sigma^2 / n_j$$

Thus:

$$\mathrm{Var}(\hat{\psi}_w) = \sum_j w_j^2 (g_{00} + \sigma^2 / n_j)$$

The optimal estimator is obtained by taking weights **inversely proportional** to $(g_{00} + \sigma^2 / n_j)$.

Consider the implications:

1. If g_{00} is much larger than σ^2, the weights will be nearly constant and $\hat{\psi}_w$ will be close to $\bar{Y}_{Schools}$.
2. Conversely, if g_{00} is much smaller than σ^2, the weights will be nearly proportional to n_j and the estimator will be close to $\bar{Y}_{Students.}$.

If it is not reasonable to treat the β_{0j}s as a random sample from the same $N(0, g_{00})$ distribution then these two estimators could estimate two quantities with very different meanings. Consider, for example, what would happen if there is a strong relationship between β_{0j} and n_js. What gets estimated is governed by the ratio $g_{00} / \sigma^2 - a$ purely statistical consideration quite disconnected from any interpretation of the estimator. It is important to appreciate that your estimator is determined by considerations that might not be relevant.

In R the command is:

```
lme ( y ~ 1 , hs, random = ~ 1 | school )
```

In SAS, the (minimal) commands would be[6]:

```
PROC MIXED DATA = MIXED.HS;
    CLASS SCHOOL;
    MODEL Y = ;
    RANDOM INTERCEPT / SUBJECT=SCHOOL;
    RUN;
```

[6] To use the HS data set, download the self-extracting file following the link at the course website. Save it in a convenient directory. Click on its icon to create the SAS data set HS.SD2. From SAS, create a library named MIXED that points to this directory. You can then use the data set using the syntax in this example.

EBLUPs

This interesting topic can, alas, be skipped. It played a central role in the early development of mixed models for animal husbandry where an important practical problem was estimating the reproductive qualities of a bull from the characteristics of its progeny. In most applications of mixed models in the social sciences, the focus is on the estimation of the fixed parameters and much less so on the 'prediction' of the random effects.

Estimating the u_{0j}s involves using two sources of information: the data and their distribution as random variables. First consider the OLS estimator for β_{0j}:

$$\hat{\beta}_{0j} = \bar{Y}_{.j}$$

Now, to get the *Empirical Best Linear Unbiased Predictor* of u_{0j}s, we pretend that the estimated values of γ_{00} and σ^2 are the ''true'' values and we calculate the conditional expectation of u_{0j}s given y_{ij}s. This is done most easily using the matrix formulation of the model and a formula for the conditional expectation in the multivariate case. We use partitioned matrices to express the joint distribution of **Y** and **u**:

$$\begin{bmatrix} \mathbf{Y} \\ \mathbf{u} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{X\gamma} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{Z\ddot{G}Z'} + \sigma^2 \mathbf{I} & \mathbf{Z\ddot{G}} \\ \mathbf{\ddot{G}Z'} & \mathbf{\ddot{G}} \end{bmatrix} \right)$$

A ‘‘well-known’’ formula gives:

$$\hat{\mathrm{E}}(\mathbf{u} | \mathbf{Y}) = \mathbf{\ddot{G}ZV}^{-1}(\mathbf{Y} - \mathbf{X\gamma})$$

where $\mathbf{V} = \mathbf{Z\ddot{G}Z'} + \sigma^2 \mathbf{I}$. This formula with a bit more mechanical work will give us the EBLUP below, but we will derive it intuitively:

1. We could estimate u_{0j} with the ''obvious'' OLS estimate:

$$\hat{u}_{0j} = \hat{\beta}_{0j} - \hat{\gamma}_{00} = \bar{Y}_{.j} - \hat{\gamma}_{00}$$

as an estimate of u_{0j} this has variance $\sigma^2 / n_{j.}$

2. We could also guess that u_{0j} is equal to 0 (the mean of its distribution) and our guess would have variance g_{00}.

How can we ''best'' combine these independent sources of information? By using weights proportional to inverse variance! This gives us the **EBLUP** of u_{0j}:

$$\tilde{u}_{0j} = \frac{\frac{1}{\sigma^2 / n_j}\hat{u}_{0j} + \frac{1}{g_{00}}0}{\frac{1}{\sigma^2 / n_j} + \frac{1}{g_{00}}} = \frac{\hat{u}_{0j}}{1 + \frac{\sigma^2 / n_j}{g_{00}}}$$

This has the effect of **shrinking** $\hat{u}_{0j}$ towards 0 by a factor of

$$\frac{\frac{1}{\sigma^2/n_j}}{\frac{1}{\sigma^2/n_j}+\frac{1}{\tau_{00}}}=\frac{1}{1+\frac{\sigma^2/n_j}{\tau_{00}}}$$

Consider how the amount of shrinking depends on the relative values of σ^2, g_{00} and n_j. There will be more shrinkage if

1. g_{00} is small: i.e. the distribution of u_{0j} is known to be close to 0.
2. σ^2 is large: i.e. $\bar{Y}_{0j}$ has large variation as an estimate of β_{0j}.
3. n_j is small: ditto.

The EBLUP estimator of β_{0j} (we'll call it $\tilde{\beta}_{0j}$ works exactly the same way with the OLS estimator (analyzing each school separately) which gets shrunk towards the overall estimator $\hat{\gamma}_{00}$. This is in exactly the same spirit as shrinkage estimators derived from Bayesian, Empirical Bayes or frequentist approaches. Bradley Efron and Carl Morris wrote an interesting article on the topic in *Scientific American*, Efron and Morris(1977).

Slightly more complex models

Means as outcomes regression

Level 1 model:

$$Y_{ij} = \beta_0 + \beta_{0j} + r_{ij}$$

Level 2 model:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j}$$

Combined model:

$$\begin{aligned} Y_{ij} = \gamma_{00} &+ \gamma_{01} W_j \\ &+ u_{0j} + r_{ij} \end{aligned}$$

Note that

$$\mathrm{Var}(Y_{ij}) = \mathrm{Var}(u_{0j} + r_{ij})$$

as above but, in this model, $\mathrm{Var}(Y_{ij})$ is a conditional variance, conditional given W.

In R the command is:

```
lm ( y ~ w , hs, random = ~ 1 | school )
```

In SAS, the commands for the means as outcomes model would be:

```
PROC MIXED DATA = MIXED.HS;
   CLASS SCHOOL;
   MODEL Y = W ;
   RANDOM INTERCEPT / SUBJECT = SCHOOL;
   RUN
```

One-way ANCOVA with random effects

Level 1 model:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij}$$

Level 2 model:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$
$$\beta_{1j} = \gamma_{10}$$

Combined model:

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + u_{0j} + r_{ij}$$

In R the command is:

```
lm ( y ~ x, hs, random = ~ 1 | school )
```

In SAS, the commands for one-way ANCOVA with random effects are:

```
PROC MIXED DATA = MIXED.HS;
    CLASS SCHOOL;
    MODEL Y = X ;
    RANDOM INTERCEPT / SUBJECT = SCHOOL;
    RUN;
```

Random coefficients model

Level 1 model:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij}$$

Level 2 model:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

with:

$$\mathrm{Var}\left(\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix}\right) = \mathbf{T} = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix}$$

Combined model:

$$Y_{ij} = \tau_{00} + \tau_{10} X_{ij}$$

$$+ u_{0j} + u_{1j} X_{ij} + r_{ij}$$

In R the command would be:

```
lm ( y ~ x , hs, random = ~ 1 + x | school )
```

In SAS, the commands for the random coefficients model are:

```
PROC MIXED DATA = MIXED.HS;
   CLASS SCHOOL;
   MODEL Y = X ;
   RANDOM INTERCEPT X / SUBJECT = SCHOOL TYPE =
UN;
   RUN;
```

Intercepts and Slopes as outcomes

This corresponds to the full model presented in 0 above.

In R the command would be:

```
lme ( y ~ x * w , hs, random = ~ 1 + x | school )
```

The SAS commands for this model are:

```
PROC MIXED DATA = MIXED.HS;
   CLASS SCHOOL;
   MODEL Y = X W X*W;
   RANDOM INTERCEPT X / SUBJECT = SCHOOL TYPE = UN;
   RUN;
```

Note the *X**W* term. It is called a ***cross-level interaction***. It has the function of allowing the mean slope with respect to *X* to vary with *W*. Note that R automatically generates the marginal terms, x and w.

Nonrandom slopes

Consider the full model but with $\tau_{11} = 0$ (hence $\tau_{01} = 0$ also, otherwise $\mathbf{T}$ would not be a variance matrix). This is a model in which the variation in

$\hat{\beta}_{1j}$ from school to school is wholly consistent with the expected variation within schools and there is no need to postulate that $\tau_{11} > 0$.

In R the command would be:

```
lm ( y ~ x * w , hs, random = ~ 1 | school )
```

The SAS commands are left as an exercise.

Contextual effects

A major – and underexploited – advantage of multilevel models is that it is easy to separately estimate the between-cluster and the within-cluster effects of a variable. The advantages of this approach are:

1. Including both effects in the model allows each to be estimated without contamination from the other. Many classical applications of mixed models are based on the assumption that the between effect and the within effect are equal. If the assumption is not satisfied the estimate is biased.

2. Effects at both levels can be estimated simultaneously with SEs that allow inference to appropriate populations. In contrast, the fixed effects model only allows generalization to new samples from the same clusters. The between-cluster model did not provide an estimate of the within-cluster effect.

3. Both between-cluster and within-cluster variables as well as cross-level interactions can be included in the same model.

Fixed part of the model with contextual cluster mean variable:

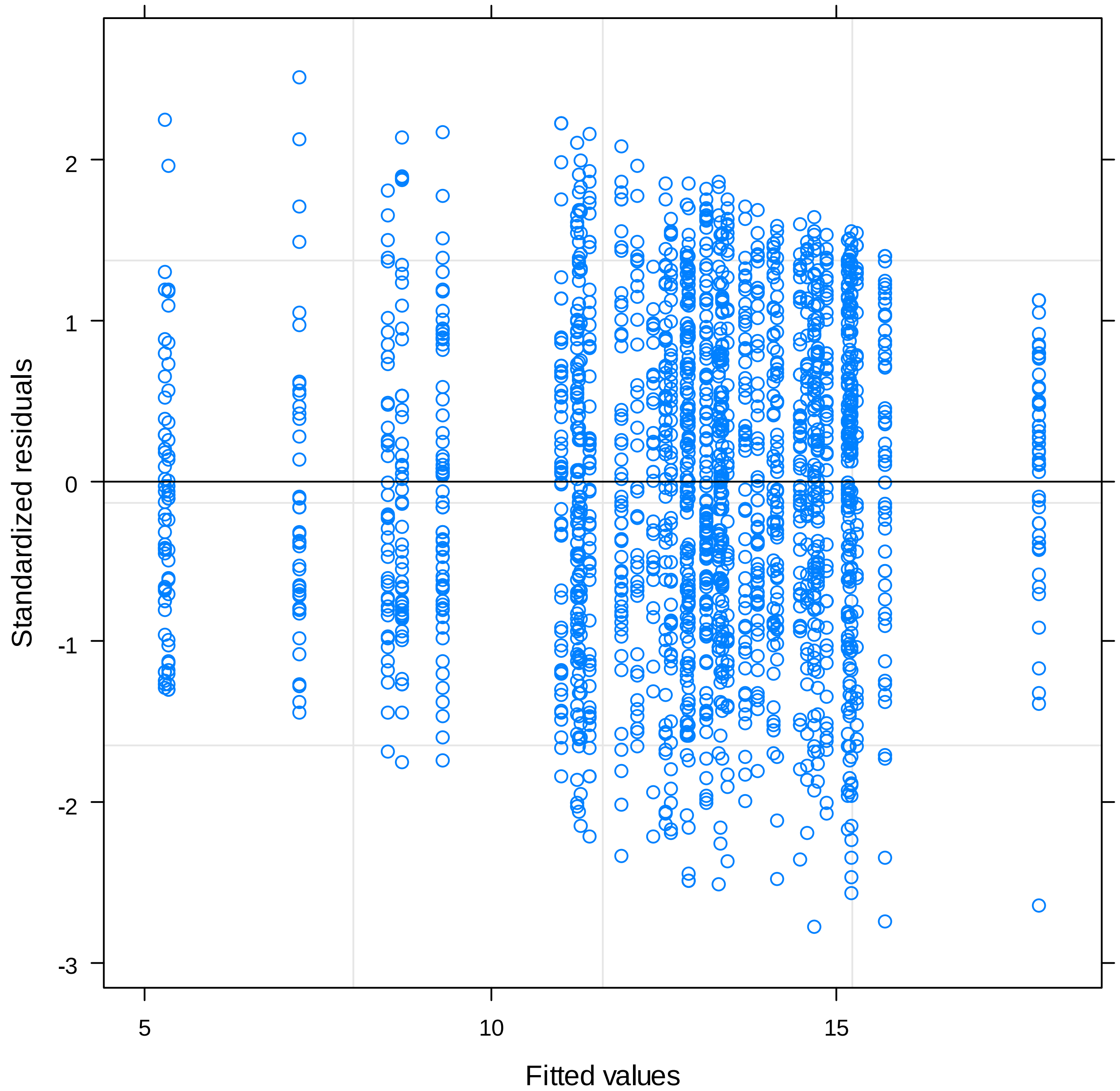
Fitting the models

One way anova with random effect

```
> fit.oneway.re <- lme( mathach ~ 1, hs, random = ~ 1 | sid)
> summary(fit.oneway.re)

Linear mixed-effects model fit by REML
 Data: hs
       AIC      BIC    logLik
  12985.94 13002.71 -6489.969

Random effects:
 Formula: ~1 | sid
        (Intercept) Residual
StdDev:    2.836278 6.296759

Fixed effects: mathach ~ 1
               Value Std.Error   DF  t-value p-value
(Intercept) 12.60468 0.4711941 1937 26.75049       0

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-2.78262694 -0.74562760  0.03825124  0.78826675  2.51105403

Number of Observations: 1977
Number of Groups: 40
```

```
>
> intervals( fit.oneway.re )
Approximate 95% confidence intervals

 Fixed effects:
               lower     est.    upper
(Intercept) 11.68057 12.60468 13.52878
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: sid
                   lower     est.    upper
sd((Intercept)) 2.214072 2.836278 3.633338

 Within-group standard error:
   lower     est.    upper
6.101522 6.296759 6.498242

> glh( fit.oneway.re )
 numDF denDF F.value p.value
     1  1937 715.589 <.00001

Coefficients  Estimate Std.Error   DF  t-value p-value Lower 0.95 Upper 0.95
  (Intercept) 12.60468   0.47119 1937 26.75049 <.00001   11.68057   13.52878
```

Note: this could use a better approximation for degrees of freedom, e.g. the Satterthwaite algorithm that SAS uses.



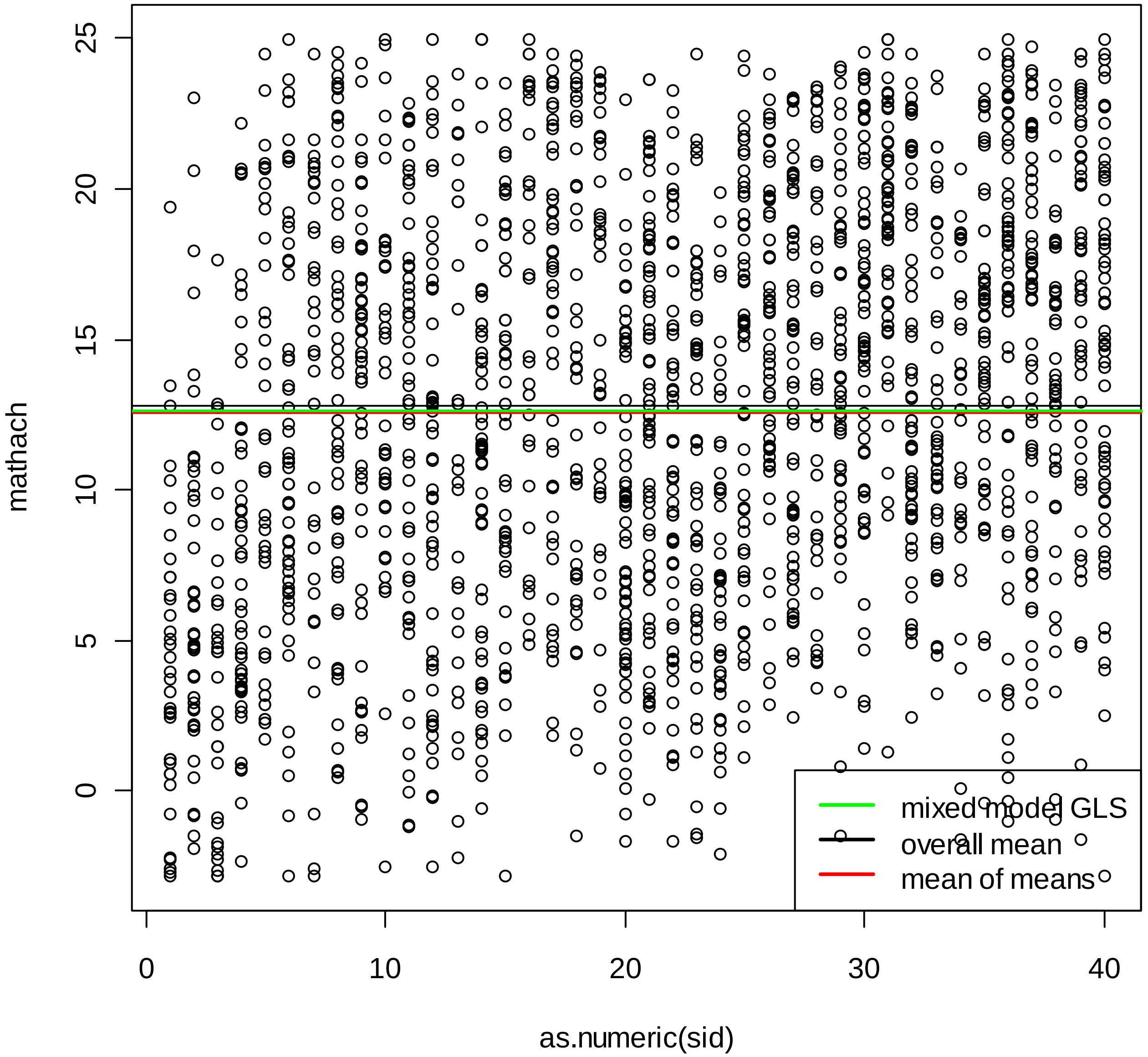

```
plot(fit.oneway.re )
```

Note pattern in fitted residuals in contrast with OLS

```
> fixef( fit.oneway.re )   # estimation of fixed part of model
(Intercept)
   12.60468

> ranef( fit.oneway.re )   # BLUP of error in random portion
       (Intercept)
P5762C -7.30651445
P2639C -5.36017663
P8854C -7.24846197
P6484C  0.26973942
. . .
C2208C  2.58744359
C2658C  0.71334861
C1906C  3.09104215
C9586G  2.08485465

> coef( fit.oneway.re )    # BLUP combining fixed and random parts
       (Intercept)
P5762C    5.298161
P2639C    7.244499
P8854C    5.356214
. . .
C5619C   15.220870
C2208C   15.192119
```

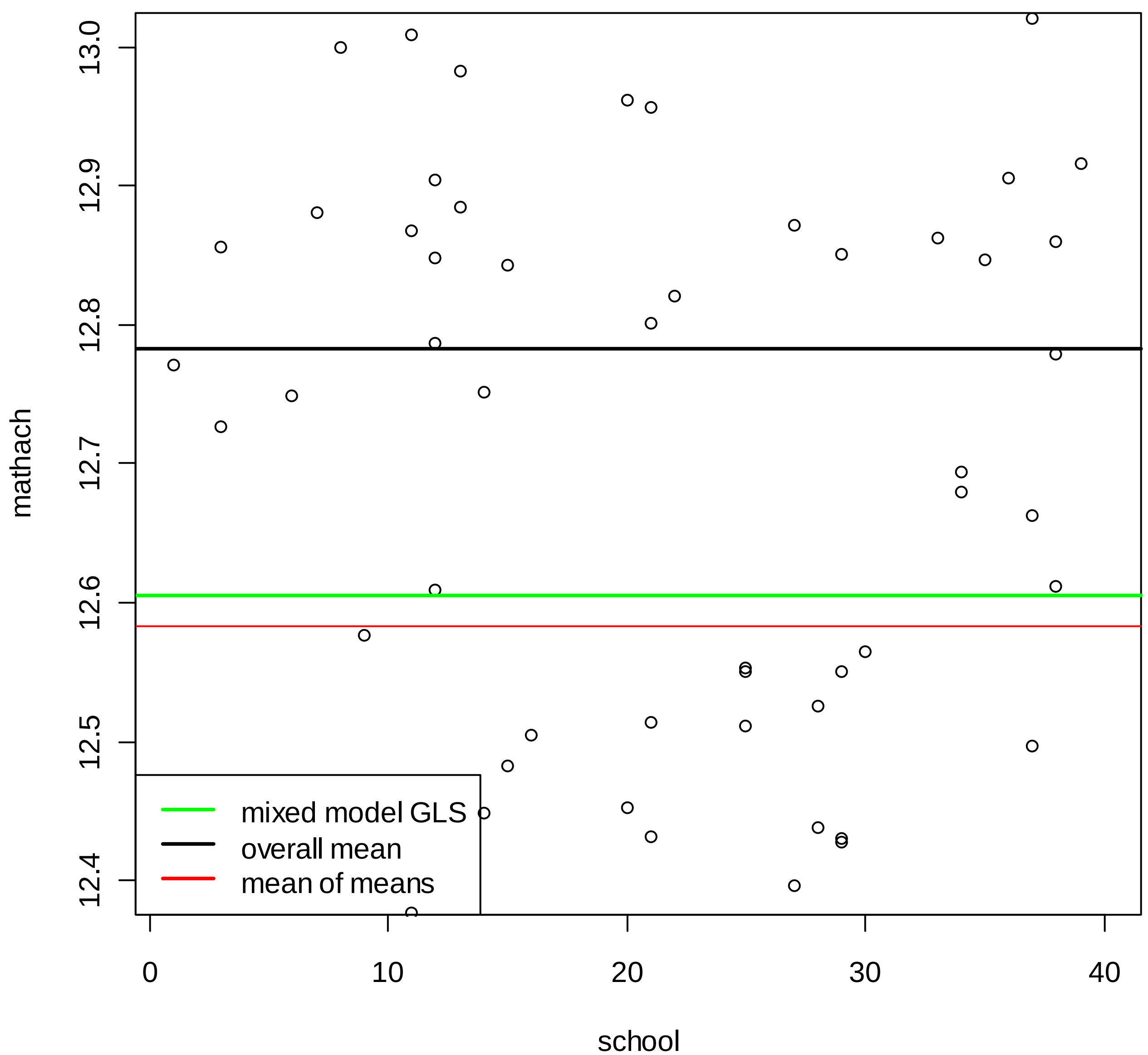
```
C2658C   13.318024
C1906C   15.695718
C9586G   14.689530


> coef( fit.oneway.re) == ( ranef(fit.oneway.re) + fixef( fit.oneway.re ))
       (Intercept)
P5762C        TRUE
P2639C        TRUE
P8854C        TRUE
. . .
C3992C        TRUE
C5619C        TRUE
C2208C        TRUE
C2658C        TRUE
C1906C        TRUE
C9586G        TRUE
```

```
> plot( mathach ~ as.numeric(sid) , hs)
> abline( h = fixef( fit.oneway.re), col = 'green',
lwd = 2)
> abline( h = mean( hs$mathach), col = 'black')
> abline( h = mean( c(tapply( hs$mathach, hs$sid,
mean))), col = 'red')
>
> ?legend
> legend( 'bottomright', c('mixed model GLS',
'overall mean', 'mean of means'),
+                        col =
c('green','black','red'), lty = 1, lwd = 2)
```



mathach
as.numeric(sid)
mixed model GLS
overall mean
mean of means
0
5
10
15
20
25
0
10
20
30
40

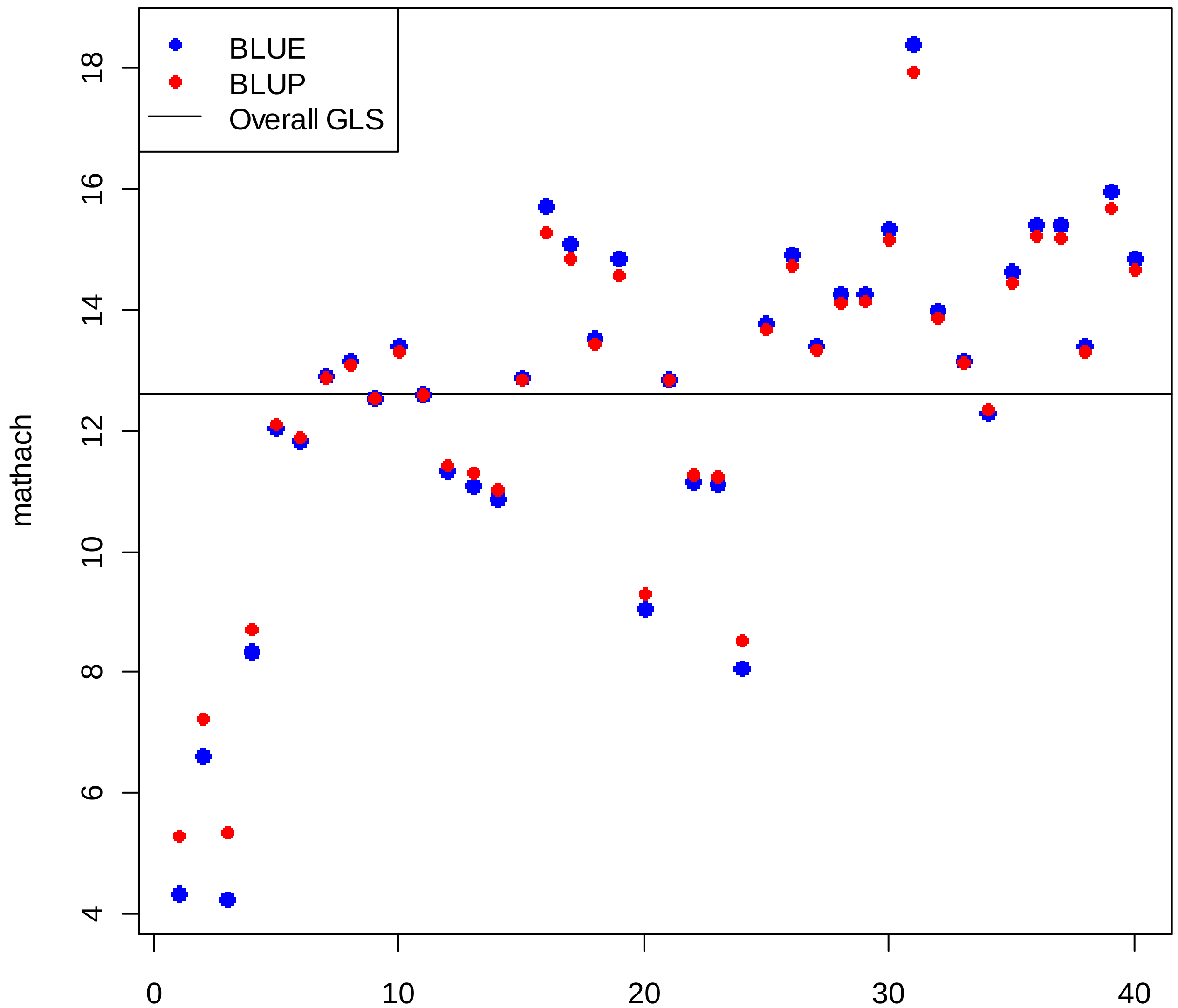
```
plot( mathach ~ as.numeric(sid) , hs, ylim = c(12.4,13), xlab = 'school')
abline( h = fixef( fit.oneway.re), col = 'green', lwd = 2)
abline( h = mean( hs$mathach), col = 'black', lwd = 2)
abline( h = mean( c(tapply( hs$mathach, hs$sid, mean))), col = 'red')
legend( 'bottomleft', c('mixed model GLS', 'overall mean', 'mean of means'),
                      col = c('green','black','red'), lty = 1, lwd = 2)
```



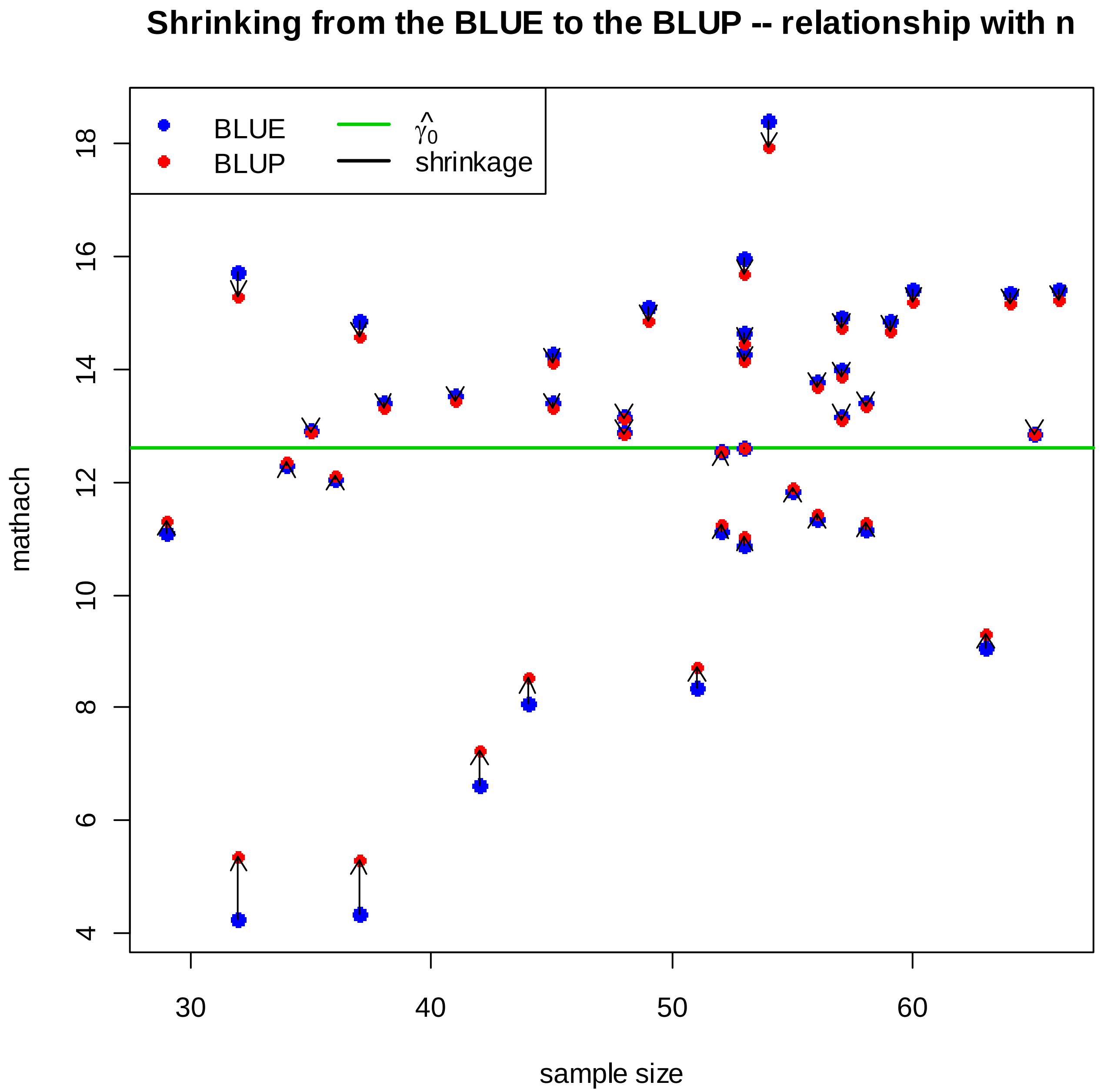
mathach
school
13.0
12.9
12.8
12.7
12.6
12.5
12.4
0
10
20
30
40
mixed model GLS
overall mean
mean of means

```
> hs$n <- capply( hs$sid, hs$sid, length)  # sample size in each school
> hs$mathach.ols <- capply( hs$mathach, hs$sid, mean)
>
> hs1.sid <- up(hs, ~ sid)
> rownames(hs1.sid) == rownames( coef( fit.oneway.re))  # check that they
match
 [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
[18] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
[35] TRUE TRUE TRUE TRUE TRUE TRUE
>
> hs1.sid$blup <- coef( fit.oneway.re ) [,1]
> some( hs1.sid)
       school sector Size   Sector PRACAD DISCLIM HIMINTY  Sex.comp Sex.cat    sid what
P8854C   8854      0  745   Public   0.18  -0.228       0 0.5312500    Coed P8854C    p
P2771C   2771      0  415   Public   0.24   1.048       0 0.5090909    Coed P2771C    p
P5640C   5640      0 1152   Public   0.41   0.256       0 0.4210526    Coed P5640C    p
P7345C   7345      0  978   Public   0.64   0.336       1 0.5178571    Coed P7345C    p
P6897C   6897      0 1415   Public   0.55  -0.361       0 0.5918367    Coed P6897C    p
C4530G   4530      1  435 Catholic   0.60  -0.245       1 1.0000000   Girls C4530G    p
C7342B   7342      1 1220 Catholic   0.46   0.380       1 0.0000000    Boys C7342B    p
C5720C   5720      1  381 Catholic   0.65  -0.352       0 0.4528302    Coed C5720C    p
C7688B   7688      1 1410 Catholic   0.65  -0.575       0 0.0000000    Boys C7688B    p
C1906C   1906      1  400 Catholic   0.87  -0.939       0 0.5094340    Coed C1906C    p
           ses.sch mathach.sch  n mathach.ols      blup
P8854C -0.75675000    4.239781 32    4.239781  5.356214
P2771C -0.33945455   11.844109 55   11.844109 11.906661
P5640C -0.17659649   13.160105 57   13.160105 13.115900
P7345C  0.03325000   11.338554 56   11.338554 11.440975
```

```
P6897C  0.34955102   15.097633 49   15.097633 14.869792
C4530G -0.59688889    9.055698 63    9.055698  9.313204
C7342B -0.44782759   11.166414 58   11.166414 11.279062
C5720C  0.03256604   14.282302 53   14.282302 14.139565
C7688B  0.18588889   18.422315 54   18.422315 17.935733
C1906C  0.51162264   15.983170 53   15.983170 15.695718
>
>

>
> plot( c(1,40), range( hs1.sid$mathach.ols), xlab = '', ylab = 'mathach',
type = 'n')
> abline( h = fixef( fit.oneway.re ), col = 'black', lwd = 1.5)
> points( 1:40, hs1.sid$mathach.ols, col = 'blue', pch = 16, cex = 1.2)
> points( 1:40, hs1.sid$blup, col = 'red', pch = 16)
> legend( 'topleft', c('BLUE', 'BLUP', 'Overall GLS'),
+         col = c('blue','red','black'),
+         pch = c(16,16,NA),
+         lty = c(NA,NA, 1))
>
```


BLUE
BLUP
Overall GLS
mathach
4
6
8
10
12
14
16
18
0
10
20
30
40

```
> ##  by sample size + a few more plotting bells and whistles
>
> plot( range( hs1.sid$n), range( hs1.sid$mathach.ols),
+       xlab = 'sample size', ylab = 'mathach', type = 'n',
+       main = 'Shrinking from the BLUE to the BLUP -- relationship with n')
> abline( h = fixef( fit.oneway.re ), col = 'green3', lwd = 2)
> points( hs1.sid$n, hs1.sid$mathach.ols, col = 'blue', pch = 16, cex = 1.2)
> points( hs1.sid$n, hs1.sid$blup, col = 'red', pch = 16)
> arrows( hs1.sid$n, hs1.sid$mathach.ols, hs1.sid$n, hs1.sid$blup, length=
.1)
Warning message:
In arrows(hs1.sid$n, hs1.sid$mathach.ols, hs1.sid$n, hs1.sid$blup,  :
  zero-length arrow is of indeterminate angle and so skipped
> legend( 'topleft',
+         # c('BLUE', 'BLUP', 'Overall GLS', 'shrinkage'),
+         expression(BLUE, BLUP, hat(gamma[0]), shrinkage),
+         ncol = 2,
+         col = c('blue','red','green3', 'black'),
+         lwd = c(NA, NA, 2,2),
+         pch = c(16,16,NA, NA),
+         lty = c(NA,NA, 1, 1))
>
```



Shrinking from the BLUE to the BLUP -- relationship with n
BLUE
BLUP
$\hat{\gamma}_0$
shrinkage
mathach
sample size
4
6
8
10
12
14
16
18
30
40
50
60

Note how shrinkage is roughly proportional to the distance of the BLUE from the overall GLS estimate (green line) and smaller as *n* gets larger. Note also that the spread of the BLUE is greater with with smaller n, illustrating the notion that the BLUE is not as good an estimate in this case.

The GLS estimate is an 'optimal' estimate that takes all these issues into account. What is being estimated is the overall mean of the population from which schools are drawn. This ***mean*** (as a ***parameter*** of the population of schools) is defined to give the same weight to all schools, regardless of sample size.

The GLS mixed model ***estimator*** gives less weight to schools with smaller *n* but only because their data gives an estimate with larger variance.

The ***BLUP*** is a reasonable estimator for a particular school as long at the information from other schools deserves the weight it gets in shrinking the ***BLUE.*** If a school is not ***'exchangeable'*** in the sample with other schools, i.e. if some known characteristic distinguishes it so that it can't be thought of as 'just another school in the sample' then the ***BLUE*** should be preferred to the ***BLUP.***

Intercepts and slopes as outcomes

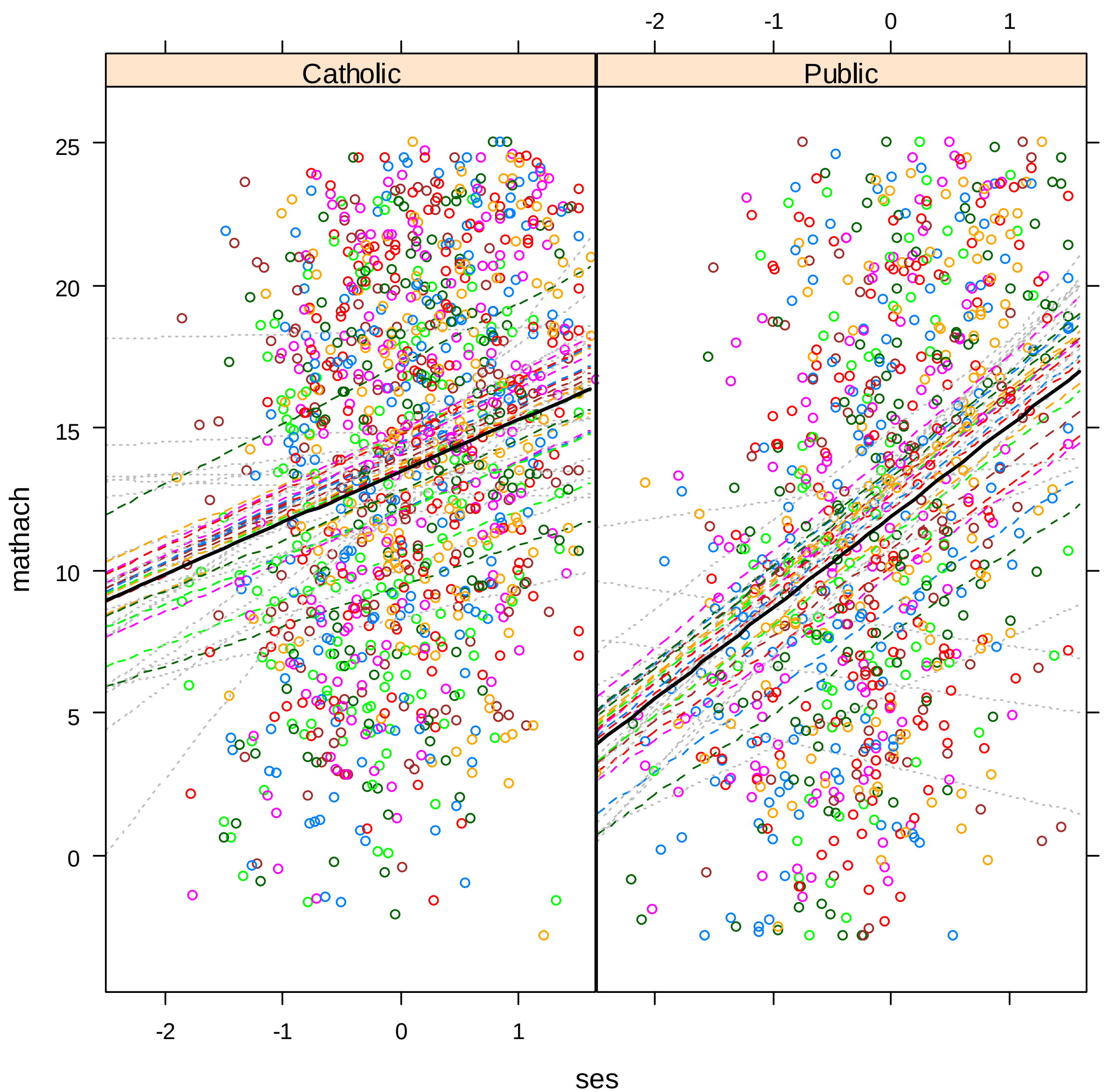


Figure 18: BLUPS from a model with random slopes

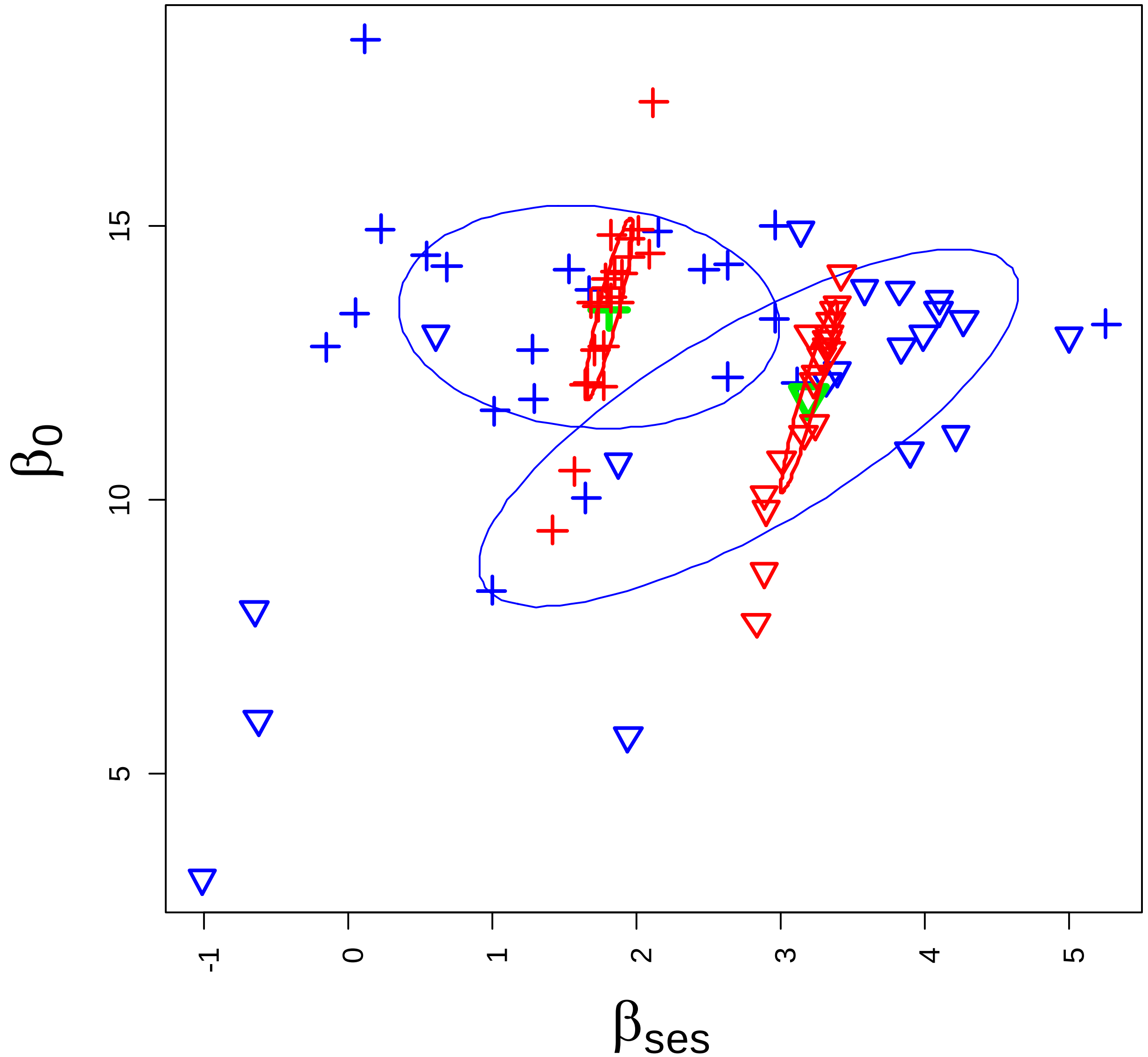
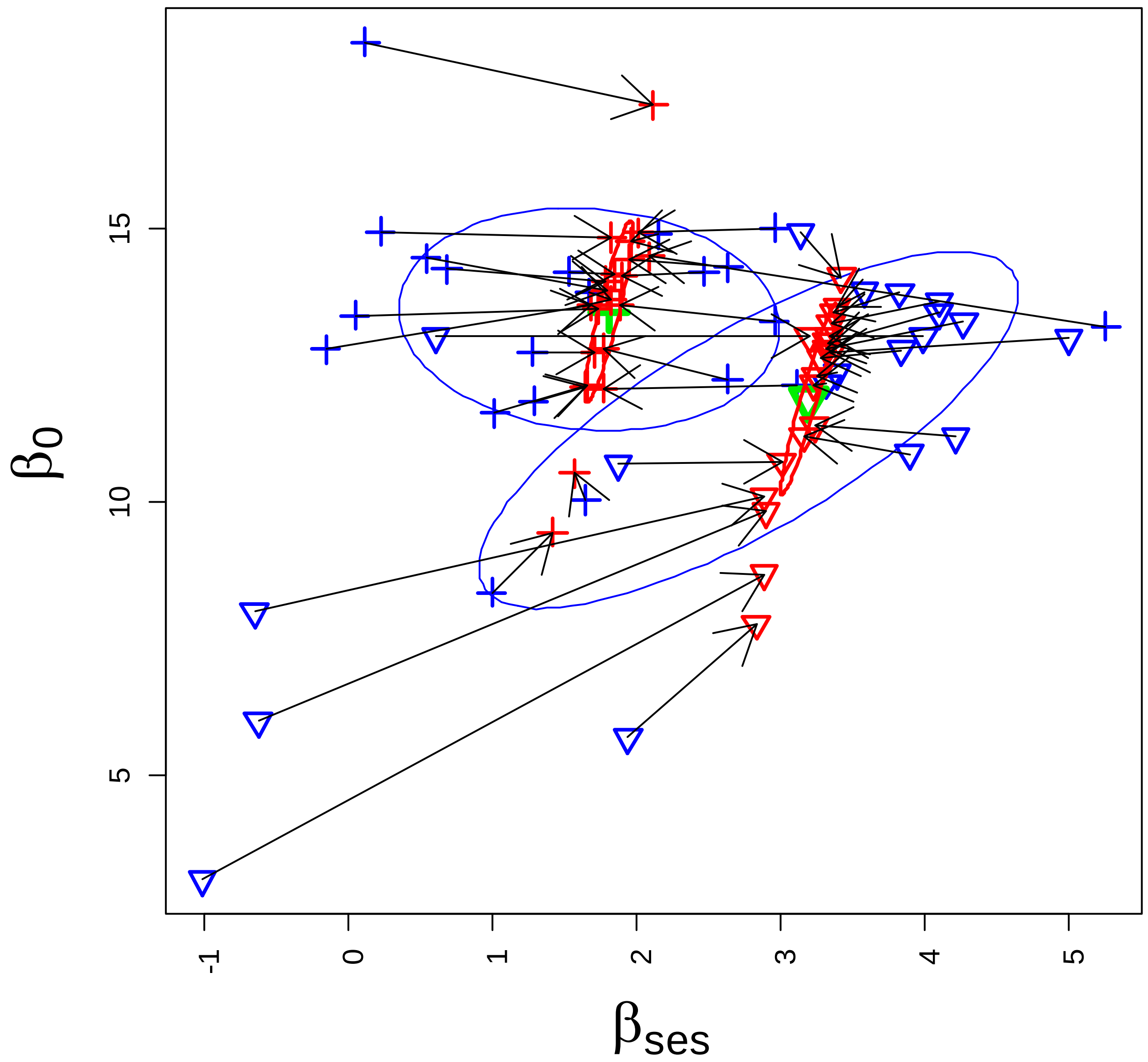
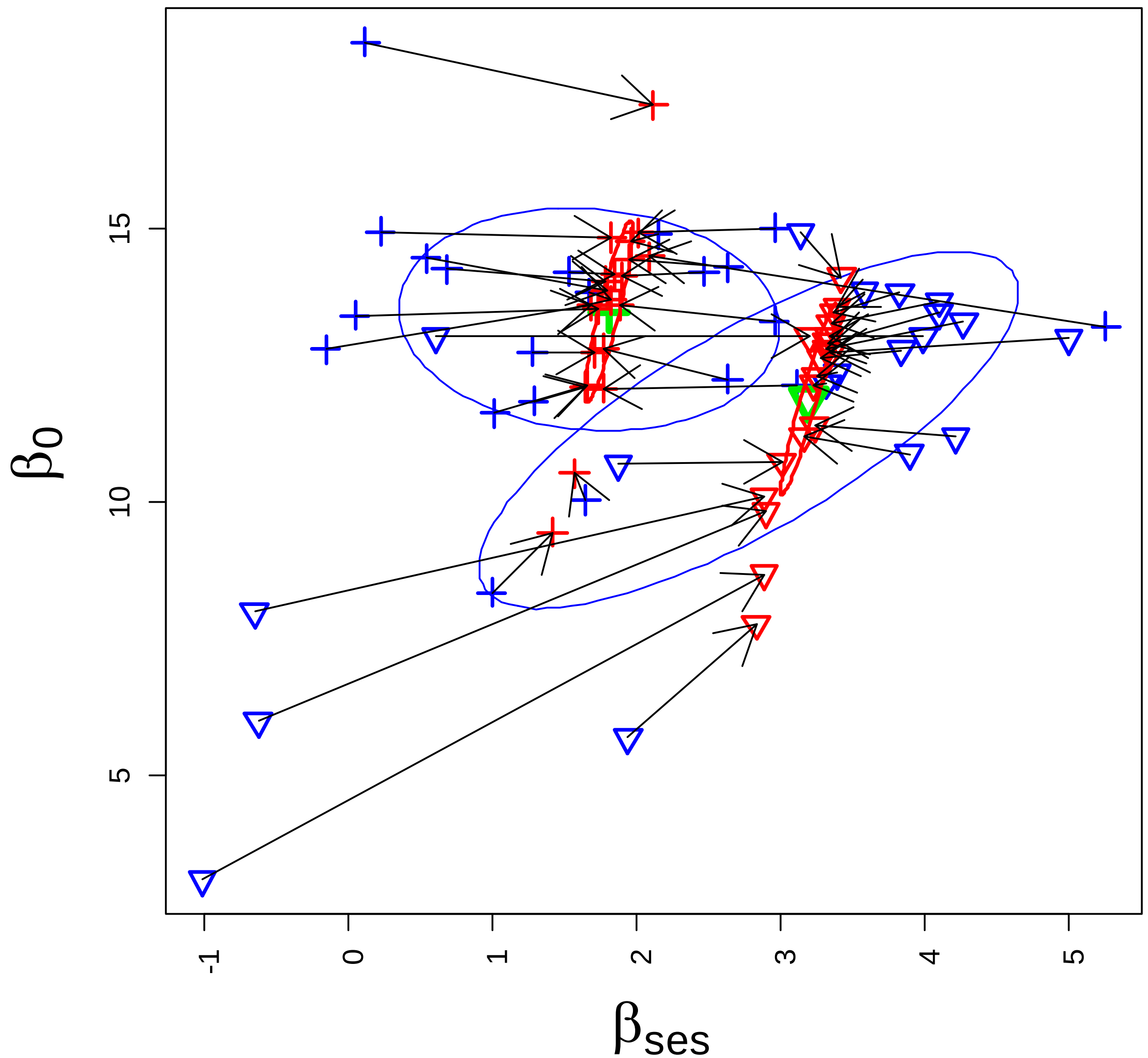


Figure 19: BLUEs in blue and BLUPs in red. Mean value in green.

The BLUPS show much less variability wrt beta.ses than the BLUES. This is because the BLUPS recognize that much of the variability in beta.ses is explicable by the large variablility in beta.hats.ses due to the samples. It does not interpret that variability as indicative of a variability in the slopes of the 'true' lines. The variability in intercepts, on the other hand, IS preserved in the BLUPS.



Why BLUPs are called 'shrinkage' estimators.
It is an inverse variance weighted combination of the BLUE and of the population estimate.



If we knew the population mean line γ, the between cluster variance, G and the the within-cluster variance, σ^2, the best predictor of β_j, the line for school *j*, combines γ and the BLUE, $\hat{\beta}_j$:

$$\tilde{\beta}_j = (\cdots)^{-1} \times$$

$$\left\{ G^{-1}\gamma + \left[\sigma^2 \left(X_j' X_j \right)^{-1} \right]^{-1} \hat{\beta}_j \right.$$

where

$$(\cdots) = G^{-1} + \left[\sigma^2 \left(X_j' X_j \right)^{-1} \right]^{-1}$$

Note that

$$\mathrm{Var}(\beta_j) = G$$

$$\mathrm{Var}(\hat{\beta}_j \mid \beta_j) = \sigma^2 \left(X_j' X_j \right)^{-1}$$

$$\mathrm{Var}(\hat{\beta}_j) = G + \sigma^2 \left(X_j' X_j \right)^{-1}$$

$$\mathrm{Var}(\tilde{\beta}_j) = \left\{ G^{-1} + \left[\sigma^2 \left(X_j' X_j \right)^{-1} \right]^{-1} \right\}^{-1}$$

Note: the BLUPS vary less than G and the BLUES vary more than G.

$$\mathrm{Var}(\tilde{\beta}_j) \le \mathrm{Var}(\beta_j) = G \le \mathrm{Var}(\hat{\beta}_j)$$

Note the estimated population lines for each sector are much closer to the centre of the BLUP ellipse than to the BLUE ellipse. Why?

The estimated population lines can be expressed as weighted combination of either the BLUES or of the BLUPS. However the weights VARY LESS when using the BLUPS than the BLUES.

How can both BLUEs and BLUPs be 'best'?

How can that be?
They are best for different things.

Recall the regression paradox: the best prediction of son's heights are best individually but they don't look like the distribution of son's heights. Best locally is not necessarily best at reproducing overall criteria.

BLUE is best for resampling from the same school over and over again. The BLUP is **best on average** for resampling from the population of schools and students.

If I'm a heartless bureaucrat and I want to be close on average I'll use the BLUP.

It's a bit like the basis of discrimination. If I don't have much information about you, I might use what I think I know about the group you seem to come from (here Catholic or Public) and I'll combine the two sources of information in an 'optimal' way.

If I really care to get a particular special school right, I would use the BLUE. The BLUP is justified only if the school is ***exchangeable*** with other schools in the sample and population conditional on the contextual variables.

Lab 1

Lab 1, which will probably take almost 2 days to complete, covers the implementation of concepts seen in these slides as well as many complementary concepts that seem to be better presented in the context of a actual analysis. Some of the ideas covered in Lab 1
Lab 1:

- First example: Between Sector gap in Math Achievement
  - Randomly selecting a subsample of clusters (schools)
  - Having a first look at multilevel data
  - Creating new Level 2 variables from Level 1 data
  - Seeing data in 3d
  - A second look at multilevel data: targeted to a model
  - Seeing fitted lines in beta space
  - Between and within cluster effects
  - Fitting a mixed model
  - Handling NAs (simplest considerations)

- Non-convergence
- First diagnostics: Hausman test
- Contextual variables to the rescue
- Interpretation of models with contextual effects
- Estimating the compositional (= between) effect
- Alternative equivalent parametrizations for the FE (fixed effects) model.
- Alternative non-equivalent parametrizations for the RE (random effects) model
- Diagnostics based on Level 1 residuals
- Diagnostics based on Level 2 residuals (REs)
- Influence diagnostics
- Plotting the fitted model: hand-made effect plots
- Linking the picture and the numbers
- Formulating and testing linear hypotheses
- Graphs to show confidence bounds for hypotheses

- Second example: Minority status and Math Achievement
  - Preliminary diagnostics using Level 1 OLS model

- OLS influence diagnostics
- Scaling Level 1 variables
- Fitting a mixed model
- Dealing with non-convergence
- Building the RE model with a forward stepwise approach
- Simulation to adjust p-values
- Test for contextual effects II
- Simplifying the model
- Using regular expression for easy tests of complex hypotheses
- Some Level 2 diagnostics
- Near-singularity: a pancake in 3D
- Visualizing the model: hand-made effect plots II
- The minority-majority gap
- Comparing different RE models
- More diagnostics
- Marginal and conditional models
- Refining the FE model
- Multilevel R Squared

- Visualizing the model to construct hypotheses